CÔNG TY CỔ PHẦN VCCORP

3rd Week Report

Leader: Mr. Ngô Văn Vĩ

Rag and VectorDB

Lê Văn Hậu

Contents

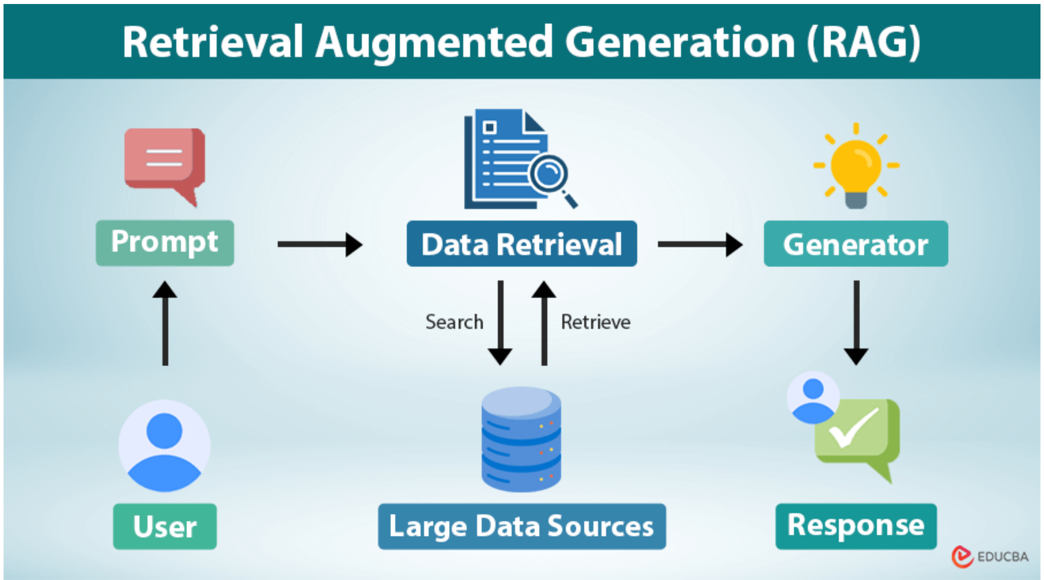
Third Week Report

FireFly

July 3, 2025

1 What is Retrieval Augmented Generation (RAG)?

Retrieval-Augmented Generation refers to an advanced natural language processing technique that combines the strengths of both retrieval models and generative models.



1.1 Retrieval

- **Embedding-based search:** Convert query and documents into dense vectors (e.g. via a transformer encoder) and perform k-nearest-neighbors (kNN) search in vector space to find the most semantically relevant chunks.
- **Sparse retrieval:** Traditional methods (BM25) that match based on token overlap; can be hybridized with dense methods.

1.2 Augmentation

- Retrieved passages are prepended or otherwise injected into the prompt fed to the generator.
- This grounding reduces hallucinations, ensures up-to-date content (as long as the corpus is refreshed), and supports long-term memory use cases.

1.3 Generation

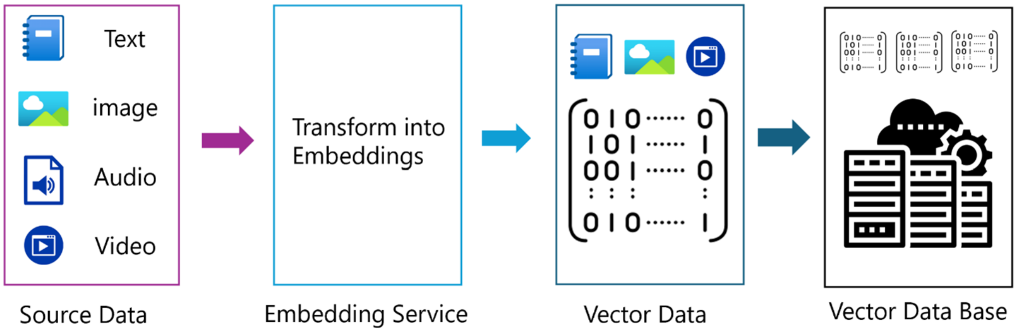
- A large language model (LLM) like GPT-4, Llama-2, or an open-source alternative consumes both the user query and the retrieved context to generate its response.
- Often implemented as a "retrieve-then-read" or "retrieve-then-generate" flow.

1.4 Optional Reranking / Fusion

- Multiple retrieved candidates can be reranked by relevance or relevance-to-query.
- "Fusion-in-Decoder" architectures let the generator attend to all retrieved chunks simultaneously for more holistic answers.

2 Vector Databases

A vector database (vector DB) is a specialized data store optimized for managing and querying high-dimensional embedding vectors. These embeddings typically represent semantic meanings of text, images, audio, or other modalities.



2.1 Core Features

- **Efficient kNN Search:** Uses approximate nearest neighbor (ANN) algorithms (eg. HNSW, IVF-PQ) to quickly find the top-k vectors closest to a query embedding.

- **Scalability:** Can handle millions to billions of vectors, often distributed across clusters.
- **Metadata Filtering:** Supports hybrid queries combining vector similarity with structured filters (eg. data ranges, tags).
- **Real-time Updates**: Allows inserts, updates, and deletes of vectors with low latency.

2.2 Popular Vector Databases

- **Pinecone:** Fully managed services, simple API, auto-scaling.
- **Milvus:** Open-source, cloud-service, rich indexing options.
- **Weaviate:** Schema-driven, supports hybrid graph and vector search, GraphQL API.

3 How RAG and VectorDB work together

Indexing

- Preprocess your corpus (documents, FAQs, knowledge bases) into chunks.
- Encode each chunk with an embedding model (e.g. OpenAI's `text-embedding-ada-002`, Sentence Transformers).
- Store embeddings (and associated chunk metadata) in the vector database.

Query-time Retrieval

- Encode the user's query into an embedding.
- Issue a kNN search to the vector DB to retrieve the top-k most semantically relevant chunks.
- Optionally apply metadata filters (e.g. only search within "Product Manuals" or "Legal Docs").

Augmented Generation

- Concatenate the retrieved chunks with the original query in a prompt template.
- Send the prompt to the generative model to produce a grounded answer.

4 Use Cases

- **Question Answering:** Accurate answers from company-specific knowledge bases or document repositories.
- **Chatbots & Virtual Assistants:** Provide contextual responses that reference up-to-date policies, manuals, or logs.
- **Content Generation:** Generate blog posts or summaries based on large corpora while grounding in actual source material.
- **Semantic Search:** Beyond keyword matching, find documents by meaning (e.g. “find all contracts related to data privacy”).