Analytics Vidhya

What is a RAG System?

Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

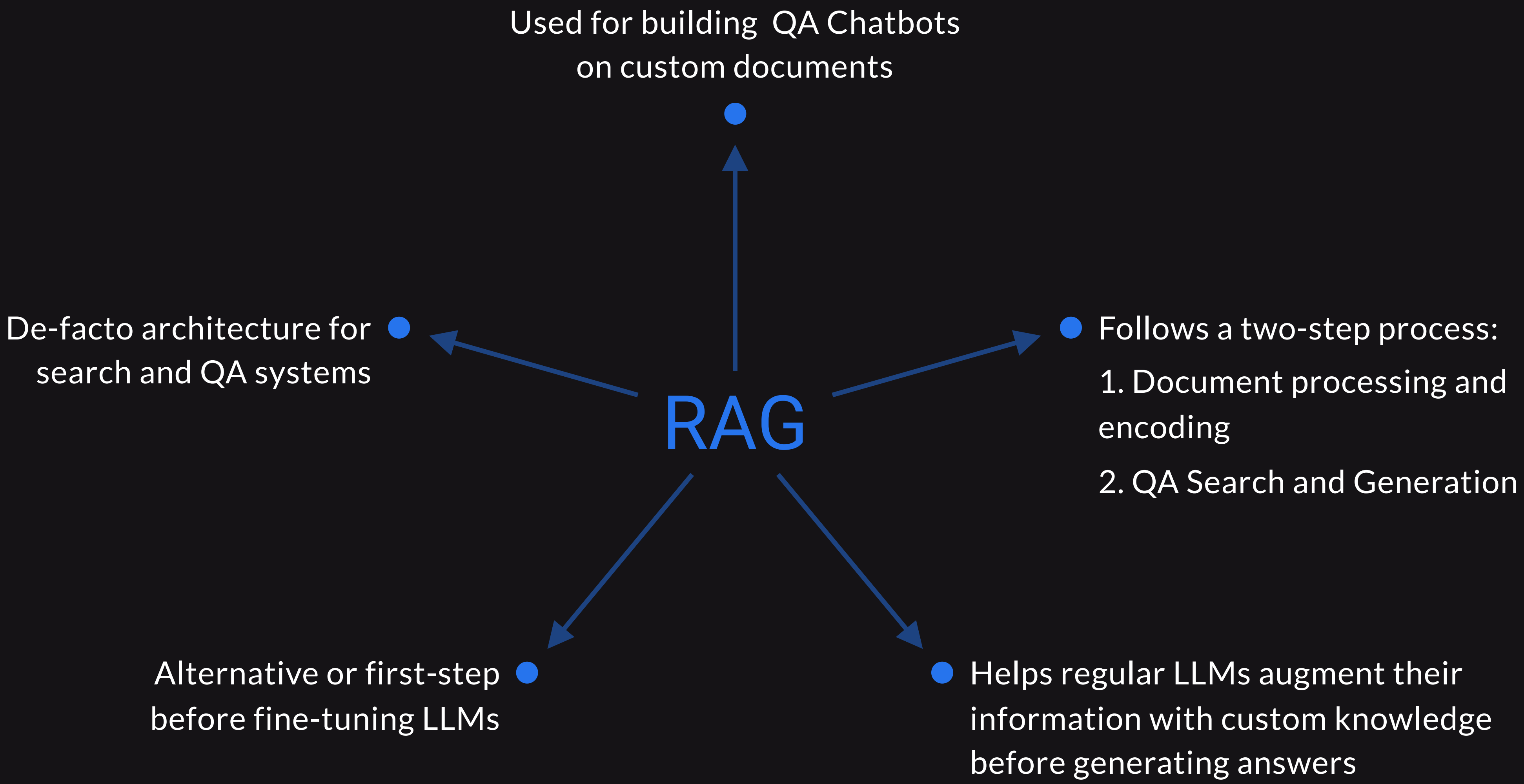
Published Author

Outline

- What is a RAG System
- RAG Workflow
- RAG System Architecture

What is a RAG System

RAG stands for Retrieval Augmented Generation



RAG

