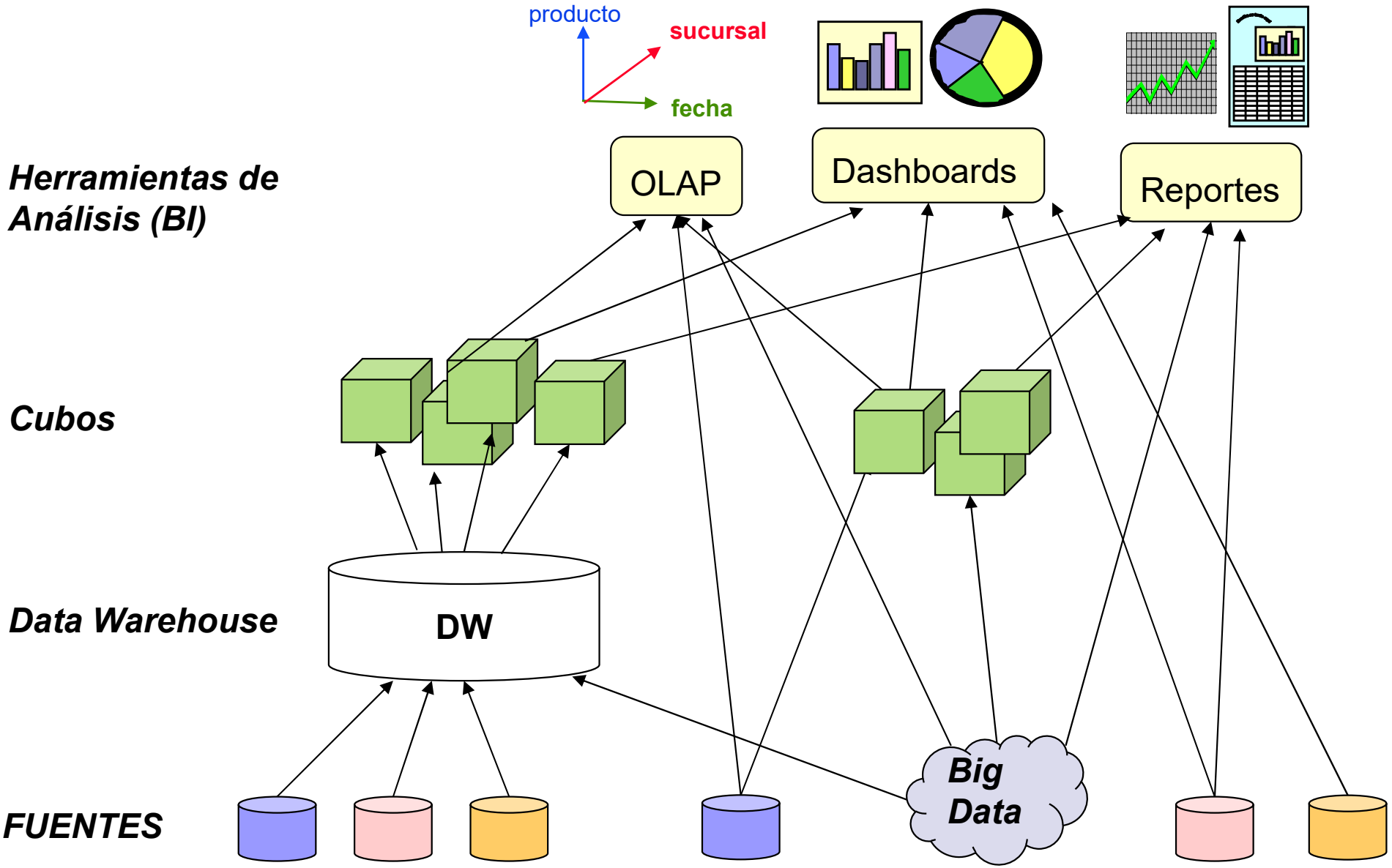
# Sistemas de Información para el Análisis de GVDatos

