Vector Databases

THE COMPLETE GUIDE TO MODERN AI STORAGE

@SAMIR_SAIYED

What are Vector Databases?

- Store high-dimensional vectors (embeddings)
- Enable semantic similarity search
- Power AI applications like RAG & recommendations
- Handle text, images, audio data
- Built for modern ML workflows

Why Vector Databases Matter

- **Speed** - Optimized for similarity search at scale
- **Scalability** - Handle millions of vectors efficiently
- **Accuracy** - Advanced indexing algorithms
- **Real-time** - Fast inference for production
- **Flexible** - Support various distance metrics

2 Main Types of Vector Databases

🎯 **Purpose-Built**

Built specifically for vectors:

- Pinecone
- Weaviate
- Qdrant
- Chroma

🔧 **Extensions**

Traditional DBs + vector capabilities

- PostgreSQL (pgvector)
- Redis Vector Search
- Elasticsearch
- MongoDB Atlas

Purpose-Built Vector Databases

- **Pinecone**
  - Fully managed & serverless
  - Auto-scaling
  - Best for: RAG applications
- **Weaviate**
  - Open-source with GraphQL
  - Built-in ML models
  - Best for: Knowledge graphs
- **Qdrant**
  - Rust-based, high performance
  - Advanced filtering
  - Best for: Production scale

Vector Extensions

- **PostgreSQL + pgvector**
  - ACID compliance
  - Familiar SQL interface
  - Best for: Existing Postgres users
- **Redis Vector Search**
  - In-memory operations
  - Ultra-fast queries
  - Best for: Real-time applications
- **Elasticsearch**
  - Full-text + vector search
  - Analytics capabilities
  - Best for: Hybrid search

Vector DB Use Cases

🛒 E-commerce

- Product - recommendations
- Visual search

💰 Finance

- Fraud detection
- Document analysis

🎬 Media

- Content - recommendations
- Duplicate - detection

🏥 Healthcare

- Drug discovery
- Medical imaging

How to Choose Your Vector Database

CONSIDER THESE FACTORS:

- **Scale** - How many vectors & queries?
- **Budget** - Managed vs self-hosted
- **Integration** - Existing tech stack
- **Features** - Filtering, metadata, updates
- **Performance** - Latency requirements
- **Compliance** - Data governance needs

Quick Performance Overview

🏃‍♂️ Speed Leaders

- 🥇 Redis (in-memory)
- 🥈 Qdrant (Rust-based)
- 🥉 Pinecone (managed)

📈 Scalability Champions

- 🏆 Pinecone (serverless)
- 🏆 Weaviate (distributed)
- 🏆 Elasticsearch (sharding)

💡 Cost-Effective Options

pgvector (open-source) • Chroma (local dev)

Ready to Get Started?

- **Step 1**: Define your use case
- **Step 2**: Estimate scale & budget
- **Step 3**: Try free tiers
- **Step 4**: Build your first vector search!

What's your vector database experience?
Comment below! 👇