Como incorporar IA generativa e pesquisa avançada aos seus aplicativos com MongoDB

Construindo aplicativos baseados em IA

Dezembro de 2023

EUA 866-237-8815 • Exterior +1-650-440-4474 • info@mongodb.com.

Índice

Introdução

A introdução de uma nova tecnologia nunca chamou tão rapidamente a atenção de empresas, governos e consumidores. A chegada do ChatGPT em novembro de 2022 demonstrou o potencial da IA generativa com grandes modelos de linguagem (LLMs) para lidar com uma vasta gama de novos casos de uso. Tais casos eram inimagináveis com a computação convencional e a IA analítica (atualmente às vezes descrita como IA "tradicional" ou "clássica").

Agora bastam alguns prompts bem elaborados para automatizar uma série de coisas. Por exemplo, é possível gerar texto, imagens, áudio, vídeo e código de programação com qualidade profissional ou melhorar o atendimento aos clientes, executar modelagem das mudanças climáticas, descobrir novos medicamentos, projetar novos materiais, prever movimentos dos mercados financeiros e muito mais.

Da noite para o dia, uma pergunta apareceu no topo de todas as pautas de reuniões: *como podemos usar a IA generativa para revolucionar nossos mercados sem sofrermos turbulências?*

No entanto, os líderes de tecnologia logo reconheceram que, apesar dos benefícios potenciais da GenAI, também há riscos da imaturidade da tecnologia. Eles não podem descartar anos de melhores práticas operacionais e conhecimento institucional. Em vez disso, é preciso garantir que seus sistemas existentes, bem como novos aplicativos em desenvolvimento, sejam capazes de empregar a IA generativa de maneiras seguras, confiáveis e precisas.

Neste documento, discutiremos como o MongoDB pode colocar você no caminho para alcançar esses objetivos enquanto usa seus próprios dados para criar aplicativos e experiências interessantes com GenAI.

Contexto é tudo

Quando todo mundo tem acesso aos modelos de GenAI, o grande diferencial do seu "superpoder" é dar a esses modelos acesso a um dos ativos mais importantes da sua empresa: seus dados. Alguns desses dados são de propriedade da organização e alguns são públicos, mas mais recentes do que os usados para treinar os modelos de base originais. Juntos, esses dados fornecem respostas que refletem melhor a "verdade real" de hoje.

O fornecimento de modelos com seus próprios dados é realizado por um novo padrão de arquitetura chamado de geração aumentada por recuperação, ou RAG. Usar o

RAG apresenta aos desenvolvedores uma combinação potente. Eles podem aproveitar os incríveis recursos de conhecimento e raciocínio de modelos GenAI pré-treinados e de finalidade geral e alimentá-los com dados precisos e atualizados da empresa.

Como resultado, as produções da GenAI são precisas, atualizadas e relevantes e fazem uso de todos os seus dados, independentemente da estrutura. Seus aplicativos com GenAI atendem melhor seus clientes, aumentam a produtividade dos funcionários e superam as inovações da concorrência. Seus desenvolvedores podem desvendar todos esses resultados sem precisar recorrer a equipes especializadas de ciência de dados para treinar ou ajustar os modelos, evitando um processo complexo, demorado e caro.

Usar suas próprias fontes de dados é uma peça importante para fazer com que a IA generativa funcione para a empresa. Mas só isso não basta. Como discutiremos mais adiante, os desenvolvedores também precisam considerar como implantar seus aplicativos em torno de um grande modelo de linguagem informado com os controles de segurança corretos em vigor e na escala e no desempenho que os usuários esperam.

A ascensão da pesquisa vetorial e da pesquisa de similaridade

Para alimentar modelos de IA com nossos próprios dados, primeiro precisamos transformá-los em incorporações vetoriais. Esses vetores fornecem codificações numéricas multidimensionais dos nossos dados que capturam seus padrões, relacionamentos e estruturas. As incorporações vetoriais dão aos nossos dados um significado semântico; o cálculo da distância entre vetores ajuda nossos aplicativos a entenderem com facilidade os relacionamentos e as semelhanças entre diferentes objetos de dados. Isso abre nossos dados para uma nova gama de aplicativos que discutimos abaixo.

Dados em qualquer formato digital e estrutura, como texto, vídeo, áudio, imagens, códigos e tabelas, podem ser transformados em vetor ao serem processados com um modelo de incorporação vetorial adequado. Por exemplo, o `text-embedding-ada-002` da OpenAI é um dos modelos mais populares de vetorização de conteúdo textual. A vantagem das incorporações vetoriais é que os dados não estruturados e, portanto, completamente opacos a um computador, agora podem ter seu significado e estrutura inferidos e representados por meio dessas incorporações. Isso significa que podemos começar a pesquisar e calcular dados não estruturados da mesma maneira que sempre conseguimos com dados comerciais estruturados. Mais de 80% dos dados que criamos todos os dias não são estruturados, então os benefícios da pesquisa vetorial transformacional combinada com a GenAI começam a ficar mais evidentes.

Conforme mostrado na Figura 1 abaixo, depois que nossos dados são transformados em incorporações vetoriais, eles são persistidos e indexados em um armazenamento vetorial, como o MongoDB Atlas Vector Search. Para recuperar vetores semelhantes, o armazenamento é consultado com um algoritmo do vizinho mais próximo (ANN) para realizar uma pesquisa de K vizinhos mais próximos (KNN) usando um algoritmo como Hierarchical Navigable Small Worlds (HNSW).

