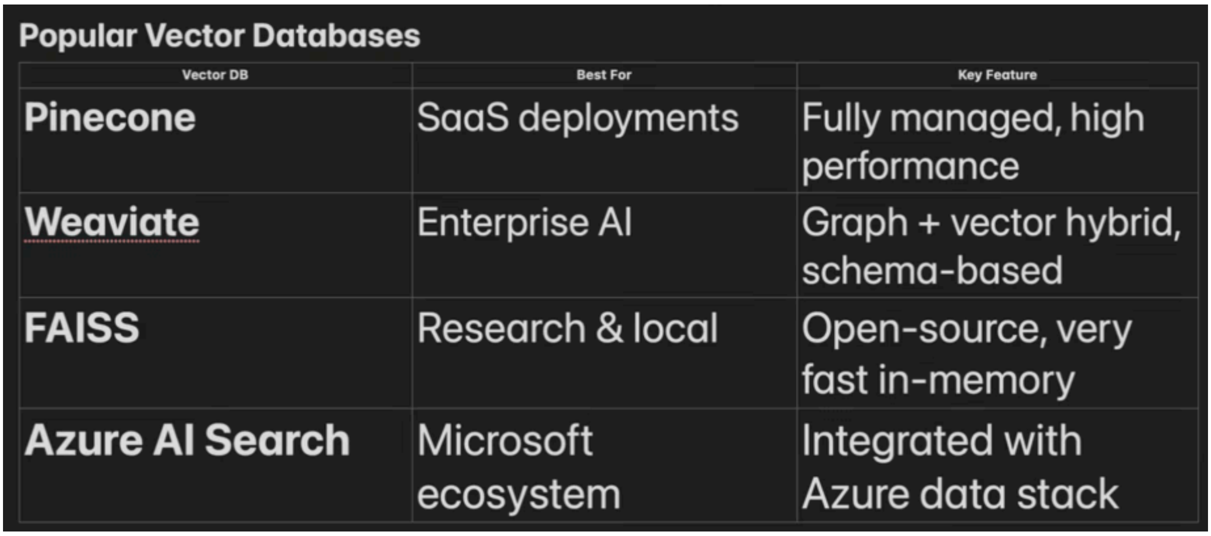
Vector Databases



Popular Vector Databases

Vector DB	Best For	Key Feature
Pinecone	SaaS deployments	Fully managed, high performance
Weaviate	Enterprise AI	Graph + vector hybrid, schema-based
FAISS	Research & local	Open-source, very fast in-memory
Azure AI Search	Microsoft ecosystem	Integrated with Azure data stack

Detailed Study on all the vector databases

1. Pinecone

Pinecone is a **fully managed, cloud-native vector database** designed for high-performance, low-latency semantic search at scale. It's built with ease of use in mind, offering developers an API-first experience with minimal infrastructure overhead.

The key strengths of Pinecone are:

- **Zero DevOps**: You don't manage any infrastructure. Just push vectors, metadata, and query.
- **High scalability**: Built for production-grade workloads, supporting billions of vectors.
- **Low latency**: Optimized for fast vector retrieval with built-in vector indexing (based on HNSW and proprietary tech).
- **Metadata filtering**: You can attach metadata to vectors and use simple filters during search.

However, it has some **limitations**:

- It's **not open source** – fully proprietary and cloud-only.
- Some users report **limited filtering flexibility**, especially when complex conditions are involved.
- Costs can rise significantly at scale depending on query load and storage.

Best For: Teams needing a reliable, scalable, and low-maintenance solution for production RAG or semantic search apps, especially when time-to-market matters more than full control.

2. Weaviate

Weaviate is a **powerful open-source vector database** that also offers a managed cloud option. It supports a broad range of features out of the box, including:

- **Hybrid search** (vector + keyword/structured filtering)
- **Dynamic schema** (supports adding data types and classes at runtime)
- **Built-in classification and vectorization** with multiple backends (OpenAI, Hugging Face, Cohere)
- **Strong metadata filtering** and structured query capabilities
- **Self-hosted or managed** – gives you deployment flexibility

The **trade-offs** with Weaviate include:

- **Operational overhead** if self-hosted – you must manage scaling, backups, performance tuning
- Slightly **higher latency** than Pinecone in some benchmarks (though this varies by workload)
- Requires more **initial setup and understanding** compared to managed services

Best For: Teams that want full control over their vector database, need hybrid search, or operate in environments where open-source or on-prem hosting is required (e.g., healthcare, defense, regulated industries).

3. FAISS

FAISS (Facebook AI Similarity Search) is a **low-level, high-performance library** for efficient similarity search and clustering of dense vectors. It's widely used in research, experimentation, and custom ML pipelines.

FAISS excels at:

- **Raw speed and flexibility** — you control everything (index type, memory usage, quantization, GPU/CPU).

- **Offline or embedded use** — runs locally or in containers, perfect for edge or custom AI stacks.
- **Highly optimized** for ANN (Approximate Nearest Neighbor) search and can run on GPU for high throughput.

But FAISS is **not** a database:

- It lacks built-in **persistence**, **metadata support**, and **query interfaces**.
- You must build all the supporting infrastructure yourself (for CRUD, updates, backups, security, etc.).
- No native support for **real-time updates** — inserting new data may require full reindexing.

Best For: Research teams, ML engineers, and advanced users who need raw performance, want to build a custom system, or are deploying vector search in embedded/offline applications.

4. Azure AI Search

Azure AI Search (formerly Azure Cognitive Search with vector capabilities) is a **fully managed search service** by Microsoft, now enhanced with native **vector search** support. It is particularly well-suited for enterprises already invested in the Azure ecosystem.

