

<https://delta.io/>

Data Lakehouse na prática

INTRODUÇÃO AO CONCEITO, DISCUSSÃO SOBRE O
TEMA E APRESENTAÇÃO DA DEMO

Apresentador



LEANDRO HUMBERTO

7 anos de experiência em dados

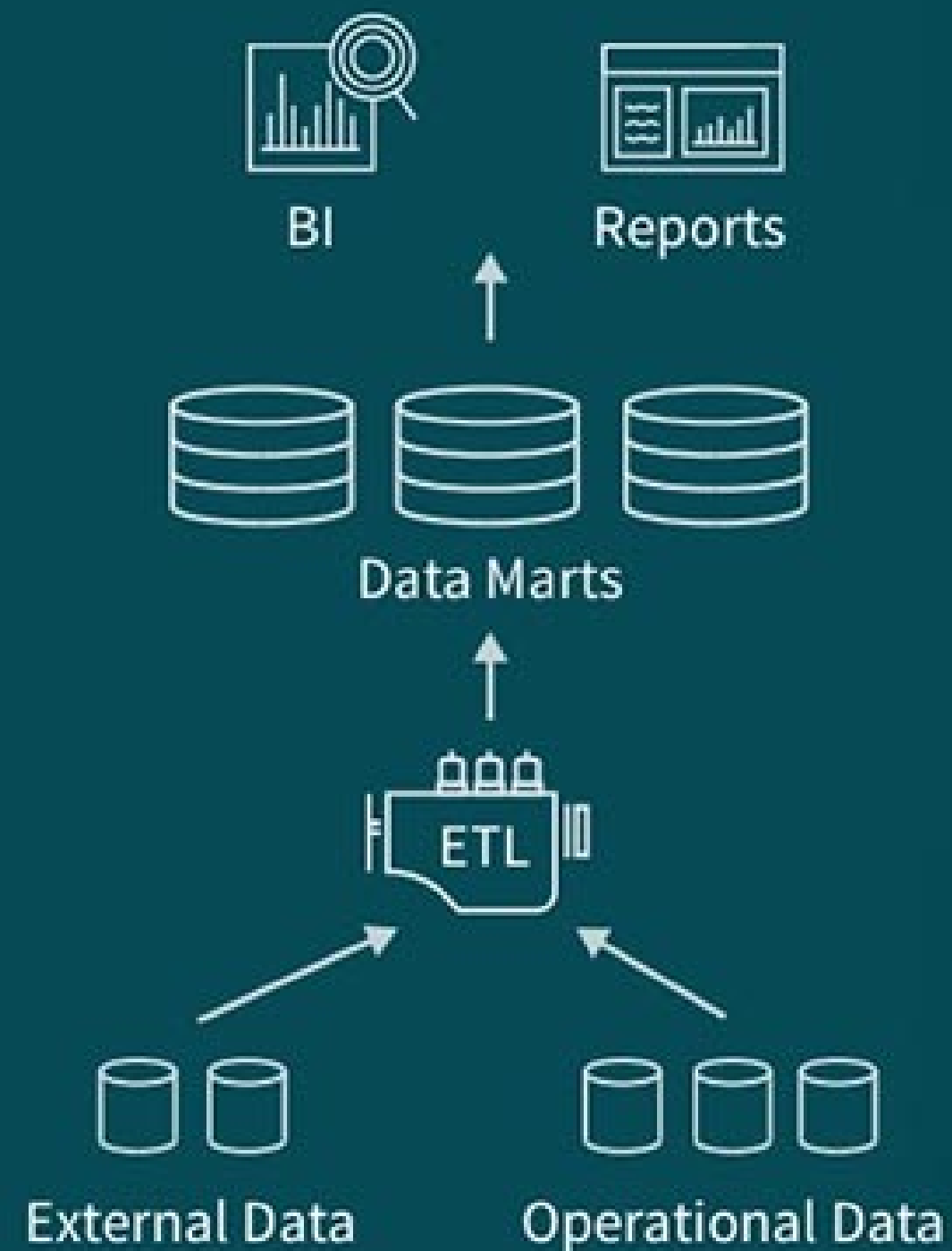
Google Professional Data Engineer

Especialista em dados @ BBTS

Agenda

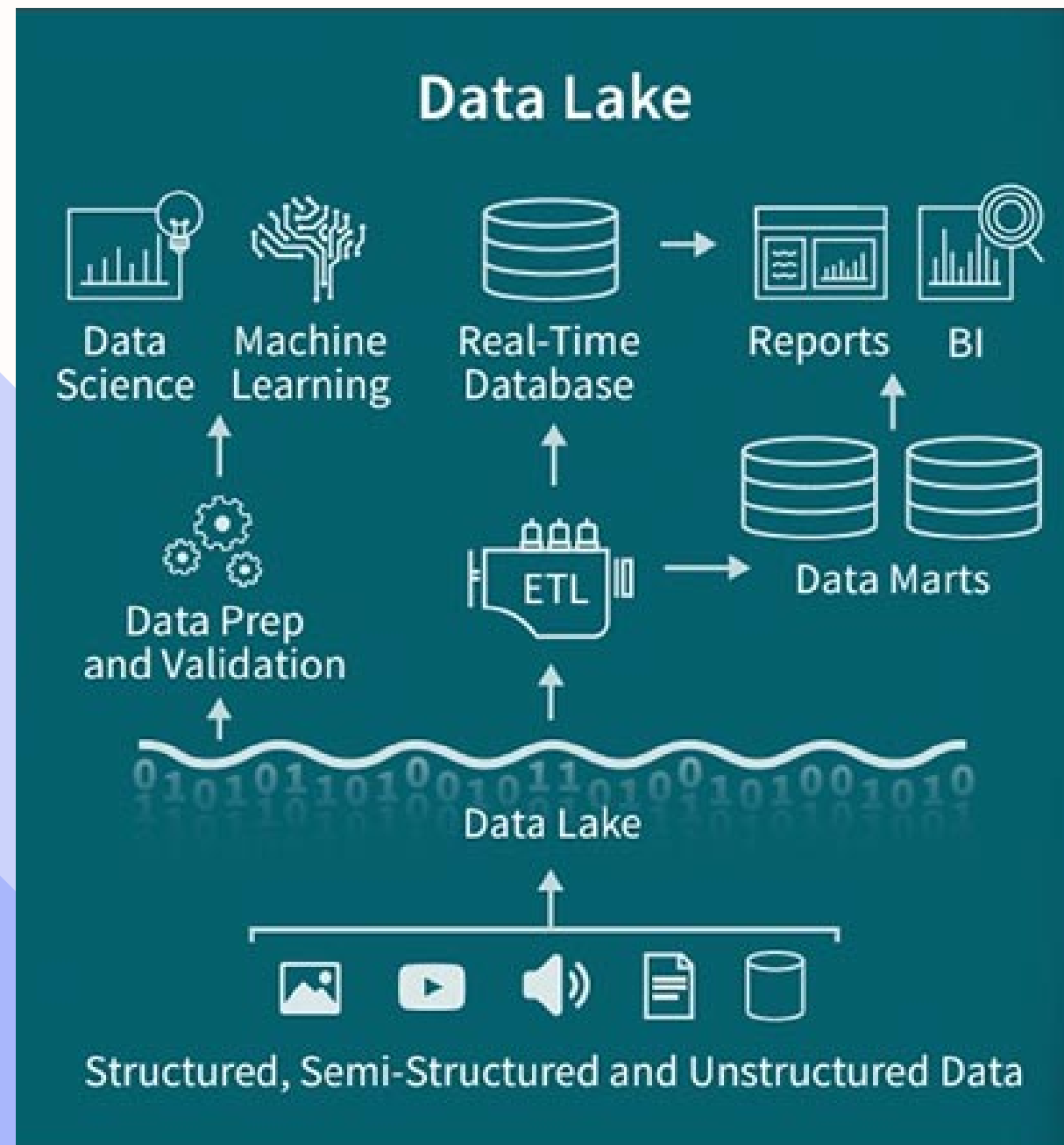
- CONTEXTO E PROBLEMA
- DATA LAKE VS DATA WAREHOUSE
- DATA LAKEHOUSE
- DELTA LAKE
- DEMO
- IMPRESSÕES
- DÚVIDAS

Data Warehouse



DATA WAREHOUSE

- Usados desde a década de 80
- Geração de métricas
- Dados transformados por ETL
- Construído em bases relacionais



DATA LAKE

- Necessidade de um repositório central
- Big Data Analytics
- Cloud Computing
- Dados carregados por ELT
- Construído em sistemas de arquivos, sendo o mais comum o HDFS



