

RAG

- ① Pinecone
- ② MMRAG

Data → Text, Image, Audio, Table, Video

- ① RAG
- ② Langchain
- ③ VDB

- ✓ ① FAISS ⇒ Inmemory
- ✓ ② ChromaDB ⇒ Disk

Text, Image, Table, Audio, video

MMR4

Data

Extraction

Parse

embedding

Data ingestion

USER

query

Response (generation)

Similarity search

Retriever

LLM

Document (PDF, Docx, PPT, XML)

PDF

Vector DB

Relevant doc

TEXT
IMAGE
TABLE

Extract

Parse

Python code

- ① Docx
- ② Pypdf
- ③ PyMuPDF

Docx

Langchain

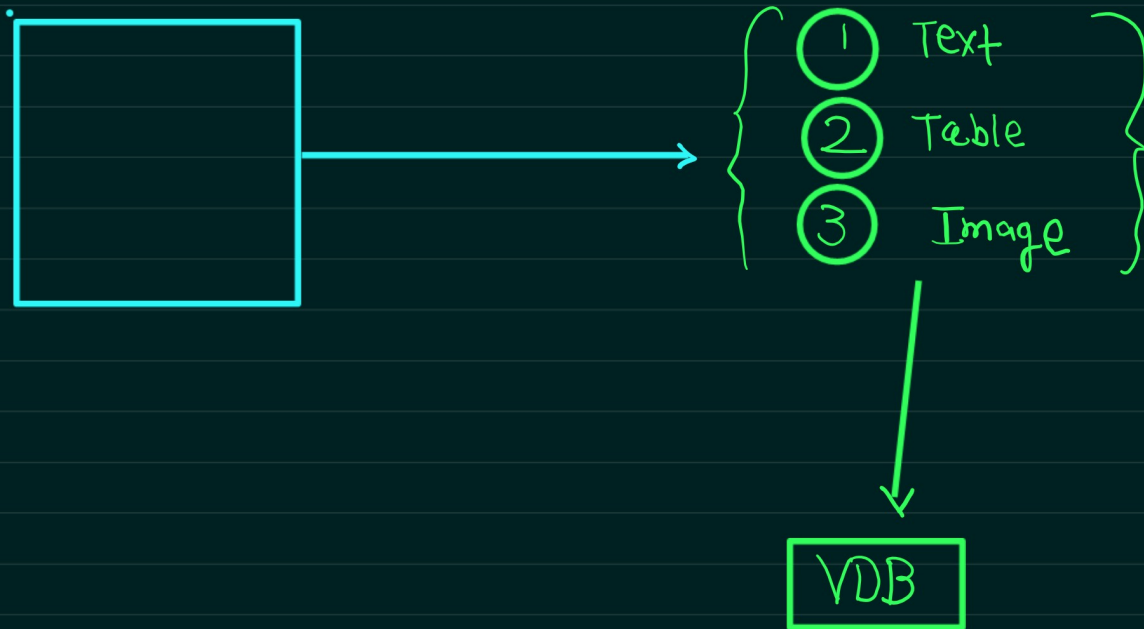
