

# Overview of RAG Framework

## Instructors

Prashant Sahu

Manager - Data Science, Analytics Vidhya

Ravi Theja

Developer Advocate Engineer, LlamalIndex



# Objective

To provide the right answer for the user query from the external knowledge base

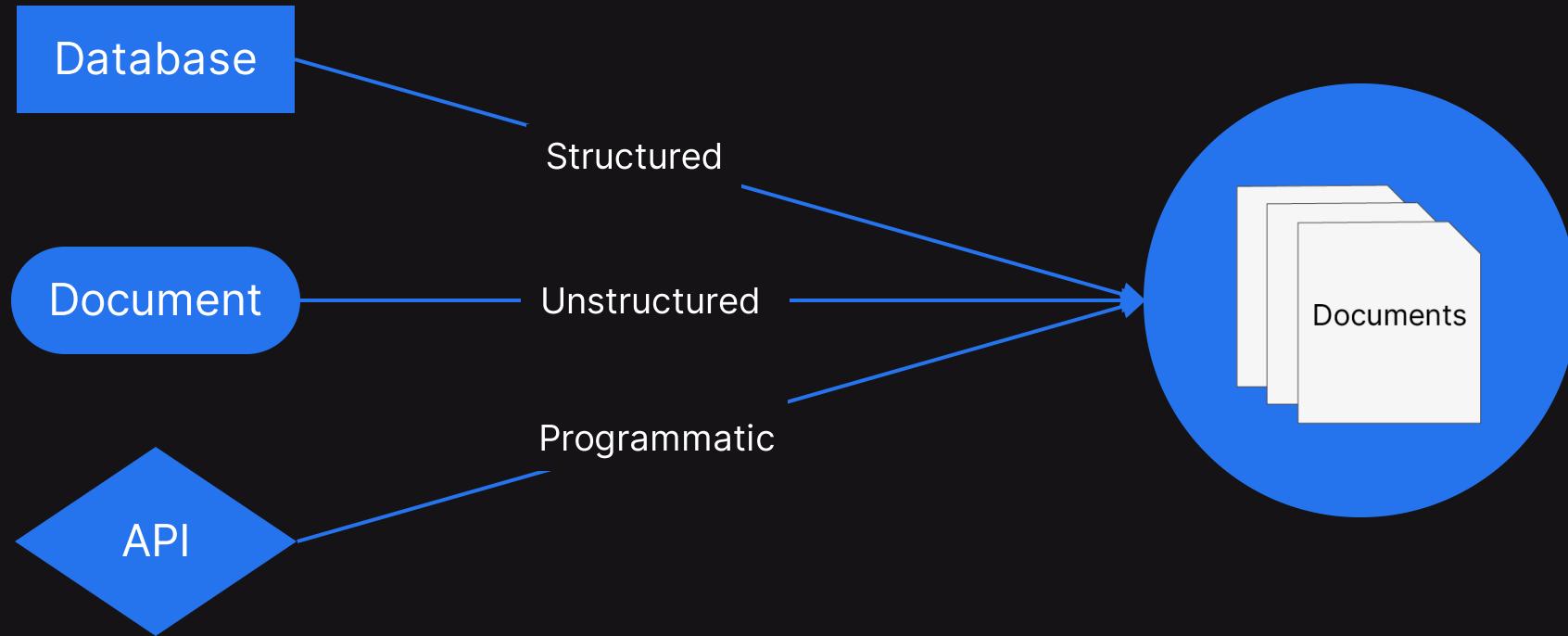
