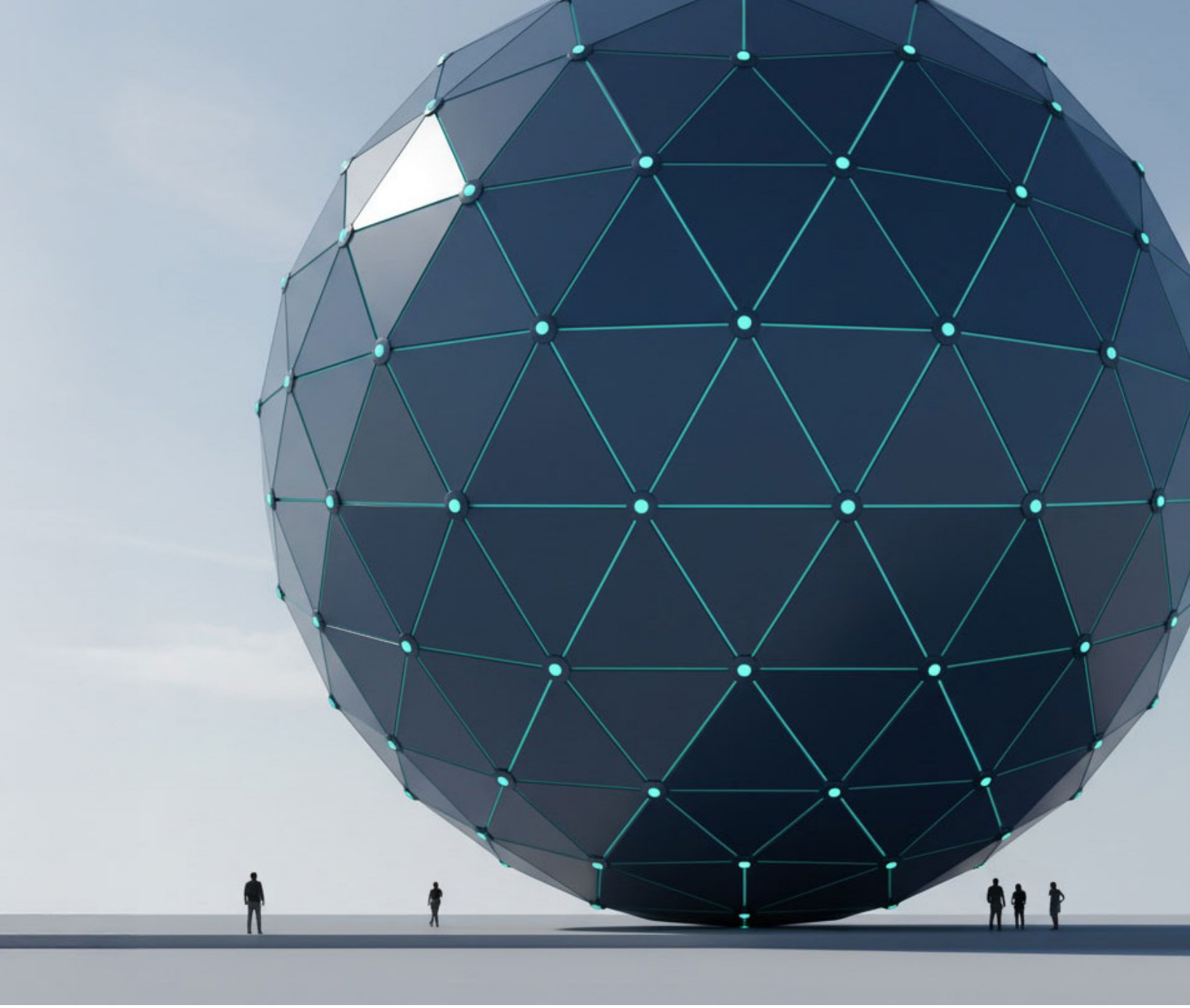


elastic

Busca vetorial avançada para ambientes air-gapped

O setor público está em um momento crítico na adoção da IA generativa (GenAI). As organizações estão sendo pressionadas a lançar iniciativas maiores de IA, estimuladas por estratégias em todo o país, bem como pelas expectativas dos stakeholders. Ao mesmo tempo, GenAI e busca se tornaram interligados. Os usuários estão recorrendo a GenAI interativa para obter respostas rápidas às suas perguntas, em vez de um mecanismo de busca que fornece uma lista de links para filtrar.

Atender a essas expectativas no setor público traz alguns desafios:

1. Existem opções para uso de GenAI, mas não há desenvolvedores suficientes. As organizações precisam investir recursos estrategicamente para gerar os resultados mais eficientes.