A consulta desses vetores nos permite fazer coisas com os dados que antes só podíamos realizar com habilidades e infraestruturas caras de ciência de dados. Em primeiro lugar, é possível ampliar a pesquisa e a descoberta de informações além da correspondência de palavras-chave para uma semantic search sensível ao contexto, que é capaz de inferir significado e intenção do termo de pesquisa de um usuário. Em segundo lugar, podemos recuperar nossos próprios dados, codificados como vetores, para fornecer ao modelo de GenAI o contexto necessário para gerar saídas mais confiáveis e precisas. Essas saídas podem incluir:

- processamento de linguagem natural (PLN) para tarefas como chatbots e tarefas de pergunta e resposta para resumo de textos e análise de sentimentos;
- visão computacional e processamento de áudio para classificação de imagens e detecção de objetos por meio de reconhecimento de fala e tradução;
- geração de conteúdo, como documentação baseada em texto, páginas da web otimizadas para SEO e código de computador, ou conversão de texto para imagem ou vídeo.

Pesquisa vetorial e fluxo de trabalho com LLM

A Figura 1 reúne o fluxo de trabalho que permite a "geração aumentada por recuperação" para um LLM.

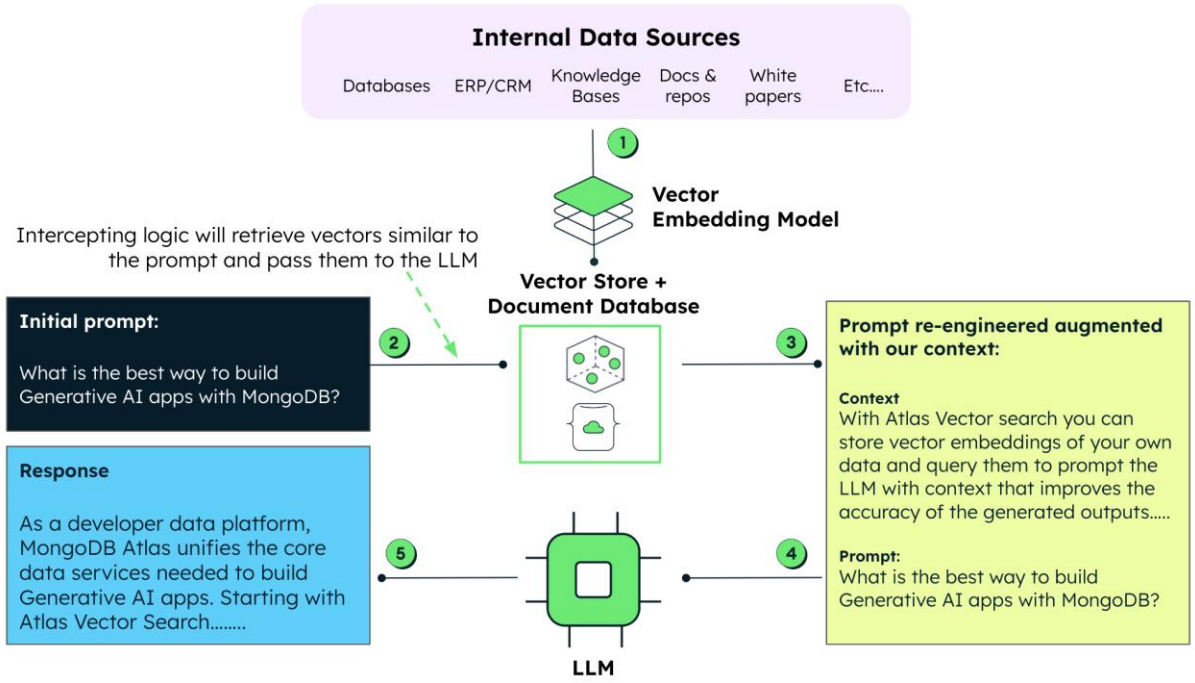

Figura 1: *faça uma combinação dinâmica dos seus dados personalizados com o LLM para gerar resultados confiáveis e relevantes*

Com antecedência, nossos dados são transformados por um modelo de incorporação vetorial e armazenados em um armazenamento de vetor. Idealmente, os metadados dos vetores e os dados brutos "agrupados" são armazenados com os próprios vetores em um banco de dados flexível de documentos, que também armazena nossos dados regulares de aplicativos. Isso permite que o aplicativo consulte os dados de várias maneiras, melhore a relevância (por exemplo, ao atribuir mais pontos aos dados mais recentes) e forneça memória de longo prazo para o LLM. Os prompts para o LLM são interceptados pela lógica que recupera vetores semelhantes do armazenamento vetorial. Em seguida, eles são usados para reprojetar o prompt inicial, e o novo prompt é enviado para o LLM, que pode usar o contexto fornecido para gerar respostas precisas e de maior qualidade usando dados mais detalhados.

Posteriormente neste artigo, você encontrará exemplos que demonstram o fluxo de trabalho acima e mostram como os recursos resultantes podem ser aplicados a diferentes classes de aplicativos.

A promessa e a realidade de um ecossistema vibrante de IA

Os armazenamentos de vetores fazem parte de um ecossistema de tecnologias de IA, que abrange criação de incorporações, engenharia por prompts, LLMs, ajuste fino de modelos, orquestração, criação de log, automação de infraestrutura e muito mais.

Nesse ecossistema, há inúmeras opções de projetos e fornecedores interessantes e promissores. Alguns mostram a "arte do possível" com demonstrações e protótipos. Mas o receio dos tomadores de decisões e dos desenvolvedores é a facilidade com que esses protótipos podem ser adaptados às suas necessidades comerciais específicas. Eles também questionam se algumas das tecnologias mais recentes podem realmente sustentar a carga de produção com confiabilidade, escalabilidade e segurança, dia após dia, em qualquer ambiente operacional. Outra consideração é como integrar os próprios bancos de dados da organização para alimentar o modelo com dados comerciais reais e verdadeiros.

O ecossistema de IA não existe isoladamente. Todas essas tecnologias precisam ser incorporadas a aplicativos do mundo real para serem verdadeiramente úteis para a empresa. Por exemplo, os armazenamentos de vetores são essenciais para viabilizar a IA generativa sensível ao contexto e a semantic search. Mas essas são apenas uma parte de um aplicativo mais amplo que também precisa gerenciar dados comerciais regulares e não vetorizados.

Esses dados podem ser qualquer coisa: registros de clientes, pedidos e inventário, negócios e transações, cotações, coordenadas geoespaciais, detalhes e preços de produtos, medições de time-series e leituras de sensores, fluxos de cliques e feeds sociais, descrições de texto e muito mais.

Todos esses dados precisam ser consultados para potencializar a funcionalidade do aplicativo. O objetivo não é apenas para recuperar vizinhos aproximados entre vetores, mas também realizar operações regulares, como recuperar registros específicos, lidar com uma série de atualizações nos dados e executar agregações e transformações sofisticadas para apoiar o processamento de análises. Essas consultas aprimoram os recursos do aplicativo fora de qualquer caso de uso da IA generativa, mas elas se tornam ainda mais importantes quando podemos usá-los com prompts de contexto para nossos modelos, melhorando a precisão e a relevância dos resultados do modelo de GenAI.

Além de trabalhar com nossos dados de aplicativos e incorporações vetoriais, precisamos fazer as coisas não funcionais – cumprir SLAs de tempo de atividade, desempenho e escalabilidade, integrar novos recursos, proteger e fazer backup de dados e auditá-los. Algumas dessas coisas podem parecer chatas. É só quando acontece um problemão daqueles que entendemos a importância dessas ações.

Combinar as tecnologias para potencializar novas experiências orientadas por IA e fundi-las nos seus aplicativos corre o risco de criar uma variedade de produtos pontuais e uma complexidade que sobrecarrega suas equipes. Todos esses desafios se somam a experiências de desenvolvedor fragmentadas e ineficientes, uma infinidade de modelos operacionais e de segurança, uma montanha de dados emaranhados e trabalho de integração e muita duplicação de dados. Tudo isso atrasa

o lançamento das suas novas experiências orientadas por IA no mercado, enquanto aumenta seus custos e riscos.

O uso de uma plataforma de dados para desenvolvedores construída no MongoDB Atlas é uma estratégia melhor.

Uma plataforma de dados do desenvolvedor: a maneira ideal de criar aplicativos inteligentes

A plataforma de dados para desenvolvedores do MongoDB, desenvolvida com base no MongoDB Atlas, unifica serviços de dados de IA operacional, analítica e generativa para simplificar a criação de aplicativos inteligentes. No entanto, como você usa a IA – desde o treinamento e o fornecimento de seus próprios modelos de aprendizado de máquina até a incorporação da IA generativa mais recente aos seus aplicativos –, o Atlas é uma parte crítica da sua pilha. Do protótipo à produção, o Atlas garante que seus aplicativos estejam fundamentados na verdade com os dados operacionais mais recentes, ao mesmo tempo em que atendam à escala, à segurança e ao desempenho que os usuários esperam.

No centro do MongoDB Atlas, estão o modelo flexível de dados de documentos e a API de consulta nativa do desenvolvedor. Juntos, eles permitem que seus desenvolvedores acelerem drasticamente a velocidade da inovação, superando os concorrentes e aproveitando as novas oportunidades de mercado apresentadas pela IA generativa.

Documentos são a melhor maneira de os desenvolvedores trabalharem com dados porque mapeiam objetos em código, tornando-os intuitivos e fáceis de racionalizar. Os documentos podem modelar dados de qualquer estrutura, como a vasta diversidade de dados regulares de aplicativos que discutimos anteriormente e incorporações vetoriais compostas por milhares de dimensões. Qualquer uma dessas estruturas pode ser modificada a qualquer momento para oferecer suporte à adição de novos tipos de dados e recursos de aplicativos. Os documentos oferecem a flexibilidade de racionalizar e aproveitar esses dados de maneiras que os modelos de dados tabulares tradicionais de relational databases não permitem.