# Steps Involved in RAG



# Steps Involved in RAG



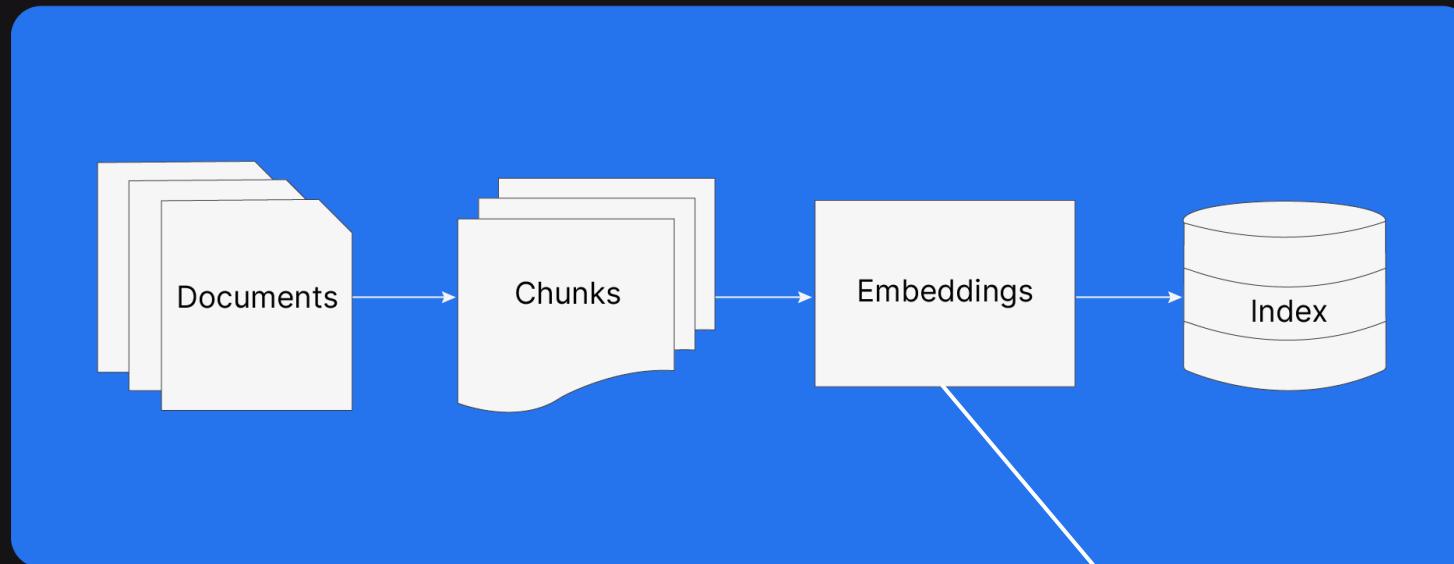
# 1. Data Ingestion



# Steps Involved in RAG

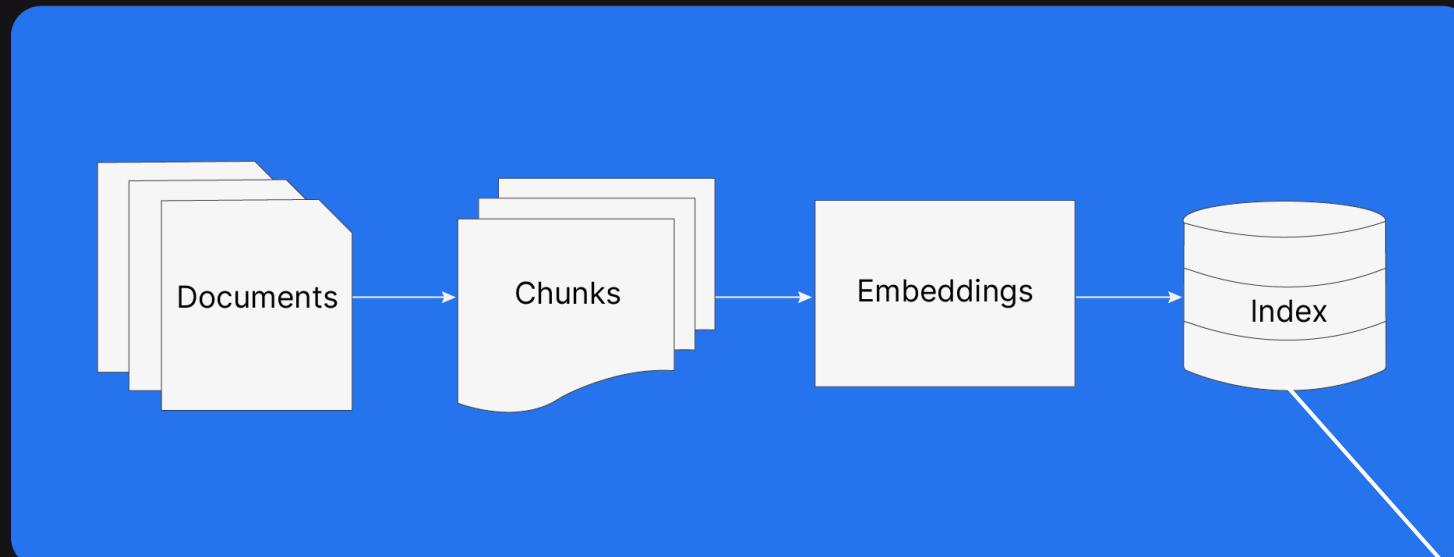


## 2. Indexing & Storing



Numerical vector representations of textual chunks  
that capture the meaning and context of the text