Used for building QA Chatbots
on custom documents

De-facto architecture for
search and QA systems

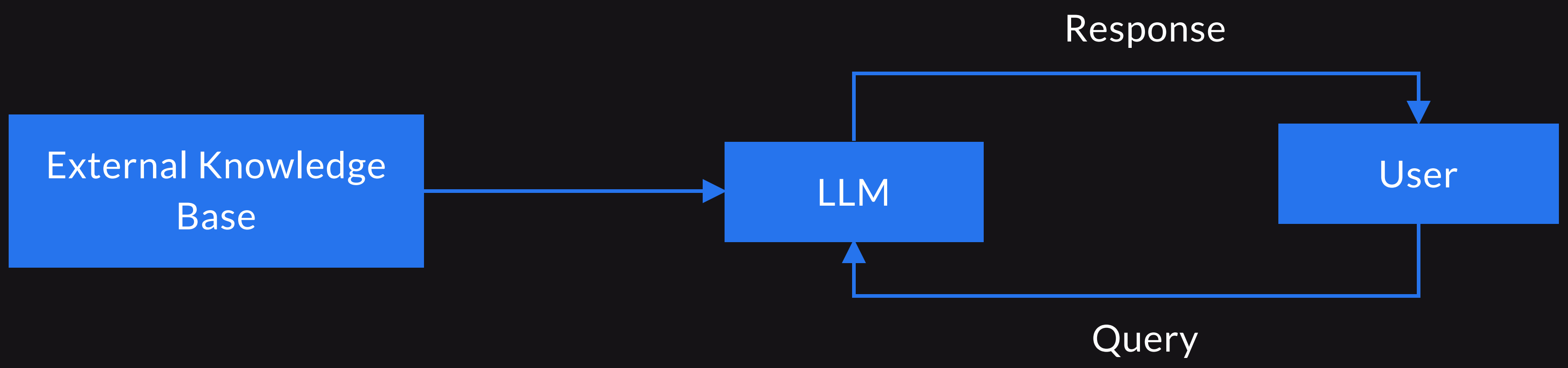
Follows a two-step process:

1. Document processing and encoding
2. QA Search and Generation

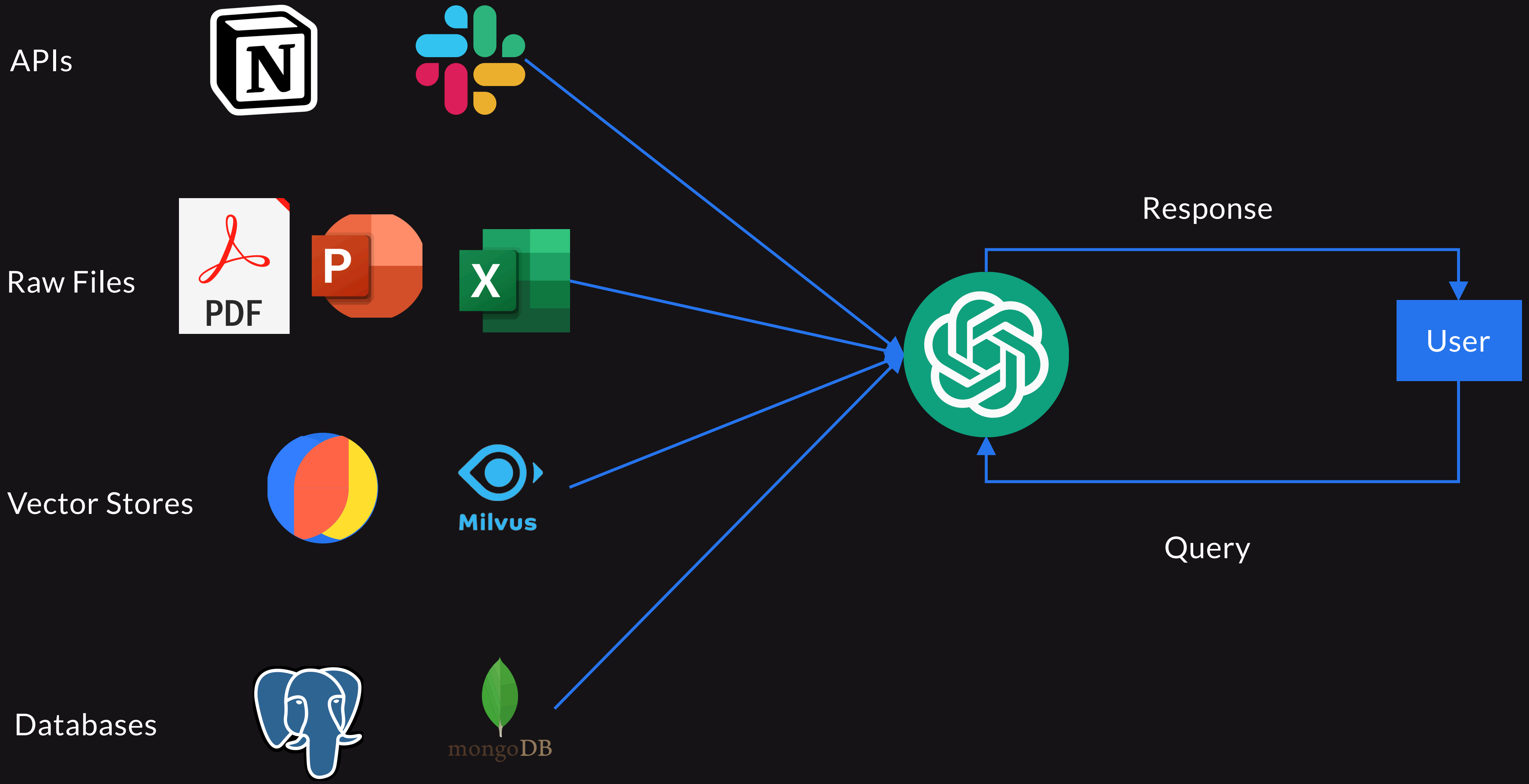
Alternative or first-step
before fine-tuning LLMs

Helps regular LLMs augment their
information with custom knowledge
before generating answers

What is a RAG System



What is a RAG System



Analytics Vidhya

What is a RAG System

Step 1

Data Processing & Indexing

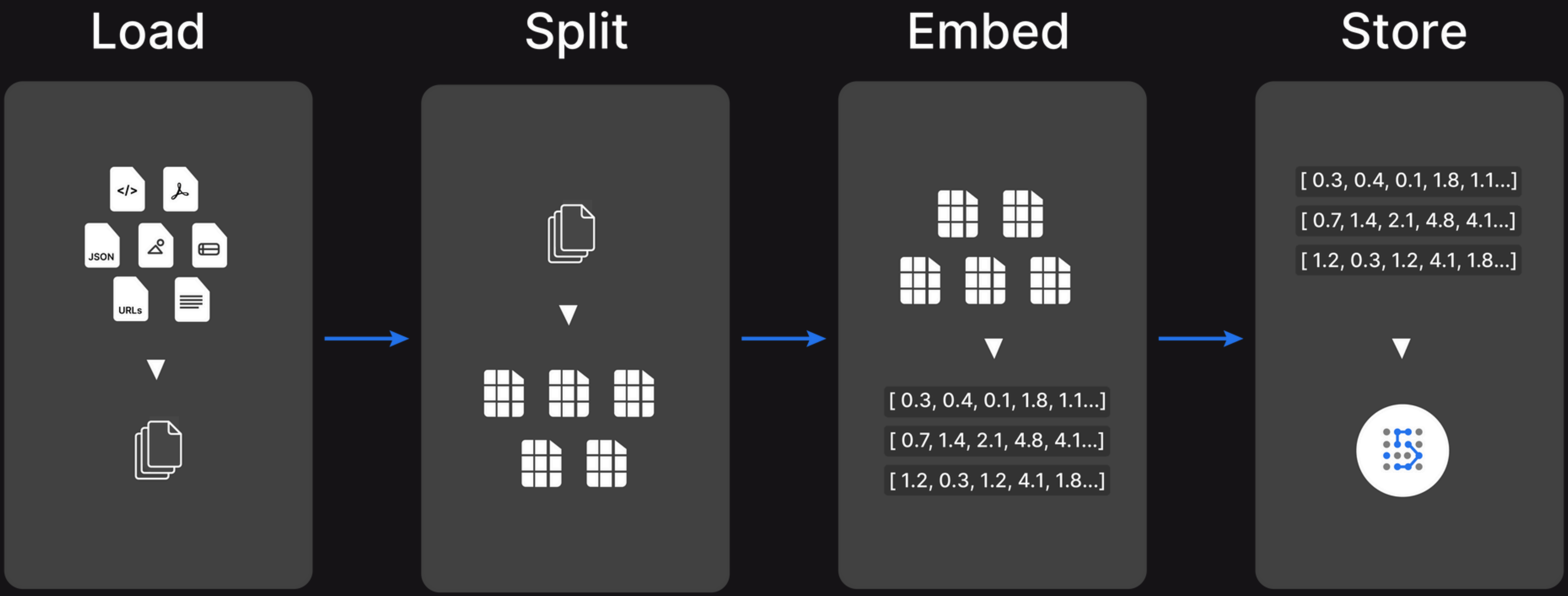
- Custom documents are processed and chunked
- These chunks are converted to embeddings with a LLM (transformer)
- Chunks and embeddings are stored in a Vector Database index (along with metadata)

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



Load

Split

Embed

[0.3, 0.4, 0.1, 1.8, 1.1...]

[0.7, 1.4, 2.1, 4.8, 4.1...]

[1.2, 0.3, 1.2, 4.1, 1.8...]

Store

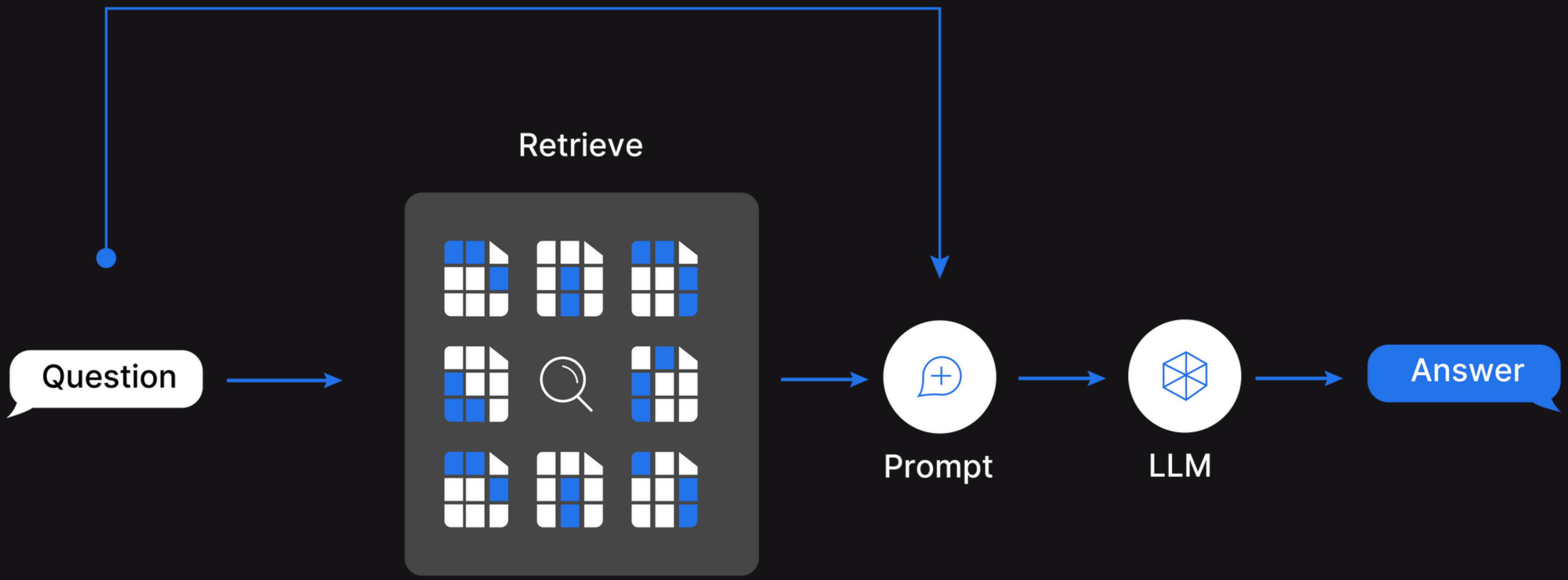
[0.3, 0.4, 0.1, 1.8, 1.1...]

