Introdução:

A Retrieval-Augmented Generation (RAG) representa uma inovação significativa na área de inteligência artificial, combinando a potência dos modelos de linguagem de grande porte (LLMs) com sistemas eficientes de recuperação de informações. Esses LLMs são treinados com uma quantidade massiva de dados, o que lhes confere a capacidade de entender e gerar linguagem natural de forma eficaz para diversas tarefas. O objetivo principal do RAG é assegurar que o conteúdo gerado seja verificável e baseado em dados concretos fornecidos pelo usuário, reduzindo assim a incidência de "alucinações" — informações incorretas ou enganosas que podem surgir em sistemas de IA generativa.

Fontes:

<https://www.oracle.com/br/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/?utm_source=chatgpt.com>

<https://www.ibm.com/br-pt/think/topics/large-language-models?utm_source=chatgpt.com>

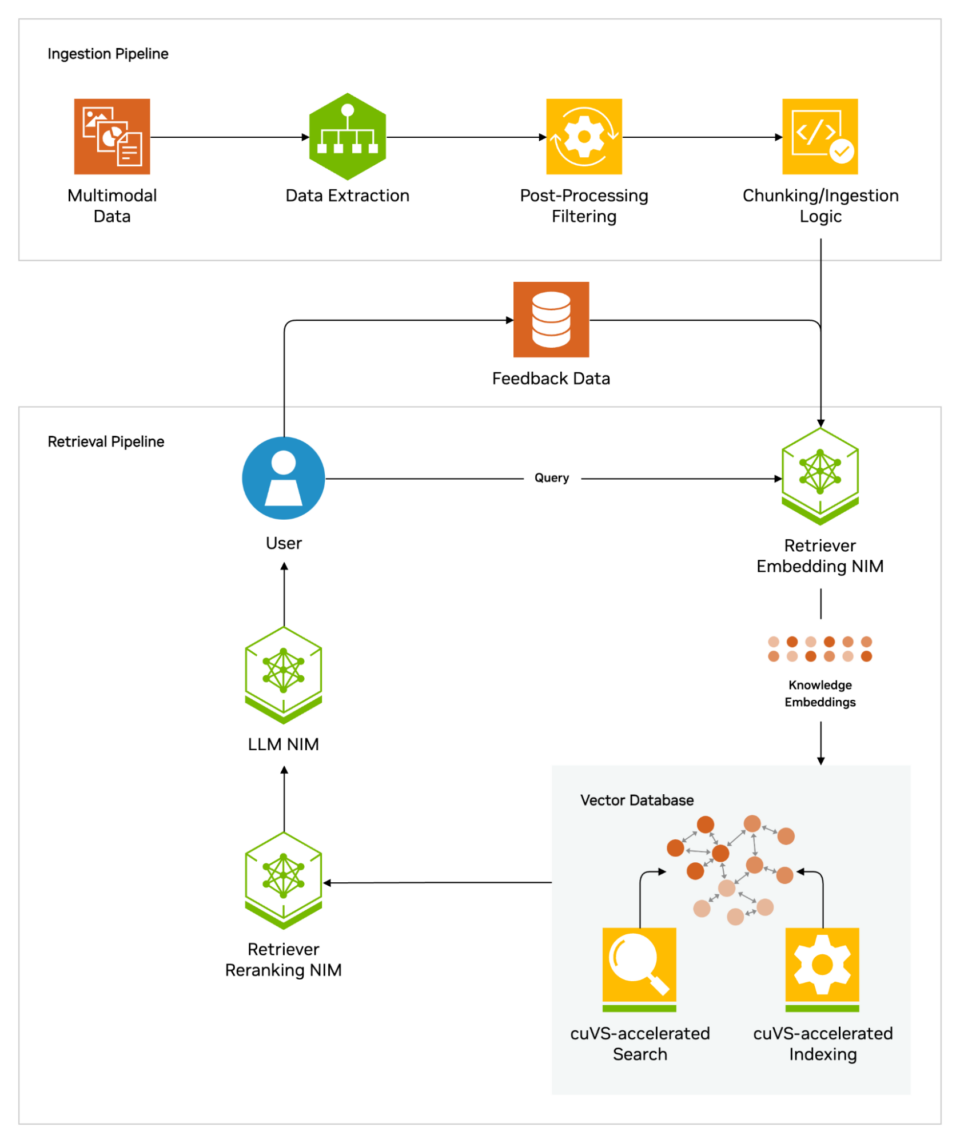
<https://docs.vectara.com/docs/learn/grounded-generation/grounded-generation-overview>

Fundamentos:

Os fundamentos essenciais do Retrieval-Augmented Generation (RAG) possuem dois pilares principais: Data Retrieval e Content Generation. O Data Retrieval consiste no processo pelo qual o modelo de IA extrai dados de um corpus estruturado. Esses dados servem como uma base para geração de respostas. O Content Generation, por sua vez, é a etapa na qual o modelo cria respostas para o usuário, utilizando como base as informações recuperadas pelo Data Retrieval, o que aumenta a precisão e a veracidade das respostas geradas.

Essa abordagem de RAG é crucial para minimizar as "alucinações", termo utilizado para descrever informações imprecisas ou enganosas que podem surgir em sistemas de IA generativa. Ao implementar o RAG, é possível transformar a maneira como a recuperação de informações e as interações com a IA são conduzidas, particularmente em contextos onde a integridade das informações é de suma importância.

Fonte:  
<https://docs.vectara.com/docs/learn/grounded-generation/grounded-generation-overview>



Pipeline:  
  
Ingestion Pipeline: O Ingestion Pipeline é a primeira etapa do processo RAG, onde dados estruturados e não estruturados são extraídos de diversos formatos. Esses dados são convertidos em texto, passam por um processo de limpeza e filtragem para reter apenas as informações relevantes, e são então divididos em unidades menores, como parágrafos. Por fim, essas unidades são indexadas em um banco de dados vetorizado, facilitando a recuperação eficiente das informações.

Retrieval Pipeline: Na fase do Retrieval Pipeline, a consulta do usuário é processada por um modelo que gera um embedding — uma representação numérica da consulta. Este embedding é comparado com os vetores de dados indexados, utilizando métricas como similaridade de cosseno ou distância euclidiana. Os resultados são ordenados por relevância, e os mais pertinentes são enviados ao LLM, que pode então gerar respostas informadas pela fonte de dados ou incorporar essas informações adicionais. O feedback do usuário é valioso, pois pode ser utilizado para aprimorar a filtragem, o ranking de resultados ou até mesmo para melhorar o treinamento dos modelos de embedding.

Fontes:  
<https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/>

<https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/?utm_source=chatgpt.com>