APRIL 2024

MongoDB Atlas AI Cookbook

6 AI recipes to kickstart innovation

MongoDB

Table of Contents

Introduction

"We see this as a driving force in three ways: organizations are looking to generative AI to build AI-powered applications and deliver powerful new experiences to customers, gain operational efficiencies, and modernize traditionally difficult-to-replatform legacy applications."

– DEV ITTYCHERIA, CEO, MONGODB

Artificial intelligence, encompassing advanced analytics, machine learning, deep learning, generative AI, across a wide range of use cases is expected to have a potential impact of up to **$25.6 Trillion** on the global economy1. AI, especially generative AI, has the potential to reshape societies - the way we work, the way we create, and the way we communicate. According to McKinsey, 40% of C-suite executives anticipate spending more on AI in the coming years. To tap into this rising tidal wave of opportunity, businesses must act. And they must act now. Speed is of the essence.

However, with great opportunity comes great risks, and considerable challenges. For organizations looking to lead the charge, making sure that they have the right strategy, skills, and data quality is the top priority. For the teams that are building AI applications, there is a host of challenges to overcome including the complexity of integrating across systems, data engineering, and security/ privacy concerns.2

In this paper, we will talk about the technical and architectural requirements for building AI-powered enterprise applications. We will provide 6 reference architectures and solutions to get you started on the path towards AI-powered applications and experiences with MongoDB Atlas and Vector Search.

Top 5 requirements for winning with AI

As Bill Gates said, "AI can be our friend." But the path to trusted systems is anything but simple. In the past year, we have seen AI applications sell a car for $1, fabricate refund policies, and several prompt injection techniques highlighting security loopholes. However, with the right technology and architecture, we can simplify the path from concept to business value. Here is our the top 5 requirements to keep in mind as you plan your next project.

Unified Data Management: AI requires vast amounts of structured and unstructured data and complex data transformations to ensure the data quality is solid. Integrated models into production will require real-time processing capabilities.

Enhanced Search Functionality: More intelligent, context-aware search mechanisms that can significantly improve the accuracy and relevance of search results are critical for applications like recommendations, semantic search, Retrieval augmented Generation(RAG), etc.

Event-driven Architecture and Serverless Compute: Allows for efficient data flow management and real-time response to data changes, which is pivotal for applications that require instantaneous adjustments based on user interactions or other inputs.

Scale, Security and Compliance: Ensuring scalability and availability through cloud-native choices, maintaining rigorous security protocols, and complying with regulatory standards are indispensable table stakes.

Seamless Integrations: Powering AI applications requires seamless integration with external services. Integrating with data providers, analytics providers, and endpoints for foundational models is vital for the development of and rapidly iterating AI applications.

In summary, building enterprise-grade AI applications is a multifaceted endeavor that blends advanced data management, real-time processing, machine learning, and intuitive search capabilities.

Simplify with a modern, multi-cloud database

Building modern applications need not be a treacherous game of Jenga, with weeks spent on integrating disparate systems, debugging fragile ETL pipelines, and trying to get to market as fast as possible while managing costs. MongoDB's modern database platform, Atlas, unifies operational, AI, and analytical data services to streamline building AI-powered applications.

