Intégration de l’IA générative et de la recherche avancée dans vos applications avec MongoDB

Créer des applications basées sur l'IA

Décembre 2023

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

Table des matières

Introduction

Jamais aucune nouvelle technologie n'avait attiré aussi rapidement l'attention des entreprises, des gouvernements et des consommateurs. L'arrivée de ChatGPT en novembre 2022 a montré le potentiel de l'IA générative alimentée par les grands modèles de langage (LLM) pour répondre à de nouveaux cas d'utilisation. Ces cas d'usage étaient auparavant inimaginables avec l'informatique conventionnelle et l'IA analytique (désormais parfois qualifiée d'IA « traditionnelle » ou « classique »).

Il semble qu'il suffise de quelques prompts bien conçus pour automatiser toute une série de choses. Générez du texte, des images, du son, des vidéos et du code de programmation de qualité professionnelle. Offrez un meilleur service client. De la modélisation du changement climatique à la découverte de nouveaux médicaments en passant par la conception de nouveaux matériaux et la prévision des mouvements des marchés financiers, de nombreuses possibilités s'offrent à vous.

Du jour au lendemain, une question était sur toutes les lèvres au sein des conseils d'administration : *« Comment utiliser l'IA générative pour perturber nos marchés tout en évitant d'être nous-mêmes perturbés ? »*

Toutefois, les leaders technologiques ont rapidement reconnu qu'à côté des avantages potentiels de l'IA générative, l'immaturité de la technologie présentait également des risques. Ils ne peuvent pas seulement abandonner des années de bonnes pratiques opérationnelles et de connaissances institutionnelles. Ils doivent au contraire s'assurer que leurs systèmes existants et les nouvelles applications en cours de développement sont capables de l'exploiter de manière sûre, fiable et précise.

Dans cet article, nous verrons comment MongoDB peut vous aider à atteindre ces objectifs tout en utilisant vos propres données pour alimenter de nouvelles applications et expériences convaincantes qui reposent sur l'IA générative.

Le contexte est essentiel

Lorsque tout le monde a accès aux modèles d'IA générative, votre « superpuissance » réside dans le fait de donner à ces modèles l'accès à l'un des actifs les plus importants de votre entreprise : vos données. Certaines de ces données sont privées et d'autres sont publiques, mais plus pertinentes que celles utilisées pour former les modèles de fondation. Ensemble, ces données fournissent des réponses qui reflètent mieux la réalité.

Fournir des modèles avec vos propres données est possible grâce à un nouveau modèle architectural appelé génération augmentée de récupération (RAG). Elle offre une puissante combinaison à vos développeurs. Ils peuvent exploiter les incroyables capacités de connaissance et de raisonnement des modèles généraux pré-entraînés et les alimenter avec des données précises et actualisées propres à l'entreprise.

Les résultats sont précis, à jour et pertinents. Ils exploitent toutes vos données, quelle que soit leur structure. Ainsi, vos applications répondent mieux aux besoins de vos clients, améliorent la productivité de vos employés et devancent vos concurrents. Vos développeurs peuvent obtenir tous ces résultats sans avoir à se tourner vers des data scientists pour entraîner ou affiner les modèles dans le cadre d'un processus complexe, chronophage et coûteux.

Utiliser vos propres sources de données est essentiel pour que l'IA générative s'adapte aux besoins de votre entreprise. Mais ce n'est pas suffisant. Comme expliqué dans cet article, les développeurs doivent également réfléchir à la manière de déployer leur application autour d'un LLM bien renseigné, avec les contrôles de sécurité adéquats, à l'échelle et avec les performances attendues par les utilisateurs.

L’essor des vecteurs et de la recherche de similarités

Pour alimenter les modèles d’IA avec nos propres données, nous devons d’abord les transformer en vector embeddings. Ces vecteurs fournissent des encodages numériques multidimensionnels de nos données qui saisissent leurs modèles, relations et structures. Les vector embeddings donnent une signification sémantique à nos données. Le calcul de la distance entre les vecteurs permet à nos applications de comprendre facilement les relations et les similitudes entre les différents objets de données. Cela ouvre nos données à une toute nouvelle gamme d'applications que nous aborderons ci-dessous.

Les données dans n’importe quel format numérique et de n’importe quelle structure (texte, vidéo, audio, images, code, tableaux) peuvent être transformées en vecteur en les traitant avec un modèle de vector embedding approprié. Par exemple, le modèle `text-embedding-ada-002` d'Open AI est l’un des modèles les plus populaires pour vectoriser le contenu textuel. L'intérêt des vector embeddings réside dans le fait que les données non structurées et donc complètement opaques pour un ordinateur peuvent désormais voir leur signification et leur structure déduites et représentées via ces intégrations. Cela signifie que nous pouvons commencer à rechercher et à calculer des données non structurées de la même manière que nous l'avons toujours fait avec des données d’entreprise structurées. Si l’on considère que plus de 80 % des données que nous créons chaque jour ne sont pas structurées, nous commençons à

comprendre à quel point la recherche vectorielle combinée à l'IA générative transforme la réalité.

Comme le montre le schéma n° 1 ci-dessous, une fois que nos données ont été transformées en vector embeddings, elles sont conservées et indexées dans un vector store tel que MongoDB Atlas Vector Search. Pour récupérer des vecteurs similaires, le store est interrogé à l'aide d'un algorithme ANN (« Approximate Nearest Neighbor ») afin d'effectuer une recherche KNN (« K Nearest Neighbor ») à l'aide d'un algorithme tel qu'HNSW (« Hierarchical Navigable Small Worlds » (HNSW).

Effectuer des requêtes sur ces vecteurs nous permet de faire des choses avec les données que nous ne pouvions auparavant accomplir qu'avec des compétences et une infrastructure coûteuses en science des données. Tout d'abord, nous pouvons étendre la recherche et la découverte d'informations au-delà de la correspondance de mots-clés à la recherche sémantique contextuelle, qui est capable de déduire le sens et l'intention à partir du terme de recherche d'un utilisateur. Deuxièmement, nous pouvons récupérer nos propres données, encodées sous forme de vecteurs, afin de fournir au modèle GenAI le contexte nécessaire pour générer des sorties plus fiables et plus précis. Voici des exemples :

- Traitement du langage naturel (TAL) pour des tâches telles que les chatbots et les réponses aux questions, la synthèse de texte et l'analyse des sentiments.
- Vision par ordinateur et traitement audio pour la classification d'images et la détection d'objets jusqu'à la reconnaissance vocale et la traduction.
- Génération de contenu, y compris la création de documents textuels et de pages web optimisées pour le référencement, de code informatique ou la conversion de texte en image ou vidéo.

Recherche vectorielle et flux de travail LLM

Le schéma n° 1 présente le flux de travail permettant la « génération augmentée de récupération » pour un LLM.

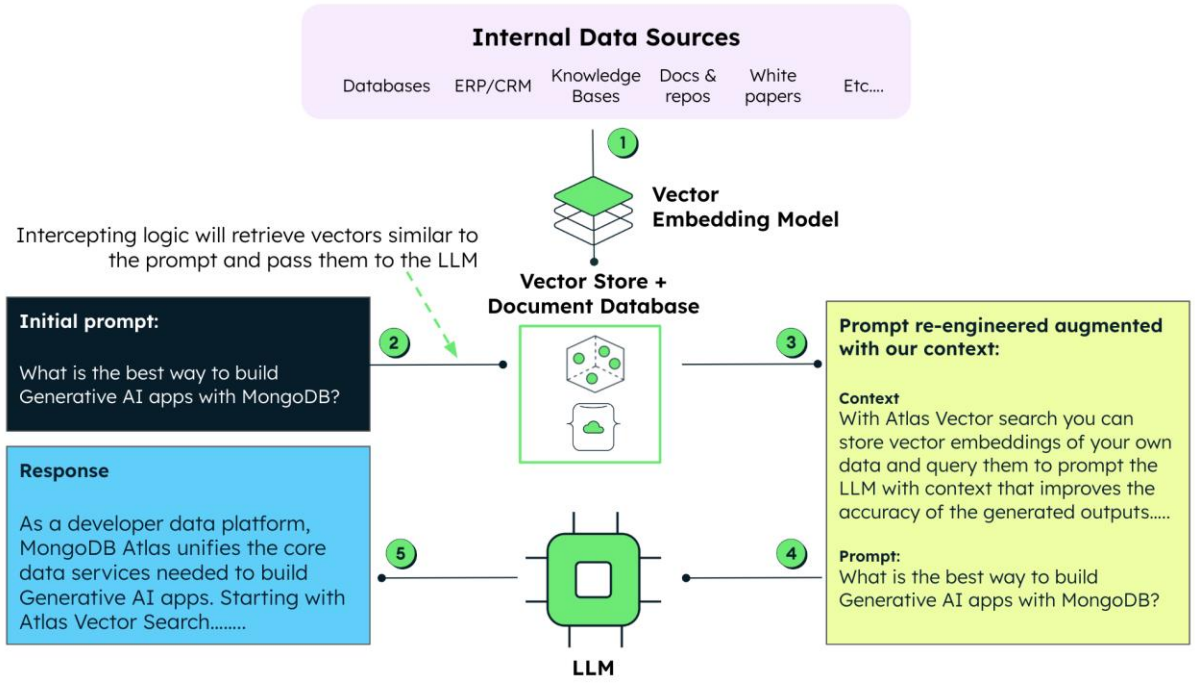

Schéma n° 1 : *combinaison dynamique de vos données personnalisées avec le LLM pour générer des résultats fiables et pertinents*

Nos données sont préalablement transformées par un modèle de vector embedding et stockées dans un vector store. Idéalement, les métadonnées des vecteurs et les données brutes « morcelées » sont stockées avec les vecteurs eux-mêmes dans une base de données documentaire flexible qui stocke également les données régulières de notre application. Cela permet à notre application d'interroger les données de plusieurs manières, d'améliorer leur pertinence (par exemple, en attribuant un score plus élevé aux données les plus récentes) et de fournir une mémoire à long terme pour le LLM. Les invites adressées aux LLM sont interceptées par une logique qui récupère des vecteurs similaires dans le vector store. Ces données sont ensuite utilisées pour remanier l'invite initiale. La nouvelle invite est envoyée au LLM qui peut utiliser le contexte fourni pour générer des réponses plus précises et de meilleure qualité à l'aide de données plus récentes.

Dans la suite de cet article, vous trouverez des exemples qui illustrent le flux de travail ci-dessus et montrent comment les capacités qui en résultent peuvent être appliquées à différentes catégories d'applications.

La promesse et la réalité d'un écosystème d'IA dynamique

Les vector stores font partie d'un écosystème de technologies d'IA en pleine expansion, allant de la création d'intégrations à l'ingénierie d'invite, en passant par les

LLM, l'affinement des modèles, l'orchestration, la journalisation, l'automatisation de l'infrastructure, etc.

Au sein de cet écosystème, il existe une multitude de projets et de fournisseurs intéressants et prometteurs avec lesquels travailler. Certains montrent « l'art du possible » à travers des démonstrations et des prototypes. Mais la crainte des décideurs et des développeurs est de savoir si ces prototypes peuvent être facilement adaptés aux besoins spécifiques de leur entreprise. Et si certaines des technologies les plus récentes peuvent réellement soutenir la charge de production avec fiabilité, évolutivité et sécurité, jour après jour, dans n'importe quel environnement opérationnel. Ils se demandent également comment intégrer les bases de données de l'entreprise afin d'alimenter le modèle avec des données commerciales fiables.

L'écosystème de l'IA est interdépendant. Toutes ces technologies doivent être intégrées dans des applications réelles pour répondre aux besoins de l'entreprise. Par exemple, les vector stores sont essentiels pour pouvoir tirer parti de l'IA générative et la recherche sémantique contextuelles. Mais il ne s'agit là que d'une partie d'une application plus large qui doit également gérer des données commerciales régulières et non vectorisées.

Ces données peuvent être de toute nature : dossiers client, commandes et stocks, échanges et transactions, devis, coordonnées géospatiales, fiches produit et tarification, mesures de time-series et relevés de capteur, flux de clics et flux sociaux, descriptions textuelles, etc.

Toutes ces données doivent être requêtées pour alimenter les fonctionnalités de l'application. Il ne s'agit pas seulement de récupérer les voisins les plus proches entre les vecteurs, mais aussi d'effectuer des opérations régulières telles que la récupération d'enregistrements spécifiques, la gestion des nombreuses mises à jour des données et l'exécution d'agrégations et de transformations sophistiquées prenant en charge le traitement analytique. Ces requêtes alimentent les fonctionnalités des applications en dehors de tout cas d'utilisation de l'IA générative. Mais elles deviennent encore plus importantes lorsque nous pouvons les utiliser avec des invites en contexte pour nos modèles, améliorant ainsi la précision et la pertinence des résultats du modèle GenAI.

En plus de travailler avec les données de nos applications et les vector embeddings, nous devons réaliser des tâches non fonctionnelles : respecter les accords de niveau de service en matière de disponibilité, de performance et d'évolutivité, intégrer de nouvelles fonctionnalités, sécuriser et sauvegarder les données, et les auditer. Certaines de ces tâches peuvent sembler ennuyeuses. Jusqu'à ce qu'elles échouent. Et soudain, elles ne sont plus si ennuyeuses…

Réunir les technologies qui alimentent les nouvelles expériences basées sur l'IA et les intégrer dans vos applications risque de créer une prolifération de produits ponctuels

et une complexité qui impose une énorme charge de travail à vos équipes. Tous ces défis se traduisent par des expériences fragmentées et inefficaces pour les développeurs, une multitude de modèles opérationnels et de sécurité à gérer, un énorme travail de manipulation et d'intégration des données, et de nombreuses duplications des données. Tout cela ralentit la vitesse de mise sur le marché de vos nouvelles expériences basées sur l'IA, tout en augmentant vos coûts et vos risques.

L'utilisation d'une Developer Data Platform construite sur MongoDB Atlas vous offre une meilleure solution.

Une developer data platform : la meilleure solution pour créer des applications intelligentes

La developer data platform de MongoDB, construite sur MongoDB Atlas, unifie les services de données d'IA opérationnels, analytiques et génératifs pour rationaliser la création d'applications intelligentes. Quelle que soit la manière dont vous exploitez l'IA, qu'il s'agisse d'entraîner et de servir vos propres modèles de machine learning ou d'intégrer l'IA générative la plus récente dans vos applications, Atlas est un élément essentiel de votre pile. Du prototype à la production, avec Atlas, vous pouvez garantir que vos applications reposent sur la vérité avec les données opérationnelles les plus récentes, tout en répondant aux attentes des utilisateurs en matière d'évolutivité, de sécurité et de performances.

Le modèle de données documentaire flexible et l'API de requête native pour les développeurs se trouvent au cœur de MongoDB Atlas. Ensemble, ils permettent à vos développeurs de stimuler l'innovation, de devancer leurs concurrents et de tirer parti des nouvelles opportunités commerciales offertes par l'IA générative.

Les documents sont le meilleur moyen pour les développeurs de travailler avec les données, car ils correspondent à des objets dans le code, ce qui les rend intuitifs et faciles à comprendre. Ils peuvent modéliser des données de n'importe quelle structure – de la grande diversité des données d'application régulières, que nous avons évoquée, aux vector embeddings composés de plusieurs milliers de dimensions. Ces structures peuvent être modifiées à tout moment pour ajouter de nouveaux types de données et fonctionnalités d'application. Contrairement aux modèles de données tabulaires traditionnels des bases de données relationnelles, ils vous offrent la flexibilité nécessaire pour rationaliser et exploiter ces données.

