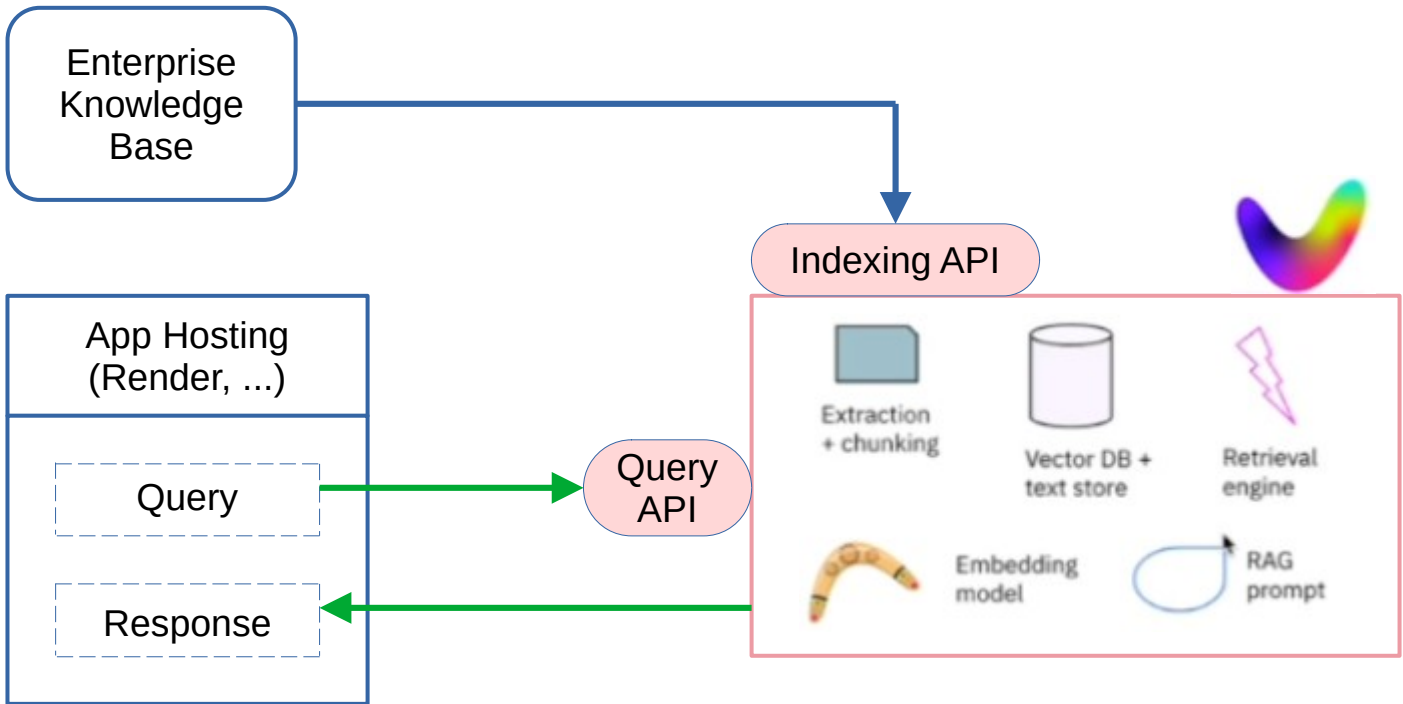# RAG as a Service with Vectara (reference architecture)

Enterprise Knowledge Base

Indexing API

Extraction + chunking

Vector DB + text store

Retrieval engine

Embedding model

RAG prompt

App Hosting (Render, ...)

Query

Query API

Response

QBRAG: QnA Boosted RAG

QnA generation (AnyScale Llama2)

Knowledge Base QnA

Enterprise Knowledge Base

Common Knowledge (Books)

Indexing API

Extraction + chunking

Vector DB + text store

Retrieval engine

Embedding model

RAG prompt

App Hosting (Render, ...)

Query

Query API

Response