*Instituto de Computación - Facultad de Ingeniería*

*Abril 2024*

# Arquitecturas de Big Data

# Análisis GVD - Arquitectura

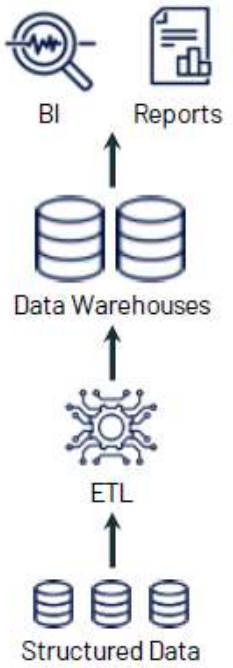


# Arquitecturas para análisis GVD

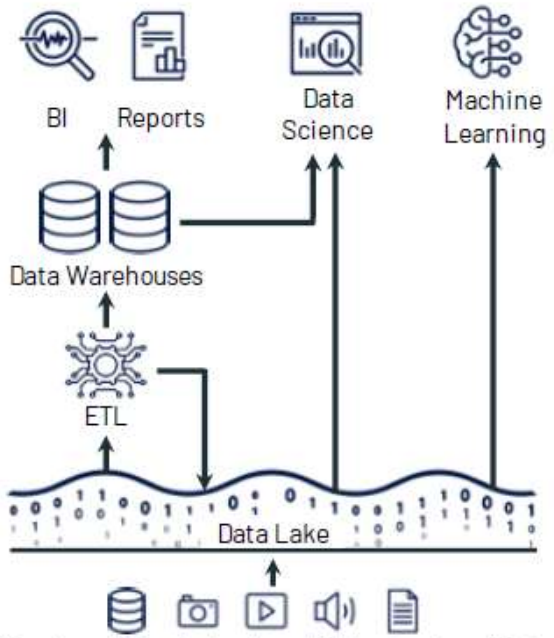
Lakehouse: A New Generation of Open Platforms that Unify DataWarehousing and Advanced Analytics

CIDR '21, Jan. 2021, Online

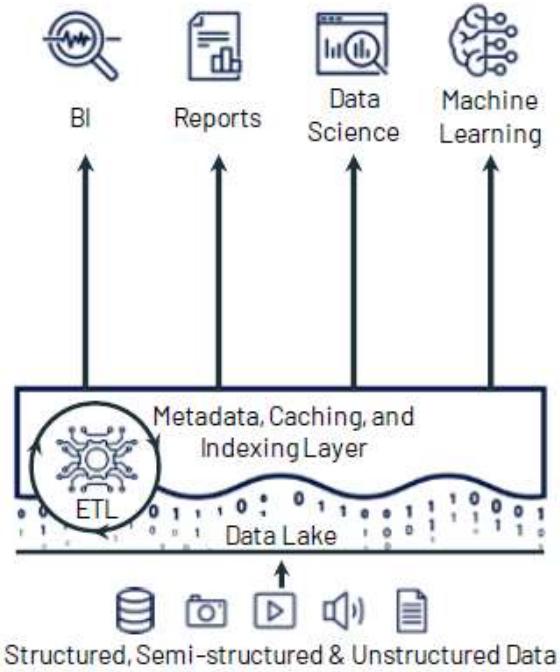
Michael Armbrust, Ali Ghodsi, Reynold Xin, and Matei Zaharia




(a) First-generation platforms.




(b) Current two-tier architectures.




(c) Lakehouse platforms.

# Definiciones

- Data Warehouse
  - Base de datos donde los datos, que provienen de diversas fuentes, fueron cuidadosamente transformados y corregidos, para conformar formatos, esquemas y semántica estándares de la organización.
- ETL – Extraction, transformation and loading
  - Proceso de transformación y carga de datos
- BI (Business Intelligence)
  - Análisis de datos orientada al negocio. En general, utilizando OLAP (On-line Analytical Processing), sistemas basados en el Modelo Multidimensional.

