

Vector database

A **vector database**, **vector store** or **vector search engine** is a <u>database</u> that uses the <u>vector space model</u> to store vectors (fixed-length lists of numbers) along with other data items. Vector databases typically implement one or more <u>approximate nearest neighbor</u> 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 <u>feature</u> of the data, with the number of dimensions ranging from a few hundred to tens of thousands, depending on the complexity of the data being represented. A vector's position in this space represents its characteristics. Words, phrases, or entire documents, as well as 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 $\underline{\text{feature extraction}}$ algorithms, $\underline{\text{word embeddings}}^{[4]}$ or $\underline{\text{deep learning}}$ networks. The goal is that semantically similar data items receive feature vectors close to each other.

Vector databases can be used for <u>similarity search</u>, <u>semantic search</u>, <u>multi-modal search</u>, recommendations engines, large language models (LLMs), object detection, etc. [5]

Vector databases are also often used to implement <u>retrieval-augmented generation</u> (RAG), a method to improve domain-specific responses of large language models. The retrieval component of a RAG can be any search system, but is most often implemented as a vector database. Text documents describing the domain of interest are collected, and for each document or document section, a feature vector (known as an "<u>embedding</u>") 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]

Techniques

The most important techniques for similarity search on high-dimensional vectors include:

- Hierarchical Navigable Small World (HNSW) graphs
- Locality-sensitive Hashing (LSH) and Sketching
- Product Quantization (PQ)
- Inverted Files

and combinations of these techniques.

In recent benchmarks, HNSW-based implementations have been among the best performers. [7][8] Conferences such as the International Conference on Similarity Search and Applications, SISAP and the Conference on Neural Information Processing Systems (NeurIPS) host competitions on vector search in large databases.

Implementations

Name	License
Aerospike ^{[9][10]}	Proprietary
AllegroGraph ^{[11][12]}	Proprietary (Managed Service)
Apache Cassandra ^{[13][14]}	Apache License 2.0
Azure Cosmos DB ^[15]	Proprietary (Managed Service)
Chroma ^{[16][17]}	Apache License 2.0 ^[18]
ClickHouse ^[19]	Apache License 2.0
Couchbase ^{[20][21]}	BSL 1.1 ^[22]
CrateDB ^[23]	Apache License 2.0
DataStax ^[24]	Proprietary (Managed Service)
Elasticsearch ^[25]	Server Side Public License, Elastic License ^[26]
HAKES ^[27]	Apache License 2.0 ^[28]
HDF5 Query Indexing ^[29]	BSD 3-Clause ^[30]
JaguarDB ^{[31][32]}	Proprietary
LanceDB ^[33]	Apache License 2.0 ^[34]
Lantern ^[35]	BSL 1.1 ^[36]
LlamaIndex ^[37]	MIT License ^[38]
MariaDB ^{[39][40]}	<u>GPL v2^[41]</u>
Marqo ^[42]	Apache License 2.0 ^[43]
Meilisearch ^[44]	MIT License ^[45]
Milvus ^{[46][47]}	Apache License 2.0
MongoDB Atlas ^[48]	Server Side Public License (Managed service)
Neo4j ^{[49][50]}	GPL v3 (Community Edition) ^[51]
ObjectBox ^[52]	Apache License 2.0 ^[53]
OpenSearch ^{[54][55][56]}	Apache License 2.0 ^[57]
Oracle Database ^[58]	Proprietary (Managed Service or License)
Pinecone ^[59]	Proprietary (Managed Service)
Pixeltable (Incremental Embedding) ^[60]	Apache License 2.0 ^[61]
Postgres with pgvector ^[62]	PostgreSQL License ^[63]
Qdrant ^[64]	Apache License 2.0 ^[65]

Redis Stack ^{[66][67]}	Redis Source Available License (https://redis.io/docs/about/licens $e/)^{[68]}$
Snowflake ^[69]	Proprietary (Managed Service)
SurrealDB ^[70]	BSL 1.1 ^[71]
Typesense ^[72]	GPL v3 (Community Edition) ^[73]
Vespa ^[74]	Apache License 2.0 ^[75]
Weaviate ^[76]	BSD 3-Clause ^[77]

See also

- Curse of dimensionality Difficulties arising when analyzing data with many aspects ("dimensions")
- Machine learning Study of algorithms that improve automatically through experience
- Nearest neighbor search Optimization problem in computer science
- Recommender system System to predict users' preferences

