Enterprise Strategy Group by TechTarget

SHOWCASE

AUGUST 2024

MongoDB Atlas and Amazon Bedrock Make It Easy and Safe to Bring Enterprise Data to Generative AI

Stephen Catanzano, Senior Analyst

Abstract: MongoDB Atlas and Amazon Bedrock make it incredibly easy and safe to simplify bringing enterprise data to generative AI using retrieval-augmented generation (RAG). RAG enables organizations to build hyper-personalized experiences uniquely tailored to business needs using enterprise data. MongoDB Atlas securely unifies real-time, operational, unstructured, and vectorized data, removing the friction of integrating the essential data components required to give large language models (LLMs) the context they need in a RAG workflow.

At the same time, Amazon Bedrock offers a fully managed, end-to-end RAG workflow feature alongside a range of foundational models (FM) and tools for creating, training, and deploying generative AI solutions. Here, we explore how they work together to accelerate time to value for enterprises looking to build generative AI experiences that take advantage of their proprietary data.

Enterprise Challenges Are Driving the Need for the Right Partners

Data leaders in organizations are inundated with new use cases from every line of business seeking to apply generative AI to existing data to increase productivity, enhance the quality of data-driven decision-making, and create more personalized experiences for customers. However, transitioning from experimental to enterprise-ready generative AI has proven challenging for many organizations. Nearly 40% of organizations reported that they struggle or anticipate struggling with validating and evaluating generated results, employee hesitancy to trust recommendations, and ethical concerns about generated content when integrating generative AI with AI infrastructure.[1]

Market Insight

Nearly 40% of organizations cited experiencing or anticipating difficulty validating and evaluating generated results, employee hesitancy to trust recommendations, and ethical considerations and biases in generated content when deploying generative AI in production.

Generative AI systems, while promising, still need the right data and data management practices to achieve the response reliability that mission-critical applications require. Hallucinations, where foundational models (FM) generate false information, hinder widespread enterprise adoption. Although these models are trained on large data sets and produce grammatically correct responses, they cannot deliver accurate business information without enterprise knowledge. Additionally, they may generate technically correct responses that lack domain-specific knowledge, making them less relevant and less specific to the use case. For example, an FM assisting in legal research should be built on existing regulations and include the most recent cases to provide the most up-to-date information.

[1] Source: Enterprise Strategy Group Research Report, *Navigating the Evolving AI Infrastructure Landscape*, September 2023.

Furthermore, most FMs do not include mechanisms to provide information about the data sources behind their responses. Without source details, understanding how responses are generated is impossible. Poor explainability will hinder the full adoption of these models and may become a dealbreaker for organizations in industries with stringent reporting requirements, such as finance or law. Achieving true explainability requires citing sources to substantiate information, explaining the reasoning behind information retrieval, and understanding why a FM selects certain information to include.

These challenges emphasize the need for organizations to build generative AI solutions using their enterprise data and trusted technology partners like MongoDB and Amazon Web Services (AWS) to maximize the value of the solutions being created.

How Do I Use My Enterprise Data With Generative AI?

One of the biggest challenges when working with generative AI is avoiding hallucinations or erroneous results returned by the FM being used. FMs are trained on public information that quickly becomes outdated and cannot take advantage of enterprises' proprietary information.

One way to tackle hallucinating FMs is to supplement a query with the organization's own data using RAG. In a RAG workflow, the FM will retrieve specific data—for instance, a customer's previous purchase history—from a designated database that acts as a "source of truth" to augment the results returned by the FM.

For a generative AI to search for, locate, and augment its responses, the relevant data needs to be turned into a vector and stored in a vector database. A vector is an ordered list or sequence of numbers used to represent data, including unstructured data like text, images, audio, or video. Vector embeddings are a way to represent such words and other data points into numbers, where each data point is represented by a vector in high-dimensional space. A vector database stores, retrieves, and searches for vectors.

An organization's data is the main differentiator for building unique generative AI solutions. Organizations can choose to build their own RAG and vectorization workflows or work with companies like MongoDB and AWS to simplify the process.

5 Key Considerations for Building a Generative AI Solution Using RAG

Organizations should consider five important areas when looking to use enterprise data with RAG to build generative AI solutions.

