Arquitetura de Referência: RAG Enterprise

Atlantyx | Documento técnico | Data: 05/11/2025

1. Visão geral

A arquitetura RAG enterprise separa ingestão, indexação, serving e observabilidade.

Objetivo: responder perguntas corporativas com base em conhecimento interno, reduzindo alucinação via citações.

2. Componentes obrigatórios

Origem de documentos: repositórios corporativos (Blob/SharePoint/Confluence).

Ingestão: extração, normalização, chunking, enriquecimento de metadados (owner, classificação, data, versão).

Indexação: busca híbrida (BM25 + vetorial) com filtros por permissão (RBAC/ABAC).

Serving: API com autenticação, retrieval, re-ranking opcional e geração com citações.

Observabilidade: métricas de latência, custo, qualidade e erros, com correlação por conversation_id.

3. Padrões de qualidade

A resposta deve incluir: (a) conclusão, (b) evidências citadas, (c) limitações quando aplicável.

Se a evidência for insuficiente, retornar 'não encontrei base suficiente' e sugerir fonte humana.

4. Estratégias de resiliência

Fallback: se o LLM estiver indisponível, retornar trechos recuperados com sumário curto.

Cache: respostas para perguntas frequentes e cache de embeddings para reduzir custo.

Rate limit: proteger o serviço contra abuso e picos.

Tabela 1 - Campos mínimos de metadados no índice

Campo	Tipo	Uso
doc_id	string	Rastrear documento e versão
source	string	origem (Blob/SharePoint/HTML)
owner_area	string	filtro por área/dono
classification	string	publico/interno/restrito
effective_date	date	priorizar versões recentes
rbac_tags	array	permissões e grupos