Pinecone, Weaviate & Milvus : Mise en Œuvre de Bases Vectorielles pour le RAG et le Semantic Search

Lien : https://innov-systems.com/formation/pinecone-weaviate-milvus-mise-en-oeuvre-de-bases-vectorielles-pour-le-rag-et-le-semantic-search

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI81

CATÉGORIE
Intelligence Artificielle Générative (IAG)

Innov Systems

OBJECTIFS DE LA FORMATION

À l'issue de cette formation, vous serez capable de :

- Comprendre les principes des bases vectorielles
- Maîtriser Pinecone, Weaviate et Milvus
- Concevoir une architecture RAG complète
- Implémenter un moteur de recherche sémantique
- Optimiser l’indexation et la recherche de vecteurs
- Intégrer les bases vectorielles avec des LLM
- Améliorer la pertinence des systèmes de recherche IA
- Déployer des solutions scalables en production
- Gérer la performance et la latence des requêtes vectorielles
- Exploiter les embeddings pour les cas d’usage métier
- Mettre en place des architectures de données modernes
- Étudier des cas concrets de RAG en entreprise

POUR QUI ?

- Data Engineers
- Machine Learning Engineers
- Data Scientists
- Architectes data
- Développeurs IA
- MLOps Engineers
- Backend Developers
- Architects cloud
- Product Managers techniques
- Consultants IA
- SRE / DevOps Engineers
- Toute personne impliquée dans les systèmes de recherche IA et RAG

Innov Systems

Programme détaillé

1 / Introduction aux bases vectorielles

- Évolution des bases de données vers le vectoriel
- Cas d’usage du RAG (Retrieval-Augmented Generation)
- Recherche sémantique vs recherche classique

2 / Fondements de l’embedding et du vector search

- Principe des embeddings (texte, image, audio)
- Représentation vectorielle des données
- Mesures de similarité (cosine, distance euclidienne)

3 / Pinecone : architecture et usage

- Base vectorielle managée cloud
- Indexation et requêtes de similarité
- Scalabilité et performance en production

4 / Weaviate : base vectorielle open source

- Architecture modulaire et schema-based
- Modules NLP et intégrations IA
- Déploiement cloud et on-premise

5 / Milvus : base vectorielle haute performance

- Architecture distribuée
- Gestion de très grands volumes de vecteurs
- Optimisation GPU et scalabilité horizontale

Innov Systems

6 / Indexation et stockage des embeddings

- Techniques d’index (HNSW, IVF, PQ)
- Compression et optimisation mémoire
- Gestion du cycle de vie des données vectorielles

7 / Mise en œuvre du RAG (Retrieval-Augmented Generation)

- Pipeline ingestion → embedding → retrieval → génération
- Intégration avec LLM (Claude, GPT, etc.)
- Amélioration de la précision des réponses IA

8 / Semantic Search avancé

- Recherche contextuelle et hybride
- Filtrage metadata + vector search
- Cas d’usage entreprise

9 / Performance, scalabilité et monitoring

- Optimisation des requêtes vectorielles
- Latence et throughput en production
- Monitoring des index et qualité des résultats

10 / Études de cas et implémentation réelle

- Déploiement d’un moteur RAG complet
- Architecture de recherche sémantique enterprise
- Bonnes pratiques de production

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

16 au 20 Nov. 2026	Distanciel
15 au 19 Fév. 2027	Présentiel - Casablanca
12 au 16 Avr. 2027	Présentiel - Casablanca
07 au 11 Juin 2027	Présentiel - Casablanca
02 au 06 Août 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés