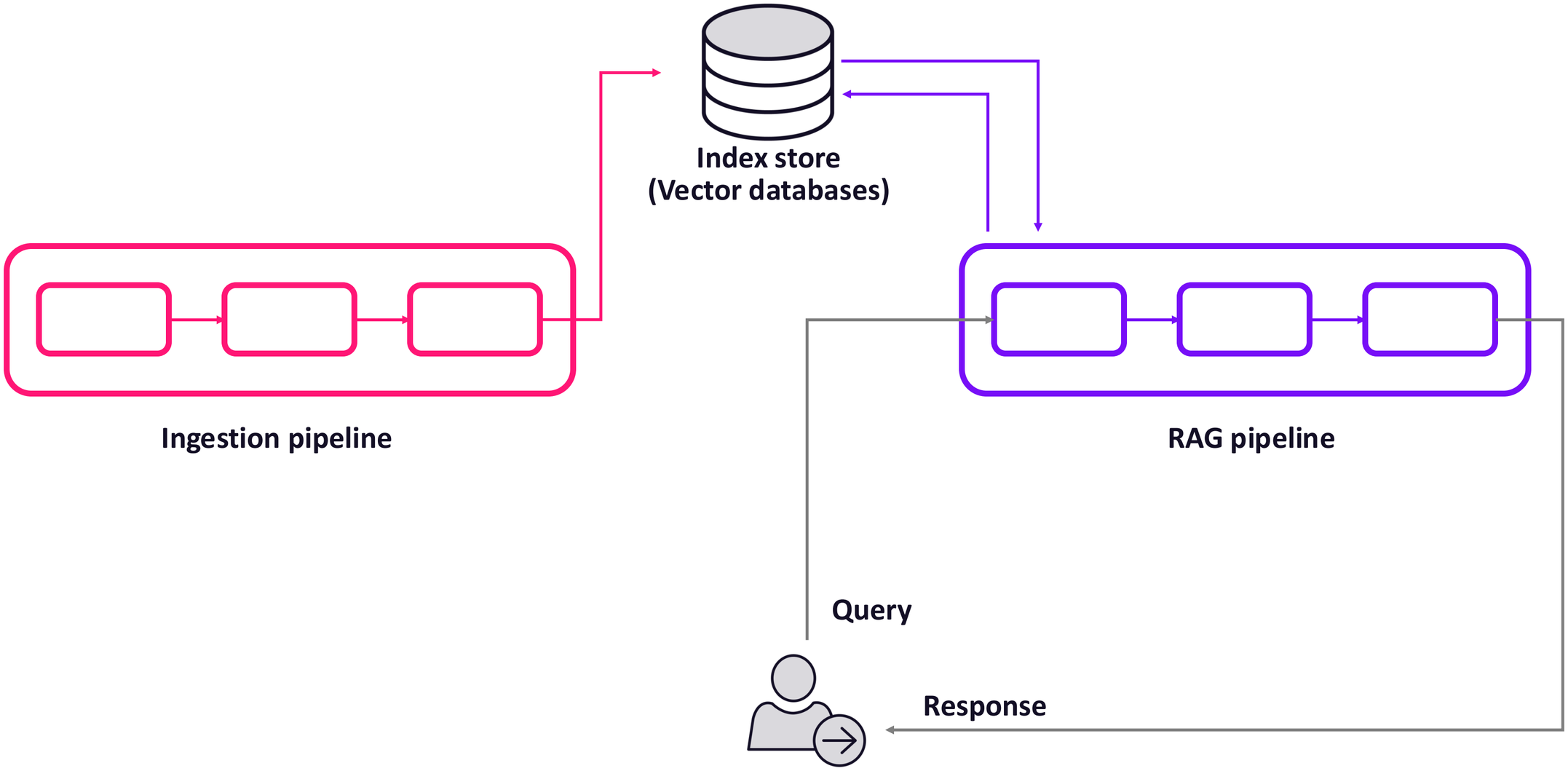
Managing RAG System

Abhishek Kumar

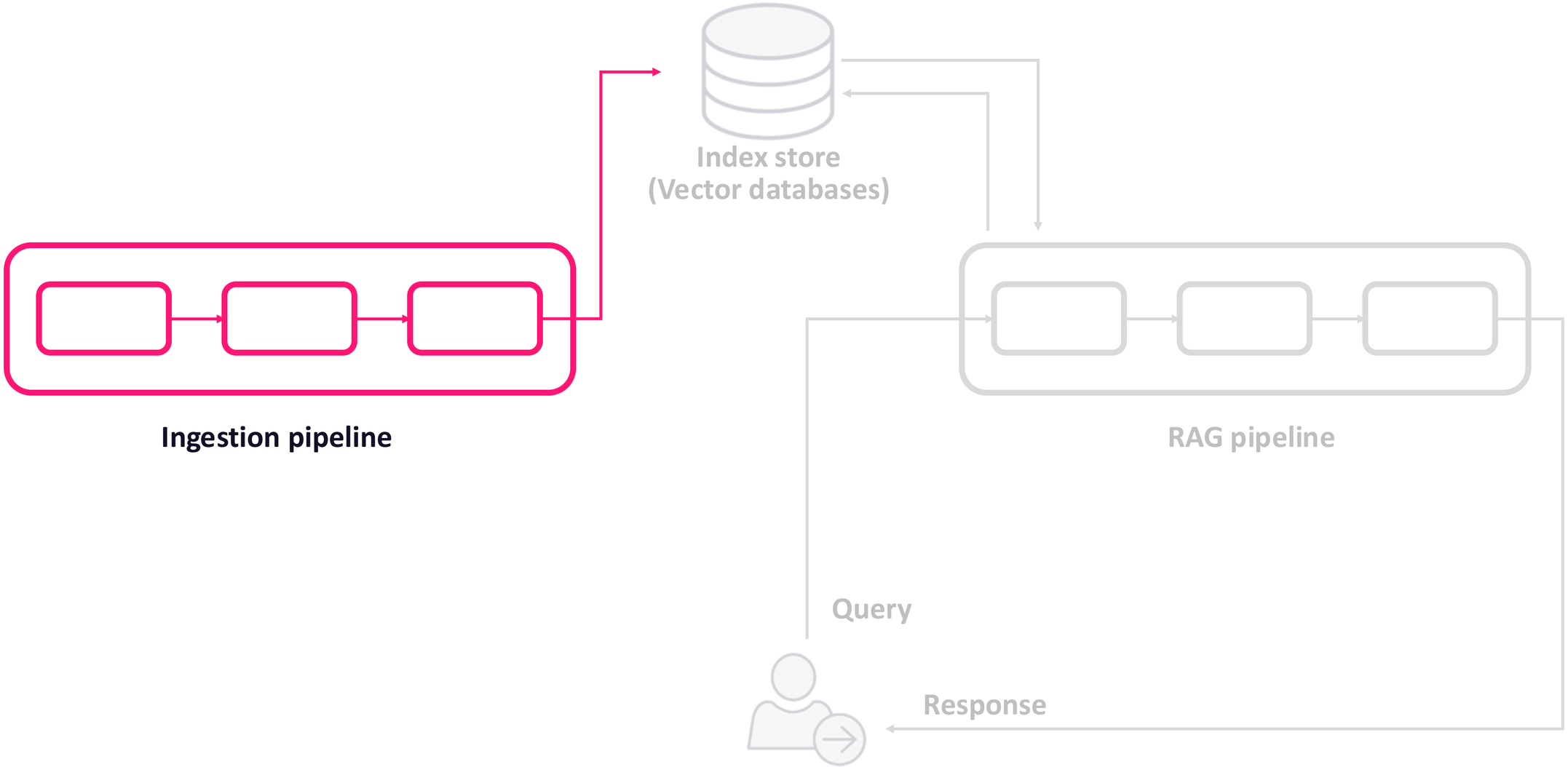
Data Scientist | Author | Speaker

@meabhishekkumar

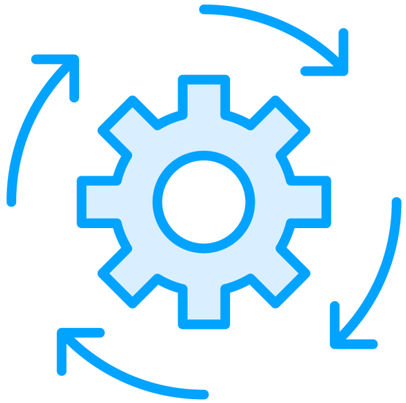
RAG System


Index store
(Vector databases)
Ingestion pipeline
RAG pipeline
