Introdução:

A Retrieval-Augmented Generation (RAG) representa uma inovação significativa na área de inteligência artificial, pois combina a potência dos modelos de linguagem de grande porte (LLMs) com sistemas eficientes de recuperação de informações. Embora os LLMs sejam treinados com uma quantidade massiva de dados, o que lhes confere uma notável capacidade de entender e gerar linguagem natural, eles têm limitações para acessar e manipular conhecimentos específicos de forma precisa. O RAG surge para mitigar esses problemas, assegurando que o conteúdo gerado seja verificável e baseado em dados concretos, o que reduz a incidência de "alucinações" – informações incorretas ou enganosas.

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>

https://proceedings.neurips.cc/paper\_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

Fundamentos:

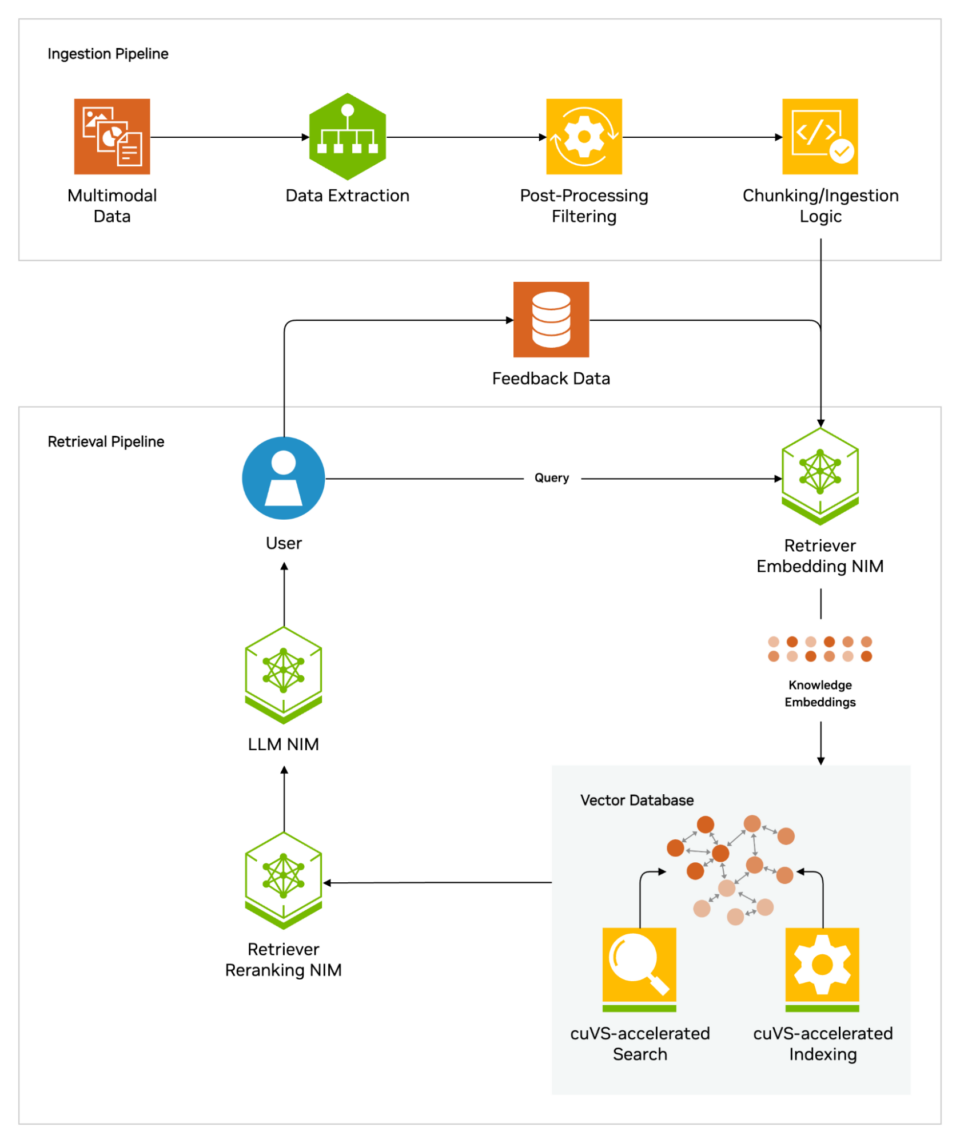
Os fundamentos do RAG estão alicerçados em dois pilares principais:

* **Data Retrieval:** A etapa de recuperação, onde o modelo extrai informações relevantes de um corpus estruturado (por exemplo, um índice denso de documentos da Wikipedia, utilizando técnicas como o Dense Passage Retriever – DPR).
* **Content Generation:** A etapa de geração, na qual um modelo seq2seq pré-treinado (como o BART) utiliza os dados recuperados para formular respostas mais precisas e factuais.

Essa abordagem híbrida permite que a memória paramétrica (conhecimento armazenado nos parâmetros do modelo) seja complementada por uma memória não-paramétrica (dados externos), aumentando assim a veracidade e a especificidade das respostas geradas. O artigo de Patrick Lewis et al. aprofunda essas ideias e explora variantes como o RAG‑Sequence e o RAG‑Token, que diferem na forma de integrar os documentos recuperados na geração da resposta.

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

https://proceedings.neurips.cc/paper\_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf



Pipeline:  
  
**Ingestion Pipeline:** Na primeira etapa, dados estruturados e não estruturados são extraídos de diversas fontes e convertidos em texto. Esse texto passa por processos de limpeza, filtragem e segmentação (por exemplo, em parágrafos), sendo posteriormente indexado em um banco de dados vetorizado. Esse índice facilita a recuperação eficiente das informações relevantes.

**Retrieval Pipeline:** Quando uma consulta é realizada, ela é transformada em uma representação numérica (embedding) que é comparada com os embeddings dos dados indexados, utilizando métricas de similaridade (como similaridade de cosseno ou distância euclidiana). Os documentos mais relevantes são selecionados e enviados ao modelo gerador, que os utiliza para compor uma resposta informada. O artigo detalha ainda como essa integração pode ser feita de maneira probabilística, utilizando técnicas de marginalização sobre os documentos recuperados.

Durante o treinamento, o retriever e o gerador podem ser ajustados conjuntamente (fine‑tuning end‑to‑end), o que permite que o sistema aprenda a recuperar e a integrar a informação de maneira otimizada.

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>

https://proceedings.neurips.cc/paper\_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf