MongoDB®

MongoDB로 생성형 AI와 고급 검색을 앱에 임베딩하기

차세대 지능형 소프트웨어 구축

2023년 6월

미국 866-237-8815 • 국제 전화 +1-650-440-4474 • info@mongodb.com

목차

도입

새로운 기술 출시가 기업과 소비자 모두의 상상력을 이렇게 급속하게 자극한 적은 없습니다. 2022년 11월에 출시된 ChatGPT는 대규모 언어 모델(LLM)로 구동되는 생성형 AI가 방대한 새로운 사용 사례를 처리하는 잠재력을 보여주었습니다. 이러한 사용 사례는 현재 '전통적인' 또는 '클래식' AI라고도 하는 기존 컴퓨팅과 분석 AI로는 상상할 수 없었습니다.

잘 만들어진 프롬프트 몇 가지만 있으면 전체 범위를 자동화할 수 있습니다. 전문가 수준의 텍스트, 이미지, 오디오, 비디오 및 프로그래밍 코드를 생성하고 고객을 더 잘 지원합니다. 기후 변화 모델링, 신약 발견이나 신소재 설계, 금융 시장의 움직임 예측 등 매우 다양한 분야에 적용할 수 있습니다.

밤사이 *"어떻게 해야 생성형 AI를 사용하여 우리가 분열하지 않고 시장을 분열할 수 있을까?"* 라는 질문이 모든 이사회 안건의 상단에 등장했습니다.

기술의 미성숙과 위험에 대한 갑작스러운 인식, 더불어 잠재적 이점과 함께 리더는 수년간의 운영 모범 사례와 제도적 지식을 그냥 버릴 수 없다는 사실을 빠르게 알아챘습니다. 대신 기존 시스템과 개발 중인 새로운 애플리케이션이 모두 안전하고 신뢰할 수 있으며 정확한 방식으로 생성형 AI와 LLM을 활용할 수 있는지 확인해야 합니다.

콘텍스트가 핵심이다

조직은 상용 및 오픈 소스 LLM 모두에서 최신 혁신 방법을 사용하기를 원하지만, 자체 데이터로 트레이닝하고 촉발해야 합니다. 모두가 LLM에 액세스할 수 있을 때, LLM이 풍부하고 잘 관리된 대량의 자체 데이터에 액세스할 수 있도록 하는 데서 '초강력' 차별화가 이루어집니다.

이 데이터 중 일부는 조직의 독점 소유로, 원래 기본 모델 교육에 사용된 것보다 더 최신의 공개 데이터일 수 있습니다. 이 데이터와 함께 현재의 '실측 정보'를 더 잘 반영하는 응답을 제공할 수 있습니다. '검색 증강 생성(RAG)'이라고 불리는 이 데이터는 비즈니스와 관련성이 높고 차별화되는 AI 생성 결과물 제작에 필요한 콘텍스트를 제공하여 품질과 데이터의 최신성을 점점 더 개선합니다. 이러한 콘텍스트를 통해 고객에게 더 나은 서비스를 제공하고 조직의 생산성을 높이며 경쟁업체를 뛰어넘는 혁신을 이룰 수 있습니다.

관련 내부 데이터 소스에서 이러한 콘텍스트를 얻는 것은 비즈니스에서 생성형 AI를 이용하기 위한 한 가지 방법입니다. 하지만 그것만으로는 충분하지 않습니다. 뒷부분에서 논의하는 것처럼 개발자는 적절한 보안 제어 기능을 갖춘 정보화된 대규모 언어 모델을 중심으로 사용자가 기대하는 규모와 성능으로 애플리케이션을 배포하는 방법 역시 고려해야 합니다.

벡터와 유사성 검색의 등장

자체 데이터로 AI 모델을 촉발하려면 먼저 데이터를 벡터 임베딩으로 전환해야 합니다. 이러한 벡터는 데이터의 패턴, 관계, 구조를 포착하는 다차원적 숫자 인코딩을 제공합니다. 벡터 임베딩은 데이터에 의미론적 의미를 부여합니다; 벡터 간의 거리를 계산하면 애플리케이션이 서로 다른 데이터 객체 간의 관계와 유사성을 쉽게 이해할 수 있습니다. 이에 따라 아래에서 논의하는 완전히 새로운 범주의 애플리케이션에 데이터를 이용할 수 있습니다.

텍스트, 비디오, 오디오, 이미지 등 모든 구조와 디지털 형식의 데이터는 적절한 벡터 임베딩 모델을 통해 벡터로 변환할 수 있습니다. 예를 들자면 OpenAI의 `text-embedding-ada-002`는 텍스트와 코드 벡터화에 가장 많이 사용하는 모델입니다. 벡터 임베딩은 구조화되지 않아 컴퓨터에서 완전히 불투명한 데이터도 이제 임베딩을 통해 구조를 추론하고 표현할 수 있다는 장점을 가집니다. 즉, 정형화된 비즈니스 데이터와 동일한 방식으로 비정형 데이터를 검색하고 계산할 수 있다는 뜻입니다. 매일 생성하는 데이터의 약 80-90%가 비정형이라는 점을 고려하면 LLM과 결합된 변형 벡터 검색이 얼마나 대단한지 알 수 있습니다.

아래 그림 1에서 볼 수 있듯, 데이터가 벡터 임베딩으로 변환되면 MongoDB Atlas Vector Search (미리 보기) 같은 벡터 저장소에서 유지되고 색인 됩니다. 유사한 벡터를 검색하고자 HNSW(Hierarchical Navigable Small Worlds) 같은 알고리듬을 사용하는 KNN(K Nearest Neighbor) 검색을 수행하기 위해 ANN(Approximate Nearest Neighbor) 알고리듬으로 이 저장소를 쿼리합니다.

