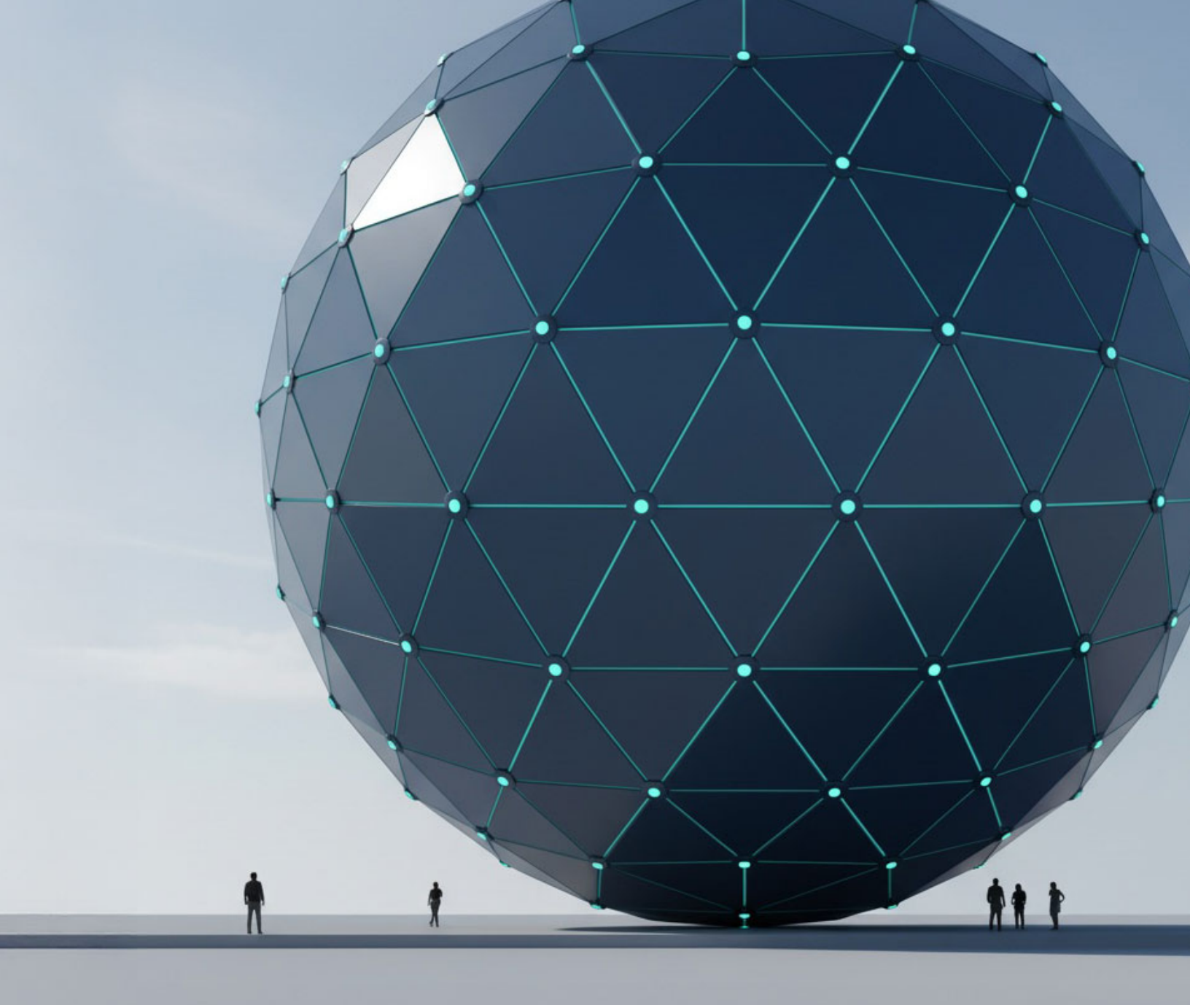


elastic

Búsqueda vectorial avanzada para entornos aislados

El sector público se encuentra en medio de un momento crítico en la adopción de la AI generativa (GenAI). Las organizaciones se están viendo empujadas a implementar iniciativas de AI más grandes, impulsadas por estrategias a nivel nacional así como por las expectativas de las partes interesadas. Al mismo tiempo, la GenAI y la búsqueda se han entrelazado. Los usuarios recurren a GenAI interactiva para obtener respuestas rápidas a sus búsquedas, en lugar de un motor de búsqueda que ofrece una lista de enlaces para examinar.

Cumplir con estas expectativas en el sector público conlleva algunos desafíos:

1. Hay opciones para el uso de GenAI, pero no hay suficientes desarrolladores. Las organizaciones necesitan invertir estratégicamente recursos para producir los resultados más eficientes.

