University of Stuttgart

Institute for Parallel and Distributed Systems (IPVS)
Applications of Parallel and Distributed Systems

Jan Schneider
jan.schneider@ipvs.uni-stuttgart.de
Universitätsstraße 38, D-70569 Stuttgart, Germany

The Data Platforms Landscape: An Overview

Poster Session

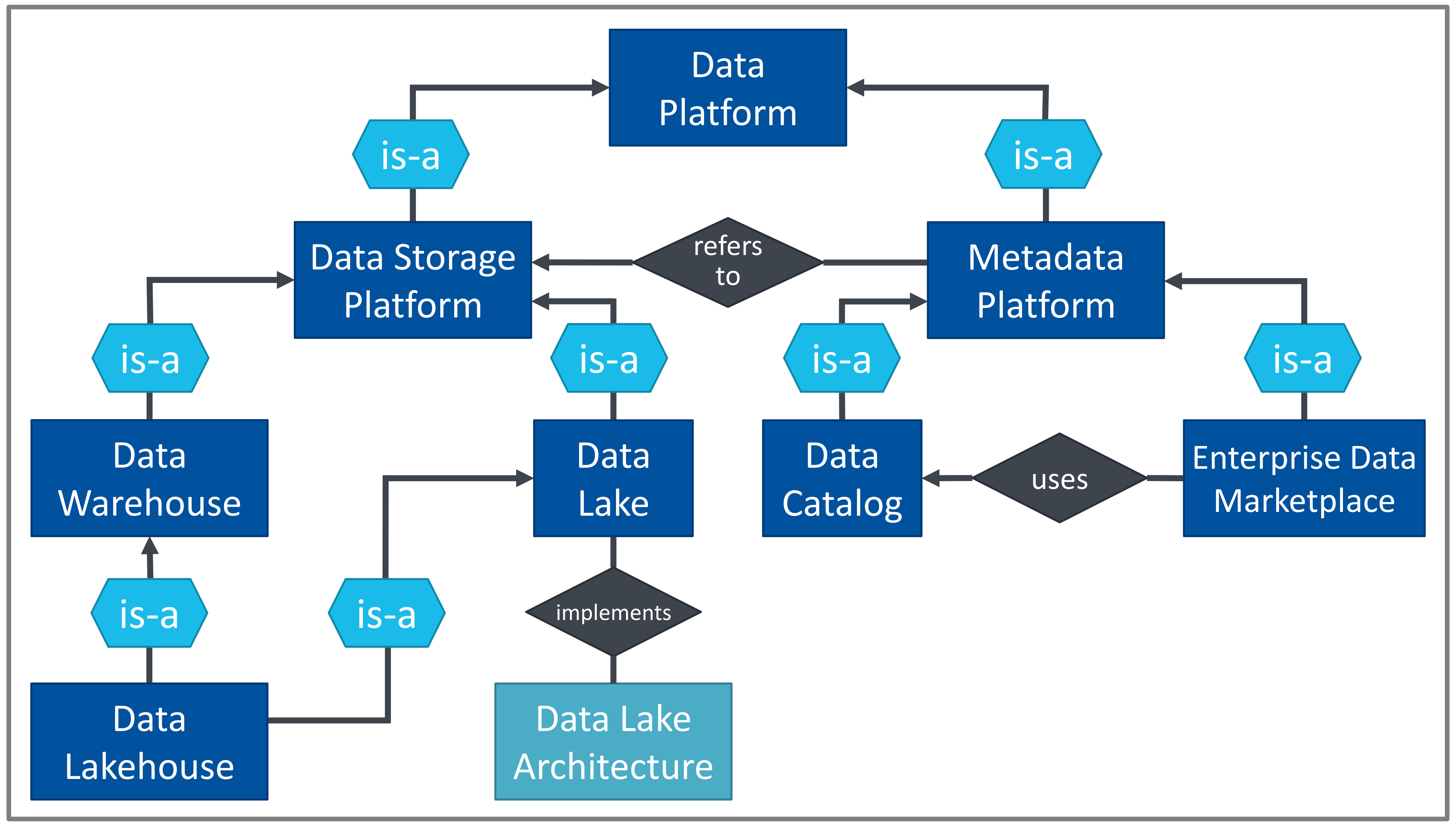
Jan Schneider

Motivation:

- Digital transformation and AI have arrived → holistic view on value chains to enable cross-phase optimizations
- Enterprises need to collect, organize, process and analyze huge amounts of data
- Different types of data platforms have emerged → large, diverse landscape

Data Platforms [1]:

- Store and manage data as well as related metadata from all sources for analytical purposes
- Use-case independent and re-usable
- Include: Data **extraction** from sources, **ingestion**, **storage**, **processing** and **provisioning**



Data Warehouse [2]:

- Subject-oriented, integrated, non-volatile and time-variant collection of data for analytical purposes
- Used for Business Intelligence, Reporting/OLAP
- Use-case specific model design upfront → inflexible
- Proprietary, mgtm. features, ACID, high performance

Data Lake [3]:

- Scalable and flexible management of all kinds of data in their raw format for analytical purposes
- Basis for Advanced Analytics (Data Science, ML, AI, ...)
- No upfront model design → „schema on read"
- Open data formats, direct access, lower performance
- All types of data and related metadata

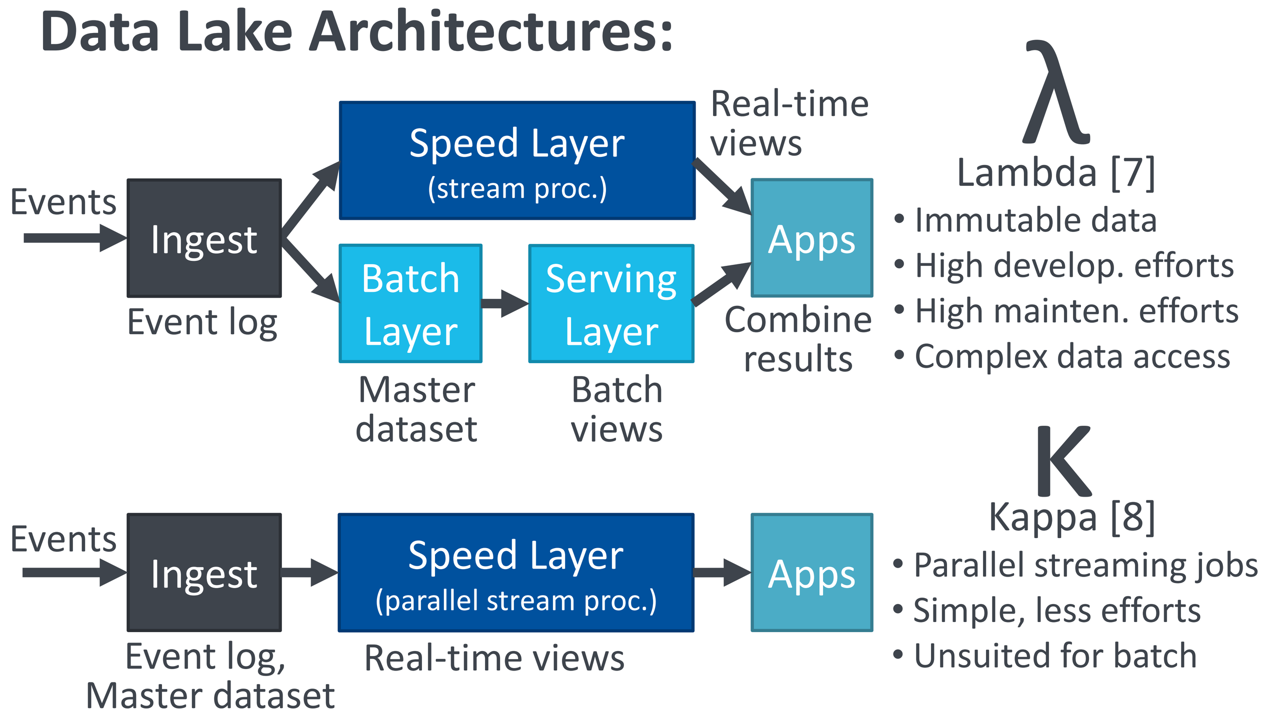
Data Catalog [4]:

- Metadata-based inventory of the available data
- „Search engine for data"
- Acquisition, storage, integration, search and provisioning of metadata
- Goals: Data transparency, discovery, understanding, revealing the interconnection of sources and processes