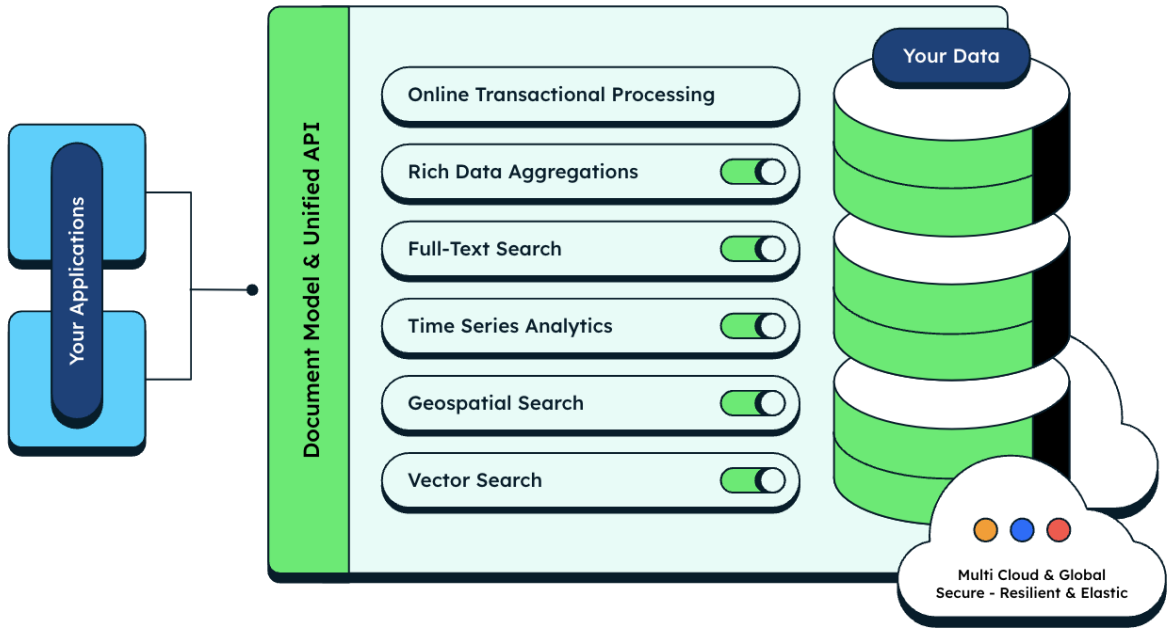

Figura 2: *o MongoDB Atlas integra os serviços de dados necessários para trazer IA para seus aplicativos*

Em conjunto com o modelo de documento, o MongoDB Query API oferece aos desenvolvedores uma maneira unificada e consistente de trabalhar com dados em qualquer serviço de dados. De operações CRUD simples a pesquisa por palavra-chave e semelhança vetorial, passando por pipelines de agregação sofisticados para análise e processamento de fluxo, o MongoDB Query API oferece aos desenvolvedores a flexibilidade de consultar e computar dados do modo que o aplicativo precisar.
No contexto da GenAI, isso viabiliza formas extremamente flexíveis e poderosas de definir filtros adicionais para consultas baseadas em vetores, por exemplo ao:

- Combinar com metadados para filtragem: "Encontre conteúdo que corresponda à consulta do usuário, mas somente conteúdo publicado nos anos X, Y e Z".
- Combinar com agregações: "Encontre todas as imagens semelhantes à imagem de consulta e agrupe-as por ID do fotógrafo".
- Combinar com pesquisa geoespacial: "Encontre anúncios imobiliários de casas semelhantes à casa desta fotografia e que estejam dentro de N quilômetros da minha localização".

Nenhum outro banco de dados é capaz de oferecer uma variedade de funcionalidades de consulta em uma única experiência de consulta unificada. Isso permite que os desenvolvedores criem funcionalidades de usuário final com mais facilidade e menos complexidade. Os desenvolvedores não precisam mais unir manualmente os resultados de consultas de vários bancos de dados, o que é um processo complicado, propenso a erros, caro e lento. Ao mesmo tempo, o sistema mantém a pegada tecnológica compacta e ágil.

"O MongoDB já estava armazenando metadados sobre artefatos em nosso sistema. Com a introdução do Atlas Vector Search, agora temos um banco de dados abrangente de metadados vetoriais que foi testado durante mais de uma década e que soluciona nossas densas necessidades de recuperação. Não precisamos implementar um novo banco de dados para gerenciar e aprender. Nossos vetores e metadados de artefatos podem ser armazenados um ao lado do outro."

Pierce Lamb, Engenheiro Sênior de Software da equipe de Dados e Machine Learning da VISO TRUST.

Mostre, não conte. Aplicativos aprimorados por IA generativa em uma plataforma de dados de desenvolvedor

Vamos nos concentrar em três casos de uso populares para mostrar como os desenvolvedores usam o MongoDB Atlas para criar aplicativos enriquecidos com IA:

- chatbot e perguntas e respostas (Q-A) para autoatendimento do cliente;
- pesquisa avançada de ecommerce e recomendações de usuários;
- análise e geração de rich media (multimodal).

Cada um desses exemplos conta com IA generativa e semantic search avançada para criar experiências de usuário incríveis e desbloquear recursos que antes estavam fora do alcance da maioria das organizações. Para serem verdadeiramente transformadores, no entanto, esses aprimoramentos de IA precisam ser entregues como parte de um aplicativo maior que está alimentando a funcionalidade crítica dos negócios.

Analisaremos cada caso de uso, mostrando um padrão de projeto de arquitetura compatível e os recursos relevantes fornecidos pelo MongoDB Atlas.

Chatbot e Q-A para autoatendimento do cliente

O MongoDB é o cérebro de muitos aplicativos de atendimento ao cliente, pois seu modelo de dados flexível facilita a criação de uma visão única de 360 graus do cliente. Isso é feito por meio da ingestão dinâmica de dados de clientes diversos e que mudam rapidamente a partir da miríade de sistemas de origem de back-end em silos, comuns na maioria das organizações. Portanto, a visão única e consolidada do cliente em tempo real, alimentada pelo MongoDB, é a plataforma ideal para treinar e fornecer recursos de chatbot e assistência de Q-A para o autoatendimento do cliente.

No exemplo mostrado na Figura 2, o banco de dados do cliente armazenado no MongoDB é exportado como arquivo JSON para um modelo de incorporação que fragmenta os dados (usando ferramentas como LangChain ou LlamaIndex) e cria incorporações vetoriais a partir dele. Outras fontes de dados internas, como bases de conhecimento e documentação, também podem ser vetorizadas para uso no aplicativo. Os dados são importados de volta para o banco de dados MongoDB.