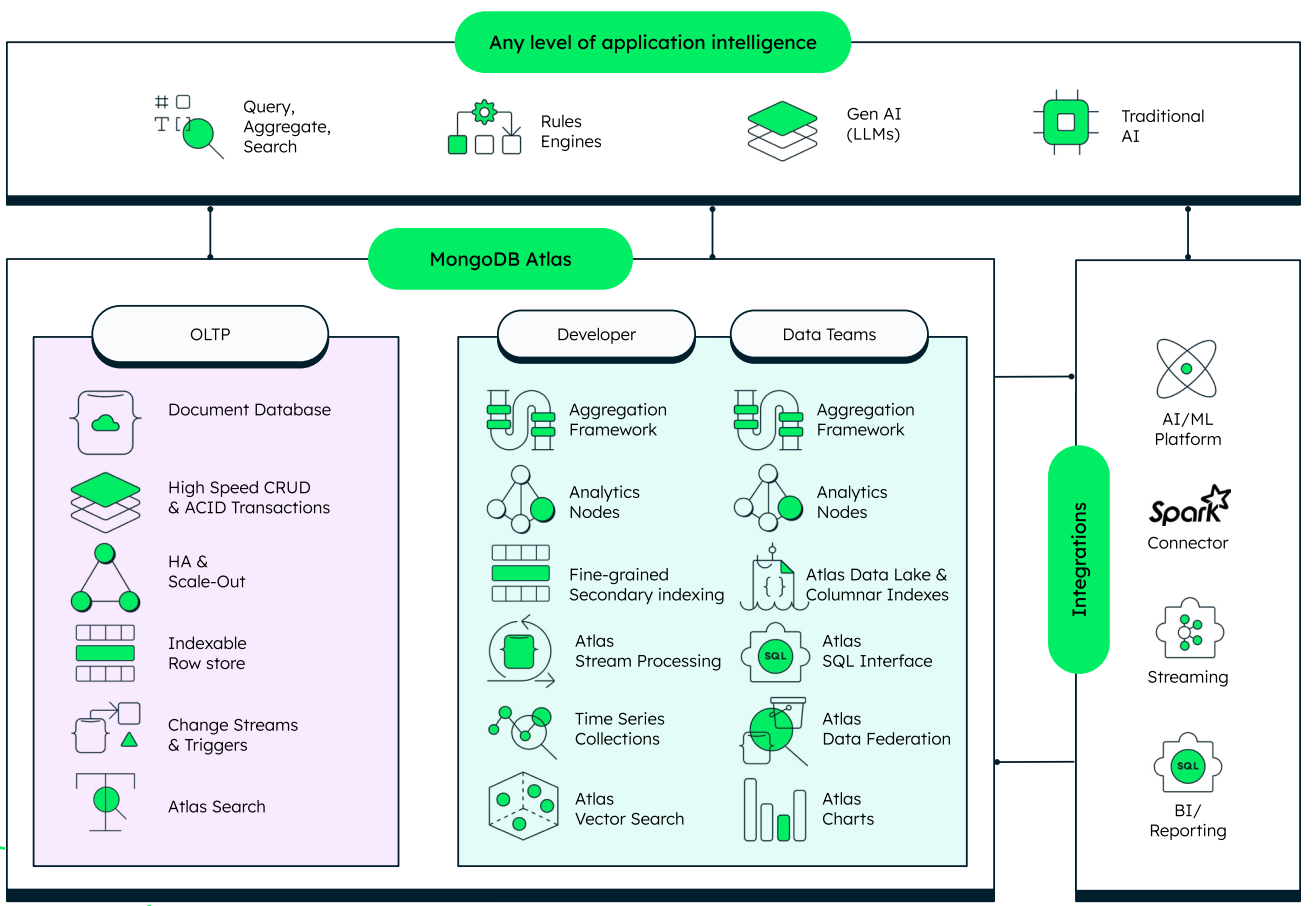

Figure 1. MongoDB Atlas unifies transactional, AI, and analytical processing with a unified, developer-native API and flexible document data model in a multi-cloud data platform.

Atlas puts powerful AI and analytics capabilities directly into the hands of developers in ways that fit their workflows, frameworks, and languages. With Atlas, they land data of any structure, index, query, search, and analyze it to provide model context and inference in any way the app needs, and then archive prompts and reasoning steps for long-term model memory. All while working with a unified API and flexible data model, without having to build their own data pipelines or duplicate data. Here's how MongoDB Atlas simplifies building AI applications

Unified Data Management: MongoDB Atlas, thanks to the document model and horizontal sharding capabilities, excels in managing vast volumes of structured and unstructured data. It supports real-time and complex data processing through capabilities such as Stream processing, Aggregation Framework, Atlas Triggers, Time Series collections, and Analytics Nodes.

Enhanced Search Functionality: Data stored in MongoDB Atlas can leverage both Atlas relevance Search and Atlas Vector Search to query the data in multiple ways, improving relevance and flexibility while building AI applications.

Event-Driven Architecture and Serverless Compute: MongoDB Atlas allows for efficient management of data flows and real-time responses to data changes through capabilities such as Atlas Stream Processing, Change Streams, Atlas Triggers and Time Series Collections.

Scale, Security, and Compliance: MongoDB Atlas supports true multi-cloud hybrid deployments, auto-scaling, and horizontal scaling through sharding to ensure your application is always available and can scale as required. MongoDB also provides client-side field-level encryption, encryption at rest, TLS/SSL encryption and queryable encryption to ensure data is always secure.

Seamless Integrations: MongoDB Atlas has connectors for LangChain, LlamaIndex, Apache Spark, Apache Kafka, Tableau, PowerBI, and more. Integrating Atlas with services such as Snowflake and Databricks is simple with over 58 third-party integrations.

In essence, MongoDB Atlas acts as the backbone for AI applications, addressing the intricate requirements of real-time data processing, scalability, security, and seamless integration. By leveraging MongoDB Atlas, developers and businesses can simplify the complex path to delivering AI applications that offer real business value.

Getting Started

We will focus on 6 common use cases to demonstrate and help you untangle the complexities around building AI applications:

- Gen-AI powered video summarization
- Automating digital underwriting with Machine Learning
- Real-time Card Fraud detection
- AI-enhanced claims adjustment for auto insurance
- Real-time dynamic pricing
- Personalized product search

To learn more about these and other solutions, access reference materials and code samples, visit the MongoDB Solutions Library.

Gen-AI powered video summarization

Utilizing Generative AI, this solution simplifies the process of distilling essential insights from videos, enabling users to quickly grasp critical content. It targets industries inundated with video data, offering a streamlined approach to video summarization and analysis. Access resources and learn more about this solution here.

Challenges

- **Volume and Complexity of Video Data:** Managing and processing extensive collections of video content to extract meaningful information poses significant computational and logistical challenges.
- **Accuracy and Contextual Relevance of Summaries:** Ensuring the generated summaries accurately reflect the video content and context, avoiding misinterpretation or loss of crucial information.
- **Integration of OCR and AI for Enhanced Analysis:** Combining Optical Character Recognition (OCR) and AI to analyze video frames for text, further complicating real-time analysis and accuracy.

Benefits

- **Efficient Information Consumption:** By condensing videos into summaries, users can consume information more efficiently, saving time and enhancing productivity.
- **Improved Accessibility of Content:** Makes video content more accessible by providing text-based summaries, aiding in content discovery and comprehension.
- **Enhanced Learning and Decision-making:** Facilitates rapid knowledge acquisition and informed decision-making by highlighting key video segments and insights.