2. Organizações do setor público lidam com vastos volumes de dados. Elas precisam de uma base de conhecimento escalonável, rápida e eficiente para usar como pilar para a IA.

3. Os dados nem sempre estão prontos para a IA. As informações críticas geralmente são armazenadas em formatos como PDFs, diagramas e vídeos, dos quais pode ser difícil extrair informações relevantes.

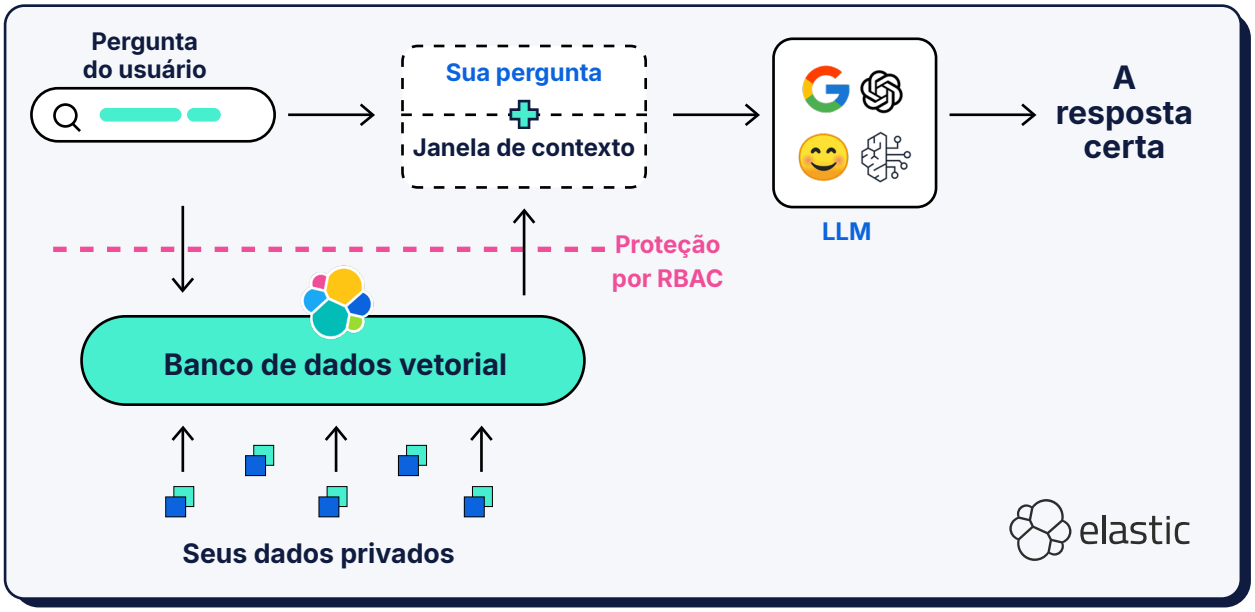
As agências de defesa e da comunidade de inteligência enfrentam o desafio adicional de operar em um ambiente air-gapped, ou mesmo em um kit tecnológico.

RAG: um modelo eficiente de recuperação de dados para IA

Muitas organizações do setor público estão achando difícil treinar ou até mesmo ajustar grandes modelos de linguagem (LLMs) para o tipo de consultas que seus stakeholders estão realizando. Elas querem usar seus próprios dados exclusivos diretamente com IA — mesmo quando esses dados são complexos.

 © 2025 Elasticsearch B.V. Todos os direitos reservados.

Retrieval-Augmented Generation (RAG) é uma solução de escalabilidade que aborda desafios maiores quanto a tempo e recursos limitados, bem como redimensionar grandes quantidades de dados discrepantes. No modelo RAG da Elastic, antes que a pergunta de um usuário seja enviada a um LLM, ele primeiro consulta nosso banco de dados vetorial, que pode revelar o contexto relevante dos dados exclusivos de uma organização por meio de busca vetorial. Essa etapa, ou camada de contexto, reduz significativamente o universo de dados do qual um LLM depende para sua saída. A IA agêntica muda ainda mais a natureza desse modelo, pois uma consulta interage com agentes de IA que ficam entre o usuário e o RAG.