É preciso assegurar que os vetores sejam constantemente atualizados com os dados mais recentes dos clientes, então usamos o Atlas Triggers para observar se houve alterações de dados na nossa visão unificada. Assim que novos registros de clientes são inseridos ou os registros existentes são atualizados no banco de dados, o Atlas Triggers chama a API do modelo de incorporação para gerar os vetores correspondentes e carregá-los novamente no Atlas.

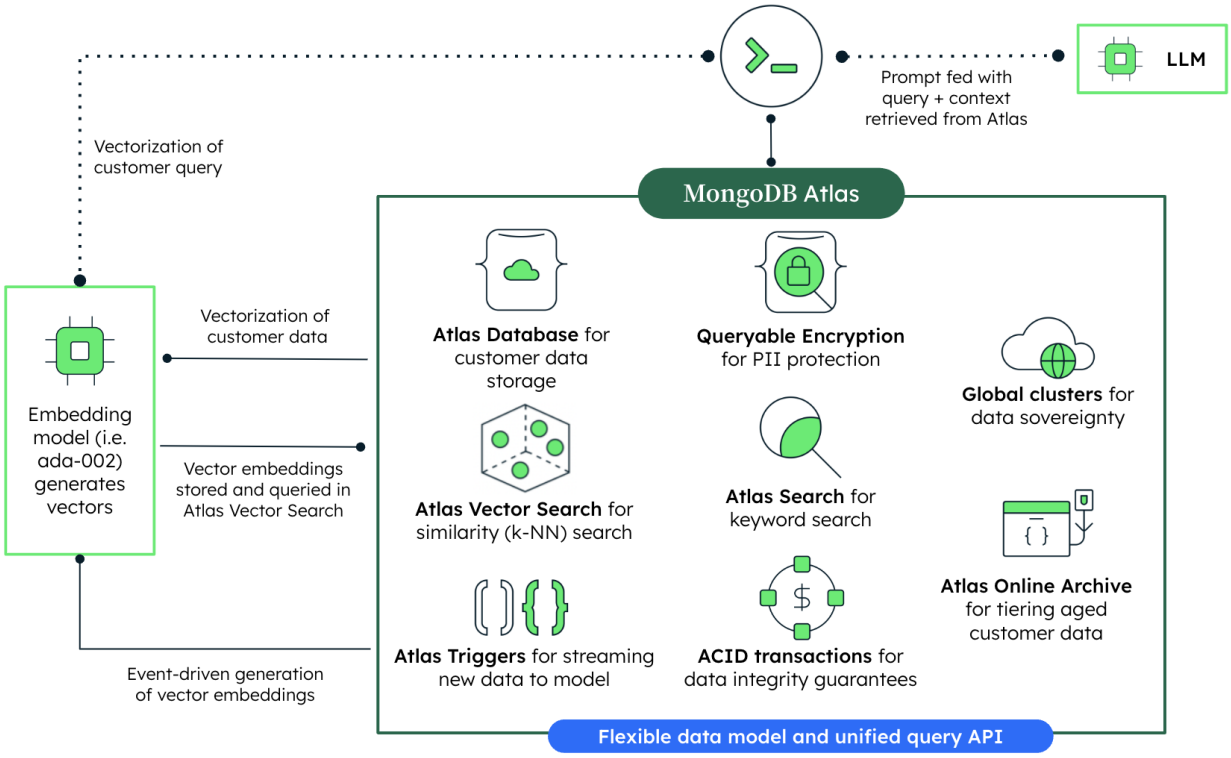

Figura 3: *chatbot e Q-A com recursos de IA generativa integrados a um aplicativo de autoatendimento do cliente com tecnologia MongoDB Atlas*

Ao utilizar o Atlas, os desenvolvedores obtêm os benefícios do modelo de dados flexível do MongoDB. Eles podem armazenar os dados, metadados e os chunks do cliente de origem juntamente com as incorporações do vetor, todos sincronizados e salvos lado a lado em uma única camada de armazenamento e acessados por uma só API de consulta e driver.

As consultas podem filtrar dados de forma eficiente usando os vetores indexados com índices de palavras-chave de campos regulares em seus documentos. Essa integração significa que o aplicativo pode oferecer suporte a uma gama muito mais ampla de funcionalidades do usuário com menos sobrecarga do desenvolvedor:

- O Atlas Vector Search retorna documentos correspondentes realizando uma pesquisa de similaridade em seus dados de incorporação indexados. Para reduzir o risco de retornar dados desatualizados, nossas consultas podem usar os metadados de um vetor armazenados no banco de dados do Atlas, como "data de criação", para excluir conteúdo antigo do filtro.
- O Atlas Search retorna resultados com base nas palavras-chave correspondentes na origem e nos dados do cliente fragmentados. Ele usa recursos como pesquisa difusa para corrigir erros de digitação na entrada do usuário e preenchimento automático para fornecer termos de pesquisa sugeridos, assim como intersecção de índice para atender de forma eficiente consultas ad hoc complexas em relação aos dados do cliente.

As consultas de banco de dados do Atlas, a pesquisa vetorial e o Atlas Search utilizam a mesma interface de consulta e driver, simplificando drasticamente o fluxo de trabalho do desenvolvedor. Os dados recuperados do MongoDB Atlas são fornecidos como contexto para aumentar o prompt para o LLM, permitindo a geração de respostas relevantes para chats e perguntas. Contexto e prompts, além de quaisquer etapas de raciocínio associadas utilizadas para responder a perguntas complexas, são persistidos para o Atlas. Isso fornece ao LLM memória de longo prazo e melhora continuamente suas saídas.