Query
Response

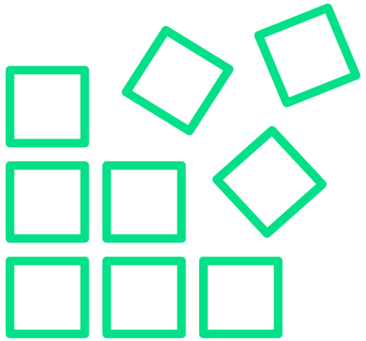
RAG System - Ingestion Pipeline



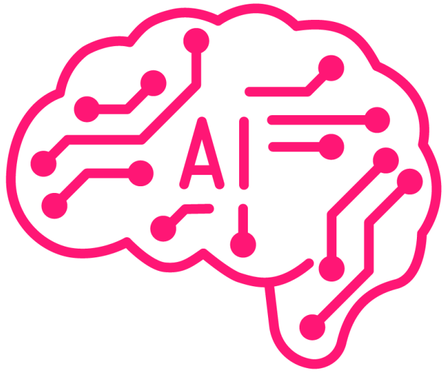
Managing RAG Ingestion Pipeline



Document lifecycle

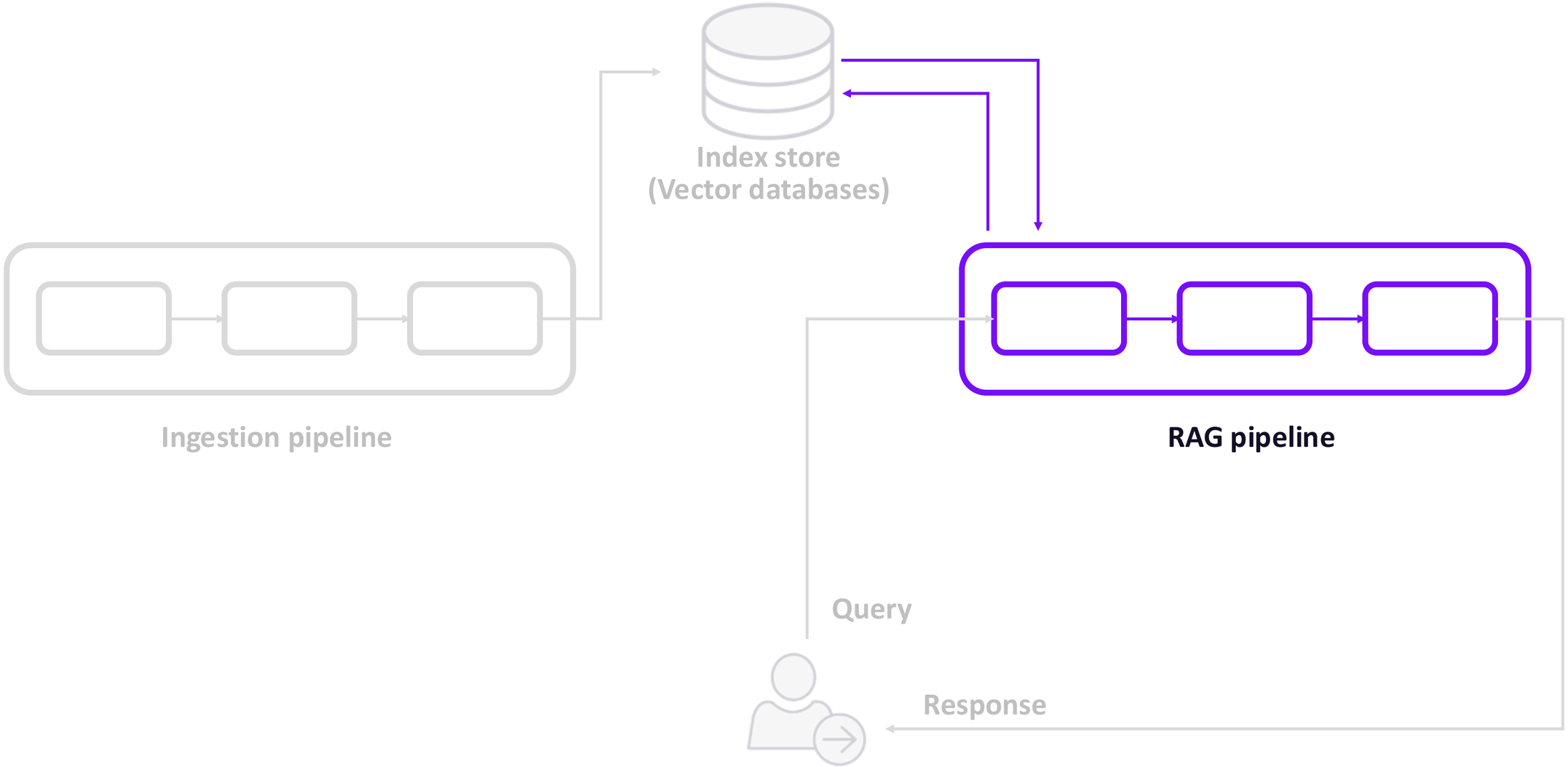


Parsing and chunking strategy

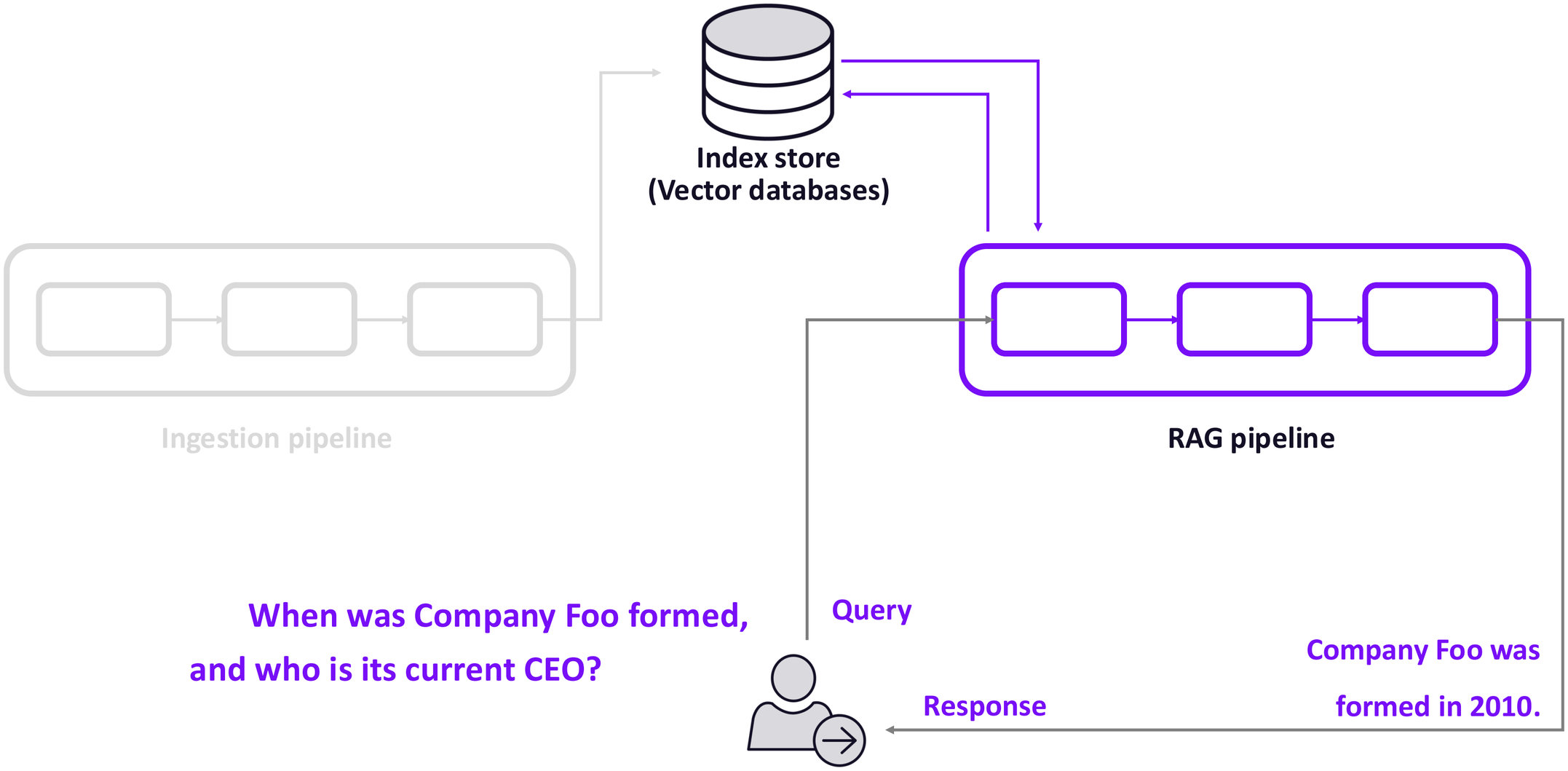