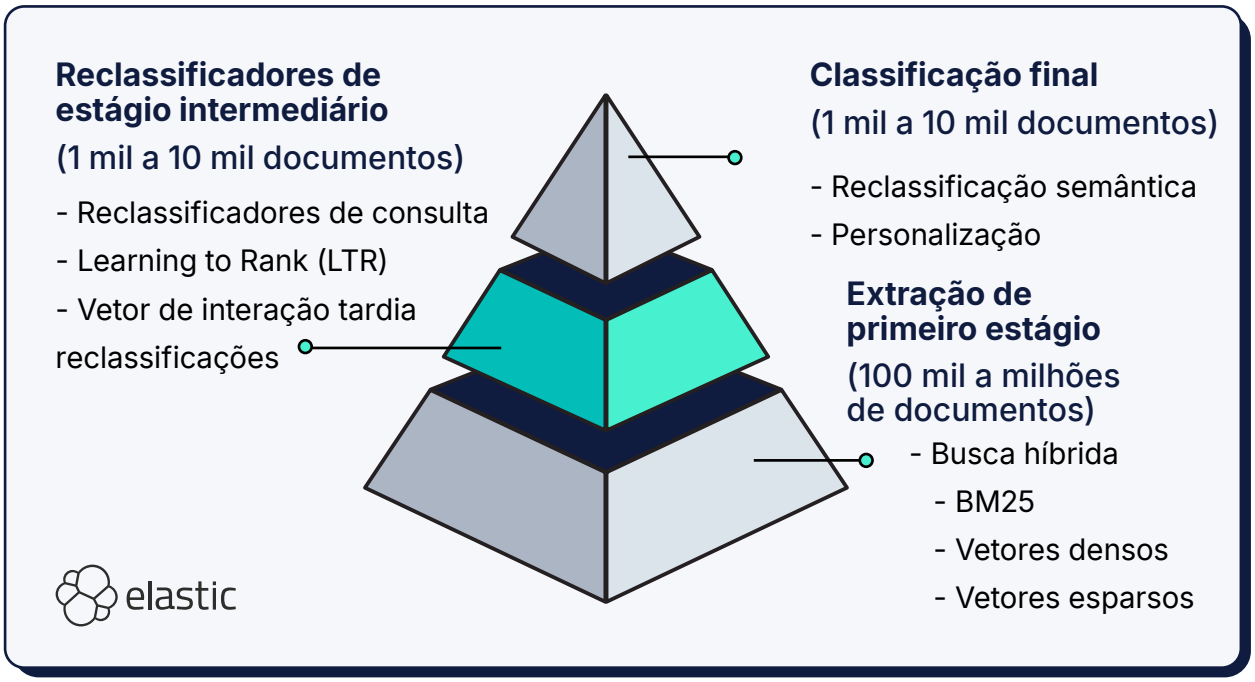


Recuperação em vários estágios e busca multimodal para tipos de dados complexos

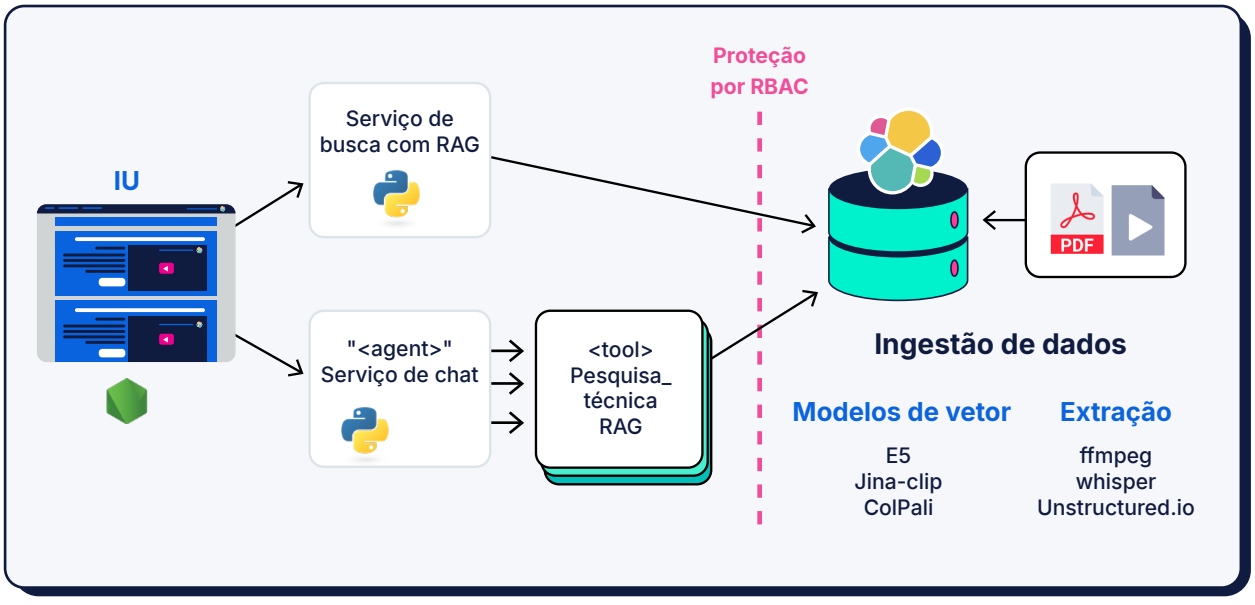
Para ajudar as organizações do setor público a tornar seus dados complexos utilizáveis para GenAI na abordagem RAG, a Elastic utiliza **recuperação em múltiplos estágios** e **busca multimodal**.

