10 de junho

1º TECH CONNECTION

Balneário Camboriú

TALK FLORIPA

Realização

TALK FLORIPA

Apoio

unialyan

INTELIGÊNCIA ACADÊMICA

Patrocinadores

DATA SIDE

CDB Data Solutions

power tuning

API4COM

cd2 RETAIL TECH

Sinapses Smart Office

Oberdan Schaider
Formado em Engenharia de Computação;
MCSE: Data Management and Analytics;
Microsoft Certified: Azure Database Administrator;
MLSA: Microsoft Learn Student Ambassadors;
DBA na empresa CDB Data Solutions.

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

TALK FLORIPA

Luciano Gambato
Arquiteto de Dados na CDB - Data Solutions;
Formado em Engenharia de Computação;
Certificações Azure Data/AI Engineer;
MCSE: Data Management and Analytics.

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

TALK FLORIPA

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

TALK FLORIPA

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

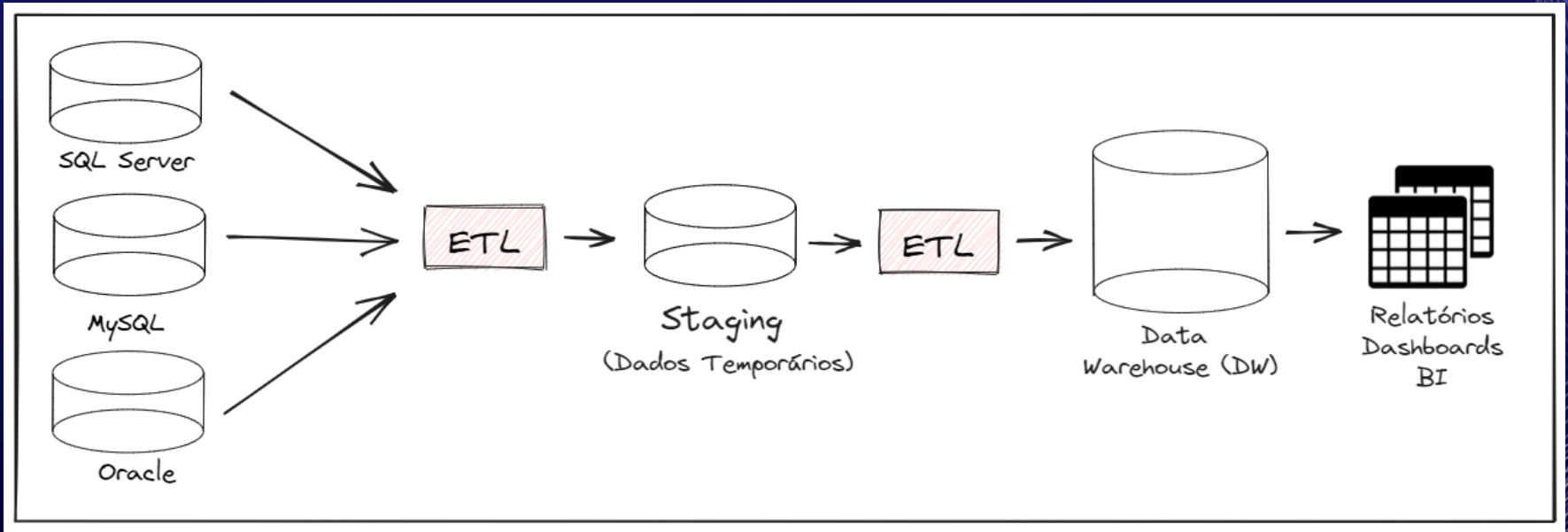
Data Warehouse (DW)

- Conceito surgiu por volta de 1980;
- Repositório de dados central, geralmente com volume de GB ou TB;
- Construído em SGBDs relacionais (SQL Server, PostgreSQL, Oracle etc.);
- Dados armazenados com viés das áreas de negócio;
- Camada de Staging (dados transientes);
- ETL (*Extract – Transform – Load*);
- Visão histórica;
- Data Mart;
- Fatos e Dimensões;
- *Schema on Write;*

- **Pontos positivos:**
  - Consultas performáticas;
  - Fonte da verdade;

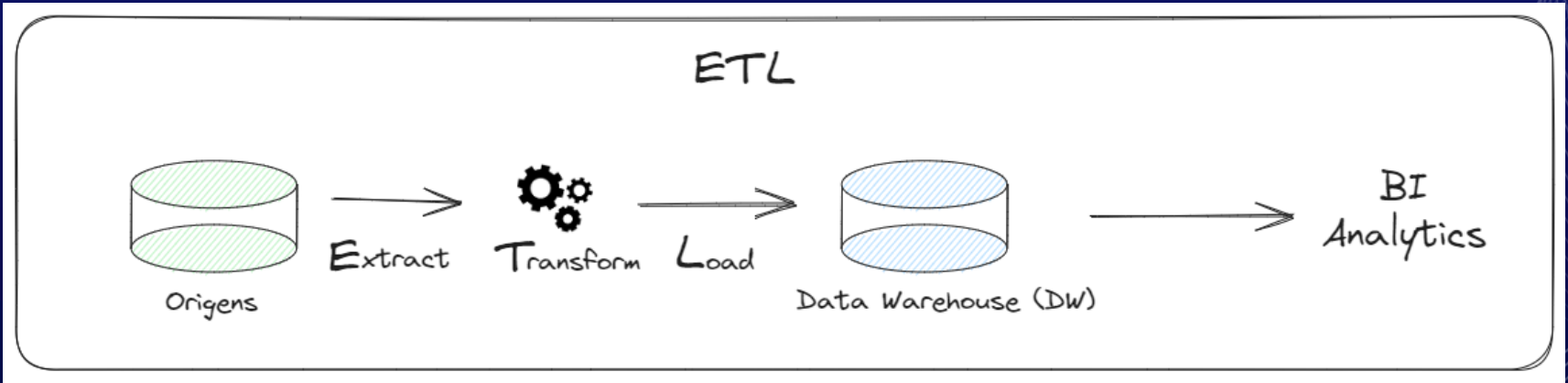
- **Pontos negativos:**
  - Custo operacional alto para criação e ajuste na modelagem e integrações (ETL);
  - Dados desatualizados.