DW

Dados de alta
qualidade

SQL

Schema Enforce

Ecossistema
maduro

DL vs DW



DATA LAKE

Facilidade de
ingestão

Alto volume
e baixo custo

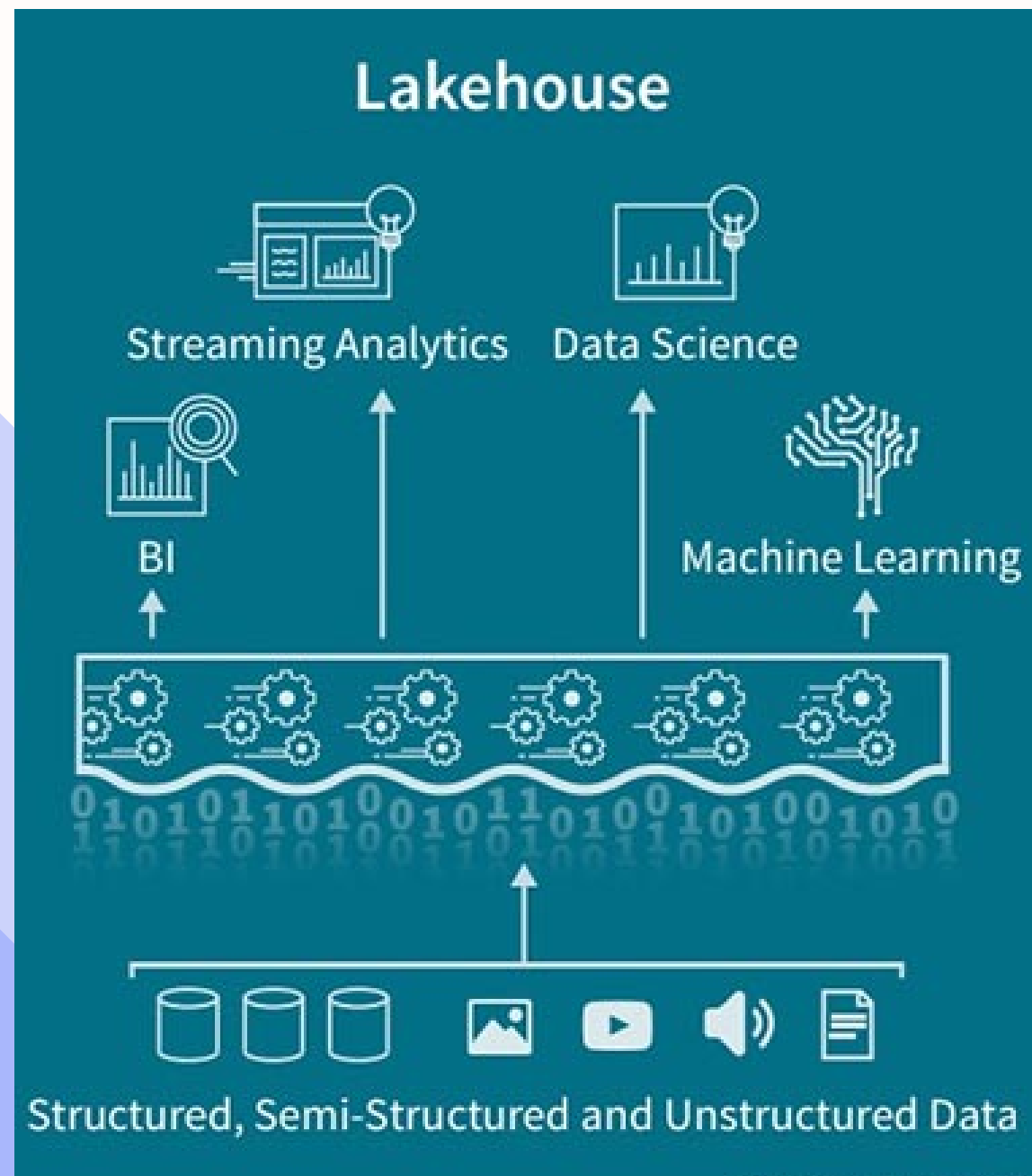
Democratização
dos dados

Problema

"Já tenho um DW e agora tenho que manter duas estruturas para fazer analytics?"