2. Las organizaciones del sector público se enfrentan a grandes volúmenes de datos. Necesitan una base de conocimientos con escalabilidad, rápida y eficiente para usar como base para la AI.

3. Los datos no siempre están listos para la AI. La información crítica a menudo se almacena en formatos como PDF, diagramas y videos, de los que puede ser difícil extraer información relevante.

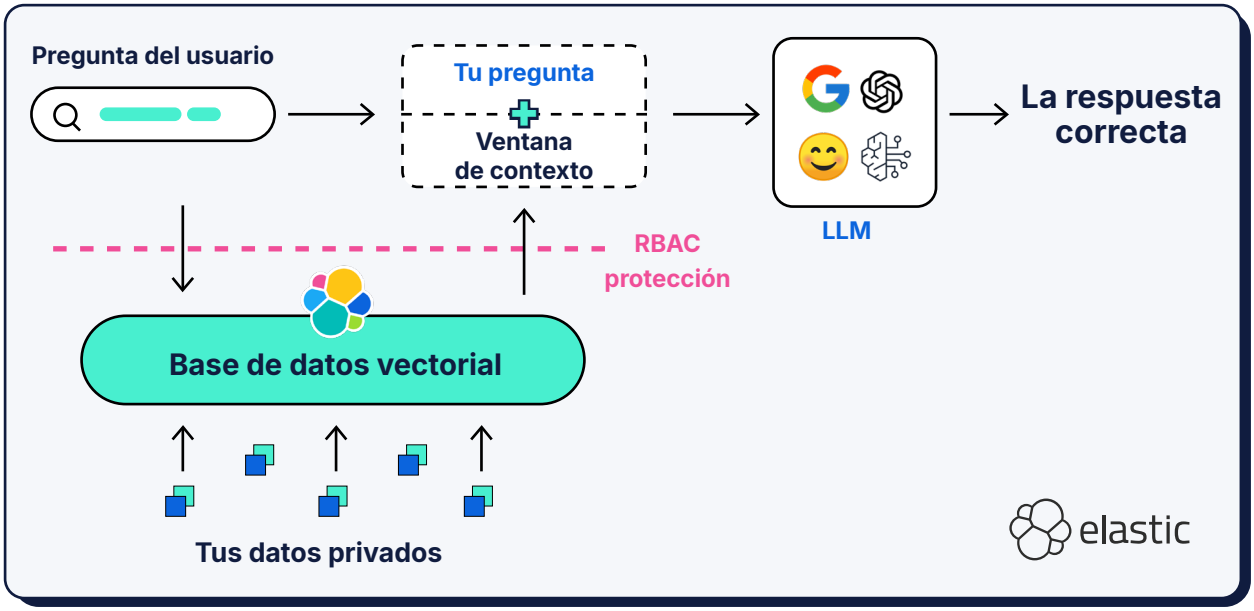
Las agencias de la comunidad de defensa e inteligencia enfrentan el desafío adicional de operar en un entorno aislado, o incluso en un kit tecnológico.

RAG: Un modelo eficiente de recuperación de datos para IA

A muchas organizaciones del sector público les resulta difícil entrenar o incluso ajustar los modelos de lenguaje grande (LLM) para el tipo de búsquedas que sus partes interesadas están realizando. Quieren usar sus datos propietarios de forma inmediata con la AI, incluso cuando esos datos son complejos.

 © 2025 Elasticsearch B.V. Todos los derechos reservados.

Retrieval-Augmented Generation (RAG) es una solución de escalabilidad que aborda desafíos mayores en torno al tiempo y los recursos limitados, así como la forma de escalar grandes cantidades de datos dispares. En el modelo RAG de Elastic, antes de que la pregunta de un usuario se envíe a un LLM, primero busca en nuestra base de datos vectorial, que puede mostrar contexto relevante de los datos propietarios de una organización a través de búsqueda vectorial. Este paso, o capa de contexto, reduce significativamente el universo de datos en el que se basa un LLM para su salida.
La AI agéntica cambia la naturaleza de este modelo aún más, ya que una búsqueda interactúa con agentes de AI que se sitúan entre el usuario y RAG.

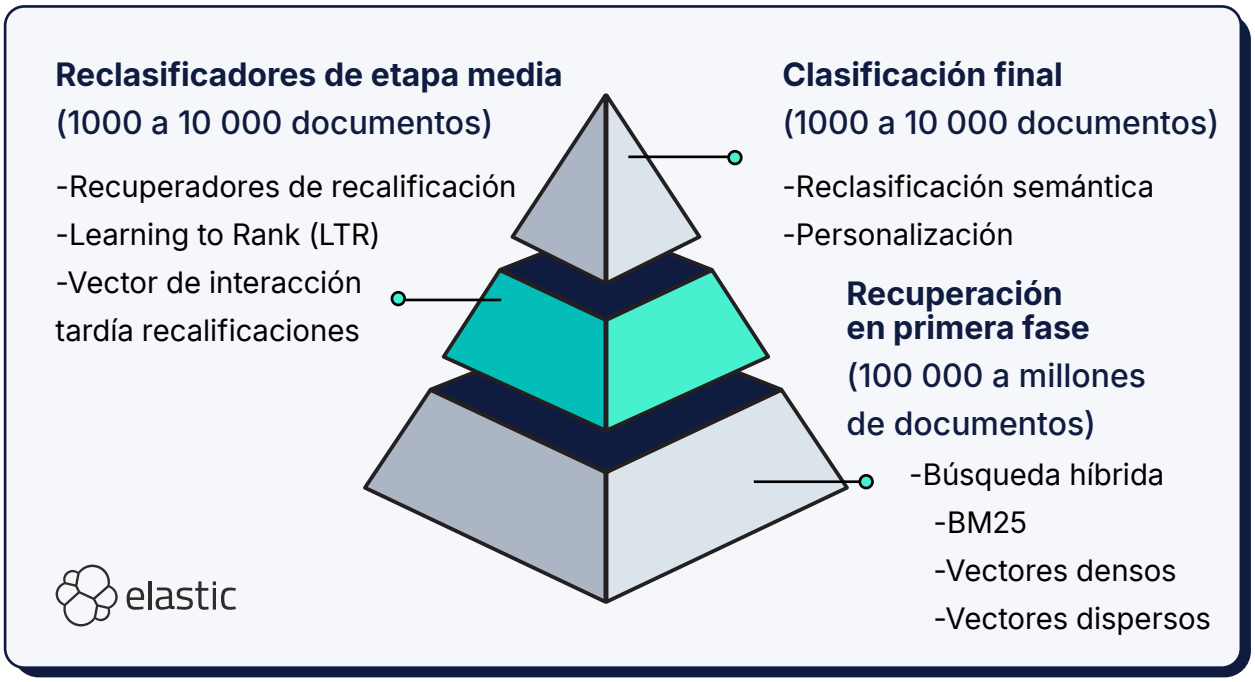


Recuperación en múltiples etapas y búsqueda multimodal para tipos de datos complejos

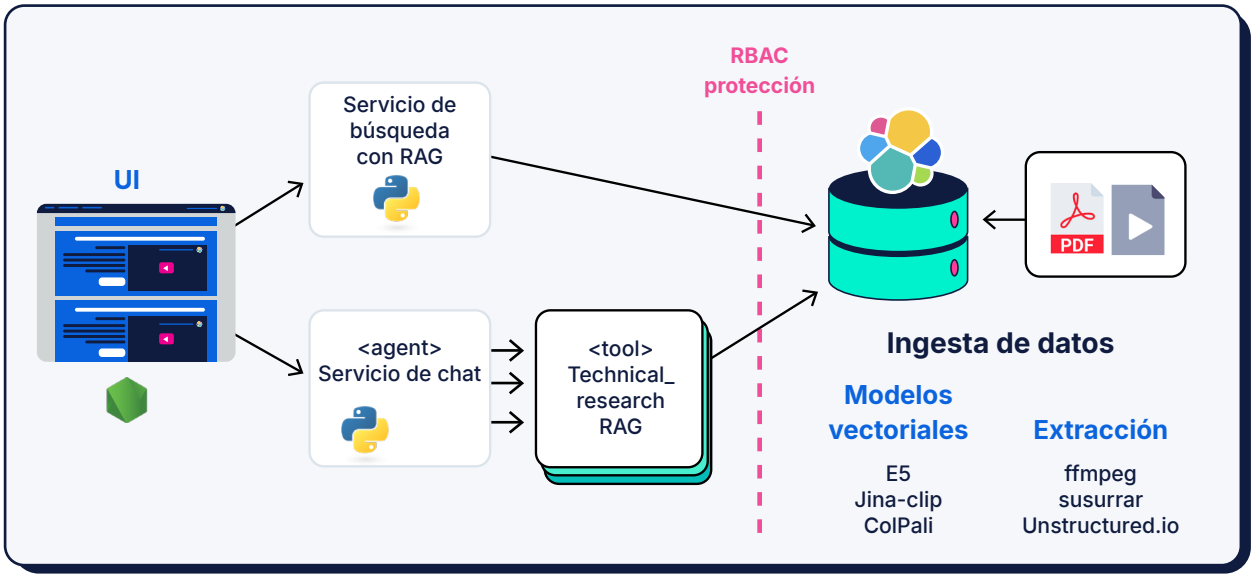
Para ayudar a las organizaciones del sector público a hacer que sus datos complejos sean utilizables para GenAI en el enfoque RAG, Elastic hace uso de **la recuperación en múltiples etapas** y **la búsqueda multimodal**.

