Chunking — Embedding

Embedding

User Query

Vector Database

Retrieved Data

{ }

Prompt+Query+Retrieved Data

eraser

Chunking

Embedding

Embedding

Vector Database

User Query

Retrieved Data

Prompt+Query+Retrieved Data

eraser

Anything

Langchain

Chunking

Embedding

Collect

User Query

Final output

Embedding

(1) Data ingestion

Vector based algorithm

Vector Database

Vector op

Diff

Local

{vendor}

Retrieved Data

K

Prompt+Query+Retrieved Data

		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		

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"
  }
}
```

Multi model data

bs4 *embedding*

Vector DB	Verdict	Use-case Fit
Milvus (Zilliz)	good for multimodal	Image + Text + Video + Audio, production scale
Qdrant	Excellent, very flexible	Text + Image + Metadata

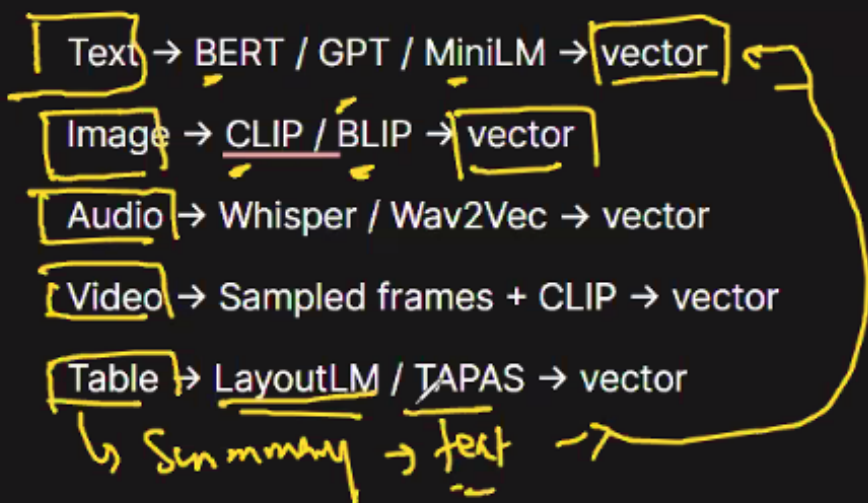|	
YouTube caption search	Video + Audio + Text
Resume screening from PDF + profile	Text + Table + Image (logos etc.)
Search from screenshots	Image → OCR → Text → Semantic Search
Voice-commanded document search	Audio → Text + Query understanding
Medical Report Q&A	Document (PDF) = Text + Charts + Tables

Text → BERT / GPT / MiniLM → vector

Image → CLIP / BLIP → vector

Audio → Whisper / Wav2Vec → vector

Video → Sampled frames + CLIP → vector

Table → LayoutLM / TAPAS → vector

↳ Summary → text →

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

Video → moving visuals with audio

Documents → combination of text + images

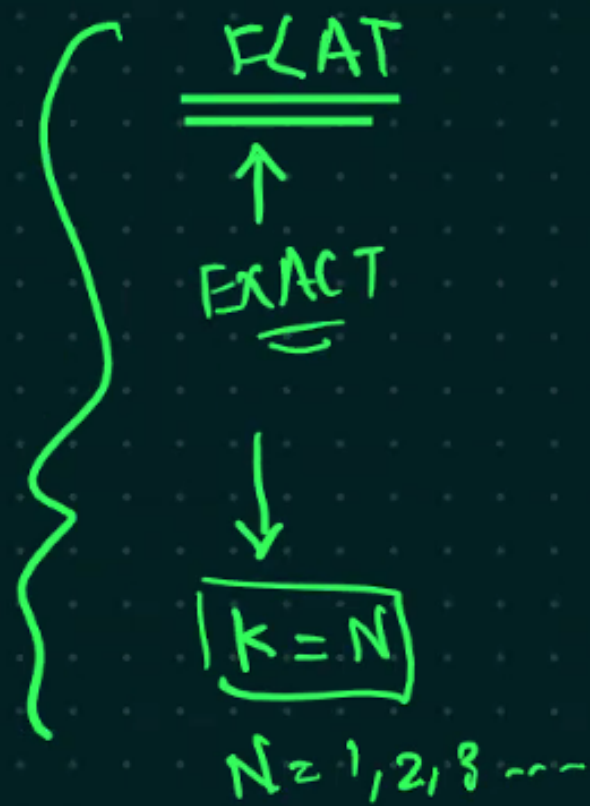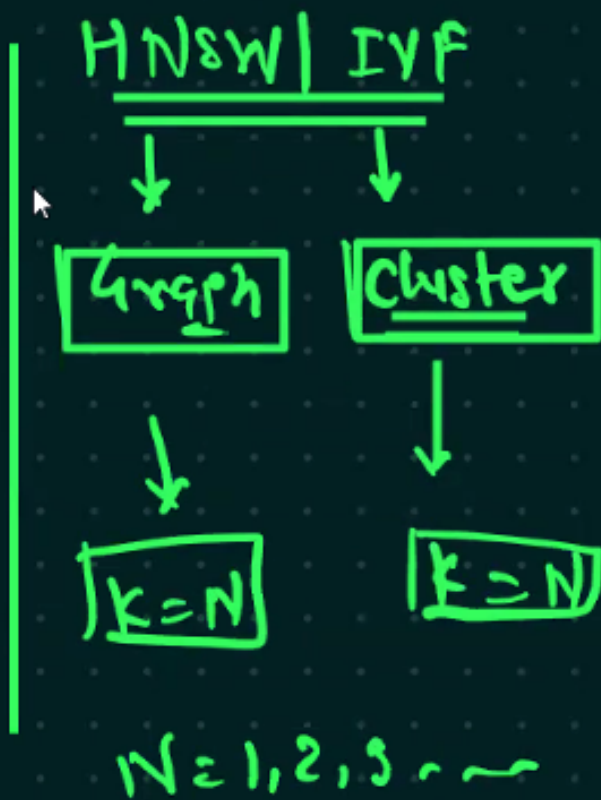| Use-case	Modalities Involved