## 2. Indexing & Storing

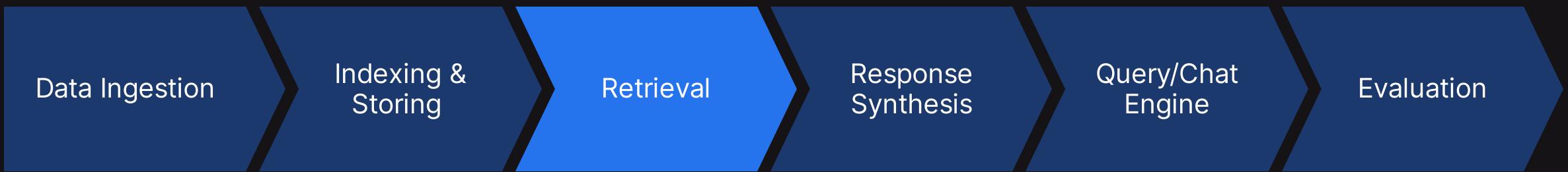


Chunks along with the embeddings and metadata are indexed

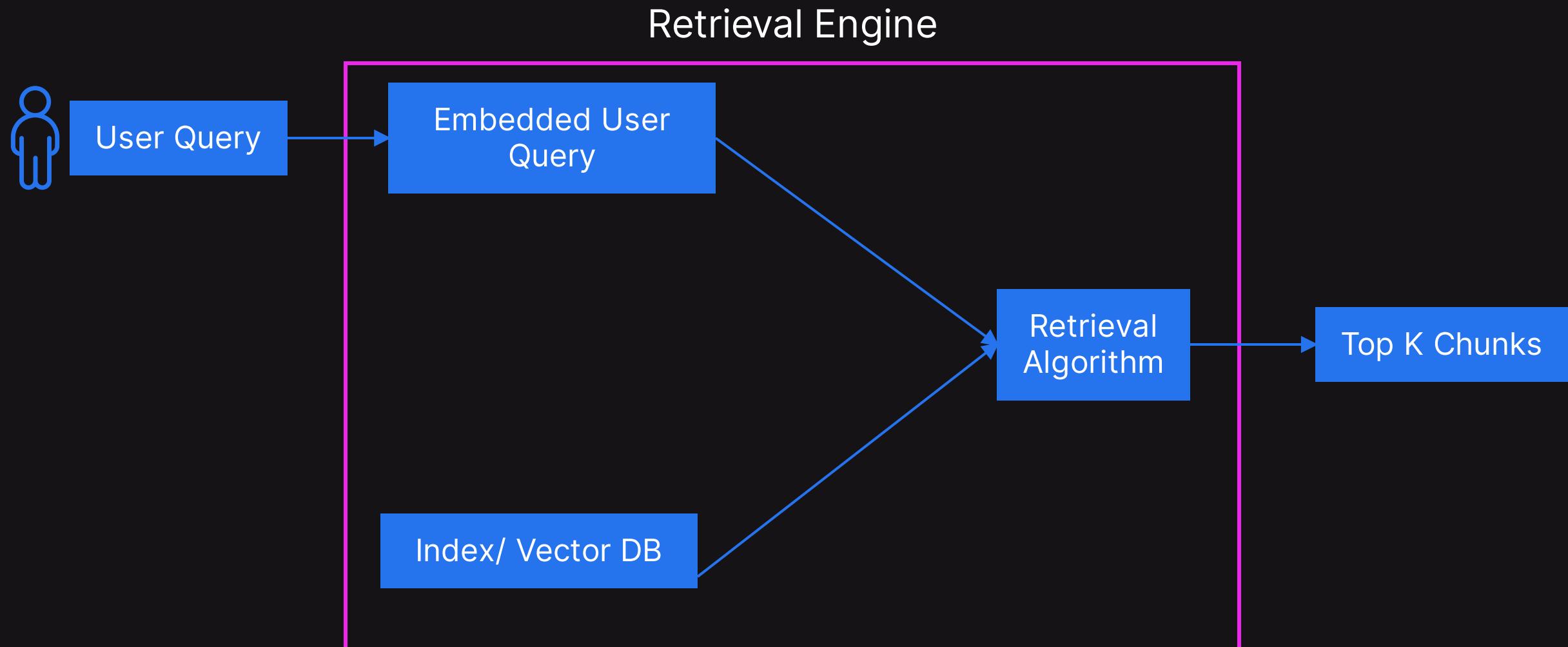
# Why Indexing?

- **Quick Retrieval:** Speeding up the process of finding relevant information
- **Enhanced Accuracy:** Improves the relevance and quality of information retrieved.
- **Scalability:** Allows the system to efficiently handle large data volumes.

