INTEGRAÇÃO DE DADOS I:

DATA LAKE

BIBLIOGRAFIA

- SIMON, Alan. Data Lakes For Dummies. 1st edition. 2021 (livro eletrônico)
- GORELIK, Alex. What Is a Data Lake? 1st edition. 2020 (livro eletrônico)

- material da Profa. Sheila Dias
- material do Prof. Claudiney Ramos

DATA LAKE

- conceito relativamente recente (2010), criado por James Dixon, então CTO (*Chief Technical Officer*) do Pentaho.

- a ideia é ter um único repositório dentro da empresa, para que todos os dados brutos estejam disponíveis a qualquer pessoa que precise fazer análise sobre eles.

https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

DATA LAKE

- é como um reservatório.

https://www.shutterstock.com

  - primeiro você cria a estrutura (um cluster) e depois enche de água (dados);

  - depois que o lago estiver pronto, você começa a usar a água (dados) para várias finalidades, como geração de energia e consumo (análises preditivas, etc.)

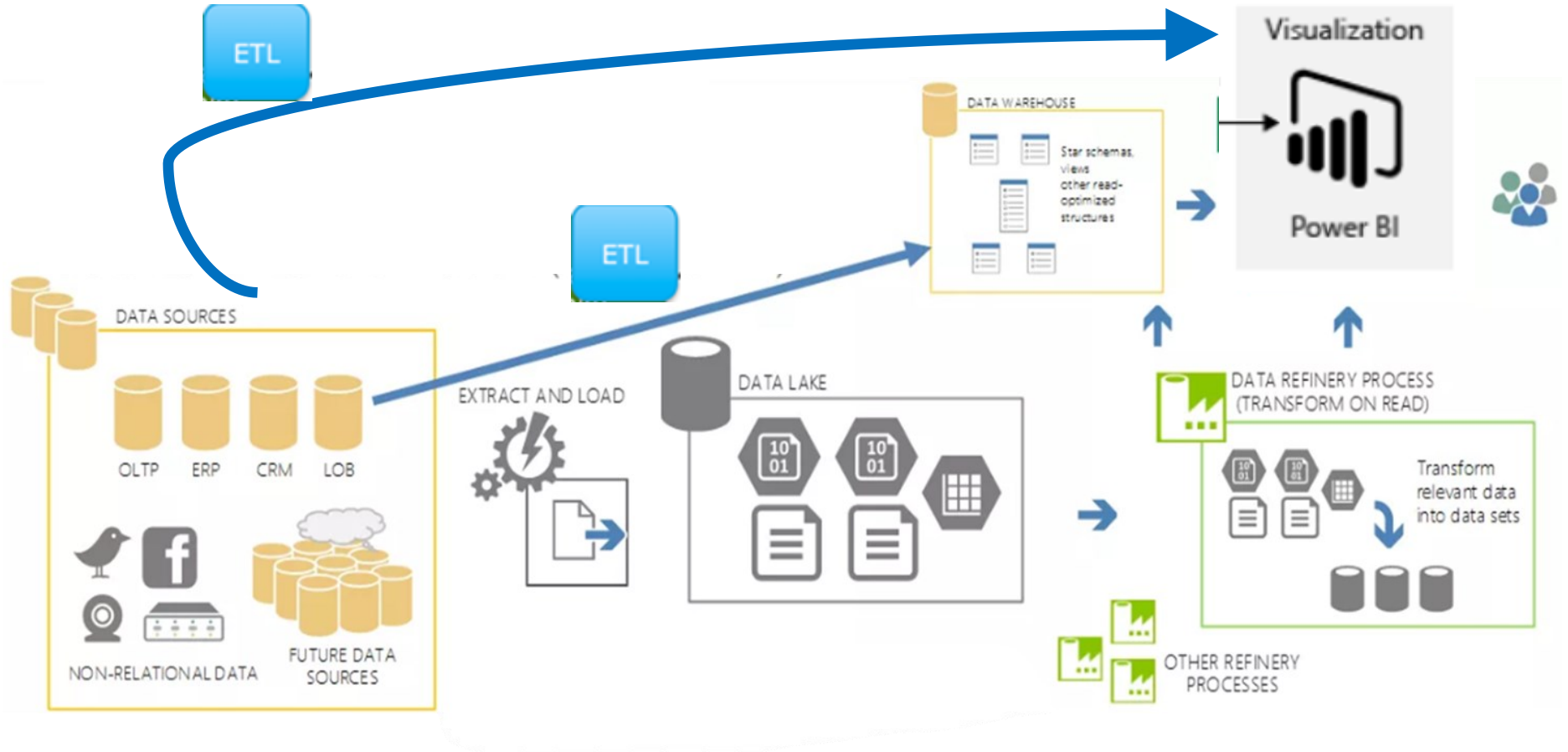
DATA LAKE

- “pode ser definido como armazenamento centralizado, consolidado e persistente de dados brutos, não modelados e não transformados, de múltiplas fontes, sem um esquema pré-definido explícito e sem metadados definidos externamente.”

