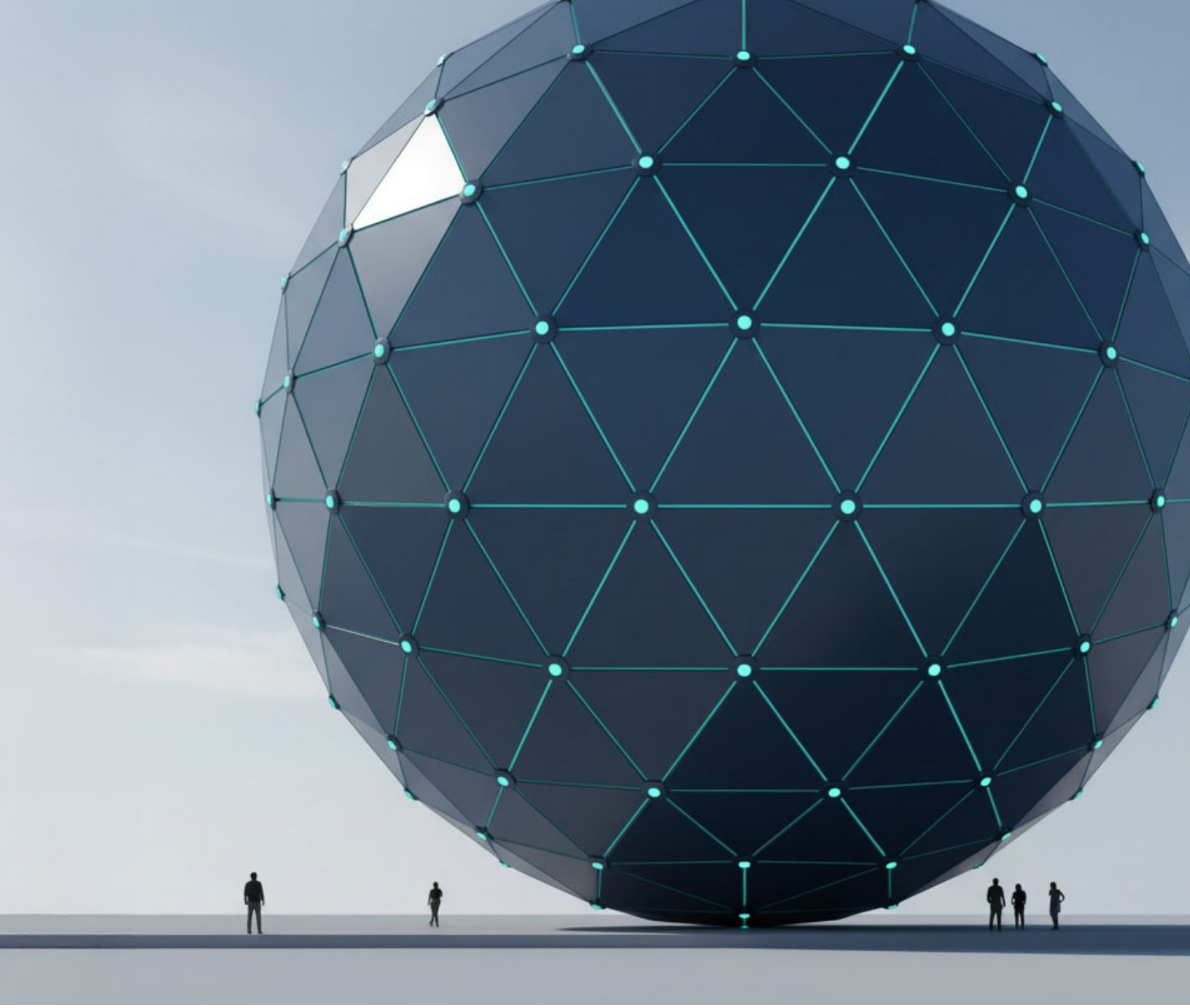


elastic

에어갭 환경을 위한 고급 벡터 검색

공공 부문은 생성형 AI(GenAI) 도입에 있어 중요한 시점에 있습니다. 조직에서는 국가 차원의 전략과 이해관계자의 기대에 힘입어 더 큰 규모의 AI 프로젝트를 추진해야 하는 상황에 직면해 있습니다. 동시에 GenAI와 검색은 긴밀히 연결되고 있습니다. 사용자는 링크 목록을 제공하는 검색 엔진 대신 빠른 답변을 제공하는 대화형 GenAI를 선호하고 있습니다.