# Steps Involved in RAG



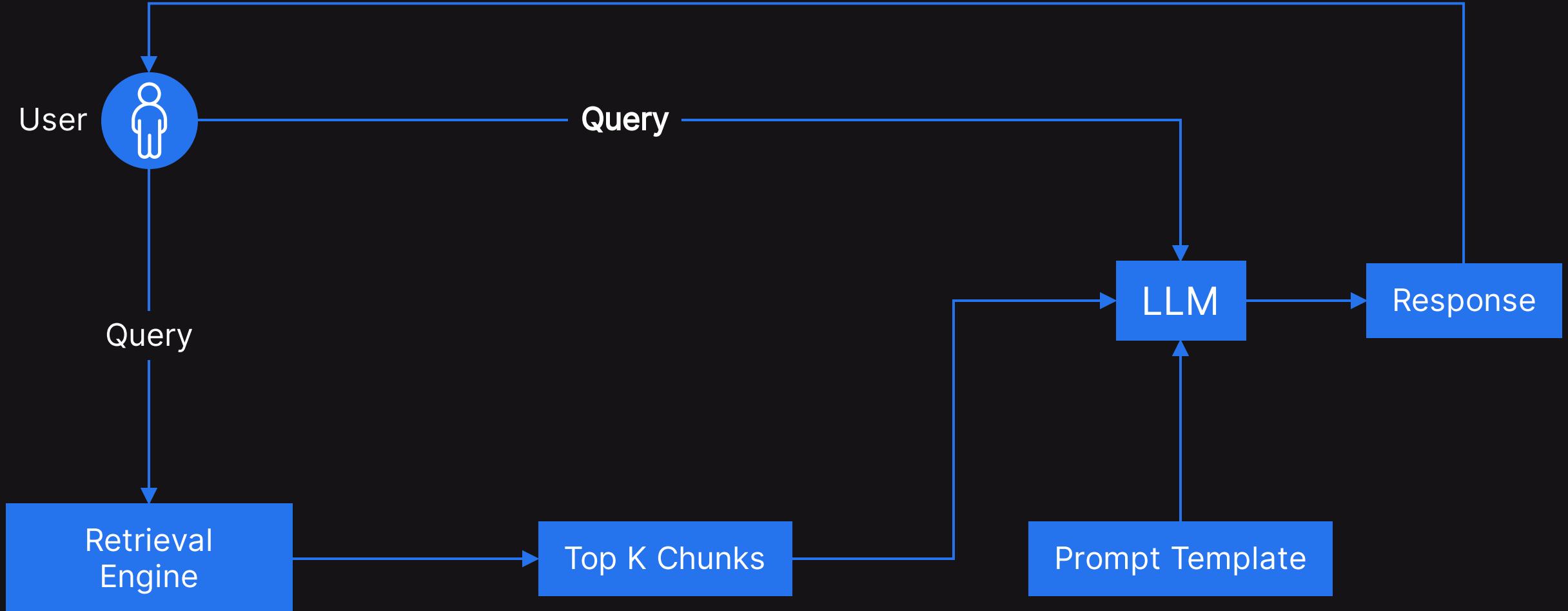
### 3. Retrieval



# Steps Involved in RAG



