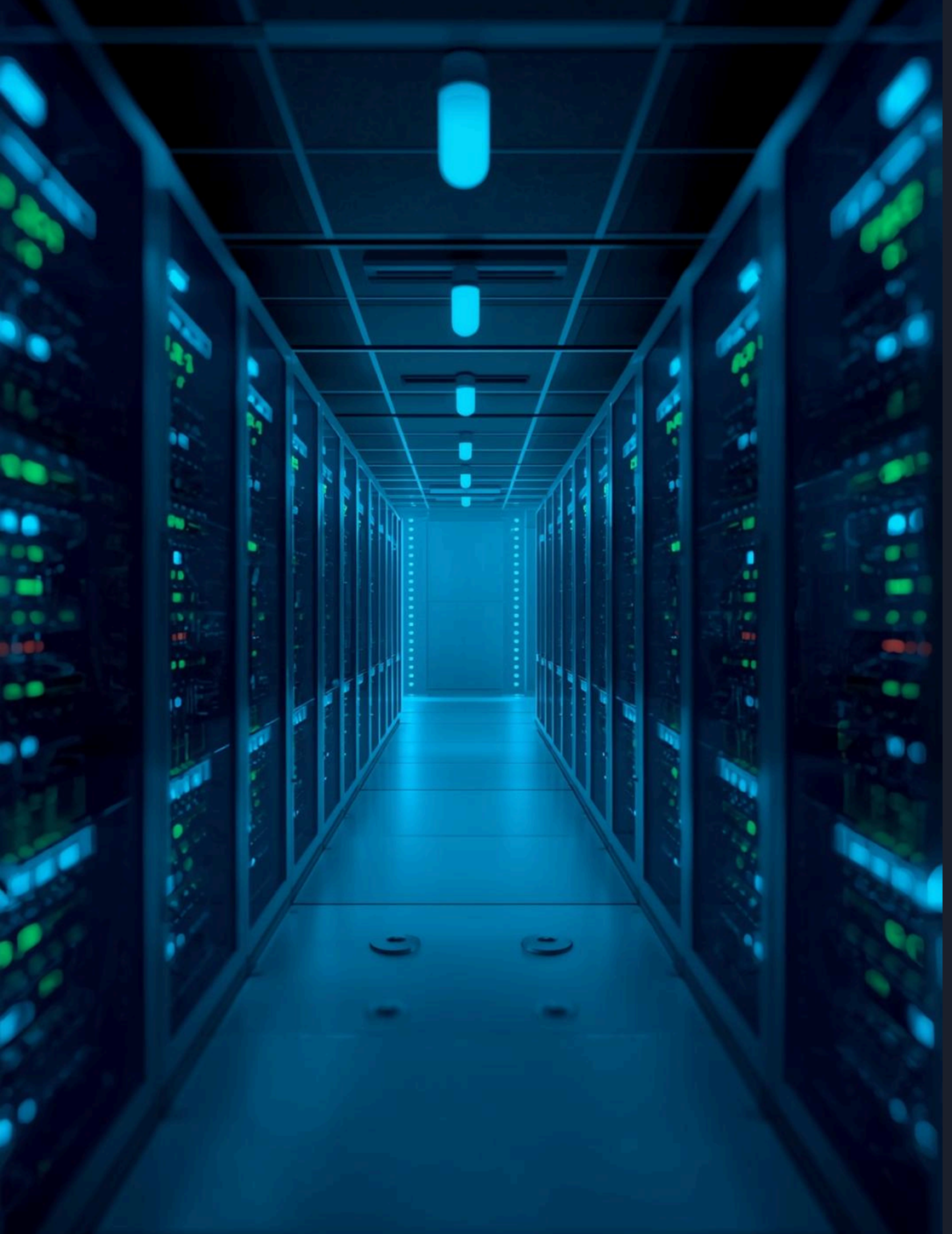
Integração de Dados e ETL

Banco de dados II

Jean Arnhold
Otávio João Maldaner



Roteiro de Apresentação

Principais tópicos abordados

- O que é ETL e por que é usado
- Etapas: extração, transformação, carga
- Ferramentas de ETL
- Desafios
- Importância para Data Warehousing e BI

O que é ETL

- Sigla para Extração, Transformação e Carregamento. Processo de mover dados de diversas fontes, limpá-los e consolidá-los em um Data Warehouse.

- Usado para transformar dados brutos (confusos, incompletos) em informações estruturadas, precisas e confiáveis. Prepara dados para análise, BI e ML.

ETL vs. Integração de Dados

- Integração de Dados (Conceito): Processo geral de combinar fontes para uma visão unificada. É a disciplina ("o quê").
- ETL (Método): Subconjunto da Integração de Dados. Metodologia para carregar Data Warehouses.
- Resumo: ETL é uma forma de fazer Integração de Dados.

Etapas do Processo ETL

Etapa 1: Extração (Extract)

- Processo inicial de coleta de dados.
- Dados brutos são copiados/extraídos das fontes de origem (CRM, mídias sociais).
- Destino: Movidos para uma "área de preparação" (staging area).

