+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated on 02/27/2026

Register

Milvus Training: Vector Database for AI Applications

2 days (14 hours)

Overview

Milvus is an open-source vector database designed to efficiently store, index, and search large-scale embeddings. It is a central building block of modern AI architectures, particularly for semantic search and RAG systems.

Our training will help you understand how a vector database works, configure indexing, and optimize similarity search in real-world contexts.

You will learn how to structure your data (vectors + metadata), choose the right indexes, integrate Milvus into a RAG pipeline, and explore other use cases such as recommendation, multimodal, and anomaly detection.

By the end of this training, you will be able to design a reliable, high-performance, and industrializable vector search architecture adapted to production constraints.

Like all our training courses, this one will introduce you to **the latest stable version** of the technology and its new features.

Objectives

- Understand how a vector database works.
- Configure indexing and optimize similarity searches with Milvus.
- Set up a hybrid search (vector + metadata).
- Integrate Milvus into a RAG pipeline and semantic search.
- Prepare for robust deployment in a production environment.

Target audience

- Data Engineers
- Backend developers
- AI/ML engineers
- Data architects
- AI-oriented DevOps

Prerequisites

- Basic Python and API usage
- Understanding of embeddings

Milvus training: Vector Database for AI Applications

[Day 1 - Morning]

Fundamentals of vector databases

- Understanding embeddings and their uses (semantic search, recommendation, RAG)
- Difference between relational databases and vector databases
- Principles of similarity search (ANN)
- Metrics: cosine, L2, inner product
- Milvus' position in the vector database ecosystem
- Hands-on workshop: Create a collection, insert vectors, and perform a top-k search.

[Day 1 - Afternoon]

Milvus internal architecture and data model

- Key components and services of Milvus
- Collections, partitions, scalar fields, and metadata
- Persistence, segments, and data lifecycle
- Index management: objectives, constraints, trade-offs
- Scalability and performance concepts

Indexing and query optimization

- Choosing an index based on use case
- Index and search parameters (recall/latency/cost)
- Hybrid search: vector + metadata filters
- Batching, pagination, and sizing strategies
- Benchmark methodology and quality indicators
- Hands-on workshop: Comparing two indexes and optimizing a query (latency vs. recall).

[Day 2 - Morning]

Milvus for RAG and semantic search

- Reminders: RAG pipeline (chunking, embeddings, retrieval, generation)
- Chunking strategies and context management
- Metadata storage: sources, permissions, scoring
- Integration with frameworks (e.g., LlamaIndex, LangChain)
- Evaluation: relevance, hallucinations, retrieval quality
- Hands-on workshop: Building a mini-RAG with Milvus.

[Day 2 - Afternoon]

Other use cases: recommendation, multimodal, anomalies

- Recommendation: item-to-item similarity, cold start, and business filters
- Multimodal: text + image (CLIP/vision embeddings)
- Anomaly detection: neighborhood, thresholds, and drift
- Data organization: metadata strategy and versioning
- Common anti-patterns and best practices

Industrialization and production deployment

- Deployment: standalone, cluster, containerized environments
- High availability, backup, and recovery
- Observability: metrics, logs, SLO, alerting
- Security: isolation, access control, data governance
- Evolution strategy: re-indexing, scaling, costs
- Hands-on workshop: Production checklist + target architecture plan.

Target companies

This training is intended for both individuals and companies, large or small, wishing to train their teams in a new advanced IT technology or to acquire specific business knowledge or modern methods.

Positioning at the start of training

The positioning at the start of the training complies with Qualiopi quality criteria. Upon final registration, the learner receives a self-assessment questionnaire that allows us to assess their estimated level of proficiency in different types of technologies, as well as their expectations and personal objectives for the upcoming training, within the limits imposed by the selected format. This questionnaire also allows us to anticipate certain connection or internal security issues within the company (intra-company or virtual classroom) that could be problematic for the monitoring and smooth running of the training session.

Teaching methods

Practical training: 60% practical, 40% theory. Training materials distributed in digital format to all participants.

Organization

The course alternates between theoretical input from the trainer, supported by examples and discussion sessions, and group work.

Assessment

At the end of the session, a multiple-choice questionnaire is used to verify that the skills have been correctly acquired.

Certification

A certificate will be issued to each trainee who has completed the entire training course.