La recuperación de múltiples etapas descompone las búsquedas complejas en el recorrido RAG en tareas más pequeñas para entregar una respuesta final. Este enfoque se basa en la necesidad de utilizar tecnologías diferentes para obtener resultados aproximados frente a resultados específicos. Cabe destacar que Elastic es la primera base de datos vectorial que incluye una opción completamente compatible aislada para la reclasificación semántica.

 © 2025 Elasticsearch B.V. Todos los derechos reservados.



Del mismo modo, los casos de uso de búsqueda multimodal están ganando terreno rápidamente a medida que las organizaciones buscan formas más intuitivas y poderosas de recuperar información en diversos tipos de datos, como archivos PDF con tablas o videos con componentes de audio y visuales. La búsqueda multimodal combina entradas de texto, imagen, audio y video, lo que en última instancia permite que las aplicaciones de IA respondan preguntas más rápido y con el contexto adecuado.

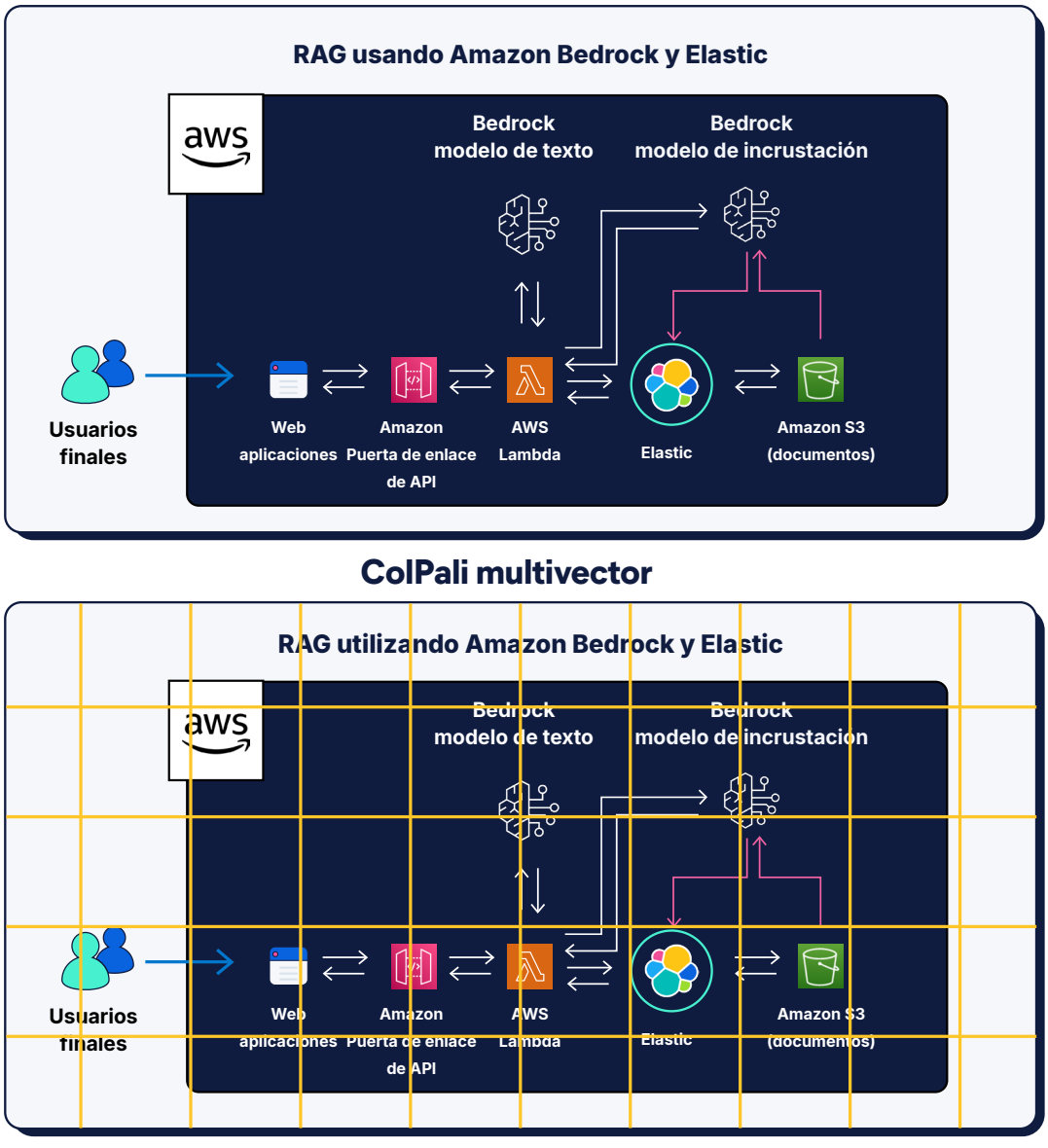


Usando la estructura anterior, las organizaciones pueden encontrar información visual incrustada en lo profundo de un archivo .MOV de 120 minutos o una frase en un diagrama en la página 63 de un PDF de 1997. Esta capacidad garantiza que los datos complejos y propietarios sean integralmente buscables y accesibles para que

 © 2025 Elasticsearch B.V. Todos los derechos reservados.

las herramientas de AI generativa los utilicen en sus salidas. Elastic también permite la búsqueda de subimágenes para garantizar que, incluso cuando algo ocupa una porción relativamente pequeña de una imagen o fotograma general de un video, aún pueda clasificarse con alta prioridad en los resultados de búsqueda según la relevancia general para la búsqueda del usuario. Ya sea que algo se discuta en la pista de audio y no en la pantalla, o viceversa, Elastic aún puede encontrarlo y clasificarlo de manera destacada en los resultados de búsqueda si es relevante para la pregunta de un usuario.

Búsqueda avanzada de imágenes con RAG



Veamos específicamente un ejemplo de cómo un enfoque RAG puede mostrar información relevante de un archivo de imagen. Por ejemplo, queremos hacer una búsqueda de “Diagramas de arquitectura Elastic utilizando Amazon Bedrock”. Para obtener la información de la primera diapositiva anterior, es probable que necesitemos

 © 2025 Elasticsearch B.V. Todos los derechos reservados.

hacer algo de OCR y extracción de texto y crear algunos flujos para retener las rutas que toman los datos. Eso puede volverse complicado, especialmente porque cada documento será diferente. Y cuando buscas en archivos PDF gubernamentales largos con varias columnas y tablas que abarcan más de 8 páginas, esto se convierte en una tarea mucho mayor. ColPali puede ayudar con esto porque crea una cuadrícula de vectores densos en la parte superior de la imagen.

Podemos ver en las imágenes de arriba que ColPali ha identificado ciertas áreas de la imagen donde Amazon Bedrock y Elastic prevalecen tanto en forma de texto como de logotipo. El logotipo de Elastic y las referencias son en realidad bastante pequeños, pero nuestra consulta aún puede encontrar y presentarnos esta diapositiva debido a la combinación de todo el vector denso de la imagen y la reclasificación de las imágenes utilizando el vector de clasificación ColPali. Elastic ha incorporado estas capacidades en el producto para simplificar este proceso y facilitar a los desarrolladores la creación de aplicaciones que resuelvan las necesidades de sus organizaciones.

Elastic es tu solución para la búsqueda vectorial abierta y escalable y RAG en el sector público

- Soporte interoperable e independiente de LLM
- Un enfoque RAG, que elimina la necesidad de entrenamiento y ajuste de LLM
- Una arquitectura bien documentada y basada en API que facilita la personalización e integración completas con las herramientas de tu pila de aplicaciones
- Controles de seguridad integrados basados en roles y atributos (RBAC y ABAC)
- La base de datos vectorial más descargada del mundo
- API de inferencia para registrar servicios de IA en cualquier entorno
- Dos modelos vectoriales listos para usar: E5 para codificado denso y ELSER para codificado disperso, y la capacidad de usar cualquier otro modelo de incrustación de tu elección
- Capacidades de búsqueda vectorial optimizadas para la calidad de clasificación, los recursos informáticos y la reducción de la latencia de las búsquedas
- Better Binary Quantization (BBQ) para comprimir vectores, haciéndolos más rápidos de buscar y procesar.
- Una base de escalabilidad para la AI agéntica, que aprovecha los modelos de recuperación avanzados y las capacidades de búsqueda vectorial

 © 2025 Elasticsearch B.V. Todos los derechos reservados.

elastic

Da los siguientes pasos

Ponte en contacto con nuestro equipo técnico del sector público para ver una demostración de cómo la búsqueda vectorial puede funcionar con tus datos: elastic.co/es/contact/public-sector.

