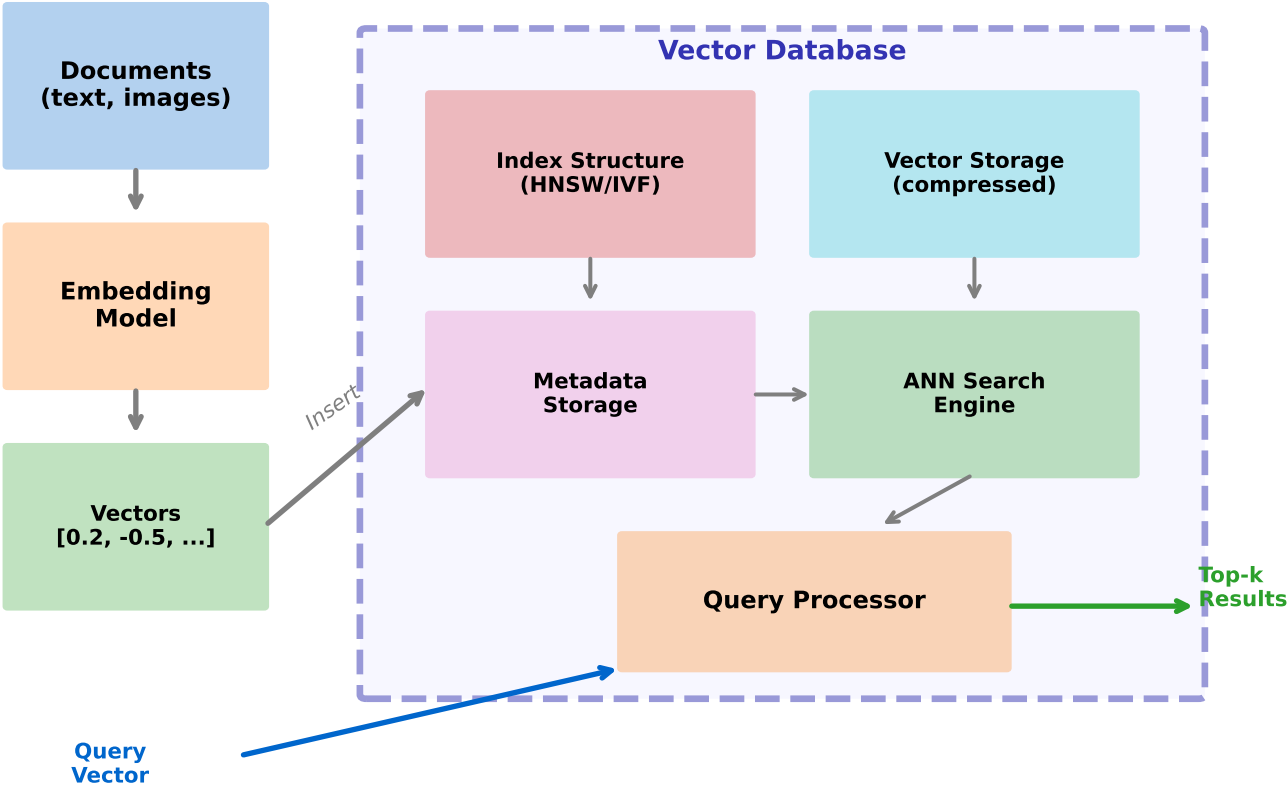# Vector Database: Architecture and Role in RAG

How Vector DB Works Internally

Documents (text, images)

↓

Embedding Model

↓

Vectors [0.2, -0.5, ...]

Query Vector

Vector Database

Index Structure (HNSW/IVF)

Vector Storage (compressed)

↓ (Insert)

Metadata Storage → **ANN Search Engine**

Query Processor

↓

Top-k Results

Vector DB in RAG Pipeline

User Query

"What is the capital of France?"

↓

Embed Query

↓

Vector Database

ANN Search: find similar docs

↓

Retrieved Documents (top-k)

z_1 — Paris is capital...

z_2 — France in Europe...

z_3 — Eiffel Tower...

Query context

Key Role: Fast retrieval of relevant context

LLM Generator

$p(y|x, z_1, z_2, z_3)$