Incorpora l'IA generativa e la ricerca avanzata nelle tue app con MongoDB

Creazione di applicazioni basate sull'IA

Dicembre 2023

USA 866-237-8815 • INTERNAZIONALE + 1-650-440-4474 • info@mongodb.com.

Sommario

Introduzione **3**

Il contesto è tutto **3**

L'ascesa dei vettori e la ricerca per analogia **4**

Ricerca vettoriale e flusso di lavoro LLM **5**

La promessa e la realtà di un ecosistema IA dinamico **6**

Una piattaforma dati per sviluppatori: il modo intelligente per creare applicazioni intelligenti **7**

Mostrare, non raccontare. App generative potenziate dall'IA generativa su una piattaforma di dati per sviluppatori **9**

Chatbot e QA per il self-service del cliente 10

Ricerca e consigli avanzati per l'e-commerce 13

Analisi e generazione di rich media (multimodali) 15

MongoDB Vector Search in azione **15**

Per iniziare **17**

Introduzione

Mai prima d'ora l'introduzione di una nuova tecnologia ha attirato così rapidamente l'attenzione di imprese, governi e consumatori. L'arrivo di ChatGPT nel novembre 2022 ha mostrato il potenziale dell'IA generativa basata su Large Language Models (LLM) nell'affrontare una vasta gamma di nuovi casi d'uso. Questi casi d'uso erano precedentemente inimmaginabili con l'informatica convenzionale e l'IA analitica (ora talvolta descritta come IA "tradizionale" o "classica").

Sembra che bastino alcuni prompt ben congegnati per automatizzare un'intera serie di elementi. Genera testo, immagini, audio, video e codice di programmazione di qualità professionale. Supporta meglio i clienti. Puoi arrivare alla modellazione del cambiamento climatico, alla scoperta di nuovi farmaci o alla progettazione di nuovi materiali, alla previsione dei movimenti dei mercati finanziari e molto, molto altro ancora.

Da un giorno all'altro nell'ordine del giorno di tutti i consigli d'amministrazione è emersa una domanda: *"come possiamo utilizzare l'IA generativa per rivoluzionare i nostri mercati senza esserne travolti"?*

Tuttavia, i leader tecnologici hanno rapidamente riconosciuto che, oltre ai potenziali vantaggi dell'IA generativa, esistono anche rischi derivanti dall'immaturità della tecnologia. Non possono semplicemente accantonare anni di migliori pratiche operative e conoscenze istituzionali. Devono invece assicurarsi che sia i sistemi esistenti che le nuove applicazioni in sviluppo siano in grado di sfruttare l'IA generativa in modi sicuri, affidabili e accurati.

In questo articolo discuteremo di come MongoDB può aiutarti a raggiungere gli obiettivi utilizzando i tuoi dati per alimentare nuove e avvincenti applicazioni ed esperienze basate sull'IA generativa.

Il contesto è tutto

Quando tutti hanno accesso ai modelli di IA generativa, la differenza deriva dall'accesso di questi modelli a una delle risorse aziendali più importanti: i dati. Alcuni di questi dati saranno di proprietà dell'organizzazione e altri saranno pubblici, ma più recenti di quelli utilizzati per addestrare i modelli di base originali. Insieme, questi dati offrono risposte che riflettono meglio la "verità fondamentale" di oggi.

Fornire ai modelli i propri dati è possibile grazie a un nuovo modello architetturale chiamato Retrieval-Augmented Generation o RAG. L'utilizzo di RAG offre agli sviluppatori una potente combinazione. Possono sfruttare le incredibili conoscenze e

capacità di ragionamento dei modelli di IA generativa pre-addestrati per scopi generali e alimentarli con dati specifici dell'azienda accurati e aggiornati.

I risultati sono output di IA generativa accurati, aggiornati, pertinenti e che sfruttano tutti i tuoi dati, indipendentemente dalla struttura. Le tue app basate sull'IA generativa servono meglio i clienti, aumentano la produttività dei dipendenti e superano la concorrenza. Gli sviluppatori possono ottenere tutti questi risultati senza doversi rivolgere a team specializzati di data science per addestrare o perfezionare i modelli, un processo complesso, dispendioso in termini di tempo e denaro.

L'utilizzo delle proprie fonti di dati è un elemento importante per far funzionare l'IA generativa per l'azienda. Ma da solo non è sufficiente. Come discuteremo più avanti nel documento, gli sviluppatori devono anche valutare come distribuire la propria applicazione attorno a un LLM informato con i controlli di sicurezza adeguati in atto e con la scala e le prestazioni che gli utenti si aspettano.

L'ascesa dei vettori e la ricerca per analogia

Per alimentare i modelli di IA con i nostri dati, dobbiamo prima trasformarli in vettori. Questi vettori forniscono codifiche numeriche multidimensionali dei dati che ne colgono modelli, relazioni e strutture. I vettori danno ai dati un significato semantico. Il calcolo della distanza tra i vettori fa sì che le applicazioni comprendano agevolmente le relazioni e le analogie tra i diversi oggetti dati. Ciò consente di utilizzare i dati in una gamma completamente nuova di applicazioni di cui parleremo di seguito.

I dati in qualsiasi formato digitale e di qualsiasi struttura, ad esempio testo, video, audio, immagini, codice, tabelle, possono essere trasformati in un vettore elaborandoli con un modello vettoriale adatto. Ad esempio, `text-embedding-ada-002` di OpenAI è uno dei modelli più popolari per la vettorializzazione di contenuti testuali. Il bello dei vettori è che i dati non strutturati e quindi completamente opachi per un computer possono ora assumere significato e struttura dedotti e rappresentati tramite questi incorporamenti. Ciò significa che possiamo iniziare a cercare ed elaborare dati non strutturati come con i dati aziendali strutturati. Considerando che oltre l'80% dei dati che creiamo ogni giorno non sono strutturati, iniziamo ad apprezzare quanto sia realmente trasformativa la ricerca vettoriale combinata con GenAI.