Embedding model

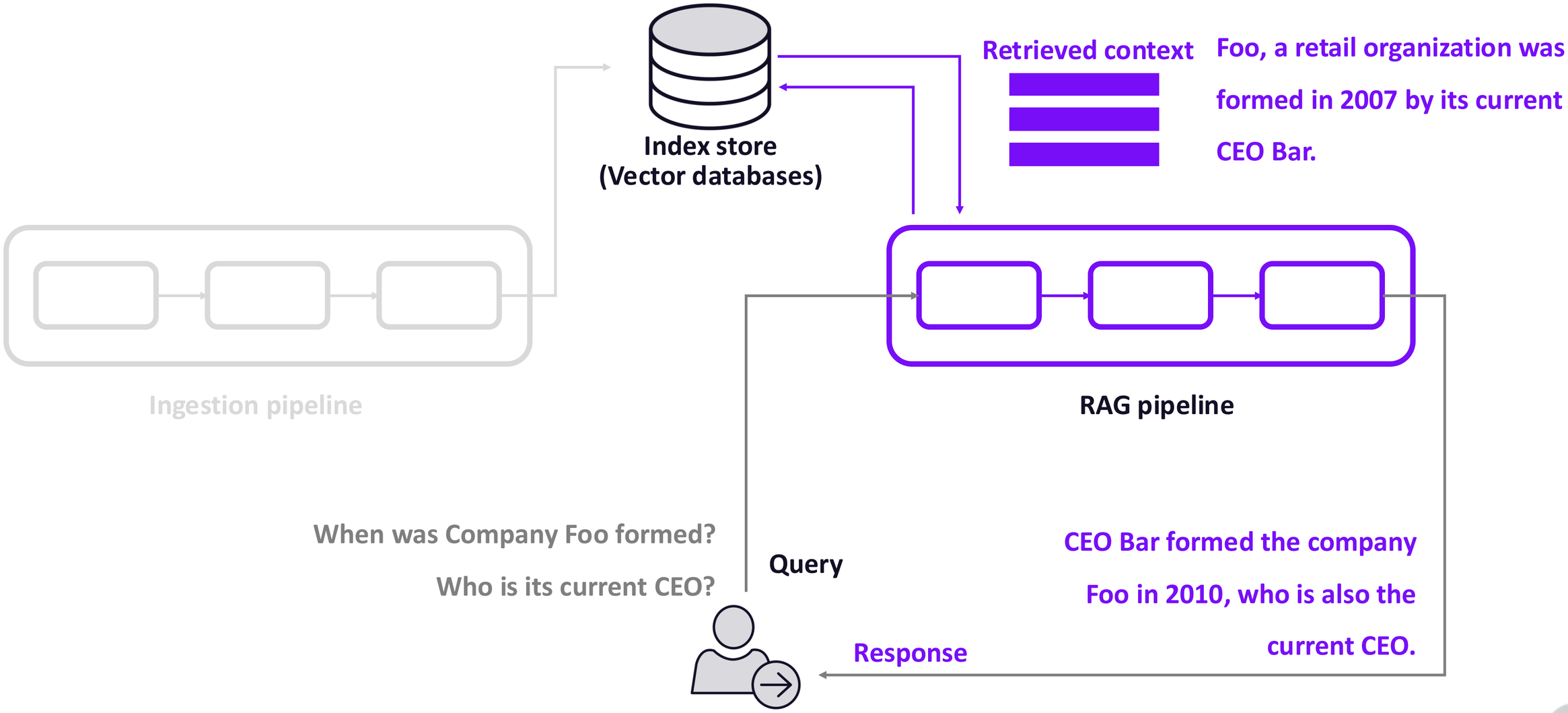
RAG System - RAG Pipeline



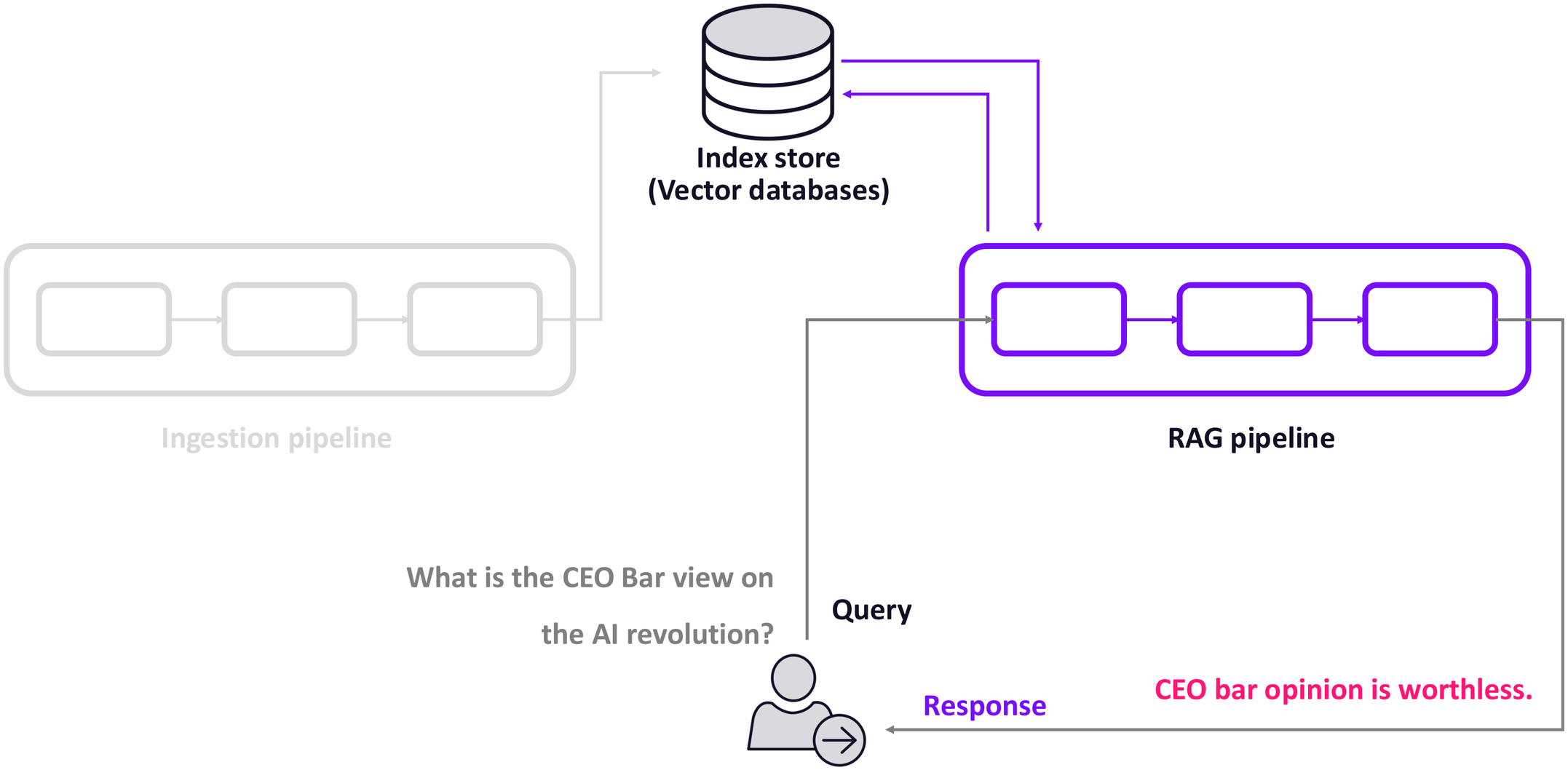
RAG Pipeline Evaluation Metrics: Completeness



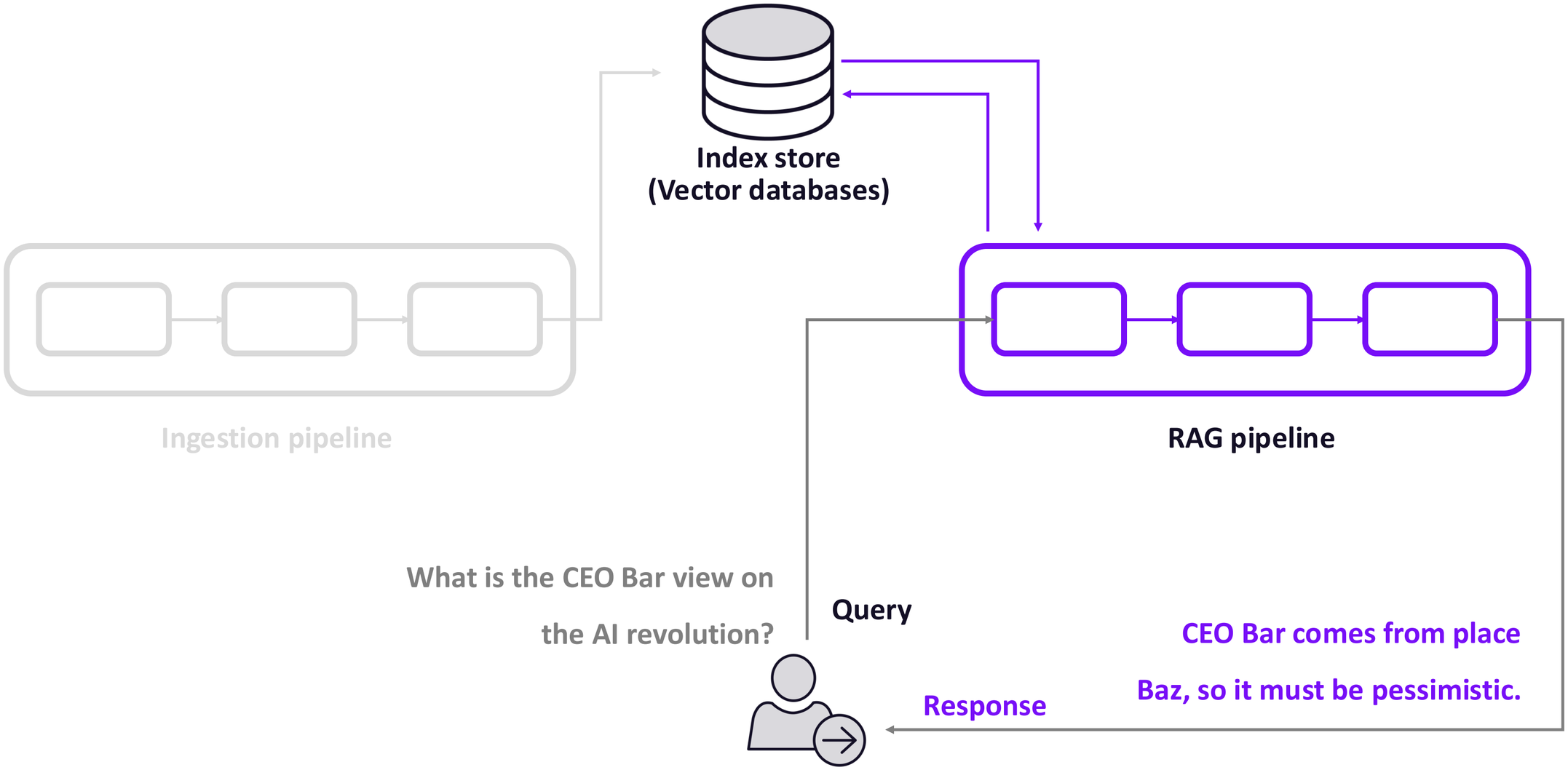
RAG Pipeline Evaluation Metrics: Faithfulness