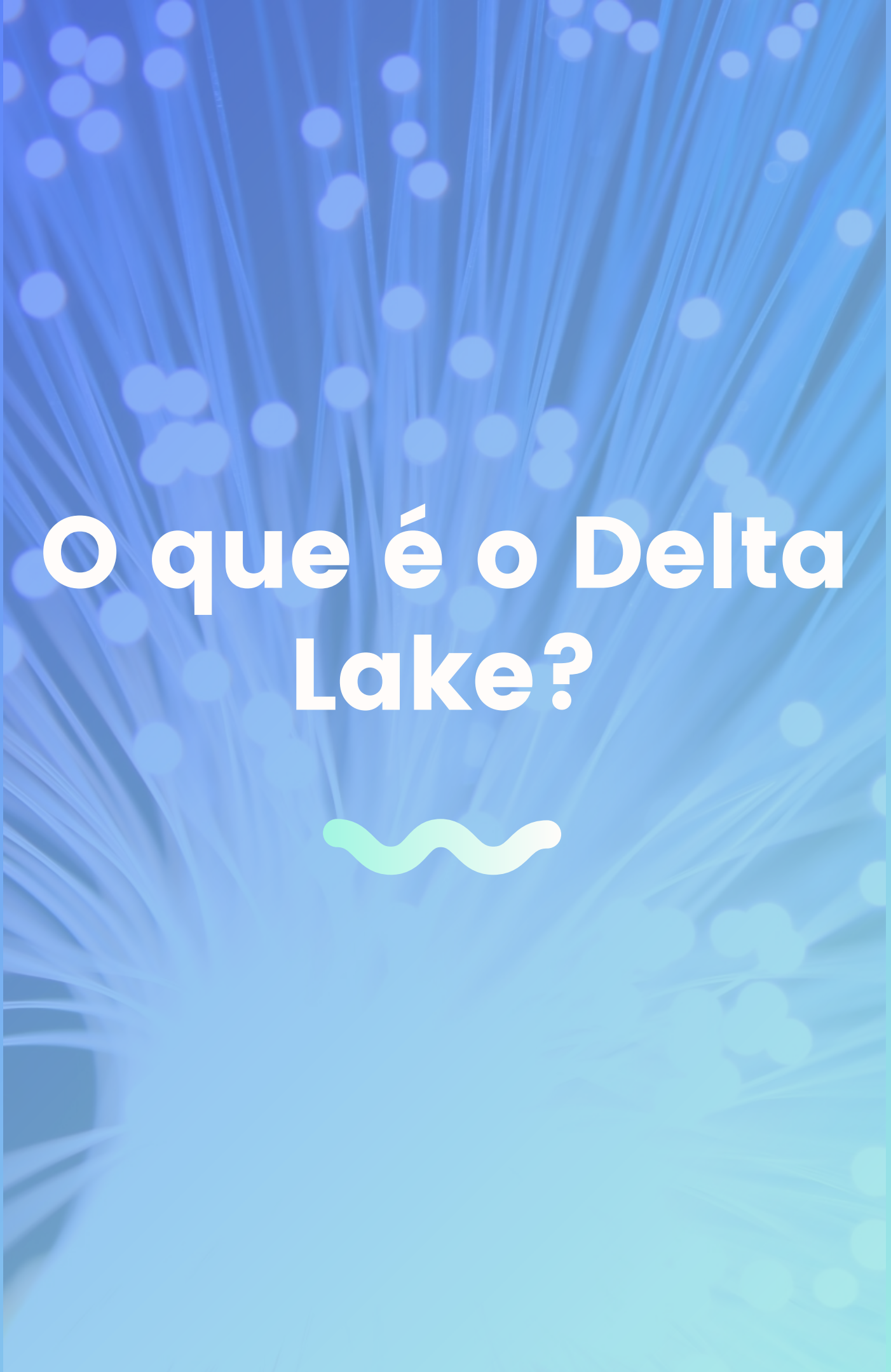


"Comprei um cluster enorme para meu Data Lake, agora tenho que adquirir um DW também?"



DATA LAKEHOUSE

- Implementa os benefícios do Data Warehouse sobre o Data Lake
- Uso do mesmo compute para diversos workloads
- Plataforma homogênea, sem problemas de integração
- "Junção das vantagens"



O que é o Delta Lake?



"DELTA LAKE É UMA CAMADA DE ARMAZENAMENTO QUE TRAZ TRANSAÇÕES ACID PARA O APACHE SPARK" – [DELTA.IO](https://delta.io)



ACID TRANSACTIONS

Atomicidade,
Consistência,
Isolamento,
Durabilidade



RUN ON SPARK

Usa um dos frameworks
de big data mais
adotados do mercado



TIME TRAVEL

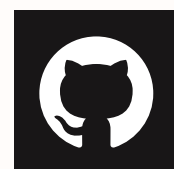
É possível realizar um
rollback da tabela para
versões anteriores



