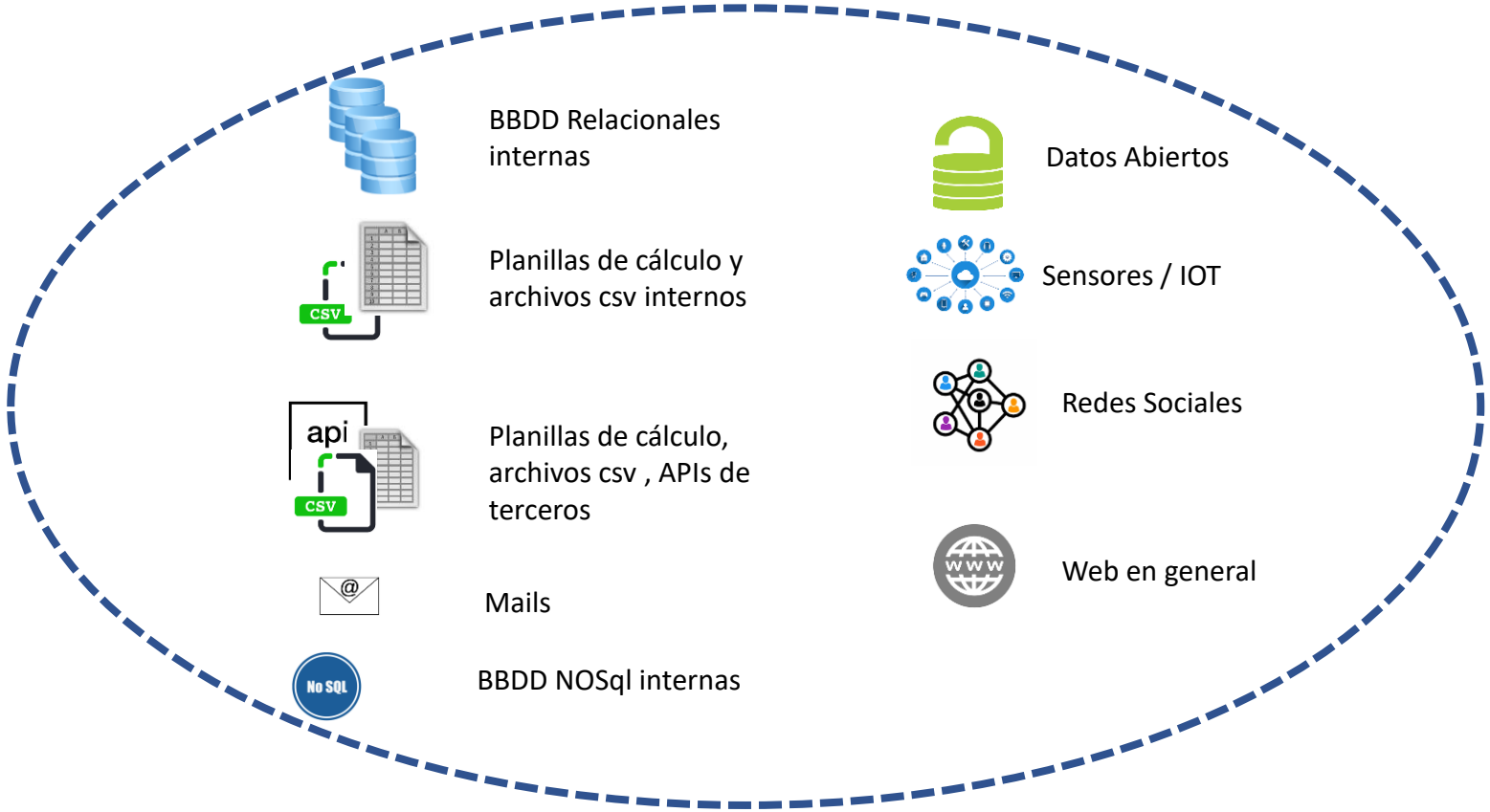
Procesamiento Transaccional y Analítico

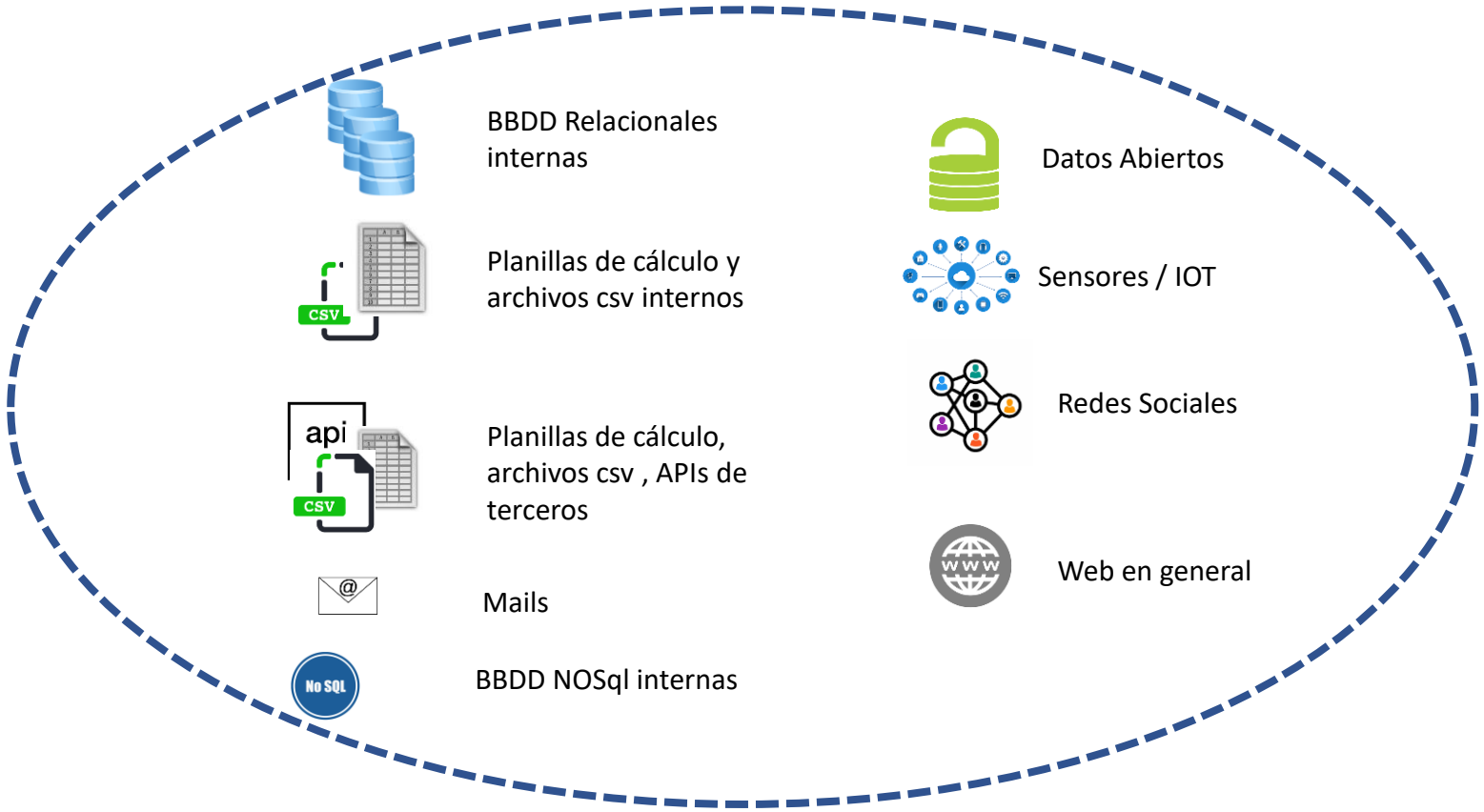
Autor: Sergio D'Arrigo

Orígenes de datos

- Orígenes de datos comunes disponibles en una organización



Modos básicos de uso



Modos básicos de uso de los datos y las bases de datos

Procesamiento Transaccional

Analítica de Datos

Transaccional vs Analítico

En líneas generales, hay dos modos de utilización de las bases de datos

- Muchos usuarios de la base de datos
- Acceso R/W a poco volumen de datos, en general un único origen
- Requiere performance en actualizaciones frecuentes de bajo volumen
- Por lo general no se requiere mucha historia
- Prioridad: funcionamiento de una organización

Procesamiento Transaccional

- Pocos usuarios, mayormente analistas de datos y gerenciales
- Acceso RO a mucho volumen de datos, múltiples orígenes
- Requiere performance en consultas no tan frecuentes de gran volumen y alta carga de procesamiento
- Se requiere historia de datos
- Prioridad: soporte a toma de decisiones, estratégicas y tácticas

Analítica de datos

Transaccional vs Analítico

¿Estos modos requieren plataformas diferentes?

Diferentes funciones

- Requieren tuneo diferente…
  - ¿Con nuevas tecnologías una misma plataforma podría tener performance satisfactoria para ambos paradigmas?