Data Warehouse (DW)

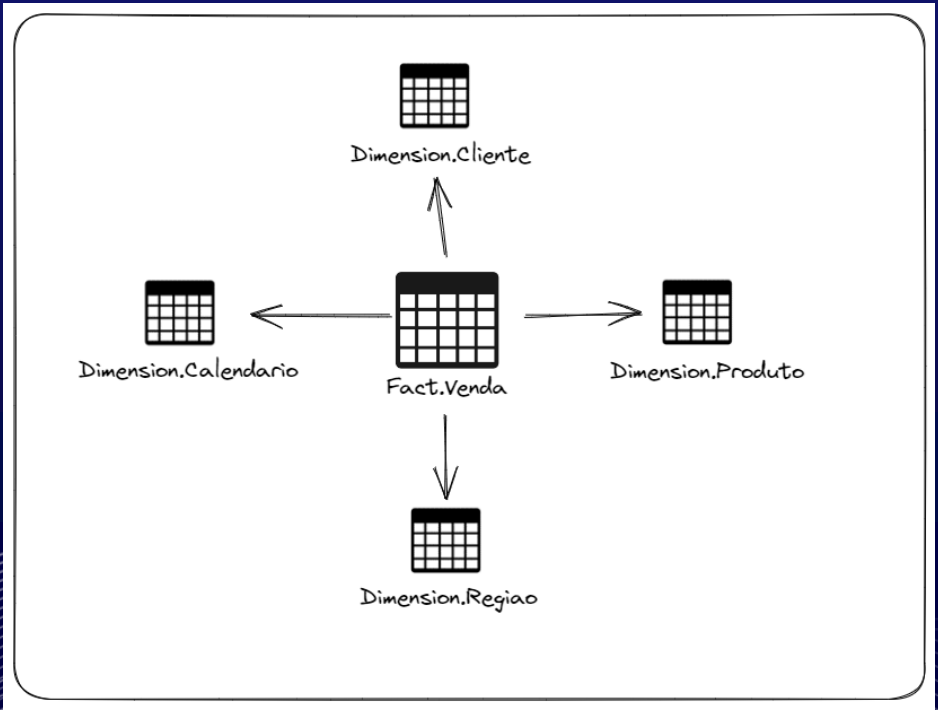


Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

ETL

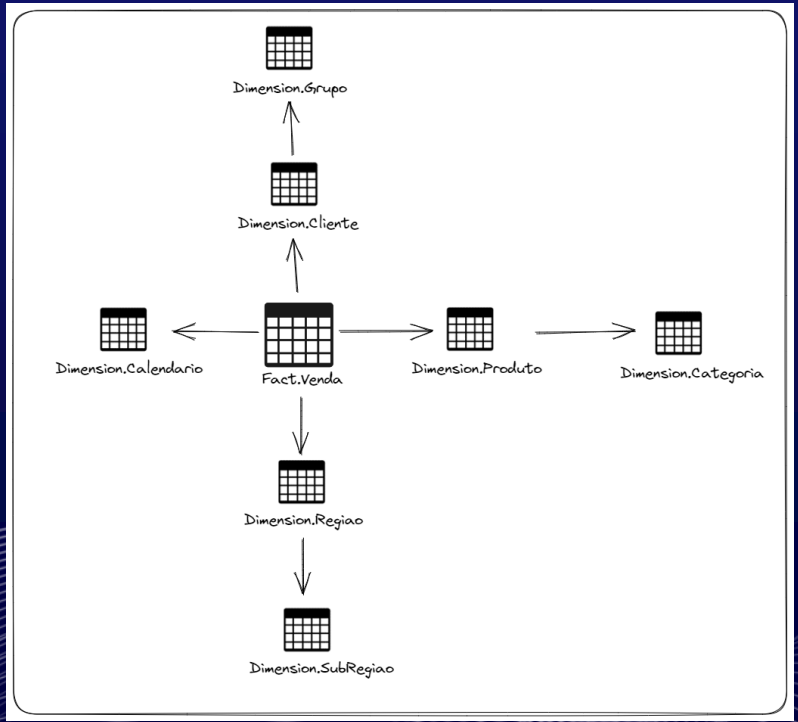


Star Schema


Dimension.Cliente
Dimension.Calendario
Fact.Venda
Dimension.Produto
Dimension.Regiao