Come mostrato in Figura 1 di seguito, dopo che i nostri dati sono stati trasformati in vettori, vengono mantenuti e indicizzati in un archivio vettoriale come MongoDB Atlas Vector Search. Per recuperare vettori simili, l'archivio viene interrogato con un algoritmo di vicinanza più vicina (ANN) approssimativo per eseguire una ricerca K Nearest Neighbor(KNN) utilizzando un algoritmo come "Hierarchical Navigable Small Worlds" (HNSW).

L'esecuzione di query su questi vettori ci consente di eseguire operazioni con i dati che in precedenza potevamo realizzare solo con competenze e infrastrutture di data science costose. In primo luogo, possiamo estendere la ricerca e la scoperta di informazioni oltre la corrispondenza delle parole chiave alla ricerca semantica sensibile al contesto, in grado di dedurre significato e intento dal termine di ricerca di un utente. In secondo luogo, possiamo recuperare i dati, codificati come vettori, per fornire al modello di IA generativa il contesto necessario per generare output più affidabili e accurati. Questi output possono includere:

- Elaborazione del linguaggio naturale (NLP) per attività quali chatbot e risposte alle domande, riepilogo del testo e analisi del sentiment.
- Visione artificiale ed elaborazione audio per la classificazione delle immagini e il rilevamento di oggetti fino al riconoscimento vocale e alla traduzione.
- Generazione di contenuti, tra cui la creazione di documentazione basata su testo e pagine web ottimizzate per SEO, codice computer o la conversione di testo in un'immagine o un video.

Ricerca vettoriale e flusso di lavoro LLM

La Figura 1 riunisce il flusso di lavoro che consente il "Retrieval Augmented Generation" per un LLM.

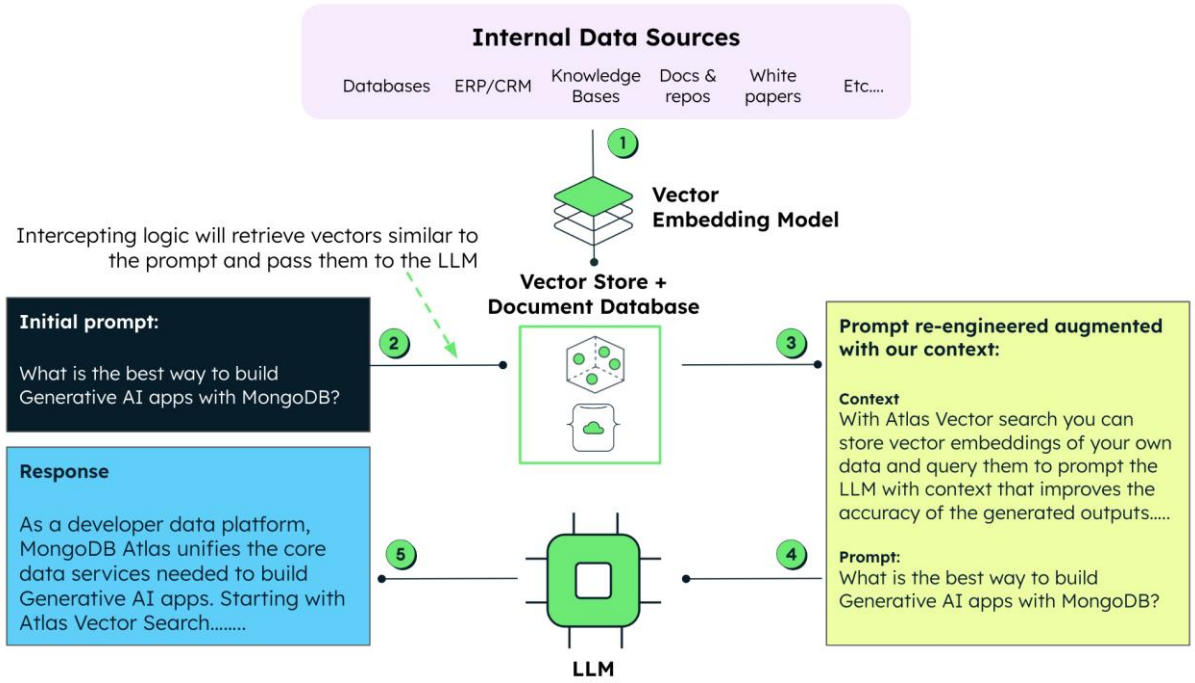

Figura 1: *Combinazione dinamica dei dati personalizzati con l'LLM per generare risultati affidabili e pertinenti*

In anticipo, i nostri dati vengono trasformati da un modello di incorporamento vettoriale e archiviati in un archivio vettoriale. Idealmente i metadati dei vettori e i dati grezzi "in blocchi" vengono archiviati insieme ai vettori stessi in un database di

documenti flessibile che memorizza anche i nostri normali dati applicativi.
Ciò consente alla nostra applicazione di interrogare i dati in diversi modi, migliorarne la pertinenza (ad es. ottenere un punteggio più alto per i dati più recenti) e fornisce memoria a lungo termine per il LLM. Le richieste all'LLM vengono intercettate dalla logica che recupera vettori simili dall'archivio vettori. Questi vengono quindi utilizzati per riprogettare il prompt iniziale. Il nuovo prompt viene inviato all'LLM che è in grado di utilizzare il contesto fornito per generare risposte accurate e di qualità superiore utilizzando dati più aggiornati.

Più avanti in questo documento troverai esempi che dimostrano il flusso di lavoro sopra riportato e mostrano come le funzionalità risultanti possono essere applicate a diverse applicazioni.

