

A Nova Aliança: Fundamentos de Dados na Era da IA Generativa

Como uma base de dados estruturada e consistente alimenta
análises avançadas e aplicações de IA e ML.

Baseado no Capítulo 9 da Apostila de Banco de Dados.

O Cenário Atual e o Nosso Foco



Você Sabia?

- O ChatGPT-3 foi treinado com mais de **570 GB** de texto, armazenados e indexados em bancos de dados especializados.
- Bancos de dados vetoriais (como Pinecone e pgvector) são a base da busca semântica em IA generativa.
- **70%** das empresas planejam integrar IA em suas análises de dados até 2025.



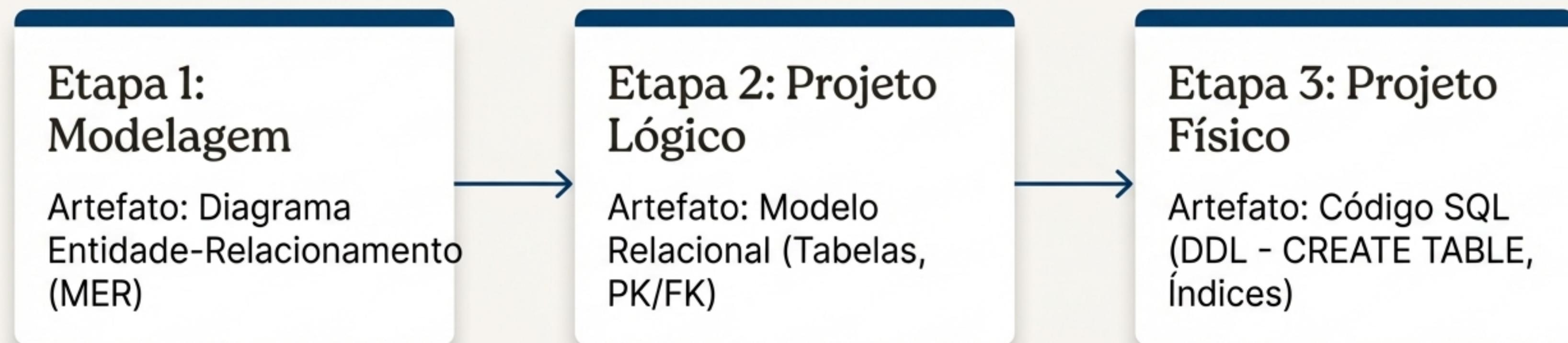
Aviso Importante

- Este material foca nos **fundamentos de Banco de Dados** que preparam e viabilizam o uso de tecnologias de IA Generativa.
- Não detalharemos ferramentas específicas (ChatGPT, Gemini, etc.), mas sim como a disciplina de BD é o alicerce para seu sucesso.

*Fonte: OpenAI, Gartner. Dados de 2023/2024.

O Alicerce de Tudo: O Pipeline de Projeto de um Banco de Dados

O projeto de um BD robusto segue etapas claras e produz artefatos específicos. A qualidade do resultado final depende da excelência em cada fase.