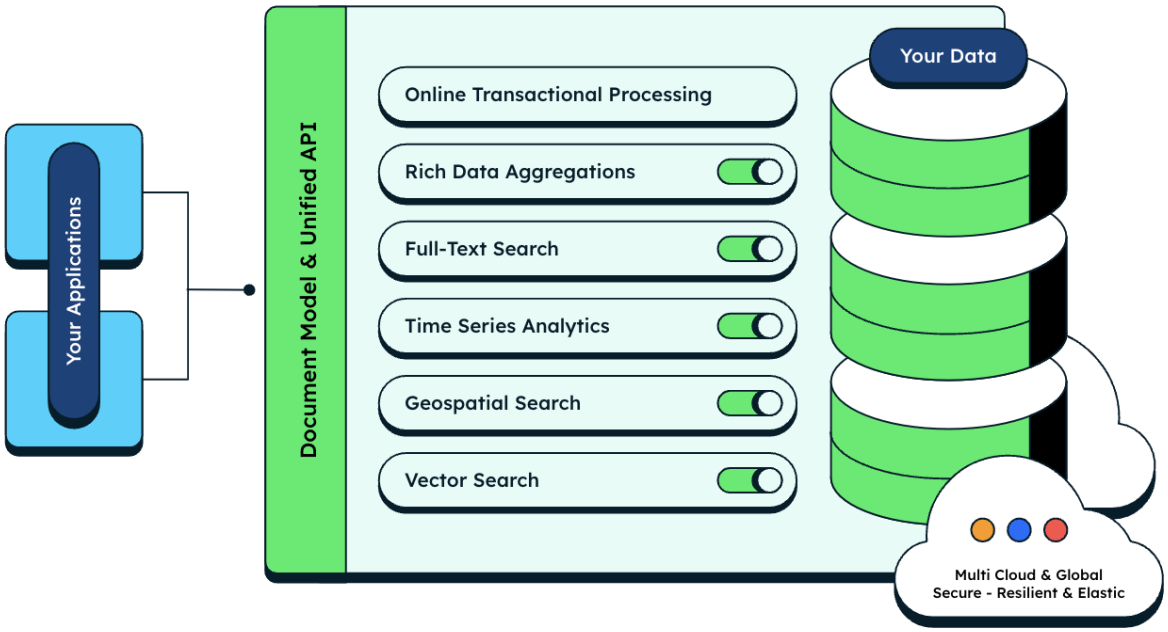

Schéma n° 2 : *MongoDB Atlas intègre les services de données nécessaires pour intégrer l'IA dans vos applications*

Combinée avec le modèle documentaire, le MongoDB Query API offre aux développeurs une méthode unifiée et cohérente pour travailler avec les données sur n'importe quel service de données. Des simples opérations CRUD à la recherche de similarité par mot-clé et par vecteur, en passant par les pipelines d'agrégation sophistiqués pour l'analyse et le traitement des flux, l'API de requête MongoDB offre aux développeurs la flexibilité nécessaire pour interroger et calculer les données selon les besoins de l'application. Dans le contexte de l'IA générative, cela offre des moyens extrêmement flexibles et puissants de définir des filtres supplémentaires sur les requêtes vectorielles, par exemple :

- Combinaison avec des métadonnées pour le filtrage : « Trouve-moi le contenu correspondant à la requête de l'utilisateur, mais uniquement publié au cours des années X, Y et Z ».
- Combinaison avec des agrégations : « Trouve-moi toutes les images similaires à l'image de la requête et regroupez-les par identifiant du photographe ».
- Combinaison avec la recherche géospatiale : « Trouve-moi des annonces immobilières pour des maisons similaires à la maison de cette photo qui se trouve à X kilomètres de ma localisation ».

Aucune autre base de données n'est en mesure d'offrir un tel éventail de fonctionnalités de requête dans une expérience de requête unique et unifiée.
Les développeurs peuvent ainsi créer plus facilement des fonctionnalités pour les utilisateurs finaux. Ils n'ont plus besoin de regrouper manuellement les résultats des requêtes de plusieurs bases de données, un processus complexe, sujet aux erreurs, coûteux et lent. En outre, votre empreinte technologique reste compacte et agile.

« MongoDB stockait déjà des métadonnées relatives aux artefacts de notre système. Avec l'introduction d'Atlas Vector Search, nous disposons désormais d'une base de données complète de métadonnées vectorielles qui a été testée pendant plus de dix ans et qui répond à nos besoins considérables en termes de récupération. Il n'est pas nécessaire de déployer une nouvelle base de données que nous devrions gérer et apprendre. Nos vecteurs et les métadonnées de nos artefacts peuvent être stockés les uns à côté des autres. »

Pierce Lamb, Senior Software Engineer de l'équipe Data and Machine Learning de VISO TRUST

Montrer plutôt que raconter : applications optimisées par l'IA générative sur une developer data platform

Nous nous concentrerons sur trois cas d'utilisation populaires pour montrer comment les développeurs utilisent MongoDB Atlas pour créer des applications enrichies par l'IA :

- Chatbot et questions-réponses (Q&A) pour le self-service client.
- Recherche avancée et recommandations aux utilisateurs dans le secteur du e-commerce.
- Analyse et génération de médias interactifs (multimodaux).

Chacun de ces exemples s'appuie sur l'IA générative et la recherche sémantique avancée pour créer des expériences utilisateur exceptionnelles et débloquer des fonctionnalités qui étaient auparavant hors de portée de la plupart des entreprises. Cependant, pour induire une profonde mutation, ces améliorations doivent s'inscrire dans le cadre d'une application plus vaste qui alimente elle-même des fonctionnalités commerciales clés.

Nous passerons en revue chaque cas d'utilisation en présentant le modèle de conception architecturale qui le sous-tend, ainsi que les fonctionnalités pertinentes fournies par MongoDB Atlas.

Chatbot et Q&A pour le self-service client

MongoDB est au cœur de nombreuses applications d'assistance client. En effet, son modèle de données flexible facilite la création d'une vue unique et à 360 degrés du client. Pour ce faire, il ingère dynamiquement des données client diverses et évoluant rapidement à partir d'une myriade de systèmes sources back-end cloisonnés

caractéristiques de la plupart des entreprises. La vue client unique et consolidée en temps réel optimisée par MongoDB est donc la plateforme idéale sur laquelle nous pouvons entraîner et fournir des fonctionnalités de chatbot et d'assistance qualité pour le self-service client.

Dans l'exemple du schéma n° 2, la base de données client stockée dans MongoDB est exportée sous forme de fichier JSON vers un modèle d'intégration qui fragmente les données (à l'aide d'outils tels que LangChain ou LlamaIndex) et crée des vector embeddings. D'autres sources de données internes telles que les bases de connaissances et la documentation peuvent également être vectorisées pour être utilisées dans l'application. Les données sont ensuite réimportées dans la base de données MongoDB.