공공 부문에서 이러한 기대를 충족하는 데에는 다음과 같이 몇 가지 어려움이 따릅니다.

1. GenAI 사용 옵션은 있지만 개발자가 충분하지 않습니다. 조직은 가장 효율적인 결과를 창출하기 위해 전략적으로 자원을 투자해야 합니다.

2. 공공 부문 조직은 방대한 양의 데이터를 처리해야 합니다. 따라서 AI의 기반으로 활용할 수 있는 확장성 있고 빠르며 효율적인 지식 기반이 필요합니다.

3. 데이터가 AI를 위해 항상 준비되어 있는 것은 아닙니다. 중요한 정보는 종종 PDF, 다이어그램, 비디오와 같은 형식으로 저장되며, 이러한 형식에서 관련 정보를 추출하기 어려울 수 있습니다.

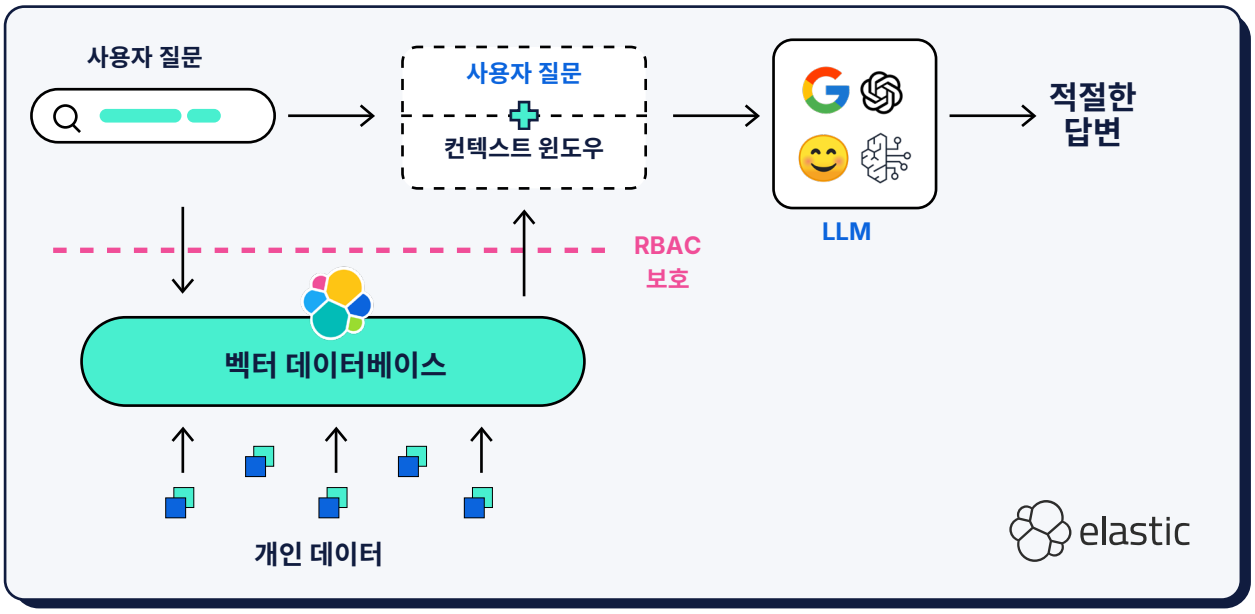
국방 및 정보 커뮤니티 기관은 에어갭 환경 또는 심지어 기술 키트에서 운영해야 하는 추가적인 과제에 직면해 있습니다.

