INSIGHTS · INTELIGÊNCIA ARTIFICIAL

RAG não é magia — é engenharia: do chunking ao serving no Microsoft Fabric

Desliza →

INTELIGÊNCIA ARTIFICIAL

01

O que é RAG?

Combina recuperação vetorial com LLM para respostas contextualizadas. Ideal quando os dados já residem em OneLake/Lakehouse.

INTELIGÊNCIA ARTIFICIAL

02

Arquitectura essencial

Quatro blocos: ingestão, embeddings, índice vetorial e orquestração retrieval+LLM. Use tabelas Delta e compute nativo em Fabric.

INTELIGÊNCIA ARTIFICIAL

03

Gerar embeddings

Gere em batch via notebooks Spark para paralelizar. Registe modelo, versão e hash do documento numa tabela de auditabilidade.

INTELIGÊNCIA ARTIFICIAL

04

Chunking faz a diferença

500–1 000 tokens para documentação técnica; 150–300 para FAQs. Remover boilerplate e deduplicar melhora recall e reduz custos.

05

Indexação e latência

Opções: Faiss/HNSW em compute ou vector DB gerido. 100k embeddings (1 536 dims) ocupam ≈600 MB; buscas em 30–150 ms.

INTELIGÊNCIA ARTIFICIAL

06

Integração com Power BI

Tabelas pré-computadas para SLAs previsíveis ou endpoints em tempo real para exploração. Trace os IDs das fontes para auditoria.

07

Mini-caso em números

25 000 documentos → 80k chunks; 80k embeddings em 90 min; top-10 ANN ≈45 ms; tempo até resposta útil caiu para 18 min.

CONTINUA NO BLOG

Por que adoptar RAG no Microsoft Fabric?

www.bconcepts.pt/insights

bConcepts
Consulting Services

Artigo completo → link nos comentários