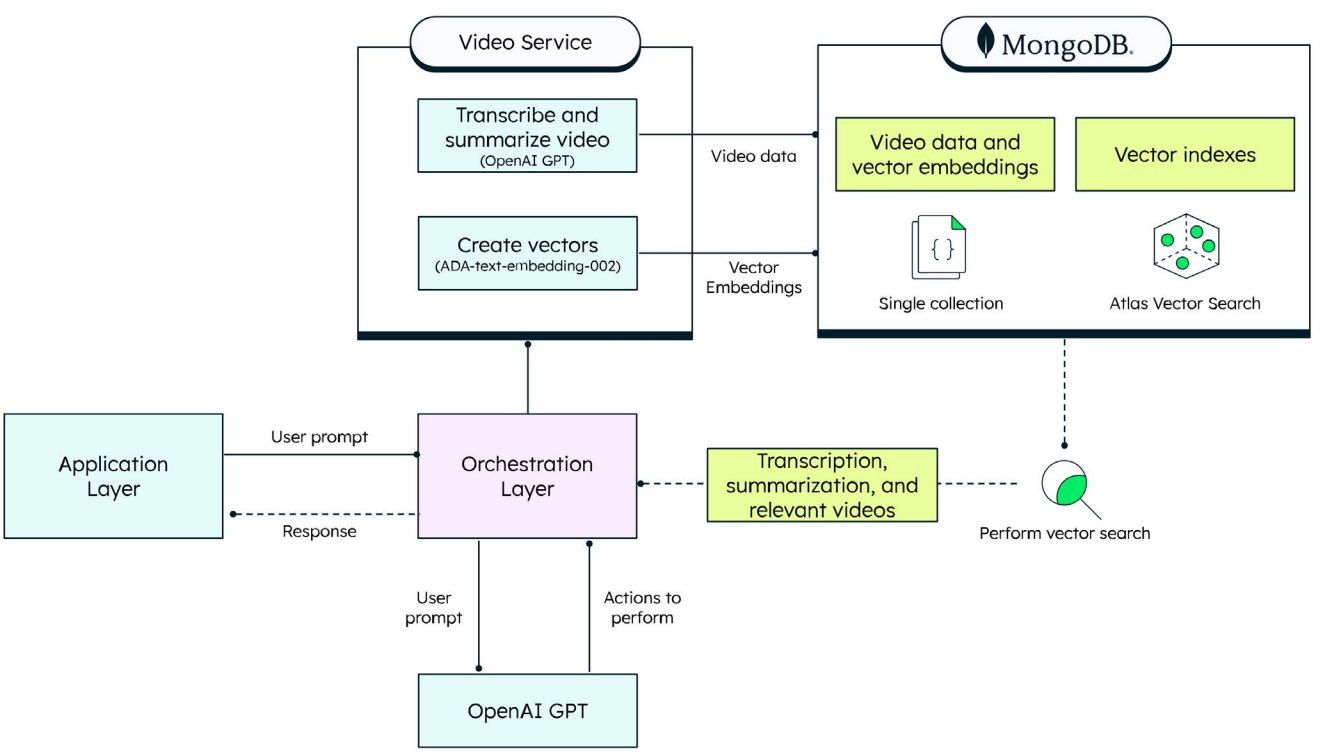

Figure 2. Reference architecture for a video summarization solution built on MongoDB Atlas

Products

- **Atlas Database:** Utilized for storing video metadata, transcripts, and summaries in a scalable and flexible manner.
- **Atlas Vector Search:** Enables semantic search across video summaries, improving content discoverability.

Partners

- **Langchain:** Collaboration enhances the solution's capabilities in language processing and AI-driven summarization.

Data Model

This solution employs a comprehensive data model that includes direct YouTube video links, detailed metadata, full transcripts, AI-generated summaries, and code analysis in a structured JSON format. This model facilitates effective data storage, retrieval, and semantic analysis, supporting advanced search functionalities.

Automating Digital Underwriting with Machine Learning

This solution automates the underwriting process in the insurance and financial sectors using machine learning, enabling real-time decision-making and efficiency. It leverages MongoDB Atlas and Databricks to process and analyze data for instant underwriting outcomes. Access resources and learn more about this solution here.

Challenges

Complexity of Real-time Data Processing: Integrating and analyzing real-time data from various sources to make instant underwriting decisions.

Accuracy in Risk Assessment: Ensuring the machine learning models accurately assess risk based on current and historical data, minimizing errors.

Scalability of the Solution: Adapting to the growing volume of underwriting requests without compromising performance or accuracy.

Benefits

Increased Efficiency: Significantly reduces the time required for underwriting processes, enhancing operational efficiency.

Improved Risk Management: Utilizes advanced analytics to provide more accurate risk assessments, improving portfolio quality.

Dynamic Adaptability: Enables the underwriting process to dynamically adjust to new data and trends, maintaining competitiveness and relevance.

