Data Lake

O data lake é um **repositório central de dados, utilizado como uma fonte única e de verdade para a empresa**. De maneira simples, Data Lake trata como um ambiente de armazenamento centralizado para dados brutos e não processados. O Data Lake é um verdadeiro ecossistema de informações. Um ambiente que contém desde arquivos diversos até registros de navegação na web e informações de IoT. O Data Lake visa a agilidade, disponibilizando dados em vários formatos sem estruturação prévia, em contraste com um Data Warehouse (DW) mais estruturado.

De maneira simples, o Data Lake poderia ser comparado a um grande lago, no qual convergem "rios" de dados de diferentes fontes. Todavia, com esse grande fluxo de dados não estruturados, o maior desafio é evitar que o "lago" se torne um "pântano" desorganizado e confuso. Tendo em vista esse desafio, a governança desempenha um papel fundamental. Um catálogo detalhado dos dados, com proprietários definidos, é essencial para manter a ordem. O processo de organização pode ser dividido em camadas progressivas:

- **Transient:** O ponto de entrada dos dados.
- **Raw:** O armazenamento inicial, onde os dados são mantidos em seu estado bruto.
- **Trusted:** Nesta etapa, os dados começam a ser tratados e classificados.
- **Refined:** Dados mais processados e prontos para consumo.

Portanto, podemos dizer que o Data Lake é uma reunião de dados brutos em um repositório centralizado, oferecendo agilidade, mas demandando uma estratégia sólida de governança para garantir usabilidade.

Data Mesh

Data mesh, que também pode ser chamado de malha de dados, é uma abordagem de gestão de dados indicada para o gerenciamento de dados analíticos.

Enquanto o Data Lake centraliza informações, o conceito de Data Mesh introduz uma abordagem descentralizada. No coração do Data Mesh está a ideia de que a responsabilidade pela extração de valor dos dados deve ser compartilhada por toda a empresa. Cada unidade de negócios (domínio) se torna responsável por seus próprios dados, desde sua qualidade até sua governança e disponibilidade. Isso distribui a expertise e a responsabilidade, mas também exige um alto grau de governança para garantir que os padrões sejam mantidos. Para que os dados sejam compreendidos e interpretados em toda a organização, é necessário padronizar metadados, semântica e interfaces. Além disso, o estabelecimento de processos e a automação são fundamentais para garantir que o Data Mesh seja eficaz e coeso.

Data Warehouse x Data Lake x Data Mesh

Data Warehouse: Ideal para empresas que lidam com dados altamente estruturados e exigem análises consistentes, como em relatórios financeiros, painéis de desempenho ou operações baseadas em BI.

Vantagens:

- Alta performance para consultas estruturadas.
- Governança e qualidade de dados bem controladas.

Desvantagens:

- Menos flexível para dados semi ou não estruturados.
- Custos altos para escalabilidade.

Data Lake: Indicado para empresas que precisam armazenar grandes volumes de dados de diversos formatos e realizar análises avançadas (como machine learning e IA), mas sem uma estrutura rígida inicial.

Vantagens:

- Flexibilidade para armazenar e processar dados em qualquer formato.
- Escalabilidade para grandes volumes.

Desvantagens:

- Risco de se tornar um "data swamp" (desorganizado e difícil de usar).
- Requer ferramentas específicas para processamento (e.g., Spark, Presto).

Data Mesh: Apropriado para organizações grandes e complexas, com vários times ou domínios que precisam consumir e gerenciar dados de forma autônoma, como e-commerce ou empresas SaaS.

Vantagens:

- Escalabilidade organizacional com equipes independentes.
- Promove colaboração e entrega rápida de dados úteis.

Desvantagens:

- Difícil implementação inicial devido à necessidade de maturidade organizacional e técnica.
- Depende de uma forte cultura de governança e comunicação entre equipes.

Diferenças entre ETL e ELT

ETL: é uma sigla para Extract, Transform, Load (Extrair, Transformar, Carregar).

- Melhor controle de qualidade: Dados são limpos e estruturados antes de chegar ao destino.
- Maior latência inicial: O processo é mais lento, pois as transformações ocorrem antes do carregamento.
- Menor dependência do destino: Não exige que o sistema de destino tenha alto poder de processamento.
- Custos intermediários: Precisa de ferramentas ou servidores para realizar transformações.

ELT: Por outro lado, representa Extract, Load, Transform (Extrair, Carregar, Transformar). A sua principal diferença é a ordem das operações

- Maior flexibilidade: Os dados brutos ficam disponíveis para várias análises ou transformações futuras.
- Carregamento rápido: Os dados são armazenados imediatamente sem esperar a transformação.
- Dependência do destino: Exige que o sistema de destino tenha alta capacidade de processamento para lidar com as transformações.
- Riscos de qualidade: Dados carregados sem transformação inicial podem ser inconsistentes ou não úteis até serem tratados.

ETL é ideal para organizações que precisam garantir a qualidade dos dados antes do carregamento, têm processos de transformação complexos e exigem alta integridade dos dados.

ELT é mais adequado para empresas que lidam com grandes volumes de dados não estruturados, necessitam de análise em tempo quase real, e possuem data warehouses modernos com alta capacidade de processamento.

ETL e ELT – Uso Mercado Atual

ETL

Cenários Comuns:

- Empresas tradicionais: Usado em setores como finanças, saúde e telecom, onde os dados precisam ser precisos, regulados e altamente estruturados.
- Sistemas de BI e relatórios: Ideal para construir dashboards e análises gerenciais com dados bem organizados.
- Armazenamento em Data Warehouses tradicionais: Ferramentas como Oracle, IBM Db2 e Microsoft SQL Server ainda dependem de ETL.

Ferramentas Populares:

- Informatica PowerCenter, Talend, Pentaho Data Integration, SSIS (SQL Server Integration Services).

Tendências:

- ETL modernizado com serviços na nuvem, como AWS Glue e Google Dataflow, que otimizam a transformação antes do carregamento.

ELT

Cenários Comuns:

- Big Data e análise avançada: Empresas de tecnologia, mídia, e-commerce e IoT que lidam com grandes volumes de dados sem estrutura fixa.
- Data Lakes e Warehouses modernos: Usado em plataformas como Snowflake, Amazon Redshift, Google BigQuery e Azure Synapse, que realizam transformações diretamente no destino.
- Machine Learning e IA: Necessário para análises preditivas e treinamento de modelos com dados não processados.

Ferramentas Populares:

- Apache Spark, Databricks, dbt (Data Build Tool), ferramentas nativas de nuvem como AWS Redshift Spectrum e Google BigQuery.

Tendências:

- Integração com arquiteturas serverless e pipelines como serviços, que permitem carregar e transformar dados em tempo real.