Cómo incorporar IA generativa y búsqueda avanzada en sus aplicaciones con MongoDB

La creación de aplicaciones basadas en IA

Diciembre de 2023

Estados Unidos 866-237-8815 • INTL + 1-650-440-4474 • info@mongodb.com

Índice

Introducción

Nunca antes la introducción de una nueva tecnología había captado tan rápidamente la atención de empresas, gobiernos y consumidores por igual. La llegada de ChatGPT en noviembre de 2022 mostró el potencial de la IA generativa basada en modelos de lenguaje grandes (LLM) para abordar una amplia gama de nuevos casos de uso. Estos casos de uso antes eran inimaginables con la IA analítica y la computación convencional (ahora descrita a veces como IA "tradicional" o "clásica").

Parece que bastan unas cuantas instrucciones bien elaboradas para automatizar toda una serie de cosas. Genere texto, imágenes, audio, video y código de programación de calidad profesional. Bríndele mejor asistencia técnica a los clientes. Modele el cambio climático, descubra nuevos fármacos o diseñe nuevos materiales, prediga los movimientos de los mercados financieros y mucho, mucho más.

De la noche a la mañana, una pregunta se instaló primera en el orden del día de todos los directorios: *"¿cómo podemos usar la IA generativa para revolucionar nuestros mercados sin que nosotros mismos nos veamos afectados?"*

Sin embargo, los líderes tecnológicos han reconocido rápidamente que, además de los beneficios potenciales de la IA generativa, también existen riesgos derivados de la inmadurez de la tecnología. No se puede descartar sin más años de mejores prácticas operativas ni conocimientos institucionales. En su lugar, hay que asegurarse de que tanto los sistemas existentes como las nuevas aplicaciones en desarrollo sean capaces de aprovechar la IA generativa de manera segura, confiable y precisa.

En este documento, analizaremos cómo MongoDB puede encaminar el logro de esos objetivos mientras utiliza sus propios datos para impulsar nuevas aplicaciones y experiencias atractivas basadas en la IA generativa.

El contexto lo es todo

Cuando todos tienen acceso a los modelos de IA generativa, su "superpoder" diferencial proviene de darles a esos modelos acceso a uno de los activos empresariales más importantes: sus datos. Algunos de estos datos serán propiedad de la organización y otros serán públicos (pero más recientes) que los utilizados para entrenar los modelos básicos originales. Juntos, estos datos proporcionan respuestas que reflejan mejor el "marco verdadero" de hoy.

Proporcionar modelos con sus propios datos se logra mediante un nuevo patrón arquitectónico llamado Generación aumentada de recuperación o RAG. El uso de RAG

le presenta a sus desarrolladores una combinación potente. Pueden tomar el increíble conocimiento y las capacidades de razonamiento de los modelos de la IA generativa de propósito general previamente entrenados y alimentarlos con datos precisos y actualizados específicos de la empresa.

Los resultados de la IA generativa son precisos, actualizados, relevantes y aprovechan todos sus datos, sin importar su estructura. Sus aplicaciones basadas en la IA generativa sirven mejor a sus clientes, aumentan la productividad de los empleados y mejoran la innovación de la competencia. Sus desarrolladores pueden desbloquear todos estos resultados sin tener que recurrir a equipos de ciencia de datos especializados para entrenar o ajustar modelos, un proceso complejo, costoso y que requiere mucho tiempo.

Usar sus propias fuentes de datos es una pieza importante para hacer que la IA generativa funcione para el negocio. Sin embargo, no es suficiente por sí solo. Como veremos más adelante en el documento, los desarrolladores también deben considerar cómo implementar su aplicación en torno a un LLM informado con los controles de seguridad adecuados implementados y a la escala y con el rendimiento que los usuarios esperan.

El auge de los vectores y la búsqueda de similitudes

Para alimentar los modelos de IA con nuestros propios datos, primero debemos convertirlos en incrustaciones vectoriales. Estos vectores proporcionan codificaciones numéricas multidimensionales de nuestros datos que capturan sus patrones, relaciones y estructuras. Las incrustaciones vectoriales dan significado semántico a nuestros datos; calcular la distancia entre vectores facilita que nuestras aplicaciones comprendan las relaciones y similitudes entre diferentes objetos de datos. Esto abre nuestros datos a una gama completamente nueva de aplicaciones que analizamos a continuación.

Los datos en cualquier formato digital y de cualquier estructura (es decir, texto, video, audio, imágenes, código, tablas) se pueden transformar en un vector procesándolos con un modelo de incrustación vectorial adecuado. Por ejemplo, `text-embedding-ada-002` de OpenAI es uno de los modelos más populares para vectorizar contenido textual. Lo particular de las incrustaciones vectoriales es que los datos que no están estructurados y, por lo tanto, resultan completamente opacos para una computadora, ahora pueden inferir y representar su significado y estructura a través de estas incrustaciones. Esto significa que podemos empezar a buscar y calcular datos no estructurados de la misma manera que siempre hemos podido hacerlo con los datos empresariales estructurados. Teniendo en cuenta que más del 80 % de los datos que

creamos todos los días no están estructurados, empezamos a apreciar lo transformadora que es en realidad la búsqueda vectorial combinada con la IA generativa.

Como se muestra en la Figura 1 a continuación, una vez que nuestros datos se han transformado en incrustaciones vectoriales, se conservan y se indexan en un almacén de vectores como Atlas Vector Search de MongoDB. Para recuperar vectores similares, se consulta el almacén con un algoritmo de vecino más cercano aproximado (ANN) para realizar una búsqueda de vecino más cercano K (KNN) utilizando un algoritmo como "Mundo pequeño navegable jerárquico" (HNSW).

Consultar estos vectores nos permite hacer cosas con datos que antes solo podíamos lograr con habilidades e infraestructura de ciencia de datos muy costosas. En primer lugar, podemos ampliar la búsqueda y el descubrimiento de información más allá de la coincidencia de palabras clave hasta la búsqueda semántica consciente del contexto, que es capaz de inferir el significado y la intención del término de búsqueda de un usuario. En segundo lugar, podemos recuperar nuestros propios datos, codificados como vectores, para proporcionar al modelo la IA generativa el contexto necesario para generar resultados más confiables y precisos. Estos resultados pueden incluir:

- Procesamiento de lenguaje natural (NLP) para tareas como chatbots y respuesta a preguntas, resumen de texto y minería de opinión.
- Procesamiento de audio y visión artificial para la clasificación de imágenes y detección de objetos a través del reconocimiento de voz y la traducción.
- Generación de contenido, incluida la creación de documentación basada en texto y páginas web optimizadas para SEO, el código informático o la conversión de texto a una imagen o video.

Búsqueda vectorial y flujo de trabajo de LLM

La Figura 1 muestra el flujo de trabajo que permite la RAG para un LLM.

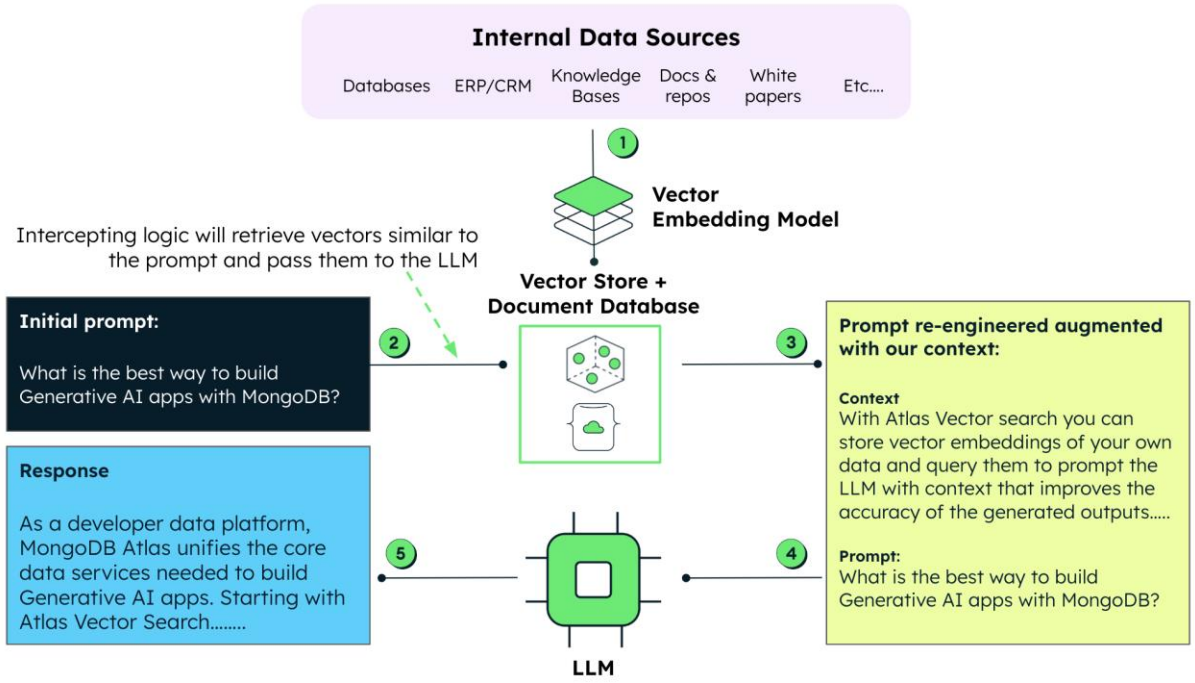

Figura 1: *Combinación dinámica de sus datos personalizados con el LLM para generar resultados confiables y relevantes*

De antemano, nuestros datos se transforman mediante un modelo de incrustación de vectores y se almacenan en un almacén de vectores. Idealmente, los metadatos de los vectores y los datos sin procesar "fragmentados" se almacenan junto con los vectores mismos en una base de datos de documentos flexible que también almacena nuestros datos de aplicaciones comunes. Esto permite a nuestra aplicación consultar datos de varias maneras, mejorar la relevancia (p. ej., que aparezcan primero los datos más recientes) y proporciona memoria a largo plazo para el LLM. A las indicaciones al LLM las intercepta lógica que recupera vectores similares del almacén de vectores. Luego, se utilizan para rediseñar la indicación inicial. La nueva indicación se envía al LLM que puede usar el contexto proporcionado para generar respuestas de mayor calidad y más precisas utilizando datos más recientes.

Más adelante en este documento, encontrará ejemplos que demuestran el flujo de trabajo anterior y muestran cómo las capacidades resultantes se pueden aplicar a diferentes clases de aplicaciones.

La promesa y la realidad de un ecosistema de IA dinámico

Los almacenes vectoriales forman parte de un ecosistema de tecnologías de IA en rápida evolución que abarca desde la creación de incrustaciones hasta la ingeniería de indicaciones, los LLM, el ajuste de modelos, la orquestación, el registro, la automatización de infraestructuras y mucho más.

En este ecosistema hay infinidad de proyectos y proveedores interesantes y prometedores con los que trabajar. Algunos muestran el "arte de lo posible" a través de demostraciones y prototipos. Sin embargo, el miedo que enfrentan los responsables de la toma de decisiones y los desarrolladores es la facilidad con la que estos prototipos se pueden adaptar a sus necesidades comerciales específicas. Y la pregunta de si algunas de las tecnologías más nuevas pueden de verdad sostener la carga de producción con confiabilidad, escalabilidad y seguridad, día tras día, en cualquier entorno operativo. Una consideración adicional es cómo integrar las bases de datos propias de la organización para ingresar datos de negocio reales y verdaderos al modelo.

El ecosistema de la IA no existe de forma aislada. Todas estas tecnologías deben integrarse en aplicaciones del mundo real para que sean útiles para el negocio. Por ejemplo, los almacenes vectoriales son esenciales para permitir la IA generativa basada en el contexto y la búsqueda semántica. Sin embargo, esto es sólo una parte de una aplicación más amplia que también tiene que gestionar datos comerciales regulares y no vectorizados.

Estos datos pueden ser cualquier registro de clientes, pedidos e inventario, negociaciones y transacciones, cotizaciones, coordenadas geoespaciales, detalles y precios de productos, mediciones de time-series y lecturas de sensores, enlaces y fuentes sociales, descripciones de texto y más.

Todos estos datos deben consultarse para potenciar la funcionalidad de la aplicación. No solo para recuperar los vecinos más cercanos aproximados entre vectores, sino también para realizar operaciones comunes, como recuperar registros específicos, manejar una gran cantidad de actualizaciones de los datos y ejecutar agregaciones y transformaciones sofisticadas que respalden el procesamiento analítico. Estas consultas potencian las características de las aplicaciones fuera de cualquier caso de uso de IA generativa. Se vuelven aún más importantes cuando podemos usarlas junto con indicaciones en contexto para nuestros modelos para mejorar la precisión y la relevancia de los resultados del modelo la IA generativa.

Más allá de trabajar con nuestros datos de aplicaciones e incrustaciones vectoriales, debemos hacer las cosas no funcionales como cumplir con el tiempo de actividad, el rendimiento y la escalabilidad de los SLA, integrar nuevas características, proteger y respaldar datos y auditarlos. Algunas cosas pueden sonar aburridas. Hasta que falla. Entonces, de repente, ya no es aburrido...

Reunir las tecnologías para impulsar nuevas experiencias basadas en IA y fusionarlas en sus aplicaciones corre el riesgo de crear una proliferación de productos puntuales y una complejidad que suponga enormes gastos generales para sus equipos. Todos estos retos se traducen en experiencias fragmentadas e ineficaces para los

desarrolladores, una multitud de modelos operativos y de seguridad con los que lidiar, una tonelada de trabajo de gestión e integración de datos y mucha duplicación de datos. Todo esto ralentiza la velocidad de comercialización de sus nuevas experiencias basadas en IA, al tiempo que aumenta sus costes y riesgos.

El uso de una plataforma de datos para desarrolladores basada en MongoDB Atlas le brinda una mejor manera de hacerlo.

Una plataforma de datos para desarrolladores: la forma inteligente de crear aplicaciones inteligentes

La plataforma de datos para desarrolladores de MongoDB, construida sobre MongoDB Atlas, unifica servicios de datos de IA operativa, analítica y generativa para agilizar la creación de aplicaciones inteligentes. Sin importar cómo aproveche la IA (ya sea que entrene y sirva sus propios modelos de aprendizaje automático o que incorpore la última IA generativa en sus aplicaciones), es fundamental que tenga Atlas en su pila. Desde el prototipo hasta la producción, con Atlas puede asegurarse de que sus aplicaciones estén basadas en la verdad con los datos operativos más actualizados y que, al mismo tiempo, cumplan con la escala, la seguridad y el rendimiento que esperan los usuarios.

En el núcleo de MongoDB Atlas se encuentra su modelo de datos de documento flexible y su API de consulta nativa para desarrolladores. Juntos, permiten a los desarrolladores acelerar drásticamente la velocidad de la innovación, y así superar a la competencia y capitalizar las nuevas oportunidades de mercado que presenta la IA generativa.

Para los desarrolladores, los documentos son la mejor forma de trabajar con datos porque se asignan a objetos en código para que sean intuitivos y fáciles de entender. Los documentos pueden modelar datos de cualquier estructura, desde la gran diversidad de datos de las aplicaciones comunes que analizamos con anterioridad hasta incrustaciones vectoriales compuestas por varios miles de dimensiones. Cualquiera de estas estructuras se puede modificar en cualquier momento para admitir la adición de nuevos tipos de datos y características de la aplicación. Los documentos le brindan la flexibilidad de racionalizar y aprovechar esos datos de una forma que no es posible de lograr para los modelos de datos tabulares tradicionales de bases de datos relacionales.

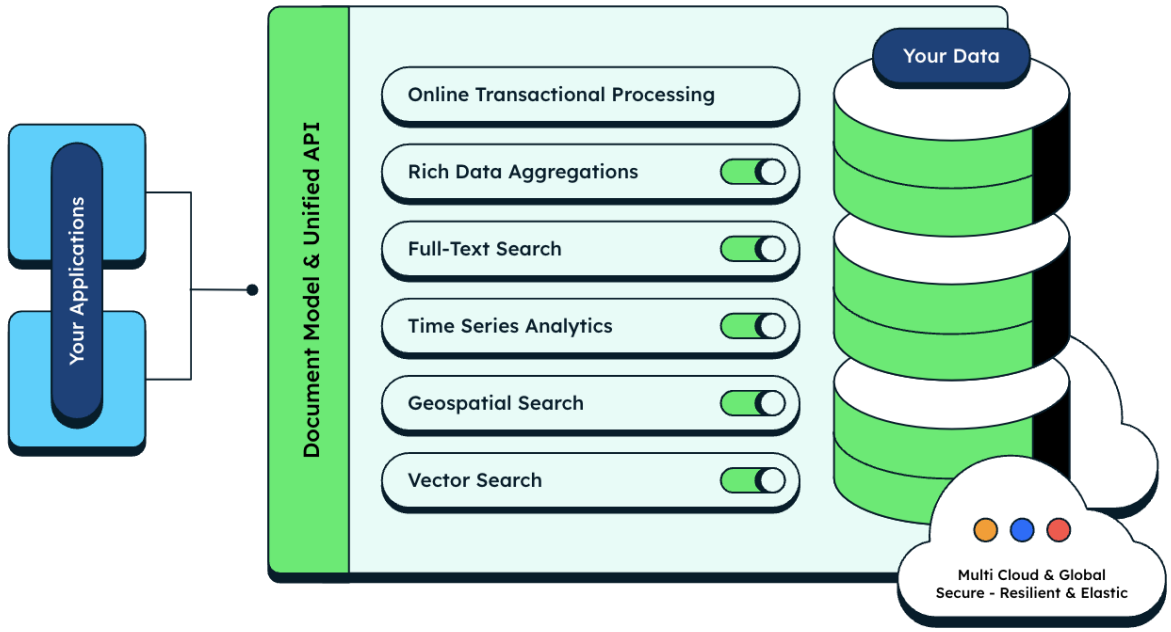

Figura 2: *MongoDB Atlas integra los servicios de datos necesarios para incorporar la IA a sus aplicaciones*

Junto con el modelo de documentos, la API consultiva de MongoDB brinda a los desarrolladores una forma unificada y coherente de trabajar con datos en cualquier servicio de datos. Desde simples operaciones CRUD hasta la búsqueda de similitud de palabras clave y vectores o sofisticadas pipelines de agregación para el procesamiento analítico y de flujos, la API consultiva de MongoDB proporciona a los desarrolladores la flexibilidad para consultar y calcular datos de cualquier manera que la aplicación necesite. En el contexto de la IA generativa, esto proporciona formas muy flexibles y poderosas de definir filtros adicionales en consultas basadas en vectores, como:

- Combinación con metadatos para filtrar: "Encuentra contenido que coincida con la consulta del usuario, pero solo si el contenido se publicó en los años X, Y y Z".
- Combinación con agregaciones: "Encuentra todas las imágenes similares a la imagen de la consulta y agrúpalas por el ID del fotógrafo".
- Combinación con la búsqueda geoespacial: " Encuentra anuncios inmobiliarios para casas que son similares a la casa de esta fotografía y que esté en un radio de N millas de mi ubicación ".

Ninguna otra base de datos puede ofrecer tal amplitud de funcionalidades de consulta en una única experiencia unificada de consulta. Esto les permite a los desarrolladores crear funciones para el usuario final con mayor facilidad y menor complejidad. Los desarrolladores ya no tienen que agrupar de forma manual los resultados de las consultas de varias bases de datos, lo cual es un proceso complejo, propenso a errores, costoso y lento. Al mismo tiempo, también mantiene su huella tecnológica compacta y ágil.

"MongoDB ya estaba almacenando metadatos sobre artefactos en nuestro sistema.Con la introducción de Atlas Vector Search, ahora contamos con una base de datos integral de metadatos vectoriales que se probó durante más de una década, y que resuelve nuestras densas necesidades de recuperación.No es necesario implementar una nueva base de datos que tendríamos que gestionar y aprender.Nuestros metadatos de artefactos y vectores se pueden almacenar juntos.".

Pierce Lamb, ingeniero de Software Senior del equipo de machine learning y datos en VISO TRUST.

No lo cuentes, muéstralo. Aplicaciones mejoradas con IA generativa en una plataforma de datos para desarrolladores

Nos enfocaremos en tres casos de uso populares para mostrar cómo los desarrolladores utilizan MongoDB Atlas para crear aplicaciones enriquecidas con IA:

- Chatbot y respuestas automáticas (Q-A) para autoservicio del cliente.
- Búsqueda avanzada de ecommerce y recomendaciones de los usuarios.
- Análisis y generación de medios enriquecidos (multimodal).

Cada uno de estos ejemplos depende de la IA generativa y de la búsqueda semántica avanzada para crear experiencias de usuario increíbles y desbloquear capacidades que antes estaban fuera del alcance de la mayoría de las organizaciones. Sin embargo, para que sean transformadoras de verdad, estas mejoras de la IA deben formar parte de una aplicación más grande que, a su vez, impulse funciones fundamentales para las empresas.

Abordaremos cada caso de uso mostrando un patrón de diseño arquitectónico compatible y las capacidades relevantes que brinda MongoDB Atlas.

Chatbot y Q-A para autoservicio del cliente

MongoDB es el corazón de muchas aplicaciones de atención al cliente. Esto se debe a que el modelo de datos flexible de MongoDB facilita la creación de una visión única de 360 grados del cliente. Para ello, ingiere de manera dinámica datos de clientes diversos y que cambian con rapidez, procedentes de la miríada de sistemas backend en silos que es típica de la mayoría de las organizaciones. La visión única y consolidada del cliente en tiempo real que ofrece MongoDB es, por tanto, la

plataforma ideal para entrenar y ofrecer características de chatbot y asistencia Q-A para el autoservicio del cliente.

En el ejemplo que se muestra en la Figura 2, la base de datos de clientes almacenada en MongoDB se exporta como un archivo de JSON a un modelo de incrustación que fragmenta los datos (mediante herramientas como LangChain o LlamaIndex) y crea incrustaciones vectoriales a partir de ellos. Otras fuentes de datos internas, como las bases de conocimientos y la documentación, también se pueden vectorizar para su uso en la aplicación. A continuación, los datos se vuelven a importar a la base de datos de MongoDB.

