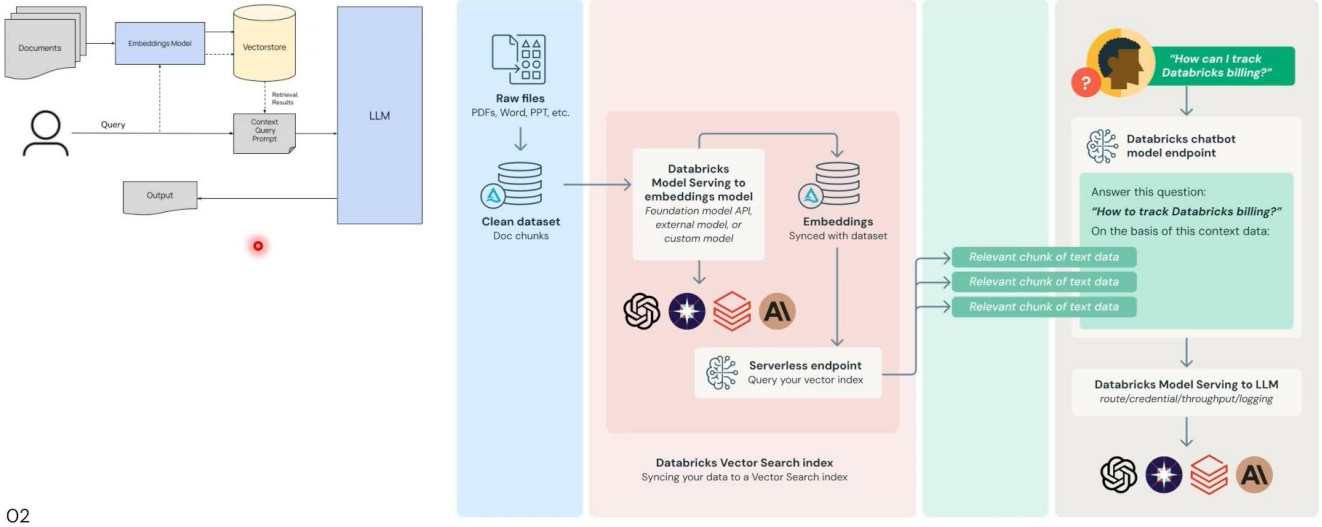
RAG Architecture:

RAG Architecture

RAG architecture combines two core modules—retrieval and generation—to create a seamless, knowledge-aware response system. It retrieves relevant context from a vector store and injects it into prompts sent to the language model, enabling accurate and grounded outputs.

01

Architecture Overview



02