이러한 벡터를 쿼리하면 이전에는 고가의 데이터 과학 기술과 인프라를 통해서만 수행할 수 있었던 작업을 데이터로 수행할 수 있습니다. 첫째, 키워드 매칭을 넘어 사용자의 검색어에서 의미와 의도를 추론할 수 있는 콘텍스트 인식 시맨틱 검색으로 정보 검색과 발견을 확장할 수 있습니다. 둘째, 벡터로 인코딩된 자체 데이터를 검색하여 더 안정적이고 정확한 출력 생성에 필요한 콘텍스트를 LLM에 제공할 수 있습니다. 이러한 출력은 다음을 포함합니다:

- 텍스트 요약과 감정 분석을 위한 질의응답 및 챗봇과 같은 작업에 필요한 자연어 처리(NLP).

- 음성 인식과 번역에 대한 이미지 분류 및 객체 감지를 위한 컴퓨터 비전과 오디오 처리.
- 텍스트 기반 문서와 SEO에 최적화된 웹 페이지, 컴퓨터 코드의 작성 또는 텍스트의 이미지 또는 비디오 변환 등을 포함한 콘텐츠 생성.

벡터 검색과 LLM 워크플로우

그림 1은 LLM의 '검색 증강 생성'을 가능하게 하는 워크플로우를 보여줍니다.

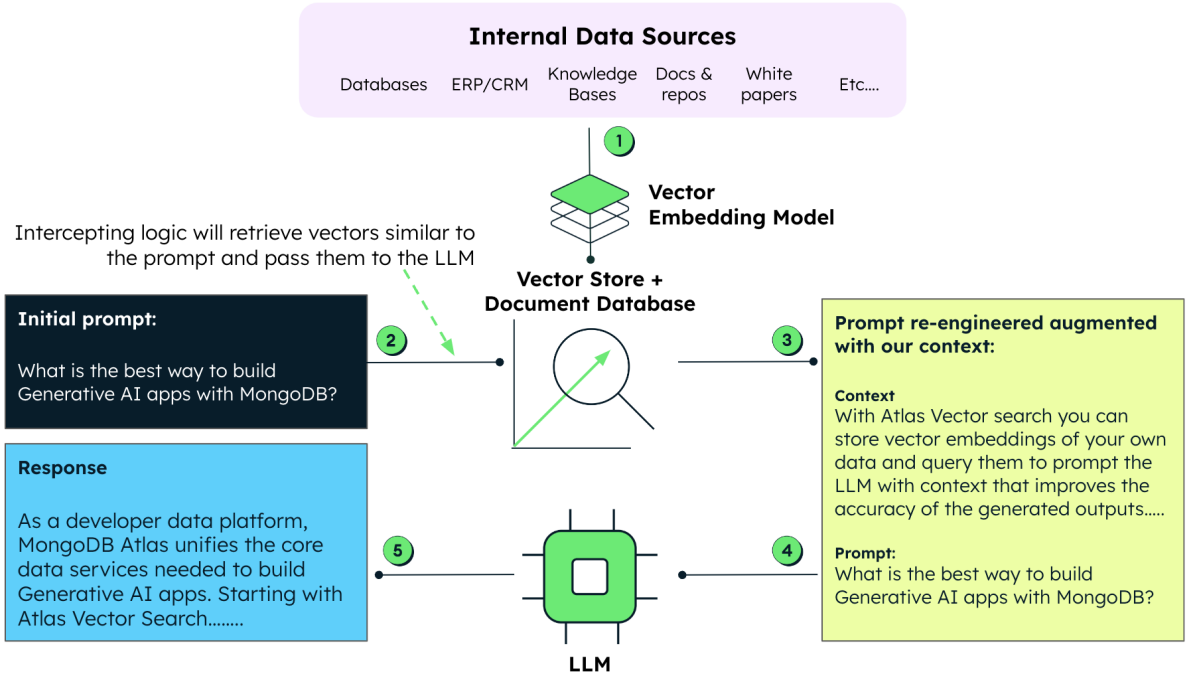

그림 1: *사용자 지정 데이터를 LLM과 동적으로 결합하여 신뢰할 수 있고 관련성 높은 결과물 생성*

빠르게도, 벡터 임베딩 모델은 데이터를 변환하여 벡터 저장소에 저장합니다. 이상적으로 벡터의 메타데이터와 원시 데이터는 벡터 자체와 함께 유연한 문서 데이터베이스에 저장됩니다. 이를 통해 애플리케이션은 다양한 방식으로 데이터를 쿼리하고 관련성을 개선하며(예: 최신 데이터에 더 높은 점수를 부여), LLM에 장기적인 메모리를 제공할 수 있습니다. 벡터 저장소에서 유사한 벡터를 검색하는 로직은 LLM에 대한 프롬프트를 차단합니다. 이후 초기 프롬프트를 재설계하는 데 사용됩니다. 새로운 프롬프트는 제공된 콘텍스트를 사용하여 최신 데이터를 활용해 더 높은 품질과 정확한 응답을 생성할 수 있는 LLM으로 전송됩니다.

이 설명서의 뒷부분에서 위의 워크플로우를 입증하는 예제를 통해 결과 기능을 다양한 애플리케이션 클래스에 어떻게 적용할 수 있는지 확인할 수 있습니다.

활기찬 AI 에코시스템의 가능성과 현실

벡터 저장소는 임베딩 생성부터 프롬프트 엔지니어링, LLM, 모델 미세 조정, 로깅, 인프라 자동화 등에 이르기까지 빠르게 진화하는 AI 지원 기술 에코시스템의 일부입니다.

이 에코시스템에는 흥미롭고 유망한 프로젝트와 협력할 벤더가 무수히 많습니다. 일부는 데모와 프로토타입을 통해 '가능성의 예술'을 보여줍니다. 하지만 기업의 의사 결정자와 개발자는 이러한 프로토타입이 특정 비즈니스의 요구에 맞춰 얼마나 쉽게 조정될 수 있을지를 염려합니다. 그리고 최신 기술 중 일부가 어떤 운영 환경에서도 매일 매일 신뢰성과 확장성, 보안을 유지하면서 생산 부하를 견딜 수 있는지도 고려해야 합니다. 추가로 고려해야 할 사항은 조직의 자체 데이터베이스를 통합하여 실제 비즈니스 데이터를 모델에 공급하는 방법입니다.

