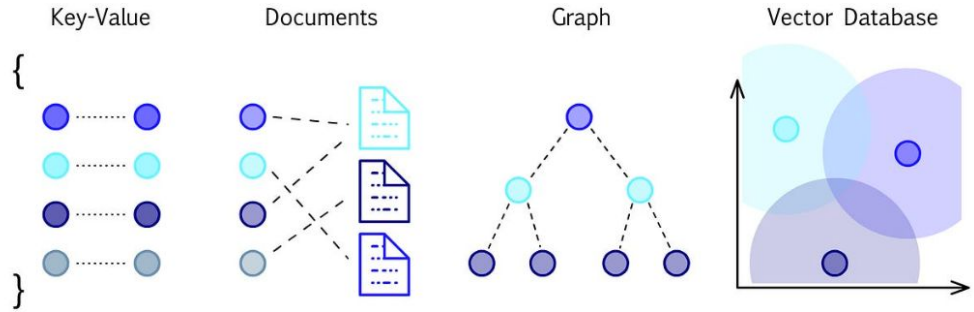
Vector Databases

Aditya Joshi
Aaditya Mankar
Grant Nickell
Atharva Sagale
Bipul Bishal Singh



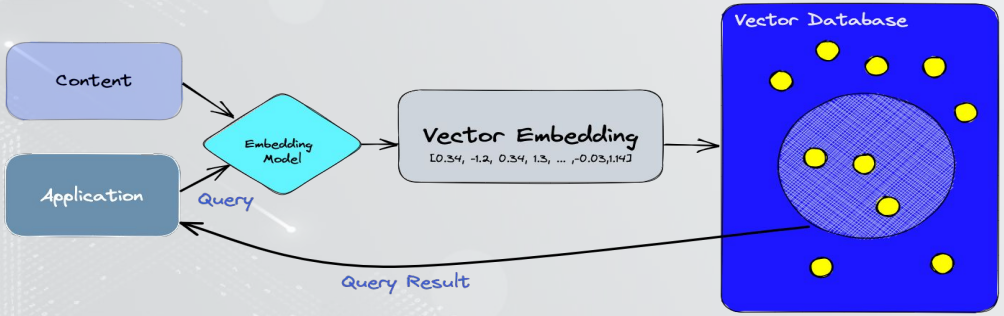
What are Vector Databases?

Storage Method: Stores data as multi-dimensional vectors that represent specific attributes, suitable for complex data including text, images, audio, and video.

Vector Creation: Uses machine learning, word embeddings, or feature extraction to generate vectors.

Advantages:

- Fast Retrieval: Quickly locates data based on vector similarity, enabling contextually relevant searches.
- Semantic Searches: Moves beyond exact matches, typical of traditional databases.



How does it work? How is it different?

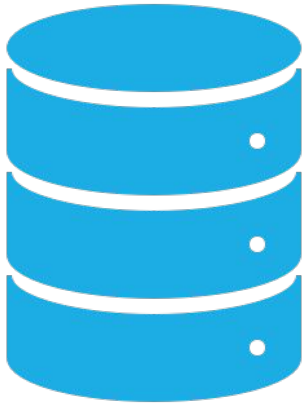
Data Type: Traditional databases store simple data like words and numbers in tables. Vector databases handle complex data (vectors) from embeddings of text, images, and audio.

Search Method:

- Traditional Databases: Search for exact matches using structured queries.
- Vector Databases: Use Approximate Nearest Neighbor (ANN) search techniques (e.g., hashing, graph-based searches) to find the closest matches based on similarity.

Key Concept: Embeddings

- Converts unstructured data into numerical vectors representing items' meanings or essences.
- Allows for efficient comparison and understanding of data by machine learning models.



Vector Databases in AI Applications

Purpose: Store high-dimensional vectors for fast, accurate similarity searches.

AI Integration: Essential for managing outputs from AI models in natural language processing and computer vision.

