**🧩 1. Environment Setup ✅ (You’ve already done this)**

Setting up the development environment — installing all the tools, frameworks, and libraries you need for the project (like Python, Docker, Node.js, etc.).

**📥 2. Ingestion & Processing**

Collect and process your documents.  
This means:

* Loading raw documents from various sources (PDFs, text files, websites…)
* Cleaning and splitting them into smaller chunks (to make them easier to handle later)

**🧠 3. Embedding + Storage in a Vector Database**

* Convert each text chunk into an **embedding** (a numerical representation of meaning) using a model like Sentence Transformers.
* Store all embeddings in a **vector database** such as Milvus or Weaviate so you can search them later.

**🔍 4. Semantic Search**

* When a user asks a question, convert it into an embedding
* Search in the vector database for the most similar text chunks
* Retrieve those chunks as **context** to help the LLM answer accurately

**🤖 5. Combine with an LLM**

* Feed the retrieved chunks + user question to a Large Language Model (like OpenAI GPT or LLaMA)
* The LLM uses the retrieved knowledge to **generate the final answer**

This whole process is called **RAG (Retrieval-Augmented Generation)**.

**💻 6. User Interface (UI)**

* Build a simple web app (for example with Streamlit or React) where users can type their questions and see the LLM’s answers.

**📦 7. Containerization & Deployment**

* Use Docker to containerize your project (pack it with all dependencies)
* Deploy it on a free cloud server or local machine using Kubernetes or Docker Compose to run it in production