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 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 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 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

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 Architecture:

2-Tier Architecture [9]:
- All kinds of data
- Structured data only
- Direct data access
- Advanced analytics
- ETL
- Batch does 
  - Two tiers to manage and maintain → higher efforts, error-prone
  - Additional ETL leads to delays when data becomes stale
  - No single source of truth

Delta Architecture:
- 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

Data Lakehouse [9]:
- Unified batch & stream data with ACID
- Concurrent read/write
- Delta
- Streaming, Reporting and Adv. Analytics

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

References: