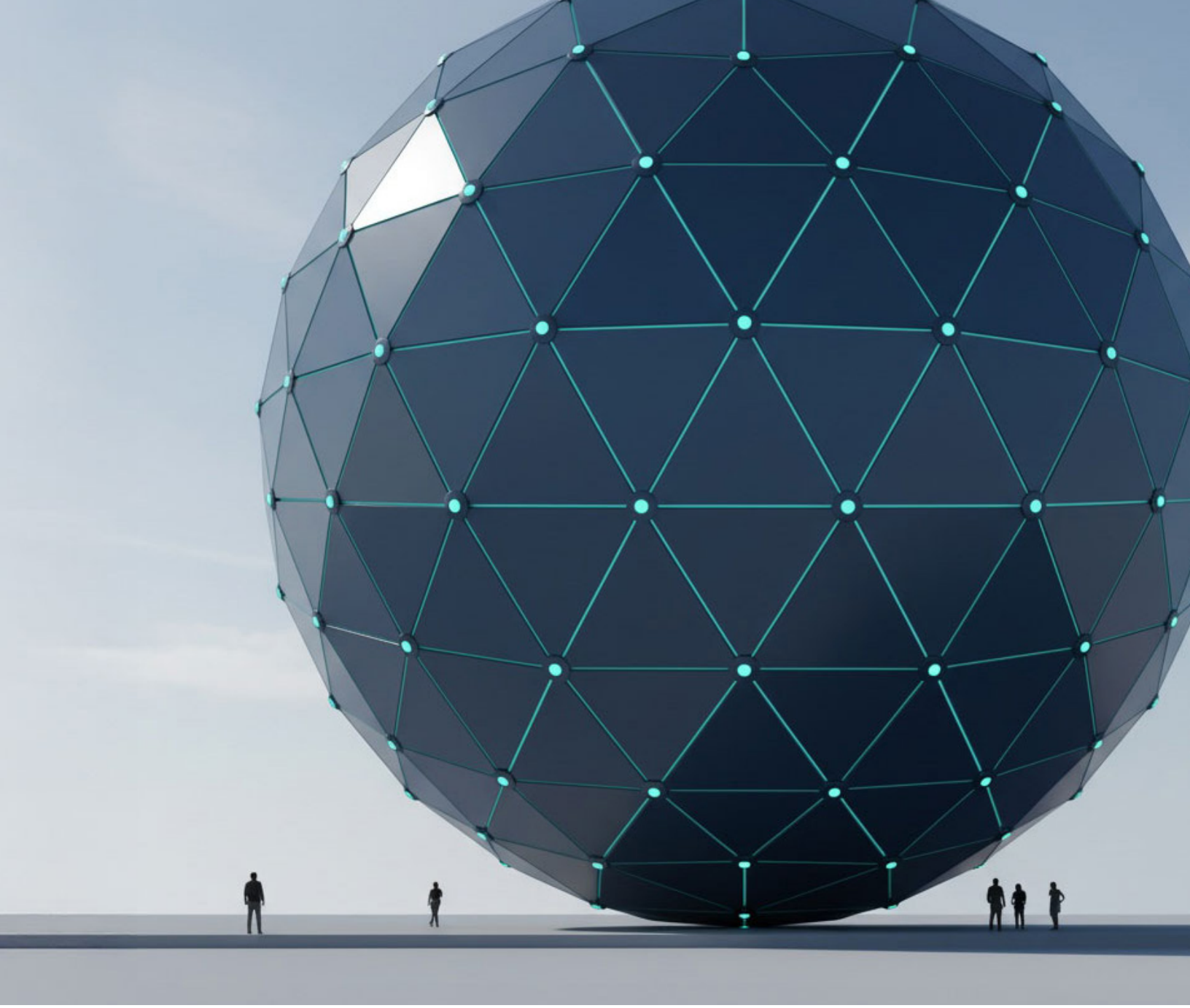


elastic

Rechercher vectorielle avancée pour les environnements air gap

Le secteur public se trouve au cœur d'une période critique en matière d'adoption de l'IA générative (GenAI). Les organisations sont incitées à déployer des initiatives d'IA à plus grande échelle, stimulées par des stratégies nationales et les attentes des parties prenantes. Parallèlement, l'IA générative et la recherche sont devenues indissociables. Les utilisateurs se tournent vers l'IA générative interactive pour obtenir des réponses rapides à leurs questions, plutôt que vers un moteur de recherche qui fournit une liste de liens à parcourir.