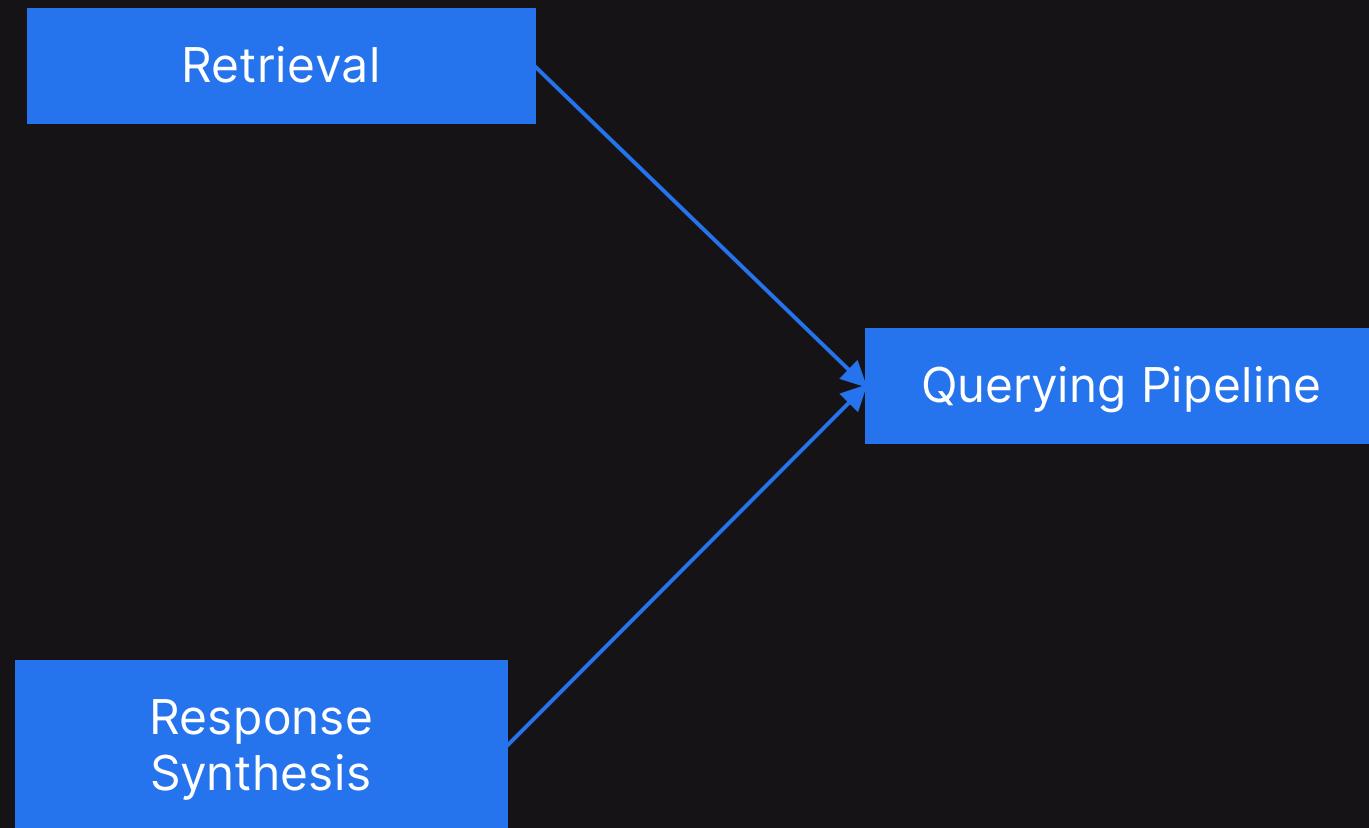
# 4. Response Synthesis



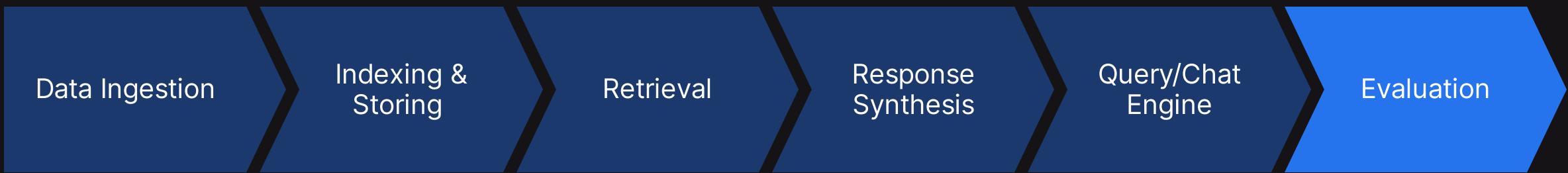
# Steps Involved in RAG