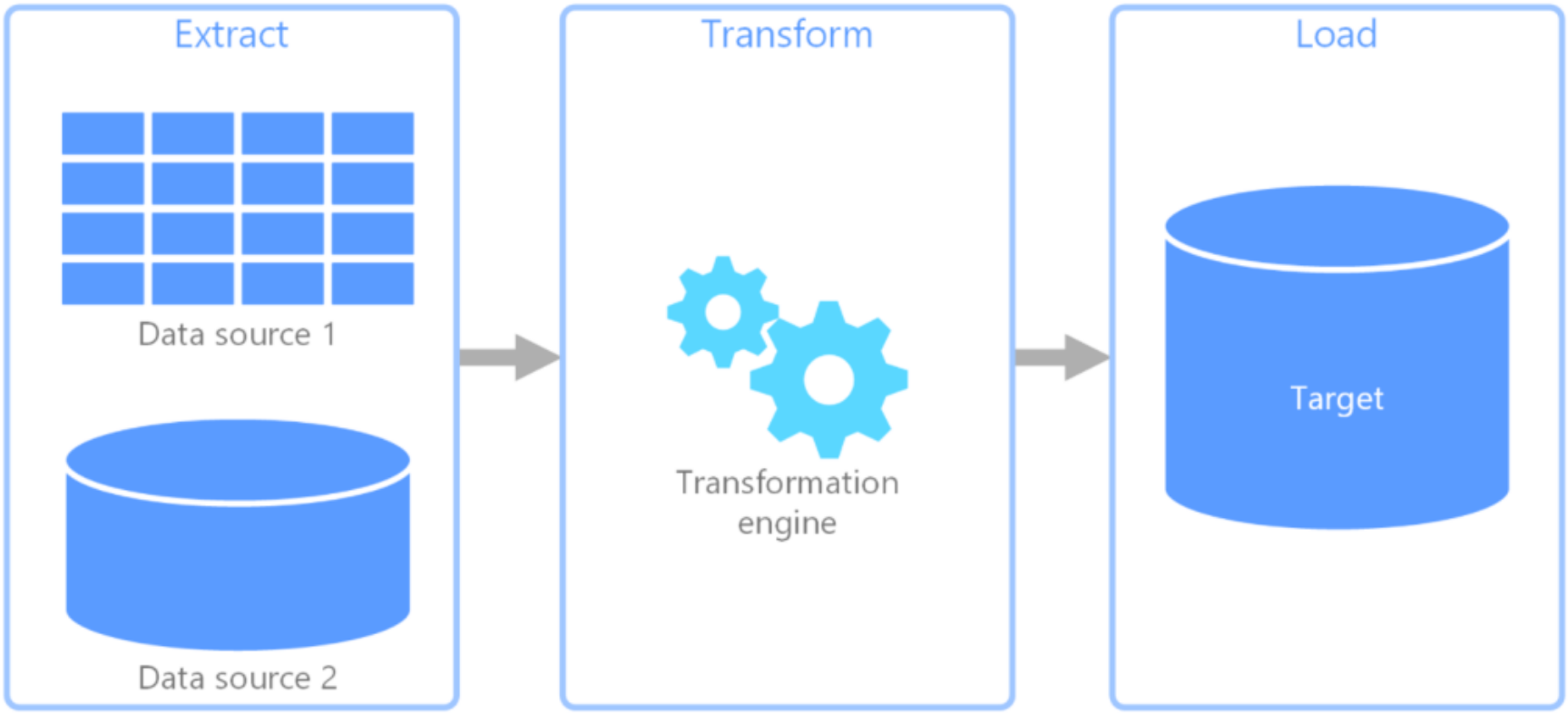
DATA LAKE

- Características:
  - “suporta aquisição de dados de forma ágil;
  - modelo de armazenamento natural para dados complexos e multi estruturados;
  - suporte para computação não relacional eficiente; e
  - fornecimento de armazenamento econômico de grandes e variados conjuntos de dados.”