Répondre à ces attentes dans le secteur public comporte quelques défis :

1. Il existe des options pour l'utilisation de l'IA générative, mais il n'y a pas suffisamment de développeurs. Les organisations doivent investir leurs ressources de manière stratégique pour obtenir les résultats les plus efficaces.

2. Les organisations du secteur public doivent faire face à d'énormes volumes de données. Ils ont besoin d'une base de connaissances évolutive, rapide et efficace qui servira de fondement à l'IA.

3. Les données ne sont pas toujours prêtes pour l'IA. Les informations critiques sont souvent stockées dans des formats tels que les PDF, les diagrammes et les vidéos, à partir desquels il peut être difficile d'extraire les informations pertinentes.

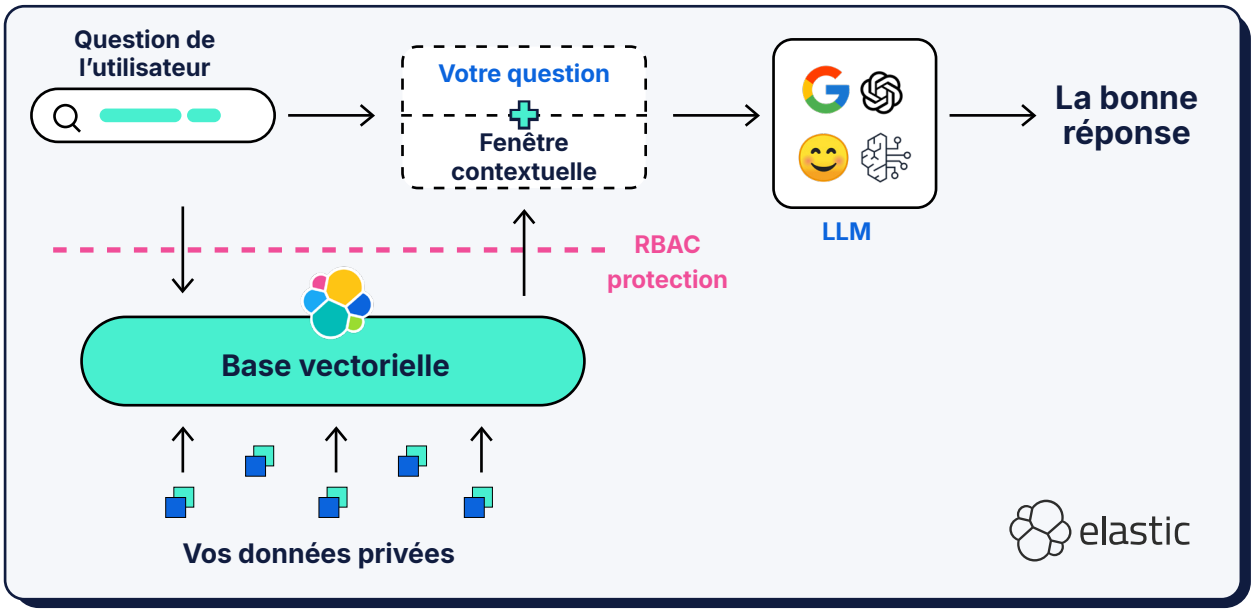
Les agences de défense et de renseignement sont confrontées au défi supplémentaire de devoir fonctionner dans un environnement isolé, voire sur un kit technologique.

RAG : un modèle de récupération de données efficace pour l'IA

De nombreuses organisations du secteur public rencontrent des difficultés pour former ou même affiner les grands modèles linguistiques (LLM) afin qu'ils répondent aux requêtes formulées par leurs parties prenantes. Elles souhaitent utiliser leurs données propriétaires telles quelles avec l'IA, même lorsque ces données sont complexes.

 © 2025 Elasticsearch B.V. Tous droits réservés.

Génération augmentée de récupération (RAG) est une solution de scalabilité qui répond aux défis majeurs liés au temps et aux ressources limités, ainsi qu'à la mise à l'échelle de vastes quantités de données disparates. Dans le modèle RAG d'Elastic, avant qu'une question d'utilisateur ne soit soumise à un LLM, elle interroge d'abord notre base vectorielle, qui peut faire apparaître le contexte pertinent à partir des données propriétaires d'une organisation via une recherche vectorielle. Cette étape, ou couche contextuelle, réduit considérablement l'univers de données sur lequel s'appuie un LLM pour sa sortie. L'IA agentique modifie encore davantage la nature de ce modèle, car une requête interagit avec des agents IA qui se trouvent entre l'utilisateur et la RAG.

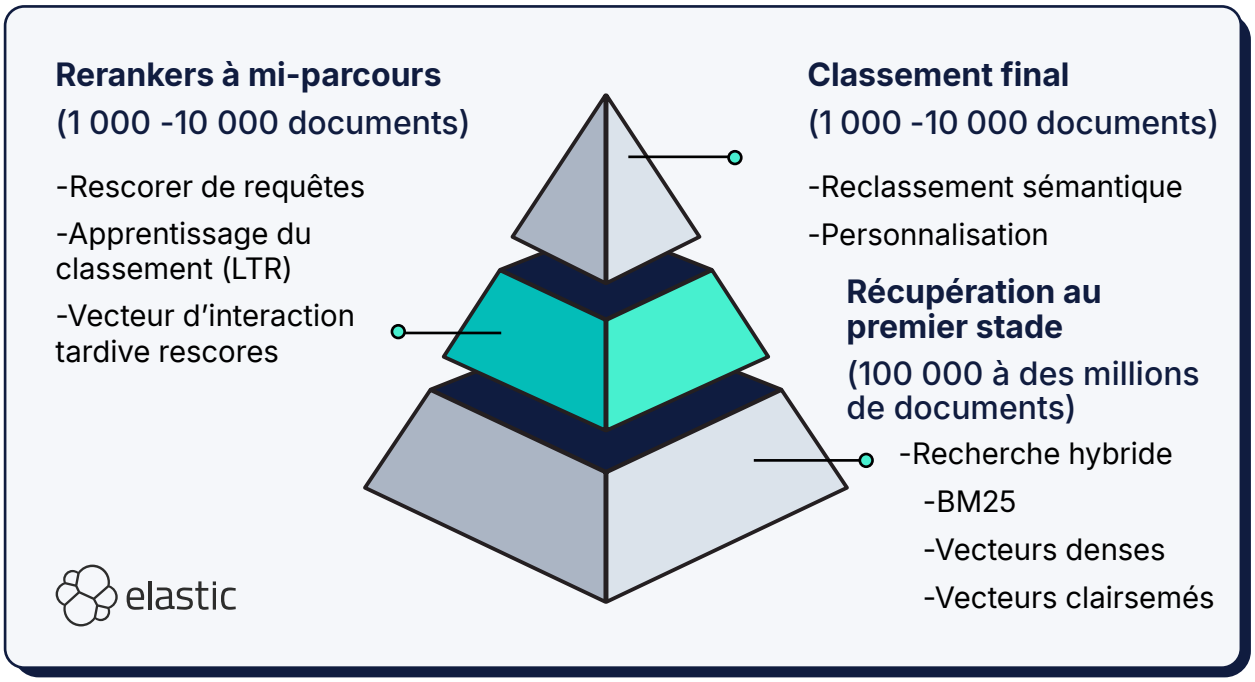


Récupération en plusieurs étapes et recherche multimodale pour les types de données complexes