# Data Lake - definiciones

- Colección masiva de *datasets* que pueden
  - estar almacenados en distintos sistemas
  - tener **formatos variados**
  - no estar acompañados de metadatos
  - cambiar en forma autónoma

- Sistema de almacenamiento y gestión de datos
  - flexible, escalable, que ingiere y almacena datos crudos de fuentes heterogéneas en su formato original
  - permite consultas y análisis de datos *on-the-fly*

# Data Lake

- Objetivos
  - Obtener los datos rápidamente de las fuentes y darle a los usuarios la posibilidad de manejar la heterogeneidad y la escalabilidad
  - Muy fácil almacenar datos allí

- Riesgo
  - Convertirse en un pantano de datos, *data swamp*
    - Datos “olvidados” en ese lugar
    - Desconocimiento de lo que hay
    - Acumulación de datos que no se usan

# Data Lake

- Raw data (enseguida de Ingestión)
  - Desconocidos
    - Semántica
    - Calidad
    - Provenance
    - Conexión con otros datasets
- Metadata o Data governance
  - Evitar data swamp
  - Puede obtenerse activamente desde el DL y a través de interacción con los usuarios

# Data Lake

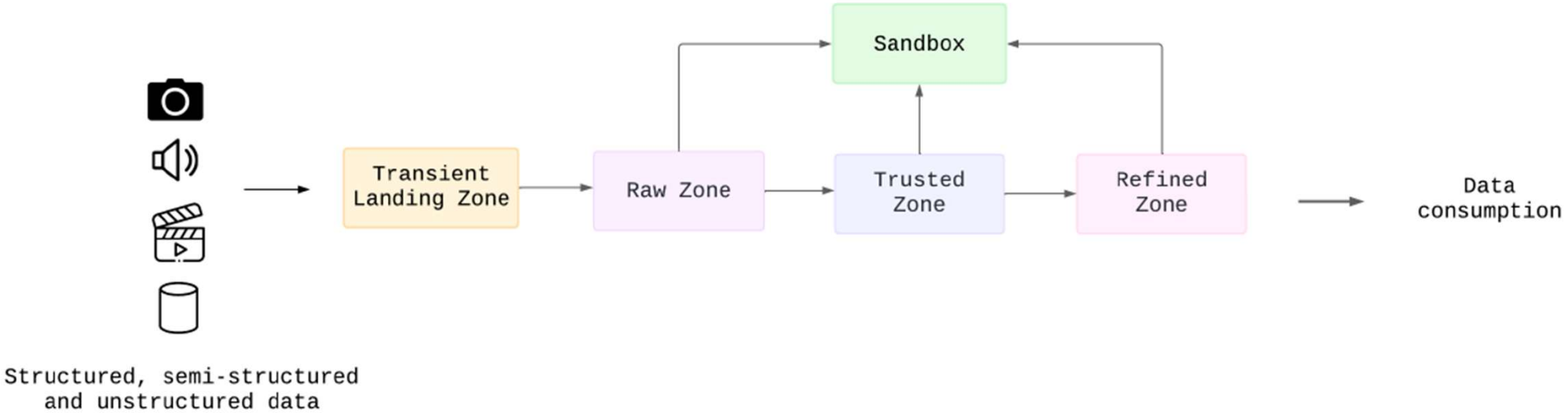
- Herramientas comerciales
  - Google cloud BigLake, Azure Data Lake, AWS Lake, Delta Lake de Databricks.

- Investigación - Desafíos
  - metadatos, calidad de datos, provenance, data preparation, organización de los datasets, modelos de datos, integración de datos, descubrimiento de relaciones entre datasets

# Arquitecturas

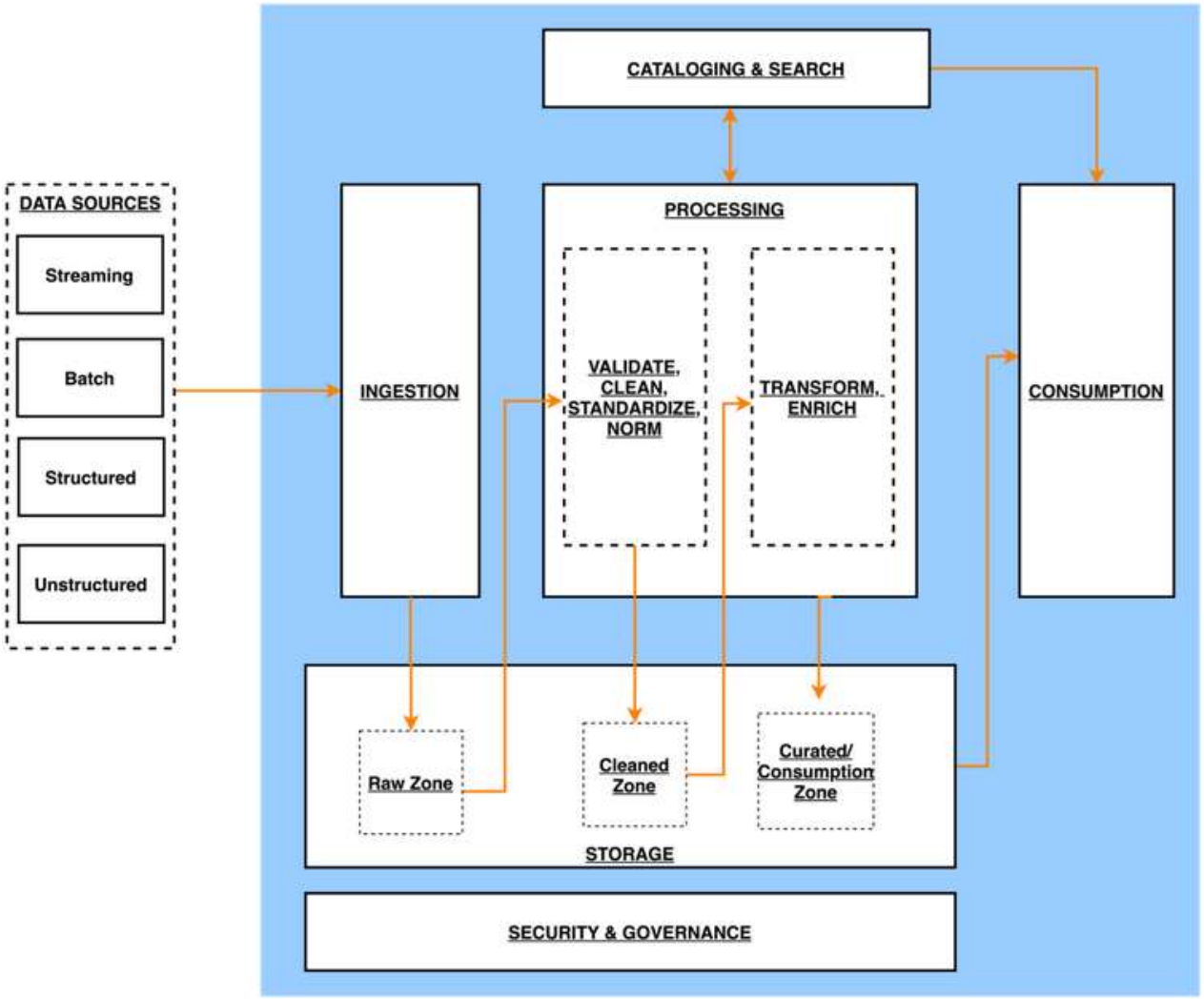
- 3 tipos de arquitecturas
  - Basadas en *data ponds* (estanque)
    - Los datos van pasando de un estanque a otro
    - Según características de los datos (estructura y uso)
  - Basadas en *data zones*
    - Zonas según el grado de procesamiento de los datos (siempre está la zona *Raw*)
    - Todas pueden ser utilizadas según las necesidades
  - Arquitecturas *lambda*
    - Organiza según datos *batch* o *streaming*
    - 2 ramas, una con los datos procesados periódicamente y otra con los datos procesados en tiempo real

# Arquitectura en zonas - Ejemplo




Basada en: Zaloni Architecture Architecting Data Lakes (2nd ed.). O'Reilly Media, Inc. What is Data Governance?