La promessa e la realtà di un ecosistema IA dinamico

Gli archivi vettoriali fanno parte di un ecosistema in rapida evoluzione di tecnologie abilitanti l'IA che si estende dalla creazione di incorporamenti, al prompt engineering, LLM, messa a punto dei modelli, orchestrazione, registrazione, automazione dell'infrastruttura e altro ancora.

All'interno di questo ecosistema vi sono molti progetti e fornitori interessanti e promettenti con cui lavorare. Alcuni mostrano "l'arte del possibile" tramite demo e prototipi. Ma il timore che devono affrontare i decision maker e gli sviluppatori aziendali è la facilità con cui questi prototipi possono essere adattati alle loro specifiche esigenze aziendali. E se alcune delle tecnologie più recenti possano davvero sostenere il carico di produzione con affidabilità, scalabilità e sicurezza, giorno dopo giorno, in qualsiasi ambiente operativo. Un'ulteriore considerazione riguarda il modo in cui integrare i database dell'organizzazione per inserire nel modello dati aziendali reali e concreti.

L'ecosistema dell'IA non esiste in isolamento. Tutte queste tecnologie devono essere integrate nelle applicazioni del mondo reale per essere veramente utili al business.
Ad esempio, gli store vettoriali sono essenziali per abilitare l'IA generativa sensibile al contesto e la ricerca semantica. Ma queste sono solo una parte di un'applicazione più ampia che deve gestire anche dati aziendali regolari e non vettorializzati.

Questi dati possono essere costituiti da qualsiasi elemento: record dei clienti, ordini e inventario, negoziazioni e transazioni, preventivi, coordinate geospaziali, dettagli e prezzi dei prodotti, misurazioni di time-series e letture dei sensori, clickstream e social feed, descrizioni di testo e altro ancora.

Tutti questi dati devono essere interrogati per potenziare la funzionalità dell'applicazione. Non solo per recuperare i nearest neighbors approssimativi tra i vettori, ma anche per eseguire operazioni consuete, come il recupero di record specifici, la gestione di una serie di aggiornamenti ai dati e l'esecuzione di aggregazioni e trasformazioni sofisticate che supportano l'elaborazione analitica. Queste query alimentano le funzionalità delle applicazioni al di fuori di qualsiasi caso d'uso dell'IA generativa. Ma diventano ancora più importanti quando possiamo usarli insieme ai prompt contestuali dei nostri modelli, migliorando l'accuratezza e la pertinenza degli output del modello di IA generativa.

Oltre a lavorare con i dati delle nostre applicazioni e gli incorporamenti di vettori, dobbiamo eseguire attività non funzionali: rispettare gli SLA in termini di tempi di attività, prestazioni e scalabilità, integrare nuove funzionalità, proteggere ed eseguire il backup dei dati e controllarli. Alcune di queste attività possono sembrare noiose. Questo finché non si raggiunge il risultato. E all'improvviso non è poi così noioso.....

Riunire le tecnologie per alimentare nuove esperienze basate sull'IA e riversarle nelle applicazioni rischia di creare una proliferazione di prodotti specifici e complessità che genera un enorme sovraccarico sui team. Tutte queste sfide si sommano a esperienze frammentate e inefficienti per gli sviluppatori, a un'ampia serie di modelli operativi e di sicurezza da gestire, a un'enorme mole di lavoro di gestione e integrazione dei dati e a un'estesa duplicazione dei dati. Tutto ciò rallenta la velocità di immissione sul mercato delle nuove esperienze basate sull'IA, aumentando al contempo costi e rischi.

L'utilizzo di una piattaforma dati per sviluppatori basata su MongoDB Atlas ti offre un modo migliore.

Una piattaforma dati per sviluppatori: il modo intelligente per creare applicazioni intelligenti

La piattaforma di dati per sviluppatori di MongoDB, basata su MongoDB Atlas, unifica i servizi di dati di IA operativi, analitici e generativa per semplificare la creazione di applicazioni intelligenti. Comunque tu stia sfruttando l'IA, dall'addestramento e dalla gestione dei tuoi modelli di machine learning all'integrazione della più recente IA generativa nelle tue app, Atlas è una parte fondamentale del tuo stack. Dal prototipo alla produzione, con Atlas puoi assicurarti che le tue app siano fondate sulla verità con i dati operativi più aggiornati, soddisfacendo al contempo la scalabilità, la sicurezza e le prestazioni che gli utenti si attendono.

Al centro di MongoDB Atlas c'è il suo document data model flessibile e l'API di query nativa per gli sviluppatori. Insieme, consentono agli sviluppatori di accelerare

notevolmente la velocità di innovazione, superando la concorrenza e sfruttando le nuove opportunità di mercato offerte dall'IA generativa.

I documenti sono il modo migliore per gli sviluppatori di lavorare con i dati, perché corrispondono agli oggetti del codice, rendendoli intuitivi e facili da interpretare. I documenti possono modellare dati di qualsiasi struttura, dalla grande varietà di dati applicativi regolari di cui abbiamo discusso in precedenza ai vettori composti da diverse migliaia di dimensioni. Tutte queste strutture possono essere modificate in qualsiasi momento per supportare l'aggiunta di nuovi tipi di dati e funzionalità applicative. I documenti offrono la flessibilità per razionalizzare e sfruttare quei dati in modi che i tradizionali modelli di dati tabulari dei relational database non riescono a fare.