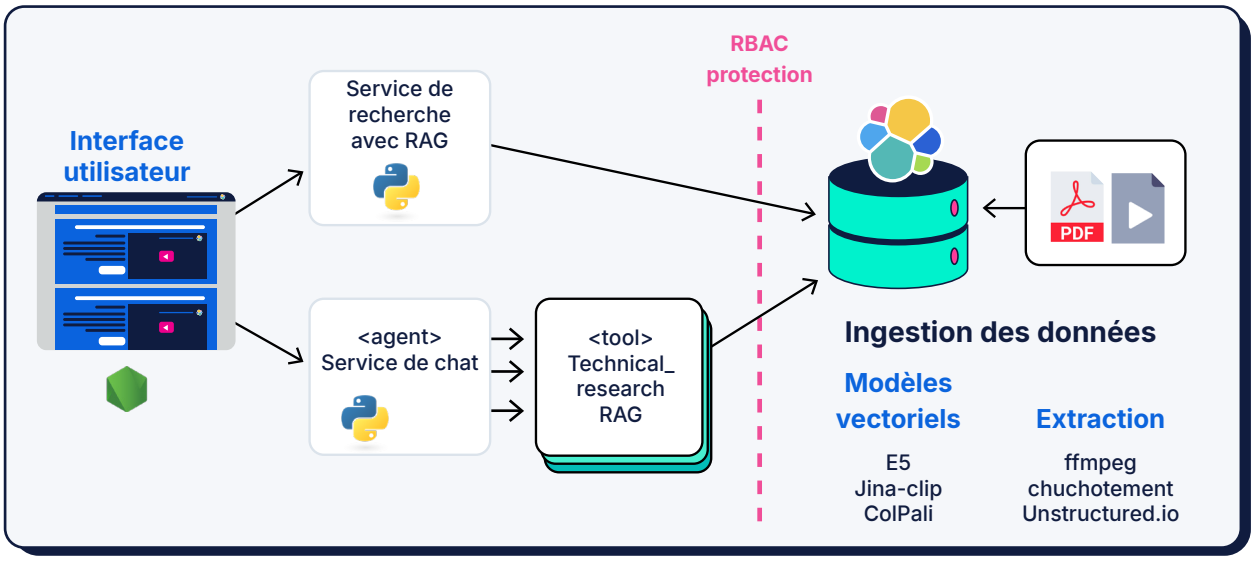
Afin d'aider les organisations du secteur public à rendre leurs données complexes exploitables pour l'IA générique dans le cadre de l'approche RAG, Elastic exploite **la recherche en plusieurs étapes** et la **recherche multimodale**.

La recherche en plusieurs étapes décompose les requêtes complexes du parcours RAG en tâches plus petites afin de fournir une réponse finale. Cette approche repose sur la nécessité d'une technologie différente pour obtenir des résultats approximatifs par rapport à des résultats spécifiques. Il est à noter qu'Elastic est la première base vectorielle à inclure une option air-gapped entièrement prise en charge pour le reclassement sémantique.

 | © 2025 Elasticsearch B.V. Tous droits réservés.



De même, les cas d'utilisation de la recherche multimodale gagnent rapidement en popularité, les organisations recherchant des moyens plus intuitifs et plus puissants pour extraire des informations à partir de divers types de données, tels que des fichiers PDF contenant des tableaux ou des vidéos avec des composants audio et visuels. La recherche multimodale combine des entrées texte, image, audio et vidéo, permettant ainsi aux applications d'IA de répondre plus rapidement aux questions, en tenant compte du contexte approprié.