RAG: AI를 위한 효율적인 데이터 검색 모델

많은 공공 부문 조직은 이해관계자가 수행하는 쿼리 유형에 맞게 대규모 언어 모델(LLM)을 훈련하거나 미세 조정하는 데 어려움을 겪고 있습니다. 데이터가 복잡한 경우에도 AI를 통해 자체 독점 데이터를 바로 활용하고자 합니다.

 © 2025 Elasticsearch B.V. All rights reserved.

Retrieval-Augmented Generation(RAG)은 제한된 시간과 리소스, 그리고 방대한 양의 서로 다른 데이터를 확장하는 방법을 둘러싼 더 큰 과제를 해결하는 확장성 있는 솔루션입니다. Elastic의 RAG 모델에서는 사용자 질문이 LLM에 제출되기 전에 먼저 벡터 데이터베이스를 쿼리합니다. 이 벡터 데이터베이스는 벡터 검색을 통해 조직의 자체 독점 데이터에서 관련 컨텍스트를 도출할 수 있습니다. 이 단계, 즉 컨텍스트 레이어는 LLM이 출력에 의존하는 데이터 범위를 크게 줄입니다. 에이전틱 AI는 쿼리가 사용자와 RAG 사이에 있는 AI 에이전트와 상호작용을 함에 따라 이 모델의 특성을 더 변화시킵니다.