[0.7, 1.4, 2.1, 4.8, 4.1...]

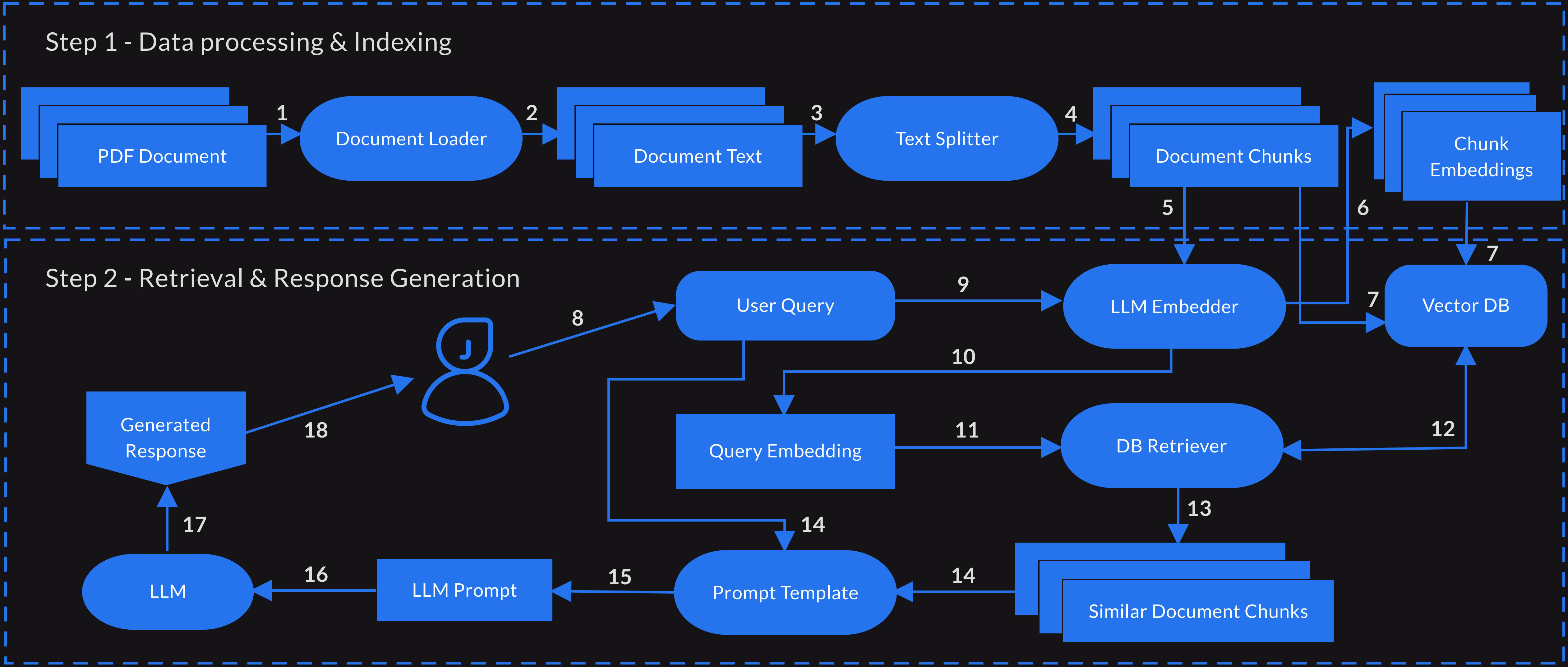
[1.2, 0.3, 1.2, 4.1, 1.8...]

Analytics Vidhya

RAG Workflow - Step 2 - Retrieval and Response Generation



RAG System Architecture



Analytics Vidhya

Thank You