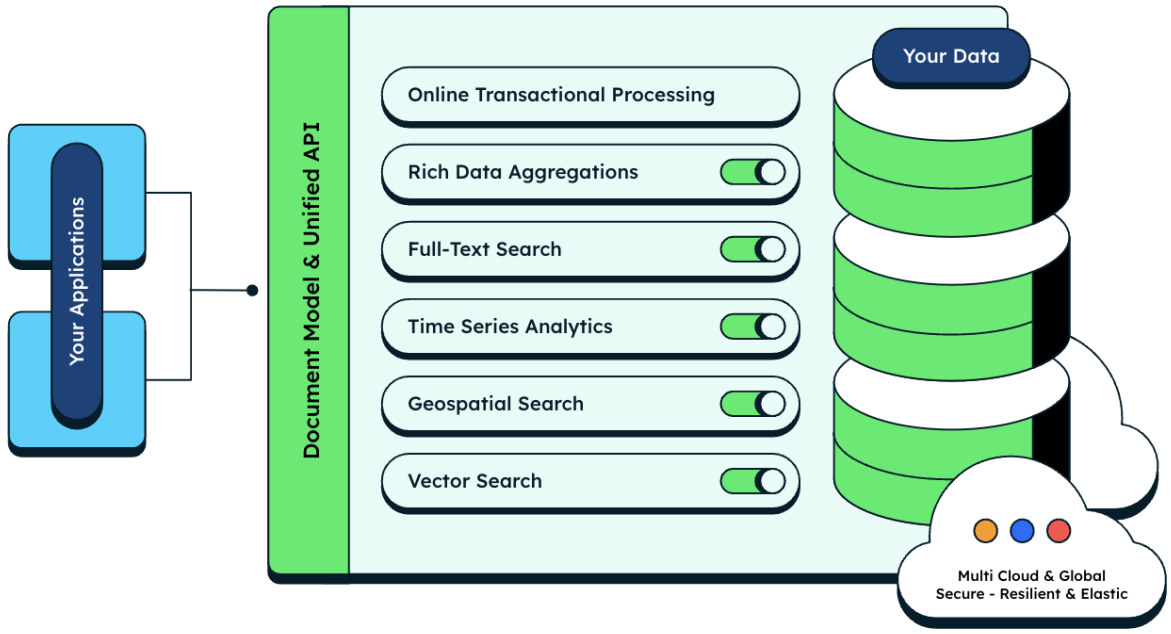

Figura 2: *MongoDB Atlas integra i servizi dati necessari per introdurre l'IA nelle tue applicazioni*

Insieme al modello di documento, l'API MongoDB Query offre agli sviluppatori un modo unificato e coerente di lavorare con i dati su qualsiasi servizio di dati. Dalle semplici operazioni CRUD alla ricerca per somiglianza di parole chiave e vettori fino a sofisticate pipeline di aggregazione per analisi ed elaborazione di stream, l'API MongoDB Query offre agli sviluppatori la flessibilità di interrogare ed elaborare i dati secondo i requisiti dell'applicazione. Nel contesto di GenAI, questo offre modi estremamente flessibili e potenti per definire filtri aggiuntivi sulle query basate su vettori, ad esempio:

- Combinazione con i metadati per il filtraggio: "Trovami i contenuti corrispondenti alla query dell'utente ma solo i contenuti pubblicati negli anni X, Y e Z".

- Combinazione con aggregazioni: "Trovami tutte le immagini simili all'immagine della query e raggruppale in base all'ID del fotografo".
- Combinazione con la ricerca geospaziale: "Trovami annunci immobiliari per case simili alla casa in questa fotografia che si trova entro N chilometri dalla mia posizione".

Nessun altro database è in grado di offrire un'ampia gamma di funzionalità di query in un'unica esperienza di query unificata. Ciò consente agli sviluppatori di creare funzionalità per l'utente finale più facilmente e in modo meno complesso.
Gli sviluppatori non devono più unire manualmente i risultati delle query da più database, un processo complesso, soggetto a errori, costoso e lento. Al contempo, mantiene anche compatta e agile l'impronta tecnologica.

"MongoDB archiviava già i metadati sugli artefatti nel nostro sistema. Con l'introduzione di Atlas Vector Search, ora disponiamo di un database completo di metadati vettoriali testato sul campo da oltre un decennio e che risolve le nostre complesse esigenze di recupero. Non è necessario distribuire un nuovo database da gestire e apprendere. I nostri vettori e i metadati degli artefatti possono essere archiviati uno accanto all'altro".

Pierce Lamb, Senior Software Engineer del team Data and Machine Learning di VISO TRUST.

Mostrare, non raccontare. App generative potenziate dall'IA generativa su una piattaforma di dati per sviluppatori

Ci concentreremo su tre casi d'uso popolari per mostrare come gli sviluppatori utilizzano MongoDB Atlas per creare app arricchite con IA:

- Chatbot e domande e risposte (Q-A) per il self-service dei clienti.
- Ricerca e-commerce avanzata e consigli per gli utenti.
- Analisi e generazione di rich media (multimodali).

Ognuno di questi esempi si basa sull'IA generativa e sulla ricerca semantica avanzata per creare esperienze utente straordinarie e consentire funzionalità che in precedenza erano fuori dalla portata della maggior parte delle organizzazioni. Per essere veramente trasformativi, tuttavia, questi miglioramenti dell'IA devono essere offerti

come parte di un'applicazione più ampia che di per sé alimenta funzionalità aziendali critiche.

Esamineremo a turno ciascun caso d'uso, mostrando un modello di progettazione architetturale che lo supporta, insieme alle funzionalità pertinenti fornite da MongoDB Atlas.

Chatbot e QA per il self-service del cliente

MongoDB è al centro di molte applicazioni di assistenza clienti. Questo perché il modello di dati flessibile di MongoDB semplifica la creazione di un'unica visione a 360 gradi del cliente. Lo fa acquisendo dinamicamente dati dei clienti diversi e in rapida evoluzione dalla miriade di sistemi di origine backend isolati tipici della maggior parte delle organizzazioni. La visualizzazione unica e consolidata dei clienti in tempo reale basata su MongoDB è quindi la piattaforma ideale sulla quale possiamo addestrare e fornire chatbot e funzionalità di assistenza Q-A per il self-service dei clienti.

Nel nostro esempio mostrato in Figura 2, il database clienti archiviato in MongoDB viene esportato come file JSON in un modello di incorporamento che raggruppa i dati (utilizzando strumenti come LangChain o LlamaIndex) e crea i vettori. Anche altre fonti di dati interne, come knowledge base e documentazione, si possono vettorializzare per l'utilizzo nell'app. I dati vengono quindi importati nuovamente nel database MongoDB.

Dobbiamo assicurarci che i nostri vettori siano costantemente aggiornati con i dati più recenti dei clienti, quindi utilizziamo Atlas Trigger per monitorare eventuali modifiche dei dati nella nostra visualizzazione unica. Non appena vengono inseriti nuovi record cliente o i record esistenti vengono aggiornati nel database, Atlas Triggers chiama l'API del modello di incorporamento per generare i vettori corrispondenti e caricarli nuovamente in Atlas.

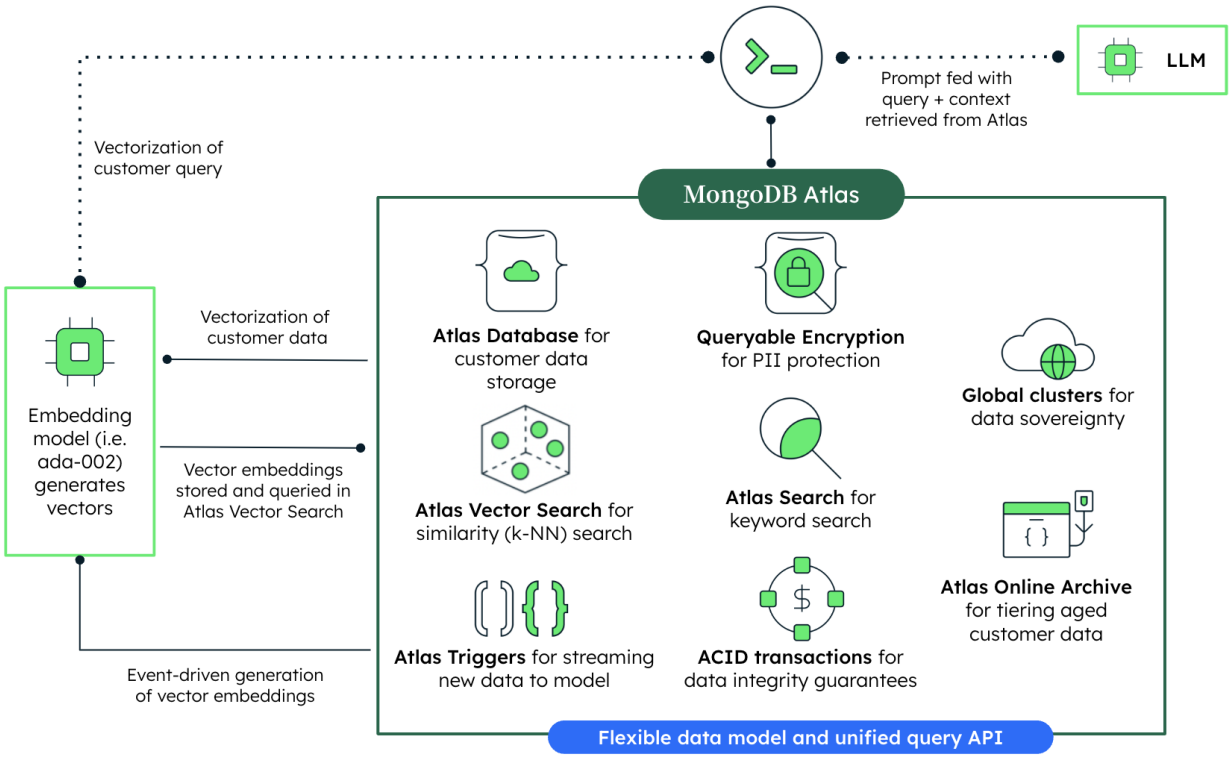

Figura 3: *funzionalità di IA generativa del Chatbot e domande e risposte integrate in un'applicazione self-service per i clienti basata su MongoDB Atlas*

Utilizzando Atlas, gli sviluppatori traggono vantaggio dal modello dati flessibile di MongoDB. Possono archiviare i dati, i metadati e i blocchi dei clienti di origine insieme ai vettori, tutti sincronizzati e affiancati in un unico livello di archiviazione e accessibili da un'unica API di query e driver.

Le query possono filtrare in modo efficiente i dati utilizzando i vettori indicizzati insieme agli indici delle parole chiave dei campi regolari nei documenti. Questa integrazione significa che l'app può supportare una gamma molto più ampia di funzionalità utente con un carico di lavoro per gli sviluppatori inferiore:

- Atlas Vector Search restituisce i documenti corrispondenti eseguendo una ricerca di analogia sui dati di incorporamento indicizzati. Per ridurre il rischio di restituire dati obsoleti, le nostre query possono utilizzare i metadati di un vettore, ad esempio la "data di creazione" memorizzata nel database Atlas, per filtrare i contenuti meno recenti.
- Atlas Search restituisce i risultati in base alle parole chiave corrispondenti nell'origine e ai dati dei clienti suddivisi in blocchi. Utilizza funzionalità come la ricerca fuzzy per correggere gli errori di battitura nell'input dell'utente e il completamento automatico per fornire termini di ricerca suggeriti. Utilizza inoltre l'intersezione degli indici per gestire in modo efficiente query ad hoc complesse sui dati dei clienti.

Le query al database Atlas, Vector Search e Atlas Search utilizzano tutte la stessa interfaccia di query e lo stesso driver, semplificando enormemente il flusso di lavoro

dello sviluppatore. I dati recuperati da MongoDB Atlas vengono forniti come contesto per aumentare le richieste al LLM, consentendogli di generare risposte pertinenti alle chat e alle domande. Il contesto e le istruzioni, insieme a tutte le fasi di ragionamento associate utilizzate per rispondere a domande complesse, vengono mantenuti in Atlas, fornendo al LLM una memoria a lungo termine e migliorando continuamente i suoi risultati.