AI 에코시스템은 고립되어 존재하지 않습니다. 이러한 모든 기술은 실제 애플리케이션에 임베딩되어야 비즈니스에서 진정으로 유용할 수 있습니다. 예를 들어 벡터 저장소는 콘텍스트 인식 생성형 AI와 시맨틱 검색 활성화에 필수적입니다. 그러나 이들은 벡터화되지 않은 일반 비즈니스 데이터도 관리해야 하는 더 광범위한 애플리케이션의 일부일 뿐입니다.

이러한 데이터는 고객 기록, 주문 및 재고, 거래 및 매매, 견적, 지리적 공간 좌표, 제품 세부 정보와 가격, 시계열 측정과 센서 판독값, 클릭스트림과 소셜 피드, 텍스트 데이터 등 무엇이든 될 수 있습니다.

애플리케이션 기능을 강화하려면 이 모든 데이터를 쿼리해야 합니다. 벡터 간의 대략적인 가장 가까운 이웃을 검색하는 것뿐만 아니라 키-값 조회와 함께 데이터에 대한 수많은 업데이트 처리하고 분석 처리를 지원하는 정교한 집계와 변환을 실행해야 합니다. 이러한 쿼리는 생성형 AI 사용 사례 외의 일반 애플리케이션 기능을 강화합니다. 그러나 모델에 대해 콘텍스트 내 프롬프트와 함께 사용할 때 더욱 중요해지며 생성하는 결과의 정확성과 관련성을 개선할 수 있습니다.

또한 생성형 AI는 혁신 속도를 높이고 새로운 애플리케이션 기능을 출시하여 모든 종류의 구조와 형식으로 새로운 유형의 데이터를 생성할 것입니다. 기존 관계형 데이터베이스에 기반한 표 형식의 경직된 구조에서는 이를 수용하기 어렵습니다.

애플리케이션 데이터와 벡터 임베딩 작업 외에도 가동 시간, 성능 및 확장성 SLA 충족, 새로운 기능 통합, 데이터 보안 및 백업, 감사 등 기능적이지 않은 일도 수행해야 합니다. 이러한 작업 중 일부는 지루하게 들릴 수 있습니다. 하지만 이 중 하나라도 오류가 나면 큰 문제가 될 것입니다.

새로운 AI 기반 경험을 제공하는 기술을 한데 모아 애플리케이션에 통합하면 수많은 포인트 제품과 복잡성이 발생하여 팀에 엄청난 오버헤드가 생길 위험이 있습니다. 이 모든 과제는 파편화되고 비능률적인 개발자 경험, 처리해야 할 다수의 운영 및 보안 모델, 엄청나게 많은 데이터 랭글링과 통합 작업, 수많은 데이터 중복으로 이어집니다. 이 모든 것이 새로운 AI 기반 경험을 시장에 출시하는 속도를 늦추면서 비용과 위험을 증가시킵니다.

MongoDB Atlas에 구축된 개발자 데이터 플랫폼을 사용하여 더 나은 방법을 찾을 수 있습니다.

개발자 데이터 플랫폼: 스마트 애플리케이션을 구축하는 현명한 방법

MongoDB Atlas에 구축된 MongoDB의 개발자 데이터 플랫폼은 운영, 분석 및 생성형 AI 데이터 서비스를 통합하여 AI 애플리케이션 구축을 간소화합니다. 자체 기계 학습 모델을 교육하고 제공하는 것부터 최신 생성형 AI를 앱에 임베드하는 것까지 어떤 방식으로 AI를 활용하든 Atlas는 스택의 중요한 부분입니다. 프로토타입에서 프로덕션까지 사용자가 기대하는 규모, 보안 및 성능을 충족하면서 최신 운영 데이터를 사용하여 앱이 사실에 기반하도록 보장하세요.

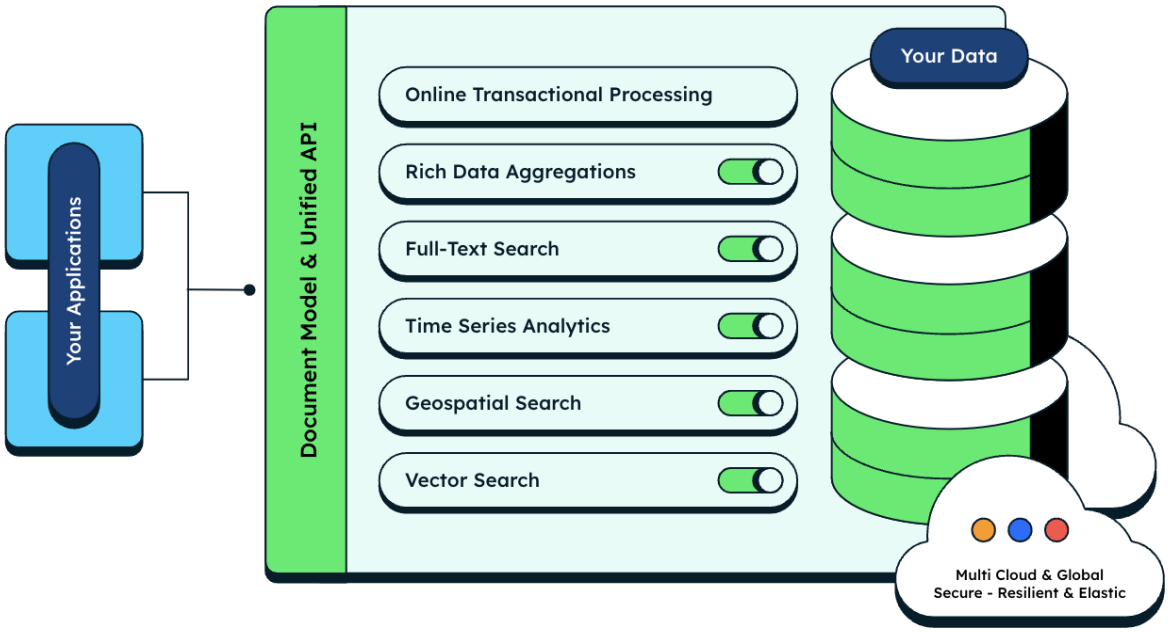

그림 2: *애플리케이션의 AI 도입에 필요한 데이터 서비스를 통합하는 MongoDB Atlas*

MongoDB Atlas의 핵심은 유연한 문서 데이터 모델과 개발자 네이티브 쿼리 API입니다.

문서는 코드의 객체에 매핑되어 직관적이고 쉽게 추론할 수 있기 때문에 개발자가 데이터로 작업하기 가장 좋은 방법입니다. 문서는 앞서 언급한 광범위한 일반 애플리케이션 데이터부터 수천 개의 차원으로 구성된 벡터 임베딩에 이르기까지 모든 구조의 데이터를 모델링할 수 있습니다. 이러한 구조는 새로운 데이터 유형과 애플리케이션 기능의 추가를 지원하기 위해 언제든지 수정될 수 있습니다.

문서 모델과 함께 MongoDB Query API는 개발자에게 모든 데이터 서비스에서 데이터로 작업할 수 있는 통일되고 일관된 방법을 선사합니다. 간단한 CRUD 작업부터 키워드와 벡터 유사성 검색, 분석 처리를 위한 정교한 집계 파이프라인까지, Query API는 개발자에게 애플리케이션에 필요한 방식으로 데이터를 쿼리하고 계산할 수 있는 유연성을 제공합니다. 따라서 개발자가 서로 다른 쿼리 언어와 드라이버 간에 지속적으로 콘텍스트를 전환해야 하는 생산성 문제를 방지하는 동시에 기술 공간을 작고 민첩하게 유지할 수 있습니다.

개발자 데이터 플랫폼에 구축된 생성형 AI 접목 앱 사용 사례

개발자가 MongoDB Atlas를 사용하여 AI가 강화된 앱을 구축하는 방법을 보여주기 위해 세 가지 인기 있는 사용 사례를 소개합니다.

- 고객 셀프서비스를 위한 챗봇과 Q-A.
- 고급 전자상거래 검색 및 사용자 추천.
- 리치 미디어(다모드) 분석과 생성.

각 사례는 생성형 AI와 고급 시맨틱 검색을 사용하여 놀라운 사용자 경험을 만들고 이전에는 대부분의 조직에서 사용할 수 없었던 기능을 활성화합니다. 그러나 진정한 변화를 위해서는 이러한 AI 향상이 중요한 비즈니스 기능을 지원하는 더 큰 애플리케이션의 일부로 제공되어야 합니다.

각 사용 사례를 차례로 살펴보면서 이를 지원하는 건축학적 설계 패턴과 더불어 MongoDB Atlas가 제공하는 관련 기능을 보여드리겠습니다.

고객 셀프서비스를 위한 Q-A와 챗봇

MongoDB는 많은 고객 지원 애플리케이션의 핵심에 자리 잡고 있습니다. MongoDB의 유연한 데이터 모델로 고객에 대한 360도 단일 뷰를 쉽게 구축할 수 있기 때문입니다. 대부분의

조직에서 사일로화된 수많은 백엔드 소스 시스템으로부터 다양하고 빠르게 변화하는 고객 데이터를 동적으로 수집하여 이를 수행합니다. 따라서 MongoDB가 지원하는 통합된 단일 실시간 고객 뷰는 고객 셀프서비스를 위한 챗봇과 Q-A(질문-응답) 지원 기능을 교육하고 제공할 수 있는 이상적인 플랫폼입니다.

그림 2의 예시에서 MongoDB에 저장된 고객 데이터베이스는 LangChain 또는 LlamaIndex 같은 도구를 사용하여 데이터를 청크하는 임베딩 모델에 JSON 파일로 내보내지고 이로부터 벡터 임베딩을 생성합니다. 이후 데이터를 MongoDB 데이터베이스로 다시 가져옵니다. 벡터는 최신 고객 데이터로 끊임없이 업데이트되어야 하므로 Atlas Triggers를 사용하여 단일 보기에서 데이터 변경을 주시합니다. 데이터베이스에 새로운 고객 레코드가 삽입되거나 기존 레코드가 업데이트되는 즉시, Atlas Triggers는 임베딩 모델의 API를 호출하여 해당 벡터를 생성하고 이를 다시 Atlas로 로드합니다.

