VECTOR DATABASE

Vector databases are specialized storage systems designed to efficiently store, index, and query high-dimensional data represented as vectors. They are crucial parts of Large Language Models (LLMs) and are used to facilitate semantic search, enhance knowledge retrieval, act as long-term memory, and enhance Retrieval-Augmented Generation (RAG). Vector databases are particularly useful in applications like semantic search, recommendation systems, image and video retrieval, and natural language processing tasks.

Popular vector databases include Pinecone, Weaviate, Milvus, Qdrant, Chroma, Faiss, and pgvector. These databases convert text data into numerical vector representations, organize and index these embeddings for efficient searching, perform similarity search to find relevant embeddings, and retrieve them for responses or decision-making.

Vector databases work by converting text data into numerical vector representations, indexing them for efficient searching, performing similarity search to find relevant embeddings, and using retrieved embeddings for responses or decision-making. They are essential infrastructure for modern AI systems focused on understanding and retrieving meaning from vast amounts of unstructured data. Traditional databases are optimized for structured data and exact matching, while vector databases excel at similarity search using distance metrics like cosine similarity or Euclidean distance.