

VEILLE TECHNOLOGIQUE · BTS SIO SLAM · THÈME : BASES DE DONNÉES

Les Bases de Données Vectorielles

— *Le Moteur Caché de l'IA Générative*

Protection des données · RAG · Embeddings · pgvector, Pinecone, Weaviate

CANIAUX Romain · 19 Mars 2026

Dans le cadre de mon BTS SIO SLAM, je poursuis ma veille sur les Bases de Données. Après les BDD en mémoire (veille n°1) et le chiffrement homomorphe (veille n°2), j'étudie aujourd'hui les bases de données vectorielles — au cœur de l'IA générative. Ces BDD répondent à un problème concret : comment une IA peut-elle retrouver de l'information pertinente parmi des millions de documents, en temps réel, sans exposer les données sensibles ?

0. MÉTHODOLOGIE DE VEILLE

Comment j'ai trouvé ces informations : Pour cette veille, j'ai défini des mots-clés ciblés : "vector database 2026", "pgvector RAG", "base de données vectorielle IA générative", "embeddings sécurité RGPD". J'ai ensuite utilisé plusieurs outils complémentaires pour collecter, filtrer et vérifier les sources.

Outil	Usage	Fréquence / rôle
Google Alerts	Alertes automatiques sur mots-clés : "vector database 2026", "pgvector", "RAG IA"	<i>Surveillance continue</i>
Feedly / RSS	Agrégation des flux de IT Social, Le Monde Informatique, InfoQ	<i>Lecture hebdomadaire</i>
Google Scholar	Recherche d'articles académiques sur les BDD vectorielles et le FHE	<i>Source académique</i>
LinkedIn / X (Twitter)	Suivi de comptes spécialisés : ingénieurs ML, chercheurs, startups IA	<i>Tendances et actualités</i>
MDN / Docs officielles	Documentation pgvector, Weaviate, Pinecone pour vérification technique	<i>Fiabilité des sources</i>

Critères de sélection des sources : J'ai conservé uniquement les articles publiés après septembre 2025 (moins de 6 mois), rédigés par des acteurs reconnus du secteur (éditeurs spécialisés, entreprises IT, chercheurs académiques). J'ai écarté les articles sans date, les forums et les contenus purement marketing. L'étude académique de Morris et al. (arXiv) a été intégrée comme référence de fond pour ancrer le sujet de la sécurité des embeddings.

1. PRÉSENTATION DU SUJET

Pourquoi ce sujet ? En explorant les BDD en mémoire (veille n°1), j'avais identifié que la RAM permettait des accès ultra-rapides aux données. Les BDD vectorielles exploitent ce même principe : elles stockent des **embeddings** (représentations mathématiques du sens d'un texte ou d'une image) en mémoire pour permettre une recherche par *similarité sémantique* plutôt que par mot-clé exact. C'est le moteur du **RAG (Retrieval-Augmented Generation)**, la technique derrière ChatGPT, Copilot ou les assistants IA d'entreprise.

Ce qu'est une BDD vectorielle : Contrairement à une base classique qui stocke des valeurs textuelles ou numériques, une BDD vectorielle stocke des vecteurs à haute dimension représentant le sens des données. Une recherche ne retourne pas le résultat exact, mais les entrées *les plus proches sémantiquement* de la requête (algorithme HNSW — Hierarchical Navigable Small World).

Figure 1 — Architecture RAG : Question → Embedding → BDD Vectorielle → Contexte → LLM → Réponse

2. RÉSUMÉ DES ARTICLES

IT Social — Base vectorielle & RAG (Mars 2026) :

IT Social dresse un état du marché : les BDD vectorielles sont valorisées à 2,65 milliards de dollars en 2025, avec une croissance de 27,5 % par an (objectif : 9 milliards d'ici 2030). L'article souligne que l'IA agentique transforme la BDD vectorielle en infrastructure critique : sans elle, les agents IA ne peuvent pas accéder à la bonne information au bon moment. Gartner confirme que la majorité des organisations en auront une en production d'ici 2026.

Diskod.ma — Guide complet BDD vectorielles 2026 (Mars 2026) :

Ce guide technique compare les solutions. pgvector (extension PostgreSQL open source) permet d'ajouter un type 'vector' dans les tables existantes sans nouvelle infrastructure. Pinecone est recommandé pour la production managée, Qdrant pour l'auto-hébergement haute performance. Les cas d'usage cités : chatbots RAG, recherche sémantique, recommandation, détection de doublons dans les documents.

Kaliop — RAG et LLM : quelle approche ? (Février 2026) :

L'article analyse les architectures RAG en 2026. La mise en oeuvre exige de maîtriser le découpage des documents (chunking), la vectorisation (embeddings) et l'orchestration des BDD vectorielles. L'approche hybride 'Self-Route' est présentée comme la plus performante : le système utilise le RAG pour les questions précises et bascule vers une analyse à long contexte pour les questions complexes.