ChatGPT Vision	Image + Text
YouTube caption search	Video + Audio + Text
Resume screening from PDF + profile	Text + Table + Image (logos etc.)
Search from screenshots	Image → OCR → Text → Semantic Search
Voice-commanded document search	Audio → Text + Query understanding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

Text → BERT / GPT / MiniLM → vector

Image → CLIP / BLIP → vector

Vector Database ···

Document | Both | Canvas

Vector Database

Modalities = Different types of input data

Text → plain sentences, paragraphs

Image → photos, diagrams, charts   } image.

Audio → voice, speech, music   = Audio

Video → moving visuals with audio   ← Frame + Audio
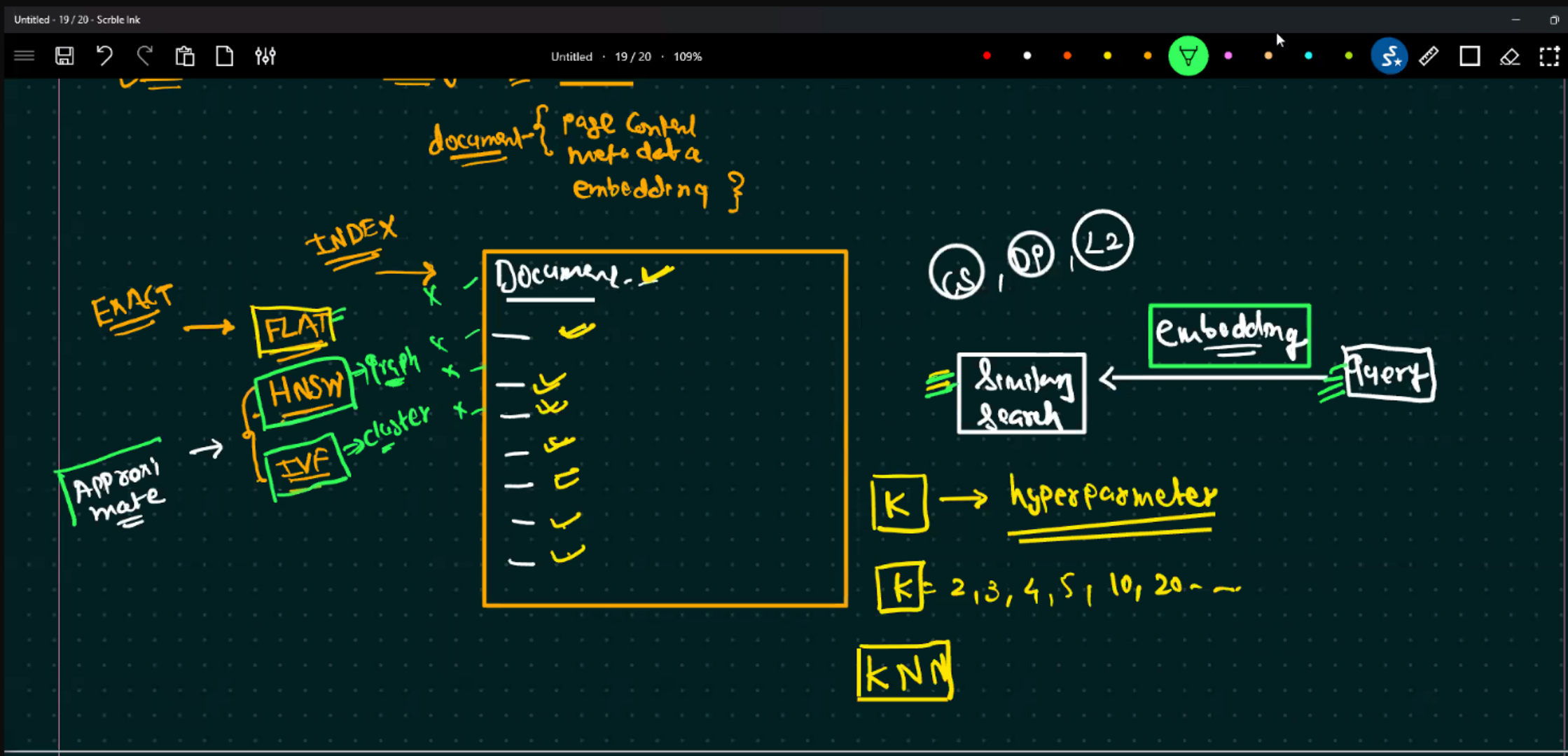