- **Trusted Data Sources.** Organizations need to identify trusted, accurate, and quality data for each generative AI use case. For example, if building a tool for the human resources department, the data might include all relevant documents, employee data, and other employee-related information.
- **Vector Database Utilization.** A vector database is crucial for generative AI applications since it stores vector embeddings that enable similarity search, providing relevant context for the FMs. similarity and relevance when used with prompts. MongoDB Atlas is a vector database with industry-leading vector search capabilities. MongoDB Atlas unifies vector embeddings with live application data in a single, fully managed database and Atlas Vector Search can index, retrieve, and supply applications that use generative AI with the vectorized data they need.

Market Insight

79% of organizations said they must use AI in business- and mission-critical processes to compete better.[2]

[2] Source: Enterprise Strategy Group Complete Survey Results, *The State of DataOps: Unleashing the Power of Data*, December 2023.

- **RAG Workflow Optimization.** It's important that organizations continuously refine the RAG workflow to improve the accuracy and efficiency of the AI model's responses. Organizations may need to experiment with different query formulations, vector embeddings, and retrieval strategies. Amazon Bedrock offers a fully managed RAG workflow at the click of a button.
- **Data Privacy and Security.** Organizations need to implement robust measures to protect sensitive information within the vector database, as compliance with data privacy regulations (e.g., GDPR, CCPA) is crucial.
- **Model Evaluation and Refinement.** Regular assessment of the RAG workflow's performance and the AI model to identify areas for improvement is necessary. Additionally, organizations should iterate on the model and data to enhance results over time. Amazon Bedrock offers easy access to many FMs for testing and deployment. Each FM has different attributes to consider, depending on the generative AI use case.

MongoDB and Amazon Bedrock Simplify the Generative AI Workflow

MongoDB Atlas and Amazon Bedrock offer a fast, easy and secure way to build RAG based generative AI applications with enterprise data. . Amazon Bedrock is a fully managed service from AWS that helps developers build generative AI applications using FMs. The service brings together models from leading AI companies and is serverless, allowing for the evaluation of FMs and seamless integration with other AWS services.

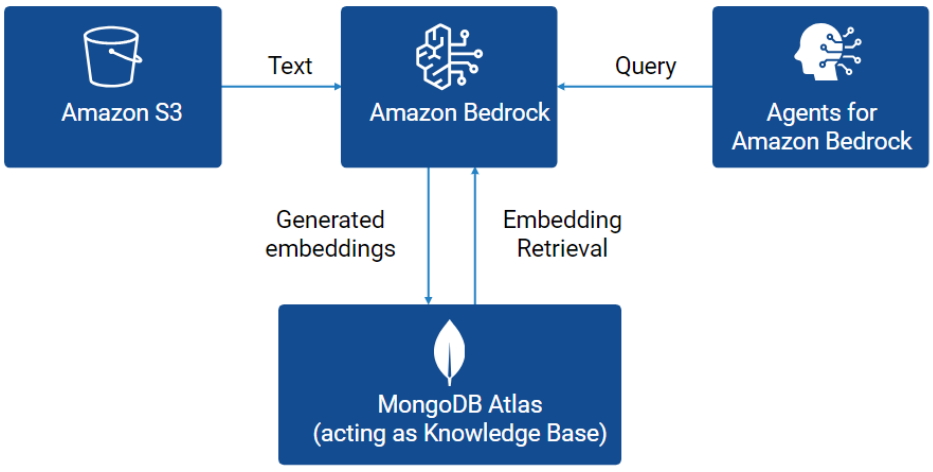
One of Amazon Bedrock's key features is its fully managed, end-to-end RAG workflow, known as Knowledge Bases. Users can easily select a Knowledge Base within the Amazon Bedrock interface, such as MongoDB, to store vector embeddings. Amazon Bedrock takes care of data ingestion, retrieval, and prompt augmentation without the need for custom integrations or data flow management, simplifying the process of implementing RAG.

With MongoDB Atlas and Amazon Bedrock, developers can rapidly deploy and scale generative AI apps grounded in up-to-date and accurate data.

How MongoDB and Amazon Web Services Work Together

Within Amazon Bedrock, developers can now "click to add" MongoDB Atlas as a Knowledge Base for their vector data store to power RAG. After selecting an embedding model and a generative model, Amazon Bedrock agents then orchestrate and use these models during their interaction with the knowledge base (see the workflow shown in Figure 1).