Os dados do cliente são alguns dos mais valiosos que qualquer organização gerencia. Embora a IA generativa nos ajude a inovar na forma como atendemos aos clientes, a proteção de seus dados permanece fundamental. O Atlas oferece uma variedade de recursos para nos ajudar a fazer isso, liberando os desenvolvedores para se concentrar nos recursos orientados por IA:

- Infraestrutura convergente que alimenta armazenamento de dados, consultas e análises, pesquisa por palavra-chave e pesquisa vetorial. Essa unificação em uma só API e modelo de dados reduz drasticamente o número de peças que os desenvolvedores precisam considerar na hora da integração e da criação.
- A Queryable Encryption é pioneira no setor para proteção de dados de clientes. Os drivers do MongoDB criptografam campos de dados confidenciais do lado do cliente, e o banco de dados só os enxerga como dados criptografados totalmente aleatórios. Mesmo com os dados criptografados, os aplicativos ainda podem executar consultas expressivas neles sem descriptografar dados no banco de dados. Na maioria dos casos, só os campos com dados de mais alta confidencialidade e que identificam exclusivamente um indivíduo, como SSN, são protegidos com Queryable Encryption. Dessa forma, a pesquisa pode ser realizada nos campos de texto não criptografado restantes.
- Transações ACID com vários documentos no banco de dados do Atlas garantem a integridade dos dados dos clientes sempre que forem acessados e modificados pelo aplicativo.

- Com o Atlas Global Clusters, os dados dos clientes podem ser fixados na região de residência deles, em conformidade com os modernos regulamentos de soberania de dados.
- O gerenciamento completo do ciclo de vida dos dados é fornecido pelo Atlas Online Archive. O serviço automaticamente transfere os dados antigos de clientes situados em bancos de dados ativos para um armazenamento de objeto de cloud de menor custo, mantendo-os acessíveis para consulta. Isso é importante para os dados dos clientes gerenciados em aplicativos que operam em setores regulamentados, que exigem a manutenção e a possibilidade de acesso por vários anos.
- Os dados do cliente são protegidos contra corrupção e ransomware com backups e restauração pontual.

O Atlas é totalmente gerenciado para você nas principais clouds de hiper-escala, com SLA de 99,995% de tempo de atividade.

Pesquisa avançada e recomendações de ecommerce

Catálogos de produtos de ecommerce são um caso de uso comum para o MongoDB:

- A diversidade dos produtos e seus atributos são naturalmente mapeados para o modelo flexível de dados de documentos do MongoDB.
- A arquitetura distribuída do Atlas com escala elástica permite que os desenvolvedores dimensionem e ajustem dinamicamente a capacidade do banco de dados em resposta à demanda dos aplicativos (por exemplo, em períodos sazonais de aumento nas vendas e em promoções).
- Com o Atlas Search, recursos de correspondência de palavras-chave, como pesquisa difusa, preenchimento automático, faceting, destaque e pontuação personalizada, permitem que os compradores naveguem rapidamente pelo catálogo de produtos, aumentando as taxas de cliques (CTRs) e garantindo conversões.

No entanto, a pesquisa por palavra-chave depende da correspondência de palavras específicas em campos de texto indexados para retornar resultados relevantes. Sem mapeamento de sinônimos extensivo e trabalhoso (por exemplo, mapeamento de bicicletas para ciclismo, ou tênis para treinadores), os usuários ficarão frustrados quando suas consultas de pesquisa não retornarem produtos relevantes. Essa frustração se traduz em perda de vendas e danos à reputação da marca.

Um desafio adicional é oferecer recomendações aos usuários. Os desenvolvedores precisam escrever mecanismos complexos baseados em regras ou recorrer a recursos especializados e escassos de ciência de dados. Normalmente, os dados devem ser ETLed (extraídos, transformados, carregados) do banco de dados operacional para um data warehouse ou data lake offline. É só depois disso que os modelos analíticos

de IA tradicionais podem gerar um conjunto de recomendações que precisam ser carregadas de volta no banco de dados operacional. O processo é complexo e caro e gera recomendações instantaneamente obsoletas, pois elas não refletem o comportamento de navegação ou as compras mais recentes do usuário.

O aprimoramento do catálogo de produtos com incorporações vetoriais elimina esses desafios.
Os vetores fornecem significado semântico para os produtos do catálogo, o que facilita a compreensão das semelhanças e dos relacionamentos entre os produtos. Isso permite que os comerciantes apresentem produtos relevantes e relacionados aos usuários com muito menos esforço, complexidade e custo. Os termos de pesquisa comuns podem ser armazenados em cache no MongoDB Atlas, fornecendo resultados relevantes aos usuários com mais agilidade.

Estender a vetorização aos dados do cliente, como demonstrado acima no aplicativo de autoatendimento do cliente, nos permite criar recomendações ainda mais sofisticadas, combinando pesquisa de similaridade de produto e cliente para ajustar as sugestões.

A Figura 4 mostra um padrão de design de alto nível para pesquisas e recomendações avançadas. A criação e a manutenção das incorporações vetoriais seguem o mesmo fluxo de trabalho descrito anteriormente para chatbots e Q-A no aplicativo de autoatendimento ao cliente.