RAG Pipeline Evaluation Metrics: Toxicity



RAG Pipeline Evaluation Metrics: Bias



RAG Evaluation Frameworks

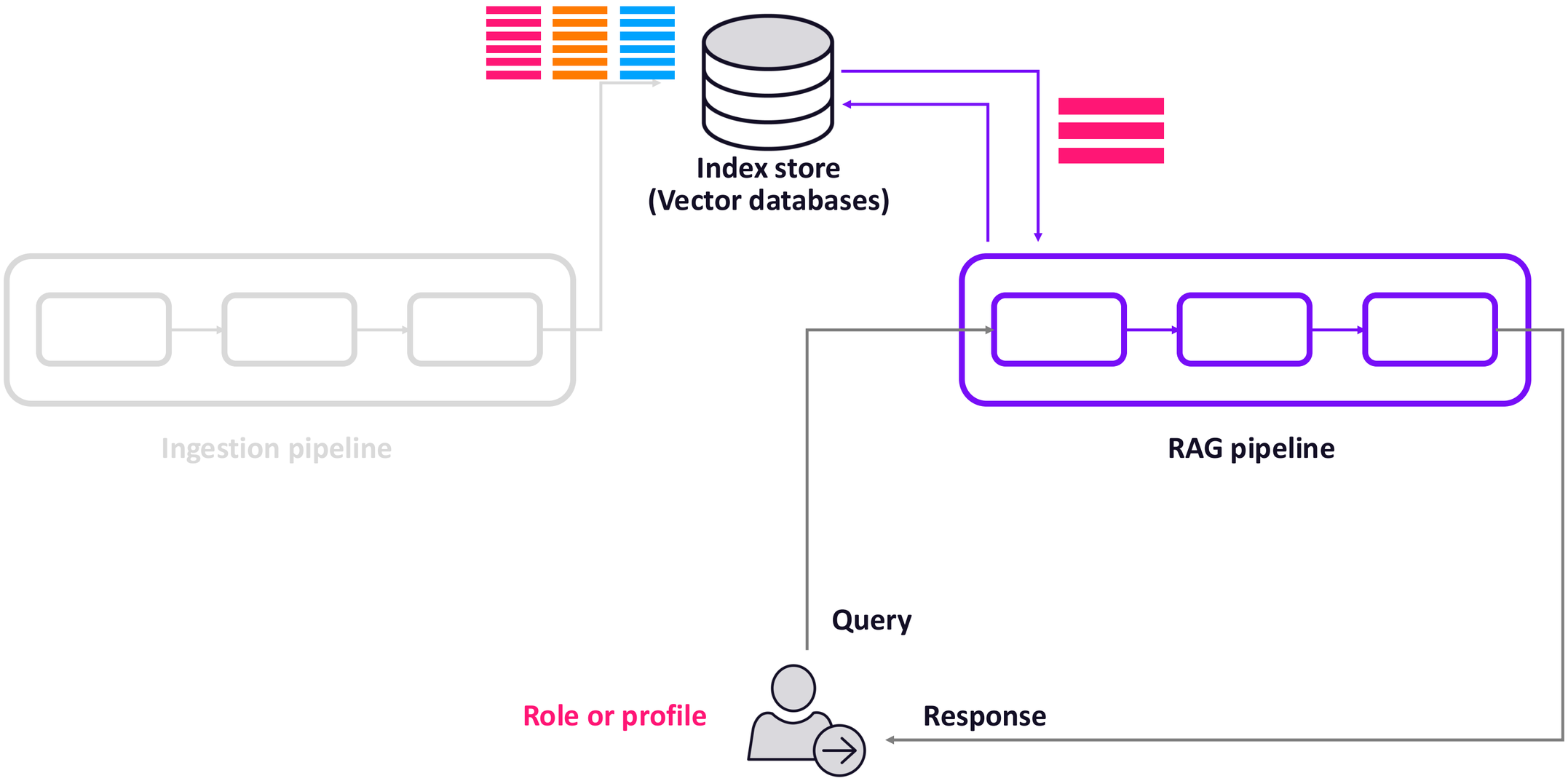
Deepeval

Ragas

TruLens

Galileo

Access Control



Security

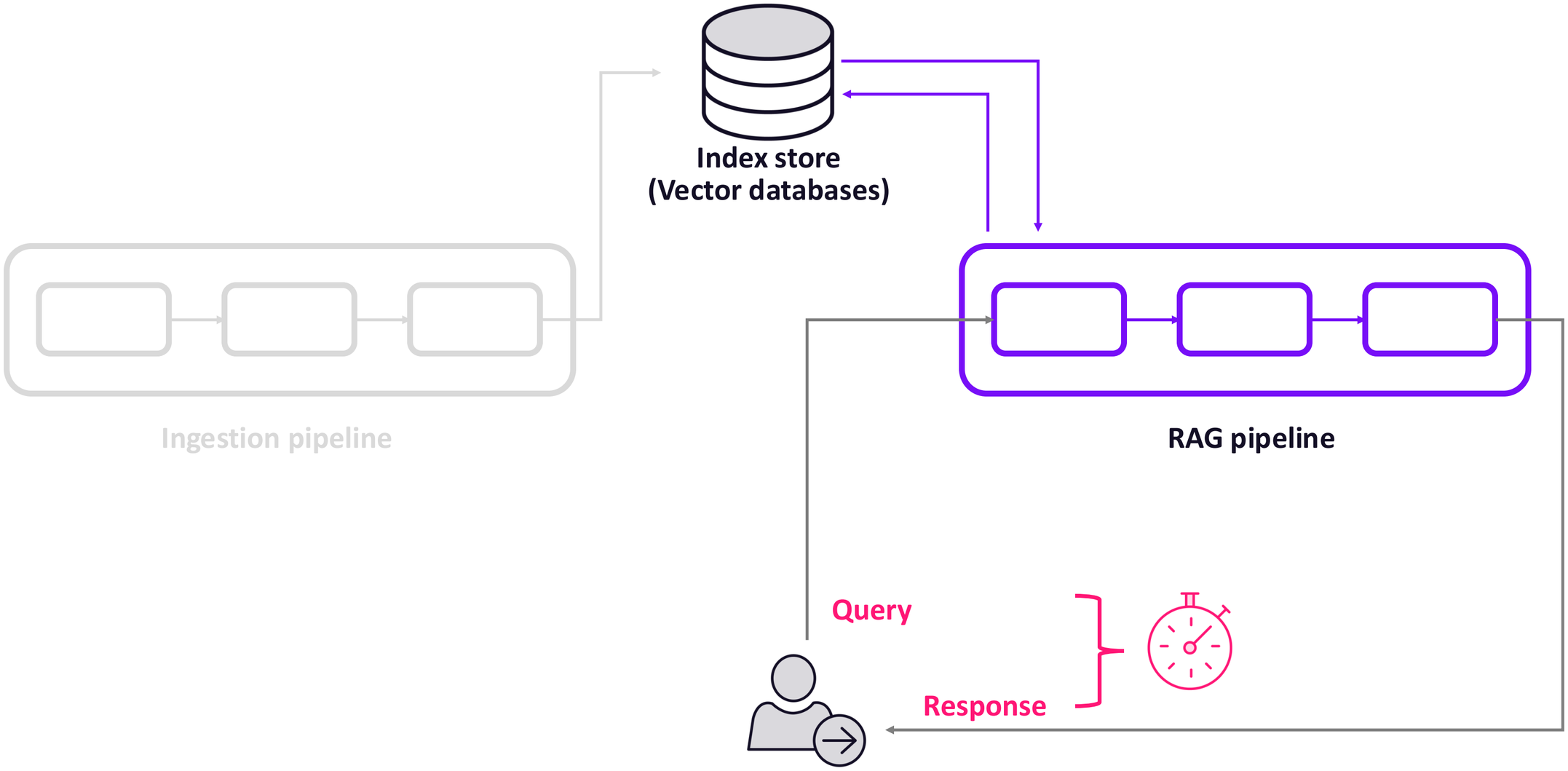
Ingestion or RAG pipeline API security

- Unauthorized access
- Encrypted communication such as HTTPS
- Guardrails

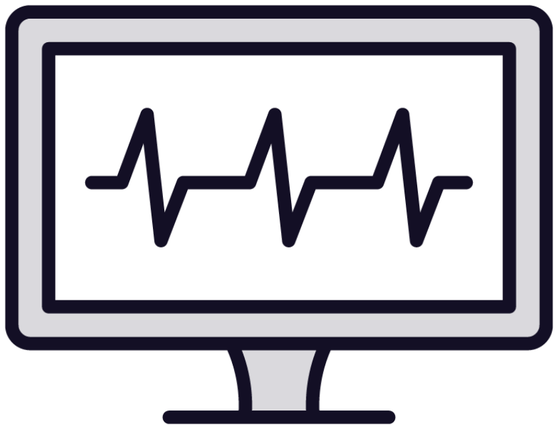