# Arquitectura en zonas - Ejemplo




https://docs.aws.amazon.com/whitepapers/latest/aws-serverless-data-analytics-pipeline/logical-architecture-of-modern-data-lake-centric-analytics-platforms.html

# Enfoques de almacenamiento

- *On-premise* (propio)
  - Se almacenan los datos en servidores propios
  - Manejo y administración de la infraestructura del almacenamiento de los datos
- *Cloud* (dado por proveedores)
  - Arquitecturas *Serverless*
    - El cliente no se tiene que preocupar por el manejo de los servidores. Escalabilidad para los datos.
  - Proveen distintos tipos de almacenamiento (costos/performance) según las funcionalidades necesarias (por ej. para las distintas zonas del DL)

# Enfoques Recuperación de datos

- Exploración – 2 tipos
  - Descubrimiento de datasets basado en relaciones entre ellos
    - Recibe una especificación de un dataset y devuelve los más relacionados.
  - Interfaz de consultas unificada para fuentes heterogéneas
    - Navegación de los datos y metadatos, consultas SQL, búsquedas, etc.

# Enfoques de carga de datos

- *Schema-on-write*
  - El esquema se define cuando se almacenan los datos
  - ETL en DW
  - Inadecuado para BigData (sensores, redes sociales, etc.), características 3Vs
- *Schema-on-read*
  - El esquema se define cuando se utilizan los datos
  - Data Lakes
  - Útil cuando se generan grandes volúmenes de datos semi/no-estructurados, a gran velocidad

# Limitaciones de los DW

- Costo de almacenamiento centralizado.
- Dificultad para manejar formatos heterogéneos, como texto, imágenes, datos de sensores.
- Obsolescencia de los datos generada por los procesos ETL y el enfoque schema-on-write.
- No adecuación para análisis de tipo aprendizaje automático y ciencia de datos.