Necesitamos asegurarnos de que nuestros vectores se actualicen de forma constante con los datos más recientes de los clientes, por lo que usamos Atlas Triggers para observar cualquier cambio en los datos en nuestra vista única. Tan pronto como se insertan nuevos registros de clientes o se actualizan los registros existentes en la base de datos, Atlas Triggers llama a la API del modelo de incrustación para generar los vectores correspondientes y volver a cargarlos en Atlas.

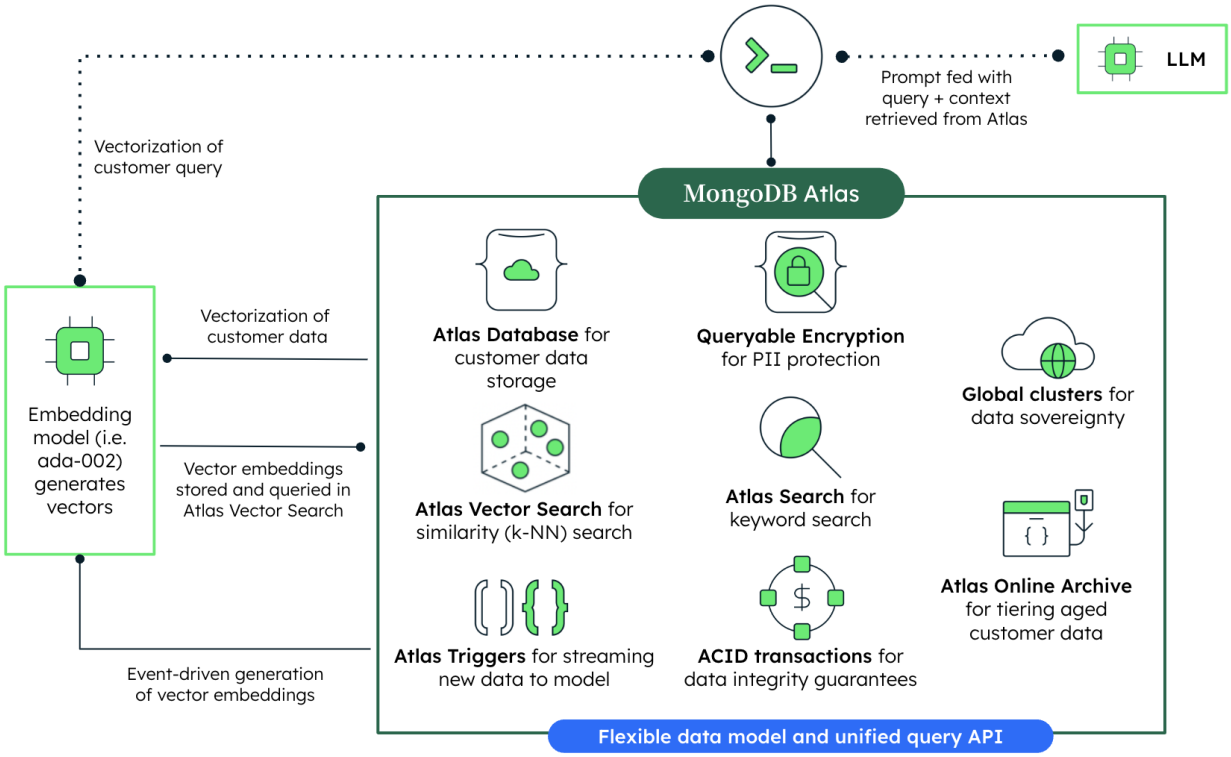

Figura 3: *Características de IA Generativa de Chatbot y respuestas automáticas (Q-A) integradas en una aplicación de autoservicio para clientes impulsada por MongoDB Atlas*

Al usar Atlas, los desarrolladores aprovechan el modelo de datos flexible de MongoDB. Pueden almacenar los datos de origen del cliente, los metadatos y los fragmentos junto con las incrustaciones vectoriales; todo ello está sincronizado y coexiste en una única capa de almacenamiento, a la que se accede mediante una única API de consulta y un único controlador.

Las consultas pueden filtrar datos de forma eficaz mediante los vectores indexados junto con los índices de palabras clave de los campos normales de los documentos. Esta integración significa que la aplicación es compatible con una gama mucho más amplia de funcionalidades de usuario con menores gastos generales para el desarrollador:

- Atlas Vector Search devuelve documentos coincidentes mediante la realización de una búsqueda de similitud en sus datos de incrustaciones indexadas. Para reducir el riesgo de que devuelva datos obsoletos, nuestras consultas pueden usar los metadatos de un vector, como la "fecha de creación" almacenada en la base de datos de Atlas, para filtrar el contenido anterior.
- Atlas Search devuelve resultados basados en palabras clave coincidentes en la fuente y datos de clientes fragmentados. Utiliza características como la búsqueda difusa para corregir errores tipográficos en el texto ingresado por los usuarios y autocompletar para sugerir términos de búsqueda. También utiliza la intersección de índices para responder de forma eficaz a consultas ad hoc complejas sobre los datos de los clientes.

Las consultas a la base de datos de Atlas, Vector Search y Atlas Search utilizan la misma interfaz de consulta y el mismo controlador, lo que simplifica en gran medida el flujo de trabajo del desarrollador. Los datos recuperados de MongoDB Atlas se proporcionan como contexto que aumenta la indicación al LLM, lo que permite generar respuestas relevantes a chats y preguntas. El contexto y las indicaciones, junto con los pasos de razonamiento asociados utilizados para responder preguntas complejas, se mantienen en Atlas, lo que le proporciona al LLM una memoria a largo plazo y mejora continuamente sus resultados.

Los datos de los clientes son de los más valioso que puede gestionar una organización. Si bien la IA generativa nos ayuda a innovar en la forma en que atendemos a nuestros clientes, la protección de sus datos sigue siendo una prioridad. Atlas proporciona una variedad de capacidades para ayudarnos a hacer esto, liberando a los desarrolladores para que puedan concentrarse en la funcionalidad impulsada por la IA:

- Infraestructura convergente que impulsa el almacenamiento de datos, las consultas y el análisis, la búsqueda de palabras clave y la búsqueda vectorial. Esta unificación detrás de una única API y modelo de datos reduce de manera drástica la cantidad de piezas móviles que los desarrolladores tienen que integrar y usar para construir.
- Queryable Encryption es una novedad en la industria a la hora de proteger los datos de los clientes. Los controladores de MongoDB cifran los campos de datos sensibles en el lado del cliente y la base de datos sólo trabaja con ellos como datos cifrados totalmente aleatorios. Incluso con los datos cifrados, las aplicaciones pueden ejecutar consultas expresivas sin tener que descifrar los datos de la base de datos. Tenga en cuenta que por lo general sólo los

campos que contienen los datos más sensibles que identifican de forma exclusiva a una persona, como el número de identificación, están protegidos con Queryable Encryption. Por lo tanto, se pueden realizar búsquedas en el resto de los campos en texto claro.
- Las transacciones ACID de varios documentos en la base de datos de Atlas garantizan la integridad de los datos de nuestros clientes siempre que la aplicación acceda a ellos y los modifique.
- Con Atlas Global Clusters, los datos de los clientes se pueden anclar a su región de residencia, en cumplimiento con la normativa moderna de soberanía de datos.
- La gestión completa del ciclo de vida de los datos se logra gracias a Atlas Online Archive. El servicio separa automáticamente los datos antiguos de los clientes de las bases de datos activas y los coloca en un almacenamiento de objetos de cloud de menor costo, al tiempo que mantiene los datos accesibles para consultarlos. Esto es importante para los datos de los clientes gestionados dentro de la aplicación que opera en industrias donde la normativa indica que se deben conservar durante varios años y tener acceso a ellos.
- Los datos de los clientes están protegidos contra la corrupción y el ransomware con copias de seguridad y punto de restauración.

Atlas está totalmente gestionada en las principales nubes de los hiperescaladores, respaldadas por un SLA de tiempo de actividad del 99.995 %.

Búsqueda avanzada de ecommerce y recomendaciones

Los catálogos de productos de ecommerce son un caso de uso común para MongoDB:

- La variedad de diferentes productos y sus atributos se asignan naturalmente al modelo de datos de documento flexible de MongoDB.
- La arquitectura de Atlas distributed con escalado elástico les permite a los desarrolladores dimensionar y ajustar dinámicamente la capacidad de la base de datos en respuesta a la demanda de la aplicación (por ejemplo, para la estacionalidad de las compras y las promociones de las ventas).
- Con Atlas Search, las características de coincidencia de palabras clave como búsqueda difusa, autocompletar, facetado, resaltado y puntuación personalizada les permiten a los compradores hacer una navegación rápida por el catálogo de productos, impulsando las tasas de clics (CTRs) y comprando conversiones.

Sin embargo, la búsqueda de palabras clave se basa en hacer coincidir palabras específicas en campos de texto indexados para devolver resultados relevantes. Sin un mapeo de sinónimos exhaustivo y laborioso (por ejemplo, mapear a las bicicletas con el ciclismo o a las zapatillas con las zapatillas de deporte), los usuarios se frustrarán rápido cuando sus consultas de búsqueda no devuelvan productos relevantes. Esta frustración se traduce en pérdidas de ventas y daños en la reputación de la marca.

Un desafío adicional es proporcionarles recomendaciones a los usuarios. Los desarrolladores tienen que escribir motores complejos basados en reglas o recurrir a recursos especializados y escasos de ciencia de datos. Por lo general, los datos primero tienen que ser ETLed (Extraídos, Transformados, Cargados) de la base de datos operativa en un data Lake o almacén de datos sin conexión. Solo entonces los modelos analíticos tradicionales de IA pueden generar un conjunto de recomendaciones que luego deben cargarse de nuevo en la base de datos operativa. El proceso es complejo, costoso y genera recomendaciones obsoletas al instante, ya que no reflejan el comportamiento de navegación o las compras más recientes del usuario.

Mejorar nuestro catálogo de productos con incrustaciones vectoriales elimina estos desafíos. Los vectores les proporcionan un significado semántico a los productos de nuestro catálogo, lo que facilita la comprensión de las similitudes y las relaciones entre los productos. Esto les permite a los comerciantes mostrarles a los usuarios productos relevantes que significan un esfuerzo, una complejidad y un coste mucho menores. Los términos de búsqueda habituales pueden almacenarse en el caché de MongoDB Atlas, lo que permite ofrecerles resultados relevantes a los usuarios con mayor rapidez.

Extender la vectorización a los datos de los clientes, como mostramos en la aplicación de autoservicio del cliente antes, nos permite crear recomendaciones aún más sofisticadas al combinar la búsqueda de similitud de productos y clientes para ajustar las sugerencias.

La Figura 4 muestra un patrón de diseño de alto nivel para la búsqueda avanzada y las recomendaciones. La creación y el mantenimiento de nuestras incrustaciones vectoriales sigue el mismo flujo de trabajo descrito con anterioridad para los chatbots y las respuestas automáticas (Q-A) en nuestra aplicación de autoservicio para clientes.

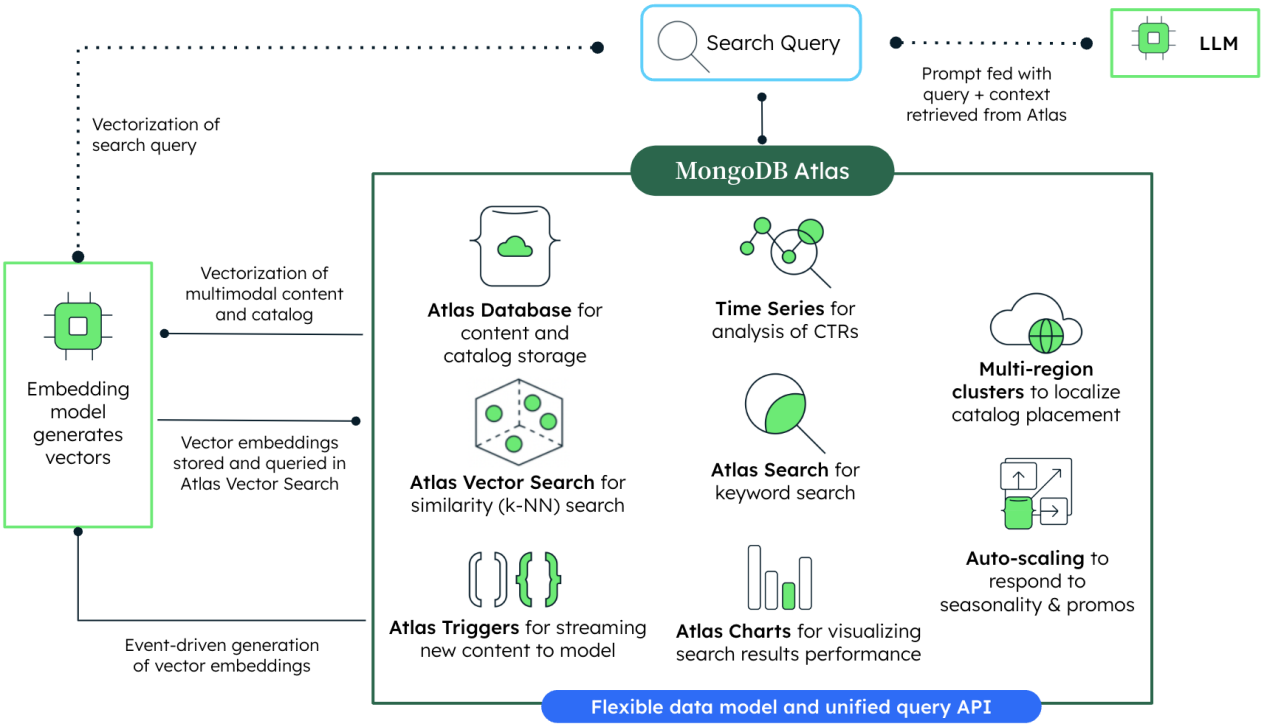

Figura 4: *La búsqueda semántica avanzada en nuestro catálogo de productos impulsa conversiones de ventas más altas y ventas adicionales*

Es fácil ver cómo la búsqueda vectorial mejora de forma drástica la búsqueda de productos y las recomendaciones. Integrar un LLM lleva esta experiencia aún más lejos. Ahora los clientes pueden hacer preguntas en vivo y obtener respuestas instantáneas sobre los productos que están evaluando, lo que ayuda a acelerar el ciclo de compra.

Los comerciantes también pueden utilizar el LLM para una serie de tareas que antes les habrían resultado laboriosas, y así pueden usar su tiempo en desarrollar formas aún más creativas de captar clientes. Por ejemplo, el LLM se puede utilizar para generar distintas variaciones del texto del producto y de las palabras clave de SEO, que luego se pueden someter a pruebas A/B para cuantificar cuál genera mayores conversiones. El LLM podría utilizarse para resumir múltiples opiniones de usuarios e inferir sentimientos, lo que ayudaría a sintetizar los comentarios que informan las hojas de ruta de los productos.

Las organizaciones pueden usar Atlas para gestionar el ciclo de vida completo del ecommerce. Además de usar la IA para que nuestra experiencia de búsqueda sea más inteligente y predictiva, los propietarios de empresas pueden realizar un seguimiento de las tasas de clics de los usuarios y las conversiones de ventas a partir de los resultados de búsqueda. Las colecciones de time-series pueden ingerir y almacenar de forma eficiente flujos de clics voluminosos y de alta velocidad procedentes de las sesiones de los usuarios, lo que permite que los datos estén disponibles para el análisis con el fin de medir el rendimiento de la búsqueda, lo que incluye las visualizaciones en directo de los

resultados mediante Atlas Charts. Con esta información, los comerciantes pueden ajustar y optimizar continuamente los datos de productos y la puntuación de relevancia para maximizar las ventas del sitio de ecommerce.

Análisis y generación de medios enriquecidos (multimodales)

La búsqueda convencional de palabras clave brinda un buen servicio para la búsqueda de texto común. Sin embargo, trabajar con activos multimedia enriquecidos (a veces llamados multimodales) como imágenes, voz y video requiere tecnología y habilidades altamente complejas en ciencia de datos. O así fue hasta ahora.

Como se señaló antes, cualquier pieza de contenido digital se puede vectorizar con el modelo de incrustación vectorial apropiado. Los hubs de IA como Hugging Face y los de los hiperescaladores en cloud proporcionan una gran cantidad de modelos ajustados para diferentes modalidades de contenido. Las incrustaciones de estos modelos se pueden almacenar en Atlas Vector Search para potenciar toda una gama de nuevas funcionalidades. Como se analizó antes, generar imágenes a partir de texto, transcribir videos para reconocimiento de voz y análisis de sentimientos, clasificar imágenes y detectar objetos son solo algunos ejemplos de lo que es posible. Se pueden combinar vectores de diferentes medios, por ejemplo, comparar un texto y una imagen incrustada para verificar si una oración determinada describe con precisión una imagen.