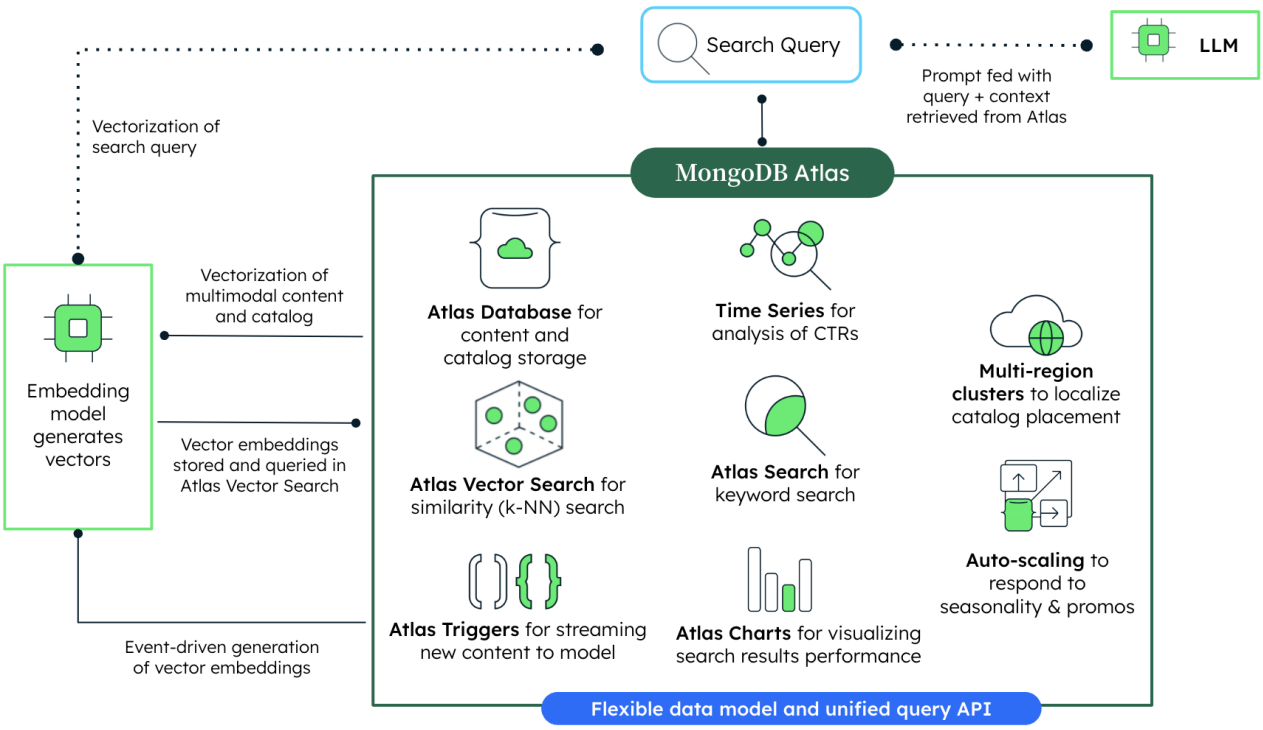

Figura 4: *a semantic search avançada no nosso catálogo de produtos gera mais conversões de vendas e vendas adicionais*

É fácil ver como a pesquisa vetorial melhora drasticamente a pesquisa de produtos e as recomendações. A integração de um LLM leva essa experiência ainda mais longe. Agora os clientes podem fazer perguntas em tempo real e obter respostas imediatas sobre os produtos que estão avaliando, o que é útil para acelerar o ciclo de compra.

Os comerciantes também podem usar o LLM para uma série de tarefas que antes seriam trabalhosas, liberando-os para desenvolver formas ainda mais criativas de envolver os clientes. Por exemplo, é possível usar o LLM para gerar diferentes variações de cópias de produtos e palavras-chave de SEO, que podem ser analisadas em um teste A/B para quantificar quais parâmetros geram mais conversões. O LLM pode ser usado para resumir várias avaliações de usuários e inferir sentimentos, ajudando a sintetizar o feedback que informa os roteiros de produtos.

As organizações podem usar o Atlas para gerenciar todo o ciclo de vida do ecommerce. Além de usar IA para tornar nossa experiência de pesquisa mais inteligente e preditiva, os proprietários de negócios podem acompanhar as taxas de cliques do usuário e as conversões de vendas dos resultados de pesquisa. Coleções de time-series podem ingerir e armazenar eficientemente dados de alta velocidade e fluxos de cliques volumosos das sessões de usuário, disponibilizando-os para medir o desempenho da pesquisa, inclusive com visualizações em tempo real dos resultados com Atlas Charts. Com esses insights, os comerciantes podem ajustar e otimizar continuamente os dados do produto e a pontuação de relevância para maximizar as vendas no site de ecommerce.

Análise e geração de rich media (multimodal)

A pesquisa de texto normal é bem servida com a pesquisa de palavras-chave convencional. No entanto, o trabalho com recursos de rich media (às vezes chamados de multimodais), como imagens, fala e vídeo, exigia tecnologia e habilidades de ciência de dados altamente complexas. Até agora.

Conforme observado, qualquer parte do conteúdo digital pode ser vetorizada com o modelo de incorporação vetorial apropriado. Os hubs de IA, como o Hugging Face e os dos hiper-escaladores de cloud, oferecem uma grande variedade de modelos ajustados para diferentes modalidades de conteúdo. As incorporações desses modelos podem ser armazenadas no Atlas Vector Search para alimentar toda uma gama de novas funcionalidades. Conforme discutido, a geração de imagens a partir de texto, a transcrição de vídeos para reconhecimento de fala e análise de sentimentos, a classificação de imagens e a detecção de objetos são apenas alguns exemplos do que é possível. Vetores de diferentes mídias podem ser combinados para, por exemplo, comparar uma incorporação de texto e imagem para atestar se uma determinada frase descreve com precisão uma imagem.