# DW vs. DL

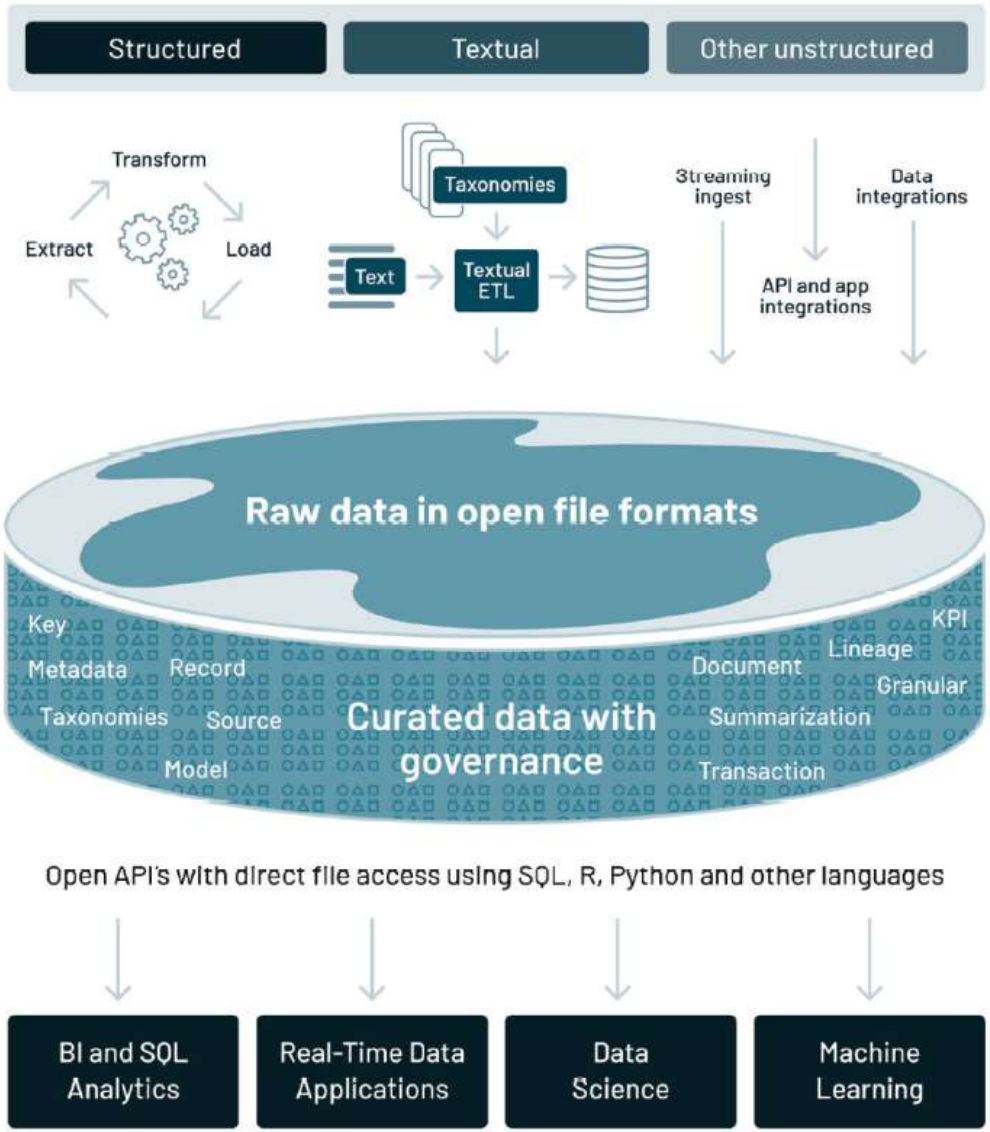
| | Data Warehouse | Data Lake |
|---|---|---|
| **Data format** | Close, proprietary format | Open format |
| **Types of data** | Structured data, with limited support for semistructured data | All types:<br>Structured data, semi-structured data, textual data, unstructured (raw) data |
| **Data Access** | SQL-only | Open APIs for direct access to files with SQL, R, Python, and others |
| **Reliability** | High quality, reliable data with ACID transactions | Low quality, data swamp |
| **Governance and security** | Fine-grained security and governance | Poor governance. Security needs applied to files |
| **Performance** | High | Low |
| **Scalability** | Scaling becomes exponentially more expensive | Scales to any amount of data at low cost, regardless of type |
| **Use case support** | BI, SQL, Decision support | Machine Learning |

Inmon 2021

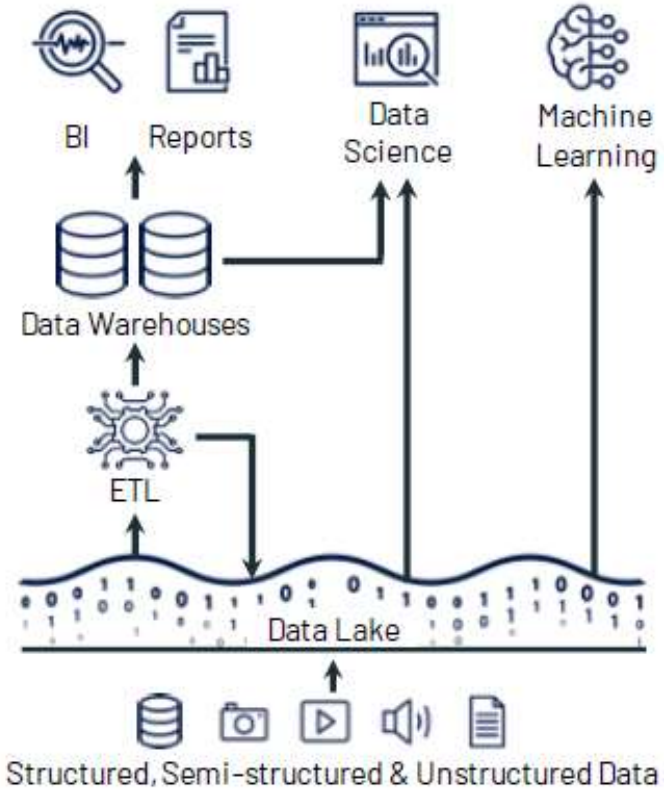
# Data Lakehouse

- Se propone para solucionar los problemas de los DW y de los DL
- Combinan almacenamiento de bajo costo en formatos abiertos y accesibles que proveen los DL, con la gran capacidad de gestión de datos y de optimización de los DW
- Herramientas comerciales
  - MS Synapse Analytics, DataBricks Delta Lake, Google BigQuery, AWS, Snowflake (multi-vendor)

