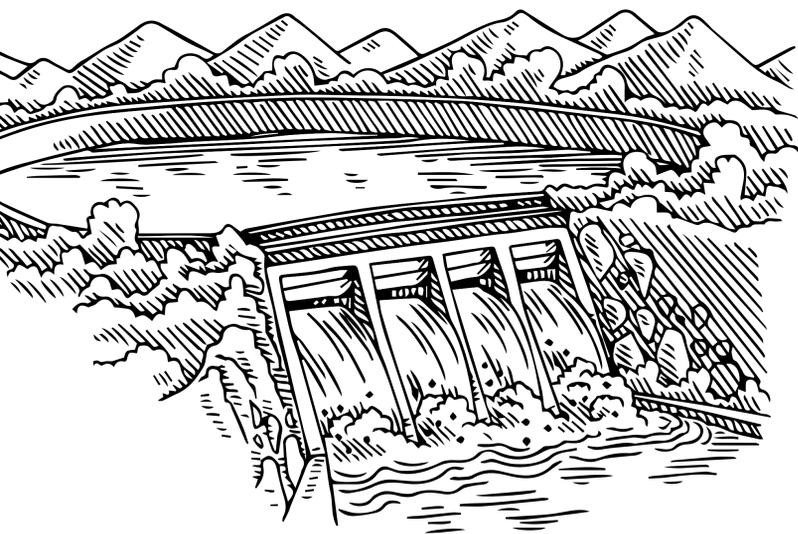


Data Lake

2010

História

Termo criado em 2010 por **James Dixon**, CTO da Pentaho



O QUE É?

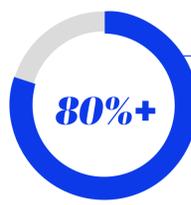
Data Lake (ou “lago de dados”) é uma **arquitetura de dados** que funciona como um repositório para integrar e disponibilizar dados brutos nas organizações. Ela resolve o **desafio de gerir e gerar valor de grandes volumes de dados**, vindos de fontes diversas – estruturados e não-estruturados – em velocidade alta e variável.

Dados Não estruturados

- ❶ Estrutura flexível, representação heterogênea.
- ❷ Ex.: XML, redes sociais, artigos, imagens, APIs, etc.

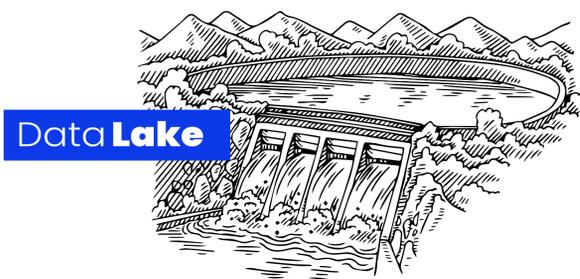
Dados Estruturados

- ❶ Estrutura rígida, projetada previamente, representação homogênea
- ❷ Ex.: planilhas, bases de dados relacionais, etc.



Estima-se que 80% ou mais dos dados gerados nas organizações sejam **não estruturados**.

LAKE vs. WAREHOUSE ?



Data Lake

- ❶ Ingera dados brutos, **estruturados, semiestruturados e não estruturados**, de diversas fontes e em qualquer escala.
- ❷ Mantém **centralizados** dados que existiam em silos.
- ❸ Conecta-se à ferramentas de **analytics e machine learning**.
- ❹ Movimenta os dados em **tempo real**.
- ❺ Estruturas mais **complexas, modulares que permitem combinar tecnologias com melhor custo-benefício**.
- ❻ Ideais para organizações que buscam criar uma cultura data driven, gerando valor a partir de **engenharia e ciência de dados**.



Data Warehouse

- ❶ Armazena de forma **centralizada dados estruturados de bases relacionais e de sistemas**, como CRM, ERPs entre outros.
- ❷ Ideais para demandas pontuais de **geração de relatórios para business users**.
- ❸ Alta performance com **custo alto**.

GANHOS



Viabiliza o uso da **Inteligência Artificial** para realização de classificações, previsões e análises complexas sobre os dados.



Estrutura central que **alimenta todas as áreas** de negócio e de analytics.



Por centralizar e integrar os dados, o data lake permite que as áreas de negócio façam o uso das **análise mais adequadas aos seus próprios desafios**.

DESAFIOS



Data Lakes exigem **governança e manutenção contínuas** para que os dados possam ser usados e acessados por **todas as áreas de negócio**. Sem esse controle, há o risco deles se tornarem **pântanos**, ou **data swamps**.

DESAFIOS DE CONSTRUÇÃO

Garantir um bom “time-to-value” ou que os dados possam rapidamente ser integrados com fluxos e ferramentas das áreas de negócio.

Visualizar de forma clara a governança: de onde os dados surgem, onde são consumidos, quem zela por eles.

Padronizar a nomenclatura dos dados para toda a organização.

DESAFIOS DE MANUTENÇÃO

Armazenar apenas os dados que **realmente entreguem valor**.

Promover o **uso pelas áreas de negócio**.

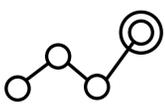
Gerenciar **custo e processamento** conforme a entrada de novos dados.

Garantir que o dado esteja **sempre disponível**.

O QUE DATA LAKES TÊM A VER COM CULTURA DATA DRIVEN?



Democratiza a informação como alavanca para autonomia e agilidade no nível organizacional;



Maior alinhamento entre a estratégia de negócio com a evolução da operação;



Fluxo automatizado de tomada de decisão.

Quer saber mais?



Video
Desbuçando a Mente Data Lake

bit.ly/3yRRDgi



Artigo
A Smarter Way to jump into data lakes
McKinsey Digital

mck.co/2pzCxfm