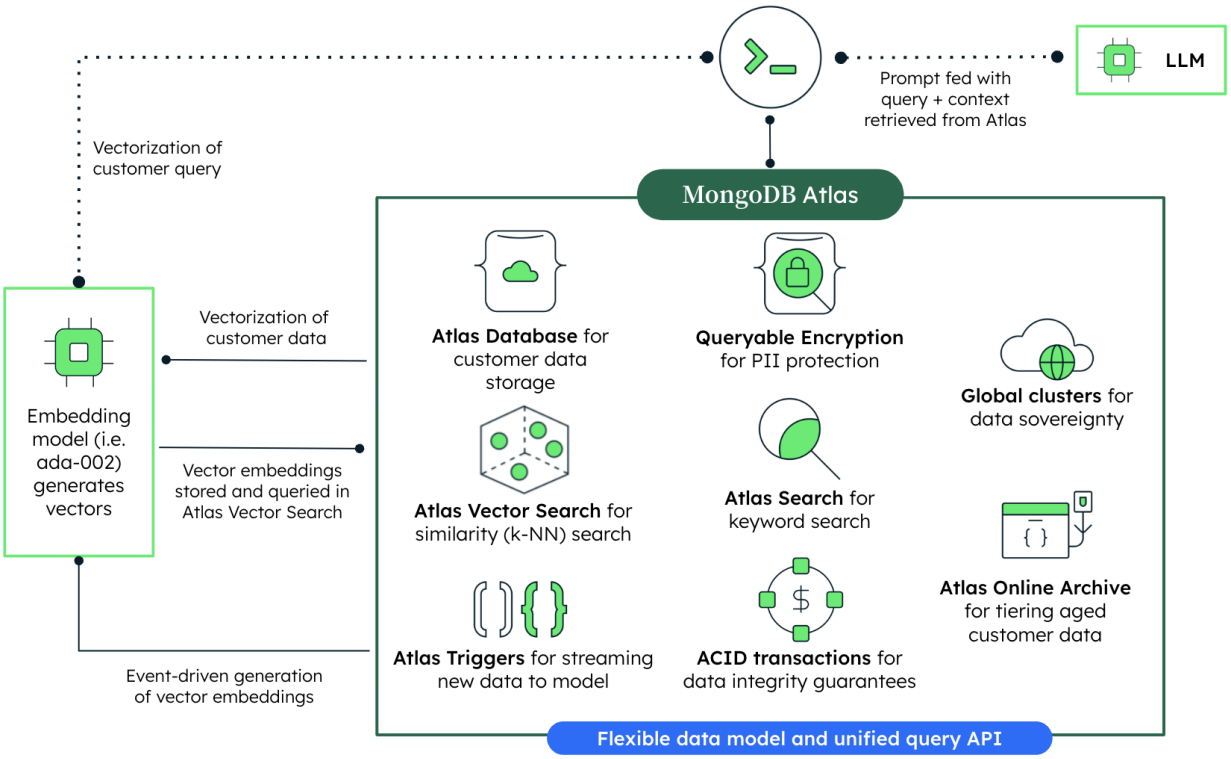

그림 3: *MongoDB Atlas로 구동되는 고객 셀프서비스 애플리케이션에 내장된 Q-A 생성 AI 기능과 챗봇*

개발자는 Atlas를 사용해 MongoDB의 유연한 데이터 모델을 활용할 수 있습니다. 소스 고객 데이터, 메타데이터, 청크를 벡터 임베딩과 함께 저장할 수 있고 모두 동기화되어 단일 스토리지 레이어에 나란히 배치되며 단일 쿼리 API와 드라이버로 액세스할 수 있습니다.

쿼리는 문서 내 일반 필드의 키워드 인덱스와 함께 인덱싱된 벡터를 사용하여 데이터를 효율적으로 필터링할 수 있습니다. 이 통합으로 앱은 개발자 오버헤드를 줄이면서 훨씬 더 광범위한 사용자 기능을 지원할 수 있습니다:

- Atlas Vector Search는 색인된 임베딩 데이터에서 유사성 검색을 수행하여 일치하는 문서를 돌려줍니다. 쿼리는 Atlas 데이터베이스에 저장된 '생성 날짜' 같은 벡터의 메타데이터를 사용해 예전 콘텐츠를 필터링하여 오래된 데이터가 반환되는 위험을 줄입니다.
- Atlas Search는 소스에서 일치하는 키워드와 청크된 고객 데이터를 기반으로 결과를 반환합니다. 퍼지 검색과 같은 기능을 사용하여 사용자가 입력한 오타를 수정하고 자동 완성으로 추천 검색어를 제공합니다. 또한 인덱스 교차를 사용하여 고객 데이터에 대한 사용자의 복잡한 임시 쿼리를 효율적으로 해결합니다.

Atlas 데이터베이스에 대한 쿼리와 Vector Search, Atlas Search는 모두 동일한 쿼리 인터페이스와 드라이버를 사용하므로 개발자의 워크플로우가 굉장히 간결해집니다. MongoDB Atlas에서 검색된 데이터는 LLM에 프롬프트를 보강하는 콘텍스트로 제공되어 채팅과 질문에 대한 관련 응답을 생성할 수 있습니다. 콘텍스트와 프롬프트, 더불어 복잡한 질문 응답에 사용되는 모든 관련 추론 단계는 Atlas를 유지시켜 LLM에 장기적인 메모리를 제공하고 출력을 지속적으로 개선합니다.

고객 데이터는 모든 조직이 관리하는 자료 중 가장 소중합니다. 생성형 AI가 고객 서비스 혁신을 돕지만 고객의 데이터 보호가 여전히 가장 중요합니다. Atlas는 이를 돕는 다양한 기능을 제공하여 개발자가 AI 기반 기능에 집중할 수 있도록 돕니다:

- 데이터 저장, 쿼리 및 분석, 키워드 검색, 벡터 검색을 지원하는 컨버지드 인프라. 단일 API와 데이터 모델을 기반으로 하는 이러한 통합은 개발자가 통합하고 구축해야 하는 움직이는 부품의 수를 극적으로 줄여줍니다.
- Queryable Encryption은 업계 최초로 고객 데이터를 보호합니다. MongoDB 드라이버는 클라이언트 측에서 데이터베이스가 완전히 무작위로 암호화된 데이터로만 작업하도록 민감한 데이터 필드를 암호화합니다. 데이터가 암호화되어도 애플리케이션은 데이터베이스의 데이터를 해독할 필요 없이 데이터에 대한 표현식 쿼리를 실행할 수 있습니다. 일반적으로 SSN처럼 개인을 고유하게 식별하는 가장 민감한 데이터가 포함된 필드만 Queryable Encryption으로 보호됩니다. 따라서 나머지 보통 텍스트 필드에서는 검색할 수 있습니다.
- Atlas 데이터베이스의 다중 문서 ACID 거래는 애플리케이션에서 액세스하고 수정할 때마다 고객 데이터의 온전함을 보장합니다.
- Atlas Global Clusters로 고객 데이터를 거주 지역에 고정하면 최신 데이터 주권 규정을 준수할 수 있습니다.
- Atlas Online Archive는 완벽한 데이터 수명주기 관리를 제공합니다. 이 서비스는 활성 데이터베이스에서 오래된 고객 데이터를 저렴한 클라우드 객체 스토리지에 자동으로 계층화하며 쿼리를 위해 데이터에 액세스할 수 있도록 유지합니다.

