30th-may- vector-dB

Sunday, June 1, 2025 6:04 PM

Doctor Appointment System using LangGraph, FastAPI, and Streamlit

This project is a multi-agent, AI-powered Doctor Appointment Booking System that simulates a realistic medical assistant capable of handling user queries regarding doctor availability, specialization, and appointment booking.

Tech Stack:

LangGraph for workflow automation between agents

LangChain for Model Loading, Prompt Creation and tool usage

FastAPI for serving the API endpoint /execute

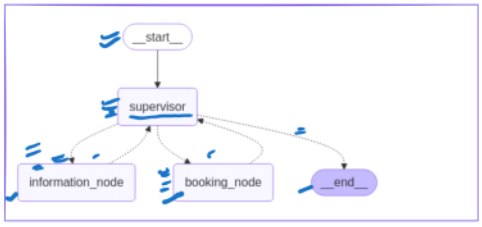
Streamlit for frontend interaction

Python + Pandas + CSV for data handling

Sample Use Case:

User: Can you check if a dermatologist is available on 5th August at 10:00 AM?

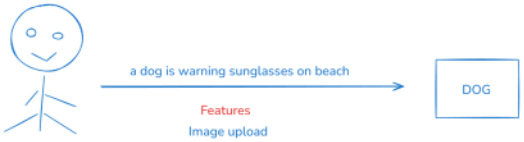
Agent Response: Dr. Smith (Dermatologist) is available at 10:00 AM on 05-08-2024. Do you want to book the appointment?



product

professional grade system design

tech stack data flow, deployment



Features

Image upload

Natural language query input

Show results with similarity scores

Optional filters (category, color, etc.)

Endpoints:

/upload_image: Accept image, generate & store embedding

/search: Accept text query, return top-k similar images

/images: Serve stored images (from S3 or local)

Embedding Layer (Multimodal Models)

Image Encoder: CLIP / BLIP / GIT (OpenAI or HF Transformers)

Text Encoder: Same model (CLIP text encoder) or sentence-transformer

Tech: transformers (Hugging Face) sentence-transformers

open-clip-torch or Salesforce/BLIP

Vector Database

Stores image embeddings

Supports similarity search (text → image)

Options: FAISS (local, fast) Qdrant/Weaviate (cloud/scale ready)

Embedding stored with: image_path meta (timestamp, uploader, etc.)

Image Storage

Local: ./data/images/

Or S3: s3://bucket-name/image-name.jpg

Serve via signed URL or FastAPI static

LLM (Optional Features)

Use-case: Query rewriting, semantic expansion

"beach dog → a dog playing on a sunny beach"

Model: GPT-3.5 (OpenAI) or phi-2 via Hugging Face

Component	Tool/Platform

"beach dog → a dog playing on a sunny beach"

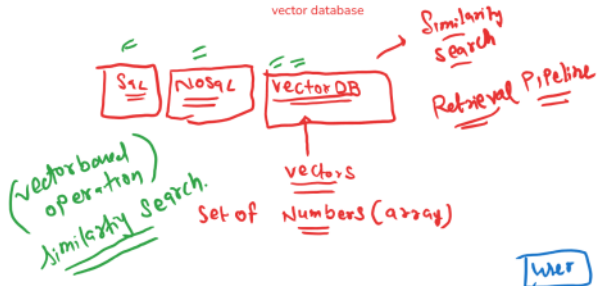
Model: GPT-3.5 (OpenAI) or phi-2 via Hugging Face

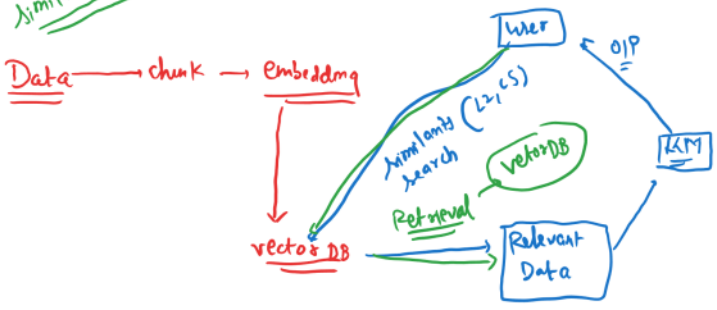
Component	Tool/Platform
Frontend	Streamlit / HuggingFace Spaces
Backend API	Docker + EC2 / Railway / Render
DB	FAISS (file-based) or Qdrant Cloud
Image Storage	AWS S3 / Local
Model Hosting	Local inference or HF Inference Endpoint
Orchestration	Docker Compose

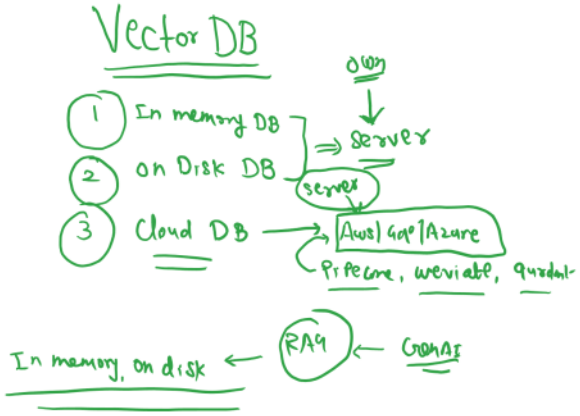
first workflow:

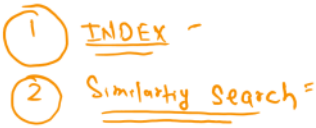
User Upload Image →
FastAPI Endpoint (/upload_image) →
CLIP Image Encoder →
Get Embedding →
Store in Vector DB (with image path)

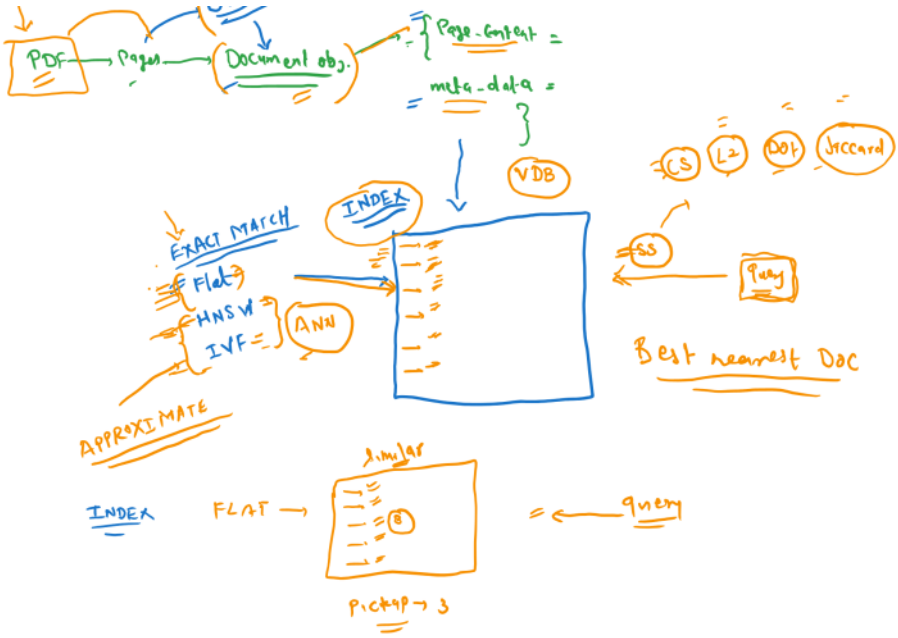
User Types Text →
FastAPI Endpoint (/search) →
Text Embedding via same CLIP encoder →
Vector DB Similarity Search →
Return top-K matching image paths →
Streamlit Displays Results