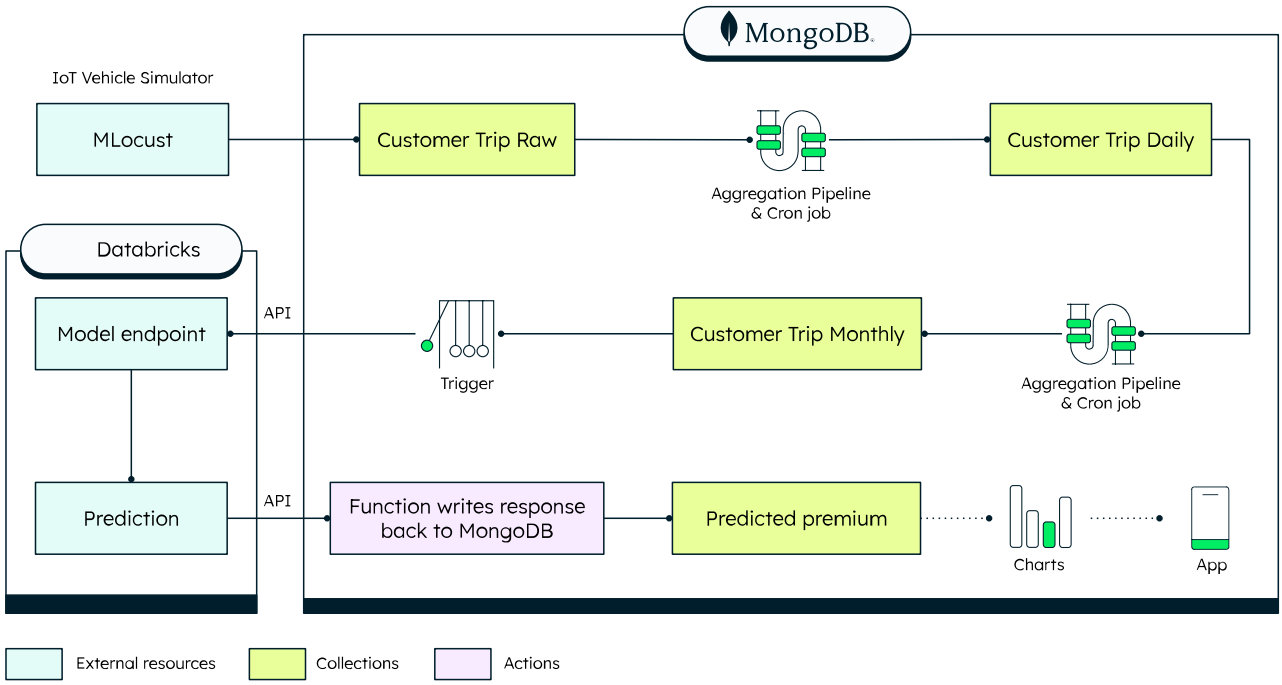

Figure 3. Reference architecture for an automated underwriting solution built on MongoDB Atlas

Products

- **Atlas Database:** Manages diverse datasets involved in the underwriting process, ensuring data integrity and accessibility.
- **Spark Connector:** Facilitates the integration of MongoDB data with Databricks for advanced analytics and machine learning model execution.

Partners

- **Databricks:** Provides the computational framework for running complex machine learning algorithms on large datasets, enhancing the solution's analytical capabilities.

Data Model

This solution focuses on time-series data collection and aggregation, supporting the underwriting process with a structured approach to capturing, storing, and analyzing transactional data. This model is crucial for understanding risk patterns and making informed decisions.

Real-time Card Fraud Detection

This solution targets real-time fraud detection in financial transactions using AI and ML, integrating MongoDB Atlas with Databricks to analyze transactional data instantaneously for suspicious activities. Access resources and learn more about this solution here.

Challenges

- **Real-time Detection and Analysis:** Rapidly identifying and analyzing potentially fraudulent transactions in real-time to prevent financial losses.
- **Integration of Diverse Data Sources:** Consolidating data from various sources to create a comprehensive view of customer transactions and behaviors.
- **Maintaining System Accuracy and Efficiency:** Ensuring the fraud detection models remain accurate over time, minimizing false positives and negatives, while handling large volumes of transactions.

Benefits

- **Enhanced Fraud Prevention:** Immediate detection and response to fraudulent activities, significantly reducing the risk of financial loss.
- **Improved Customer Trust and Security:** Provides customers with a secure transaction environment, bolstering trust and satisfaction.
- **Adaptive Fraud Detection Models:** Utilizes continuous learning to adapt fraud detection models to new patterns and techniques, staying ahead of fraudsters.