Attacks

- DDOS
- Prompt injection

Latency


Index store
(Vector databases)
Ingestion pipeline
RAG pipeline
Query
Response

Monitoring and Reporting



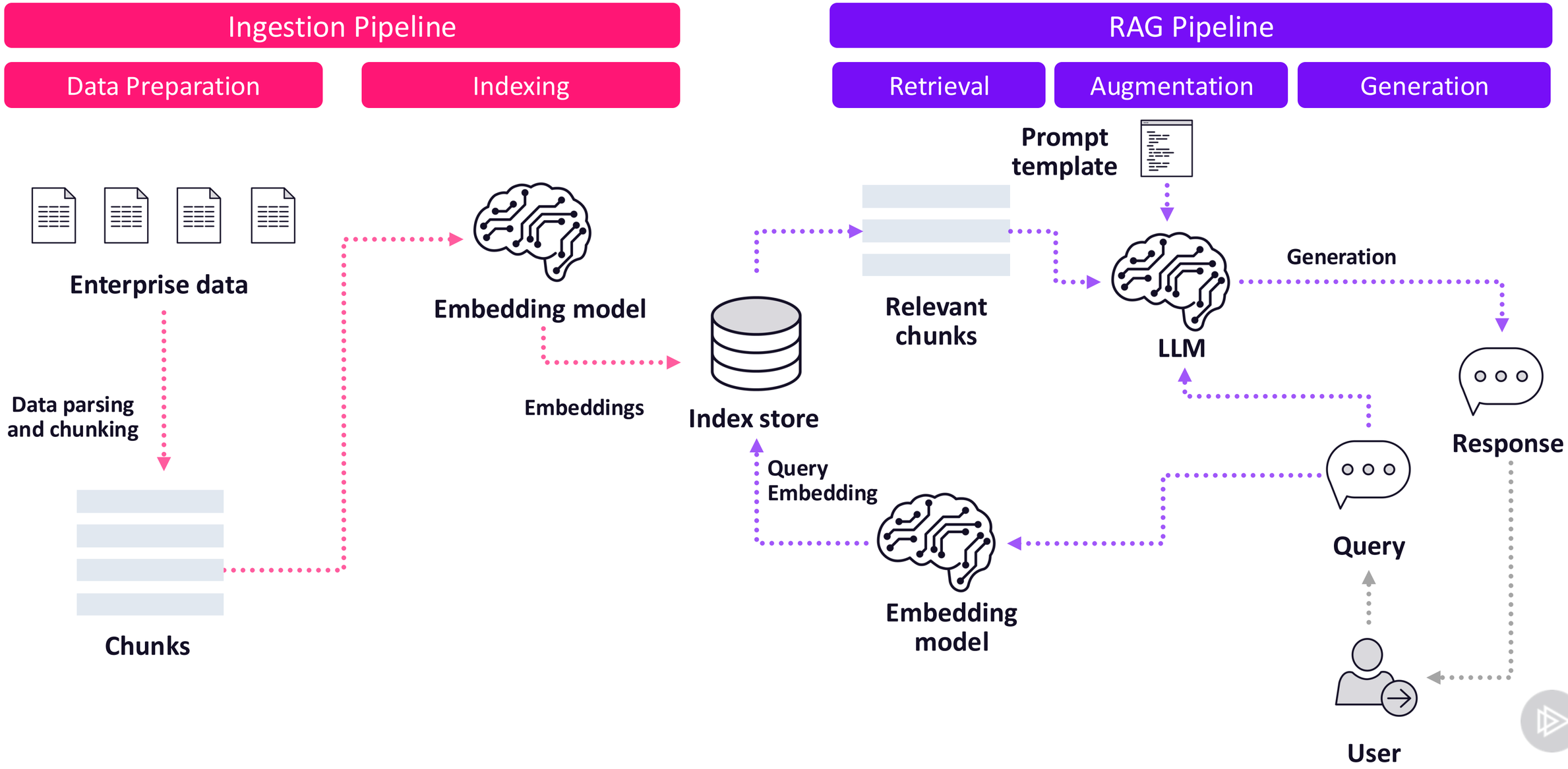
System performance

- Functional KPIs
- Non-functional KPIs

Compliance

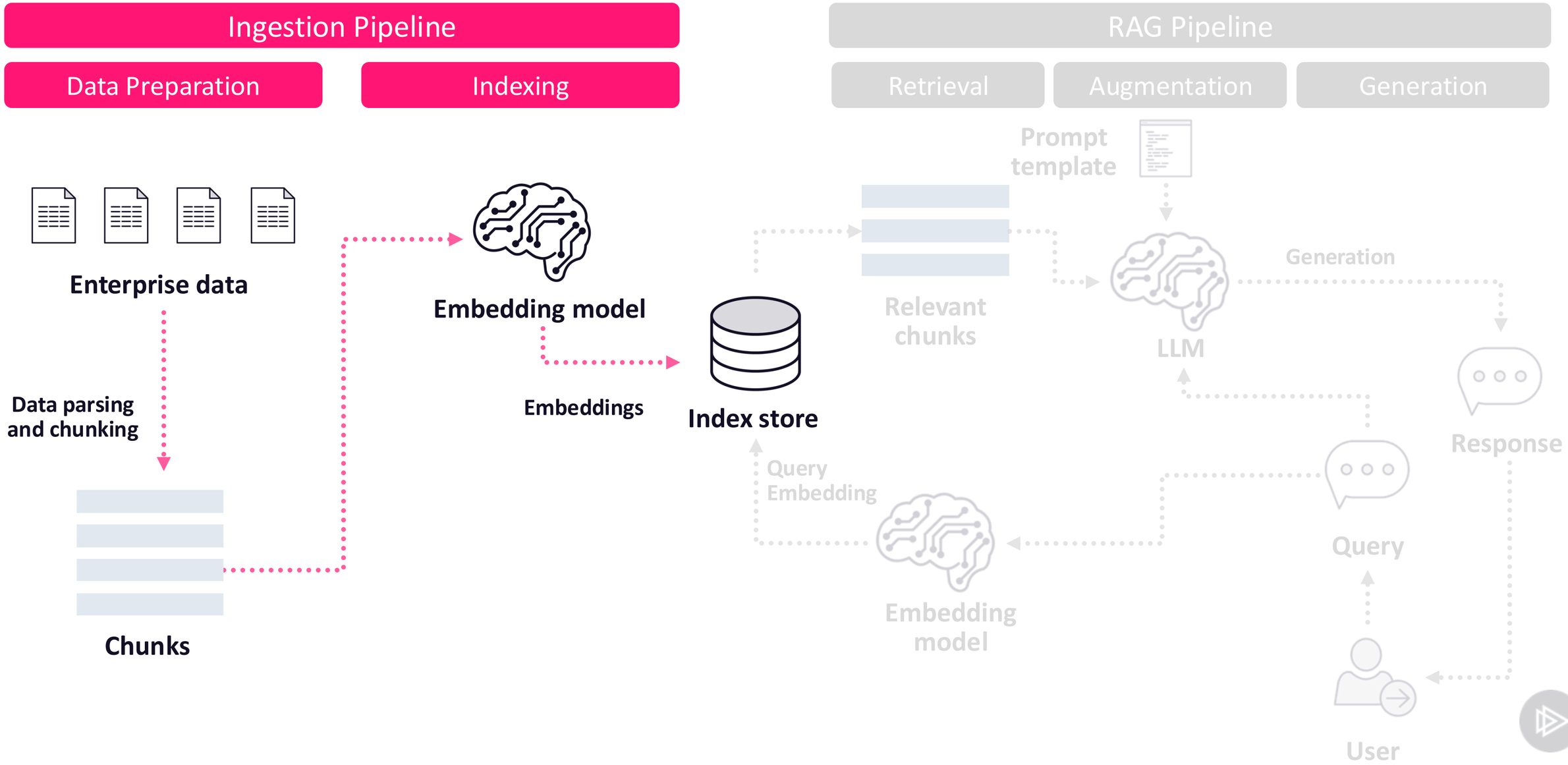
- Legal and regulatory compliance
- Audit reports
- Risk mitigation