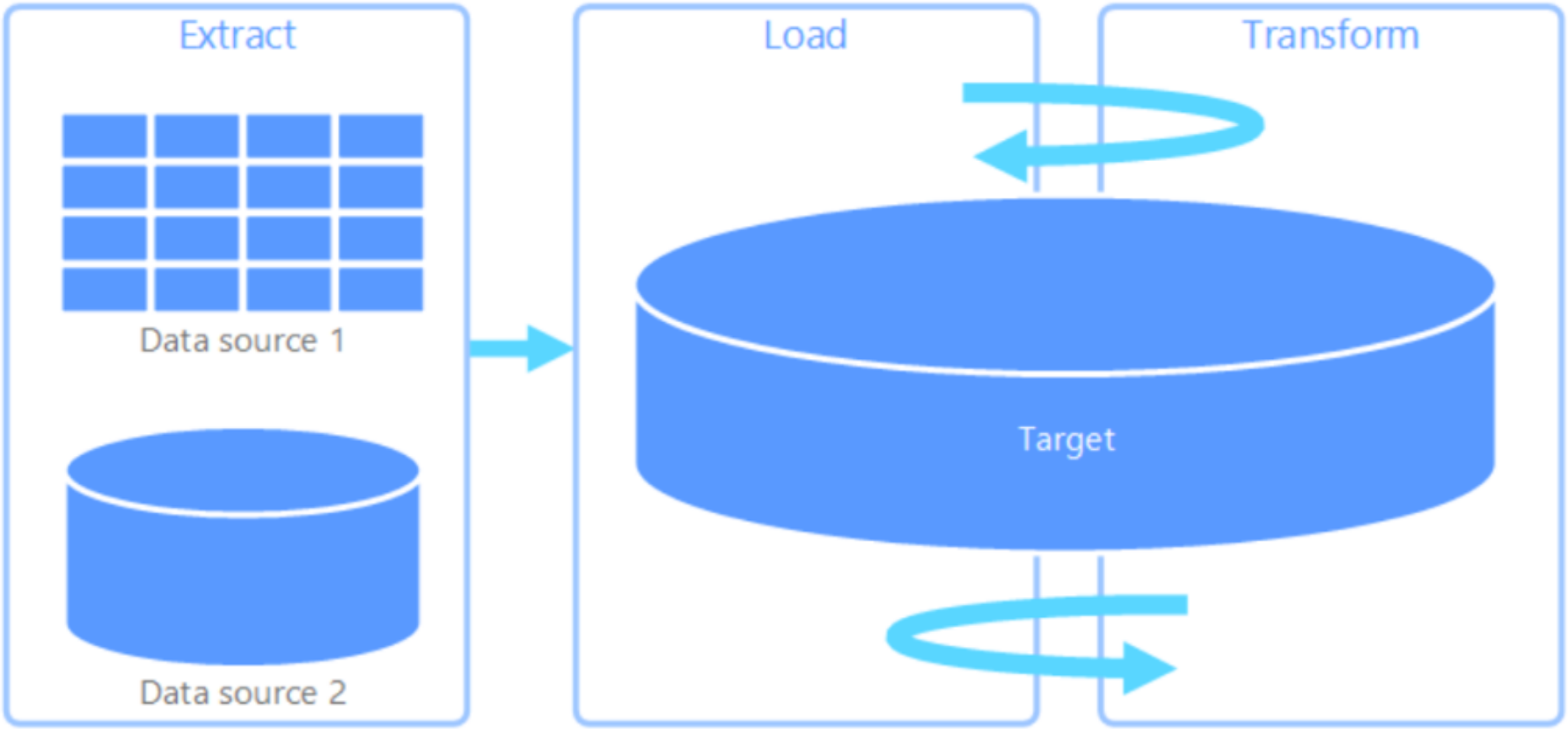
DATA LAKE, ETL, ELT E ELTL



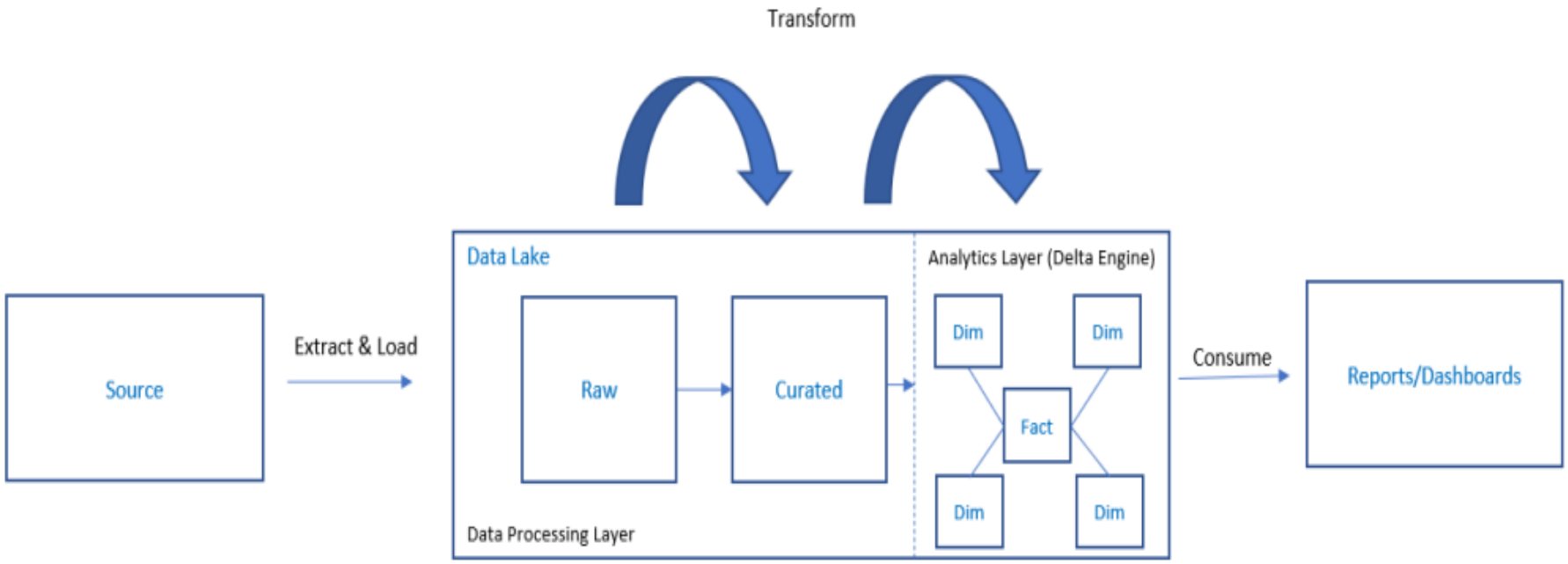
ETL


Extract
Data source 1
Data source 2
Transform
Transformation engine
Load
Target

ELT



ELT – Data Lakehouse



ELTL

- envolve a carga de dados imediatamente em um mecanismo de armazenamento escalável e de baixo custo.

- em seguida, os dados são transformados e novamente carregados em uma camada de apresentação mais avançada.

- é útil quando há uma variedade de fontes de dados que serão utilizadas para uma série de propósitos.