A recuperação em vários estágios divide as consultas complexas na jornada do RAG em tarefas menores para fornecer uma resposta final. Essa abordagem é baseada na necessidade de uma tecnologia diferente para resultados aproximados versus resultados específicos. Em especial, a Elastic é o primeiro banco de dados vetorial que inclui uma opção air-gapped totalmente compatível para reclassificação semântica.

 © 2025 Elasticsearch B.V. Todos os direitos reservados.



Da mesma forma, os casos de uso de busca multimodal estão ganhando força rapidamente à medida que as organizações procuram maneiras mais intuitivas e eficazes de extrair informações em diversos tipos de dados, como PDFs com tabelas ou vídeos com componentes de áudio e visuais. A busca multimodal combina entradas de texto, imagem, áudio e vídeo, permitindo que os aplicativos de IA respondam perguntas mais rapidamente, com o contexto certo.

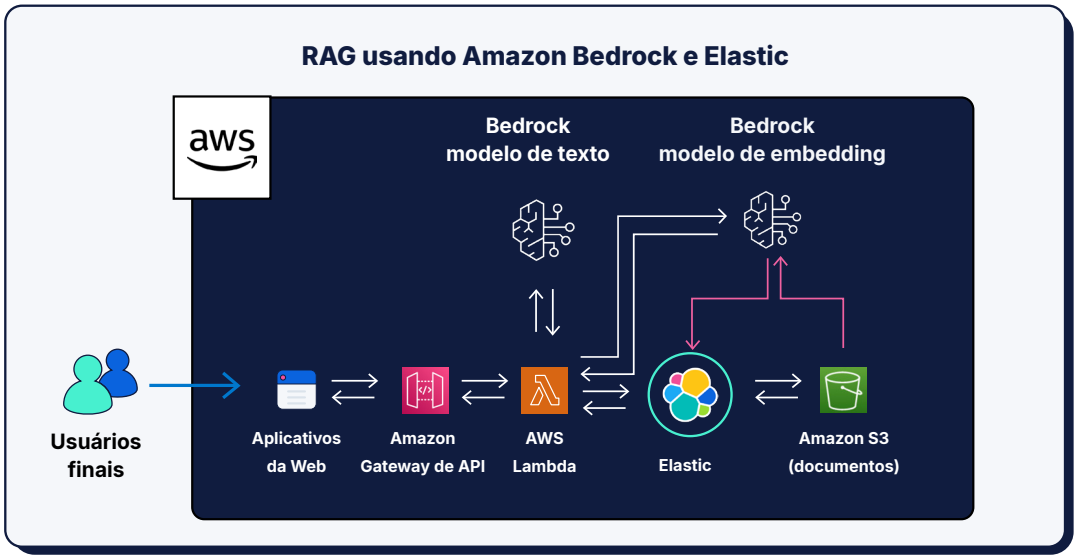


Usando a estrutura acima, as organizações podem encontrar informações visuais embutidas profundamente em um arquivo .MOV de 120 minutos ou em uma frase em um diagrama na página 63 de um PDF de 1997. Esse recurso garante que dados complexos e exclusivos sejam pesquisáveis e acessíveis de forma integral para

