Vector Databases: A Practical Guide

A Vector Database (Vector DB) is a specialized database designed to store, index, and search high-dimensional vector embeddings. These embeddings represent semantic meaning of data such as text, images, audio, video, and code.

Why Vector Databases Exist

Traditional databases are optimized for exact matches and structured queries. They struggle with similarity search over high-dimensional data. Vector databases solve this problem using Approximate Nearest Neighbor (ANN) algorithms.

How Vector Databases Work

1. Data is converted into vector embeddings using embedding models.
2. Vectors are stored and indexed efficiently.
3. Queries are embedded and compared using similarity metrics.
4. Most similar vectors are returned with relevance scores.

Types of Vector Databases

1. Vector Libraries: FAISS, Annoy, HNSWlib
2. Standalone Vector DBs: Pinecone, Weaviate, Milvus, Qdrant, Chroma
3. Hybrid DBs: PostgreSQL (pgvector), Elasticsearch, OpenSearch

Common Use Cases

1. Semantic Search and Enterprise Knowledge Bases
2. Retrieval-Augmented Generation (RAG) with LLMs
3. Recommendation Systems
4. Image, Audio, and Video Similarity Search
5. Fraud Detection and Anomaly Detection

Vector Databases in RAG Systems

In RAG systems, vector databases retrieve relevant context that is injected into LLM prompts. This allows LLMs to answer questions using private or domain-specific data without retraining.