Diferentes datos

- Soporte a decisiones requiere datos históricos…
  - ¿con nuevas tecnologías podrían mantenerse datos históricos en transaccional sin afectar su funcionamiento?

Orígenes de datos

- Soporte a decisiones requiere integración de múltiples orígenes…
  - Con nuevas tecnologías orientadas a microservicios se diversifican más los orígenes…

Calidad de datos

- Soporte a decisiones requiere datos de calidad y homogeneidad en su tratamiento y conceptos…
  - Las BBDD de los sistemas transaccionales por lo general no garantizan este requerimiento de acuerdo a la necesidad analítica…

Repositorios Analíticos de Datos

Arquitectura Trazo Grueso

Orígenes de datos

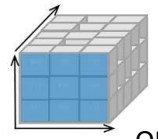
Orígenes internos

Orígenes externos

Big Data

Repositorio Analítico de Datos

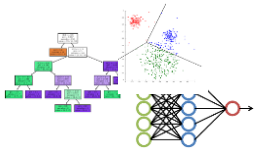
Reporting y Analítica



OLAP

Reporting y BI

Estadísticas



Minería de Datos

Ecosistema de un Data Warehouse Clásico

Orígenes de datos

Orígenes internos

Orígenes externos

Big Data

Repositorio Analítico de Datos

Capa Back-end

Data Stage

Herramientas ETL

Réplica Transaccional

Capa Data Warehouse

Metadata

Data Warehouse Corporativo

Data Marts

ODS (Operational Data Store)

Capa OLAP

Servidor OLAP

Capa Front-end

Herramientas OLAP

Herramientas Reporting y BI

Herramientas Estadísticas

Herramientas Minería de Datos

Este es un esquema genérico, puede haber otras variantes…

Variantes Data Warehouses Clásicos

No todas las capas y/o componentes están presentes en todas las implementaciones.

Podrían no tener servidor OLAP.

Podrían tener Data Warehouse corporativo y no data marts, o viceversa.

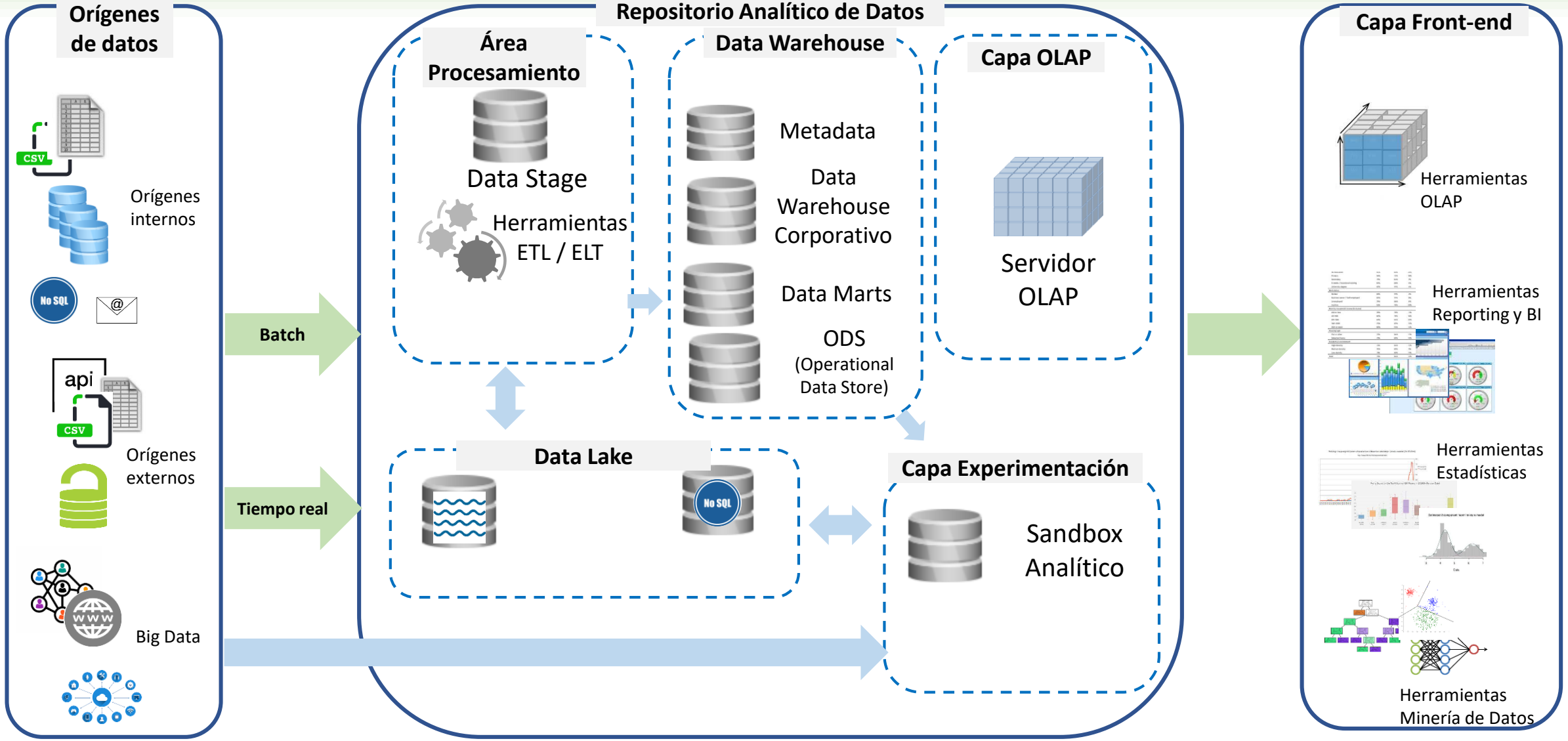
Podrían no tener un ODS o no tener sandbox analítico.