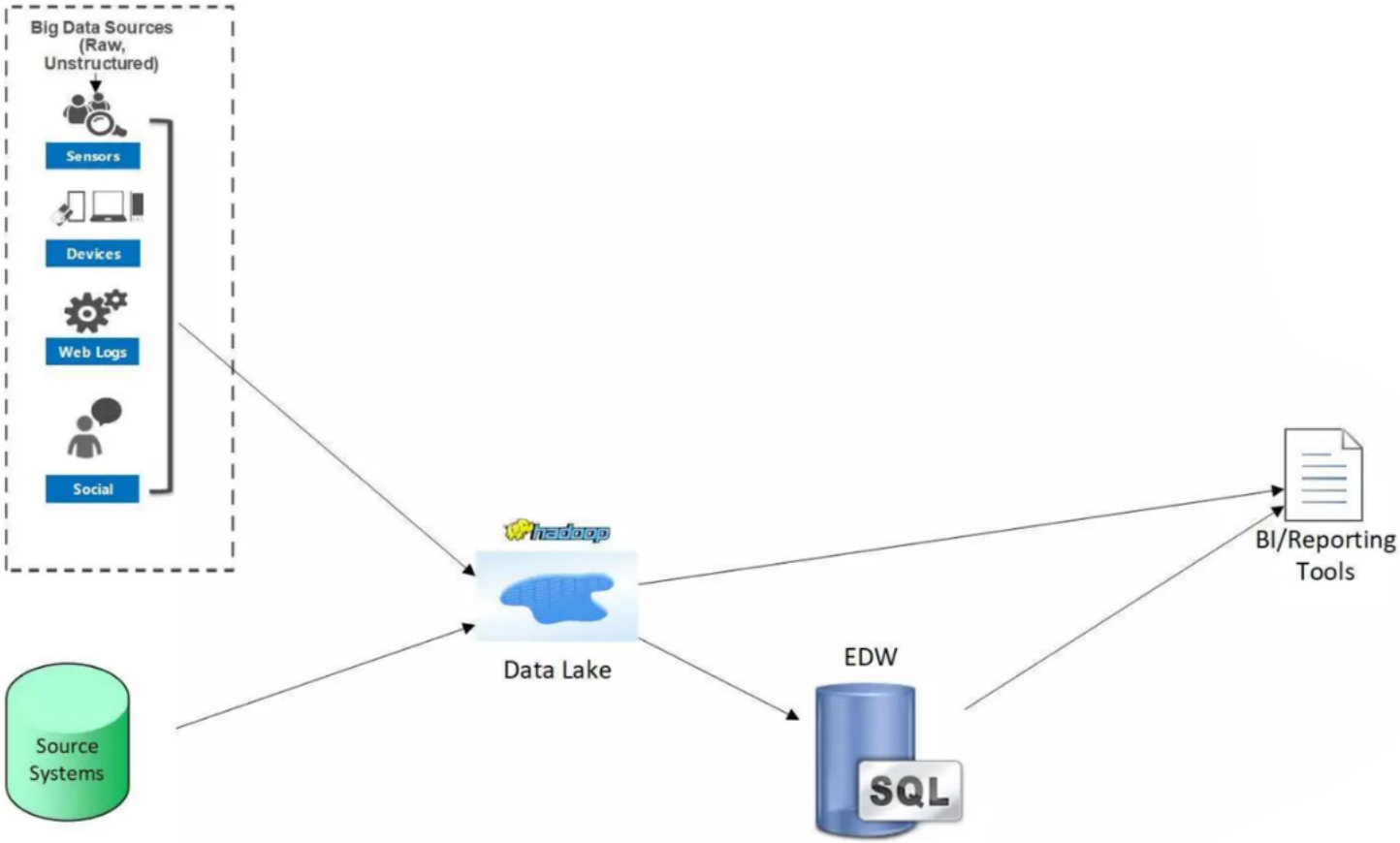
ELTL

- é particularmente útil ao implementar uma solução de integração de dados em uma plataforma de nuvem.

- um Data Lake pode ser estabelecido para propósitos de data science e data discovery, com diferentes segmentos dos dados sendo processados.