Nous devons nous assurer que nos vecteurs sont constamment mis à jour avec les données clients les plus récentes, c'est pourquoi nous utilisons Atlas Triggers pour surveiller toute modification des données dans notre vue unique. Dès que de nouveaux enregistrements clients sont insérés ou que des enregistrements existants sont mis à jour dans la base de données, Atlas Triggers appelle l'API du modèle d'intégration pour générer les vecteurs correspondants et les charger dans Atlas.

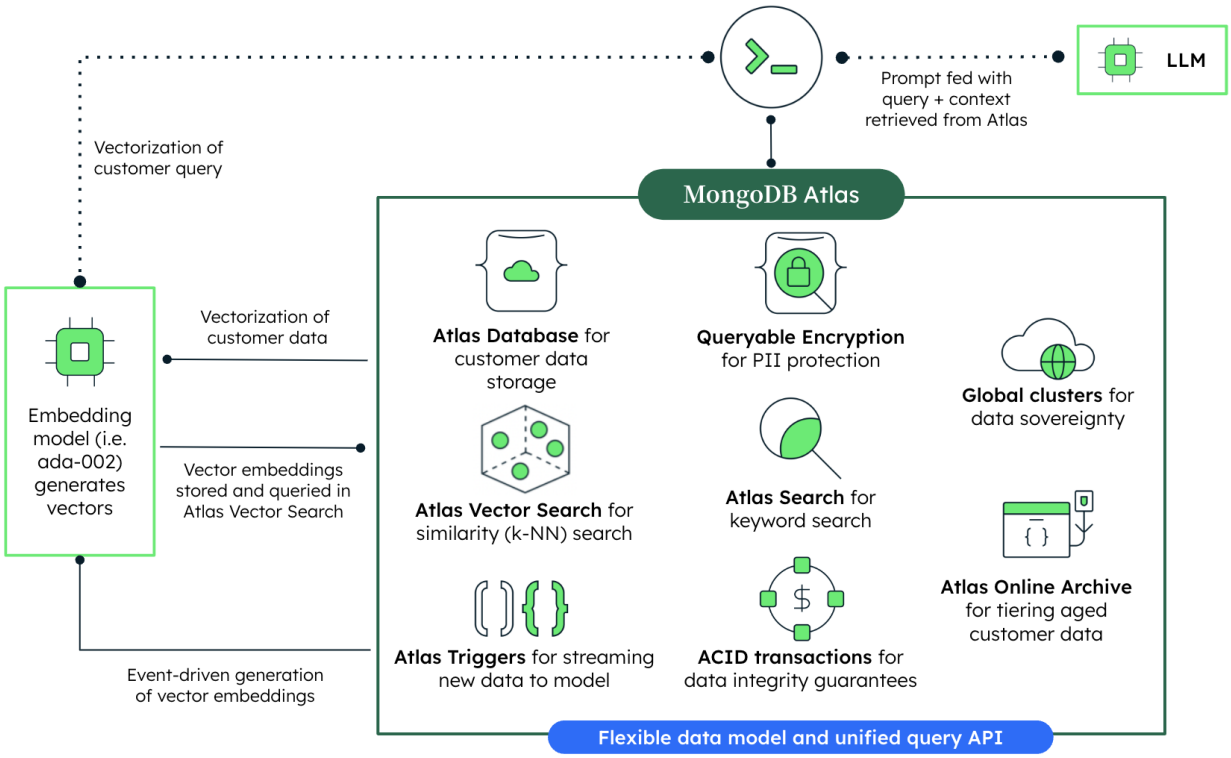

Schéma n° 3 : *Chatbot et Q&A intégrés à une application en self-service client optimisée par MongoDB Atlas*

En utilisant Atlas, les développeurs tirent parti du modèle de données flexible de MongoDB. Ils peuvent stocker les données clients sources, les métadonnées et les segments ainsi que les vector embeddings, tout en les synchronisant et en les plaçant

côte à côte dans une couche de stockage accessible via une API de requête et un pilote uniques.

Les requêtes peuvent filtrer efficacement les données à l'aide des vecteurs indexés ainsi que des index de mots-clés des champs réguliers de vos documents. Cette intégration signifie que l'application peut prendre en charge un nombre de fonctionnalités utilisateur beaucoup plus étendu avec des frais généraux réduits pour les développeurs :

- Atlas Vector Search renvoie des documents correspondants en effectuant une recherche de similarité sur les données d'intégration indexées. Pour réduire le risque de renvoyer des données obsolètes, nos requêtes peuvent utiliser un filtre de métadonnées vectorielles, tel que la « date de création » stockée dans la base de données Atlas, pour filtrer les contenus les plus anciens.
- Atlas Search renvoie des résultats en fonction des mots clés correspondants dans la source et les données clients fragmentées. Il utilise des fonctionnalités telles que la recherche floue pour corriger les fautes de frappe des l'utilisateur et la saisie semi-automatique pour proposer des termes de recherche. Il utilise également l'intersection des index pour répondre efficacement à des requêtes ad hoc complexes portant sur les données des clients.

Les requêtes adressées à la base de données Atlas, Vector Search et Atlas Search utilisent toutes la même interface de requête et le même pilote, ce qui simplifie considérablement le flux de travail des développeurs. Les données extraites de MongoDB Atlas sont fournies sous forme de contexte augmentant l'invite au LLM, ce qui lui permet de générer des réponses pertinentes aux chats et aux questions. Le contexte et les invites, ainsi que toutes les étapes de raisonnement associées utilisées pour répondre aux questions complexes, sont conservés dans Atlas, fournissant au LLM une mémoire à long terme et améliorant continuellement ses résultats.