- **Example: Large Language Models (LLMs)**
  - **Models like GPT-3 transform massive data sets into high-dimensional vectors.**
  - **Vector databases handle the immense volume and complexity of this data, enabling efficient querying and management.**
- **Advantages Over Conventional Databases:**
  - **Conventional databases cannot effectively process the volume and complexity of vectorized data from models with billions of parameters.**
  - **Vector databases provide optimized environments for AI-driven applications, facilitating better performance and scalability.**

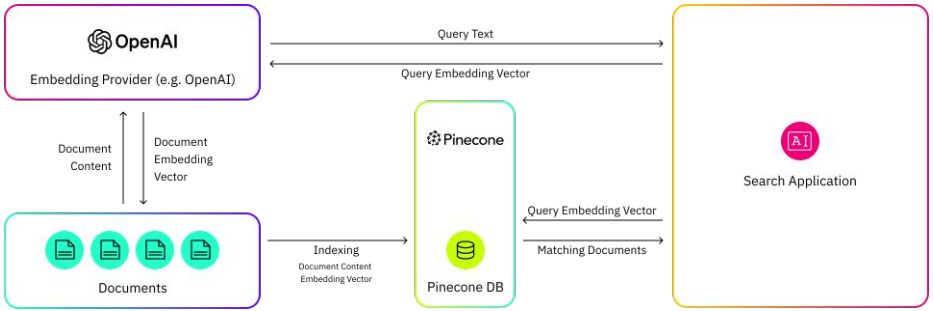
Practical Uses

Search Engines: Power advanced search functionalities in platforms like Spotify or YouTube, where users can find similar songs or videos based on content.

Recommendation Systems: Improve the relevance of recommendations in e-commerce and streaming services by understanding user preferences and item characteristics on a deeper level.

Content Discovery: Facilitate discovery in large databases of images or documents by enabling content-based retrieval rather than metadata or tags alone.

Popular Tools : Pinecone



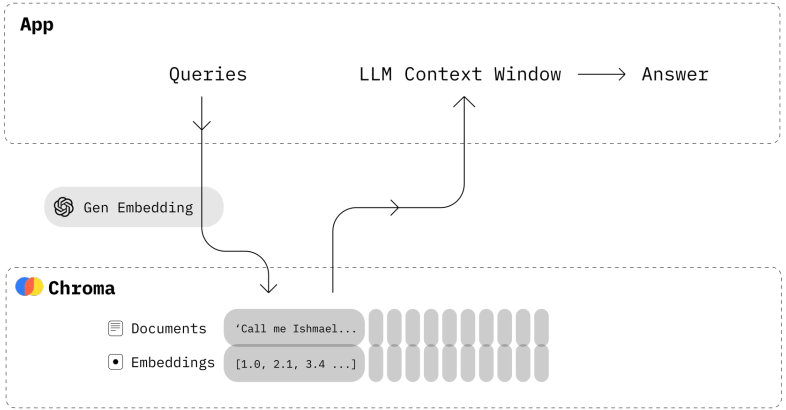
Open-source database.

Built to tackle challenges associated with high-dimensional data.

Cutting-edge indexing and search capabilities

Key feature:

- Provides Features like: Fully managed service
- Highly scalable
- Real-time data ingestion
- Low-latency search
- Integration with LangChain



Popular Tools : Chroma

Open-source embedding database.

Easy to build LLM apps

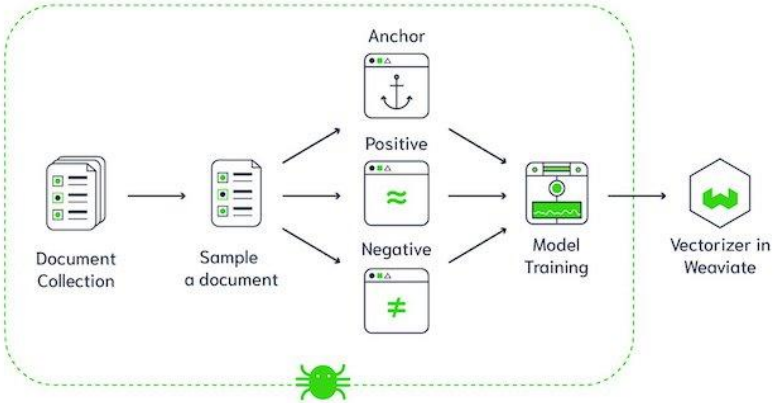
Supports LongChain (Python & JS), LlamaIndex

Python notebook can be scaled to Prod Cluster

Provides Features like:

- Queries
- Filtering
- Density estimates

Popular Tools : Weaviate



Open-source vector database.

Key Features:

- Speed
- Flexibility
- Production-ready
- Beyond search

Demo

References

https://www.datacamp.com/blog/the-top-5-vector-databases

https://www.trychroma.com/hrm4.svg

https://vectara.com/wp-content/uploads/2023/08/challenges-with-pinecone.svg

https://weaviate.io/assets/images/hero-46c2c60d3ba1ab6ff0d14cb04915e2f3.jpg