References

- 1. Roie Schwaber-Cohen. "What is a Vector Database & How Does it Work" (https://www.pinec one.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" (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/s emantic-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" (https://learn.microsoft.com/en-us/azure/cosmos-db/vector-database). *learn.microsoft.com*. 2023-12-26. Retrieved 2024-01-11.
- 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. Aumüller, Martin; Bernhardsson, Erik; Faithfull, Alexander (2017), Beecks, Christian; Borutta, Felix; Kröger, Peer; Seidl, Thomas (eds.), "ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms" (http://link.springer.com/10.1007/978-3-319-68474-1_3), Similarity Search and Applications, vol. 10609, Cham: Springer International Publishing, pp. 34–49, arXiv:1807.05614 (https://arxiv.org/abs/1807.05614), doi:10.1007/978-3-319-68474-1_3 (https://doi.org/10.1007%2F978-3-319-68474-1_3), ISBN 978-3-319-68473-4, retrieved 2024-03-19

- 8. Aumüller, Martin; Bernhardsson, Erik; Faithfull, Alexander (2017). "ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms" (https://link.springer.com/chapter/10.1007/978-3-319-68474-1_3). In Beecks, Christian; Borutta, Felix; Kröger, Peer; Seidl, Thomas (eds.). Similarity Search and Applications. Lecture Notes in Computer Science. Vol. 10609. Cham: Springer International Publishing. pp. 34–49. arXiv:1807.05614 (https://arxiv.org/abs/1807.05614). doi:10.1007/978-3-319-68474-1_3 (https://doi.org/10.1007/978-3-319-68474-1_3). ISBN 978-3-319-68474-1.
- 9. "Aerospike Recognized by Independent Research Firm Among Notable Vendors in Vector Databases Report" (https://www.morningstar.com/news/globe-newswire/9111790/aerospike-recognized-by-independent-research-firm-among-notable-vendors-in-vector-databases-report). *Morningstar*. 2024-05-07. Retrieved 2024-08-01.
- 10. "Aerospike raises \$109M for its real-time database platform to capitalize on the AI boom" (ht tps://techcrunch.com/2024/04/04/aerospike-raises-100m-for-its-real-time-database-platform-to-capitalize-on-the-ai-boom/). *TechCrunch*. 2024-04-04. Retrieved 2024-08-01.
- 11. "AllegroGraph 8.0 Incorporates Neuro-Symbolic AI, a Pathway to AGI" (https://thenewstack.i o/allegrograph-8-0-incorporates-neuro-symbolic-ai-a-pathway-to-agi/). *TheNewStack*. 2023-12-29. Retrieved 2024-06-06.
- 12. "Franz Inc. Introduces AllegroGraph Cloud: A Managed Service for Neuro-Symbolic Al Knowledge Graphs" (https://www.datanami.com/this-just-in/franz-inc-introduces-allegrograp h-cloud-a-managed-service-for-neuro-symbolic-ai-knowledge-graphs/). *Datanami*. 2024-01-18. Retrieved 2024-06-06.
- 13. "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.
- 14. "Vector Search quickstart" (https://cassandra.apache.org/doc/latest/cassandra/vector-search/overview.html). Retrieved 2023-11-21.
- 15. "Vector database" (https://learn.microsoft.com/azure/cosmos-db/vector-database). *learn.microsoft.com*, 26 December 2023. Retrieved 2024-01-10.
- 16. 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.
- 17. 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.
- 18. "chroma/LICENSE at main · chroma-core/chroma" (https://github.com/chroma-core/chroma/blob/main/LICENSE). *GitHub*.
- 19. "Can you use ClickHouse for vector search? | ClickHouse Docs" (http://web.archive.org/web/20250622222726/https://clickhouse.com/docs/knowledgebase/vector-search). 2023-10-26. Archived from the original (https://clickhouse.com/docs/knowledgebase/vector-search) on 2025-06-22. Retrieved 2025-07-02.
- 20. "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.
- 21. "Investor Presentation Third Quarter Fiscal 2024" (https://investors.couchbase.com/static-files/551e5b96-5307-4119-b225-19cfd8540242). *Couchbase Investor Relations*. 2023-12-06.
- 22. Anderson, Scott (2021-03-26). "Couchbase Adopts BSL License" (https://www.couchbase.c om/blog/couchbase-adopts-bsl-license/). *The Couchbase Blog*. Retrieved 2024-02-14.
- 23. "Open Source Vector Database" (https://cratedb.com/blog/open-source-vector-database). *CrateDB Blog.* 16 November 2023. Retrieved 2024-11-06.