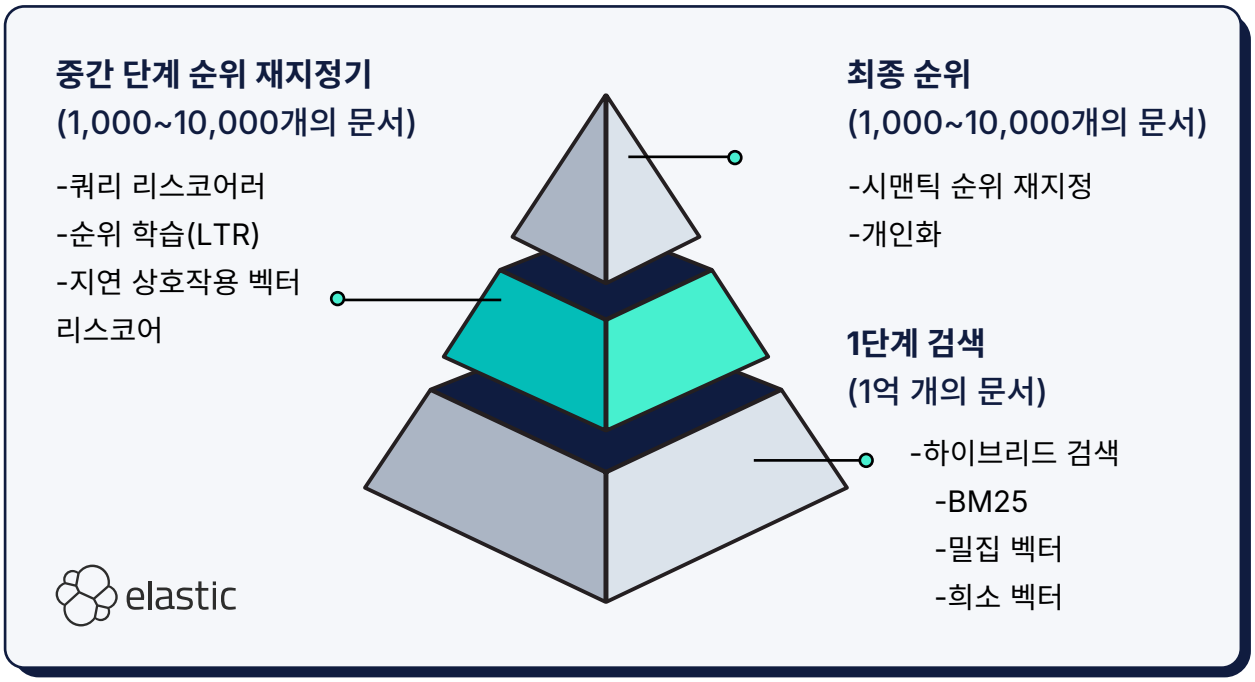


복잡한 데이터 유형을 위한 다단계 검색 및 멀티 모달 검색

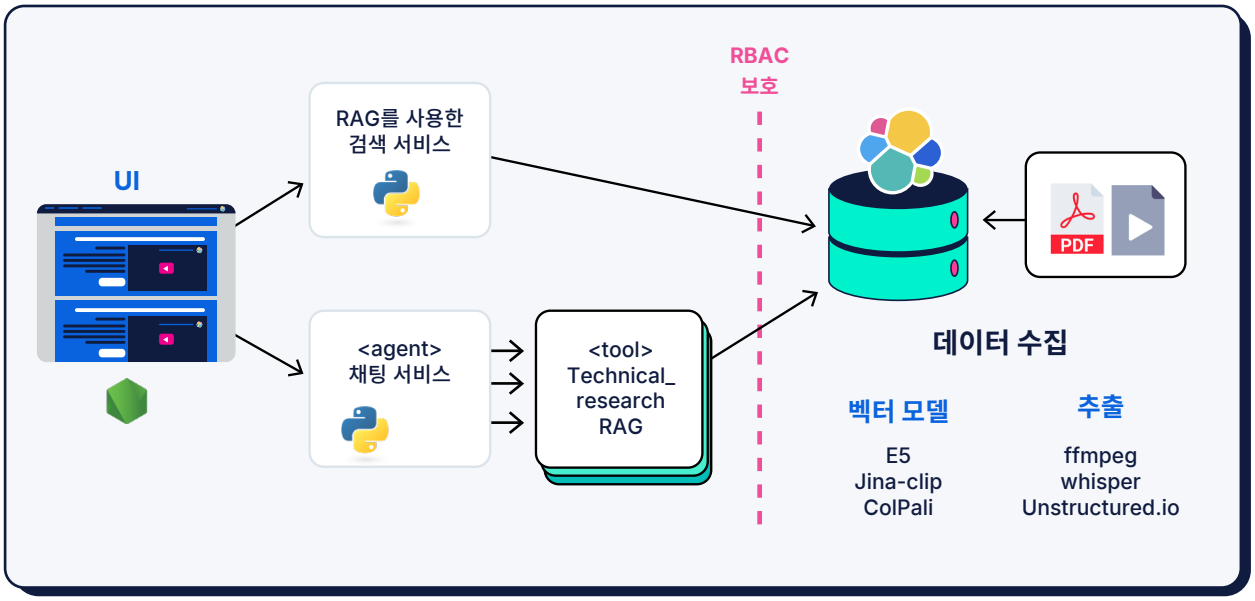
Elastic은 공공 부문 조직이 RAG 접근 방식을 통해 복잡한 데이터를 GenAI에 사용할 수 있도록 지원하기 위해 **다단계 검색**과 **멀티 모달 검색**을 활용합니다.

다단계 검색은 RAG 여정의 복잡한 쿼리를 작은 작업으로 나누어 최종 답변을 제공합니다. 이 접근 방식은 대략적인 결과와 구체적인 결과를 얻기 위해 서로 다른 기술이 필요하다는 점을 기반으로 합니다. 특히 Elastic은 시맨틱 순위 재지정을 위해 완전히 지원되는 에어갭 옵션을 포함한 최초의 벡터 데이터베이스입니다.

 © 2025 Elasticsearch B.V. All rights reserved.



마찬가지로 표가 포함된 PDF나 오디오 및 시각적 구성 요소가 포함된 비디오 등 다양한 데이터 유형에서 정보를 검색할 수 있는 보다 직관적이고 강력한 방법을 찾는 조직이 증가하면서 멀티 모달 검색 사용 사례가 빠르게 주목받고 있습니다. 멀티 모달 검색은 텍스트, 이미지, 오디오 및 비디오 입력을 결합하여 궁극적으로 AI 애플리케이션이 올바른 컨텍스트에서 질문에 더 빠르게 답변할 수 있도록 합니다.

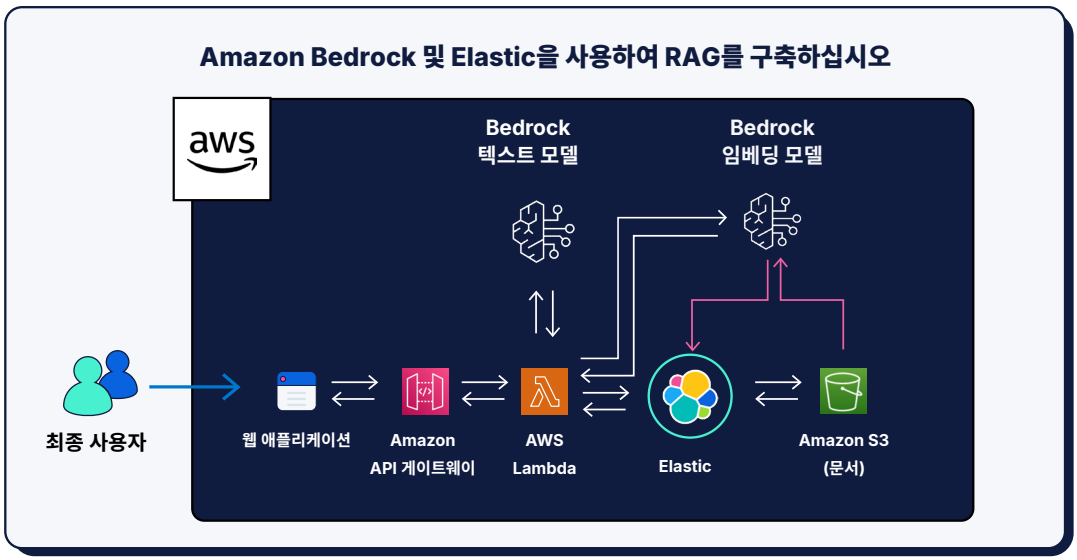


위의 구조를 사용하면 조직에서는 120분짜리 .MOV 파일 깊숙이 포함된 시각적 정보나 1997년 PDF 63페이지의 다이어그램에 있는 문구를 찾을 수 있습니다. 이 기능을 사용하면 복잡하고 독점적인 데이터를 전체적으로 검색하고 액세스하여 생성형 AI 도구가 출력에 사용할 수 있습니다. Elastic은 하위 이미지 검색 기능을 제공하여 전체 이미지나 비디오 프레임에서 상대적으로 작은

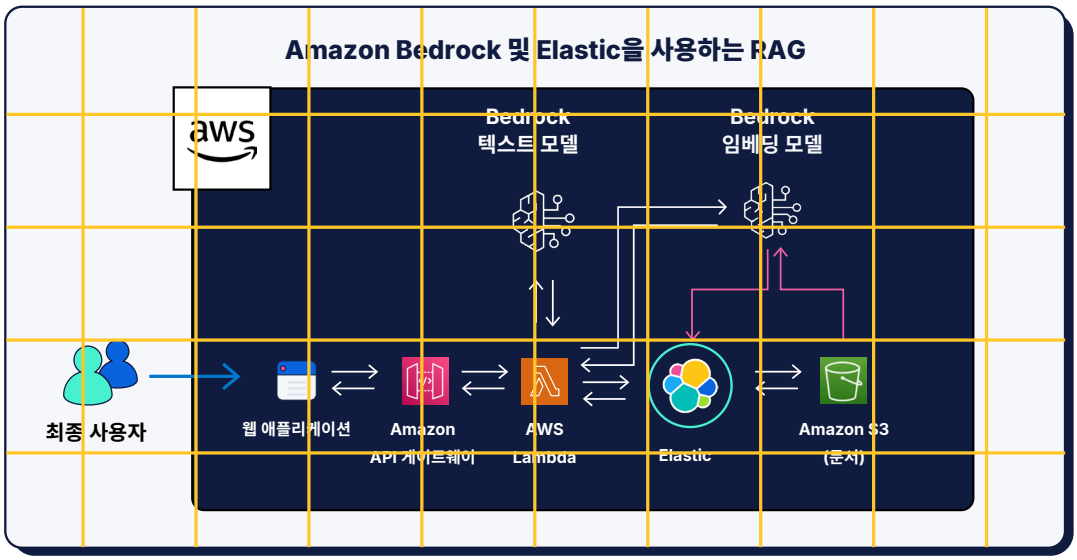
 © 2025 Elasticsearch B.V. All rights reserved.