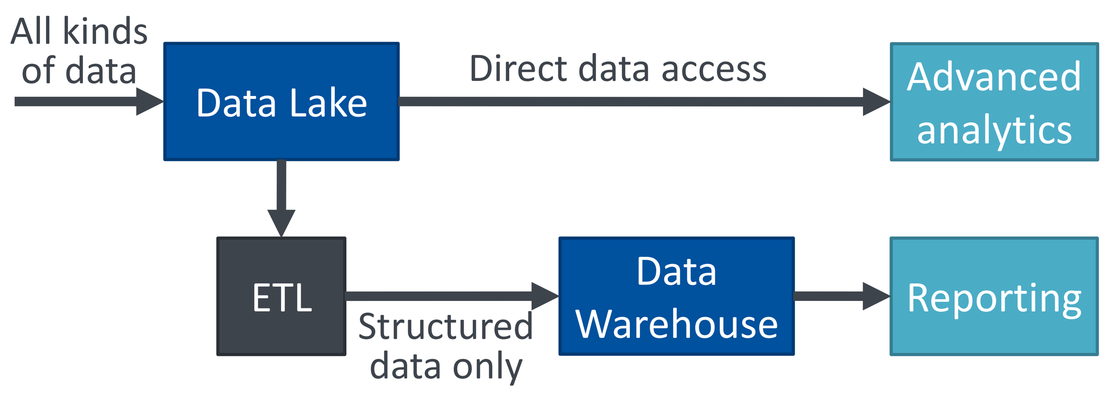
Enterprise Data Marketplace [5,6]:

- Metadata based self service platform connecting data producers and consumers to match supply and demand
- Goals: Data democratization, incentivisation, enforcing compliance → covering the entire data lifecycle
- Producer self-services: publishing, curation
- Consumer self-services: discovery, preparation

Data Lake Architectures:



2-Tier Architecture [9]:



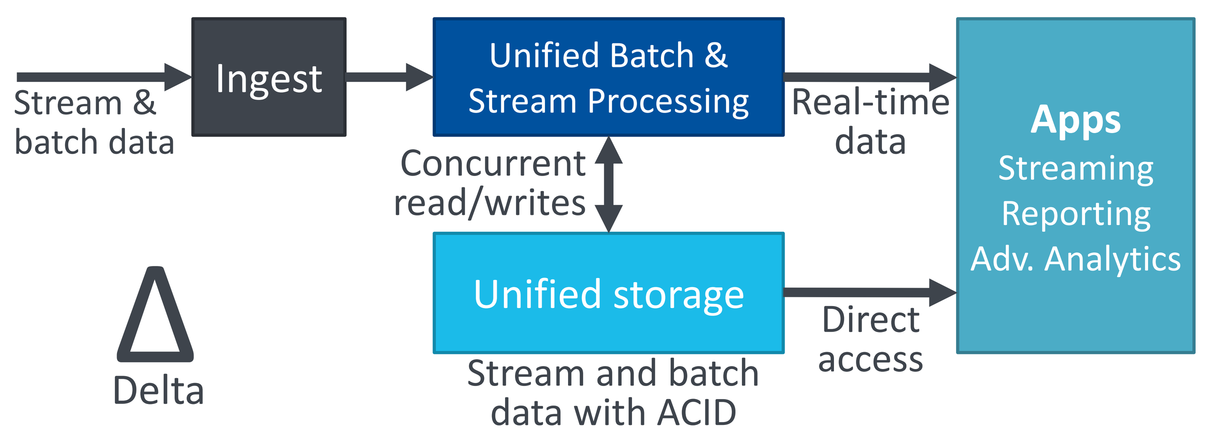
- Two tiers to manage and maintain → higher efforts, error-prone
- Additional ETL leads to delays → data becomes stale
- No single source of truth

Data Lakehouse [9]:

- Integrated platform combining the benefits of DWs and DLs → Streaming, Reporting and Adv. Analytics
- Low-cost data storage with metadata layer on top
- Open formats and direct data access
- Metadata enable ACID transactions, high performance, time travel and management features
- Concurrent batch and stream processing from and to data collections → enables Delta Architecture

Delta Architecture:

(partly [10])



- Processing reads from/writes to storage concurrently
- Processing engine uses same code for batch and streaming (e.g. Spark, Flink)
- Intermediate results are materialized in shared storage

References:

[1] Gröger, C. (2022). Industrial analytics–An overview. it-Information Technology.
[2] Inmon, W. H. (2005). Building the data warehouse. John Wiley & sons.
[3] Giebler, C. et al. (2019). Leveraging the data lake: Current state and challenges. In DaWak 2019.
[4] Zaidi, E. et al. (2017). Data catalogs are the new black in data management and analytics. Gartner Report.
[5] Gröger, C. (2021). There is no AI without data. Communications of the ACM, 64(11), 98-108.
[6] Eichler, R. et al. (2021). Enterprise-wide metadata management. In Business Information Systems.
[7] Warren, J. et al. (2015). Big Data: Principles and best practices of scalable realtime data systems. Manning.
[8] Kreps, J. (2014). Questioning the lambda architecture. Online article, oreilly.com .
[9] Armbrust, M. et al. (2021). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR
[10] Leano, H. (2020). Delta vs. Lambda: Why Simplicity Trumps Complexity for Data Pipelines. Online article.

www.ipvs.uni-stuttgart.de