Wikipedia
The Free Encyclopedia

Vector database

A **vector database management system** (**VDBMS**) or simply **vector database** or **vector store** is a database that can store vectors (fixed-length lists of numbers) along with other data items. Vector databases typically implement one or more Approximate Nearest Neighbor (ANN) algorithms,[1][2] so that one can search the database with a query vector to retrieve the closest matching database records.

Vectors are mathematical representations of data in a high-dimensional space. In this space, each dimension corresponds to a feature of the data, and tens of thousands of dimensions might be used to represent sophisticated data. A vector's position in this space represents its characteristics. Words, phrases, or entire documents, and images, audio, and other types of data can all be vectorized.[3]

These feature vectors may be computed from the raw data using machine learning methods such as feature extraction algorithms, word embeddings[4] or deep learning networks. The goal is that semantically similar data items receive feature vectors that are close to each other.

Vector databases can be used for similarity search, multi-modal search, recommendations engines, large languages models (LLMs), etc.[5]

Vector databases are also used to implement Retrieval-Augmented Generation (RAG), a method to improve domain-specific responses of large language models. Text documents describing the domain of interest are collected and for each document a feature vector (known as an "embedding") is computed, typically using a deep learning network, and stored in a vector database. Given a user prompt, the feature vector of the prompt is computed and the database is queried to retrieve the most relevant documents. These are then automatically added into the context window of the large language model and the large language model proceeds to create a response to the prompt given this context.[6]

List of vector databases

Name	License
Apache Cassandra[7][8]	Apache License 2.0
Azure Cosmos DB Vector Database Extension[9]	N/A (Managed Service)
LlamaIndex [10]	MIT License[11]
Milvus [12][13]	Apache License 2.0
MongoDB Atlas [14]	N/A (Managed service)
Couchbase[15][16]	Unknown (Preview)
Pinecone [17]	Closed source
Postgres with pgvector [18]	PostgreSQL License[19]
Qdrant [20]	Apache License 2.0[21]
Weaviate [22]	BSD 3-Clause[23]
Chroma[24][25]	Apache License 2.0[26]
Elasticsearch[27]	Server Side Public License, Elastic License [28]
Vespa [29]	Apache License 2.0[30]
SurrealDB[31]	Business Source License & Apache License (After 4 years)[32]

References

1. Roie Schwaber-Cohen. "What is a Vector Database & How Does it Work" (https://www.pinecone.io/learn/vector-database/). Pinecone. Retrieved 18 November 2023.
2. "What is a vector database" (https://www.elastic.co/what-is/vector-database). Elastic. Retrieved 18 November 2023.
3. "Vector database - Azure Cosmos DB" (https://learn.microsoft.com/en-us/azure/cosmos-db/vector-database). *learn.microsoft.com*. 2023-12-26. Retrieved 2024-01-11.
4. Evan Chaki (2023-07-31). "What is a vector database?" (https://learn.microsoft.com/en-us/semantic-kernel/memories/vector-db). Microsoft. "A vector database is a type of database that stores data as high-dimensional vectors, which are mathematical representations of features or attributes."
5. "Vector database - Azure Cosmos DB" (https://learn.microsoft.com/en-us/azure/cosmos-db/vector-database). *learn.microsoft.com*. 2023-12-26. Retrieved 2024-01-11.
6. Lewis, Patrick; Perez, Ethan; Piktus, Aleksandra; Petroni, Fabio; Karpukhin, Vladimir; Goyal, Naman; Küttler, Heinrich (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks". *Advances in Neural Information Processing Systems 33*: 9459–9474. arXiv:2005.11401 (https://arxiv.org/abs/2005.11401).