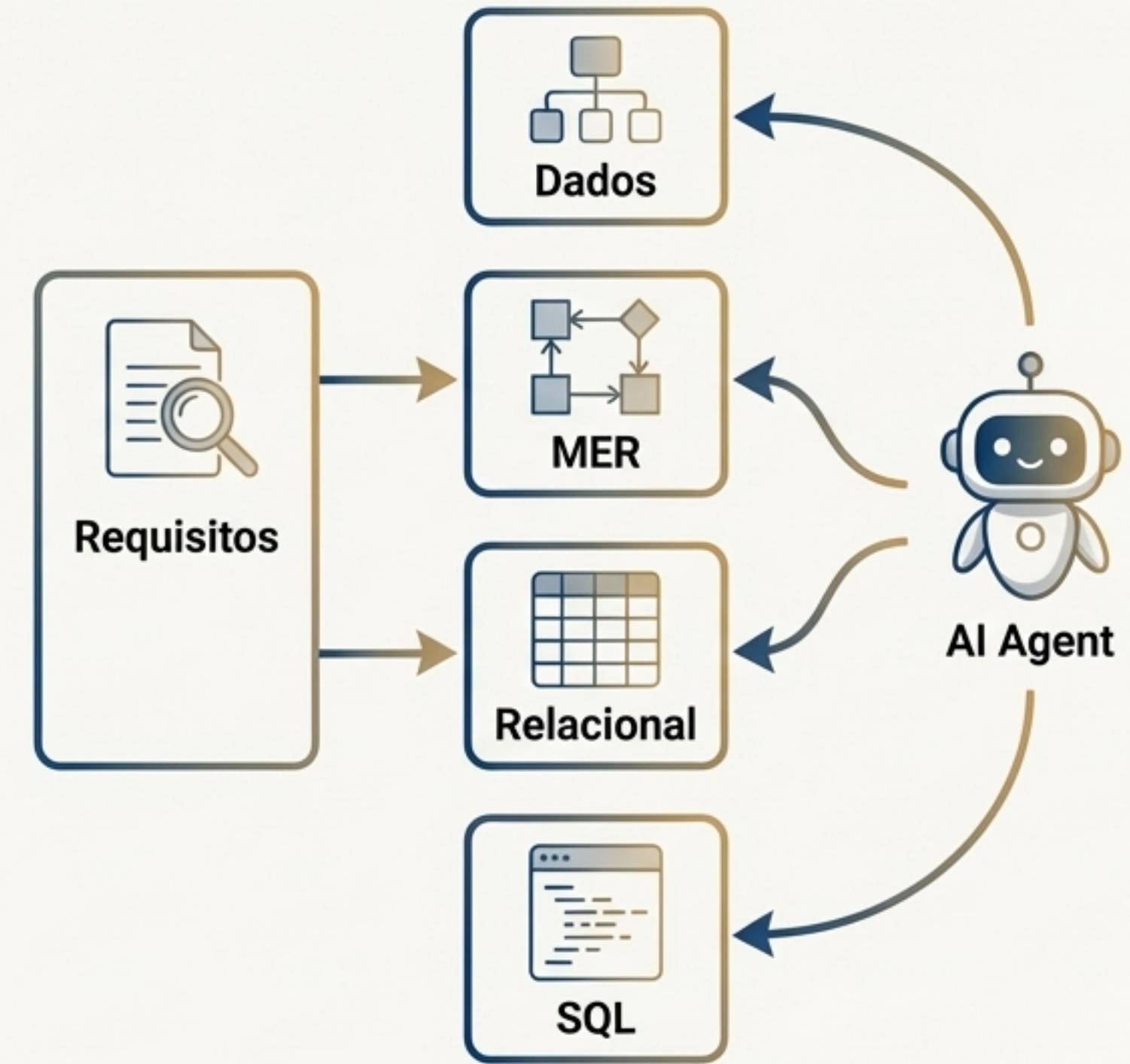
Cada etapa constrói sobre a anterior, transformando requisitos de negócio em uma estrutura de dados executável e eficiente.

IA Generativa como Co-piloto no Projeto de BD

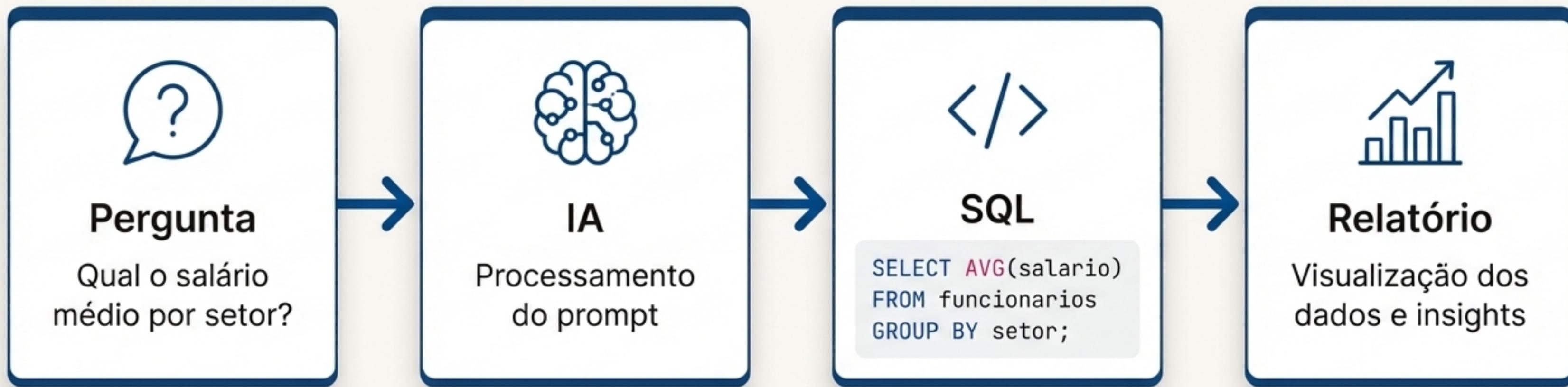
A IA atua como um assistente poderoso, capaz de:

- Transcrever e organizar requisitos de negócio.
- Sugerir modelos Entidade-Relacionamento (MER).
- Converter o modelo lógico em tabelas relacionais.
- Gerar o código SQL (DDL) para a criação física do banco.

Mensagem Chave: A IA acelera a tradução entre os artefatos, permitindo que o profissional foque na validação, otimização e nos ajustes finos.



Da Pergunta de Negócio ao Relatório: Análise Potencializada por IA



Exemplos de Prompts Adicionais

Total por categoria

Top 10 clientes

Média mensal

Traduzindo Objetivos de Negócio em SQL

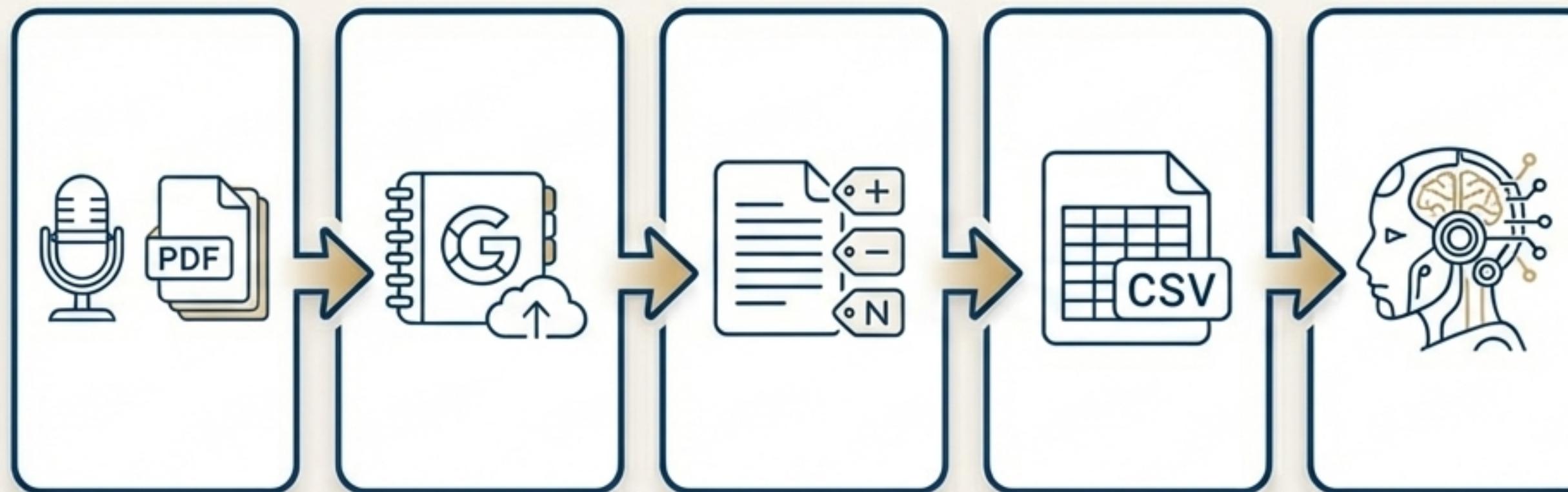
Consultas avançadas (**DQL**) transformam dados brutos em informação. A IA pode ajudar a construir junções, agregações e subconsultas complexas.

Objetivo do Relatório	Comando SQL Base
Salário médio por setor	SELECT AVG(Salario), ID_Setor FROM FUNCIONARIO GROUP BY ID_Setor;
Cientes com vendas acima de um determinado valor	JOIN + Agregação (SUM ou COUNT) + HAVING por cliente
Listar professores por departamento	JOIN PROFESSOR ↔ DEPARTAMENTO

****Mensagem Chave:**** O domínio de conceitos como **GROUP BY**, **HAVING** e **JOIN** é fundamental para validar e refinar as consultas geradas pela IA.

Estudo de Caso: Usando NotebookLM para Gerar Datasets

Contexto: Gerar um dataset limpo para análise de sentimentos a partir de fontes não estruturadas como áudio e PDFs.



1. Entrada
Áudio/PDF

2. NotebookLM
Processamento com Google AI

3. Texto
Texto limpo com rótulos de sentimento
(positivo, negativo, neutro)

4. CSV
Dataset estruturado

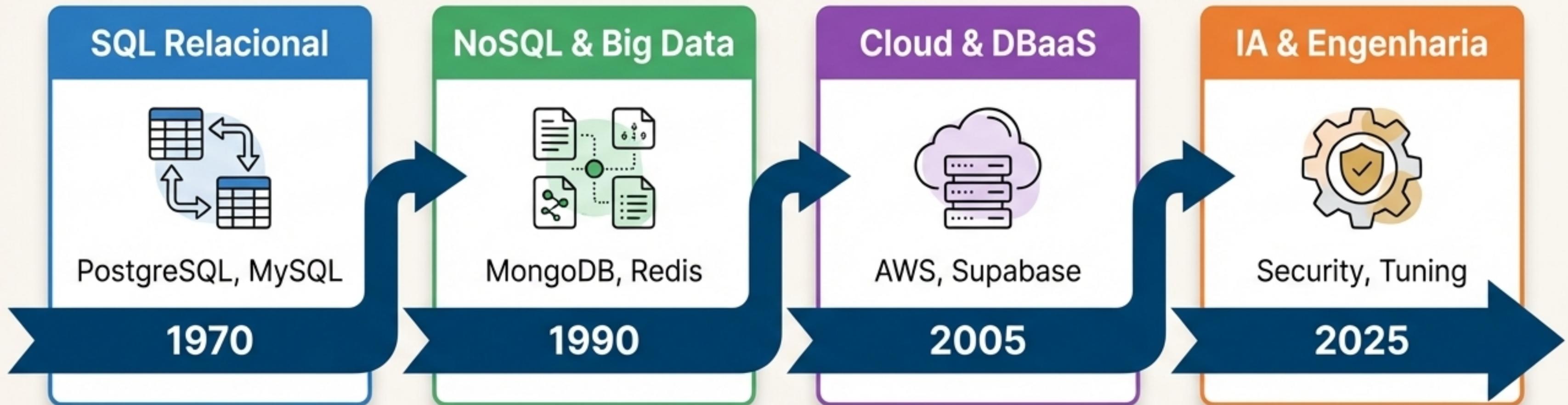
5. ML
Pronto para o modelo de Machine Learning

Conexão com BD

A qualidade do dataset final depende de **princípios de banco de dados** aplicados durante a estruturação:

- **Normalização (1FN, 2FN, 3FN)** para minimizar redundância.
- **Integridade (PK/FK, domínios)** para garantir consistência.

A Evolução do Mercado de Dados: Onde Estamos Hoje



Cada era construiu sobre a anterior. O modelo relacional trouxe independência de dados. O NoSQL lidou com formatos não tradicionais. A nuvem trouxe escalabilidade. Agora, a IA integra-se a essa infraestrutura para gerar inteligência.

A Fronteira da IA: RAG (Recuperação Aumentada por Geração)

Inter

O que é?

RAG é uma técnica que combina a **busca em fontes de dados confiáveis** com a capacidade de geração de linguagem dos LLMs.

Por que é importante?

Inter Regular

Reduz drasticamente as “alucinações” da IA ao forçá-la a basear suas respostas em um contexto relevante e verificado, extraído diretamente de seus dados.

Inter

Componentes-Chave do Pipeline:



Indexação/Embeddings:

Documentos são representados como vetores.



Retriever:

O sistema recupera os trechos mais relevantes para a pergunta.



Prompt Template:

O contexto recuperado é injetado no prompt do LLM.



LLM:

Gera a resposta final com base no contexto fornecido.

RAG na Prática: Bancos de Dados Vetoriais

A Conexão:

A “recuperação” no RAG acontece dentro de um banco de dados otimizado para busca de similaridade, como bancos vetoriais.