I dati dei clienti sono tra i più preziosi gestiti da un'organizzazione. Sebbene l'IA generativa ci aiuti a innovare la modalità con la quale serviamo i nostri clienti, proteggere i loro dati rimane fondamentale. Atlas offre una gamma di funzionalità per aiutarci a raggiungere l'obiettivo, consentendo agli sviluppatori di concentrarsi sulle funzionalità basate sull'IA:

- Infrastruttura convergente che supporta l'archiviazione dei dati, le query e l'analisi, la ricerca per parole chiave e la ricerca vettoriale. Questa unificazione basata su un'unica API e un modello di dati riduce drasticamente il numero di parti mobili che gli sviluppatori devono integrare e utilizzare.
- Queryable Encryption è una novità assoluta nel settore della protezione dei dati dei clienti. I driver MongoDB crittografano i campi di dati sensibili lato client e il database li utilizza solo come dati crittografati completamente randomizzati. Anche con i dati crittografati, le applicazioni possono comunque eseguire query espressive su di essi senza dover mai decrittografare i dati nel database. Tieni presente che in genere solo i campi che contengono i dati più sensibili che identificano in modo univoco un individuo, come un SSN, sono protetti con Queryable Encryption. La ricerca può quindi essere eseguita sui restanti campi di testo in chiaro.
- Transazioni ACID multi-documento nel database Atlas garantiscono l'integrità dei dati dei clienti ogni volta che si accede e si modifica dall'applicazione.
- Con Atlas Global Clusters, i dati dei clienti possono essere bloccati nella loro regione di residenza, in conformità con le moderne normative sulla sovranità dei dati.
- La gestione completa del ciclo di vita dei dati è fornita da Atlas Online Archive. Il servizio consente di spostare automaticamente i dati dei clienti obsoleti dai database attivi all'archiviazione di oggetti nel cloud a basso costo, mantenendo i dati accessibili per le query. Questo è importante per i dati dei clienti gestiti all'interno di app che operano in settori regolamentati, dove i dati devono essere conservati e accessibili per diversi anni.
- I dati dei clienti sono protetti da danneggiamento e ransomware con backup e ripristino point-in-time.

Atlas è completamente gestito per te sui principali cloud hyperscale, con il supporto di uno SLA con tempi di attività del 99,995%.

Ricerca e consigli avanzati per l'e-commerce

I cataloghi di prodotti e-commerce sono un caso d'uso comune per MongoDB:

- La varietà dei diversi prodotti e i loro attributi si adattano naturalmente al data model doc flessibile di MongoDB.
- L'architettura distribuita di Atlas con scalabilità elastica consente agli sviluppatori di dimensionare e regolare dinamicamente la capacità del database in risposta alla domanda delle applicazioni (ad esempio per la stagionalità degli acquisti e le promozioni delle vendite).
- Con Atlas Search, le funzionalità di corrispondenza delle parole chiave come la ricerca fuzzy, il completamento automatico, la ricerca sfaccettata, l'evidenziazione e il punteggio personalizzato consentono agli acquirenti di sfogliare e navigare rapidamente nel catalogo dei prodotti, aumentando le percentuali di clic (CTR) e le conversioni di acquisto.

Tuttavia, la ricerca per parole chiave si basa sulla corrispondenza di parole specifiche nei campi di testo indicizzati per restituire risultati pertinenti. Senza una mappatura dei sinonimi estesa e laboriosa (ad esempio, mappando le biciclette con il ciclismo o le scarpe da ginnastica con le scarpe da ginnastica), gli utenti si sentiranno rapidamente frustrati quando le loro query di ricerca non riescono a restituire prodotti pertinenti. Questa frustrazione determina una perdita di vendite e in un danno alla reputazione del marchio.

Un'ulteriore sfida consiste nel fornire consigli agli utenti. Gli sviluppatori devono scrivere motori complessi basati su regole o rivolgersi a risorse di data science specializzate e scarse. In genere, i dati devono essere prima elaborati (estrazione, trasformazione, caricamento) dal database operativo in un data warehouse o in un data lake offline. Solo a questo punto i tradizionali modelli analitici di IA possono generare una serie di raccomandazioni che devono poi essere caricate nel database operativo. Il processo è complesso, costoso e genera consigli che sono immediatamente obsoleti in quanto non riflettono il comportamento di navigazione o gli acquisti dell'utente più recenti.

Migliorare il nostro catalogo di prodotti con i vettori elimina questi problemi.
I vettori forniscono un significato semantico ai prodotti del nostro catalogo, agevolando la comprensione delle analogie e delle relazioni tra i prodotti. Ciò consente ai rivenditori di presentare agli utenti prodotti pertinenti e correlati con impegno, complessità e costi inferiori. I termini di ricerca comuni possono essere memorizzati nella cache di MongoDB Atlas, fornendo risultati pertinenti agli utenti, più velocemente.

L'estensione della vettorizzazione ai dati dei clienti, come dimostrato in precedenza nell'app self-service per i clienti, ci consente di creare consigli ancora più sofisticati combinando la ricerca di analogie tra prodotto e cliente per ottimizzare i suggerimenti.

La Figura 4 mostra un modello di progettazione di alto livello per ricerca e consigli avanzati. La creazione e il mantenimento delle nostre integrazioni vettoriali seguono lo stesso flusso di lavoro descritto in precedenza per i chatbot e la Q-A nella nostra applicazione self-service per i clienti.

