



RAG no Microsoft Fabric: ligar LLMs ao Power BI sem alucinações

Desliza →

O que é RAG?

Retrieve-and-Generate combina recuperação baseada em documentos com geração por LLMs para respostas com origem identificável.

Porquê usar com Fabric + Power BI

Reduz alucinações e produz explicações com fontes. Ex.: ver taxa de rotura por SKU e contexto extraído de comunicações dos últimos 90 dias.

Arquitetura em 4 blocos

OneLake (ficheiros), pipeline de embeddings, indexador vetorial (Azure Cognitive Search ou vector DB) e uma API que serve o Power BI.

Preparar dados: chunking

Docs técnicos: 600–1.000 tokens com 15–25% overlap; emails: 250–400 tokens; políticas: 400–800 tokens. Enriquecer com metadados (data, origem, confidencialidade).

Onde guardar embeddings

Guarde vetores e metadados em ficheiros parquet em OneLake para auditabilidade. Ex.:
200k chunks $\times$ 1.536 dims $\approx$ 3–5 GB com overhead.

Indexação: 3 opções

Azure Cognitive Search (integração), vector DB gerido (controle e throughput) ou OneLake+Faiss (mais económico). Custos: para 100k q/m podem variar ~€200–€1.000/mês.

Servir respostas e latência

Objetivos: retrieval puro <200 ms;
retrieval+geração 800–1.200 ms para modelos rápidos. Use caching (Redis) com TTL 15 min–1 h para perguntas recorrentes.

CONTINUA NO BLOG

Como vai integrar RAG no Power BI da sua organização?

www.bconcepts.pt/insights