Les données clients comptent parmi les données les plus précieuses gérées par une entreprise. Bien que l'IA générative nous aide à innover en matière de service client, la protection de leurs données reste un enjeu clé. Atlas fournit un large éventail de fonctionnalités pour nous aider à y parvenir. Les développeurs peuvent ainsi se concentrer sur les fonctionnalités basées sur l'IA :

- Une infrastructure convergente qui alimente le stockage des données, les requêtes et l'analyse, la recherche par mot-clé et la recherche vectorielle. Cette unification dans une API et un modèle de données uniques réduit considérablement le nombre d'éléments mobiles que les développeurs doivent intégrer et développer.
- Queryable Encryption est une première dans l'industrie en matière de sécurisation des données clients. Les pilotes MongoDB chiffrent les champs de données sensibles côté client, la base de données ne fonctionnant avec eux que sous forme de données chiffrées entièrement aléatoires. Même avec les

données chiffrées, les applications peuvent toujours exécuter des requêtes expressives sans avoir à déchiffrer les données de la base de données. Notez qu'en général, seuls les champs contenant les données les plus sensibles permettant d'identifier un individu, comme le numéro de sécurité sociale, sont protégés par Queryable Encryption. Il est donc possible d'effectuer des recherches dans les autres champs en texte clair.

- Les transactions ACID multi-documents dans la base de données Atlas garantissent l'intégrité des données de nos clients lorsqu'elles sont consultées et modifiées par l'application.
- Avec Atlas Global Clusters, les données client peuvent être rattachées à leur région de résidence, conformément aux réglementations actuelles en matière de souveraineté des données.
- La gestion complète du cycle de vie des données est assurée par Atlas Online Archive. Ce service extrait automatiquement les anciennes données client des bases de données actives pour les stocker dans un cloud à moindre coût, tout en maintenant l'accès aux données pour les requêtes. C'est une étape essentielle pour les données clients gérées dans des applications opérant dans des secteurs réglementés où elles doivent être conservées et accessibles pendant plusieurs années.
- Les données client sont protégées contre la corruption et les ransomwares grâce à des sauvegardes et des restaurations ponctuelles.

Atlas est entièrement géré pour vous sur les principaux clouds hyperscale, avec un SLA de disponibilité de 99,995 %.

Recherche avancée et recommandations dans le secteur du e-commerce

Les catalogues de produits dans le secteur du e-commerce sont un cas d'utilisation courant de MongoDB :

- La diversité des produits et leurs caractéristiques correspondent naturellement au modèle de données documentaire flexible de MongoDB.
- L'architecture distribuée d'Atlas avec une évolutivité élastique permet aux développeurs de dimensionner et d'ajuster dynamiquement la capacité des bases de données en réponse à la demande des applications (c'est-à-dire pour la saisonnalité des achats et les promotions commerciales).
- Avec Atlas Search, les fonctionnalités de correspondance de mots clés telles que la recherche partielle, la saisie semi-automatique, les facettes, la mise en évidence et la notation personnalisée permettent aux acheteurs de parcourir et de naviguer rapidement dans le catalogue de produits, générant ainsi des taux de clics (CTR) et des conversions d'achat.

Cependant, la recherche par mot-clé repose sur la correspondance de mots spécifiques dans les champs de texte indexés afin de renvoyer des résultats pertinents. En l'absence d'une correspondance synonymique approfondie et laborieuse (par exemple, la correspondance entre vélos et bicyclette ou baskets et tennis), les utilisateurs seront rapidement frustrés lorsque leurs requêtes de recherche n'aboutiront pas à des produits pertinents. Cette frustration se traduit par une perte de ventes et nuit à la réputation de la marque.

Autre défi : donner des recommandations aux utilisateurs. Les développeurs doivent soit écrire des moteurs complexes basés sur des règles, soit se tourner vers des ressources spécialisées et rares en science des données. En règle générale, les données doivent d'abord être ETL (extraites, transformées, chargées) de la base de données opérationnelle vers un data warehouse hors ligne ou un data lake. Ce n'est qu'ensuite que les modèles analytiques traditionnels d'IA peuvent générer un ensemble de recommandations qui doivent ensuite être réintégrées dans la base de données opérationnelle. Le processus est complexe, coûteux et génère des recommandations qui deviennent instantanément obsolètes car elles ne reflètent pas les dernières recherches ou les derniers achats de l'utilisateur.

L'amélioration de notre catalogue de produits avec des vector embeddings permet de surmonter ces obstacles.
Les vecteurs donnent une signification sémantique aux produits de notre catalogue, ce qui facilite la compréhension des similitudes et des relations entre les produits. Les commerçants peuvent ainsi présenter aux utilisateurs des produits pertinents et apparentés avec beaucoup moins d'efforts, de complexité et de coûts. Les termes de recherche courants peuvent être mis en cache dans MongoDB Atlas, ce qui permet de fournir plus rapidement des résultats pertinents aux utilisateurs.

L'extension de la vectorisation aux données clients, comme le montre l'application de self-service client précédemment évoquée, nous permet de créer des recommandations encore plus élaborées en combinant la recherche de similarités entre les produits et les clients pour affiner les suggestions.

Le schéma n° 4 présente un modèle de conception de haut niveau pour la recherche avancée et les recommandations. La création et la mise à jour de nos vector embeddings suivent le même flux de travail que celui précédemment évoqué pour les chatbots et les Q&A dans notre application de self-service pour les clients.