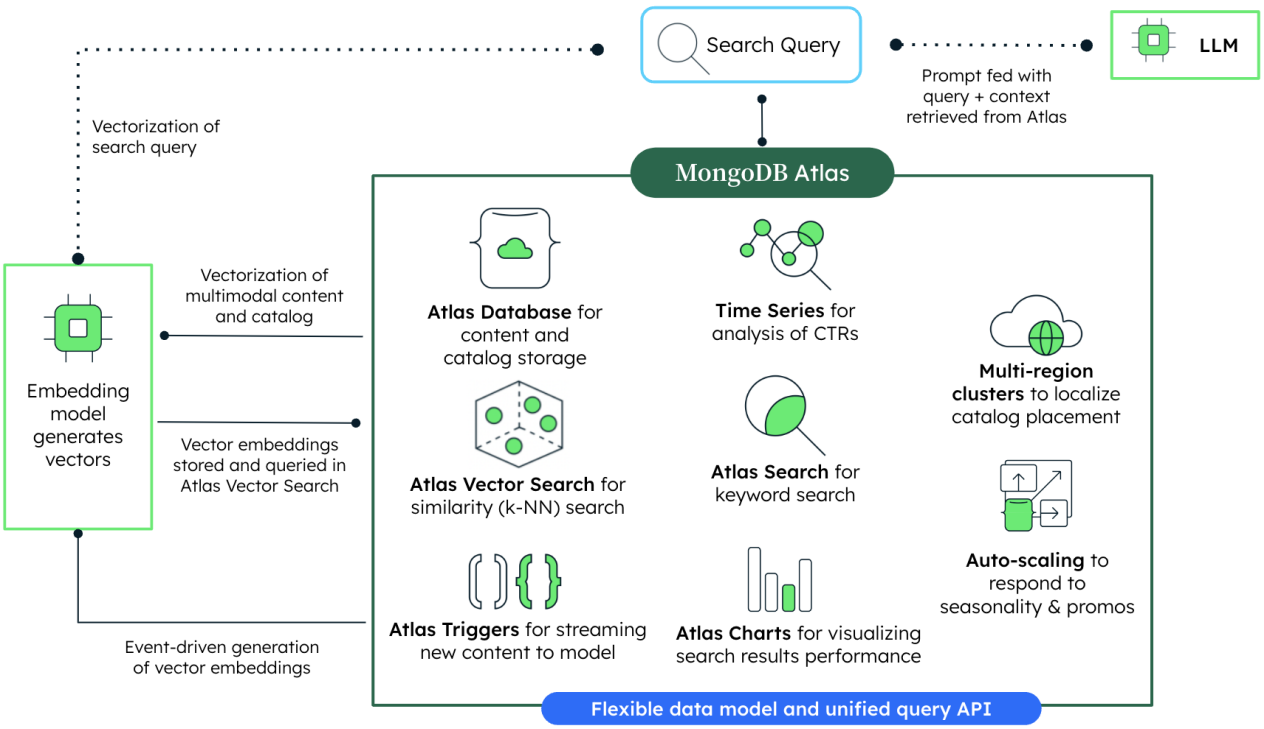

Figura 4: *la ricerca semantica avanzata nel nostro catalogo di prodotti aumenta le conversioni di vendita e l'upselling*

È facile osservare come la ricerca vettoriale migliori notevolmente la ricerca e i consigli dei prodotti. L'integrazione di un LLM migliora ulteriormente questa esperienza. Ora i clienti possono porre domande in tempo reale e ottenere risposte immediate sui prodotti che stanno valutando e questo contribuisce ad accelerare il ciclo di acquisto.

I rivenditori possono inoltre utilizzare l'LLM per una serie di compiti che in precedenza sarebbero stati laboriosi, potendo così sviluppare modi ancora più creativi per coinvolgere i clienti. Ad esempio, il LLM può essere utilizzato per generare diverse varianti del testo del prodotto e delle parole chiave SEO per fare A/B testing e quantificare così quali generano conversioni più elevate. Il LLM potrebbe essere utilizzato per riassumere più recensioni degli utenti e dedurre il sentiment, aiutando a sintetizzare il feedback che informa le roadmap dei prodotti.

Le organizzazioni possono utilizzare Atlas per gestire l'intero ciclo di vita dell'e-commerce. Oltre a utilizzare l'IA per rendere la nostra esperienza di ricerca più intelligente e predittiva, gli imprenditori possono monitorare le percentuali di clic degli utenti e le conversioni di vendita dai risultati di ricerca. Le collection di time-series

possono inserire e archiviare in modo efficiente flussi di clic voluminosi e ad alta velocità dalle sessioni utente, rendendo disponibili tali dati per l'analisi per misurare le prestazioni di ricerca, incluse le visualizzazioni in tempo reale dei risultati tramite Atlas Charts. Con queste informazioni, i rivenditori possono affinare e ottimizzare costantemente i dati dei prodotti e il punteggio di pertinenza per massimizzare le vendite dal sito di e-commerce.

Analisi e generazione di rich media (multimodali)

La ricerca testuale consueta funziona bene con la ricerca di parole chiave tradizionale. Tuttavia, l'utilizzo di risorse multimediali più avanzate (a volte dette multimodali) come immagini, voce e video richiede tecnologie e competenze di data science molto complesse. Fino a ora.

Come accennato in precedenza, qualsiasi contenuto digitale può essere vettorializzato con il modello di vettore appropriato. Gli hub di IA come Hugging Face e quelli degli hyperscaler cloud forniscono una vasta gamma di modelli ottimizzati per diverse modalità di contenuto. Gli incorporamenti di questi modelli possono essere memorizzati in Atlas Vector Search per alimentare un'intera gamma di nuove funzionalità. Come discusso in precedenza, la generazione di immagini dal testo, la trascrizione di video per il riconoscimento vocale e l'analisi del sentiment, la classificazione delle immagini e il rilevamento di oggetti sono solo alcuni esempi di ciò che è possibile. È possibile combinare vettori di diversi media, ad esempio confrontando un testo e l'incorporamento di un'immagine per verificare se una determinata frase descrive accuratamente un'immagine.