Podrían no tener área de replicación transaccional.

Podríamos tener un “DWH Virtual”, a partir de VISTAS sobre las bases transaccionales, sin tener persistencia alguna.

Caso extremo y poco frecuente

Ecosistema de un Data Warehouse / Data Lake



Este es un esquema genérico, puede haber otras variantes…

DWH y DL – Ejemplos de arquitecturas

- Arquitectura genérica DL-DWH (Vaisman-Zimanyi)

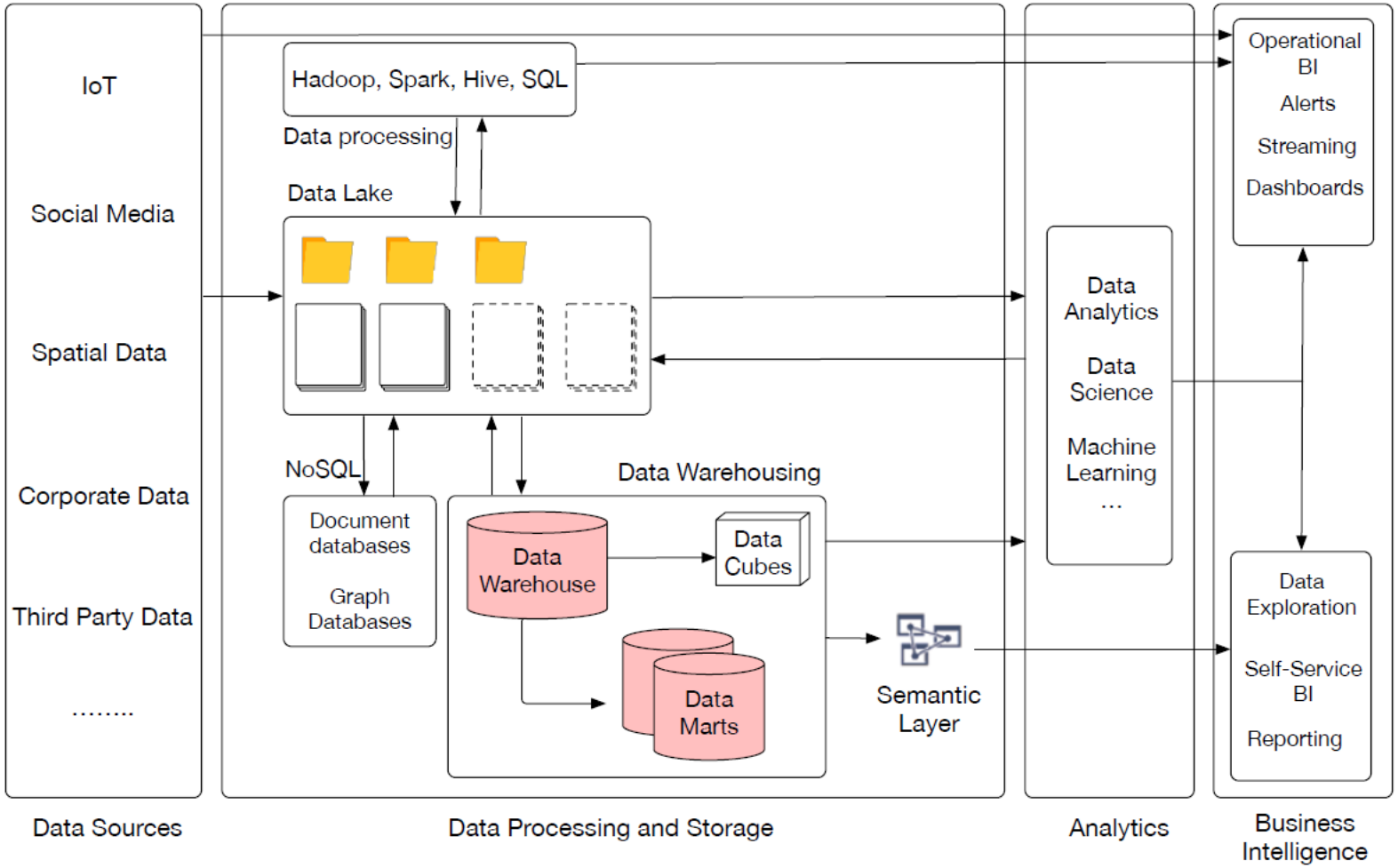


Fig. 15.21 A typical data lake architecture

DWH y DL – Ejemplos de arquitecturas

- Arquitectura DL-DWH (Google Cloud)