7. "5 Hard Problems in Vector Search, and How Cassandra Solves Them" (https://thenewstack.io/5-hard-problems-in-vector-search-and-how-cassandra-solves-them/). *TheNewStack*. 2023-09-22. Retrieved 2023-09-22.
8. "Vector Search quickstart" (https://cassandra.apache.org/doc/latest/cassandra/vector-search/overview.html). Retrieved 2023-11-21.
9. "Vector database - Azure Cosmos DB" (https://learn.microsoft.com/azure/cosmos-db/vector-database). *learn.microsoft.com*. Retrieved 2024-01-10.
10. Wiggers, Kyle (2023-06-06). "LlamaIndex adds private data to large language models" (https://techcrunch.com/2023/06/06/llamaindex-adds-private-data-to-large-language-models/). *TechCrunch*. Retrieved 2023-10-29.
11. "llama_index/LICENSE at main · run-llama/llama_index" (https://github.com/run-llama/llama_index/blob/main/LICENSE). *GitHub*. Retrieved 2023-10-29.
12. "Open Source Vector Database – Milvus – LFAI & DATA" (https://milvus.io/). Retrieved 29 October 2023.
13. Liao, Ingrid Lunden and Rita (2022-08-24). "Zilliz raises $60M, relocates to SF" (https://techcrunch.com/2022/08/24/zilliz-the-startup-behind-the-milvus-open-source-vector-database-for-ai-applications-raises-60m-and-relocates-to-sf/). *TechCrunch*. Retrieved 2023-10-29.
14. "Introducing Atlas Vector Search: Build Intelligent Applications with Semantic Search and AI Over Any Type of Data" (https://www.mongodb.com/blog/post/introducing-atlas-vector-search-build-intelligent-applications-semantic-search-ai). *MongoDB*. 2023-06-22.
15. "Couchbase aims to boost developer database productivity with Capella IQ AI tool" (https://venturebeat.com/ai/couchbase-aims-to-boost-developer-database-productivity-with-capella-iq-ai-tool/#h-next-on-the-roadmap-for-couchbase-is-vector-support). *VentureBeat*. 2023-08-30.
16. "Investor Presentation Third Quarter Fiscal 2024" (https://investors.couchbase.com/static-files/551e5b96-5307-4119-b225-19cfd8540242). *Couchbase Investor Relations*. 2023-12-06.
17. "Pinecone leads 'explosion' in vector databases for generative AI" (https://venturebeat.com/ai/pinecone-leads-explosion-in-vector-databases-for-generative-ai/). *VentureBeat*. 2023-07-14. Retrieved 2023-10-29.
18. "pgvector" (https://github.com/pgvector/pgvector). *GitHub*. Retrieved 2023-11-27.
19. "pgvector/License" (https://github.com/pgvector/pgvector/blob/master/LICENSE). *GitHub*. Retrieved 2023-11-27.
20. Sawers, Paul (2023-04-19). "Qdrant, an open source vector database startup, wants to help AI developers leverage unstructured data" (https://techcrunch.com/2023/04/19/qdrant-an-open-source-vector-database-startup-wants-to-help-ai-developers-leverage-unstructured-data/). *TechCrunch*. Retrieved 2023-10-29.
21. "qdrant/LICENSE at master · qdrant/qdrant" (https://github.com/qdrant/qdrant/blob/master/LICENSE). *GitHub*. Retrieved 2023-10-29.
22. "Weaviate reels in $50M for its AI-optimized vector database" (https://siliconangle.com/2023/04/21/weaviate-reels-50m-ai-optimized-vector-database/). *SiliconANGLE*. 2023-04-21. Retrieved 2023-10-29.
23. "weaviate/LICENSE at master · weaviate/weaviate" (https://github.com/weaviate/weaviate/blob/master/LICENSE). *GitHub*. Retrieved 2023-10-29.
24. Palazzolo, Stephanie. "Vector database Chroma scored $18 million in seed funding at a $75 million valuation. Here's why its technology is key to helping generative AI startups" (https://www.businessinsider.com/vector-database-startup-chroma-raises-seed-funding-generative-artificial-intelligence-2023-4). *Business Insider*. Retrieved 2023-11-16.
25. MSV, Janakiram (2023-07-28). "Exploring Chroma: The Open Source Vector Database for LLMs" (https://thenewstack.io/exploring-chroma-the-open-source-vector-database-for-llms/). *The New Stack*. Retrieved 2023-11-16.

26. https://github.com/chroma-core/chroma/blob/main/LICENSE
27. Kerner, Sean (23 May 2023). "Elasticsearch Relevance Engine brings new vectors to generative AI" (https://venturebeat.com/ai/elasticsearch-relevance-engine-brings-new-vectors-to-generative-ai/). *VentureBeat*. Retrieved 18 November 2023.
28. https://github.com/elastic/elasticsearch/blob/main/LICENSE.txt
29. Riley, Duncan (4 October 2023). "Yahoo spins off AI scaling engine Vespa as an independent company" (https://siliconangle.com/2023/10/04/yahoo-spins-off-ai-scaling-engine-vespa-independent-company/). *siliconANGLE*. Retrieved 18 November 2023.
30. https://github.com/vespa-engine/vespa/blob/master/LICENSE
31. Wiggers, Kyle (2023-01-04). "SurrealDB raises $6M for its database-as-a-service offering" (https://techcrunch.com/2023/01/04/surrealdb-raises-6m-startup-funding-database-as-a-service/). *TechCrunch*. Retrieved 2024-01-19.
32. "SurrealDB | License FAQs | The ultimate multi-model database" (https://surrealdb.com/license). *SurrealDB*. Retrieved 2024-01-19.

▪