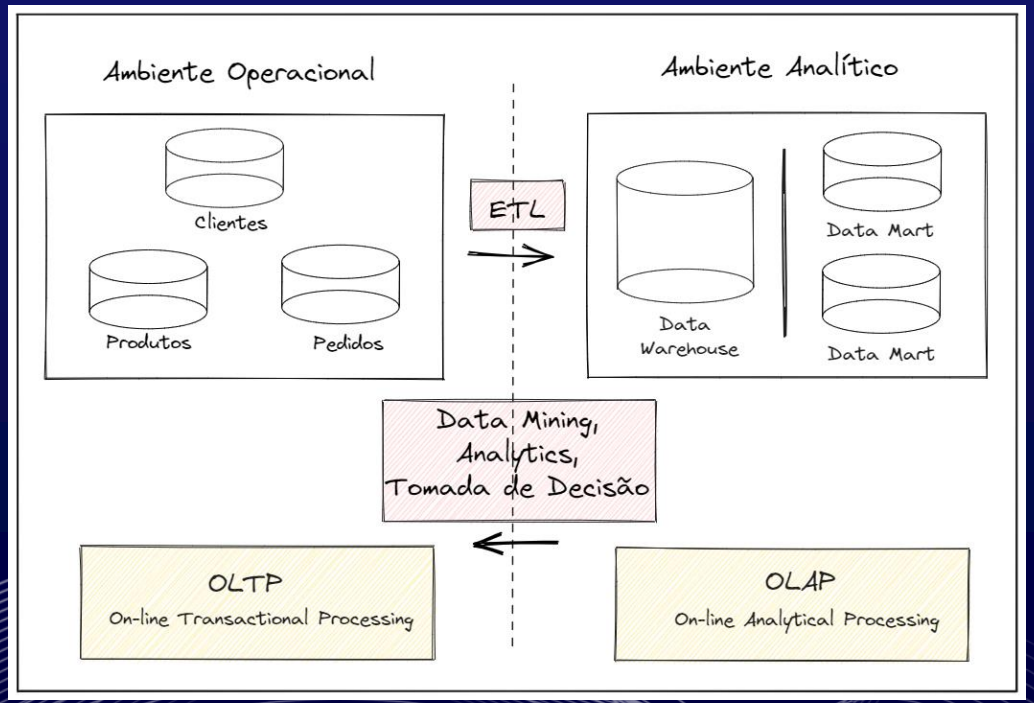
Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Snowflake Schema



Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

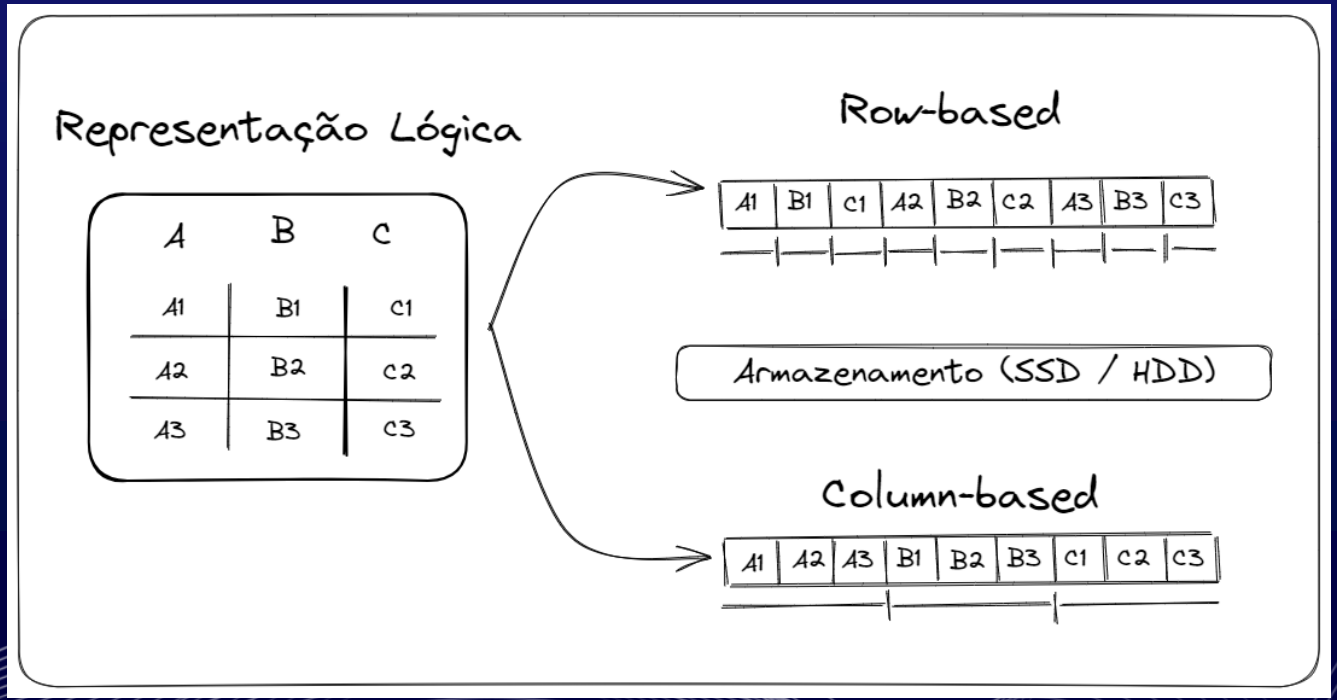
OLTP vs. OLAP



TALK FLORIPA

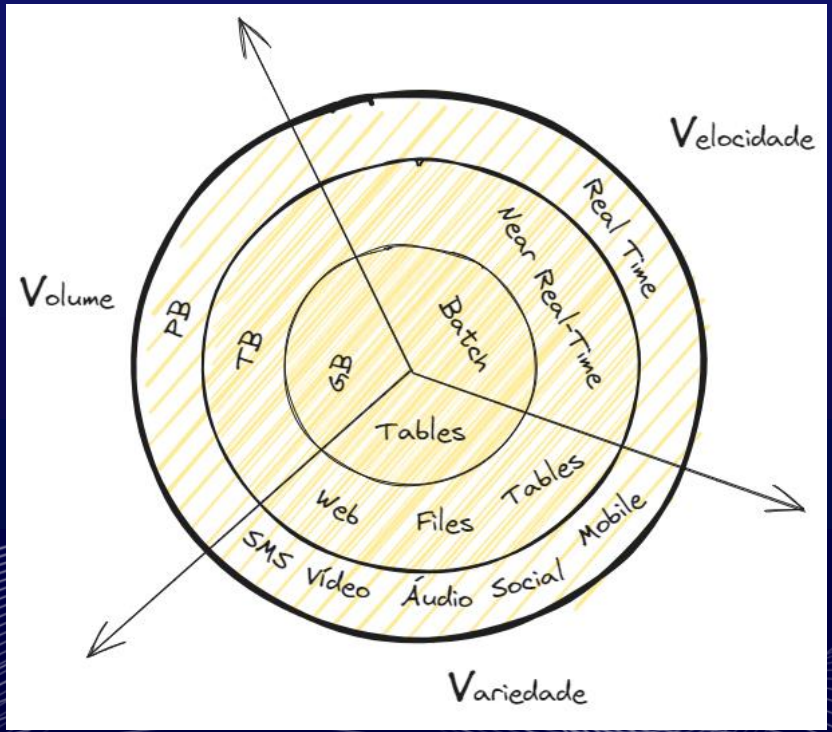
Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