Etapa 2: Transformação (Transform)

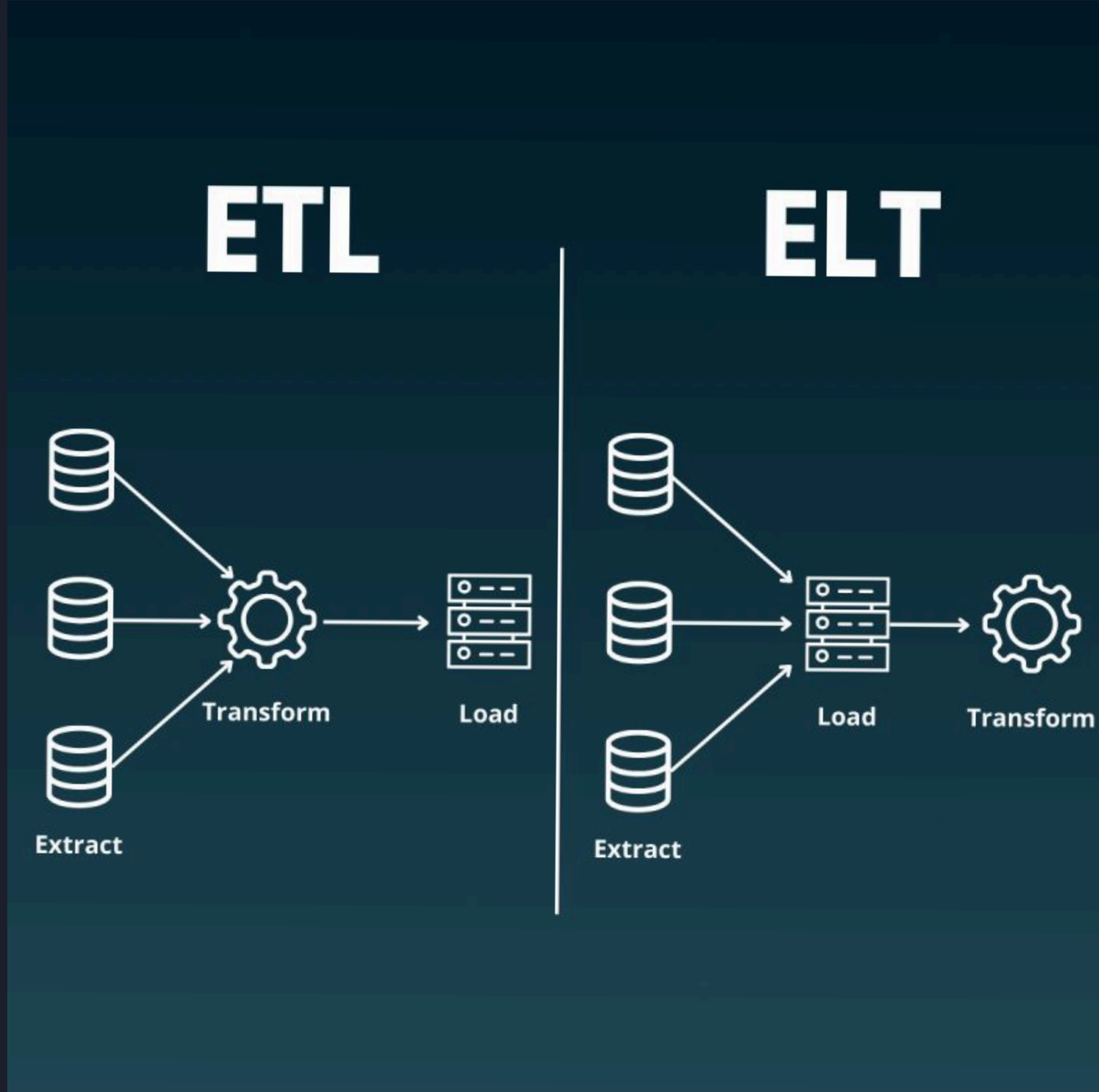
- Etapa crítica. Na staging area, os dados são limpos e organizados para garantir qualidade.
- **Processos:**
  - Limpeza (Corrigir erros, dados ausentes)
  - Padronização (Formatos de datas, moedas)
  - Desduplicação (Remover duplicatas)
  - Regras de Negócio (Aplicar cálculos)

Etapa 3: Carga (Load)

- Última etapa do processo.
- É aonde ocorre o carregamento dos dados limpos e transformados da área de preparação para o **Data Warehouse**.
- Garante que as informações estejam prontas para análise e relatórios.

Evolução

- **ETL (Tradicional):** A transformação (limpeza, regras) acontece em um servidor intermediário (staging) antes dos dados chegarem ao Data Warehouse.
- **ELT (Moderno):** Extrai, Carrega os dados brutos diretamente no Warehouse e, só então, usa o poder do Warehouse para Transformar os dados.
- **Por que ELT é importante:** Mais flexível, escalável e rápido para Big Data.



Ferramentas de ETL

Ferramentas Tradicionais (On-premise)

- Informatica PowerCenter
- Talend
- Pentaho
- IBM DataStage

Ferramentas Modernas (Nuvem)

- Hevo Data
- Fivetran
- Matillion
- Stitch

Desafios

- **Integração de Múltiplas Fontes:** Dados espalhados em sistemas e formatos diferentes. ETL consolida.
- **Qualidade de Dados:** Dados brutos "sujos" (erros, duplicatas). A "transformação" limpa e valida.
- **Escalabilidade:** Volume exponencial de dados.
- **Tempo Real:** ETL clássico é em "lotes" , mas há necessidade de streaming.
- **Silos de Dados:** Novas fontes surgem o tempo todo.

Importância para DW e BI

- **Suporte ao Data Warehousing:** ETL constrói e alimenta o DW. Sem ele, o DW teria dados brutos inutilizáveis.
- **Suporte ao BI:** ETL é "fundamental" para o BI. Garante dados limpos e consolidados para relatórios confiáveis. (Resumo: Não existe BI confiável sem um bom ETL).
- **Além:** Crucial para Data Science e IA.

Referências

Fontes de informação sobre ETL e integração de dados:

https://aws.amazon.com/pt/what-is/etl/
https://hevodata.com/learn/data-integration-vs-etl/