

RAG System Recap

Instructor

Dipanjan Sarkar

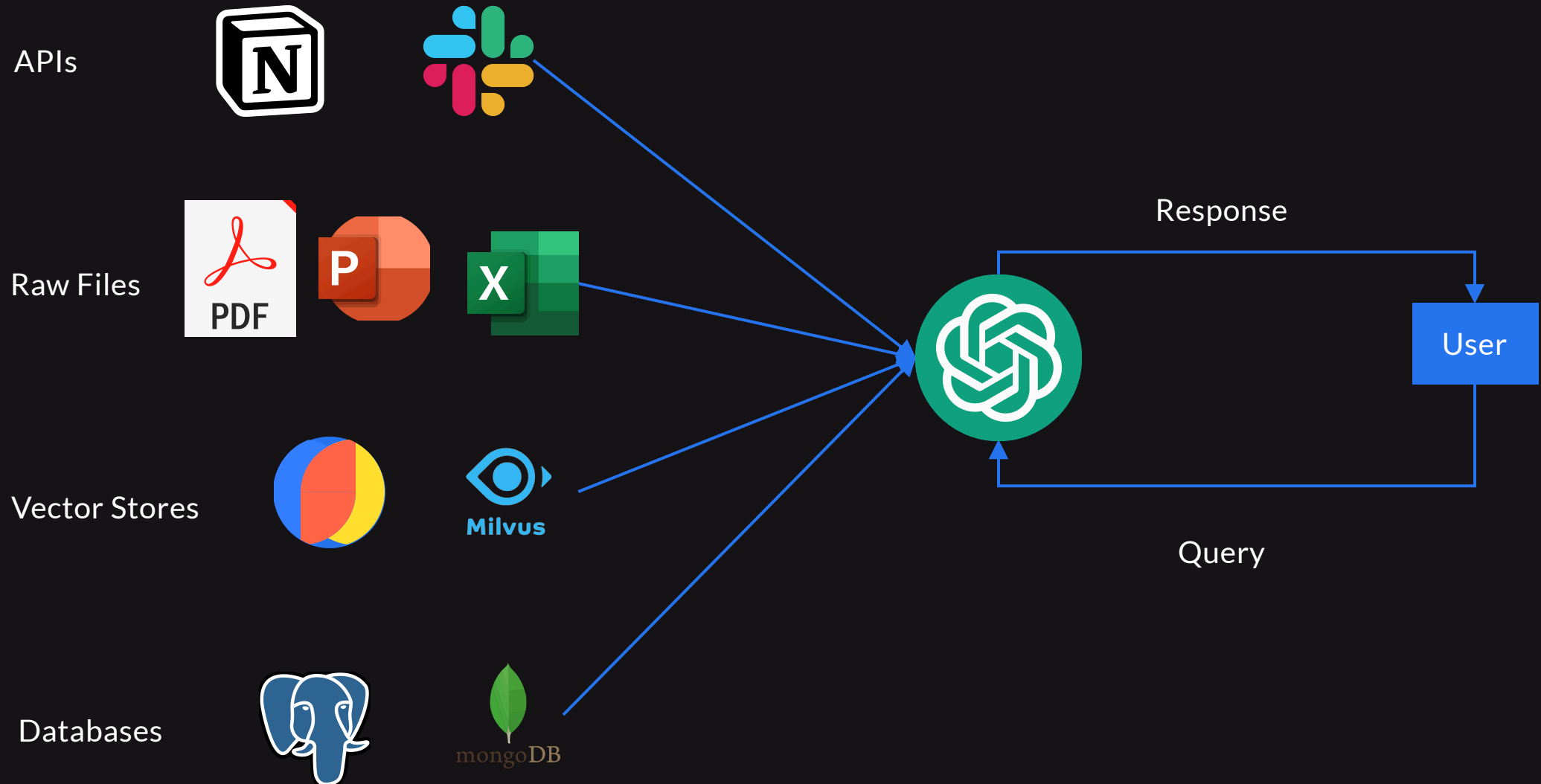
Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

