

# TRABALHO DE BIG DATA

Grupo (Os Amostradinhos)

## Conceitos de Data Lake e Data Mesh

- **Data Lake**

Um Data Lake é um repositório de dados centralizado que armazena dados brutos em seu formato original, sejam eles estruturados, semiestruturados ou não estruturados. Ele é escalável, econômico e geralmente baseado em soluções de armazenamento na nuvem.

**Características principais:**

- Flexibilidade: aceita qualquer tipo de dado.
- Escalabilidade: lida com volumes massivos de dados.
- Armazena dados para processamento futuro (não necessariamente organizado).

**Exemplo de uso:** análise de dados, machine learning, big data analytics.

- **Data Mesh**

O Data Mesh é uma abordagem descentralizada para a gestão de dados. Em vez de centralizar tudo em um Data Lake, ele promove a ideia de **produtos de dados** que são gerenciados por equipes independentes, utilizando um modelo federado.

**Características principais:**

- Governança descentralizada.
- Dados tratados como produtos, com equipes responsáveis por qualidade e acesso.
- Foco na escalabilidade organizacional.

**Exemplo de uso:** organizações com múltiplos departamentos, cada um cuidando de seu domínio de dados.

## Comparação: Data Warehouse x Data Lake x Data Mesh

Aspecto	Data Warehouse	Data Lake	Data Mesh
<b>Estrutura</b>	Altamente estruturado (dados organizados).	Flexível (dados brutos e diversos formatos).	Descentralizado (dados como produtos).
<b>Escalabilidade</b>	Limitada e custosa.	Altamente escalável.	Escalabilidade organizacional.
<b>Propósito</b>	Relatórios e BI.	Big data, machine learning, análises ad hoc.	Democratizar e descentralizar o acesso.
<b>Complexidade</b>	Simples de consultar, mas rígido.	Requer mais preparo para análise.	Alta complexidade organizacional.
<b>Exemplo</b>	Empresas focadas em relatórios financeiros.	Startups e big techs lidando com big data.	Grandes empresas com múltiplos domínios.

## Diferenças entre ETL e ELT

- **ETL (Extract, Transform, Load):**
    - Dados são extraídos, transformados e então carregados no destino (geralmente um Data Warehouse).
    - **Vantagens:**
      - Controle rígido sobre os dados.
      - Ideal para dados estruturados.
    - **Desvantagens:**
      - Pode ser lento e custoso para grandes volumes de dados.
  - **ELT (Extract, Load, Transform):**
    - Dados são extraídos, carregados em um destino (como um Data Lake) e transformados posteriormente.
    - **Vantagens:**
      - Processamento mais rápido.
      - Escalável para grandes volumes de dados.
    - **Desvantagens:**
      - Requer ferramentas robustas para transformar dados.
-

## Aplicação no mercado atual

### 1. **Data Warehouse:**

- Amplamente usado em empresas tradicionais para relatórios financeiros e análises de desempenho.
- Exemplos: Bancos, seguradoras e varejistas.

### 2. **Data Lake:**

- Utilizado por startups e big techs para machine learning, análise de grandes volumes de dados ou inovação.
- Exemplos: Amazon, Netflix, Google.

### 3. **Data Mesh:**

- Adotado por grandes empresas que precisam descentralizar a governança de dados, mantendo escalabilidade.
- Exemplos: Organizações globais como Spotify e Zalando.

### 4. **ETL e ELT:**

- ETL é comum em pipelines de Data Warehouse.
- ELT é preferido em ambientes modernos como Data Lakes e plataformas baseadas em nuvem (Snowflake, Databricks).

A escolha entre essas arquiteturas e métodos depende das necessidades da organização, tamanho da equipe e objetivos de longo prazo. Por exemplo, uma empresa de grande porte com diferentes domínios pode optar pelo Data Mesh, enquanto uma startup focada em inovação pode se beneficiar de um Data Lake.