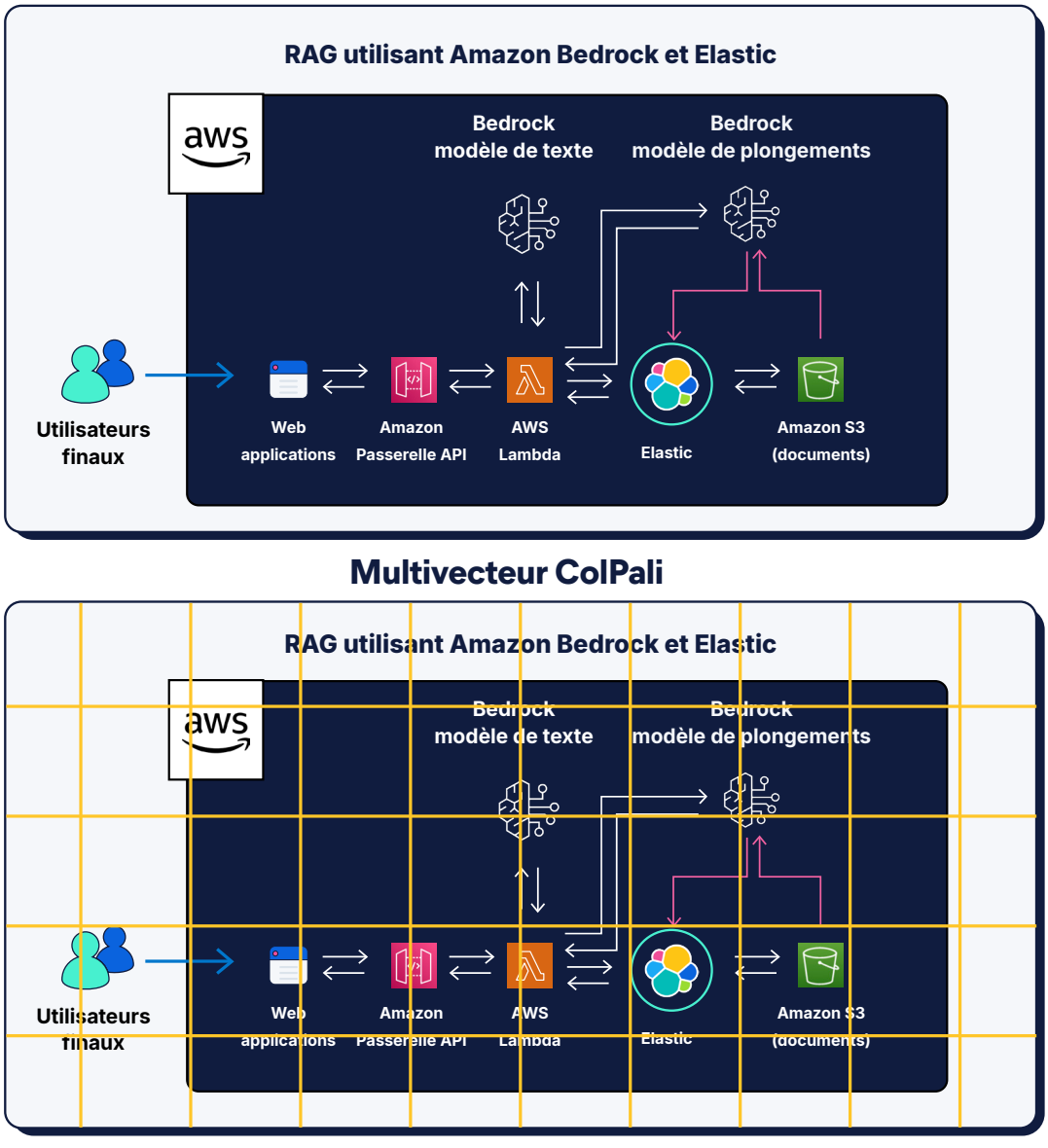


Grâce à la structure ci-dessus, les organisations peuvent trouver des informations visuelles enfouies dans un fichier .MOV de 120 minutes ou une phrase dans un diagramme à la page 63 d'un PDF datant de 1997. Cette fonctionnalité garantit que les données complexes et propriétaires sont entièrement consultables et accessibles pour