OLTP vs. OLAP



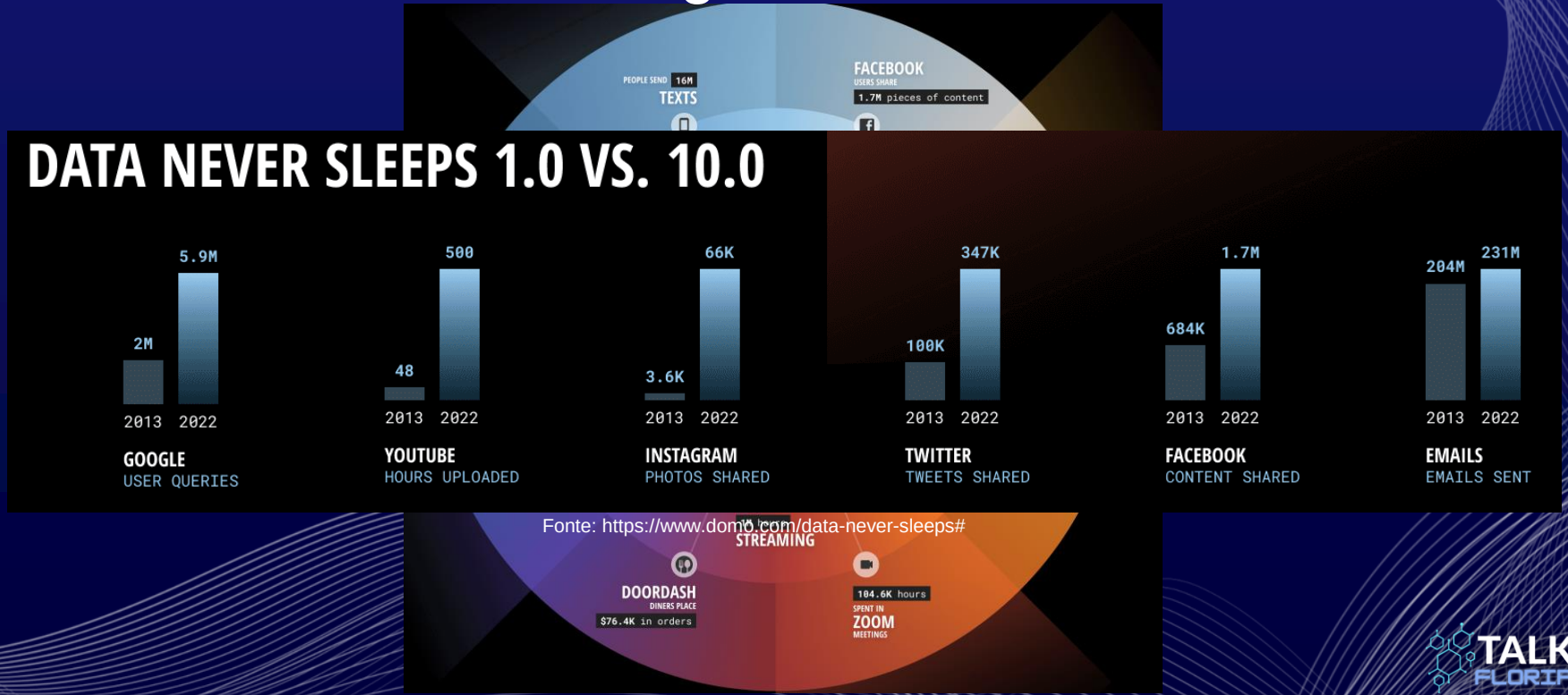
Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Big Data



Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Big Data

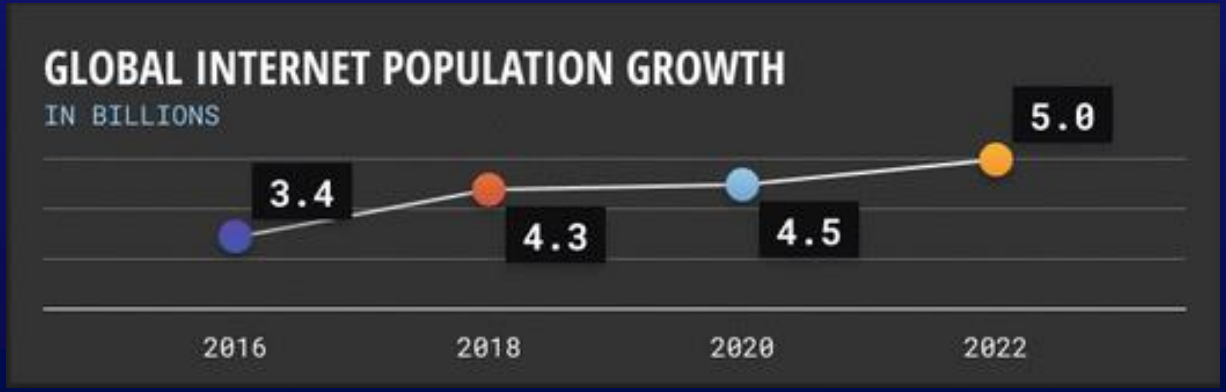


Fonte: https://www.domo.com/data-never-sleeps#

Fonte: https://www.domo.com/data-never-sleeps#

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Big Data



Fonte: https://www.domo.com/data-never-sleeps#

Data Lake

- Conceito surgiu por volta de 2011;
- Repositório de dados brutos (RAW);
- ELT (*Extract – Load – Transform*);
- Não processa dados, somente armazena;
- Dados estruturados, semiestruturados e não estruturados;
- Suporta o aumento de demanda por diferentes fontes, formatos e volume de dados;
- *Schema on read*;

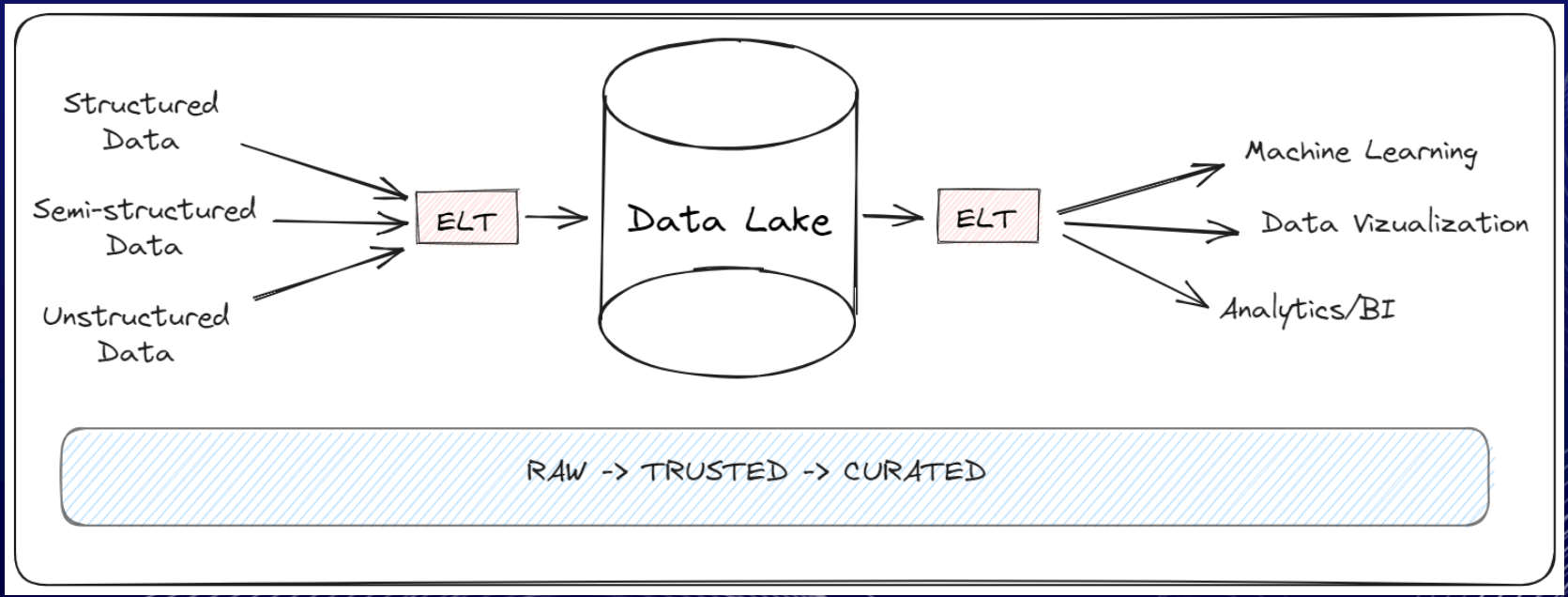
Pontos positivos:

- Escalabilidade de processamento e armazenamento;
- Armazenamento de arquivos com formatos distintos;

Pontos negativos:

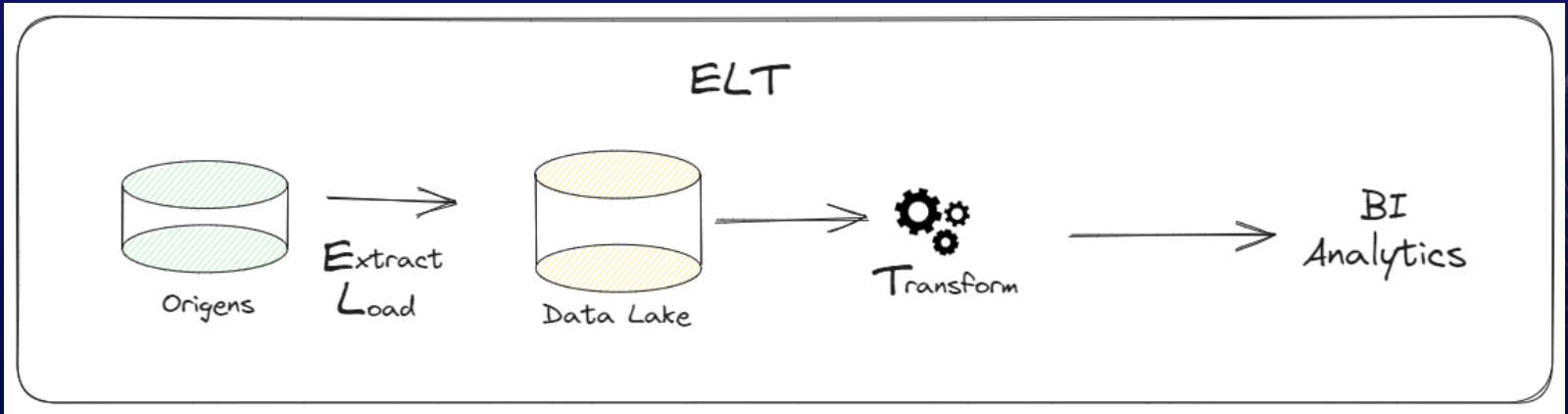
- Problemas de performance;
- Inconsistência de dados.