- 24. Sean Michael Kerner (18 July 2023). "DataStax brings vector database search to multicloud with Astra DB" (https://venturebeat.com/data-infrastructure/datastax-brings-vector-database-search-to-multicloud-with-astra-db/). Venture Beat.
- 25. Kerner, Sean (23 May 2023). "Elasticsearch Relevance Engine brings new vectors to generative AI" (https://venturebeat.com/ai/elasticsearch-relevance-engine-brings-new-vector s-to-generative-ai/). *VentureBeat*. Retrieved 18 November 2023.
- 26. "elasticsearch/LICENSE.txt at main · elastic/elasticsearch" (https://github.com/elastic/elasticsearch/blob/main/LICENSE.txt). *GitHub*.
- 27. "HAKES | Efficient Data Search with Embedding Vectors at Scale" (https://www.comp.nus.e du.sg/~dbsystem/hakes). Retrieved 8 March 2025.
- 28. "HAKES/LICENSE at main · nusdbsystem/HAKES" (https://github.com/nusdbsystem/HAKE S/blob/main/LICENSE). *GitHub*. Retrieved 8 March 2025.
- 29. "HDF5 Query Indexing" (https://github.com/HDFGroup/hdf5doc/tree/master/RFCs/HDF5/Query-Indexing). *GitHub*. 27 Sep 2019. Retrieved 3 May 2024.
- 30. "HDFGroup/COPYING at master · HDFGroup/hdf5" (https://github.com/HDFGroup/hdf5/blob/master/COPYING). *GitHub*. Retrieved 2023-10-29.
- 31. "JaguarDB Homepage" (http://jaguardb.com/). JaguarDB. Retrieved 2025-04-12.
- 32. "Vector DBMS" (https://db-engines.com/de/ranking/vector+dbms). db-engines.com. 2023-07-03. Retrieved 2025-04-12.
- 33. "LanceDB Homepage" (https://lancedb.com/). LanceDB. 2024-12-17. Retrieved 2024-12-17.
- 34. "lancedb/LICENSE at main · lancedb/lancedb" (https://github.com/lancedb/lancedb?tab=Apa che-2.0-1-ov-file). *GitHub*. Retrieved 2024-12-17.
- 35. "Lantern" (https://lantern.dev/). 2024-04-05. Retrieved 2024-04-05.
- 36. "lantern/LICENSE at main /lanterndata/lantern" (https://github.com/lanterndata/lantern/blob/main/LICENSE). *GitHub*. Retrieved 2024-04-10.
- 37. Wiggers, Kyle (2023-06-06). "LlamaIndex adds private data to large language models" (http s://techcrunch.com/2023/06/06/llamaindex-adds-private-data-to-large-language-models/). *TechCrunch*. Retrieved 2023-10-29.
- 38. "llama_index/LICENSE at main · run-llama/llama_index" (https://github.com/run-llama/llama_index/blob/main/LICENSE). *GitHub*. Retrieved 2023-10-29.
- 39. "MariaDB Vector" (https://mariadb.org/projects/mariadb-vector/). *MariaDB.org*. Retrieved 2024-07-30.
- 40. "Vector search in old and modern databases" (https://manticoresearch.com/blog/vector-sear ch-in-databases/). *manticoresearch.com*. Retrieved 2024-07-30.
- 41. "Licensing FAQ" (https://mariadb.com/kb/en/licensing-faq/). *MariaDB KnowledgeBase*. Retrieved 2024-07-30.
- 42. Sawers, Paul (2023-08-16). "Meet Marqo, an open source vector search engine for Al applications" (https://techcrunch.com/2023/08/16/meet-marqo-an-open-source-vector-searc h-engine-for-ai-applications/). *TechCrunch*. Retrieved 2024-08-20.
- 43. *marqo-ai/marqo* (https://github.com/marqo-ai/marqo?tab=Apache-2.0-1-ov-file#readme), Marqo, 2024-08-20, retrieved 2024-08-20
- 44. "Meilisearch Homepage" (https://meilisearch.com/). *Meilisearch*. 2024-10-08. Retrieved 2023-10-29.
- 45. "meilisearch/LICENSE at main · meilisearch/meilisearch" (https://github.com/meilisearch/meilisearch/blob/main/LICENSE). *GitHub*. Retrieved 2024-10-08.
- 46. "Open Source Vector Database Milvus LFAI & DATA" (https://milvus.io/). Retrieved 29 October 2023.
- 47. Liao, Ingrid Lunden and Rita (2022-08-24). "Zilliz raises \$60M, relocates to SF" (https://tech_crunch.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.

- 48. "Introducing Atlas Vector Search: Build Intelligent Applications with Semantic Search and Al 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.
- 49. "Neo4j enhances its graph database with vector search" (https://itbrief.com.au/story/neo4j-enhances-its-graph-database-with-vector-search). *itbrief*. 2023-08-22.
- 50. "Vector search indexes" (https://neo4j.com/docs/cypher-manual/current/indexes/semantic-indexes/vector-indexes). *neo4j*.
- 51. "Neo4j licensing" (https://neo4j.com/licensing/).
- 52. "Top Fifteen Vector Databases" (https://db-engines.com/de/ranking/vektor+dbms). db-engines.com. 2024-07-03. Retrieved 2024-07-03.
- 53. "ObjectBox Java license" (https://github.com/objectbox/objectbox-java/blob/main/LICENSE.t xt). *github*.
- 54. "Using OpenSearch as a Vector Database" (https://opensearch.org/platform/search/vector-d atabase.html). *OpenSearch.org*. 2023-08-02. Retrieved 2024-02-07.
- 55. Pan, James Jie; Wang, Jianguo; Li, Guoliang (2023-10-21), *Survey of Vector Database Management Systems*, arXiv:2310.14021 (https://arxiv.org/abs/2310.14021)
- 56. "AWS debuts new AI-powered data management and analysis tools" (https://siliconangle.com/2023/07/26/aws-debuts-new-ai-powered-data-management-analysis-tools/). SiliconANGLE. 2023-07-26. Retrieved 2024-02-07.
- 57. "OpenSearch license" (https://github.com/opensearch-project/OpenSearch/blob/main/LICEN SE.txt). *github*.
- 58. Hook(1) and Priyadarshi(2), Doug(1) and Ranjan(2) (May 2, 2024). "Oracle Announces General Availability of AI Vector Search in Oracle Database 23ai" (https://blogs.oracle.com/database/post/oracle-announces-general-availability-of-ai-vector-search-in-oracle-database-23ai). oracle. Retrieved July 9, 2024.
- 59. "Pinecone leads 'explosion' in vector databases for generative Al" (https://venturebeat.com/a i/pinecone-leads-explosion-in-vector-databases-for-generative-ai/). *VentureBeat*. 2023-07-14. Retrieved 2023-10-29.
- 60. "Automatic incremental embedding index" (https://www.pixeltable.com/blog/pixeltable-incremental-embedding-indexes). *Pixeltable*. Retrieved 2025-07-04.
- 61. "OpenSearch license" (https://github.com/opensearch-project/OpenSearch/blob/main/LICEN SE.txt). *github*.
- 62. "payector" (https://github.com/payector/payector). GitHub. Retrieved 2023-11-27.
- 63. "pgvector/License" (https://github.com/pgvector/pgvector/blob/master/LICENSE). *GitHub*. Retrieved 2023-11-27.
- 64. Sawers, Paul (2023-04-19). "Qdrant, an open-source vector database startup, wants to help Al 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.
- 65. "qdrant/LICENSE at master · qdrant/qdrant" (https://github.com/qdrant/qdrant/blob/master/LICENSE). *GitHub*. Retrieved 2023-10-29.
- 66. "Using Redis as a Vector Database with OpenAl | OpenAl Cookbook" (https://cookbook.openai.com/examples/vector_databases/redis/getting-started-with-redis-and-openai). cookbook.openai.com. Retrieved 2024-02-10.
- 67. "Redis as a vector database quick start guide" (https://redis.io/docs/get-started/vector-database/). *Redis*. Retrieved 2024-01-31.
- 68. "Search and query" (https://redis.io/docs/interact/search-and-query/). *Redis.* Retrieved 2024-02-10.

- 69. "Vector data type and vector similarity functions General Availability" (https://docs.snowflake.com/en/release-notes/2024/other/2024-05-16-vector-data-type-ga). Snowflake. 2024-05-17. Retrieved 2024-05-17.
- 70. 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.
- 71. "SurrealDB | License FAQs | The ultimate multi-model database" (https://surrealdb.com/license). SurrealDB. Retrieved 2024-02-14.
- 72. Martinez, Miguel (2024-06-20). "Typesense Homepage" (https://typesense.org/). *Typesense*. Retrieved 2024-06-20.
- 73. "Typesense licensing" (https://github.com/typesense/typesense/blob/main/LICENSE.txt). *GitHub*.
- 74. Riley, Duncan (4 October 2023). <u>"Yahoo spins off AI scaling engine Vespa as an independent company"</u> (https://siliconangle.com/2023/10/04/yahoo-spins-off-ai-scaling-engine-vespa-independent-company/). *siliconANGLE*. Retrieved 18 November 2023.
- 75. "vespa/LICENSE at master · vespa-engine/vespa" (https://github.com/vespa-engine/vespa/b lob/master/LICENSE). *GitHub*.
- 76. "Weaviate reels in \$50M for its Al-optimized vector database" (https://siliconangle.com/2023/04/21/weaviate-reels-50m-ai-optimized-vector-database/). SiliconANGLE. 2023-04-21. Retrieved 2023-10-29.
- 77. "weaviate/LICENSE at master · weaviate/weaviate" (https://github.com/weaviate/weaviate/bl ob/master/LICENSE). *GitHub*. Retrieved 2023-10-29.

External links

■ Sawers, Paul (2024-04-20). "Why vector databases are having a moment as the AI hype cycle peaks" (https://techcrunch.com/2024/04/20/why-vector-databases-are-having-a-mome nt-as-the-ai-hype-cycle-peaks/). *TechCrunch*. Retrieved 2024-04-23.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Vector_database&oldid=1298831203"