Conceitos e Comparações de Data Lake, Data Mesh e Data Warehouse

Leonardo Nadolny Magalhães - 01559557

Gabriel Galvão dos Anjos - 01563414

João Victor Freire de Souza - 01525280

Juliana Camilo de Barros - 01532960

Matheus Vitor Fernandes Moreira - 01523037

1. Conceitos de Data Lake e Data Mesh

Data Lake

Data Lake é uma arquitetura de armazenamento projetada para acomodar grandes volumes de dados brutos em seu formato original, sejam eles estruturados, semiestruturados ou não estruturados. Utilizando a abordagem de schema-on-read, os dados só têm sua estrutura definida no momento da consulta, o que proporciona flexibilidade para usos futuros.

Frequentemente implementados em nuvem, como Amazon S3 ou Azure Data Lake Storage, os Data Lakes armazenam dados em formatos naturais, como CSV, JSON e vídeos. Eles são ideais para análises de big data, aprendizado de máquina e inteligência artificial. No entanto, sua eficácia depende de boa governança e organização, sob o risco de se tornarem "Data Swamps".

Vantagens incluem armazenamento escalável e econômico, flexibilidade para múltiplos usos e suporte ao processamento avançado. Porém, é essencial cuidado com segurança e acesso para garantir sua eficiência.

Data Mesh

Data Mesh é uma abordagem descentralizada de arquitetura de dados que visa superar os desafios das arquiteturas centralizadas, como gargalos operacionais e falta de propriedade dos dados. Trata os dados como produtos, promovendo a autonomia de domínios de negócio, onde cada equipe é responsável pela qualidade e aplicabilidade de seus dados. Ele combina governança federada com infraestrutura reutilizável e automatizada para agilizar a entrega de insights.

Baseado em quatro princípios — **Propriedade de Domínio**, **Dados como Produto**, **Infraestrutura como Plataforma** e **Governança Federada** —, o Data Mesh é ideal para organizações grandes e distribuídas, como multinacionais e marketplaces. Essa abordagem permite escalar operações e atender rapidamente às demandas de dados, mas exige maturidade cultural, tecnologia avançada e planejamento rigoroso para ser bem-sucedida.

2. Comparação entre Data Warehouse, Data Lake e Data Mesh

Estrutura de Dados:

- **Data Warehouse:** Estruturado, com *schema-on-write* para consultas rápidas e análises.
- **Data Lake:** Flexível, suporta dados diversos com *schema-on-read*.
- **Data Mesh:** Descentralizado, dados como produtos gerenciados por domínios.

Escalabilidade:

- **Data Warehouse:** Limitada, ideal para volumes menores e estruturados.
- **Data Lake:** Altamente escalável e econômico.
- **Data Mesh:** Escalável, mas exige maturidade cultural e tecnológica.

Facilidade de Uso:

- **Data Warehouse:** Fácil com ferramentas de BI.
- **Data Lake:** Requer habilidades técnicas avançadas.
- **Data Mesh:** Exige alinhamento entre equipes e conhecimento técnico.

Governança:

- **Data Warehouse:** Centralizada.
- **Data Lake:** Flexível, mas arriscado sem boa gestão.
- **Data Mesh:** Federada, equilibrando padrões globais e autonomia.

Custo:

- **Data Warehouse:** Alto custo com licenças e infraestrutura.
- **Data Lake:** Econômico com armazenamento em nuvem.
- **Data Mesh:** Custo inicial alto devido à complexidade.

Vantagens:

- **Data Warehouse:** Estruturas robustas para relatórios precisos e rápidos.
- **Data Lake:** Flexibilidade e custo-benefício para armazenar diversos tipos de dados.
- **Data Mesh:** Promove agilidade organizacional e autonomia das equipes.

Desvantagens:

- **Data Warehouse:** Custo elevado e limitado para dados não estruturados.
- **Data Lake:** Exige habilidades especializadas e pode ter problemas de governança.
- **Data Mesh:** Implementação complexa e requer alta maturidade cultural.

3. Como estas arquiteturas são aplicadas no mercado?

Data Lake:

Os Data Lakes são amplamente utilizados para armazenar grandes volumes de dados brutos que ainda não passaram por transformação ou modelagem. Empresas que lidam com big data, como plataformas de streaming e telecomunicações, aproveitam essa arquitetura para análise de dados de consumo, detecção de tendências e treinamento de modelos de aprendizado de máquina. A flexibilidade dos Data Lakes também permite que dados diversos, como logs, imagens e vídeos, sejam armazenados e processados rapidamente, reduzindo custos e aumentando a escalabilidade

Data Warehouse:

Os Data Warehouses são uma escolha padrão para relatórios corporativos e análises históricas em setores como finanças, varejo e saúde. Com uma arquitetura estruturada, eles permitem que os dados sejam facilmente acessados por ferramentas de BI para criar relatórios detalhados e dashboards. Por serem otimizados para consultas analíticas, os DWs continuam a ser essenciais para empresas que precisam de dados confiáveis e consistentes para suportar decisões estratégicas e operacionais

Data Mesh:

O Data Mesh é mais recente e ainda emergente no mercado. Ele é adotado principalmente por organizações grandes e distribuídas que buscam descentralizar a propriedade dos dados. Empresas que utilizam Data Mesh muitas vezes operam com domínios de dados autônomos, permitindo que equipes específicas gerenciem e criem produtos de dados adaptados às suas necessidades. Essa abordagem é útil para melhorar a escalabilidade organizacional e agilizar a entrega de insights, mas sua implementação requer maturidade organizacional e mudanças culturais significativas

Referência

SERRA, James. *Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh*. 1. ed. Sebastopol: O'Reilly Media, 2024.