

warning "Atualização de backend (outubro de 2026)" A inferência do GitHub Models foi descontinuada em 30 de julho de 2026. As referências abaixo são mantidas apenas como contexto histórico enquanto este lab é migrado. Use o Lab 013 ? Backends de LLM gratuitos para configurar o Foundry Local ou o Ollama antes de continuar.

Lab 039: Comparação de Bancos de Dados Vetoriais

Nível: L300 Trilha: ? RAG Tempo: ~50 min ? Custo: Free ? Todas as opções têm planos gratuitos ou modo local

O que Você Vai Aprender

Entender as principais diferenças entre os grandes bancos de dados vetoriais Comparar pgvector, Chroma, Qdrant e Azure AI Search na mesma tarefa Avaliar cada um em: complexidade de configuração, velocidade de consulta, filtragem, integração com nuvem Escolher o banco de dados vetorial certo para os requisitos do seu agente

Introdução

Escolher um banco de dados vetorial é uma das decisões arquiteturais mais consequentes para um agente baseado em RAG. A "melhor" escolha depende da stack existente da sua equipe, requisitos de escala, necessidades de filtragem e estratégia de nuvem.

Os candidatos:

Banco de Dados | Tipo | Opção gratuita | Melhor para | pgvector | Extensão PostgreSQL | Azure free tier / local | Equipes que já usam PostgreSQL | Chroma | Embutido / servidor | Totalmente open-source | Desenvolvimento local, projetos pequenos | Qdrant | BD vetorial dedicado | Qdrant Cloud free tier | Produção em alta escala, filtragem avançada | Azure AI Search | Serviço Azure | Free tier (1 índice) | Nativo do Azure, busca híbrida, enterprise |

Pré-requisitos

[diagram or code sample]

Nenhuma chave de API necessária para Chroma (local) e Qdrant (modo local). GitHub Token necessário para embeddings: [diagram or code sample]

Configuração: Função de Embedding Compartilhada

Todos os quatro testes usam os mesmos dados de produtos OutdoorGear e o mesmo modelo de embedding:

[diagram or code sample]

Opção A: Chroma (Local, Sem Configuração)

Chroma é a maneira mais fácil de começar ? Python puro, roda em memória ou em disco:

[diagram or code sample]

Opção B: Qdrant (Modo Servidor Local)

Qdrant oferece filtragem avançada e escala para centenas de milhões de vetores:

[diagram or code sample]

[diagram or code sample]

Opção C: pgvector (Azure PostgreSQL ou Local)

Veja o Lab 031 para a configuração completa do pgvector. Comparação rápida:

[diagram or code sample]

Opção D: Azure AI Search (Busca Híbrida)

Azure AI Search suporta de forma única a busca híbrida: busca vetorial + busca por palavras-chave BM25 combinada com reranqueamento semântico:

[diagram or code sample]

Framework de Decisão

[diagram or code sample]

Tabela Comparativa

| pgvector | Chroma | Qdrant | Azure AI Search | Configuração | Média (BD necessário) |

Mínima | Fácil (Docker) | Média (Azure) | Dev local | ? (Docker) | ? (em memória) | ? (Docker) | ? (apenas Azure) | Busca híbrida | ? (manual) | ? | ? | ? (integrada) | Filtragem | SQL WHERE | Metadados básicos | Avançada | OData completo | Escala | Moderada (100M) | Alta (enterprise) | Integração Azure | ? (PG gerenciado) | ? | Parcial | ? (nativa) | Custo (plano gratuito) | Free tier PG | Gratuito | Qdrant Cloud gratuito | 1 índice gratuito |

? Verificação de Conhecimento

question "1. O que é busca híbrida e por que ela costuma ser melhor que busca vetorial pura?" Busca híbrida combina busca vetorial (semântica) com busca por palavras-chave (BM25) e classifica os resultados usando ambos os sinais. A busca vetorial se destaca na compreensão semântica ("abrigo quente" ? saco de dormir) mas pode perder correspondências exatas (um ID de produto específico). O BM25 é excelente para correspondências exatas de palavras-chave mas perde sinônimos. Combinar ambos supera qualquer um isoladamente, especialmente para nomes de produtos, SKUs e terminologia especializada.

question "2. Por que você escolheria pgvector em vez de um banco de dados vetorial dedicado como o Qdrant?" Se você já tem PostgreSQL como seu banco de dados principal, o pgvector adiciona busca vetorial sem adicionar outro serviço para operar, manter e pagar. Os dados ficam junto com seus dados relacionais ? você pode fazer JOIN de registros de produtos com seus embeddings em uma única consulta. Para a maioria das aplicações com menos de 1M de vetores, o desempenho do pgvector é excelente. Escolha Qdrant quando precisar de > 100M de vetores ou filtragem muito avançada.

question "3. O que é o índice IVFFlat no pgvector e quando você deve usar HNSW em vez dele?" IVFFlat (Inverted File Index com quantização Flat): rápido de construir, usa menos memória, bom para conjuntos de dados que não mudam frequentemente. Usa busca aproximada ? o parâmetro lists controla o trade-off entre recall e velocidade. HNSW (Hierarchical Navigable Small World): melhor recall, consultas mais rápidas, mas maior uso de memória e mais lento para construir. Use IVFFlat para conjuntos de dados < 1M que não mudam muito; use HNSW para conjuntos de dados frequentemente atualizados ou quando o recall é crítico. Ambos requerem pgvector ? 0.5.0.

Resumo

Para o cenário do OutdoorGear learning hub (< 10K produtos, infraestrutura Azure, equipe conhece SQL):

Recomendado: pgvector no Azure Database for PostgreSQL Flexible Server.

Nenhum serviço novo para aprender SQL + vetorial em uma única consulta Free tier disponível

Pronto para produção com as migrações do Lab 031

Próximos Passos

Monte a configuração pgvector: ? Lab 031 ? Busca Semântica com pgvector Aplicação RAG

completa: ? Lab 022 ? RAG com GitHub Models + pgvector RAG empresarial com avaliação: ? Lab 042 ? RAG Empresarial *(em breve)*

Kit de apoio

O build do site gera downloads estáveis em PDF e ZIP. O ZIP contém o guia do lab e todos os arquivos de apoio declarados para este lab.

Baixar PDF gerado Baixar ZIP gerado

Supporting file: CHEATSHEET.md

Arquivos de apoio: Lab 039: Comparação de Bancos de Dados Vetoriais

Comece pelo caminho local ou gratuito documentado. Registre premissas e evidências de validação. Mantenha serviços hospedados opcionais, salvo quando o lab os exigir.

Supporting file: README.md

Lab 039: Comparação de Bancos de Dados Vetoriais

Use estes arquivos com o guia do lab e o kit gerado.