

Mis à jour le 27/02/2026

S'inscrire

Formation Milvus : Base de Données Vectorielle pour Applications IA

2 jours (14 heures)

Présentation

Milvus est une base de données vectorielle open source conçue pour stocker, indexer et rechercher efficacement des embeddings à grande échelle. Elle constitue une brique centrale des architectures modernes d'IA, notamment pour la recherche sémantique et les systèmes RAG.

Notre formation vous permettra de comprendre le fonctionnement d'une base vectorielle, de configurer l'indexation et d'optimiser la recherche de similarité dans des contextes réels.

Vous apprendrez à structurer vos données (vecteurs + métadonnées), à choisir les index adaptés, à intégrer Milvus dans un pipeline RAG, et à explorer d'autres cas d'usage comme la recommandation, le multimodal et la détection d'anomalies.

À l'issue de cette formation, vous serez capable de concevoir une architecture de recherche vectorielle fiable, performante et industrialisable, adaptée aux contraintes de production.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs

- Comprendre le fonctionnement d'une base vectorielle.
- Configurer l'indexation et optimiser la recherche de similarité avec Milvus.
- Mettre en place une recherche hybride (vecteur + métadonnées).
- Intégrer Milvus dans un pipeline RAG et recherche sémantique.
- Préparer un déploiement robuste en environnement de production.

Public visé

- Data Engineers
- Développeurs backend
- Ingénieurs IA / ML
- Architectes data
- DevOps orientés IA

Pré-requis

- Bases Python et usage d'API
- Notions sur les embeddings

Formation Milvus : Base de Données Vectorielle pour Applications IA

[Jour 1 - Matin]

Fondamentaux des bases vectorielles

- Comprendre les embeddings et leurs usages (recherche sémantique, recommandation, RAG)
- Différence entre base relationnelle et base vectorielle
- Principes de la recherche de similarité (ANN)
- Métriques : cosine, L2, inner product
- Positionnement de Milvus dans l'écosystème des bases vectorielles
- Atelier pratique : Créer une collection, insérer des vecteurs et exécuter une recherche top-k.

[Jour 1 - Après-midi]

Architecture interne et modèle de données Milvus

- Composants et services clés de Milvus
- Collections, partitions, champs scalaires et métadonnées
- Persistance, segments et cycle de vie des données
- Gestion des index : objectifs, contraintes, compromis
- Notions de scalabilité et de performance

Indexation et optimisation des requêtes

- Choisir un index selon le cas d'usage
- Paramètres d'index et de recherche (recall/latence/coût)
- Recherche hybride : vecteur + filtres sur métadonnées
- Stratégies de batching, pagination, et dimensionnement
- Méthodologie de benchmark et indicateurs de qualité
- Atelier pratique : Comparer deux index et optimiser une requête (latence vs recall).

[Jour 2 - Matin]

Milvus pour le RAG et la recherche sémantique

- Rappels : pipeline RAG (chunking, embeddings, retrieval, génération)
- Stratégies de chunking et gestion du contexte
- Stockage des métadonnées : sources, permissions, scoring
- Intégration avec frameworks (ex : LlamaIndex, LangChain)
- Évaluation : pertinence, hallucinations, qualité de retrieval
- Atelier pratique : Construire un mini-RAG avec Milvus.

[Jour 2 - Après-midi]

Autres cas d'usage : recommandation, multimodal, anomalies

- Recommandation : similarité item-to-item, cold start et filtres métier
- Multimodal : texte + image (embeddings CLIP/vision)
- Détection d'anomalies : voisinage, seuils et dérive
- Organisation des données : stratégie de métadonnées et versionning
- Anti-patterns courants et bonnes pratiques

Industrialisation et mise en production

- Déploiement : standalone, cluster, environnements containerisés
- Haute disponibilité, sauvegarde et reprise
- Observabilité : métriques, logs, SLO, alerting
- Sécurité : isolation, contrôle d'accès, gouvernance des données
- Stratégie d'évolution : re-indexation, montée en charge, coûts
- Atelier pratique : Checklist production + plan d'architecture cible.

Sociétés concernées

Cette formation s'adresse à la fois aux particuliers ainsi qu'aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d'auto-évaluation nous permettant d'apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d'anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l'acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.