Data Lake


Structured Data
Semi-structured Data
Unstructured Data
ELT
Data Lake
ELT
Machine Learning
Data Vizualization
Analytics/BI
RAW -> TRUSTED -> CURATED

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

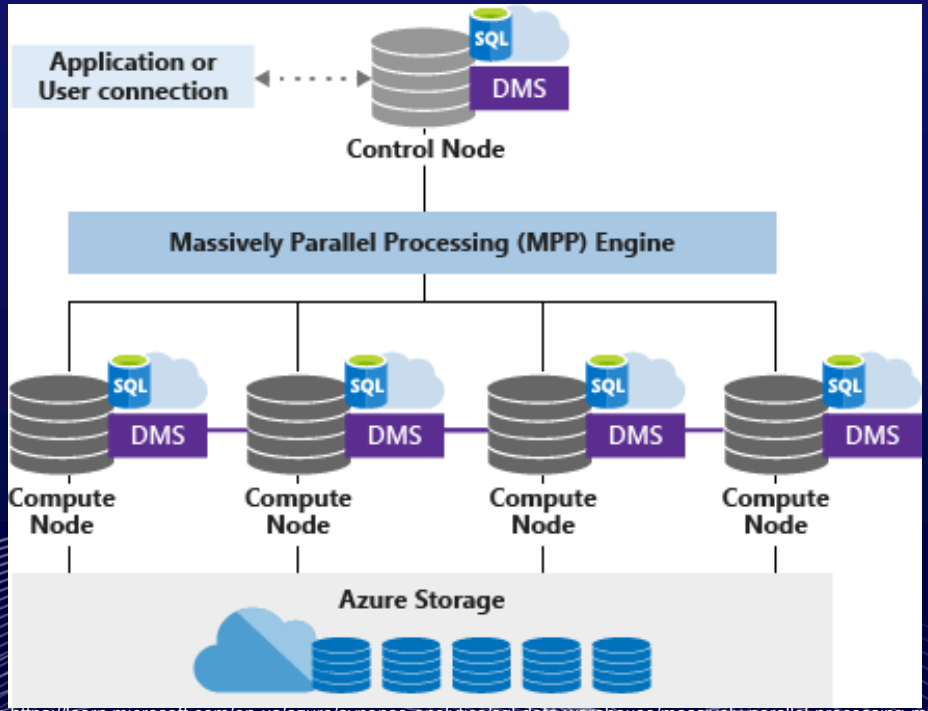
ELT



Modern Data Warehouse (MDW)

- *Massive Parallel Processing* (MPP);
- Surgiu com o advento de recursos PaaS nos provedores de *Cloud;*
- Synapse SQL Dedicated Pool (Azure) / Big Query (GCP) / Redshift (AWS);
- Desacoplamento das camadas de armazenamento e processamento;
- Armazenamento colunar;
- Processamento distribuído.

Modern Data Warehouse (MDW)



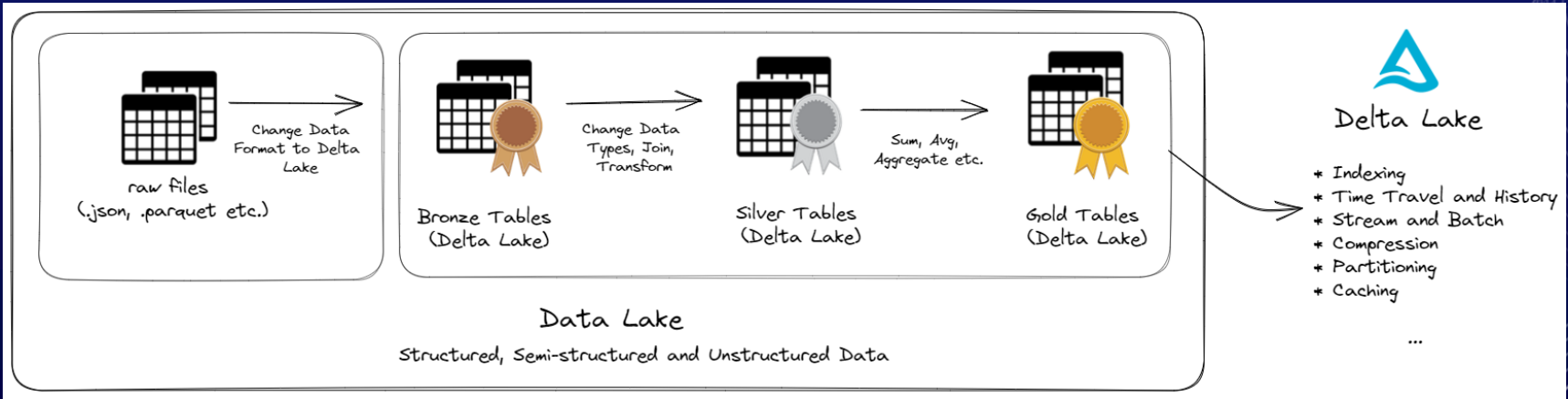
Fonte: https://learn.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/massively-parallel-processing-mpp-architecture

