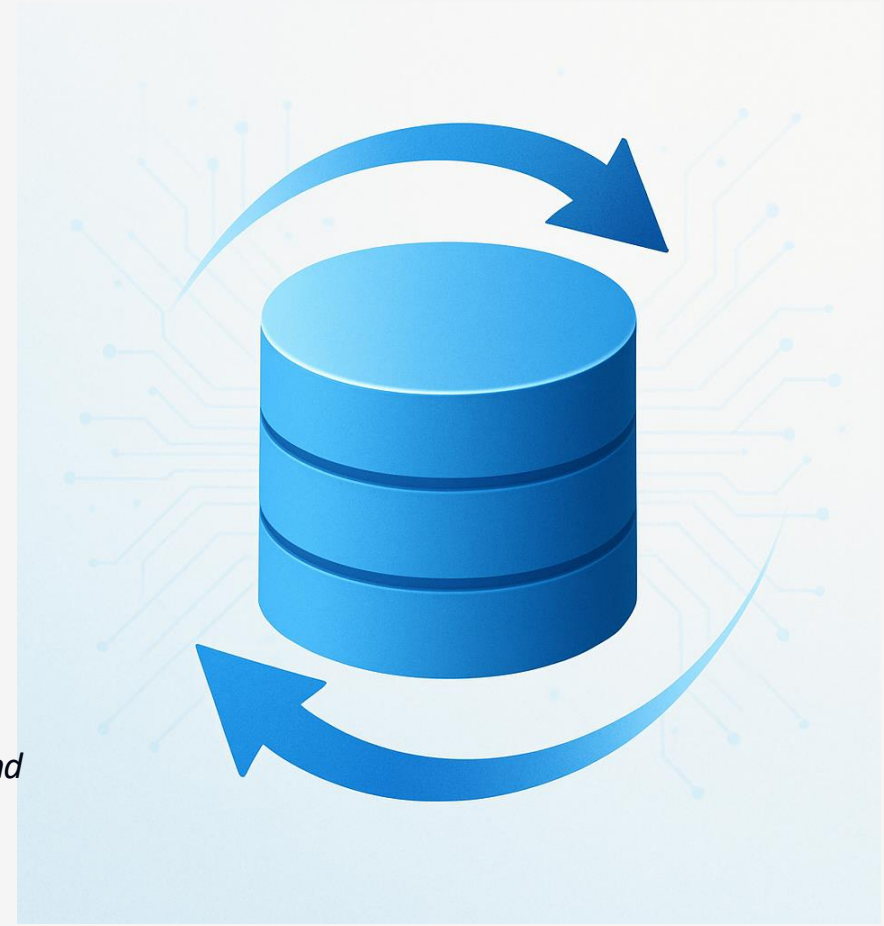
Vector Databases for Scalable Features

Leverage vector databases to power semantic search, recommendations and more.
Learn embeddings, similarity search, provider choices and design considerations.



July 30, 2025

Embeddings

Represent diverse data as numerical vectors to capture meaning and context.

Text

Words & documents compressed into vectors capturing semantic relationships.

Images

Visual content mapped to embeddings that encode objects and styles.

Audio/Video

Sound & motion represented as vectors reflecting patterns and tone.

- Embeddings compress complex data into dense vectors while preserving semantic meaning.
- Distances between vectors reflect the similarity between original objects.

[1] [2]

Similarity Search

Find the nearest neighbours of your query in high-dimensional space.

Euclidean

Straight-line distance; use when magnitude matters.

Cosine

Measures the angle between vectors; ideal for text where direction matters.

Dot Product

Assesses alignment or agreement; popular in recommendation systems.

- Encode query into a vector using the same embedding model.
- Use approximate nearest neighbour algorithms (e.g., HNSW) to efficiently locate similar vectors.
- Return the closest items ranked by distance metric.

[3] [4] [5]

Vector Database Providers

Choose the right platform based on features, flexibility and scale.

Pinecone

Managed vector DB service for ML workloads with horizontal scaling.

Weaviate

Open source vector DB combining semantic search with REST & GraphQL APIs.

Milvus

High-performance, open source DB supporting billions of vectors and modular indexes.

- Select a provider based on scalability requirements, API convenience, and deployment model (cloud vs self-hosted).

[6]

Aligning with Product Goals

Design vector solutions around user needs and business outcomes.

Identify User Problems	Define the customer goal: search, recommendation, anomaly detection or RAG.
Select Embeddings & Metrics	Choose embedding models (text, image, audio) and distance metrics suited to the data and task.
Choose DB Features	Decide on indexing options, metadata filtering and hybrid search capabilities.
Map to KPIs	Define business KPIs (e.g., click-through rate, relevance score) and evaluate results.

[7] [8]

Scalability & Cost

Balance performance, capacity and expenses when deploying vector databases.

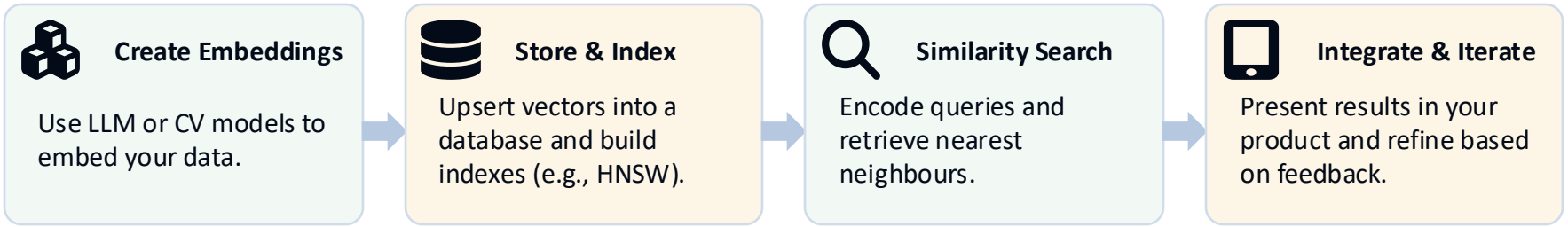
Factor	Considerations
Setup & Maintenance	Specialised skills increase initial costs; efficient vector operations lead to long-term savings.
Operational Efficiency	Vector stores accelerate AI-driven tasks, improving user experience and functionality.
Scaling	Horizontal scaling with distributed architecture is often more cost-effective than vertical scaling.

- Distributed vector stores support horizontal scaling, high availability and redundancy.
- Evaluate costs based on embedding dimensionality, number of vectors and query volume.

[9] [10]

Let's Build: Semantic Search

Turn theory into practice by building a simple semantic search feature.



- Follow four steps: embed data, insert and index, run similarity queries, and integrate results into your product.
- Refine embeddings, metrics and user interface based on real-world feedback.

[11] [12]