Essa funcionalidade multimodal pode ser usada em uma série de casos de uso, como enriquecimento de catálogos de produtos, como os descritos acima, ou aprimoramento da descoberta a partir da análise de imagens e vídeos. Ela também é útil para simplificar os processos de design, fabricação e publicação, ou para criar classes de aplicativos totalmente novas em domínios como segurança e vigilância ou realidade aumentada (AR).

O padrão de projeto de arquitetura e os recursos do MongoDB Atlas descritos para a pesquisa avançada de ecommerce e as recomendações acima se aplicam igualmente à geração de conteúdo multimodal.

Pesquisa vetorial do MongoDB em ação

O MongoDB já observou uma adoção generalizada para casos de uso de IA tradicionais. A Continental escolheu o MongoDB para a plataforma de engenharia de recursos na sua iniciativa de direção autônoma Vision Zero. Tanto a Bosch quanto a Telefonica usam o MongoDB em suas plataformas de IoT aprimoradas por IA. A Kronos negocia bilhões de dólares em criptomoedas todos os dias usando modelos de ML configurados e construídos com dados do MongoDB. A Iguazio usa o MongoDB como a camada de persistência para sua plataforma de ciência de dados e MLOps, enquanto a H2O.ai e a Featureform usam o MongoDB como repositório de recursos nas suas respectivas plataformas.

Partindo dessa fundação sólida, o MongoDB Atlas é usado hoje em uma variedade de aplicativos que estão ultrapassando os limites do possível com a GenAI. Confira nossa página de estudos de caso para saber mais sobre os diversos casos de uso atendidos pelo MongoDB Atlas. Veja alguns exemplos específicos:

- Ada: ajuda empresas como Meta, ATT e Verizon a oferecer melhor suporte aos clientes com automação orientada por IA e IA conversacional.
- ExTrac: identifica e classifica os riscos físicos e digitais emergentes da análise de fluxos de dados em tempo real.
- Eni: desbloqueia dados geológicos e os torna acionáveis para tomar decisões melhores e acelerar o caminho da empresa rumo ao net zero.
- Inovaare: monitora, extrai e classifica continuamente os dados em todo o ciclo de vida do setor de saúde para relatórios de conformidade regulamentar, auditoria e avaliações de risco.
- Source Digital: obteve redução de 7x nos custos ao migrar sua plataforma de detecção de vídeo do PostgreSQL para o MongoDB Atlas.
- Catylex: extrai, classifica e analisa automaticamente os termos do contrato para identificar direitos, obrigações e riscos.
- Robust Intelligence: protege grandes modelos de linguagem (LLMs) em produção ao validar entradas e saídas em tempo real com seu firewall de IA.

- Potion: regenera fluxos de vídeo e áudio usando modelos de áudio e visão personalizados.

Figura 5: *pesquisa State of AI da Retool – os principais bancos de dados vetoriais do setor*

Refletindo a popularidade do MongoDB entre os desenvolvedores de IA, o fornecedor de ferramentas de software Retool constatou na sua pesquisa State of AI que o MongoDB Atlas Vector Search:

1. obteve o mais alto Net Promoter Score (NPS) entre todos os bancos de dados vetoriais pesquisados;
2. tornou-se o segundo banco de dados vetoriais mais usado poucos meses depois do lançamento, ficando à frente de soluções alternativas que existem há anos.

"O Atlas Vector Search é robusto, econômico e incrivelmente rápido!" Saravana Kumar, CEO da Kovai, fala sobre o desenvolvimento do assistente de IA da sua empresa.

Começar

É indiferente se você está criando o próximo grande divisor de águas em uma startup ou em uma empresa de grande porte. Com o MongoDB Atlas, é possível:

- Acelere a construção de suas aplicações enriquecidas com IA generativa que são fundamentados na verdade dos dados operacionais.
- Simplificar sua pilha de tecnologia com uma única plataforma que permite que seu aplicativo armazene dados operacionais e incorporações vetoriais no mesmo lugar, reaja às mudanças nos dados de origem com funções serverless e pesquise em várias modalidades de dados para melhorar a relevância e a precisão das respostas geradas pelos aplicativos.
- Evoluir facilmente seus aplicativos enriquecidos com IA generativa com a flexibilidade do modelo de documento e, ao mesmo tempo, manter uma experiência de desenvolvedor simples e elegante.
- Integre perfeitamente os principais serviços e sistemas de IA, como os hiperescaladores e LLMs e estruturas de código aberto, para manter a competitividade em mercados dinâmicos.
- Construir aplicativos enriquecidos com GenAI em um banco de dados operacional de alto desempenho e altamente escalável, que tem o respaldo de uma década de validação em uma ampla variedade de casos de uso de IA.

Para saber mais sobre como criar aplicativos baseados em IA com MongoDB, visite nosso centro de recursos AI/ML.

A melhor maneira para os desenvolvedores darem seus primeiros passos é criando uma conta no MongoDB Atlas. A partir daí, eles podem criar uma instância MongoDB gratuita com o banco de dados Atlas, Atlas Vector Search e Atlas Search, carregar seus próprios dados ou nossos exemplos de conjuntos de dados, e explorar as possibilidades da plataforma.

Porto seguro

O desenvolvimento, o lançamento e o timing de quaisquer recursos ou funcionalidades descritos dos nossos produtos permanecem a nosso exclusivo critério. Esta informação destina-se apenas a delinear a direção geral do nosso produto e não deve ser invocada na tomada de uma decisão de compra nem é um compromisso, promessa ou obrigação legal de entregar qualquer material, código ou funcionalidade.

EUA 866-237-8815 • Exterior +1-650-440-4474 • INFO@MONGODB.COM.