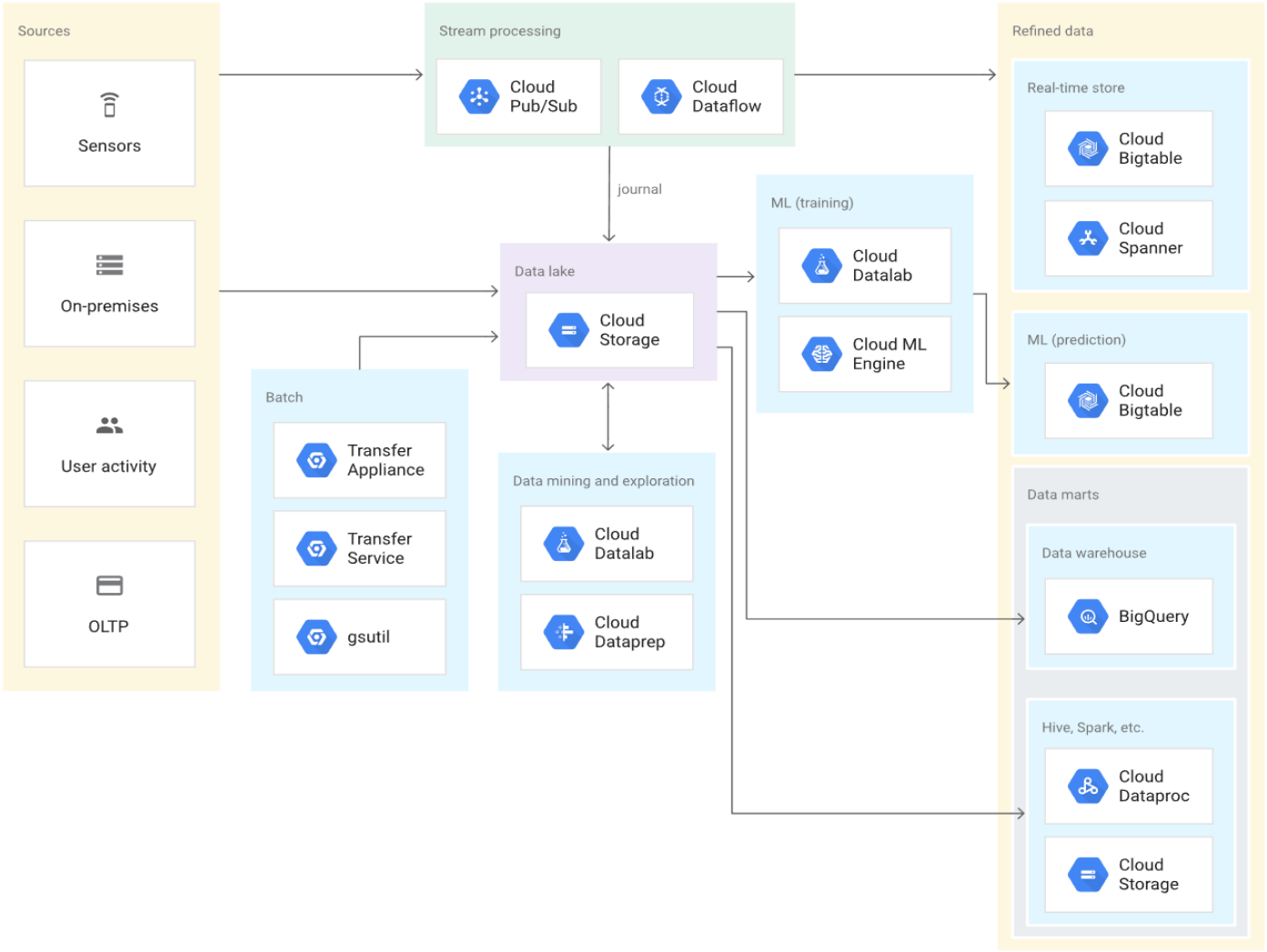
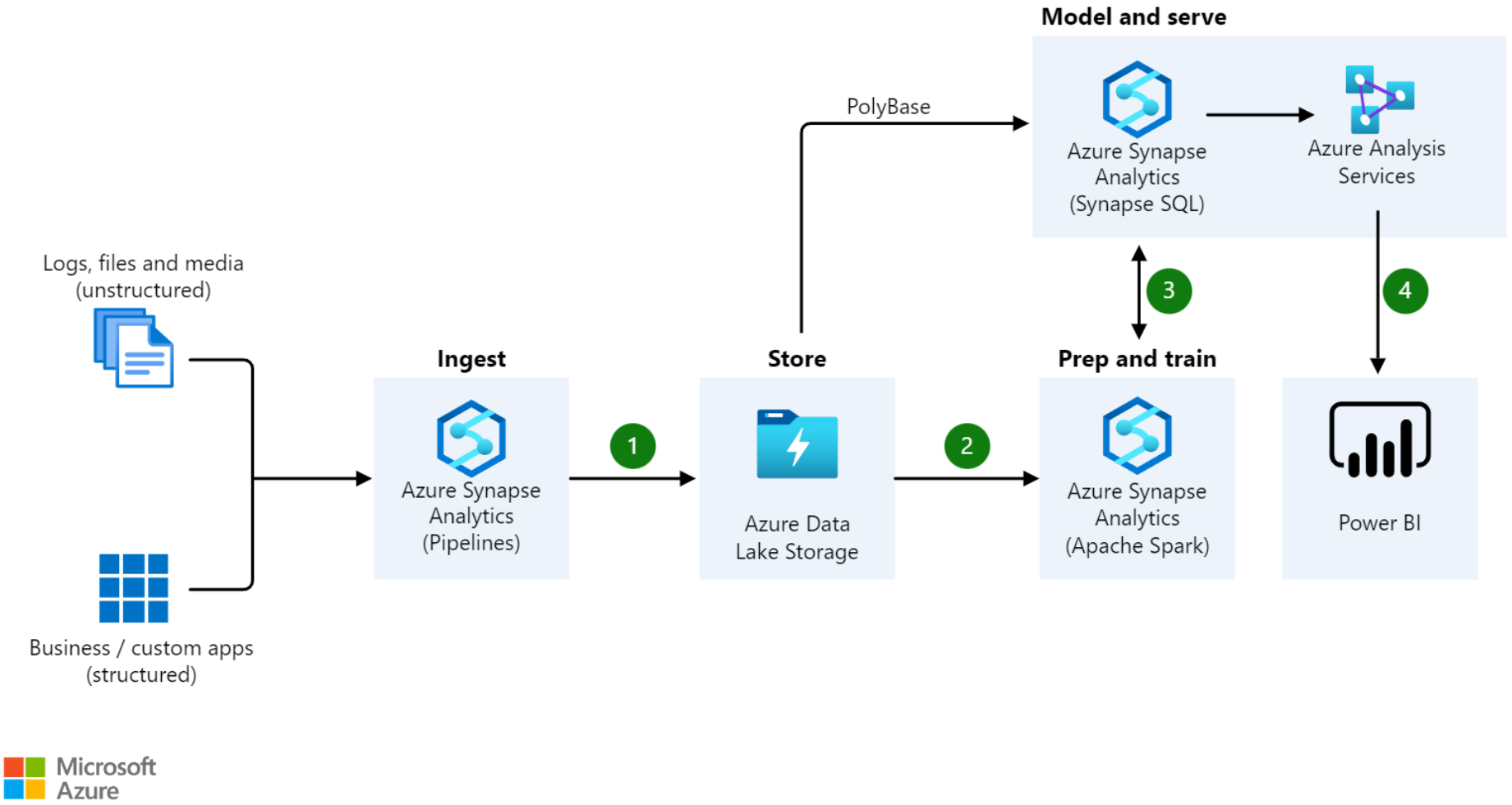


Figura de https://cloud.google.com/architecture/build-a-data-lake-on-gcp?hl=es-419

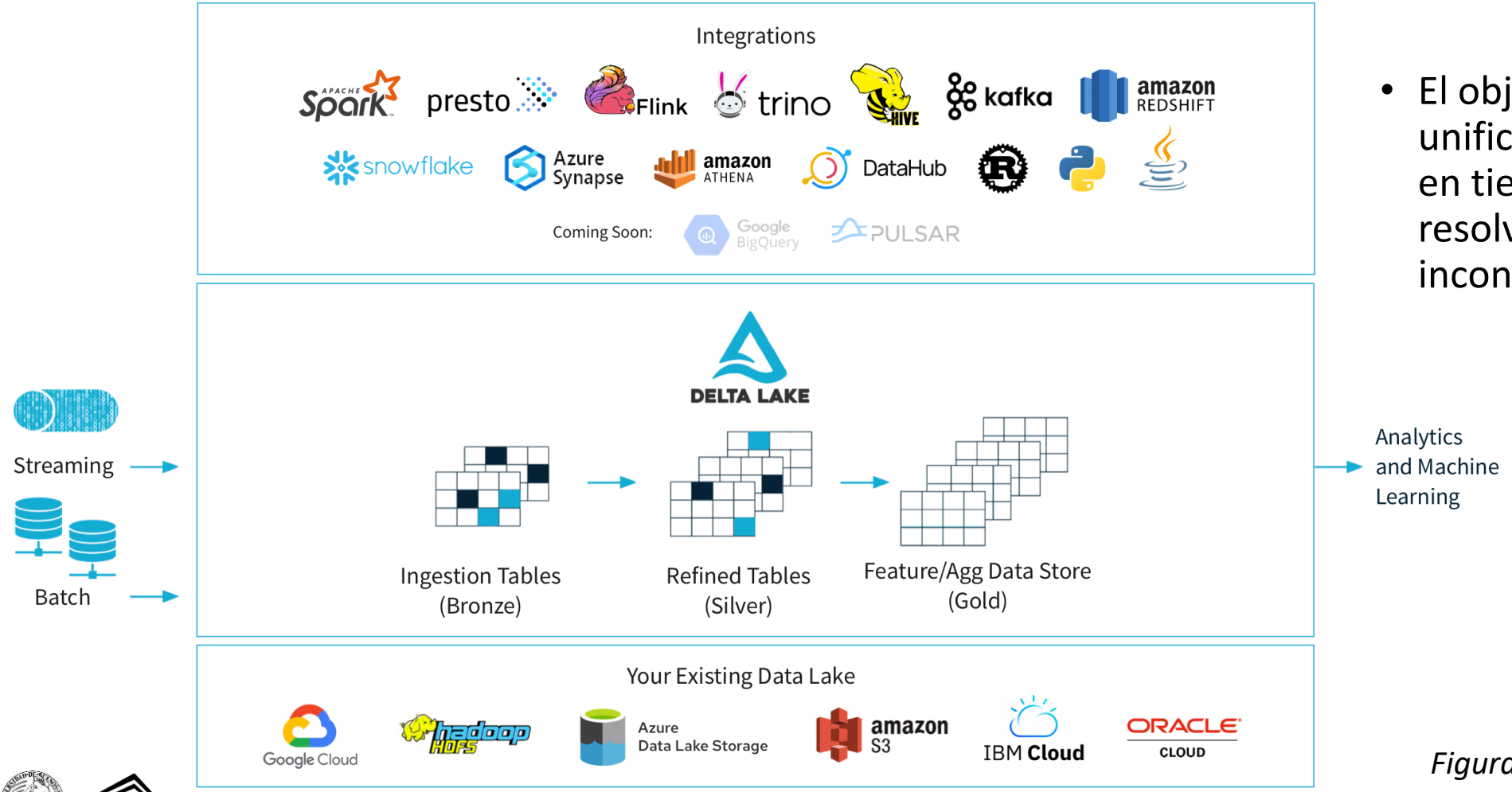
- Una arquitectura Azure Synapse

https://learn.microsoft.com/en-us/azure/architecture/solution-ideas/articles/enterprise-data-warehouse



Delta Lakes

- Existen diversas iniciativas para dar transaccionalidad y propiedades ACID a los Data Lakes.
- Una de ellas es Delta Lake. Es una capa de almacenamiento de código abierto que brinda transacciones con propiedades ACID a Apache Spark y cargas de trabajo de big data.



- El objetivo de Delta Lake es unificar el procesamiento de datos en tiempo real y por lotes para resolver los problemas de inconsistencias.