- 이는 규제 산업에서 운영하는 앱 내에서 관리되는 고객 데이터를 수년 동안 보관하고 액세스해야 하므로 중요합니다.
- 고객 데이터는 백업과 특정 시점 복원을 통해 손상과 랜섬웨어로부터 보호됩니다.

Atlas는 주요 하이퍼스케일 클라우드에서 완벽하게 관리되며 99.995%의 가동 시간 SLA로 지지됩니다.

고급 전자상거래 검색 및 추천

Ecommerce product catalogs는 MongoDB의 일반적인 사용 사례입니다:

- 서로 다른 다양한 제품과 속성은 MongoDB의 유연한 문서 데이터 모델에 자연스럽게 매핑됩니다.
- 개발자는 탄력적인 스케일 아웃이 가능한 Atlas의 분산 아키텍처로 애플리케이션 수요(예를 들어 계절성 및 판매 프로모션)에 따라 데이터베이스 용량을 동적으로 조정하고 사이즈를 조정할 수 있습니다.
- Atlas Search를 사용하면 퍼지 검색, 자동 완성, 패싯팅, 하이라이팅 및 고객 스코어링과 같은 키워드 매칭 기능을 통해 구매자가 제품 카탈로그를 빠르게 검색하고 탐색하여 클릭률(CTR)과 구매 전환을 높일 수 있습니다.

그러나 키워드 검색은 관련 결과를 반환하기 위해 색인된 텍스트 필드에서 일치하는 특정 단어에 의존합니다. 광범위하고 고된 동의어 매핑(예를 들어 자전거를 사이클링으로, 스니커즈를 트레이닝 슈즈로 매핑)이 없으면 검색 쿼리가 관련 제품을 보여주지 못할 때 사용자는 바로 실망할 것입니다. 이러한 불만은 자연스럽게 판매 손실과 브랜드 평판 손상으로 이어집니다.

추가적인 과제는 사용자에게 권장 사항을 제공하는 것입니다. 개발자는 복잡한 규칙 기반 엔진을 작성하거나 전문적이고 희소한 데이터 과학 리소스에 의존해야 합니다. 일반적으로 데이터를 먼저 운영 데이터베이스에서 오프라인 데이터 웨어하우스 또는 데이터 레이크로 ETL(추출, 변환, 로드)해야 합니다. 그래야만 기존 분석 AI 모델이 일련의 권장 사항을 생성할 수 있으며 이후 운영 데이터베이스에 다시 로드해야 합니다. 이 프로세스는 복잡하고 비용이 많이 들며 즉시 구식이 되는 권장 사항을 생성합니다.

벡터 임베딩으로 제품 카탈로그를 개선하면 이러한 문제를 해결할 수 있습니다. 벡터는 카탈로그 내 제품에 의미론적 의미를 부여하여 제품 간의 관계와 유사성을 쉽게 이해할 수 있도록 해줍니다. 이를 통해 머천다이저는 더 낮은 전자 로트와 복잡성, 비용으로 사용자에게 관련성과 연관성이 있는 제품을 보여줄 수 있습니다. 일반적인 검색어를 MongoDB Atlas에 캐시하여 사용자에게 관련성 높은 결과를 더 빠르게 제공할 수 있습니다.

앞서 고객 셀프서비스 앱에서 살펴본 것처럼 벡터화를 고객 데이터로 확장하면 제안을 미세 조정하는 고객 유사성 검색과 제품을 결합하여 더욱 정교한 권장 사항을 구축할 수 있습니다.

