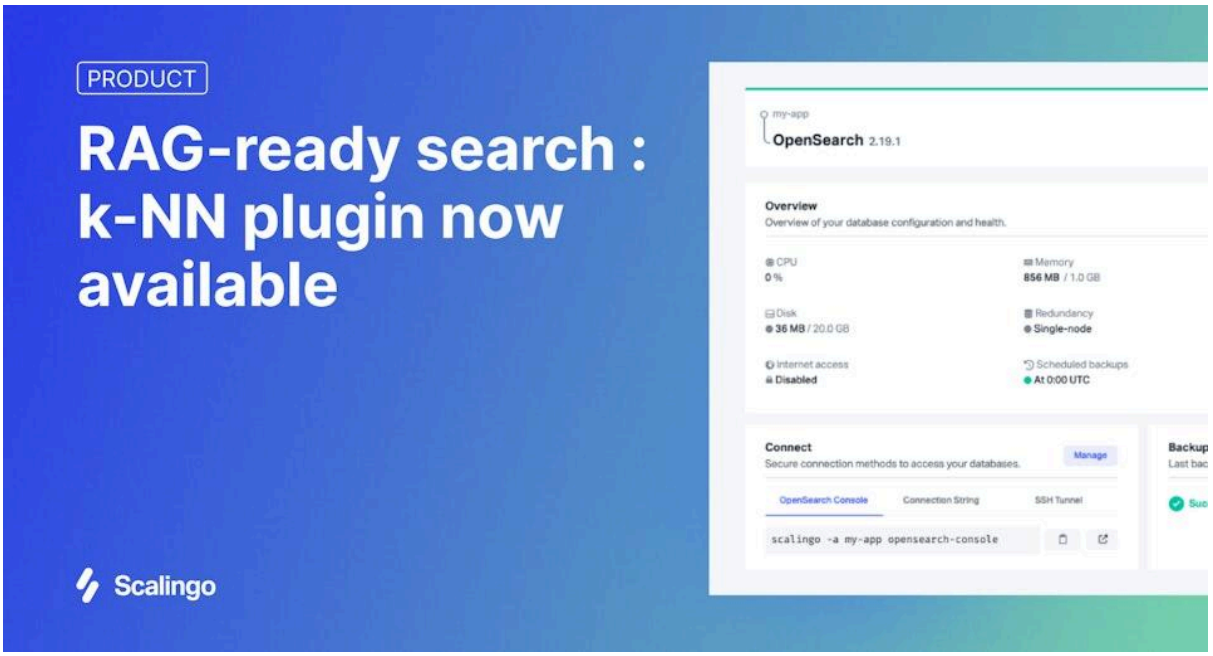
Recherche vectorielle sur Scalingo : le plugin k-NN disponible pour OpenSearch®

12 juin 2025 - 5 min de lecture



Nous sommes ravis d'annoncer que Scalingo for OpenSearch® prend désormais en charge le **plugin k-NN**, ouvrant la voie à une recherche sémantique basée sur les vecteurs. C'est une avancée majeure pour notre plateforme managée, conçue pour rendre l'IA générative et les technologies de search intelligentes à la fois accessibles, sécurisées et simples à mettre en œuvre : le tout, hébergé en France.

Pourquoi la recherche vectorielle change la donne

Les moteurs de recherche traditionnels s'appuient sur des mots-clés et la correspondance exacte de termes. La recherche vectorielle, elle, permet de **rechercher par sens**, plutôt que par mots. En représentant les contenus et les requêtes sous forme de vecteurs numériques, il devient possible de retrouver les documents et contenus les plus pertinents, même s'ils ne contiennent pas les mots exacts tapés par l'utilisateur.

Cela ouvre la voie à de nombreux cas d'usage :

- Recherche sémantique dans des outils internes ou bases de connaissances
- Récupération intelligente de documents
- Assistants IA basés sur la **génération augmentée par récupération (RAG)**

- Expériences de recherche plus naturelles, proches de la manière dont on pense et formule ses questions

Le plugin k-NN, en quelques mots

Le plugin k-NN (pour k-nearest neighbors) est la brique essentielle qui permet tout cela. Il permet à OpenSearch® de stocker et rechercher des embeddings vectoriels, ces fameuses représentations numériques de contenus générées par des modèles comme `all-MiniLM` d'OpenAI ou `text-embedding-3-small` de Hugging Face.

Résultat ? OpenSearch® peut renvoyer des résultats basés sur la **proximité sémantique**, et pas seulement sur des chaînes de caractères.

C'est rapide, scalable, et parfaitement adapté aux **applications de type GenAI**, où la pertinence contextuelle est clé.

Ce que cela change sur Scalingo

Grâce à l'intégration du plugin k-NN, Scalingo for OpenSearch® devient aussi une base de données vectorielle, en plus de ses capacités de recherche et d'observabilité déjà puissantes.