Esta funcionalidad multimodal se puede utilizar en una amplia gama de casos de uso. Por ejemplo, enriquecer catálogos de productos como los descritos antes, o mejorar el descubrimiento a partir del análisis de imágenes y vídeos. Podrían usarse para optimizar los procesos de diseño, fabricación y publicación, o para crear clases de aplicaciones completamente nuevas en dominios como la seguridad y la vigilancia o la realidad aumentada (AR).

El patrón de diseño arquitectónico y las capacidades de MongoDB Atlas descritos antes para la búsqueda avanzada en ecommerce y las recomendaciones se aplican de igual forma a la generación de contenidos multimodales.

Vector Search de MongoDB en acción

Ya muchos adoptaron MongoDB para casos de uso de IA tradicionales. Continental seleccionó MongoDB para la plataforma de ingeniería de características en su iniciativa de conducción autónoma Vision Zero. Tanto Bosch como Telefónica utilizan MongoDB en sus plataformas IoT mejoradas con IA. Kronos hace transacciones con miles de millones de dólares en criptomonedas cada día utilizando modelos ML configurados y construidos con datos de MongoDB. Iguazio utiliza MongoDB como capa de persistencia para su plataforma de ciencia de datos y MLOps, mientras que

H2O.ai y Featureform admiten MongoDB como almacenes de características en sus respectivas plataformas.

Sobre esta base, MongoDB Atlas ya se utiliza hoy en día en una serie de aplicaciones que están ampliando los límites de lo posible con la IA generativa. Eche un vistazo a nuestra página de casos para obtener más información sobre el alcance de los casos de uso atendidos por MongoDB Atlas. Esta es una selección de ejemplos específicos:

- Ada: ayuda a empresas como Meta, ATT y Verizon a brindar mejor asistencia técnica a sus clientes a través de la automatización impulsada por la IA y la IA conversacional.
- ExTrac: identifica y clasifica los riesgos físicos y digitales emergentes del análisis de flujos de datos en tiempo real.
- Eni: desbloquea datos geológicos y los hace accionables para una mejor toma de decisiones y para acelerar el camino de la empresa hacia el cero neto.
- Inovaare: monitorea, extrae y clasifica de forma continua los datos a lo largo del ciclo de vida de la atención médica para obtener informes de cumplimiento regulatorio, auditorías y evaluaciones de riesgos.
- Source Digital: logra una reducir los costos 7 veces después de migrar de PostgreSQL a MongoDB Atlas para su plataforma de detección de video.
- Catylex: extrae, clasifica y analiza automáticamente los términos del contrato para identificar derechos, obligaciones y riesgos
- Robust Intelligence: protege los modelos de lenguaje grandes (LLM) en producción al validar entradas y salidas en tiempo real con su oferta de AI Firewall.
- Potion: regenera transmisiones de vídeo y audio utilizando modelos de visión y audio personalizados.

Figura 5: *Encuesta sobre el estado de la IA de Retool: las principales bases de datos vectoriales del sector*

Como reflejo de la popularidad de MongoDB entre los desarrolladores de IA, el proveedor de herramientas de software Retool concluyó en su encuesta sobre el estado de la IA que Atlas Vector Search de MongoDB:

1. Obtuvo el Net Promoter Score (NPS) más alto de todas las bases de datos vectoriales encuestadas.
2. Llegó a ser la segunda base de datos vectorial más utilizada a los pocos meses de su lanzamiento, lo que la colocó por delante de las soluciones alternativas que existen desde hace años.

¡Atlas Vector Search es robusta, rentable y superrápida!

Saravana Kumar, CEO de Kovai, habla sobre el desarrollo del asistente de inteligencia artificial de su empresa.

Cómo empezar

Ya sea que esté construyendo la próxima gran novedad en una startup o empresa, MongoDB Atlas le permite:

- Acelerar la creación de sus aplicaciones enriquecidas con IA generativa que se basan en la verdad de los datos operativos.
- Simplificar su pila tecnológica aprovechando una única plataforma que le permite a su aplicación almacenar datos operativos e incrustaciones vectoriales en el mismo lugar, reaccionar a los cambios en los datos fuente con funciones serverless y buscar en múltiples modalidades de datos y así mejorar la relevancia y la precisión en las aplicaciones.
- Haga evolucionar con facilidad sus aplicaciones enriquecidas con IA generativa gracias a la flexibilidad del modelo de documento, y mantenga al mismo tiempo una experiencia de desarrollador sencilla y elegante.
- Integre a la perfección los principales servicios y sistemas de IA, como los hiperescaladores y los LLM y marcos de código abierto, para seguir siendo competitivo en mercados dinámicos.
- Cree aplicaciones enriquecidas con la IA generativa en una base de datos operativa de alto rendimiento y altamente escalable con una década de validación en una amplia variedad de casos de uso de IA.

Puede obtener más información sobre la creación de aplicaciones basadas en IA con MongoDB visitando nuestro centro de recursos AI/ML.

La mejor manera para que los desarrolladores comiencen a crear una cuenta en MongoDB Atlas. Desde allí, pueden crear una instancia gratuita de MongoDB con la base de datos de Atlas, Atlas Vector Search y Atlas Search, cargar sus propios datos o nuestros conjuntos de datos, y explorar qué puede hacer con la plataforma.

Puerto seguro

El desarrollo, lanzamiento y calendario de cualquier característica o funcionalidad descrita para nuestros productos queda a nuestro exclusivo criterio. Esta información tiene como único objetivo describir la dirección general de nuestro producto y no se debe confiar en ella para tomar una decisión de compra ni es un compromiso, promesa u obligación legal de entregar ningún material, código o funcionalidad.

Estados Unidos 866-237-8815 • INTL + 1-650-440-4474 • info@mongodb.com.