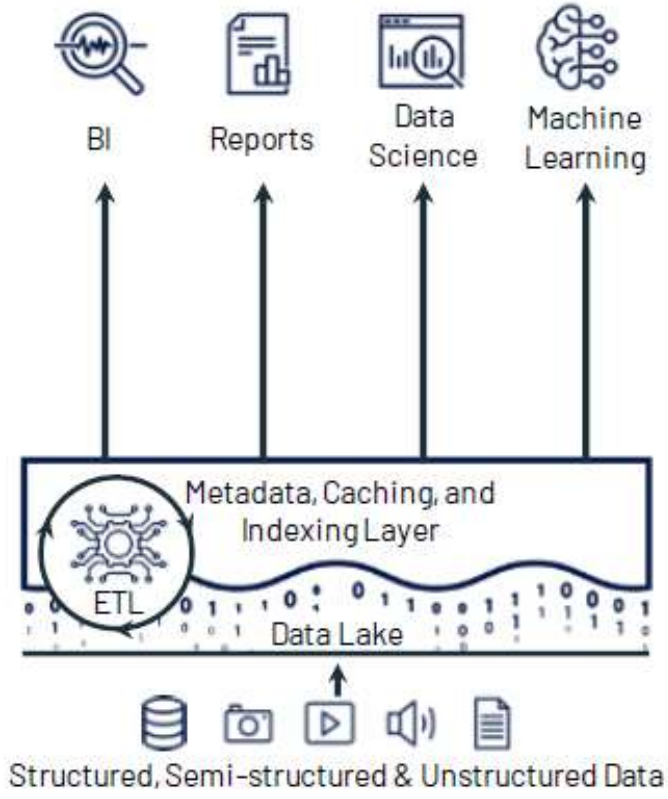
# Data Lakehouse (Inmon 2021)



# Data Lakehouse




(b) Current two-tier architectures.



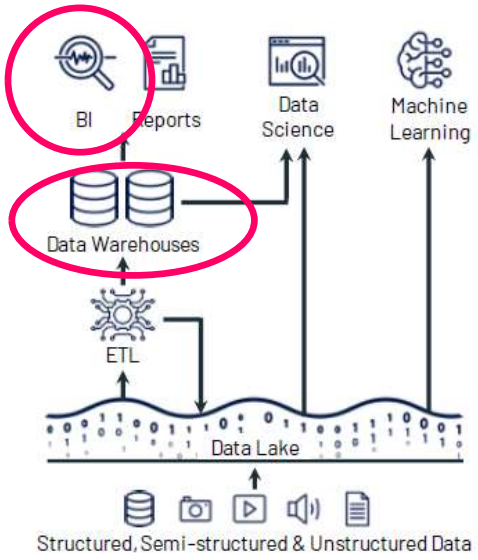

(c) Lakehouse platforms.