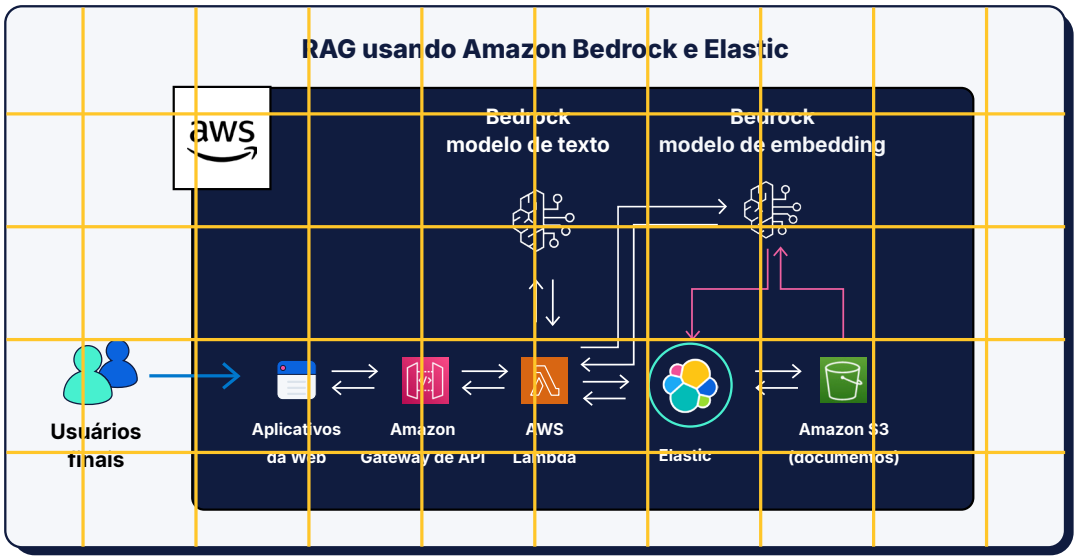
 © 2025 Elasticsearch B.V. Todos os direitos reservados.

ferramentas generativas de IA usarem em suas saídas. A Elastic também permite a busca por subimagem para garantir que, mesmo quando algo ocupa uma porção relativamente pequena de uma imagem geral ou quadro de um vídeo, ainda possa ser classificado altamente nos resultados de busca com base na relevância geral para a consulta do usuário. Se algo é discutido na faixa de áudio e não na tela, ou vice-versa, a Elastic ainda pode encontrá-lo e classificá-lo com destaque nos resultados de busca se for relevante para a pergunta de um usuário.

Busca avançada de imagens com RAG



ColPali multivector



Vejamos especificamente um exemplo de como uma abordagem RAG pode extrair informações relevantes de um arquivo de imagem. Aqui, digamos que queremos buscar por “Diagramas da arquitetura Elastic utilizando o Amazon Bedrock”. Para obter as informações do primeiro slide acima, provavelmente precisaríamos realizar algum

 © 2025 Elasticsearch B.V. Todos os direitos reservados.

OCR e extração de texto, além de criar alguns fluxos para ajudar a reter os caminhos que os dados percorrem. Isso pode se tornar complicado, especialmente porque cada documento será diferente. E ao pesquisar em PDFs governamentais extensos, com várias colunas e tabelas que abrangem mais de 8 páginas, essa tarefa se torna muito mais difícil. O ColPali consegue ajudar com isso porque cria uma grade de vetores densos sobre a imagem.

Podemos ver nas imagens acima que o ColPali identificou certas áreas da imagem onde o Amazon Bedrock e a Elastic são predominantes tanto na forma de texto quanto no logotipo. O logotipo e as referências da Elastic são, na verdade, muito pequenos, mas nossa consulta ainda é capaz de encontrar e apresentar esse slide para nós por causa da combinação de todo o vetor denso da imagem e da reclassificação das imagens utilizando o vetor de classificação ColPali. A Elastic incorporou esses recursos ao produto para simplificar esse processo e facilitar a criação de aplicativos que resolvam as necessidades de suas organizações.

Elastic é sua solução ideal para busca vetorial aberta e escalabilidade e RAG no setor público

- Suporte interoperável e agnóstico a LLM
- Uma abordagem RAG, eliminando a necessidade de treinamento e ajuste fino em LLM
- Uma arquitetura bem documentada, orientada por API, que permite personalização completa e integração com ferramentas na sua pilha de aplicativos
- Controles de segurança integrados baseados em função e em atributo (RBAC e ABAC)
- O banco de dados vetorial mais baixado do mundo
- APIs de inferência para registrar serviços de IA em qualquer ambiente
- Dois modelos de vetores prontos para uso — E5 para codificação densa e ELSER para codificação esparsa — e a capacidade de usar qualquer outro modelo de incorporação de sua escolha
- Recursos de busca vetorial otimizados para garantir qualidade de classificação, recursos computacionais e redução da latência de consulta
- Better Binary Quantization (BBQ) para compactar vetores, tornando-os mais rápidos para pesquisar e processar
- Uma base de escalabilidade para IA agêntica, usando modelos avançados de recuperação e recursos de busca vetorial

 © 2025 Elasticsearch B.V. Todos os direitos reservados.

elastic

Dê os próximos passos

Entre em contato com nossa equipe técnica do setor público para ver uma demonstração de como a busca vetorial pode funcionar com seus dados: elastic.co/pt/contact/public-sector.