Delta Lake Features

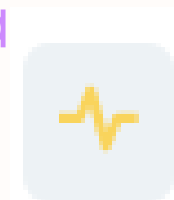
OPEN SOURCE

Pode ser usado no seu
Data Center, acoplado
no Spark OSS



BATCH E STREAMING

Log de transações
permite concorrência

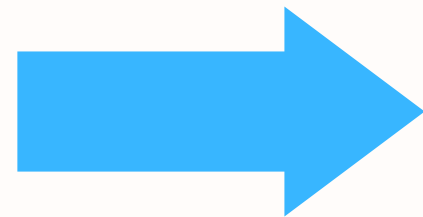


UPDATES E DELETES

Permite operações de
banco de dados sobre
os arquivos



Demo



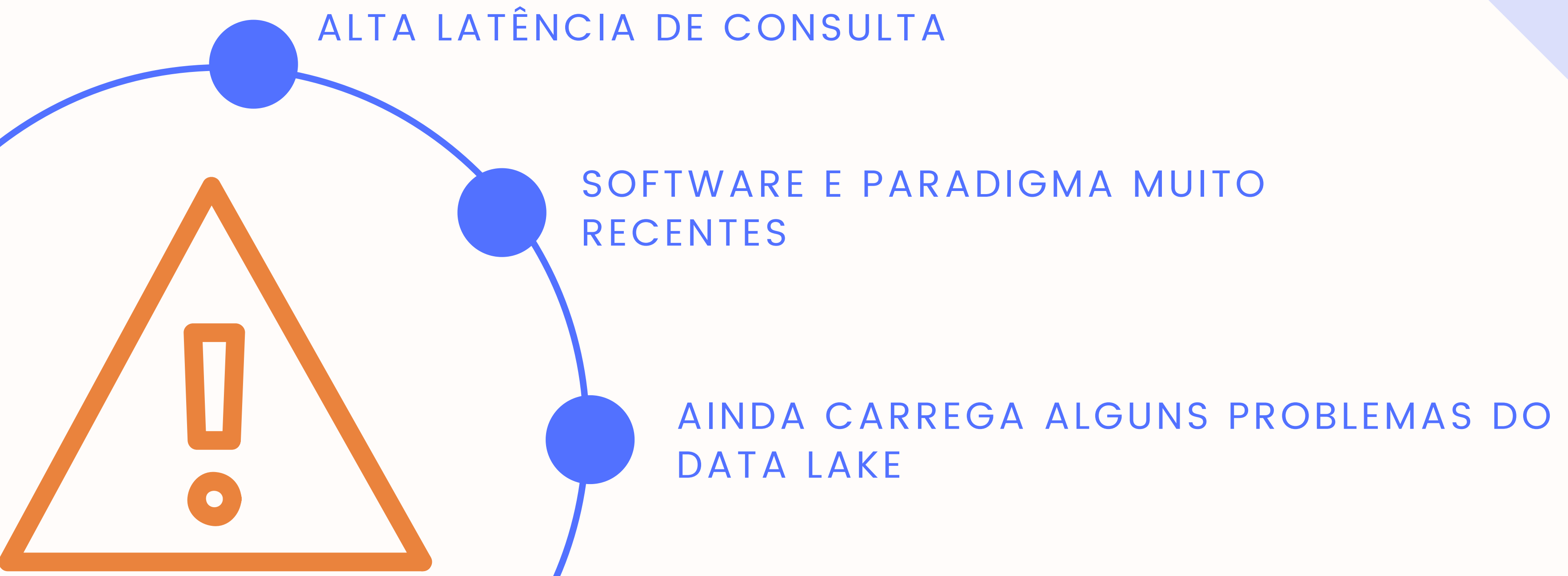
fat_chamada



dim_operador



Impressões



Dúvidas?

LEANDRO HUMBERTO

leandrohmvieira@gmail.com

CONTATO

(62) 981 936 680



/IN/BEDATADRIVEN/



@PROFISSIONALDEDADOS



/LEANDROHMOVIEIRA

Referências



<http://www.cienciaedados.com/do-data-warehouse-para-o-data-lake/>

<https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html>

<https://www.zdnet.com/article/a-standard-for-storing-big-data-apache-spark-creators-release-open-source-delta-lake/>

<https://docs.delta.io/latest/delta-intro.html>

Para mais detalhes acesse:

<https://www.slideshare.net/databricks/designing-etl-pipelines-with-structured-streaming-and-delta-lakehow-to-architect-things-right>

https://www.youtube.com/watch?v=eOhAzjf_iQ