Lakehouse

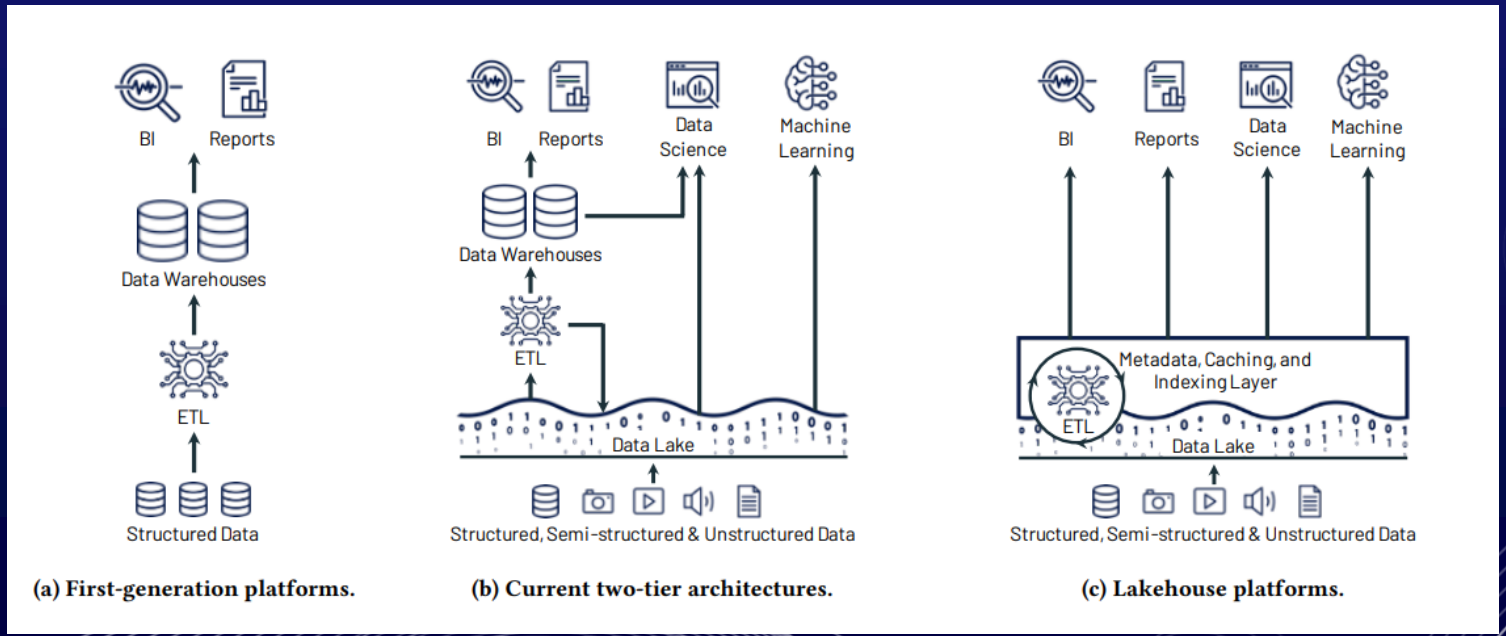
- Conceito surgiu por volta de 2019;
- Construído com a utilização de um Data Lake;
- União das melhores características do Data Warehouse e Data Lake;
- Construído com a utilização de um *Storage Layer:*
  - Delta Lake;
  - Apache Iceberg;
  - Apache Hudi;
- *The Medallion Architecture.*

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Lakehouse com Delta Lake



Data Warehouse vs. Data Lake vs. Lakehouse



Fonte: https://www.databricks.com/sites/default/files/2020/12/cidr_lakehouse.pdf

Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Data Warehouse vs. Data Lake vs. Lakehouse

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

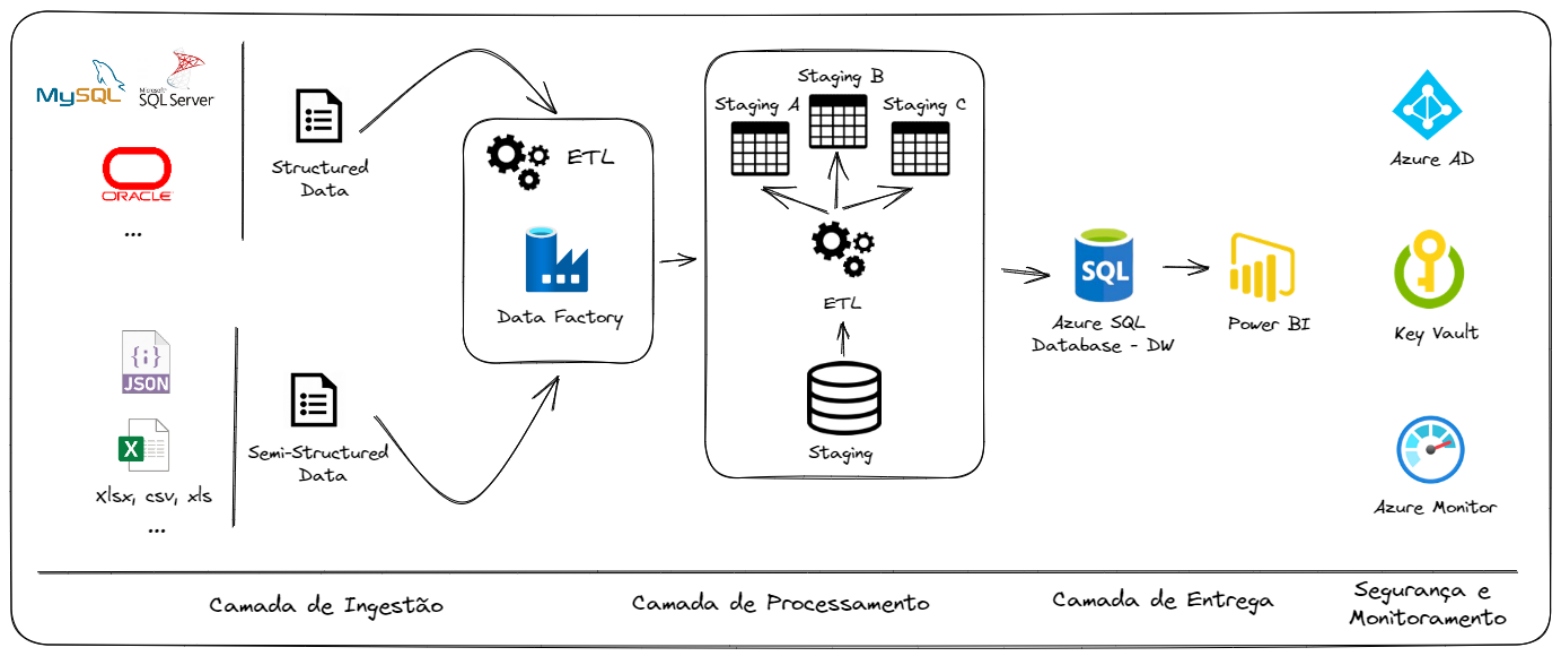
Michael Armbrust[1], Ali Ghodsi[1,2], Reynold Xin[1], Matei Zaharia[1,3]
[1]Databricks, [2]UC Berkeley, [3]Stanford University

Abstract

This paper argues that the data warehouse architecture as we know it today will wither in the coming years and be replaced by a new architectural pattern, the Lakehouse, which will (i) be based on open direct-access data formats, such as Apache Parquet, (ii) have first-class support for machine learning and data science, and (iii) offer state-of-the-art performance. Lakehouses can help address several major challenges with data warehouses, including data staleness, reliability, total cost of ownership, data lock-in, and limited use-case support. We discuss how the industry is already moving toward Lakehouses and how this shift may affect work in data management. We also report results from a Lakehouse system using Parquet that is competitive with popular cloud data warehouses on TPC-DS.

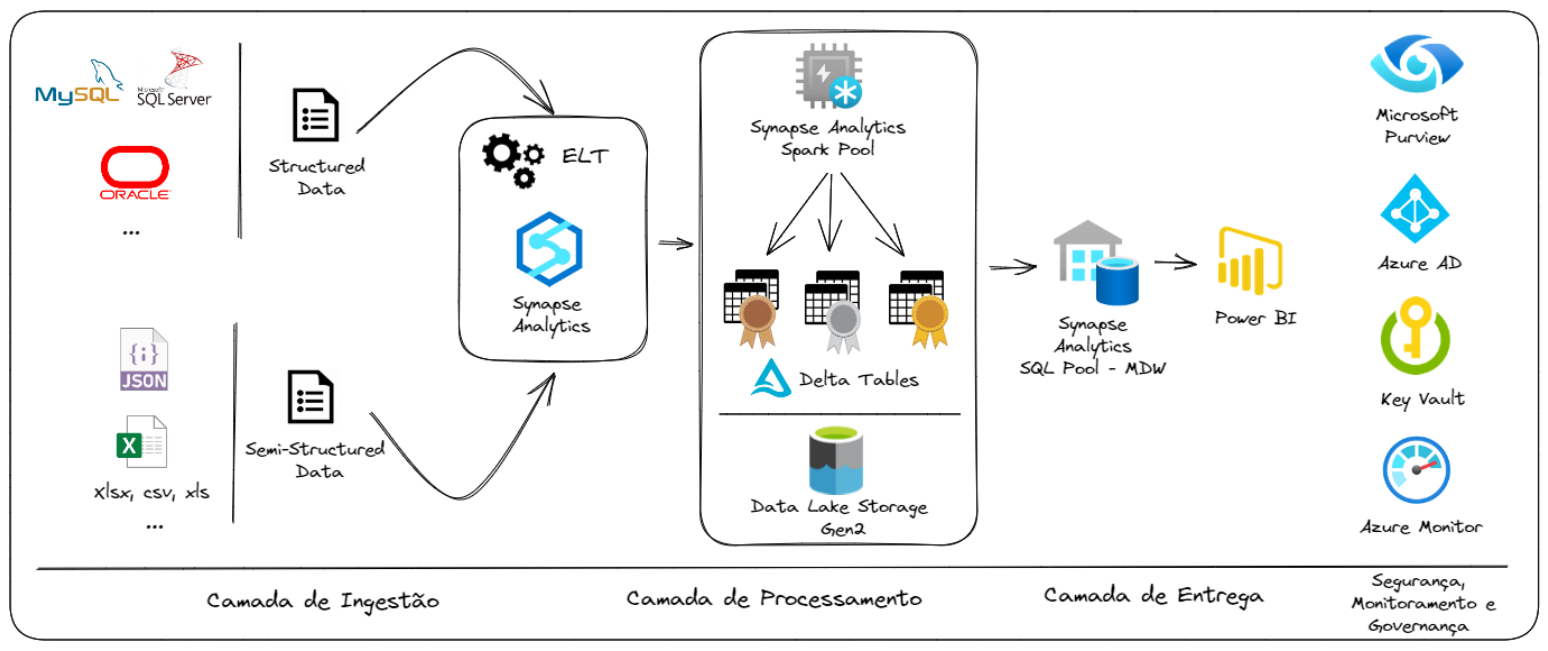
Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Arquitetura de Referência



Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

Arquitetura de Referência



Conceitos e Aplicações de Data Warehouse, Data Lake e Lakehouse no Azure

TALK FLORIPA

Agradecimentos