Exemplo com pgvector (PostgreSQL):

Extensões como `pgvector` transformam o PostgreSQL em um poderoso banco de dados vetorial.

```
-- Habilitar extensão de vetores
CREATE EXTENSION IF NOT EXISTS vector;

-- Tabela para armazenar documentos e seus vetores (embeddings)
CREATE TABLE docs (
    id SERIAL PRIMARY KEY,
    content TEXT NOT NULL,
    embedding VECT0R(1536) -- Ex: dimensão do embedding da OpenAI
);

-- Busca por similaridade semântica
SELECT id, content FROM docs
ORDER BY embedding <-> :query_embedding -- O operador <-> faz a busca
LIMIT 5;
```

Da Experimentação à Produção: A Disciplina de MLOps

Definição: MLOps organiza o ciclo de vida de modelos de Machine Learning em produção, garantindo versionamento, implantação, monitoramento e governança.

Experiment Tracking

Registro de métricas e hiperparâmetros (ex: MLflow).

Versionamento

Controle de versões de dados, código e modelos (ex: DVC).

CI/CD

Automação de testes, validação e implantação.

Monitoramento

Acompanhamento de qualidade, *data drift* e *concept drift*.

Retraining

Rotinas para re-treinamento com dados atualizados.

Feature Store

Catálogo centralizado de *features* para consistência.

Estudo de Caso Final: Análise de Risco em Empréstimos de Biblioteca

Contexto do Negócio

Gerar um dataset consolidado para prever atrasos na devolução de livros ou para criar um sistema de recomendação simples.

Fontes de Dados



Três tabelas normalizadas: `ALUNO`, `LIVRO`, `EMPRESTIMO`.

Estratégia de Preparação de Dados

- JOINS**: Utilizar as chaves primárias e estrangeiras (PK/FK) para recompor uma visão unificada dos dados.
- Feature Engineering**: Criar novas colunas (features) relevantes para o modelo, como `Dias_Atraso`, calculada a partir das datas existentes.
- Exportação**: Gerar um arquivo CSV final, pronto para ser consumido por ferramentas de Machine Learning.

A Solução: O Poder do SQL para Preparar Dados para IA

Consulta SQL para Geração do Dataset:

```
SELECT
    A.Matricula,
    A.Nome AS Nome_Aluno,
    L.Titulo AS Titulo_Livro,
    E.Data_Emprestimo,
    E.Data_Devolucao_Prevista,
    E.Data_Devolucao_Real,
    -- Feature Engineering: Calculando os dias de atraso
    (E.Data_Devolucao_Real - E.Data_Devolucao_Prevista) AS Dias_Atraso
FROM
    ALUNO AS A
JOIN
    EMPRESTIMO AS E ON E.Matricula_Aluno = A.Matricula
JOIN
    LIVRO AS L ON L.Codigo = E.Codigo_Livro
ORDER BY
    A.Nome, E.Data_Emprestimo;
```

Análise: Esta única consulta realiza a junção de dados, a renomeação de colunas para clareza e a criação de uma *feature* preditiva (`Dias_Atraso`), demonstrando o poder expressivo do SQL.

Princípios Atemporais na Era da IA

A Nova Aliança: A IA Generativa não substitui os fundamentos de dados; ela os potencializa.

Síntese da Jornada:



Começamos com o **Fundamento**: O pipeline de projeto de BD e a importância do SQL para design (DDL) e análise (DQL).



Vimos a IA como um **Co-piloto**: Acelerando a criação de código, a geração de relatórios e a preparação de datasets.



Exploramos a **Fronteira**: Tecnologias como RAG e disciplinas como MLOps que dependem de dados bem estruturados e governados.

Mensagem Central: O sucesso com IA começa com a maestria dos seus dados.

Seu Próximo Passo: A Prática Leva à Maestria

Chamada para Ação: As ferramentas de IA evoluem rapidamente. Sua vantagem competitiva reside no domínio dos princípios que as alimentam. **Continue praticando.**



Exemplo de Desafio: Você consegue escrever, sem ajuda da IA, uma consulta para encontrar os departamentos, o número de professores e o salário médio, apenas para departamentos com mais de dois professores?

A Solução Envolve: `GROUP BY` e `HAVING`.

```
SELECT DNumero, COUNT(*), AVG(Salario)
FROM PROFESSOR
GROUP BY DNumero
HAVING COUNT(*) > 2;
```

As ferramentas mudam. A lógica e a estrutura não.