CONCEITOS DE DATA LAKE E DATA MESH

Data Lake

Um Data Lake é um repositório centralizado projetado para armazenar grandes volumes de dados em qualquer formato (estruturados, semi estruturados ou não estruturados) e escala. Ele permite a ingestão de dados brutos diretamente, sem pré-processamento, sendo ideal para aplicações futuras, como análises avançadas e inteligência artificial. Entre suas principais características estão a capacidade de armazenar qualquer tipo de dado e a compatibilidade com ferramentas de análise, machine learning e big data.

Data Mesh

Data Mesh é uma abordagem arquitetural que descentraliza a responsabilidade pelos dados, organizando-os em torno de domínios de negócios. Cada domínio gerencia seus próprios dados, tratados como produtos confiáveis, acessíveis e auto descritivos, enquanto uma governança central define diretrizes gerais. Suas características incluem descentralização, dados como produto, infraestrutura de autoatendimento e governança federada.

DATA WAREHOUSE X DATA LAKE X DATA MESH

- Estrutura

Data Warehouse: Centralizada, utiliza bancos de dados relacionais.
Data Lake: Descentralizada, utiliza armazenamento de objetos.
Data Mesh: Descentralizada, por domínios, com responsabilidade distribuída.

- Processamento

Data Warehouse:ETL (Extração,Transformação, Carregamento).
Data Lake: Carregamento direto, sem transformação.
Data Mesh: Processamento distribuído.

- Foco

Data Warehouse: Análise, relatórios e Business Intelligence.
Data Lake: Armazenamento de grandes volumes, flexibilidade.
Data Mesh: Governança, qualidade e acesso.

Vantagens e Desvantagens

Data Warehouse:
Vantagens: análise fácil, relatórios, segurança
Desvantagens: rigidez, complexidade, custo

Data Lake:
Vantagens: flexibilidade, escalabilidade, baixo custo
Desvantagens: dificuldade de análise, gestão.

Data Mesh:
Vantagens: governança, acesso, flexibilidade
Desvantagens: complexidade, necessidade de coordenação

Quando usar cada um

Data Warehouse: Em análises complexas, relatórios, BI.
Data Lake: Armazenamento de grandes volumes, IoT, machine learning.
Data Mesh: Governança, qualidade, acesso distribuído.

DIFERENÇAS ENTRE ETL E ELT

Os processos de extração, transformação e carregamento (ETL) e de extração, carregamento e transformação (ELT) são duas abordagens de processamento de dados para análise. Elas precisam filtrar, classificar e limpar esse grande volume de dados para torná-lo útil para análise e inteligência de negócios.
A abordagem ETL usa um conjunto de regras de negócios para processar dados de várias fontes antes da integração centralizada. A abordagem ELT carrega os dados como estão e os transforma em um estágio posterior, dependendo do caso de uso e dos requisitos de análise. O processo de ETL requer maior definição no início.

Processo ETL

1. Você extrai dados brutos de várias fontes
2. Você usa um servidor de processamento secundário para transformar esses dados.
3. Você carrega esses dados em um banco de dados de destino.

O estágio de transformação garante a conformidade com os requisitos estruturais do banco de dados de destino. Você só move os dados quando eles são transformados e estão prontos.

Processo ELT

1. Você extrai dados brutos de várias fontes
2. Você o carrega em seu estado natural em um data warehouse ou data lake
3. Você o transforma conforme necessário enquanto está no sistema de destino

Com o ELT, toda a limpeza, transformação e enriquecimento de dados ocorrem dentro do data warehouse. Você pode interagir e transformar os dados brutos quantas vezes forem necessárias.

Cada arquitetura encontra aplicações específicas e complementares:

- **Data Warehouse:** Ideal para análises estruturadas e relatórios gerenciais. É amplamente utilizado em setores como financeiro, varejo e manufatura para:
    - **Relatórios financeiros**
    - **Análises de vendas**
    - **CRM e ERP**
- **Data Lake:** Perfeito para armazenar grandes volumes de dados em formato bruto, permitindo análises exploratórias e machine learning. É utilizado em setores como IoT, saúde e finanças para:
    - **IoT**
    - **Saúde**
    - **Finanças**
- **Data Mesh:** Promove uma abordagem descentralizada e autônoma para a gestão de dados, ideal para empresas grandes e distribuídas. É utilizada em setores como tecnologia, telecomunicações e varejo para:
    - **Acelerar a inovação**
    - **Melhorar a agilidade**
    - **Melhorar a governança de dados**