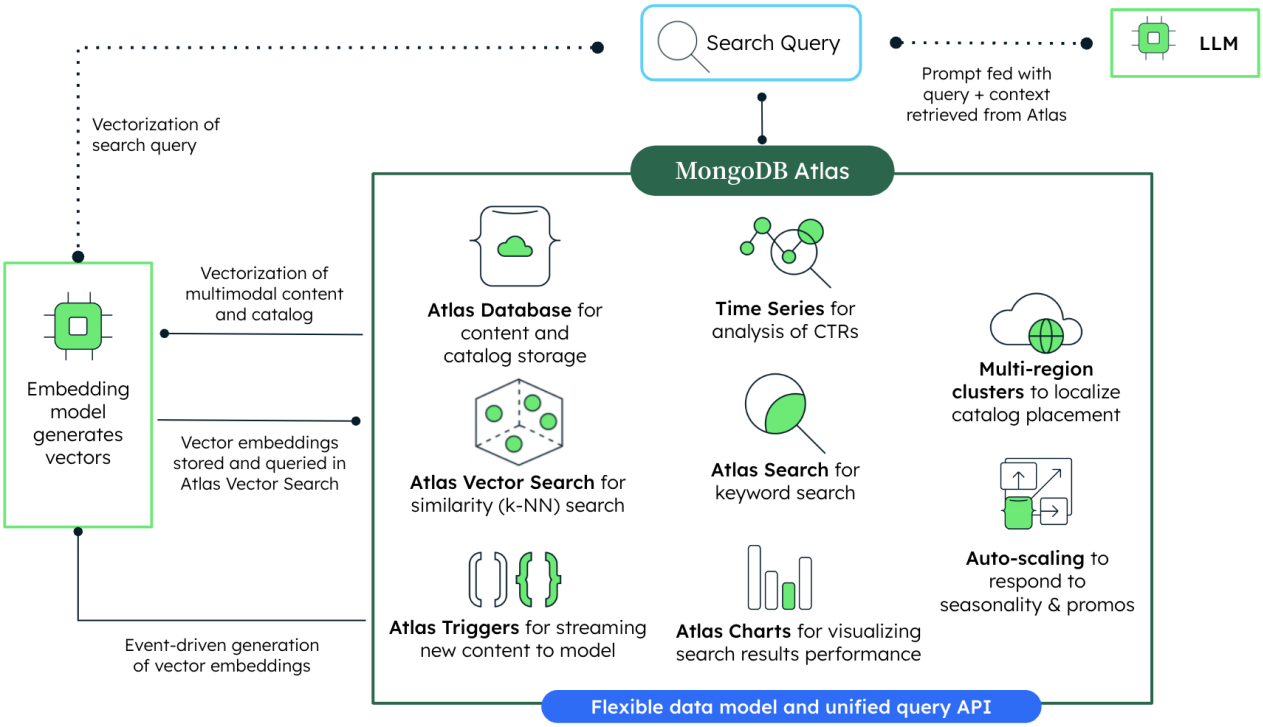

Schéma n° 4 : *la recherche sémantique avancée sur notre catalogue de produits permet d'augmenter les ventes et les conversions*

On constate que la recherche vectorielle améliore considérablement la recherche de produits et les recommandations. L'intégration d'un LLM permet d'aller encore plus loin. Désormais, les clients peuvent poser des questions en direct et obtenir des réponses instantanées sur les produits qu'ils évaluent, ce qui accélère le cycle d'achat.

Les commerçants peuvent également utiliser les LLM pour réaliser une foule de tâches laborieuses, ce qui leur permet de mettre en place de nouvelles stratégies pour attirer les clients. Par exemple, ils peuvent être utilisés pour générer différentes variantes de texte de produit et de mots-clés de référencement, qui peuvent ensuite faire l'objet d'un test A/B pour quantifier ce qui génère le plus de conversion. Le LLM peut être utilisé pour résumer plusieurs avis d'utilisateurs et déduire leurs sentiments, aidant ainsi à synthétiser les commentaires qui éclairent les feuilles de route des produits.

Les entreprises peuvent utiliser Atlas pour gérer l'ensemble du cycle de vie du e-commerce. En plus d'utiliser l'IA pour rendre notre expérience de recherche plus intelligente et prédictive, les propriétaires d'entreprise peuvent suivre les taux de clics des utilisateurs et les conversions de ventes à partir des résultats de recherche. Les collections time-series peuvent ingérer et stocker efficacement des flux de clics volumineux et à grande vitesse à partir de sessions utilisateur. Ces données peuvent donc être analysées afin de mesurer les performances de recherche et visualiser les résultats en direct à l'aide d'Atlas Charts. Grâce à ces informations, les commerçants peuvent continuellement ajuster et optimiser les données sur les produits et l'évaluation de la pertinence afin de maximiser les ventes sur le site de e-commerce.

Analyse et génération de médias enrichis (multimodaux)

La recherche par mot-clé est adaptée aux recherches textuelles les plus fréquentes. Cependant, travailler avec des médias plus riches (parfois appelés multimodaux) tels que les images, la parole et la vidéo nécessite des technologies et des compétences très pointues en matière de science des données, du moins jusqu'à présent.

Comme indiqué précédemment, tout élément de contenu numérique peut être vectorisé à l'aide d'un modèle de vector embedding approprié. Les hubs d'IA tels qu'Hugging Face et ceux des hyperscalers du cloud fournissent une multitude de modèles adaptés à différentes modalités de contenu. Les intégrations de ces modèles peuvent être stockées dans Atlas Vector Search pour alimenter une foule de nouvelles fonctionnalités. Comme évoqué précédemment, la génération d'images à partir de textes, la transcription de vidéos pour la reconnaissance vocale et l'analyse des sentiments, la classification d'images et la détection d'objets ne sont que quelques exemples de l'étendue des possibles. Les vecteurs provenant de différents médias peuvent être combinés. Par exemple, en comparant un texte et une image pour vérifier si une phrase donnée décrit correctement une image.