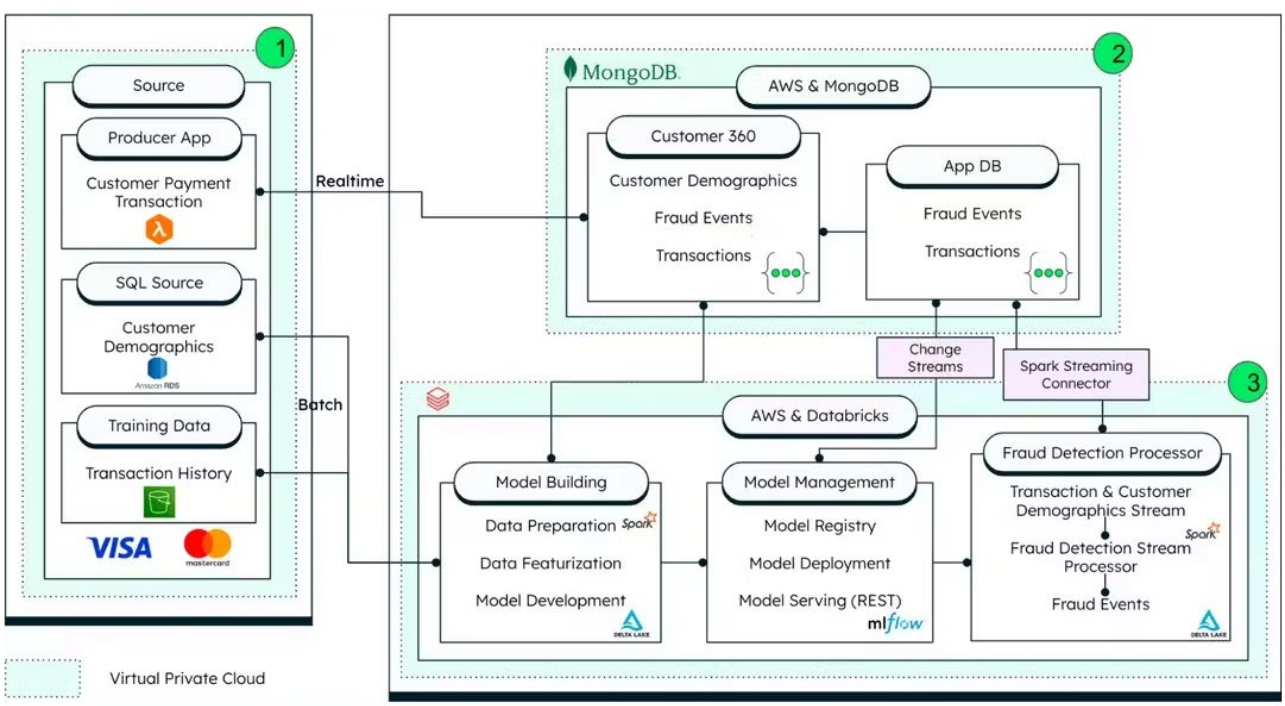

Figure 4. Reference architecture for a real-time fraud detection solution built on MongoDB Atlas

Products

- Atlas Clusters: Offers a scalable and flexible database solution for storing transactional data, facilitating real-time fraud analysis.
- Change Streams and Atlas Triggers: Enable real-time data processing and instant alerting for suspicious transactions.
- Spark Streaming Connector: Links MongoDB data with Databricks for dynamic data analysis and ML model application.

Partners

- Databricks: Powers the AI/ML analytics platform, providing the computational horsepower needed to analyze transactions and detect fraud patterns efficiently.

Data Model

This solution employs a transaction-centric data model that captures detailed information about each transaction, merchant, and customer involved. This model supports complex analyses to identify fraudulent activities based on transaction patterns and behaviors.

AI-enhanced claims adjustment for auto insurance

Revolutionizes auto insurance claim adjustments through AI-powered image analysis and vector search, streamlining the process by comparing accident photos for precise damage assessments. Access resources and learn more about this solution here.

Challenges

- **Accuracy of Damage Assessment:** Ensuring the AI accurately interprets and assesses damage from images, aligning with actual repair needs.
- **Processing High Volumes of Claims:** Efficiently handling many claims with varying degrees of complexity and damage.
- **Integrating AI with Existing Systems:** Seamlessly incorporating AI technologies into existing claim processing systems without disrupting workflows.

Benefits

- **Faster Claim Processing:** Accelerates the claims adjustment process by automating damage assessments, improving operational efficiency.
- **Increased Accuracy and Consistency:** Provides more consistent and accurate damage evaluations, leading to fairer claim settlements.
- **Enhanced Customer Satisfaction:** Reduces claim processing time, improving the overall customer experience during stressful post-accident periods.