부분을 차지하는 경우에도 사용자 쿼리와의 전반적인 관련성을 기반으로 검색 결과에서 높은 순위를 차지할 수 있도록 합니다. 어떤 내용이 오디오 트랙에서 논의되고 화면에서는 논의되지 않거나 그 반대의 경우에도, 사용자 질문과 관련이 있다면 이를 찾아 검색 결과에서 높은 순위로 지정할 수 있습니다.

RAG를 사용한 고급 이미지 검색



ColPali 멀티벡터



RAG 접근 방식으로 이미지 파일에서 관련 정보를 어떻게 도출하는지 구체적으로 살펴보겠습니다. 예를 들어 'Amazon Bedrock을 활용한 Elastic 아키텍처 다이어그램'을 검색한다고 가정해 보겠습니다. 위 첫 번째 슬라이드의 정보를 얻으려면 OCR 및 텍스트 추출 작업을 수행하고, 데이터 경로를 유지하는 데 도움이 되는 흐름을 생성해야 할 가능성이 높습니다. 특히 각 문서가 서로 다르기

 © 2025 Elasticsearch B.V. All rights reserved.

때문에 이러한 작업은 복잡해질 수 있습니다. 8페이지 이상에 달하는 여러 열과 표가 있는 긴 정부 PDF를 검색할 경우, 이 작업은 훨씬 더 복잡해집니다. ColPali는 이미지 위에 밀집 벡터 그리드를 생성하여 이러한 작업을 지원합니다.

위 이미지에서 ColPali가 Amazon Bedrock과 Elastic이 텍스트와 로고 형태로 모두 많이 등장하는 특정 영역을 식별했음을 확인할 수 있습니다. Elastic 로고와 참조는 실제로 매우 작지만, 전체 이미지 밀집 벡터와 ColPali 순위 벡터를 활용한 이미지 재순위 지정 덕분에 쿼리가 이 슬라이드를 찾아 제시할 수 있습니다. Elastic은 이러한 기능을 제품에 구축하여 프로세스를 간소화하고 개발자가 조직의 요구 사항을 충족하는 애플리케이션을 더 쉽게 개발할 수 있도록 지원합니다.

Elastic, 공공 부문의 개방적이고 확장성 있는 벡터 검색과 RAG를 위한 솔루션

- 상호 운용 가능한 LLM 독립적 지원
- LLM 교육 및 미세 조정이 필요 없는 RAG 접근 방식
- 애플리케이션 스택 내 도구와의 완전한 사용자 정의 및 통합을 지원하는 잘 문서화된 API 기반 아키텍처
- 역할 기반 및 속성 기반의 통합 보안 제어(RBAC 및 ABAC)
- 세계에서 가장 많이 다운로드된 벡터 데이터베이스
- 모든 환경에서 AI 서비스를 등록할 수 있는 추론 API
- 즉시 사용 가능한 두 가지 벡터 모델 — 고밀도 인코딩을 위한 E5와 희소 인코딩을 위한 ELSER — 및 선택한 다른 임베딩 모델을 사용할 수 있는 기능
- 품질 순위 지정, 리소스 컴퓨팅 및 쿼리 지연 시간 단축에 최적화된 벡터 검색 기능
- 벡터를 압축하여 검색 및 처리 속도를 높이는 Better Binary Quantization(BBQ)
- 고급 검색 모델과 벡터 검색 기능을 활용하는 에이전틱 AI를 위한 확장성 있는 기반

 © 2025 Elasticsearch B.V. All rights reserved.

elastic

다음 단계 수행

벡터 검색이 데이터에서 어떻게 작동하는지 데모를 보려면 공공 부문 기술 팀에 문의하세요.
elastic.co/kr/contact/public-sector.