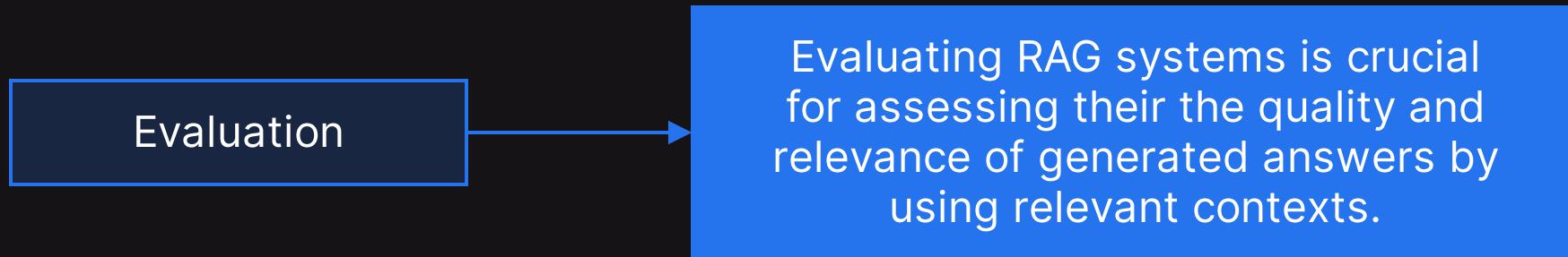
# 5. Query/Chat Engine



# Steps Involved in RAG



# 6. Evaluation



# RAG Framework



# Thank You

---