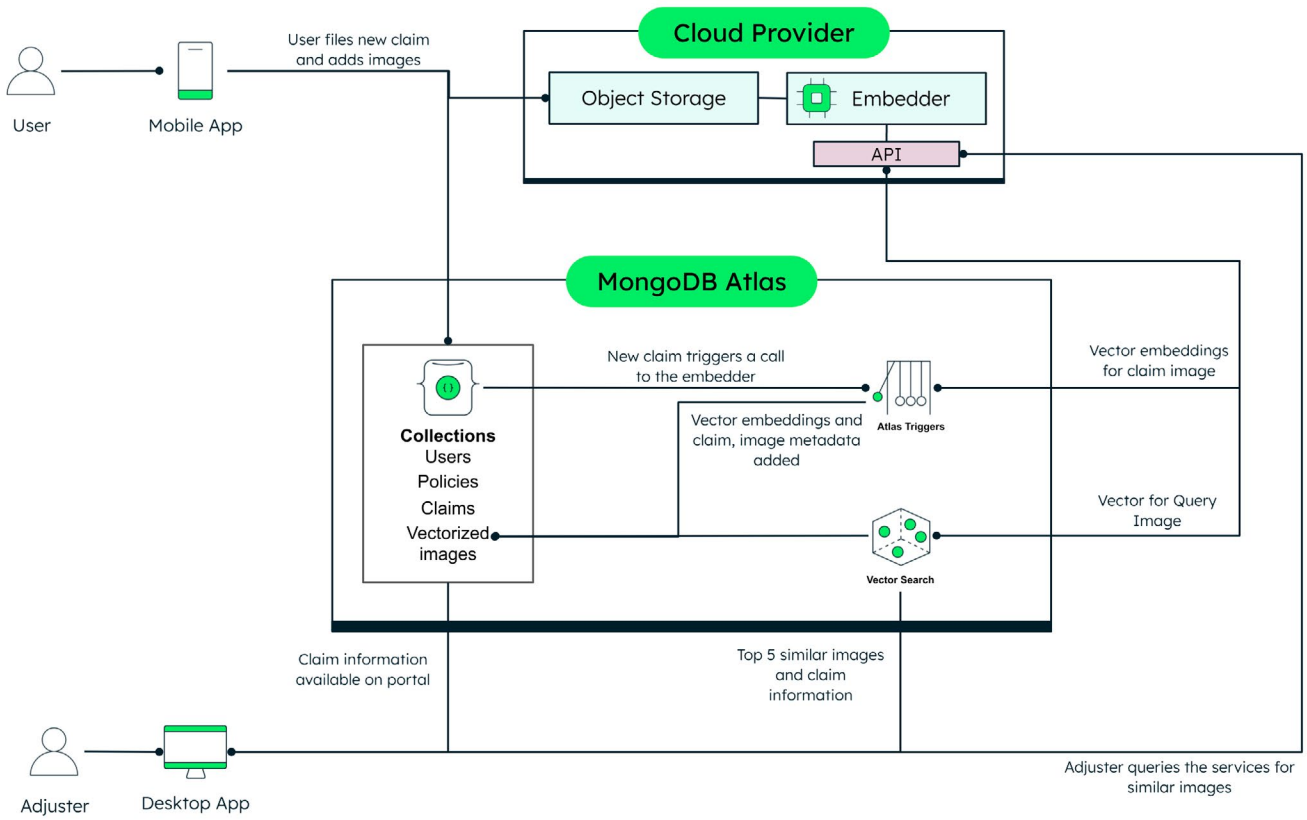

Figure 5. Reference architecture for a claims adjustment solution built on MongoDB Atlas

Products

- **Atlas Database and Atlas Vector Search:** Facilitate the storage and semantic search of accident images, enhancing the ability to find and compare relevant past claims.
- **Atlas App Services:** Integrates AI and ML models into the claim processing workflow, automating damage assessments.

Partners

- **PyTorch:** Used for developing and training the AI models responsible for analyzing and interpreting accident images.

Data Model

This solution incorporates a detailed representation of claims, including references to accident photos stored as AWS S3 links and metadata describing the accident and damage. This model supports efficient retrieval and analysis of claim data for AI-powered processing.

Real-time Dynamic Pricing

Empowers retail businesses to implement dynamic pricing strategies using real-time analytics, adjusting prices based on inventory, market trends, and consumer behavior by leveraging MongoDB Atlas and Databricks. Access resources and learn more about this solution here.

Challenges

- **Integration and Analysis of Diverse Data Sources:** Harmonizing data from various inputs such as inventory levels, sales performance, and consumer trends to inform pricing strategies.
- **Scalability of Real-time Analytics:** Scaling analytics capabilities to process and respond to data changes across extensive product catalogs instantly.
- **Accuracy in Dynamic Pricing Decisions:** Ensuring pricing adjustments are timely, market-competitive, and align with business objectives without alienating customers.

Benefits

- **Optimized Revenue Management:** Dynamically adjusts prices to maximize sales and profits while remaining competitive.
- **Enhanced Market Responsiveness:** Quickly responds to market changes, inventory levels, and consumer demand, maintaining an edge in fast-paced retail environments.
- **Improved Customer Experience:** Offers more personalized pricing and promotions, increasing customer satisfaction and loyalty.