Documents → combination of text + images   ← PI

Use-case	Modalities Involved
ChatGPT Vision	Image + Text
YouTube caption search	Video + Audio + Text
Resume screening from PDF + profile	Text + Table + Image (logos etc.)
e → OCR → Text → Semantic Search	

FLAT

↑

EXACT

↓

$k = N$

$N = 1, 2, 3 \cdots$

$= \boxed{kNN}$

HNSW | IVF

↓   ↓

Graph   Cluster

↓   ↓

$k = N$ $k = N$

$N = 1, 2, 3 \cdots$

$= \boxed{ANN}$

document → { page content
              meta data
              embedding }

INDEX

ENACT

FLAT

HNSW → graph

IVF → cluster

Approximate

Document

CS , DP , L2

Embedding

Similary Search ← Query

K → hyperparameter

K = 2, 3, 4, 5, 10, 20 ~ ~

KNN

| | (LangChain etc.) | | |
| 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

⤢ Langchain Vector DB Comparison

⤢ VectorDB Comparison

Term	Full Form	Meaning in Simple Terms
	eighbors	Exact search
		: Return top K

Vector Database

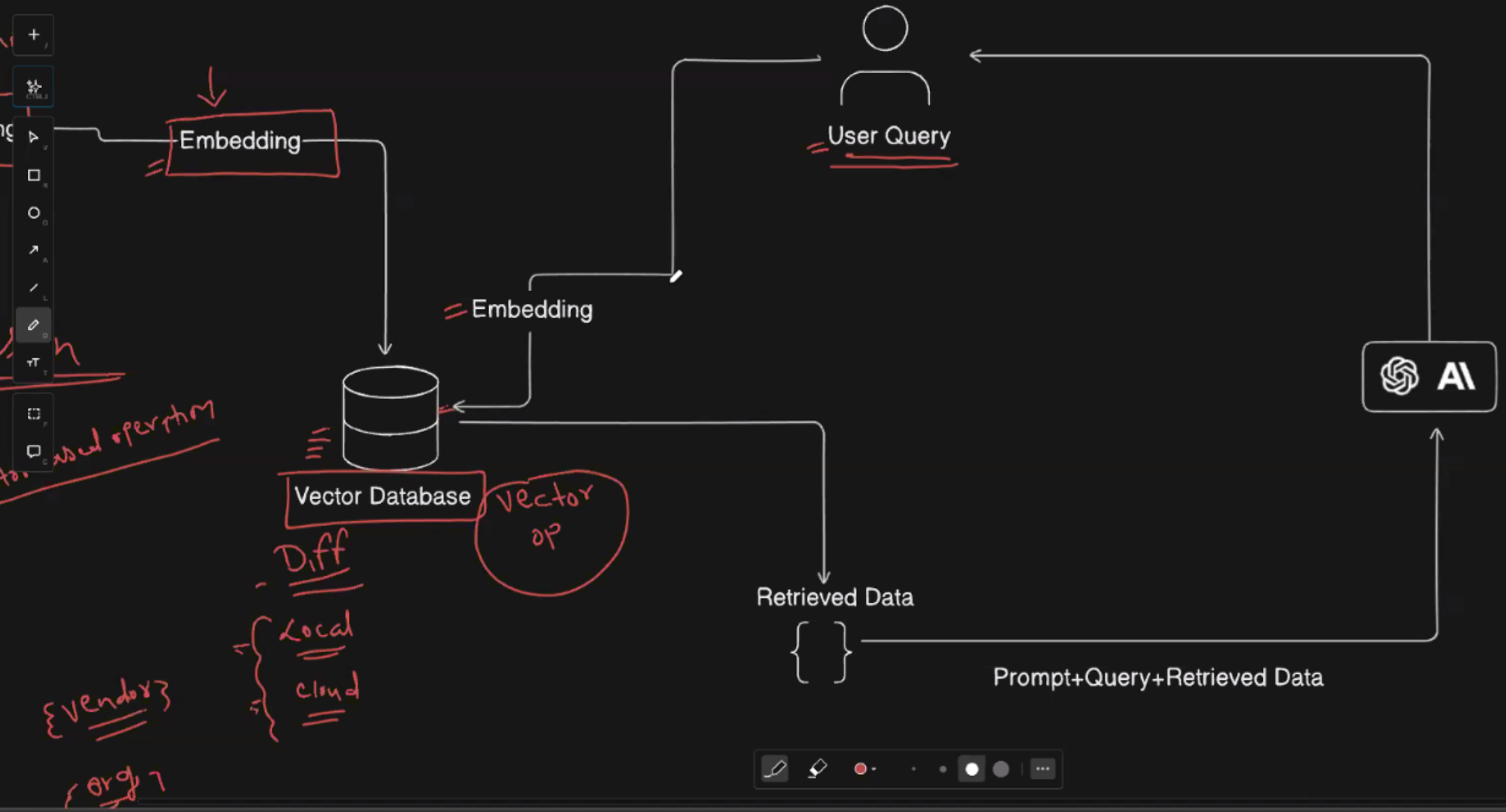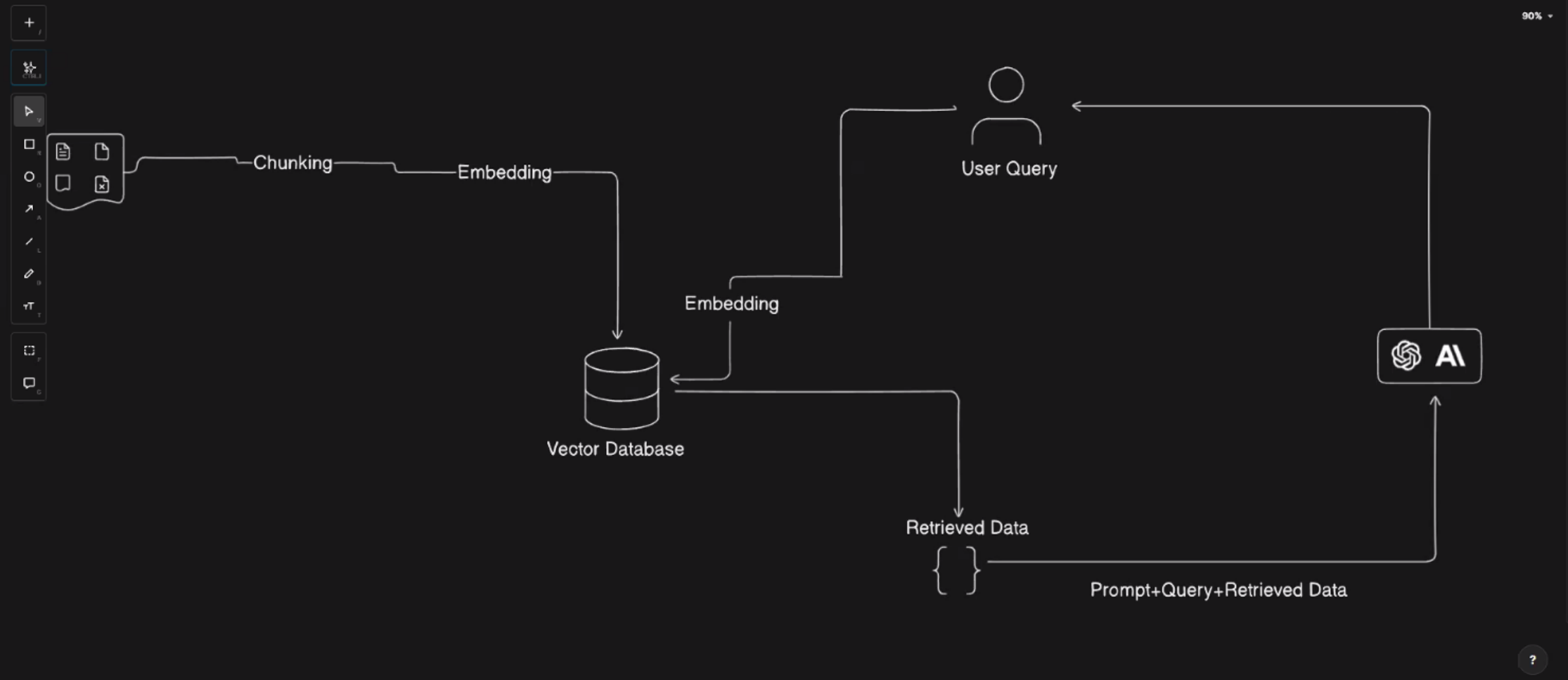| Type	Description	Example DBs
Type	Description	Example Databases
Local	Runs on your own laptop or server, either on the file system or via Docker	FAISS, Qdrant (local), Weaviate
In-Memory	Stores data only in RAM — data is lost when the system restarts	FAISS, Qdrant (local), Weaviate
Cloud Hosted	Fully 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

Vector Database

Document | Both | Canvas

Embedding

User Query

Embedding

Vector Database

Vector op

based operation

Diff

Local

Cloud

{vendor}

{org}

Retrieved Data

{ }

Prompt+Query+Retrieved Data

AI

Vector Database ···

Document | Both | Canvas

Ctrl K   Share

90%

Chunking

Embedding

User Query

Embedding

Vector Database

Retrieved Data

{ }

Prompt+Query+Retrieved Data

Text Generation Workflow

User → User Input → Embeddings model → Embedding: | 0.89 | -0.02 | -0.53 | 0.95 | 0.17 | ••• | -0.38 | → Semantic search → Vector store → Context → Prompt augmentation → Large Language Model → Response
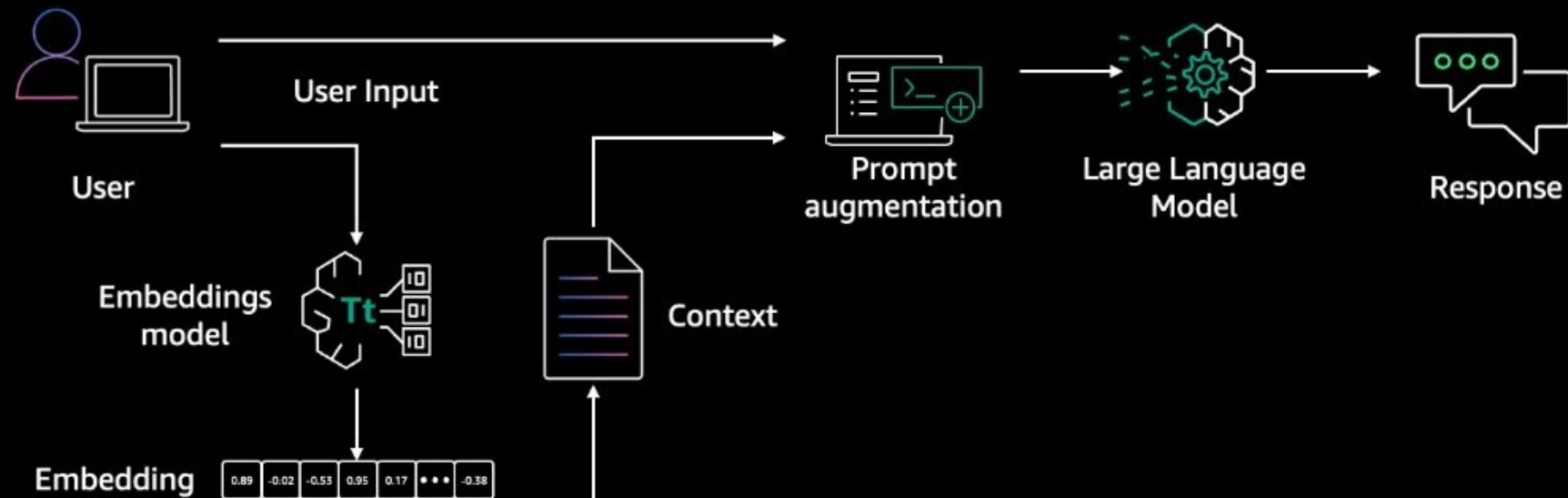
Data Ingestion Workflow
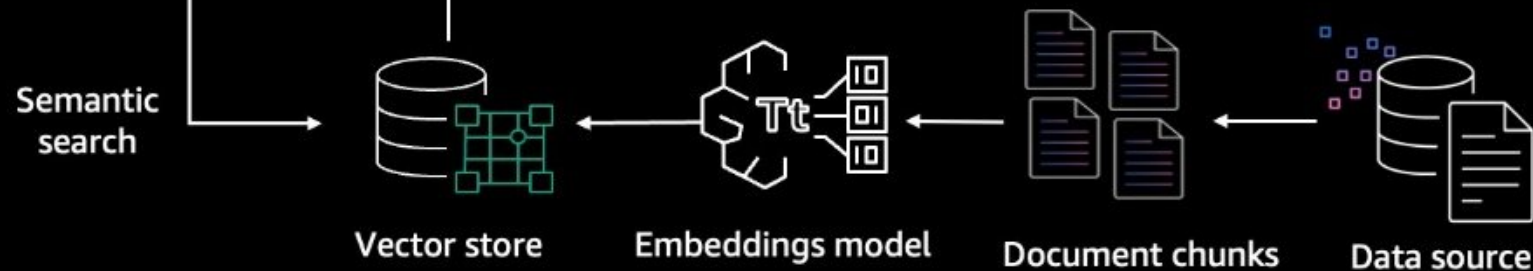
Semantic search → Vector store ← Embeddings model ← Document chunks ← Data source

Vector Database

| Document | Both | Canvas |

- Real-time updates
- Monitoring & backups

Vector Database

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