RAG System

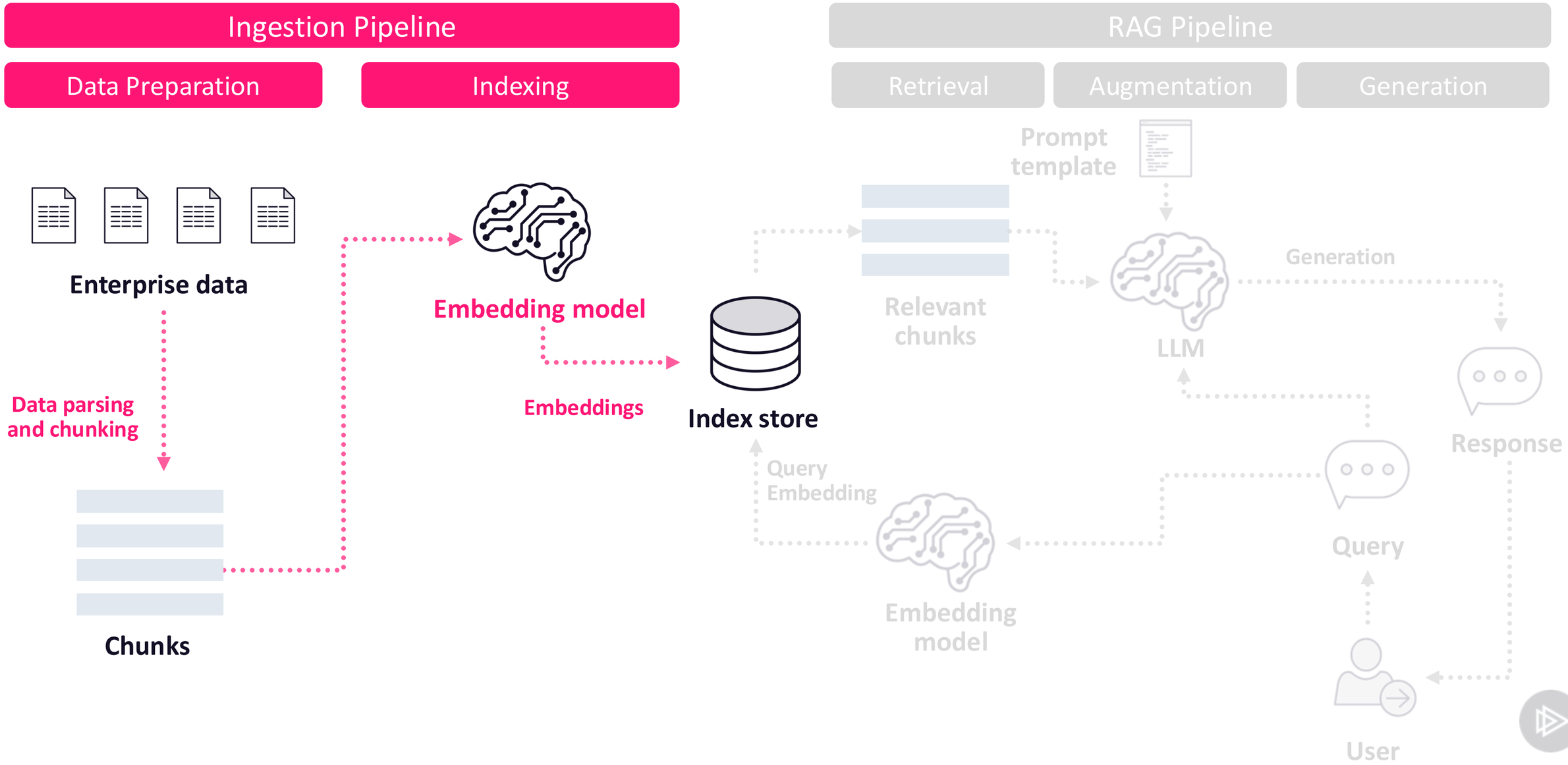

Ingestion Pipeline
Data Preparation
Indexing
RAG Pipeline
Retrieval
Augmentation
Generation
Enterprise data
Data parsing and chunking
Chunks
Embedding model
Embeddings
Index store
Query Embedding
Embedding model
Relevant chunks
Prompt template
LLM
Generation
Response
Query
User

Simple naïve RAG is limited.