Questa funzionalità multimodale può essere utilizzata in una serie di casi d'uso. Ad esempio arricchendo i cataloghi di prodotti come quelli descritti sopra o migliorando la scoperta mediante l'analisi di immagini e video. Possono essere utilizzati per semplificare i processi di progettazione, produzione e pubblicazione o per creare nuove classi di applicazioni in domini quali sicurezza e sorveglianza o realtà aumentata (AR).

Il modello di progettazione dell'architettura e le funzionalità MongoDB Atlas descritte per la ricerca avanzata di e-commerce e le raccomandazioni di cui sopra si applicano ugualmente alla generazione di contenuti multimodali.

MongoDB Vector Search in azione

MongoDB è già stato ampiamente adottato per i casi d'uso tradizionali dell'IA. Continental ha scelto MongoDB per la piattaforma di feature engineering della sua iniziativa di guida autonoma Vision Zero. SiaBosch che Telefonica utilizzano MongoDB nelle loro piattaforme IoT potenziate dall'IA. Kronos negozia miliardi di dollari di

criptovalute ogni giorno utilizzando modelli di ML configurati e costruiti con i dati di MongoDB. Iguazio utilizza MongoDB come livello di persistenza per la sua piattaforma di data science e MLOps, mentre H2O.ai e Featureform supportano MongoDB come archivi di funzionalità nelle rispettive piattaforme.

Basandosi su queste fondamenta, MongoDB Atlas è già utilizzato oggi in una serie di applicazioni che stanno ampliando i confini di ciò che è possibile con l'IA generativa. Dai un'occhiata alla nostra pagina dei case study per saperne di più sulla vasta gamma di casi d'uso offerti da MongoDB Atlas. Ecco una selezione di esempi specifici:

- Ada: aiuta le aziende come Meta, ATT e Verizon a supportare meglio i clienti grazie all'automazione basata sull'IA e all'IA conversazionale.
- ExTrac: identifica e classifica i rischi fisici e digitali emergenti dall'analisi dei flussi di dati in tempo reale.
- Eni: sblocca i dati geologici e li rende fruibili per migliorare il processo decisionale e accelerare il percorso dell'azienda verso l'azzeramento delle emissioni nette.
- Inovaare: monitora, estrae e classifica continuamente i dati lungo tutto il ciclo di vita dell'assistenza sanitaria per la rendicontazione della conformità normativa, l'audit e la valutazione del rischio.
- Source Digital: ottiene una riduzione dei costi di 7 volte dopo la migrazione da PostgreSQL a MongDB Atlas per la sua piattaforma di rilevamento video.
- Catylex: estrae, classifica e analizza automaticamente i termini contrattuali per individuare diritti, obblighi e rischi
- Robust Intelligence: protegge i modelli linguistici di grandi dimensioni (LLM) in produzione convalidando input e output in tempo reale con la sua offerta di Firewall IA.
- Potion: rigenera i flussi video e audio utilizzando modelli visivi e audio personalizzati.

Figura 5: *Sondaggio sullo stato dell'IA di Retool: i principali database vettoriali del settore*

Riflettendo la popolarità di MongoDB tra gli sviluppatori di IA, il fornitore di strumenti software Retool ha concluso il suo Sondaggio sullo stato dell'IA secondo cui MongoDB Atlas Vector Search:

1. Ha il Net Promoter Score (NPS) più alto di tutti i database vettoriali oggetto del sondaggio.
2. Era diventato il secondo database vettoriale più utilizzato in pochi mesi dalla release, posizionandosi davanti a soluzioni alternative che esistevano da anni.

"Atlas Vector Search è solido, conveniente e incredibilmente veloce!"
Saravana Kumar, CEO di Kovai parla dello sviluppo dell'assistente IA della sua azienda.

Per iniziare

Che tu stia creando la prossima grande novità in una startup o in un'impresa, con MongoDB Atlas puoi:

- Accelera la creazione delle tue applicazioni arricchite dall'IA generativa basate sulle informazioni affidabili dei dati operativi.

- Semplificare lo stack tecnologico sfruttando un'unica piattaforma che consente all'app di archiviare i dati operativi e i vettori nella stessa posizione, reagire alle modifiche dei dati di origine con funzioni serverless e cercare tra più modalità di dati, migliorando la pertinenza e l'accuratezza nelle risposte generate dalle app.
- Far evolvere facilmente le tue app arricchite dall'IA generativa con la flessibilità del document model, mantenendo un'esperienza di sviluppo semplice ed elegante.
- Integra perfettamente i principali servizi e sistemi di IA, come hyperscaler, LLM e framework open source, per rimanere competitivo nei mercati dinamici.
- Sviluppare applicazioni arricchite dall'IA generativa su un database operativo ad alte prestazioni e altamente scalabile con dieci anni di convalida su un'ampia gamma di casi d'uso dell'IA.

Per saperne di più sulla creazione di app basate sull'IA con MongoDB, visita il nostro centro risorse AI/ML.

Il modo migliore per consentire agli sviluppatori di iniziare è creare un account su MongoDB Atlas. Da lì, possono creare un'istanza MongoDB gratuita con database Atlas, Atlas Vector Search e Atlas Search, caricare i propri dati o i nostri set di dati di esempio ed esplorare ciò che è possibile fare all'interno della piattaforma.

Safe Harbor

Lo sviluppo, il rilascio, e la tempistica di qualsiasi caratteristica o funzionalità ivi descritte per i nostri prodotti sono a nostra esclusiva discrezione. Le presenti informazioni hanno il solo scopo di delineare la nostra direzione generale in termini di prodotto e non dovranno essere invocate nel prendere una decisione di acquisto, né è questo un impegno, promessa o obbligo legale a consegnare qualsivoglia materiale, codice o funzionalità.