What is a vector database, and when do you need one?

1 November 2026 · 5 min read · blogs.iictp.org/what-is-a-vector-database

A vector database stores embeddings, which are lists of numbers that represent the meaning of text, images or other data, and finds the ones closest to a query. It is the retrieval layer behind semantic search and many retrieval-augmented generation (RAG) systems.

TL;DR The short version

- A vector database stores embeddings and finds the ones closest to a query, the retrieval layer behind semantic search and RAG.
- You do not always need a dedicated one, and chunking and metadata filters matter as much as the database.
- Retrieved text must respect who may see it; measure quality with real questions that have known good passages.

Read online, listen to it, or change the text size: https://blogs.iictp.org/what-is-a-vector-database

What is an embedding?

A model turns a piece of content into a vector so that similar meanings sit close together. A search for “refund policy” can then find a passage titled “returns and money back” without sharing any words.

How does the search work?

The query is turned into a vector, and the database returns the nearest stored vectors. For large collections it uses approximate nearest-neighbour indexes, which trade a little recall for much better speed.

Do you always need a dedicated vector database?

Not always. For small or moderate collections, a vector extension in a database you already run can be enough, and it keeps data and permissions in one place. A dedicated system makes sense when scale, latency or filtering needs outgrow that.

What else matters besides the database?

How documents are split into chunks changes what can be found, and metadata filters limit results by owner, date or access rights. Hybrid search combines keyword and vector matching, and content must be re-indexed when it changes.

How do permissions work?

Retrieved text must respect who may see it. Store access rights with each chunk and filter on them at query time, or a user could be shown passages from documents they are not allowed to open.

How do you judge quality?

Build a small set of real questions with known good passages, and measure whether retrieval finds them. See RAG explained for how retrieval fits into a full system.

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. pgvector. github.com/pgvector/pgvector (an example of a vector extension for an existing database).
2. Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Published by IICTP, an independent technical credential body, on 1 November 2026. The current version of this post is at https://blogs.iictp.org/what-is-a-vector-database. More guides: blogs.iictp.org.