Published Author



What is a RAG System



What is a RAG System

Step 1

Data Processing & Indexing

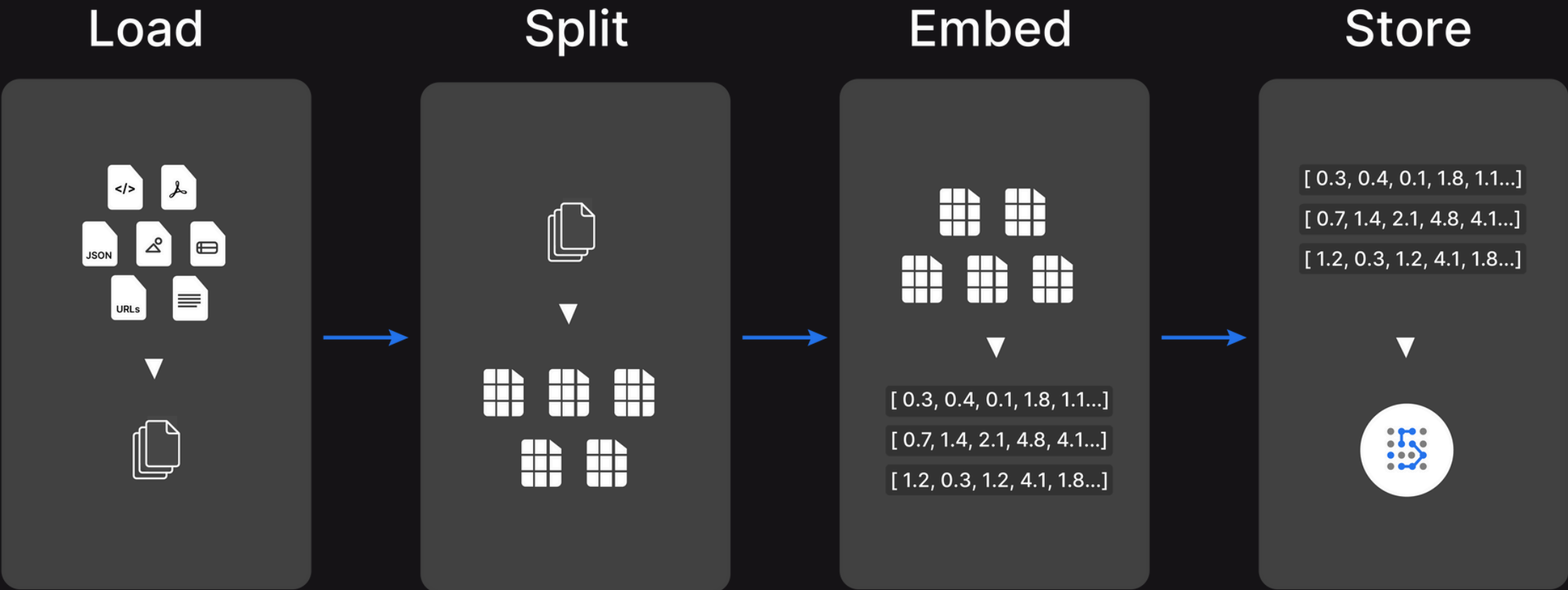
- Custom documents are processed and chunked
- These chunks are converted to embeddings with a LLM (transformer)
- Chunks and embeddings are stored in a Vector Database index (along with metadata)

Step 2

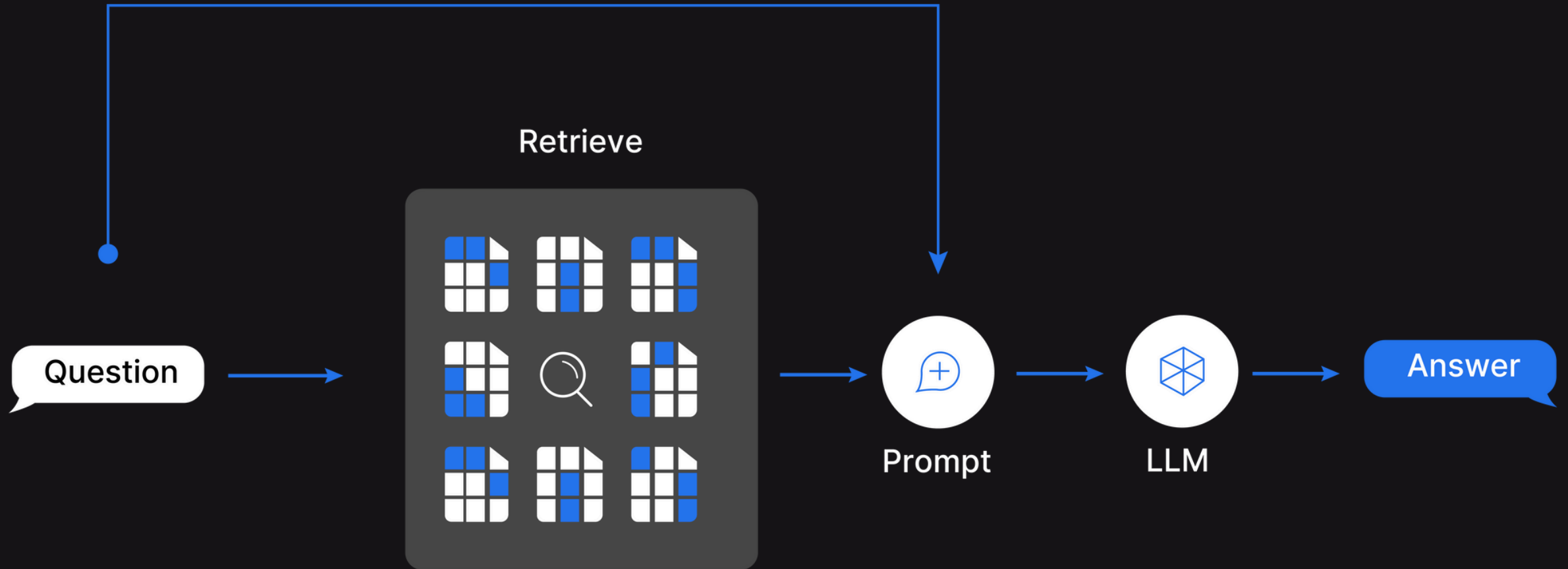
Retrieval & Response Generation

- Based on user query the system **retrieves** relevant document chunks
- Passes these chunks to the LLM along with the query to **augment** its knowledge
- The LLM **generates** a human-like response to the user query based on this information

RAG Workflow - Step 1 - Data processing and Indexing



RAG Workflow - Step 2 - Retrieval and Response Generation



Thank You
