



# Vector Databases with LangChain

Unlocking the power of semantic search and AI-driven data retrieval for modern applications

Sivakumar Rajendran

# What Are Vector Databases?

Vector databases revolutionize how we store and retrieve information by converting data into high-dimensional numerical representations called **embeddings**.

Unlike traditional databases that search for exact matches, vector databases find **semantically similar** content—understanding meaning rather than just matching keywords.







# Why Vector Databases Matter



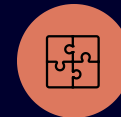
## Semantic Understanding

Search by meaning, not just keywords—find "king" when searching for "monarch"



## Lightning-Fast Retrieval

Optimized algorithms return relevant results from millions of records in milliseconds



## AI Integration

Essential foundation for RAG systems, chatbots, and LLM-powered applications

# From Text to Vectors: The Embedding Process



## Raw Text

Original documents, queries, or content



## Embedding Model

AI model converts text to numerical vectors



## Vector Representation

High-dimensional arrays capturing semantic meaning



## Storage

Indexed vectors ready for similarity search

Each word, sentence, or document becomes a point in multidimensional space where **similar meanings cluster together**.

# How Similarity Search Works

## Distance Metrics

Vector databases measure similarity using mathematical distance calculations:

- **Cosine Similarity**—measures angle between vectors
- **Euclidean Distance**—straight-line distance in vector space
- **Dot Product**—magnitude and direction comparison

The **closer two vectors** are in this space, the more semantically similar their content.





# LangChain's Vector Store Integration

1

## Choose Your Database

LangChain supports Pinecone, Chroma, Weaviate, FAISS, and 30+ other vector stores

2

## Load & Split Documents

Import content and chunk it into manageable pieces for embedding

3

## Generate Embeddings

Use OpenAI, HuggingFace, or custom models to create vector representations

4

## Query & Retrieve

Search with natural language and receive semantically relevant results



# Real-World Applications



## Conversational AI

Power chatbots with contextual memory and relevant knowledge retrieval from vast document collections



## Recommendation Systems

Suggest products, content, or services based on deep semantic similarity



## Document Intelligence

Analyze, categorize, and extract insights from large document repositories



## Semantic Search Engines

Build search systems that understand intent and context, not just keywords



## Anomaly Detection

Identify outliers and unusual patterns by measuring vector distance from normal behavior



## RAG Systems

Retrieval-Augmented Generation combines LLMs with vector search for grounded responses

# Building a RAG Pipeline

01

## Document Ingestion

Load source documents using LangChain's document loaders

02

## Text Chunking

Split documents into optimal-sized pieces with overlap for context preservation

03

## Vector Embedding

Convert chunks into embeddings using your chosen model

04

## Index Creation

Store vectors in your database with metadata for filtering

05

## Query Processing

Convert user questions into vectors and retrieve relevant chunks

06

## LLM Generation

Feed retrieved context to LLM for accurate, grounded responses



# Performance Optimization Strategies

## Indexing Methods

- HNSW graphs for speed
- IVF for memory efficiency
- Product quantization for compression

## Metadata Filtering

- Pre-filter by date, category
- Hybrid search combinations
- Dynamic query refinement

## Chunking Strategy

- Optimal chunk size: 500-1000 tokens
- Overlap: 10-20% for context
- Semantic splitting at boundaries

## Model Selection

- Balance quality vs. speed
- Domain-specific embeddings
- Regular model updates



# Start Building with Vector Databases



## Quick Start

Use LangChain's simple API to integrate vector stores in minutes



## Rich Ecosystem

Leverage 30+ supported databases and embedding providers



## Production-Ready

Scale from prototype to production with enterprise-grade solutions

Vector databases are transforming how we build **intelligent applications**. With LangChain's unified interface, you can experiment, iterate, and deploy **semantic search solutions** faster than ever before.