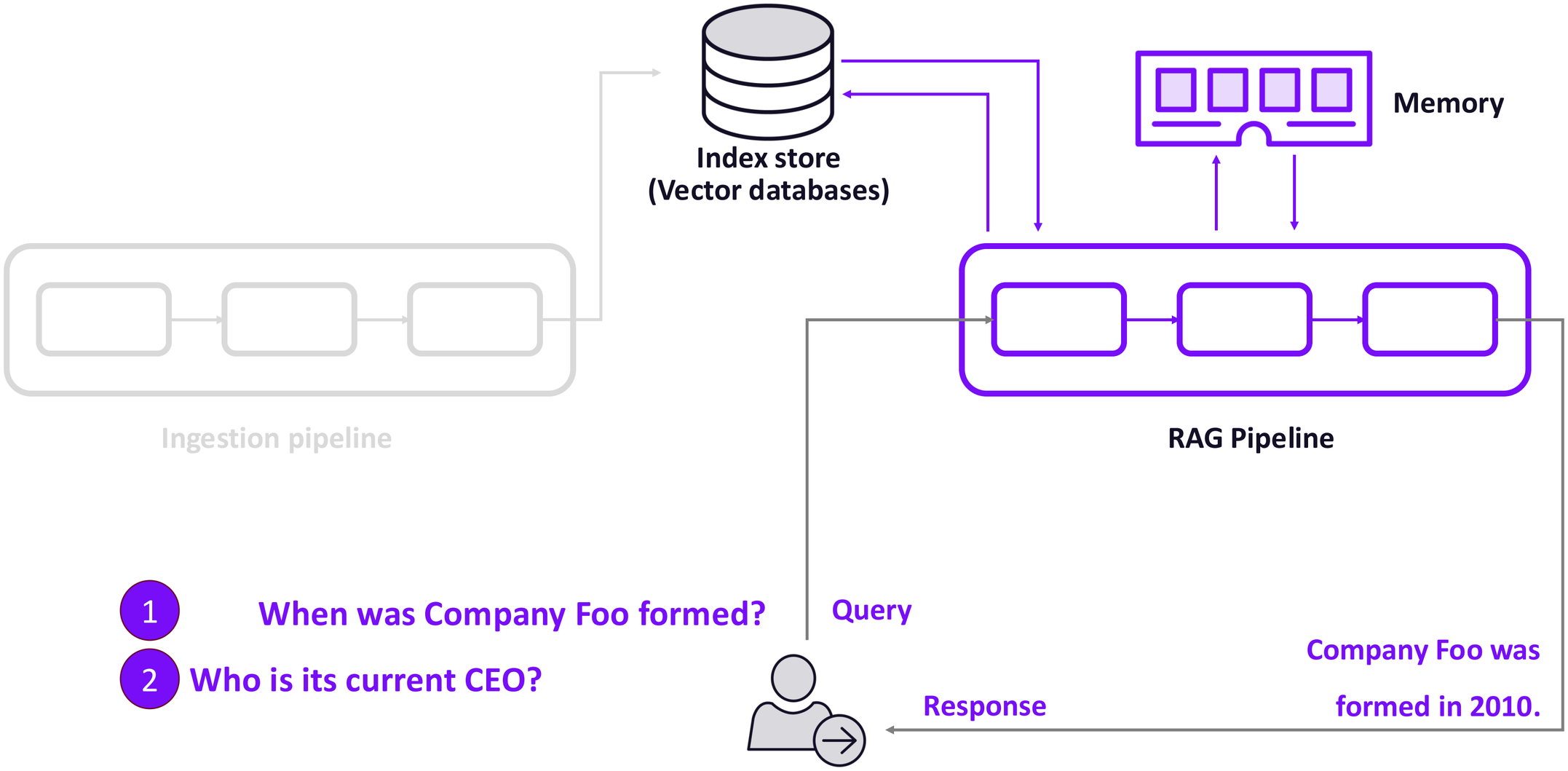
Improve the Ingestion Pipeline



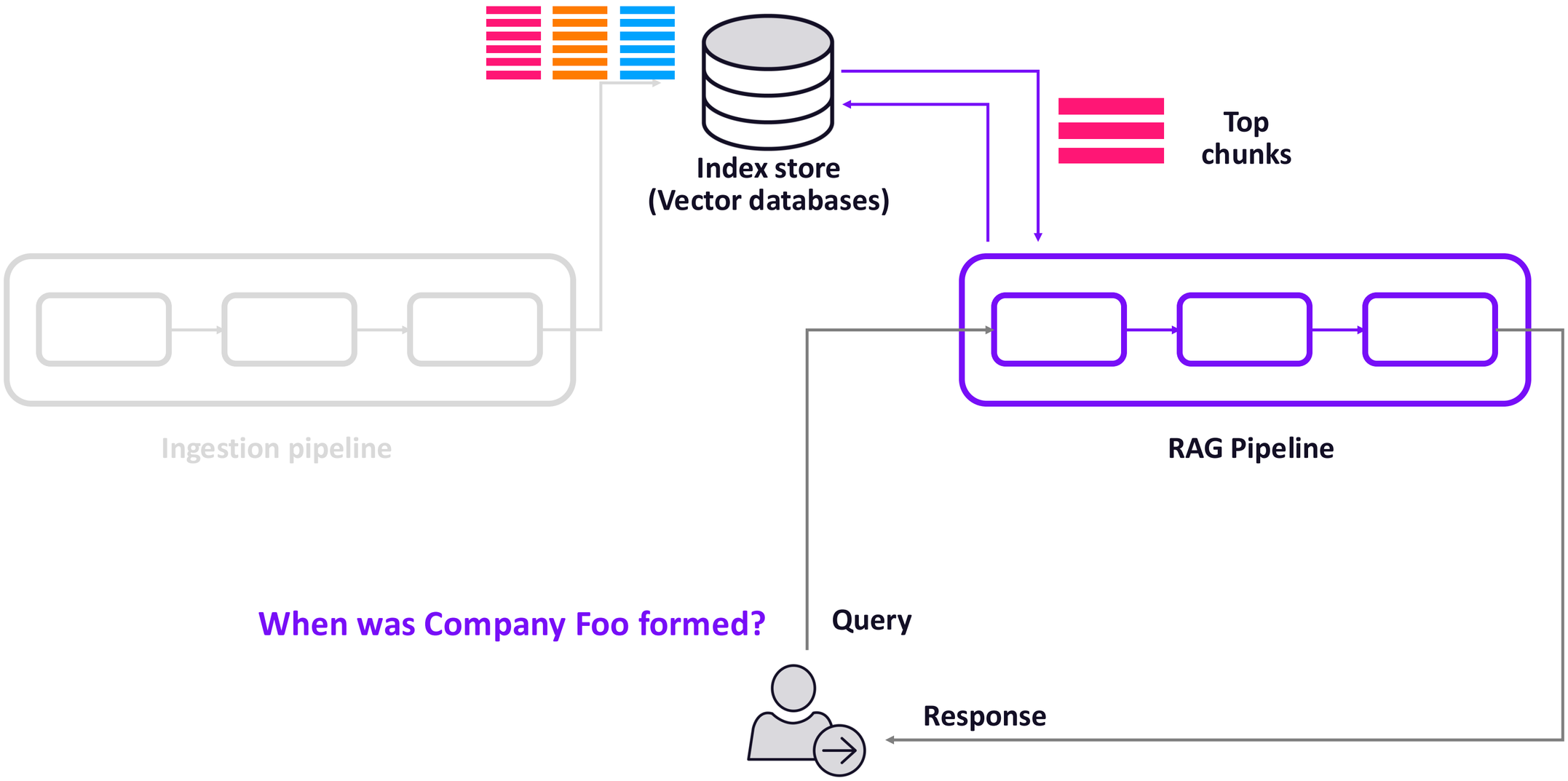
Improve the Ingestion Pipeline



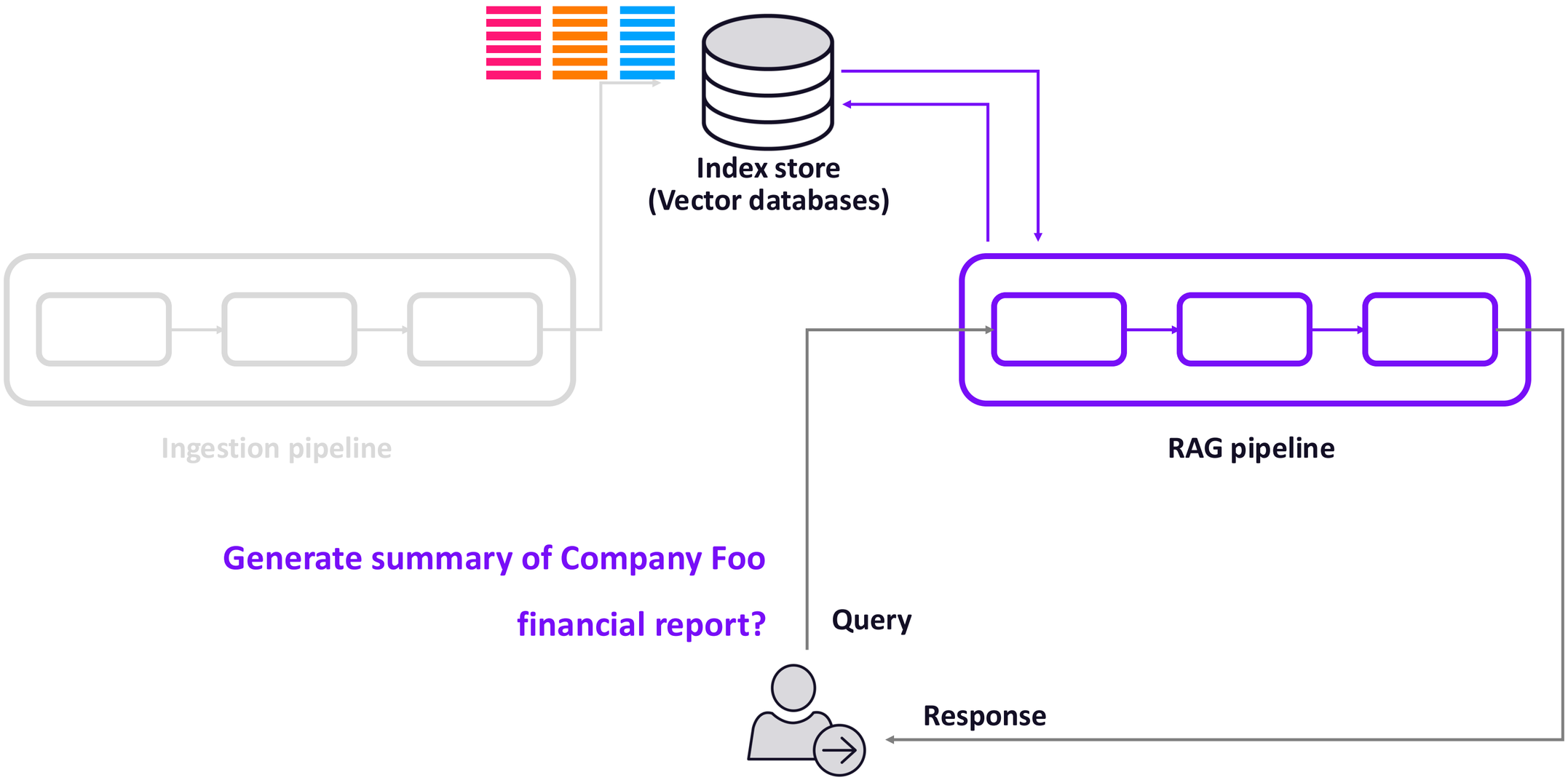
Naïve RAG Lacks Memory



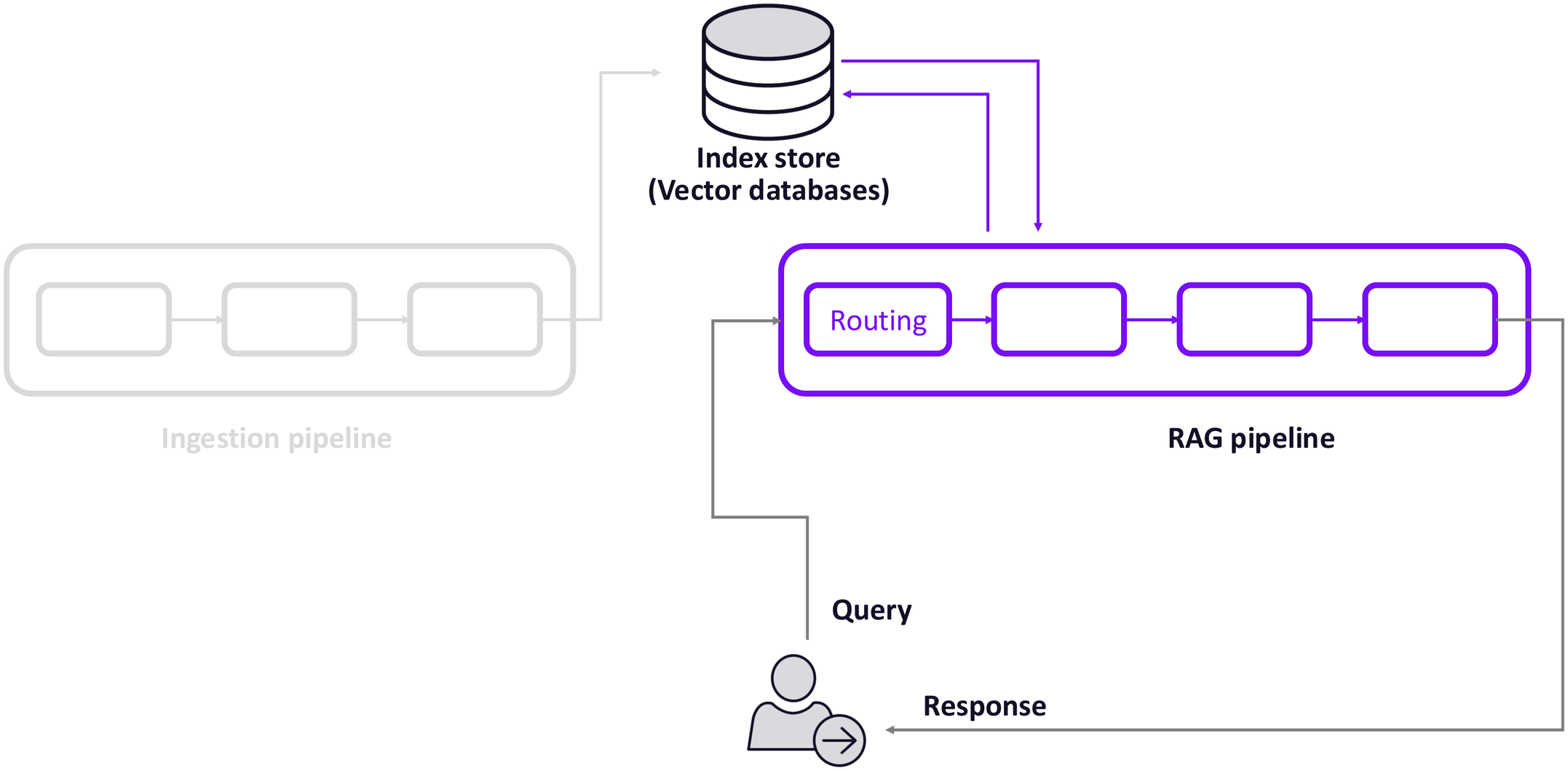
Naïve RAG Provides Semantic Search



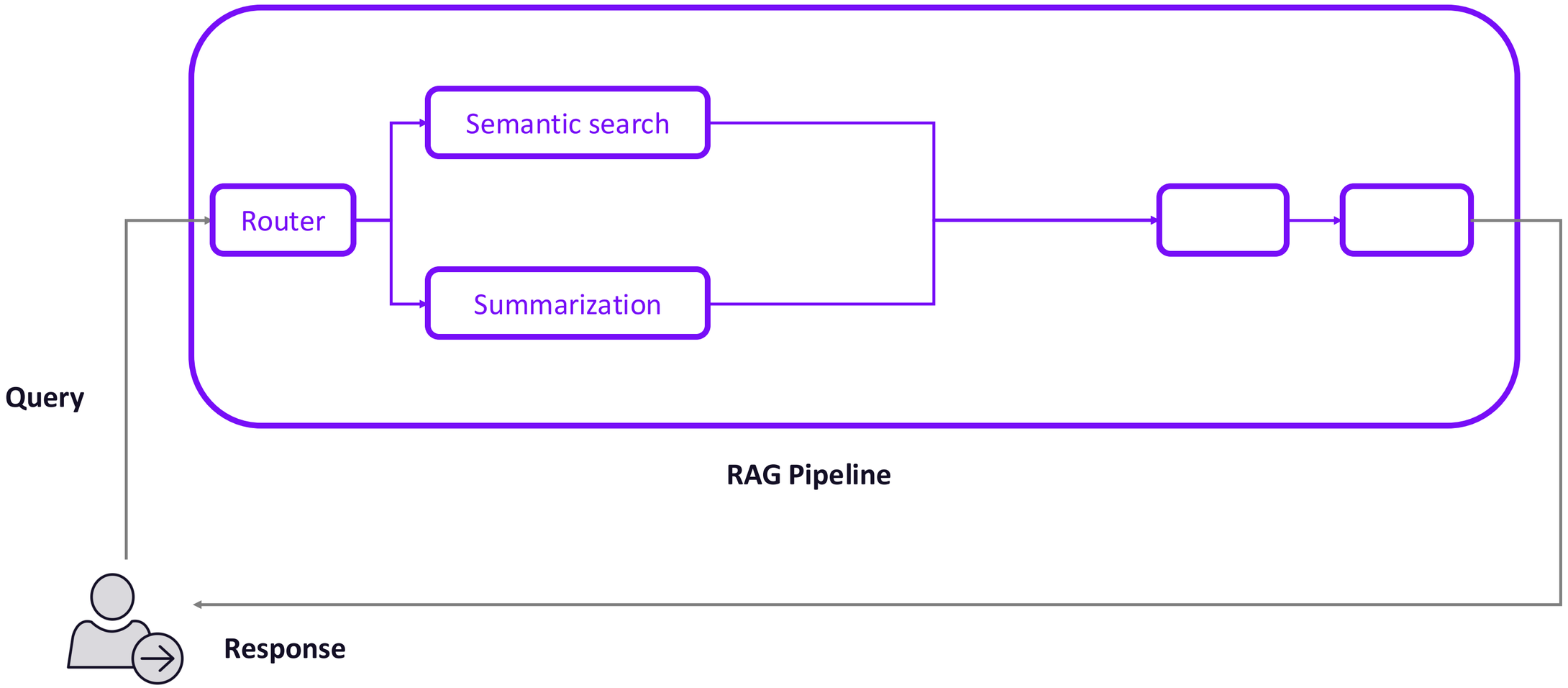
Naïve RAG Lacks Summarization Capability



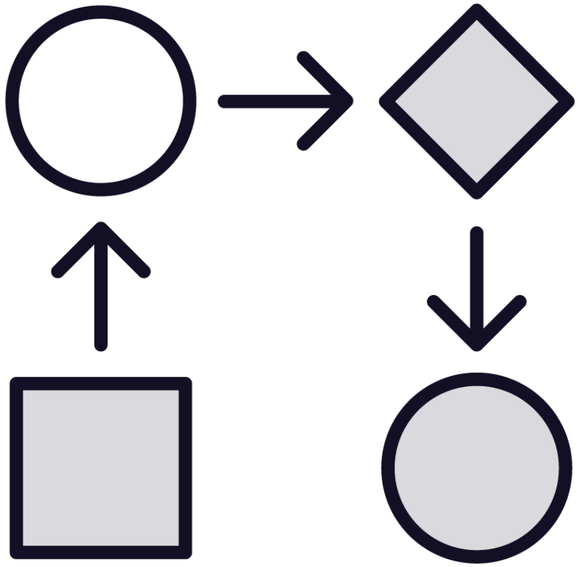
Add Routing Capability



Add Routing Capability