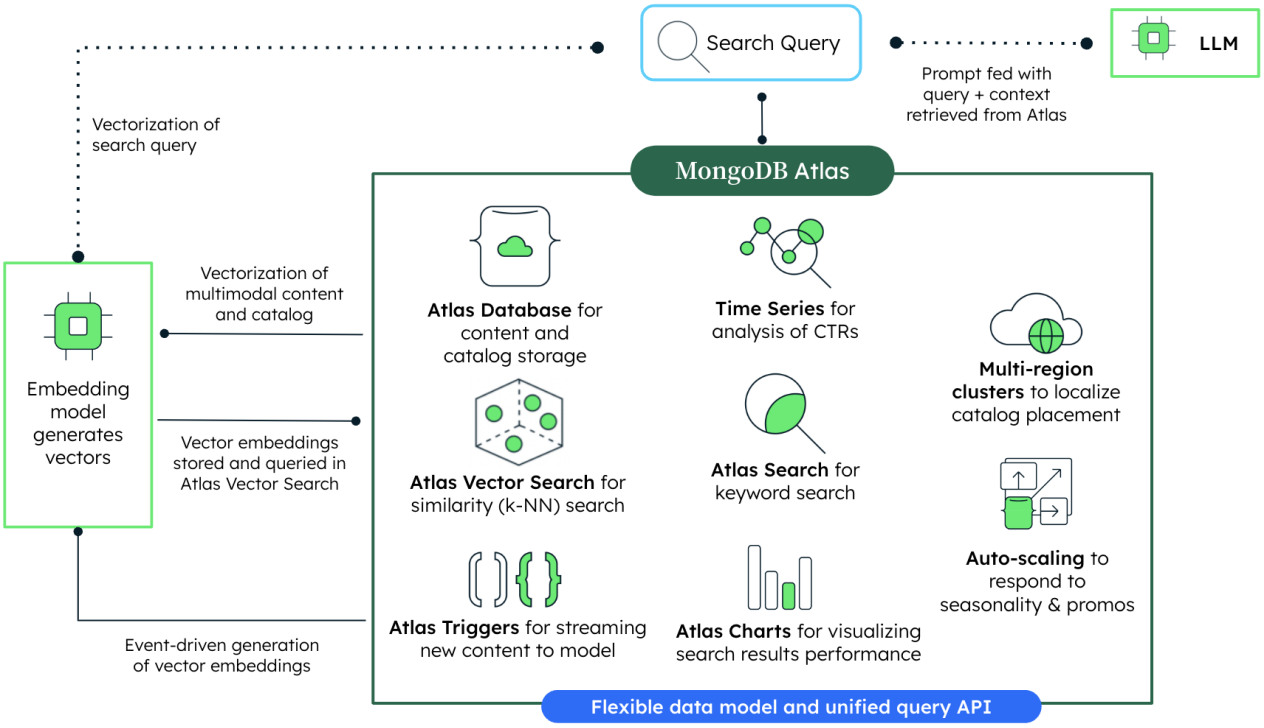

그림 4: *제품 카탈로그의 고급 시맨틱 검색으로 더 높은 판매 전환율과 상향 구매 유도*

위 그림은 고급 검색과 추천을 위한 높은 수준의 디자인 패턴을 보여줍니다. 벡터 임베딩을 만들고 유지하는 일은 앞서 설명한 고객 셀프서비스 애플리케이션의 챗봇 및 Q-A와 동일한 워크플로우를 따릅니다.

제품 검색에 LLM을 사용하는 것은 선택 사항이지만 일반 검색 환경보다 강력한 향상을 선사합니다. 이제 고객은 평가 중인 제품에 대해 실시간으로 질문하고 즉각적인 답변을 얻을 수 있어 구매 주기를 가속화할 수 있습니다.

머천다이저는 또한 이전에 힘들었던 다양한 작업에 LLM을 사용하여 더욱 창의적으로 문제를 해결할 수 있습니다. 예를 들어 LLM으로 제품 카피와 SEO 키워드의 다양한 변형을 생성한 다음 A/B 테스트를 통해 어떤 것이 더 잘 전환하는지 정량화할 수 있습니다. LLM은 여러 사용자의 리뷰 요약과 감정 추론에 사용할 수 있으며, 제품과 상업적 개선을 주도하는 피드백을 종합하는 데 도움이 됩니다.

조직은 Atlas를 사용해 전체 전자상거래 수명 주기를 관리할 수 있습니다. AI를 사용하여 검색 경험을 더 지능적이고 예측할 수 있게 만드는 것 외에도 비즈니스 소유자는 검색 결과에서 사용자 클릭률과 판매 전환을 추적할 수 있습니다.

Time series collections은 사용자 세션에서 빠른 속도와 방대한 클릭스트림을 효율적으로 수집하고 저장할 수 있으며, Atlas Charts를 사용한 결과의 실시간 시각화를 비롯한 검색 성능을 측정하기 위한 분석에 이러한 데이터를 활용할 수 있습니다. 이러한 인사이트를 통해 판매자는 제품 데이터와 관련성 점수를 지속적으로 조정하고 최적화하여 전자상거래 사이트에서 매출을 극대화할 수 있습니다.

리치 미디어 분석과 생성

일반 텍스트 검색은 기존 키워드 검색으로 충분히 가능합니다. 하지만 이미지, 음성, 동영상 등 더 풍부한 미디어(멀티모달이라고도 함) 자산으로 작업하려면 고도로 복잡한 데이터 과학 기술과 기량이 필요합니다. 지금까지는요.

앞서 언급했듯 모든 디지털 콘텐츠 자산은 적절한 벡터 임베딩 모델로 벡터화할 수 있습니다. Hugging Face 같은 AI 허브와 클라우드 하이퍼스케일러의 허브는 다양한 콘텐츠 양식에 맞게 조정된 풍부한 모델을 제공합니다. 이러한 모델의 임베딩은 Atlas Vector Search에 저장되어 다양한 새 기능을 지원합니다. 앞서 설명한 것처럼 텍스트에서 이미지를 생성하고, 음성 인식과 감정 분석을 위해 비디오를 기록하고, 이미지를 분류하고, 객체를 감지하는 것은 가능한 일의 몇 가지 예시일 뿐입니다. 다양한 미디어의 벡터를 결합할 수 있습니다. 예를 들어 텍스트와 이미지 임베딩을 비교하여 주어진 문장이 이미지를 정확하게 설명하는지 확인합니다.

이 멀티모달 기능은 다양한 사용 사례에서 사용할 수 있습니다. 예를 들자면 위에서 설명한 것과 같은 제품 카탈로그를 보강하거나 내부 지식 저장소에서 검색을 강화할 수 있습니다. 설계와 제조, 퍼블리싱 프로세스를 간소화하거나 보안 및 감시와 같은 영역에서 완전히 새로운 클래스의 애플리케이션 생성에 사용할 수 있습니다.

위의 고급 전자상거래 검색과 권장 사항에 관해 설명한 아키텍처 설계 패턴과 MongoDB Atlas 기능은 멀티모달 콘텐츠 생성에 동일하게 적용됩니다.

MongoDB Vector Search 도입 사례

MongoDB는 이미 전통적인 AI 사용 사례에 널리 채택되고 있습니다. Continental은 MongoDB를 Vision Zero 자율 주행 이니셔티브의 기능 엔지니어링 플랫폼으로 선택했습니다. Bosch와 Telefonica는 모두 AI 강화 IoT 플랫폼에서 MongoDB를 사용합니다. Kronos는 MongoDB의 데이터로 구성 및 구축된 ML 모델을 사용해 매일 수십억 달러의 암호화폐를 거래합니다. Iguazio는 MongoDB를 데이터 과학 및 MLOps 플랫폼에 대한 지속성 계층으로 사용하고 H2O.ai와 Featureform은 각 플랫폼의 기능 저장소로 MongoDB를 지원합니다.