Unstructure

Parse any doc

element

- Image
- text
- tables

- Docx
- PPT
- PDF
- XML
- CSV



Document object

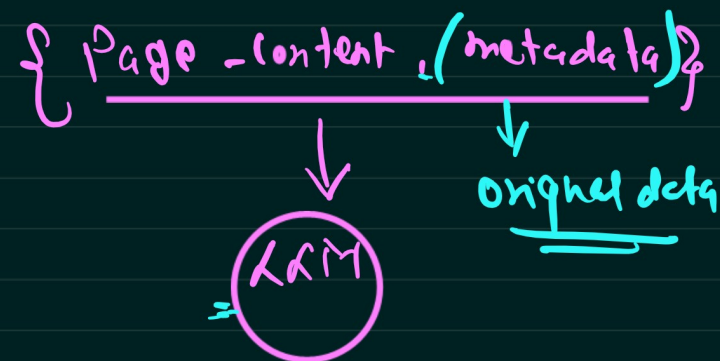
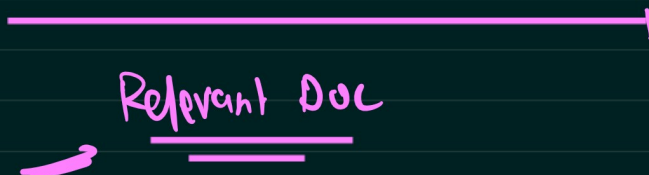
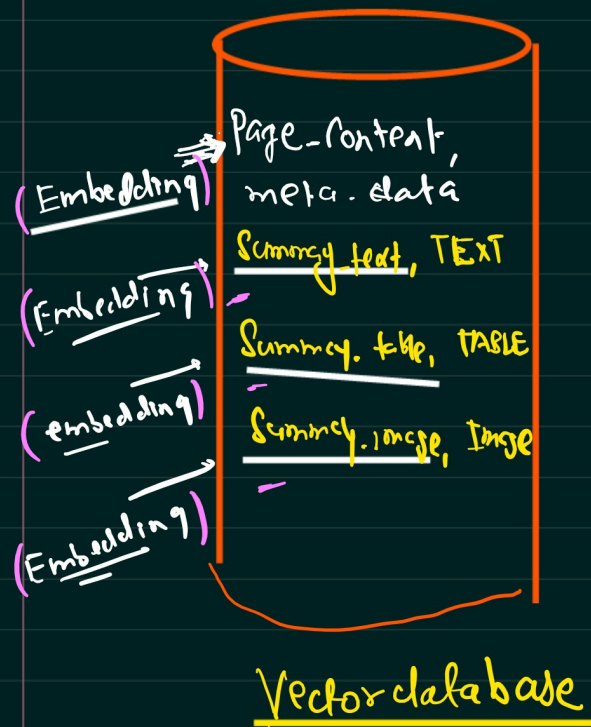
Document(Page-content, meta-data)

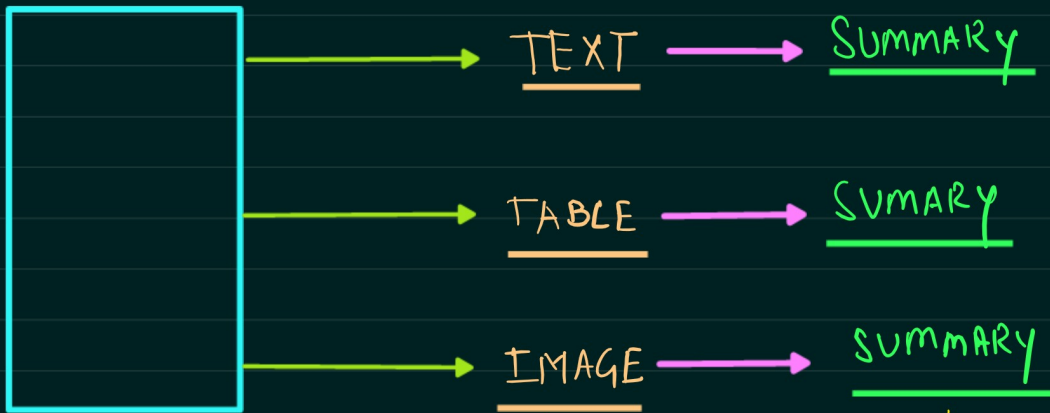
TEXT → SUMMARY
TABLE → SUMMARY
IMAGE → SUMMARY

Page-content

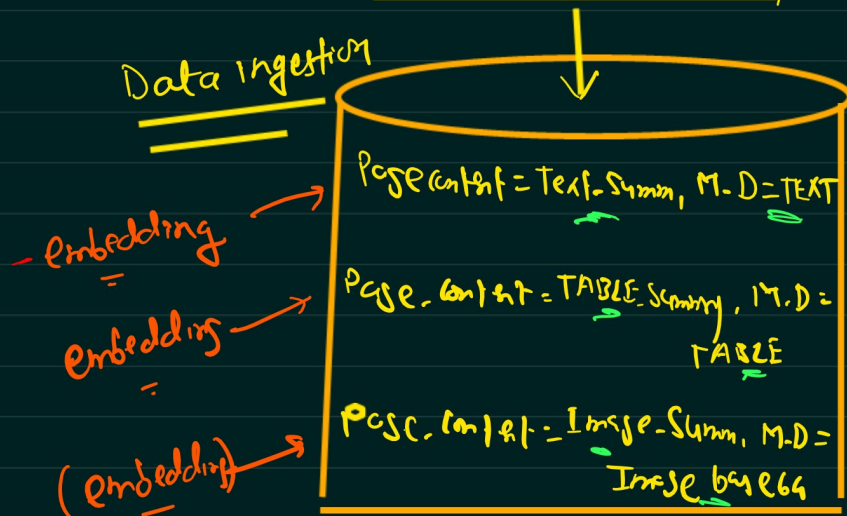
Original TEXT
Original TABLE
Original IMAGE

Base64
string





Data Ingestion



Data ing