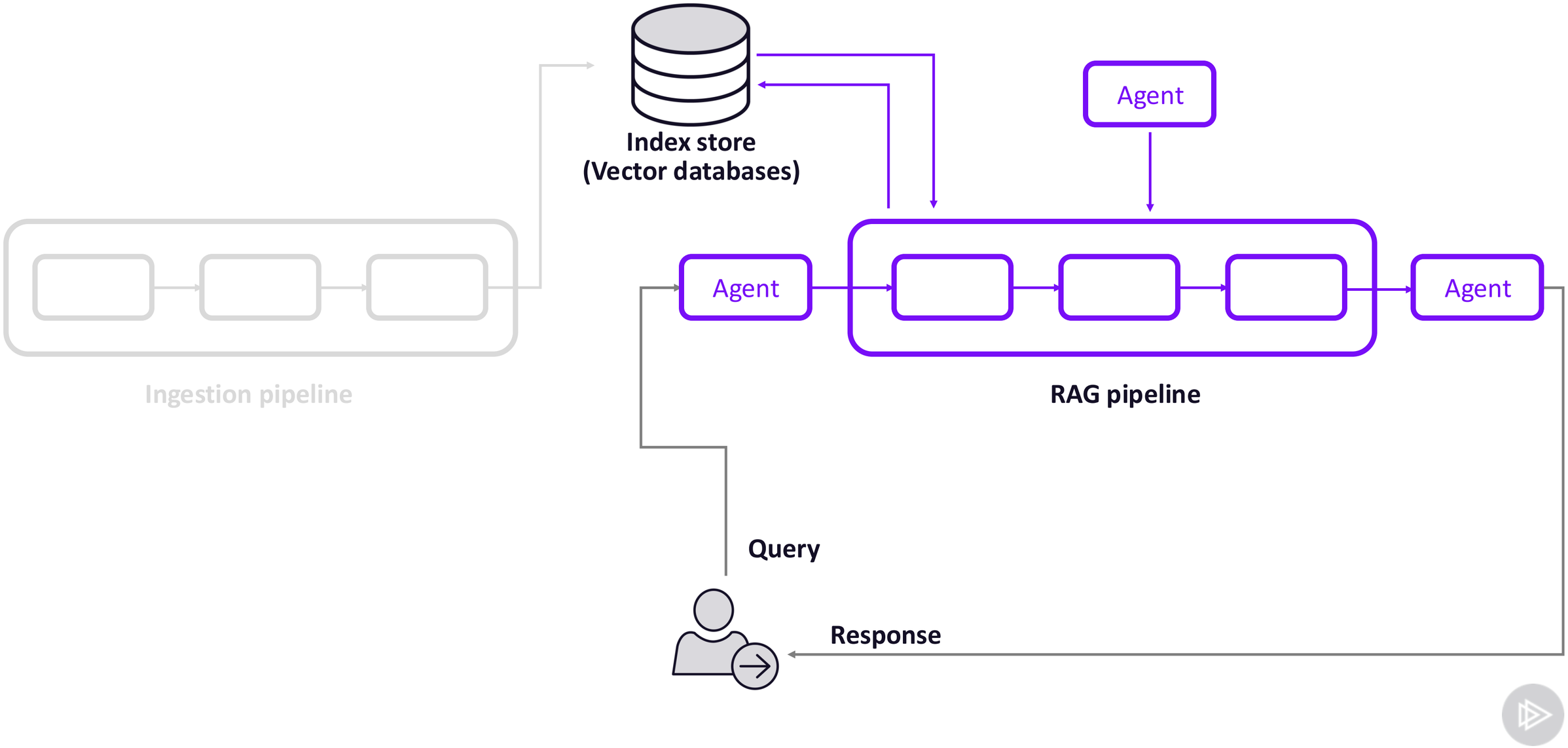
Agentic RAG



Need for agentic RAG

- Multiple tools and integrations
- Enable multi-turn questions
- Complex reasoning
- Perform reflection

Agentic RAG


Index store
(Vector databases)
Agent
Agent
Agent
Ingestion pipeline
RAG pipeline
Query
Response

The world of RAG systems are rapidly evolving.

Where to Go from Here?

Explore RAG frameworks

Keep a tab on ongoing research

Build in an incremental fashion

Make RAG systems safe