 © 2025 Elasticsearch B.V. Tous droits réservés.

les outils d'IA générative qui les utilisent dans leurs sorties. Elastic permet également la recherche d'images partielles afin de garantir que même lorsqu'un élément occupe une partie relativement petite d'une image ou d'une image vidéo, il puisse tout de même être bien classé dans les résultats de recherche en fonction de sa pertinence globale par rapport à la requête de l'utilisateur. Qu'un élément soit mentionné dans la piste audio et non à l'écran, ou vice versa, Elastic peut tout de même le trouver et le classer en tête des résultats de recherche s'il est pertinent par rapport à la question de l'utilisateur.

Recherche avancée d'images avec la RAG



Examinons plus précisément un exemple illustrant comment une approche RAG peut faire ressortir des informations pertinentes à partir d'un fichier image. Ici, supposons que nous souhaitions rechercher « Diagrammes de l'architecture Elastic utilisant Amazon Bedrock ». Afin d'obtenir les informations de la première diapositive

 © 2025 Elasticsearch B.V. Tous droits réservés.

ci-dessus, nous devrions probablement effectuer une reconnaissance optique de caractères (OCR) et une extraction de texte, puis créer des flux pour aider à conserver les chemins empruntés par les données. Cela peut devenir compliqué, d'autant plus que chaque document sera différent. Et lorsque l'on recherche dans de longs PDF gouvernementaux avec plusieurs colonnes et tableaux qui s'étendent sur plus de 8 pages, cela devient une demande beaucoup plus importante. ColPali peut nous aider dans cette tâche, car il crée une grille de vecteurs denses sur l'image.

On peut voir sur les images ci-dessus que ColPali a identifié certaines zones de l'image où Amazon Bedrock et Elastic sont prédominants, à la fois sous forme de texte et de logo. Le logo Elastic et les références sont en réalité assez petits, mais notre requête parvient tout de même à trouver et à nous présenter cette diapositive grâce à la combinaison du vecteur dense de l'image entière et du reclassement des images à l'aide du vecteur de classement ColPali. Elastic a intégré ces fonctionnalités au produit afin de simplifier ce processus et de permettre aux développeurs de créer plus facilement des applications qui répondent aux besoins de leurs organisations.

Elastic est votre solution pour la recherche vectorielle ouverte et scalabilité et la RAG dans le secteur public

- Prise en charge interopérable et indépendante du LLM
- Une approche RAG, qui élimine le besoin de formation et d'ajustement des LLM
- Une architecture bien documentée, pilotée par API, qui permet une personnalisation complète et une intégration avec les outils de votre pile d'applications
- Contrôles de sécurité intégrés basés sur les rôles et les attributs (RBAC et ABAC)
- La base vectorielle la plus téléchargée au monde
- API d'inférence pour enregistrer des services d'IA dans n'importe quel environnement
- Deux modèles vectoriels prêts à l'emploi, E5 pour l'encodage dense et ELSER pour l'encodage clairsemé, et la possibilité d'utiliser tout autre modèle d'intégration de votre choix
- Fonctionnalités de recherche vectorielle optimisées pour améliorer la qualité du classement, les ressources informatiques et réduire la latence des requêtes
- Meilleure quantification binaire (BBQ) pour compresser les vecteurs, les rendant plus rapides à rechercher et à traiter
- Une base de scalabilité pour l'IA agentique, tirant parti de modèles de recherche avancés et de capacités de recherche vectorielle

 © 2025 Elasticsearch B.V. Tous droits réservés.

elastic

Passez aux étapes suivantes

Contactez notre équipe technique du secteur public pour voir une démonstration de la façon dont la rechercher vectorielle peut fonctionner avec vos données : elastic.co/fr/contact/public-sector.