이러한 토대 위에 구축된 MongoDB Atlas Vector Search는 비즈니스를 위한 생성형 AI와 고급 검색의 가능성을 탐색하는 기업에서 수많은 프로젝트에 이미 사용하고 있습니다. 예를 들겠습니다:

- 공공 및 민간 부문 조직의 혁신을 전문으로 지원하는 한 글로벌 경영 컨설팅 회사는 Q-A 시스템에서 Atlas Vector Search를 사용합니다. 70개국에 퍼져 있는 35,000명의 직원으로 구성된 회사의 컨설턴트들은 이 시스템을 사용하여 특정 산업이나 기업에 대한 연구와 발견을 가속화하고 있습니다. 이들은 Atlas의 Q-A 시스템을 사용하여 도메인 주제별 전문가(SME)와 인터뷰 요약을 식별하고 생성합니다.
- 150,000명 이상의 직원과 300억 유로의 매출을 보유한 유럽의 주택 개조 소매 그룹은 경쟁 가격 책정 시스템에서 Atlas를 사용하고 있습니다. 이 회사는 경쟁사 웹사이트를 크롤링하여 제품 설명과 이미지를 추출하고 벡터화하여 Atlas Vector Search에 저장합니다. 머천다이저는 유사성 시맨틱 검색과 키워드 검색을 모두 실행해 경쟁 제품을 자신의 제품과 일치시키고 회사의 가격 책정 엔진에 결과를 입력하여 필요한 조정을 수행합니다.
- 한 고객 인텔리전스 스타트업은 고객 통화 기록과 지원 채팅의 쿼리와 저장에 Atlas Vector Search를 사용하고 있습니다. 조직은 스타트업의 대화형 챗봇을 사용하여 계정 상태와 제품 정서, 기능 로드맵 우선순위에 대한 더욱 정확한 인사이트를 도출할 수 있습니다.
- CRM 및 ERP 프로세스 같은 프런트 및 백오피스 업무 자동화에 중점을 둔 세계 최대의 로보틱 프로세스 자동화 소프트웨어 공급업체 중 한 곳은 헬프 포털 내에서 Atlas를 사용하고 있습니다. Atlas Vector Search와 키워드 검색은 검색 결과의 전반적인 품질을 개선하고 사용자에게 관련성이 더 높은 헬프 문서를 돌려주어 문제 해결과 고객의 생산성 경로를 가속화합니다.
- 유럽의 한 주요 자동차 제조업체는 엔진 진단을 간소화하는 방법을 모색하기 위해 Atlas를 사용합니다. 엔진에서 오디오 파일을 녹음한 다음 이를 벡터화하고 검색하여 유사한 사례를 찾을 수 있습니다. 매칭을 더욱 개선하기 위해 하이브리드 검색 방식을 사용합니다. 이는 키워드 기반 메타데이터 검색(예를 들어 자동차 모델, 제조 연도, 제조 공장)과 오디오 기반 벡터 유사성 검색을 위한 Atlas Search를 결합합니다.
- 한 다국적 디자인 소프트웨어 회사는 Atlas를 사용하여 마케터와 아티스트가 이미지 라이브러리에서 관련 디지털 자산을 찾을 수 있도록 돕습니다. 유사성 검색은 Atlas Search로 보완되어 키워드로 결과를 필터링하고 콘텐츠 검색 속도를 개선합니다.

시작하기

스타트업이든 대기업이든 MongoDB Atlas를 사용하여 차세대 거대 시장을 구축할 수 있습니다:

- 운영 데이터의 진실에 기반한 생성형 AI 강화 애플리케이션 구축을 가속화합니다.
- 앱과 벡터 데이터를 같은 위치에 저장하고, 서버리스 기능으로 소스 데이터의 변화에 대응하고, 앱이 생성하는 응답의 관련성과 정확성을 개선하기 위해 여러 데이터 양식을 검색할 수 있는 단일 플랫폼을 활용하여 기술 스택을 단순화합니다.
- 간단하고 우아한 개발자 환경을 유지하면서 문서 모델의 유연성을 통해 생성형 AI 강화 앱을 쉽게 발전시킵니다.
- 역동적인 시장에서 경쟁력을 유지하기 위해 하이퍼스케일러와 오픈 소스 LLM, 프레임워크 같은 선도적인 AI 서비스와 시스템을 원활하게 통합합니다.
- 다양한 AI 사용 사례를 통해 십 년간 검증된 확장성 높은 고성능 운영 데이터베이스에서 생성형 AI 강화 애플리케이션을 구축합니다.

AI/ML 리소스 센터를 방문하여 MongoDB로 지능형 앱을 구축하는 방법을 더 자세히 알아볼 수 있습니다.

개발자가 시작하는 최상의 방법은 MongoDB Atlas에 계정을 등록하는 것입니다. Atlas 데이터베이스와 Atlas Vector Search, Atlas Search를 사용하여 무료 MongoDB 인스턴스를 생성하고 자체 데이터 또는 샘플 데이터 세트를 로드하며 플랫폼 내에서 가능한 것을 탐색할 수 있습니다. Atlas Vector Search는 현재 Public Preview 프로그램으로 제공됩니다.

MongoDB Developer Center에서는 프로그래밍 언어와 제품으로 구성된 사용 지침서, 샘플 코드, 동영상, 문서 등 다양한 리소스를 호스팅합니다. MongoDB 전문 서비스에서 제공하는 강사 주도형 교육 및 컨설팅 서비스와 더불어 MongoDB University를 통한 자기 주도형 교육도 제공합니다. 각 속성의 지식 기반을 확장하기 위해 Vector search와 AI 콘텐츠를 지속적으로 추가합니다.

이러한 리소스는 종합적으로 생성형 AI와 고급 검색을 통해 앱을 강화하는 과정을 시작하도록 돕습니다.

회피 조항

제품에 대해 설명된 모든 특성 또는 기능의 개발, 출시 및 시기는 당사의 단독 재량을 따릅니다. 이 정보는 일반적인 제품 방향을 설명하기 위한 것일 뿐이고 구매 결정을 내리는 데 의존해서는 안 되며, 자료나 코드 또는 기능을 제공하겠다는 공약이나 약속 또는 법적 의무를 의미하지 않습니다.

미국 866-237-8815 • 국제 전화 +1-650-440-4474 • info@mongodb.com