Understanding Vector Databases

Understanding Vector Databases

Vector databases are specialized database systems designed to store, index, and query high-dimensional vector embeddings efficiently. They are essential for modern AI applications that rely on semantic similarity search.

What are Vector Embeddings?
Vector embeddings are numerical representations of data (text, images, audio) in a high-dimensional space. Similar items are positioned closer together in this space, making it possible to find semantically similar items through mathematical distance calculations.

Popular Vector Databases:
- ChromaDB: Open-source, easy to use, great for prototyping
- Pinecone: Managed service, highly scalable, commercial
- Weaviate: Open-source, GraphQL API, with ML capabilities
- Milvus: Open-source, high performance, cloud-native
- Qdrant: Open-source, written in Rust, with filtering support

Distance Metrics:
- Cosine Similarity: Measures angle between vectors (0-1)
- Euclidean Distance: Straight-line distance between points
- Dot Product: Inner product of vectors
- Manhattan Distance: Sum of absolute differences

Applications:
- Semantic search engines
- Recommendation systems
- Duplicate detection
- Anomaly detection
- Image and video search
- Chatbots and question-answering systems

Performance Considerations:
Vector databases use specialized indexing algorithms like HNSW (Hierarchical Navigable Small World) and IVF (Inverted File Index) to enable fast approximate nearest neighbor search even with billions of vectors.