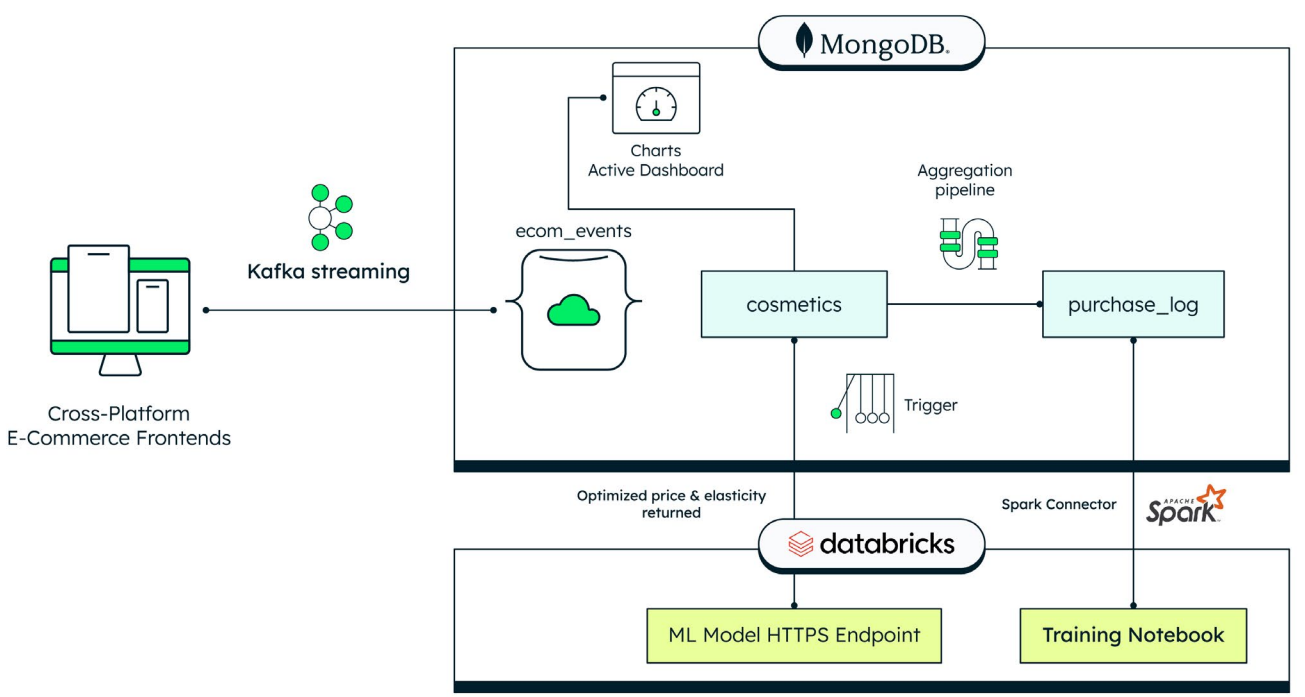

Figure 6. Reference architecture for a dynamic pricing solution built on MongoDB Atlas

Products

- **Aggregation Pipeline and Atlas App Services:** Utilized for real-time data aggregation and processing, enabling dynamic analysis of pricing factors.
- **Triggers and Functions:** Automate the pricing adjustment process based on real-time data insights, ensuring prices are always optimized and relevant.

Partners

- **Databricks:** Facilitates complex analytics and machine learning model deployment, providing the intelligence behind dynamic pricing strategies.

Data Model

Designed around event-driven architecture, capturing customer interactions (views, cart additions, purchases) in real-time. This model supports the nuanced analysis of consumer behavior and product performance to inform dynamic pricing adjustments.

Personalized Product Search

Enhances e-commerce platforms with AI-powered search, offering personalized and accurate product recommendations using machine learning and MongoDB Atlas's Vector Search capabilities. Access resources and learn more about this solution here.

Challenges

- **Understanding Complex Consumer Queries:** Interpreting the semantic context of search queries to deliver relevant results.
- **Maintaining Real-time Responsiveness:** Providing instant search results in dynamic retail environments with constantly changing inventories.
- **Seamless Integration of AI Technologies:** Incorporating AI-driven search functionalities into existing e-commerce platforms without disrupting user experience.

Benefits

- **Personalized Shopping Experiences:** Uses AI to understand and predict customer preferences, offering tailored product recommendations.
- **Increased Search Relevance and Accuracy:** Improves the precision of search results, enhancing the likelihood of customer satisfaction and purchase.
- **Competitive Advantage in E-commerce:** Sets platforms apart by providing a superior search experience, driving sales, and customer loyalty.

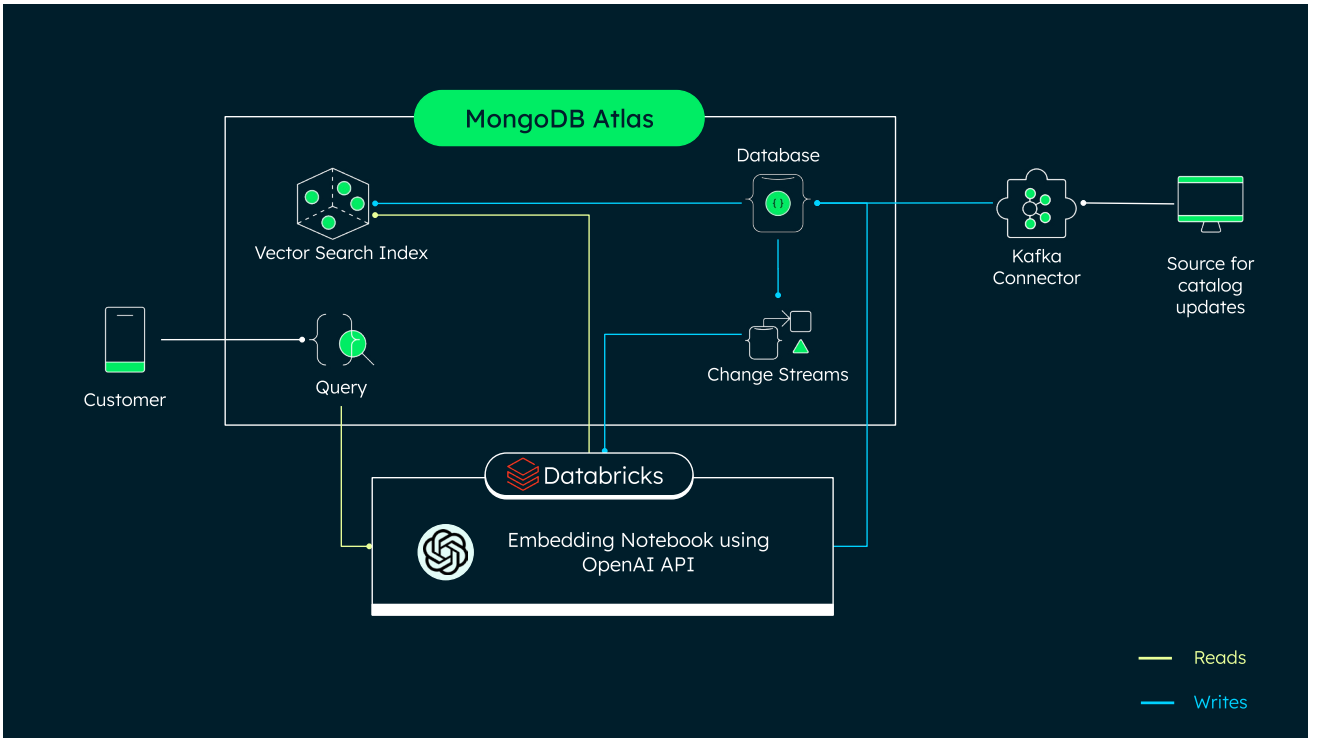

Figure 7. Reference architecture for a personalized product search solution built on MongoDB Atlas