Figure 1. MongoDB Atlas and Amazon Bedrock



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Amazon Bedrock reads an organization's text data from an Amazon Simple Storage Service (S3) bucket, chunks it, and then uses the customer-chosen embedding model to create the vector embeddings. Amazon Bedrock stores these text chunks, embeddings, and related metadata in MongoDB Atlas' vector database. An Atlas vector search index is also created as part of the setup for querying the vector embeddings.

As an Amazon Bedrock Knowledge Base and part of the Amazon Bedrock ecosystem, MongoDB Atlas solves one of the biggest challenges in deploying generative AI solutions using RAG. At the click of a button, it enables Amazon Bedrock users to store, search, and find the most relevant and up-to-date information, and augment their generative AI model outputs with that data.

Why Choose MongoDB Atlas as a Amazon Bedrock Knowledge Base?

MongoDB Atlas combines operational data, vector data, and metadata in a single platform, making it an ideal knowledge base for Amazon Bedrock users who want to augment their generative AI experiences while also simplifying their generative AI stack. In addition, MongoDB Atlas enables developers to set up dedicated infrastructure for search and vector search, optimizing compute resources to scale search and database workloads independently.

Case Study: MongoDB Atlas and Amazon Bedrock Deliver Results With Generative AI Solutions

A global healthcare company accelerated its time to value, reducing the time it takes to produce a clinical study report (CSR) from 12 weeks to 10 minutes using MongoDB Atlas and Amazon Bedrock. The company, which employs 64,000 people in 80 countries, devises treatments that benefit millions of people living with diabetes, obesity, and rare blood and endocrine diseases. The company produces 50% of the world's insulin.

In the healthcare industry, a CSR plays a pivotal role in developing any new medication. It is a comprehensive document that captures the methodology, execution, results, and analyses of a clinical trial. The report's primary purpose is to provide a detailed account of the medical trial, ensuring that regulatory authorities, healthcare professionals, and other stakeholders can assess the efficacy and safety of a new pharmaceutical product. Typically, a CSR takes around 12 weeks to compile and can be hundreds of pages. Each day of delay means patients don't get the treatments they need, and the company cannot start to recover its research and development costs.

Using generative AI, the healthcare company's digitalization team saw the opportunity to drive significant efficiencies in the production of CSRs, leading to the development of a new system.

A New System: Built on Amazon Bedrock, LangChain, and MongoDB Atlas Vector Search

Initiating the project in mid-2023, the digitalization team started to experiment with dynamically compiling the CSR by leveraging RAG to prompt state-of-the-art LLMs using both statistical outputs from the clinical trials and vector embeddings of report templates. Within a few weeks, the experiments proved successful. The new system, which uses the Anthropic's Claude 3 and Amazon Titan FMs hosted by Amazon Bedrock, was deemed ready for widespread usage, as it produced CSRs faster—i.e., the production time of CSRs dropped from 12 weeks to just 10 minutes—and more accurately, requiring fewer resources than the previous manual methods. The team can switch between models with the LangChain development and orchestration framework as needed. LangChain is a framework designed for developing applications powered by FM's. It enables the creation of applications that can understand and generate human language, making it ideal for tasks such as natural language processing, chatbots, and AI-driven content generation.

Conclusion

Generative AI has emerged as a strategic imperative for organizations worldwide, yet there are significant challenges, particularly around data trust and the accuracy of responses. Overcoming obstacles such as hallucinations, ambiguity, and lack of explainability is crucial for building game-changing solutions. RAG offers a promising solution by enhancing FMs with external data sources without requiring model retraining and by adding enterprise data context to the solution.

A robust data collection and storage solution is essential to fully realize the potential of generative AI. MongoDB Atlas, with native vector search capabilities, stands out as the optimal data structure for this purpose and accelerates time to value for generative AI solutions built with Amazon Bedrock as fully managed RAG workflow. The combination of Amazon Bedrock and MongoDB Atlas empowers any organization to build and deploy generative AI quickly and easily for its unique use cases.

Ultimately, the choice of database technology significantly impacts the capabilities of enterprise generative AI systems. By selecting MongoDB Atlas and Amazon Bedrock for RAG, organizations can unlock the full potential of generative AI, driving innovation across a wide range of industries and use cases.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

contact@esg-global.com
www.esg-global.com