# DW – DL – DLH

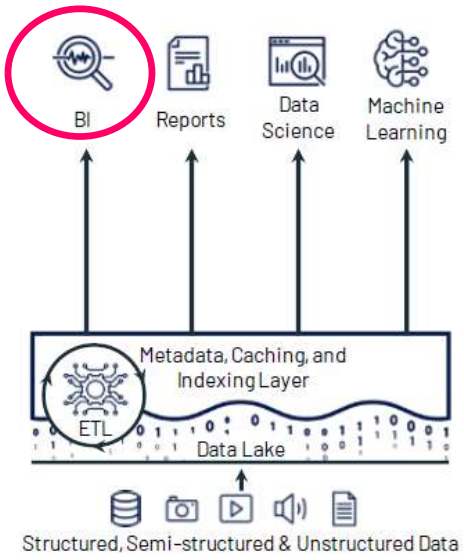
| | Data Warehouse | Data Lake | Data Lakehouse |
|---|---|---|---|
| **Data format** | Close, proprietary format | Open format | Open format |
| **Types of data** | Structured data, with limited support for semistructured data | All types: Structured data, semi-structured data, textual data, unstructured (raw) data | All types: Structured data, semi-structured data, textual data, unstructured (raw) data |
| **Data Access** | SQL-only | Open APIs for direct access to files with SQL, R, Python, and others | Open APIs for direct access to files with SQL, R, Python, and others |
| **Reliability** | High quality, reliable data with ACID transactions | Low quality, data swamp | High quality, reliable data with ACID transactions |
| **Governance and security** | Fine-grained security and governance | Poor governance. Security needs applied to files | Fine-grained security and governance |
| **Performance** | High | Low | High |
| **Scalability** | Scaling becomes exponentially more expensive | Scales to any amount of data at low cost, regardless of type | Scales to any amount of data at low cost, regardless of type |
| **Use case support** | BI, SQL, Decision support | Machine Learning | One data architecture for BI, SQL, and machine learning |

Inmon 2021

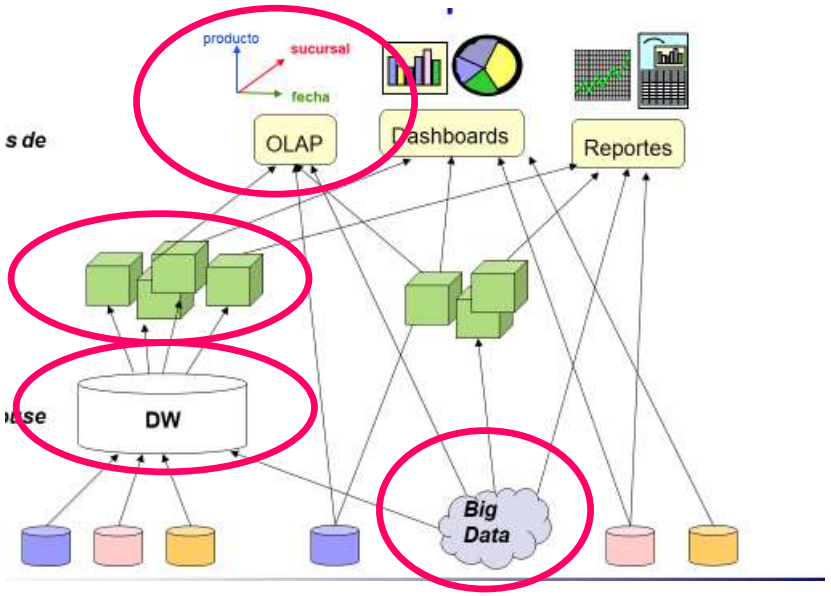
# OLAP – Modelos multidimensionales




(b) Current two-tier architectures.




(c) Lakehouse platforms.



# Bibliografía


- M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia. Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedingsof CIDR, 2021.
- C. Giebler, C. Gröger, E. Hoos, H. Schwarz, B. Mitschang. Leveraging the datalake: Current state and challenges. In Big Data Analytics and Knowledge Discovery, pages 179–188, 2019. Springer.
- R. Hai, C. Quix, and M. Jarke. Data lake concept and systems: a survey, 2021
- F. Nargesian, E. Zhu, R. J. Miller, K. Q. Pu, and P. C. Arocena. Data Lake Management: Challenges and Opportunities. Proc. VLDB Endow., 12(12): 1986–1989, 2019.
- F. Ravat and Y. Zhao. Data Lakes: Trends and Perspectives. In Database and Expert Systems Applications, pages 304–313, 2019. Springer.
- B. Inmon, M. Levins, and R. Srivastava. Building the Data Lakehouse. Technics Publications, 2021
- D. Oreanin and T. Hlupi. Data lakehouse - a novel step in analytics architecture. In 44th International Convention on Information, Communication and Electronic Technology (MIPRO), pages 1242–1246, 2021.