Ayinedjimi Consultants — Sécurité des Embeddings (Septembre 2025) :

Cet article, appuyé par l'étude académique de Morris et al. (arXiv, 2023), démontre que les embeddings ne sont pas anonymes : il est possible de récupérer jusqu'à 92 % du contenu original depuis un vecteur. La solution avancée : combiner BDD vectorielle et FHE (chiffrement homomorphe — cf. ma veille n°2) pour effectuer des recherches de similarité sans jamais déchiffrer les vecteurs. Coût actuel encore élevé : réservé aux secteurs très sensibles (santé, défense, finance).

3. CE QUE CETTE TECHNOLOGIE APPORTE ET SES LIMITES

Apports :	Limites :
Recherche par sens, pas par mot-clé exact	Coût mémoire élevé pour les index (HNSW)
Moteur du RAG et de toute l'IA générative	pgvector limité au-delà de 100 millions de vecteurs
pgvector intégré à PostgreSQL (sans nouvelle infra)	Les embeddings peuvent être inversés (risque RGPD)
Compatible LangChain, LlamaIndex, AWS Bedrock	FHE + BDD vectorielle : coût calcul encore prohibitif

4. CAS D'USAGE CONCRETS

D'après les articles consultés, les BDD vectorielles s'appliquent directement dans des secteurs soumis à de fortes contraintes réglementaires (RGPD, AI Act) :

- Santé : analyse de dossiers médicaux chiffrés sans exposer les patients, même au prestataire cloud — combinaison BDD vectorielle + FHE (lien direct avec ma veille n°2).
- Finance : scoring bancaire et détection de fraude sur données vectorisées, avec conformité RGPD garantie — cas documenté par Kaliop.
- Cloud / Entreprise : assistants IA internes (chatbots RAG) qui interrogent la base documentaire de l'entreprise sans réentraîner le LLM — cas d'usage central selon IT Social.

5. ESPRIT CRITIQUE

Ce que j'ai trouvé convaincant : Les chiffres de marché cités par IT Social (+27,5 %/an) sont cohérents avec ce qu'on observe dans les offres d'emploi et les annonces de recrutement tech en 2025-2026. pgvector comme alternative pragmatique à PostgreSQL est un argument solide : pas besoin de changer d'infrastructure, ce qui facilite vraiment l'adoption en entreprise.

Ce qui mérite d'être nuancé : Certains articles (notamment Diskod.ma) sont rédigés par des prestataires ayant un intérêt commercial à promouvoir ces technologies. Les chiffres de récupération de données depuis les embeddings (92 % selon Morris et al.) concernent des modèles spécifiques — ils ne s'appliquent pas uniformément à toutes les solutions du marché. Enfin, la combinaison FHE + BDD vectorielle reste aujourd'hui un sujet de recherche plus qu'une solution déployée en production.

6. MON ANALYSE PERSONNELLE

Cette veille m'a permis de relier mes trois sujets en un fil conducteur cohérent. Les BDD en mémoire (veille n°1) fournissent la vitesse nécessaire à la recherche vectorielle en temps réel. Le chiffrement homomorphe (veille n°2) apporte la couche de sécurité pour traiter des embeddings sensibles sans les exposer. Les BDD vectorielles constituent la brique applicative qui unit les deux. En tant que futur développeur SLAM, pgvector me permet d'intégrer de l'IA sémantique dans une application PostgreSQL existante sans changer d'infrastructure — une compétence concrète et directement applicable en stage ou en emploi.

CONCLUSION

Les bases de données vectorielles sont passées du laboratoire à la production en moins de trois ans, portées par l'explosion de l'IA générative et du RAG. Combinées aux BDD en mémoire pour la vitesse et au chiffrement homomorphe pour la confidentialité, elles forment un écosystème complet pour des applications IA respectueuses des données. Dans un contexte où le RGPD et l'AI Act imposent des standards renforcés, maîtriser ces technologies représente un atout concret pour le développeur SLAM de demain.

7. SOURCES CONSULTÉES (FORMAT APA)

Source	Titre de l'article	Année	Lien
IT Social	Base vectorielle et RAG, nouvelle infrastructure critique de l'IA agentique	Mars 2026	https://itsocial.fr/cloud-infrastructure-it/cloud-infrastructure-it-articles/base-vectorielle-et-rag-nouvelle-infrastructure-critique-de-lentreprise-agentique/
Diskod.ma	Bases de données vectorielles pour l'IA 2026 : Guide complet	Mars 2026	https://diskod.ma/bases-donnees-vectorielles-ia/
Kaliop	RAG et LLM : quelle approche pour une meilleure performance ?	Fév. 2026	https://www.kaliop.com/fr/rag-et-llm/
Ayinedjimi Consultants	Sécurité et confidentialité des Embeddings en entreprise	Déc. 2024	https://www.ayinedjimi-consultants.fr/securite-confidentialite-embeddings/
arXiv (académique)	Morris et al. — Text Embeddings Reveal (Almost) As Much As Text	2023 (réf.)	arxiv.org/abs/2310.06816