Figura de "https://delta.io/"

Data Lakehouses

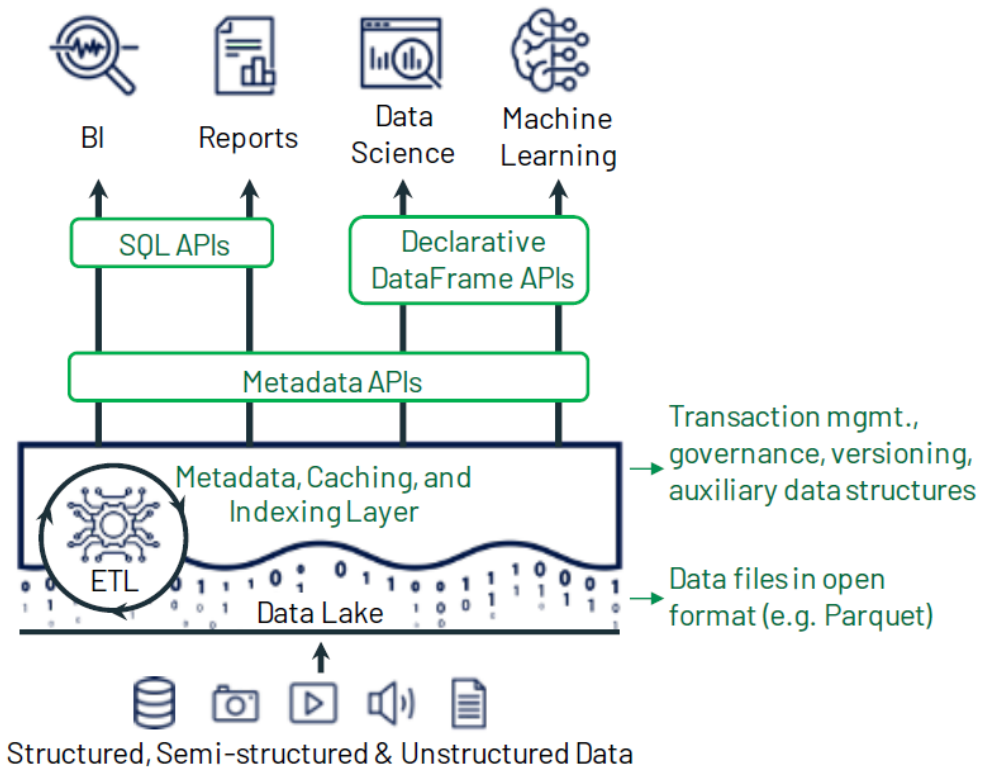


Figura de "Lakehouse: A New Generation of Open Platforms that Unify DataWarehousing and Advanced Analytics", Armbrust y otros

- Se trata de una arquitectura reciente que tiene como objetivo permitir tanto tareas de inteligencia artificial como de Data Warehousing / BI directamente a partir de los datos en los Data Lakes.

Características principales:

- Los datos se almacenan en una plataforma de almacenamiento de objetos de bajo costo (data lake)
- Sobre ese almacenamiento se incorpora una capa transaccional de metadata, que cumple propiedades ACID
- Implementa optimizaciones de acceso a datos para SQL sin necesidad de aplicarle cambios a los archivos de datos.
- Implementa APIs de DataFrames declarativos, permitiendo integración directa con librerías de ML.

Bibliografía

- *Data Warehouse Systems: Design and Implementation, 2da. Ed., Vaisman, Zimanyi, 2022*
- *Building the Data Lakehouse, Inmon, Levin, Srivstava, 2021*
- *The Enterprise Big Data Lake – Delivering the Promise of Big Data and Data Science, Gorelik, 2019*
- *Delta Lake: HighPerformance ACID Table Storage over Cloud Object Stores, Armbrust, Das, Sun, Yavuz, Zhu, Murthy, Torres, van Hovell, Ionescu, Łuszczak, Switakowski,Szafranski, Xiao Li, Ueshin, Mokhtar, Boncz, Ghodsi,Paranjpye, Senster, Xin, Zaharia, Proceedings of the VLDB Endowment, Vol. 13, No. 12, 2020*
- *Lakehouse: A New Generation of Open Platforms that Unify DataWarehousing and Advanced Analytics, Armbrust, Ghodsi, Xin, Zaharia. 11th Annual Conference on Innovative Data Systems Research (CIDR ’21), January 11–15, 2021*
- *Considering unstructured data for OLAP: a feasibility study using a systematic review. Montenegro González, Reis Lopes. Revista de Sistemas de Informação da FSMA n. 14 (2014)*

Dudas

Muchas gracias!