Abordagem de Integração	Como funciona	Quando utilizar
ETL	DWs pequenos/médios – as transformações ocorrem "*on the flight*" dentro da ferramenta ETL	• Volumes tipicamente < 1 TB
ELT (com *data warehouse*)	As transformações ocorrem em um DW em nuvem, usando SQL e processamento paralelo	• Processamento no DW corporativo, sem o uso de *data lakes*
ELTL (com *data warehouse* + *data lake*)	Carrega os dados brutos no *data lake*, usa um motor de processamento para fazer as transformações	• Suporte a *reporting*, ML, e *analytics* avançado – volumes de TB, PB e EB

Modern Data Warehouse



Fonte: https://www.slideshare.net/jamserra/data-lake-overview

DATA WAREHOUSE X DATA LAKE

Data Warehouse	Data Lake
Dados Estruturados	Dados não estruturados, semiestruturados, e estruturados.
Esquema definido na escrita	Esquema definido na leitura
BI baseado em SQL	Ciência de dados, análise preditiva, BI
Armazenamento de dados frequentemente acessados, assim como dados agregados e sumarizados	Armazenamento de dados detalhados, brutos e, também, processados
Acoplamento entre o armazenamento e o processamento	Separação entre o armazenamento e o processamento

DATA LAKE

- Governança de Dados
- Qualidade de dados
- Pântano de Dados

DATA WAREHOUSE

- **Vantagens**:

1. **Alta performance:** Estrutura otimizada para consultas complexas e análises rápidas.
2. **Qualidade dos dados:** Dados estruturados e bem organizados garantem consistência e integridade.
3. **Ferramentas de BI:** Integração com diversas ferramentas de Business Intelligence, facilitando a análise e visualização.

- **Desvantagens**

1. **Custo elevado:** Implementação e manutenção podem ser caras.
2. **Flexibilidade limitada:** Difícil de escalar para novos tipos de dados ou grandes volumes não estruturados.
3. **Tempo de preparação:** Processo de ETL (Extract, Transform, Load) pode ser demorado e complexo.

Fonte: Demetrius Magela da Mata (https://www.linkedin.com/pulse/vantagens-e-desvantagens-de-dw-data-lake-lakehouse-em-demetrius-mata-ldgwf)

DATA LAKE

- **Vantagens**:

1. **Flexibilidade:** Capaz de armazenar grandes volumes de dados estruturados e não estruturados.
2. **Custo-benefício:** Geralmente mais barato de implementar e escalar.
3. **Agilidade:** Permite o armazenamento de dados na sua forma bruta, possibilitando diferentes formas de análise posterior.

- **Desvantagens**

1. **Complexidade na gestão de dados:** Sem uma gestão adequada, pode se transformar em um "lago de dados sujos“ (pântano de dados).
2. **Performance de consultas:** Consultas podem ser mais lentas em comparação aos Data Warehouses.
3. **Governança de dados:** Dificuldade em garantir a qualidade e consistência dos dados.

Fonte: Demetrius Magela da Mata (https://www.linkedin.com/pulse/vantagens-e-desvantagens-de-dw-data-lake-lakehouse-em-demetrius-mata-ldgwf)

DATA LAKEHOUSE

- **Vantagens**:

1. **União de benefícios:** Combina a performance analítica dos DW com a flexibilidade dos Data Lakes.
2. **Economia de custos:** Redução de custos ao consolidar infraestrutura de DW e Data Lake.
3. **Escalabilidade:** Facilmente escalável para grandes volumes de dados estruturados e não estruturados.

- **Desvantagens**

1. **Complexidade de implementação:** Requer um planejamento detalhado e expertise técnica para configurar corretamente.
2. **Tecnologia emergente:** Ainda em evolução, podendo apresentar desafios de maturidade e suporte.
3. **Custo inicial:** Investimento inicial pode ser alto devido à complexidade e necessidade de novas ferramentas.

Fonte: Demetrius Magela da Mata (https://www.linkedin.com/pulse/vantagens-e-desvantagens-de-dw-data-lake-lakehouse-em-demetrius-mata-ldgwf)

Dúvidas, Perguntas ou Sugestões?