UNINASSAU

Big Data

Alan Victor Matos Cerqueira - 01536516
Amanda Gabriela Bernardo da Costa - 01566793
Breno Antônio Alexandrino da Silva - 01280628
Cibele Benício Lourenço - 01535470
Clóvis Vitor dos Santos Tavares - 01564866
Maria Christina Santos Barbosa - 01565538

Recife, Novembro de 2024

Data Lake e Data Mesh

O **Data Lake** é um repositório centralizado que pode armazenar qualquer tipo de dado, ou seja, estruturado, semi estruturado e não estruturado, de forma bruta. Ele permite que os dados sejam mantidos em seu formato original, sem a necessidade de organização prévia, e oferece para suporte vários tipos de análise, como visualizações, processamento de big data, análises em tempo real e modelos de machine learning, promovendo insights que os apoiam decisões mais embasadas.

E a **Data Mesh** é uma arquitetura que facilita o acesso imediato, a proteção e a escalabilidade dos dados. Construída para resolver desafios complexos de segurança de dados através de uma abordagem de propriedade descentralizada e distribuída, com ela, é possível unir e vincular fontes diferentes de dados com base em normas de governança de compartilhamento de dados centralizadas, algo essencial para diversas organizações. Ainda sendo possível, através das funções de negócio, cada área manter o controle sobre o acesso, o formato e a disponibilidade dos dados compartilhados.

Data Warehouse x Data Lake e Data Mesh

As principais diferenças entre Data Warehouse, Data Lake e Data Mesh estão na forma como organizam, gerenciam e utilizam os dados. O **Data Warehouse** é uma solução centralizada que organiza dados estruturados de forma altamente organizada para análises e relatórios. Ele segue um processo rígido de transformação de dados antes do armazenamento, sendo ideal para informações consolidadas e consultas rápidas, mas tem limitações em relação a dados não estruturados e escalabilidade.

Por outro lado, o **Data Lake** é um repositório mais flexível, capaz de armazenar dados em qualquer formato, incluindo estruturados, semiestruturados e não estruturados. Apesar de ser mais econômico e escalável, sua governança é mais complexa, e, sem um bom gerenciamento, pode se transformar em um repositório desorganizado.

Já o **Data Mesh** não é apenas uma tecnologia, mas uma abordagem descentralizada de gerenciamento de dados. Ele distribui a responsabilidade por diferentes equipes ou domínios, tratando os dados como produtos. Essa abordagem é altamente escalável e favorece a colaboração em organizações grandes e complexas, mas requer mudanças significativas na cultura organizacional e uma padronização rigorosa para evitar inconsistências.

Aspecto	Data Warehouse	Data Lake	Data Mesh
Arquitetura	Centralizada	Centralizada	Descentralizada
Formato de Dados	Estruturado	Qualquer formato	Focado no domínio
Custos	Altos	Baixos	Variável
Escalabilidade	Moderada	Alta	Muito alta
Governança	Alta	Baixa (se mal gerido)	Alta (quando bem implementado)
Complexidade	Baixa (bem definida)	Alta (exige gestão)	Alta (requer maturidade)

Diferenças entre ETL e ELT

O que são?

ETL (Extract, Transform, Load) é a metodologia tradicional de processamento de dados. Nela, os dados são extraídos de diversas fontes, transformados para um formato padrão e, em seguida, carregados em um data warehouse ou data mart.

ELT (Extract, Load, Transform), por sua vez, inverte a ordem das etapas de transformação. Os dados são extraídos e carregados diretamente no data warehouse, e a transformação ocorre posteriormente, quando os dados são consultados para análise.

Aspecto	ETL	ELT
Ordem das Etapas	Extrai, transforma e depois carrega.	Extrai, carrega e transforma em seguida, no ambiente de destino.
Tempo de Processamento	Mais lento, devido às transformações realizadas antes do carregamento.	Mais rápido, com transformações ocorrendo após o carregamento.
Flexibilidade	Menos flexível, já que as transformações são pré-definidas.	Mais flexível, permitindo ajustes dinâmicos conforme necessário.
Custo	Pode ser mais caro, exigindo infraestrutura dedicada para processar os dados.	Tende a ser mais econômico, aproveitando o processamento do data warehouse, especialmente em ambientes de nuvem.

Escalabilidade	Menos escalável, com transformações que podem gerar gargalos.	Mais escalável, ideal para grandes volumes de dados processados de forma distribuída.

A escolha entre ETL e ELT depende muito das necessidades específicas de cada empresa. Em ambientes onde o volume e a complexidade dos dados são enormes, o **ELT** tem ganhado destaque pela sua flexibilidade e capacidade de escalar com eficiência. Porém, o **ETL** ainda é uma escolha importante em cenários onde a **qualidade e consistência dos dados** são prioridades e precisam ser asseguradas antes do armazenamento.

Arquiteturas Aplicadas no Mercado de Trabalho

As arquiteturas de **Data Warehouse**, **Data Lake** e **Data Mesh** têm aplicações distintas no mercado, dependendo das necessidades e do contexto de cada organização. O **Data Warehouse** é amplamente utilizado em setores como finanças, seguros e saúde, onde a governança, a precisão e a consistência dos dados são essenciais.

Já o **Data Lake** é mais comum em setores que lidam com grandes volumes de dados, como tecnologia, mídia, streaming e e-commerce. Ele permite o armazenamento de dados em qualquer formato, sendo uma solução flexível para machine learning, análises preditivas e uso em tempo real.

Por sua vez, o **Data Mesh** é uma abordagem emergente, adotada principalmente por grandes organizações com múltiplos domínios de dados, como multinacionais de tecnologia e telecomunicações.

Atualmente, é comum que essas arquiteturas coexistam dentro de uma mesma organização, formando ecossistemas híbridos. Enquanto o Data Warehouse é usado para análises estruturadas, o Data Lake oferece flexibilidade para grandes volumes de dados não estruturados, e o Data Mesh promove a descentralização e a escalabilidade em ambientes complexos. Essa combinação permite que as empresas aproveitem o melhor de cada arquitetura para atender às suas necessidades de dados de forma eficiente e estratégica.