Concrètement, vous pouvez désormais :

- Stocker et interroger des vecteurs d'embedding
- Combiner recherche structurée et recherche sémantique
- Alimenter des fonctionnalités GenAI comme des chatbots, assistants documentaires ou interfaces de connaissance

Tout cela avec les garanties Scalingo :

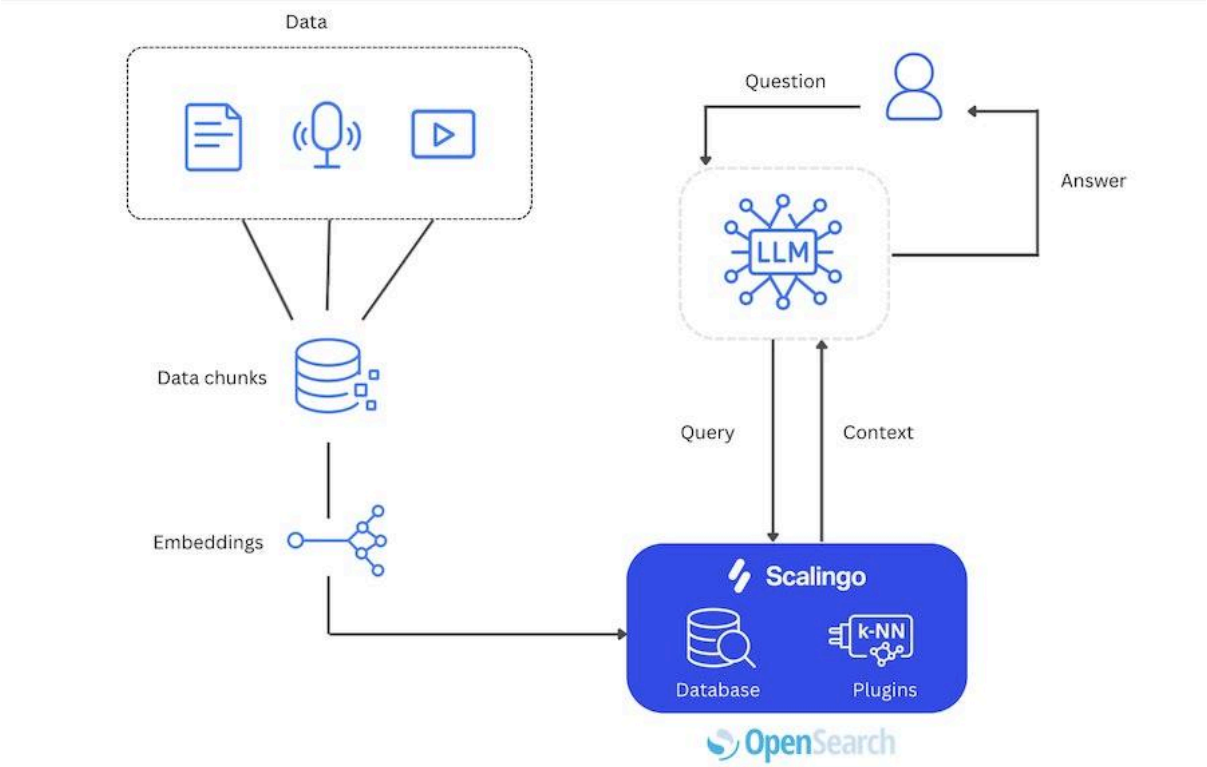
- Infrastructure 100 % managée
- Conformité ISO 27001 et HDS
- Hébergement en France
- Intégration fluide à vos applications
- Facturation transparente à la minute

Exemple : Exploitez votre base documentaire interne avec une recherche boostée par l'IA

Votre équipe dispose probablement d'une masse importante de documents internes : spécifications techniques, guides d'onboarding, procédures, notes de réunion… Jusqu'ici, les retrouver reposait sur des recherches par mots-clés, souvent longues et approximatives. Si, comme nous, vous utilisez Notion comme base documentaire, vous voyez le tableau.

Avec la recherche vectorielle et la **génération augmentée par récupération (RAG)**, vous pouvez passer à une recherche intelligente. Comment ? En vous appuyant sur l'écosystème open source pour générer et indexer vos vecteurs, et sur OpenSearch pour propulser votre recherche sémantique en quatre étapes :

1. Ingestion documentaire : Commencez par préparer vos documents (quel que soit le type de contenu que vous souhaitez exploiter) pour la recherche sémantique. Cela passe généralement par un découpage du texte (chunking), puis la génération d'embeddings vectoriels à l'aide d'un modèle comme `all-MiniLM` (OpenAI) ou `text-embedding-3-small` (Hugging Face). Des outils comme LangChain, LlamaIndex, ou un script sur-mesure peuvent vous aider à orchestrer cette étape.
2. Indexation des vecteurs : Les embeddings générés sont ensuite envoyés vers une instance OpenSearch® hébergée sur Scalingo. Grâce au plugin k-NN, ils sont stockés dans un index vectoriel conçu pour effectuer des recherches par similarité de manière performante et scalable.
3. Requêtes contextuelles : Lorsqu'un utilisateur pose une question via un LLM (modèle de langage de grande taille) comme ChatGPT ou Mistral AI, la requête est elle aussi transformée en vecteur. OpenSearch® utilise alors la recherche vectorielle pour retrouver les documents les plus proches sémantiquement.
4. Génération de la réponse: Les documents ainsi retrouvés sont transmis au LLM en tant que contexte, lui permettant de générer une réponse pertinente, fiable, et directement basée sur votre corpus documentaire.



Résultat : Un assistant IA qui hallucine rarement, qui reste connecté à votre base de connaissances interne, et qui peut répondre aux questions de vos utilisateurs en langage naturel. Le tout propulsé par OpenSearch®, directement sur Scalingo.

Lancez-vous dès aujourd’hui

La recherche vectorielle est dès à présent disponible pour tous les utilisateurs d’OpenSearch® sur Scalingo.

Que vous construisiez un moteur de recherche interne plus intelligent ou que vous exploriez des cas d’usage autour des assistants IA, cette fonctionnalité ouvre un tout nouveau champ des possibles.

👉 Pour démarrer, consultez notre documentation sur les plugins OpenSearch.

Et si vous avez besoin d’un coup de main, n’hésitez pas à contacter notre équipe !