Key benefits include:

- **Hybrid search**: Supports combining vector similarity with full-text keyword search.
- **Tight Azure integration**: Works well with Azure OpenAI, Blob Storage, Cognitive Services, etc.
- **Security and compliance**: Enterprise-ready with RBAC, encryption, multitenancy, etc.
- **Indexers and pipelines**: Easily ingest and transform data from various Azure data sources.

Limitations:

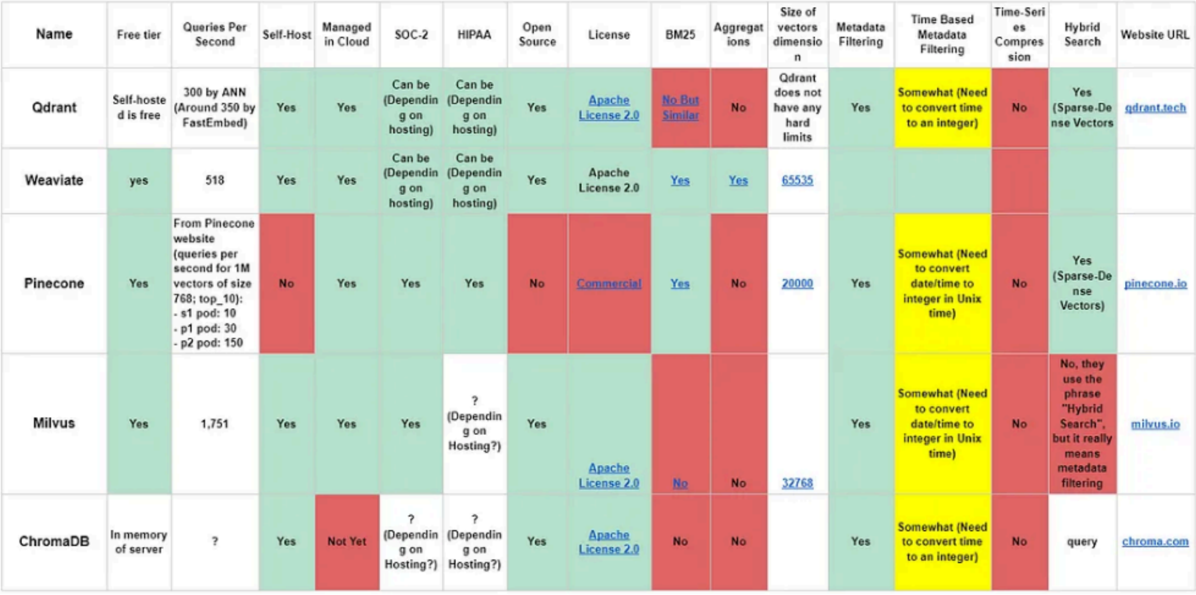
- **Cloud-only** — no self-hosted option.
- **Less flexibility** than open-source options like Weaviate for custom data models.
- **Potential cost overhead** for small/experimental projects if not already on Azure.

Best For: Enterprises building AI-powered search or RAG solutions **within the Microsoft Azure ecosystem**, where seamless integration and enterprise security are essential.

Final Summary

If you're deciding among these:

- Choose **Pinecone** if you want a **turnkey solution** with **minimal setup** and production-grade performance.
- Choose **Weaviate** if you need **flexibility**, **hybrid search**, and have the resources to self-manage or prefer open-source.
- Choose **FAISS** if you're building a **custom**, high-performance pipeline and want full control over vector handling.
- Choose **Azure AI Search** if you are already on Azure and want **enterprise-grade** vector and hybrid search features **in one integrated service**.



Name	Free tier	Queries Per Second	Self-Host	Managed in Cloud	SOC-2	HIPAA	Open Source	License	BM25	Aggregations	Size of vectors dimension	Metadata Filtering	Time Based Metadata Filtering	Time-Series Compression	Hybrid Search	Website URL
Qdrant	Self-hosted is free	300 by ANN (Around 350 by FastEmbed)	Yes	Yes	Can be (Depending on hosting)	Can be (Depending on hosting)	Yes	Apache License 2.0	No But Similar	No	Qdrant does not have any hard limits	Yes	Somewhat (Need to convert time to an integer)	No	Yes (Sparse-Dense Vectors	qdrant.tech
Weaviate	yes	518	Yes	Yes	Can be (Depending on hosting)	Can be (Depending on hosting)	Yes	Apache License 2.0	Yes	Yes	65535	🟩	🟩	🟥	🟩	
Pinecone	Yes	From Pinecone website (queries per second for 1M vectors of size 768; top_10): - s1 pod: 10 - p1 pod: 30 - p2 pod: 150	No	Yes	Yes	Yes	No	Commercial	Yes	No	20000	Yes	Somewhat (Need to convert date/time to integer in Unix time)	No	Yes (Sparse-Dense Vectors)	pinecone.io
Milvus	Yes	1,751	Yes	Yes	Yes	? (Depending on Hosting?)	Yes	Apache License 2.0	No	No	32768	Yes	Somewhat (Need to convert date/time to integer in Unix time)	No	No, they use the phrase "Hybrid Search", but it really means metadata filtering	milvus.io
ChromaDB	In memory of server	?	Yes	Not Yet	? (Depending on Hosting?)	? (Depending on Hosting?)	Yes	Apache License 2.0	No	No		Yes	Somewhat (Need to convert time to an integer)	No	query	chroma.com

https://medium.com/the-ai-forum/which-vector-database-should-you-use-choosing-the-best-one-for-your-needs-5108ec7ba133