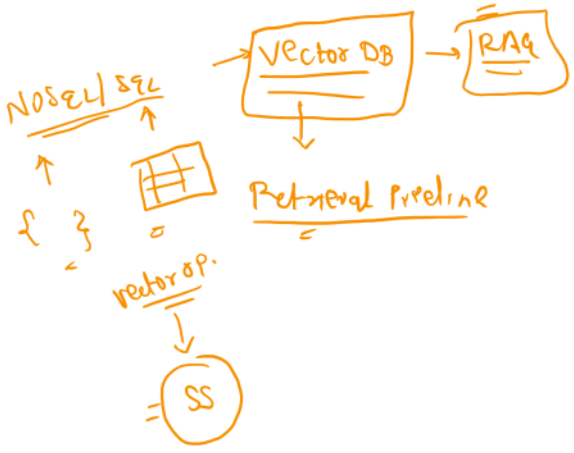


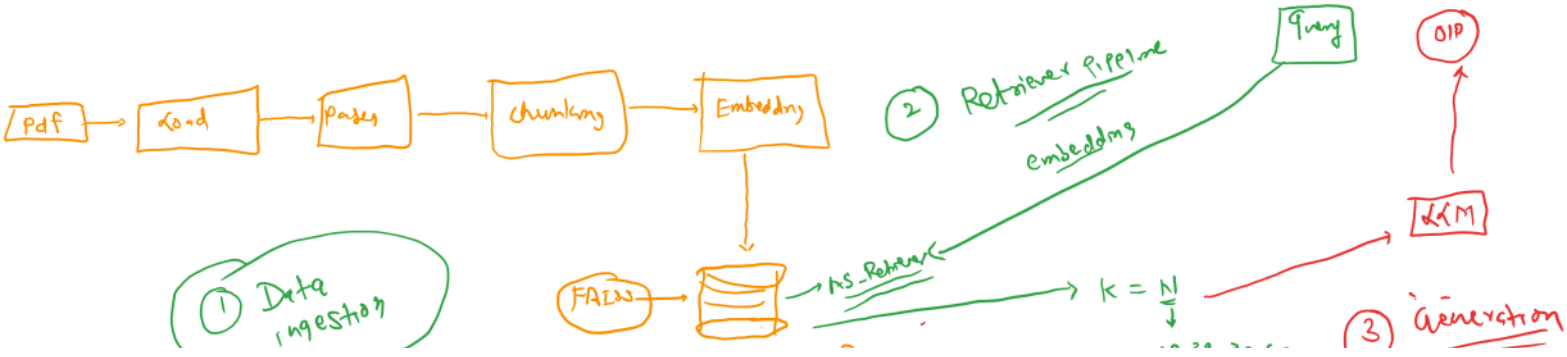


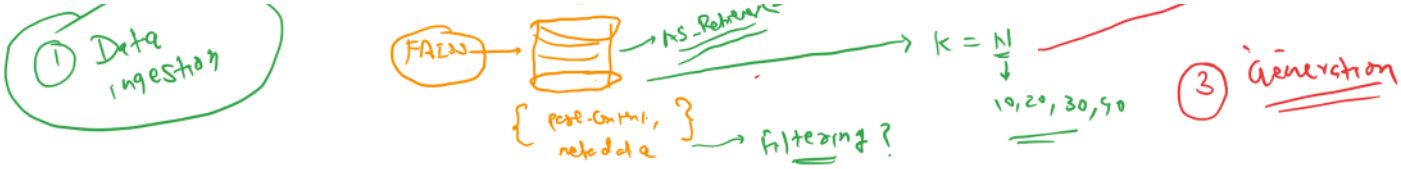


Graph based → HNSW

cluster based → IVF







Why we use VDB ?

= we use VDB for Creating Retriever Pipeline

1st-june-vector-dB

Sunday, June 1, 2025 6:06 PM

Type	Description	Example DBs
Type	**Description**	**Example Databases**
Local	Runs on your own laptop or server, either on the file system or via Docker	FAISS, Qdrant (local), Weaviate
In-Memory	Stores data only in RAM — data is lost when the system restarts	FAISS, Qdrant (local), Weaviate
Cloud Hosted	Fully hosted by vendors — supports persistence, scaling, APIs, monitoring, etc.	Pinecone, Qdrant Cloud, Weaviate Cloud, Milvus (Zilliz Cloud)

Local: useful for **experiments**, dev, prototypes, or small deployments

Cloud: needed for **prod systems** with

- Scalability
- High availability
- Multi-user/team access
- Real-time updates
- Monitoring & backups

Feature	In-Memory	Local Persistent	Cloud Hosted
Data survives restart	No	Yes	Yes
Scales across nodes	No	Limited	Yes
Team access (multi-user)	No	Manual	Yes
Easy integration (LangChain etc.)	SDKs	SDKs	SDKs/APIs ready
Use-case fit	Small, fast-only	Dev/PoC	Production scale

Vendor	Local Support	Cloud Hosted
FAISS	Yes	No
Quadrant	Yes	Yes
Weaviate	Yes	Yes
Pinecone	No (API only)	Only
Milvus	Yes	(Zilliz)
Chroma	Yes	No

https://python.langchain.com/docs/integrations/vectorstores/
https://superlinked.com/vector-db-comparison

Term	Full Form	Meaning in Simple Terms
KNN	K Nearest Neighbors	**Exact search** : Return top K **truly closest** points
ANN	Approximate Nearest Neighbors	**Faster search** : Return top K **almost closest** points (approximate)

KNN (Exact)	ANN (Approximate)
Compares query with **all vectors**	Compares with a **smart subset**
Slow but 100% accurate	**Fast but ~95–99% accurate**
Works fine for small datasets	Best for large-scale datasets (millions)
Examples: brute-force FAISS flat	Examples: HNSW, IVF, PQ, ScaNN, etc.

Modalities = Different types of input data
Text → plain sentences, paragraphs
Image → photos, diagrams, charts
Audio → voice, speech, music
Video → moving visuals with audio
Documents → combination of text + images

Use-case	**Modalities Involved**
ChatGPT Vision	Image + Text
YouTube caption search	Video + Audio + Text
Resume screening from PDF + profile	Text + Table + Image (logos etc.)
Search from screenshots	Image → OCR → Text → Semantic Search
Voice-commanded document search	Audio → Text + Query understanding
Medical Report Q&A	Document (PDF) = Text + Charts + Tables

Text → BERT / GPT / MiniLM → vector

Image → CLIP / BLIP → vector

Audio → Whisper / Wav2Vec → vector

Video → Sampled frames + CLIP → vector

Table → LayoutLM / TAPAS → vector

The **difference is not in creating the vector;** the difference lies in the **features and ecosystem of the vector database.**

Pinecone

```
{"id": "doc-1", "values": [0.22, 0.88, ...]}
```

Weaviate

```
{
  "class": "Product",
  "properties": {
    "image": "path.jpg",
    "text": "Best shirt in market"
  }
}
```

Vector DB	Verdict	Use-case Fit
Milvus (Zilliz)	**good for multimodal**	Image + Text + Video + Audio, production scale
Qdrant	Excellent, very flexible	Text + Image + Metadata hybrid search
Weaviate	Good with Transformers + hybrid	RAG, text + image + metadata combo
Pinecone	Text-only focus (no native image)	Great for text-based RAG, not for images
Vespa.ai	Enterprise-heavy multimodal infra	Multimodal + advanced query logic
AWS OpenSearch (KNN)		
Azure AI Search		
GCP Vertex AI Vector Search		