Products

- **Atlas Search and Atlas Vector Search:** Power advanced, AI-enhanced search capabilities, allowing for nuanced query understanding and accurate product matching.
- **MongoDB Connector for Spark:** Enables efficient data processing and analysis, enriching search algorithms with deep consumer insights.

Partners

- **Databricks:** Provides the analytical and machine learning backbone, processing vast amounts of data to refine search algorithms continually.

Data Model

This solution utilizes a polymorphic pattern to accommodate a wide range of product types in a single collection, facilitating efficient query processing. This flexible schema supports diverse product attributes while ensuring quick access to relevant search data.

Conclusion

In conclusion, the advent of AI heralds a transformative era for businesses and society. This whitepaper has underscored the immense potential and challenges of AI, highlighting MongoDB Atlas as a key enabler for building powerful, AI-powered applications. MongoDB Atlas simplifies the development process, providing a robust platform for real-time, scalable, and integrated solutions.

Embracing a modern, multi-cloud database platform allows businesses to efficiently manage complex data and leverage AI to drive innovation, security, and compliance. As AI becomes increasingly essential, the ability to quickly adapt and implement these technologies will set industry leaders apart.

The message is clear: for organizations to thrive in the AI-driven landscape, investment in the right technology and platforms is crucial. By doing so, businesses can unlock new opportunities, foster growth, and lead in the digital economy. In the journey towards AI transformation, the time to act is now.

Resources

For more information, please visit mongodb.com or contact us at sales@mongodb.com.

Reference Solutions (mongodb.com/solutions/solutions-library)

Case Studies (mongodb.com/customers)

Presentations (mongodb.com/presentations)

Free Online Training (university.mongodb.com)

Webinars and Events (mongodb.com/events)

Documentation (docs.mongodb.com)

MongoDB Atlas database as a service for MongoDB (mongodb.com/cloud)

MongoDB Enterprise Download (mongodb.com/download)

MongoDB Realm (mongodb.com/realm)

Legal Notice

This document includes certain "forward-looking statements" within the meaning of Section 27A of the Securities Act of 1933, as amended, or the Securities Act, and Section 21E of the Securities Exchange Act of 1934, as amended, including statements concerning our financial guidance for the first fiscal quarter and full year fiscal 2021; the anticipated impact of the coronavirus disease (COVID-19) outbreak on our future results of operations, our future growth and the potential of MongoDB Atlas; and our ability to transform the global database industry and to capitalize on our market opportunity. These forward-looking statements include, but are not limited to, plans, objectives, expectations and intentions and other statements contained in this press release that are not historical facts and statements identified by words such as "anticipate," "believe," "continue," "could," "estimate," "expect," "intend," "may," "plan," "project," "will," "would" or the negative or plural of these words or similar expressions or variations. These forward-looking statements reflect our current views about our plans, intentions, expectations, strategies and prospects, which are based on the information currently available to us and on assumptions we have made. Although we believe that our plans, intentions, expectations, strategies and prospects as reflected in or suggested by those forward-looking statements are reasonable, we can give no assurance that the plans, intentions, expectations or strategies will be attained or achieved. Furthermore, actual results may differ materially from those described in the forward-looking statements and are subject to a variety of assumptions, uncertainties, risks and factors that are beyond our control including, without limitation: our limited operating history; our history of losses; failure of our database platform to satisfy customer demands; the effects of increased competition; our investments in new products and our ability to introduce new features, services or enhancements; our ability to effectively expand our sales and marketing organization; our ability to continue to build and maintain credibility with the developer community; our ability to add new customers or increase sales to our existing customers; our ability to maintain, protect, enforce and enhance our intellectual property; the growth and expansion of the market for database products and our ability to penetrate that market; our ability to integrate acquired businesses and technologies successfully or achieve the expected benefits of such acquisitions; our ability to maintain the security of our software and adequately address privacy concerns; our ability to manage our growth effectively and successfully recruit and retain additional highly-qualified personnel; the price volatility of our common stock; the financial impacts of the coronavirus disease (COVID-19) outbreak on our customers, our potential customers, the global financial markets and our business and future results of operations; the impact that the precautions we have taken in our business relative to the coronavirus disease (COVID-19) outbreak may have on our business and those risks detailed from time-to-time under the caption "Risk Factors" and elsewhere in our Securities and Exchange Commission ("SEC") filings and reports, including our Quarterly Report on Form 10-Q filed on December 10, 2019, as well as future filings and reports by us. Except as required by law, we undertake no duty or obligation to update any forward-looking statements contained in this release as a result of new information, future events, changes in expectations or otherwise.