Cette fonctionnalité multimodale peut être utilisée dans de nombreux cas d'utilisation. Par exemple, enrichir des catalogues de produits tels que ceux mentionnés ci-dessus ou améliorer la découverte à partir de l'analyse d'images et de vidéos. Ils pourraient être utilisés pour simplifier les processus de conception, de fabrication et de publication, ou pour créer de nouvelles catégories d'applications dans des domaines tels que la sécurité et la surveillance ou la réalité augmentée (AR).

Le modèle de conception architecturale et les fonctionnalités MongoDB Atlas pour la recherche avancée et les recommandations dans le secteur du e-commerce s'appliquent également à la génération de contenu multimodal.

MongoDB Vector Search en action

MongoDB a déjà été largement adopté pour les cas d'utilisation traditionnels de l'IA. Continental a choisi MongoDB comme plateforme d'ingénierie des fonctionnalités dans le cadre de son projet de conduite autonome Vision Zero. Bosch et Telefonica utilisent MongoDB dans leurs plateformes IoT basées sur l'IA. Kronos échange des milliards de dollars de cryptomonnaie chaque jour à l'aide de modèles ML configurés et construits avec les données de MongoDB. Iguazio utilise MongoDB comme couche de persistance pour sa plateforme de science des données et de MLOps, tandis que H2O.ai et Featureform prennent en charge MongoDB en tant que magasins de fonctionnalités sur leurs plateformes respectives.

En s'appuyant sur cette base, MongoDB Atlas est déjà utilisé aujourd'hui pour de nombreuses applications qui repoussent les limites de l'IA générative. Consultez notre page d'études de cas pour en savoir plus sur l'étendue des cas d'utilisation de MongoDB Atlas. Voici quelques exemples :

- Ada aide des entreprises comme Meta, ATT et Verizon à mieux assister leurs clients grâce à l'automatisation basée sur l'IA et à l'IA conversationnelle.
- ExTrac identifie et classe les risques physiques et numériques émergents à partir de l'analyse des flux de données en temps réel.
- Eni débloque les données géologiques et les rend exploitables afin d'améliorer la prise de décision et d'accélérer la trajectoire de l'entreprise vers le zéro net.
- Inovaare surveille, extrait et classe continuellement les données tout au long du cycle de vie des soins de santé pour les rapports de conformité réglementaire, les audits et les évaluations des risques.
- Source Digital permet de diminuer par sept les coûts après la migration de PostgreSQL vers MongDB Atlas pour sa plateforme de détection vidéo.
- Catylex extrait, classe et analyse automatiquement les clauses contractuelles pour identifier les droits, les obligations et les risques.
- Robust Intelligence protège les grands modèles de langage (LLM) en production en validant les entrées et sorties en temps réel avec son offre AI Firewall.
- Potion régénère les flux vidéo et audio à l'aide de modèles de vision et d'audio personnalisés.

Schéma n° 5 : *enquête Retool sur l'état de l'IA – les meilleures bases de données vectorielles du secteur*

Afin de démontrer la popularité de MongoDB auprès des développeurs d'IA, le fournisseur d'outils logiciels Retool a conclu dans son enquête sur l'état de l'IA que MongoDB Atlas Vector Search :

1. Affiche le Net Promoter Score (NPS) le plus élevé de toutes les bases de données vectorielles interrogées.
2. Est devenue la deuxième base de données vectorielles la plus utilisée quelques mois seulement après son lancement et devance ainsi les solutions alternatives qui existent depuis des années.

« Atlas Vector Search est robuste, rentable et incroyablement rapide ! » Saravana Kumar, CEO of Kovai, évoque le développement de l'assistant IA de son entreprise.

Mise en route

Que vous meniez un projet pour une start-up ou une entreprise, MongoDB Atlas vous permet de faire ce qui suit :

- Accélérer la création de vos applications génératives enrichies par l'IA et fondées sur la vérité des données opérationnelles.
- Simplifier la pile technologique en tirant parti d'une plateforme unique qui permet à votre application de stocker des données opérationnelles et des vector embeddings en un seul endroit, de réagir aux changements dans les données sources avec des fonctions serverless et de rechercher parmi plusieurs modalités de données pour améliorer la pertinence et la précision des réponses générées par leurs applications.
- Faire évoluer facilement vos applications enrichies par l'IA générative grâce à la flexibilité du document model, tout en conservant une expérience de développement simple et élégante.
- Intégrer facilement les principaux services et systèmes d'IA tels que les hyperscalers et les LLM et frameworks open source afin de rester compétitif sur des marchés en perpétuelle évolution.
- Créer des applications enrichies par l'IA générative sur une base de données opérationnelle hautement performante et évolutive, validée depuis une dizaine d'années pour de nombreux cas d'utilisation de l'IA.

Pour en savoir plus sur la création d'applications basées sur l'IA avec MongoDB, consultez notre centre de ressources IA/ML.

La meilleure façon pour les développeurs de se lancer est de créer un compte sur MongoDB Atlas. De là, ils peuvent créer une instance MongoDB gratuite avec la base de données Atlas, Atlas Vector Search et Atlas Search, charger leurs propres données ou nos exemples d'ensembles de données, et découvrir les fonctionnalités de la plateforme.

Sphère de sécurité

Le développement, la publication et le calendrier des caractéristiques ou fonctionnalités décrites pour nos produits demeurent à notre entière discrétion. Cette information est simplement destinée à décrire l'orientation générale des produits et ne doit pas être invoquée pour prendre une décision d'achat, et n'est pas un engagement, une promesse ou une obligation légale de fournir un matériel, un code ou une fonctionnalité.

États-Unis : 866-237-8815 • INTL : +1-650-440-4474 • info@mongodb.com.