University of Stuttgart
Institute for Parallel and Distributed Systems

Industrial Data Lab
A Cooperation of
BOSCH | Universität Stuttgart

Data Platform Architectures

Holger Schwarz and Jan Schneider

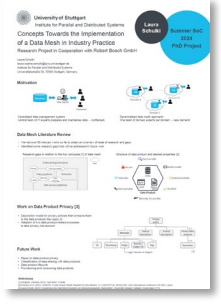
SummerSoC
Service Oriented Computing
18th Symposium and Summer School On Service-Oriented Computing
June 24th - 29th, 2024

Overview

- Data platforms – our understanding
- Data Warehouse → Data Lake → Data Lakehouse → Data Mesh
- Data Lakes
  - Architecture framework
  - Zone reference model
  - Zone implementation patterns

- Data Lakehouse (Jan)

- Data Mesh (Laura)



Background

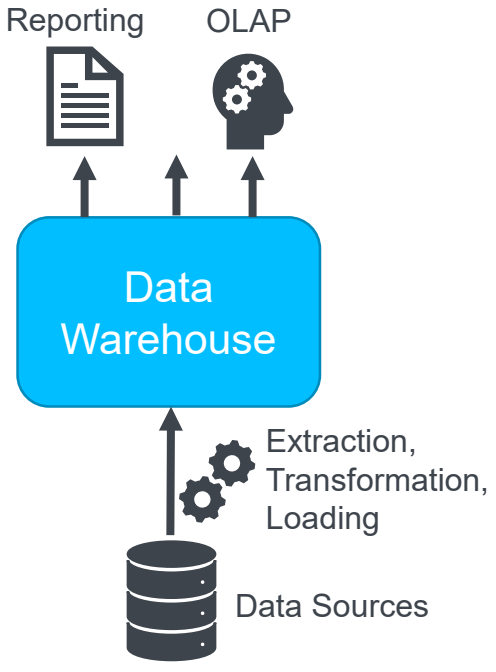
Data Platforms

- Data-driven analysis techniques allow enterprises to optimize their business processes
- Need for collecting, storing, organizing and processing huge amounts of data

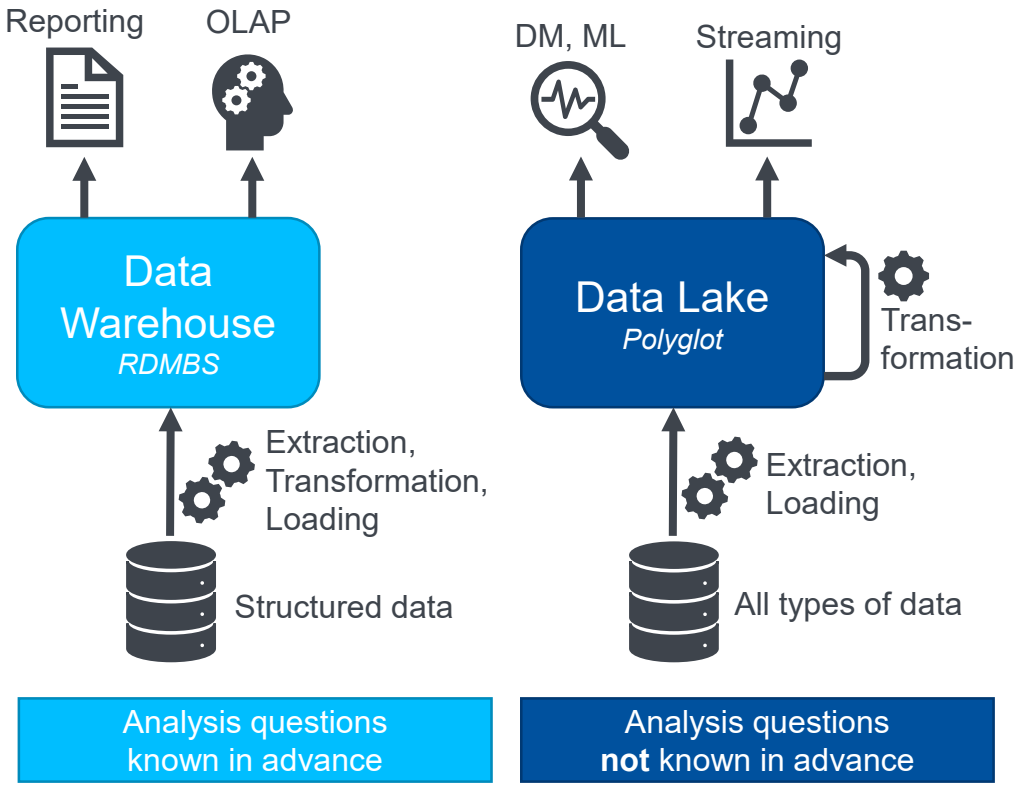
- Data Platform: Platform for managing data and metadata for analytical purposes
  → foundation for data collection, data processing and analytics applications

Data Warehouse

- Mature data platform
- Challenges
  - data volume
  - heterogeneous data sources
  - advanced analytics
  - need to define analytic use cases in advance



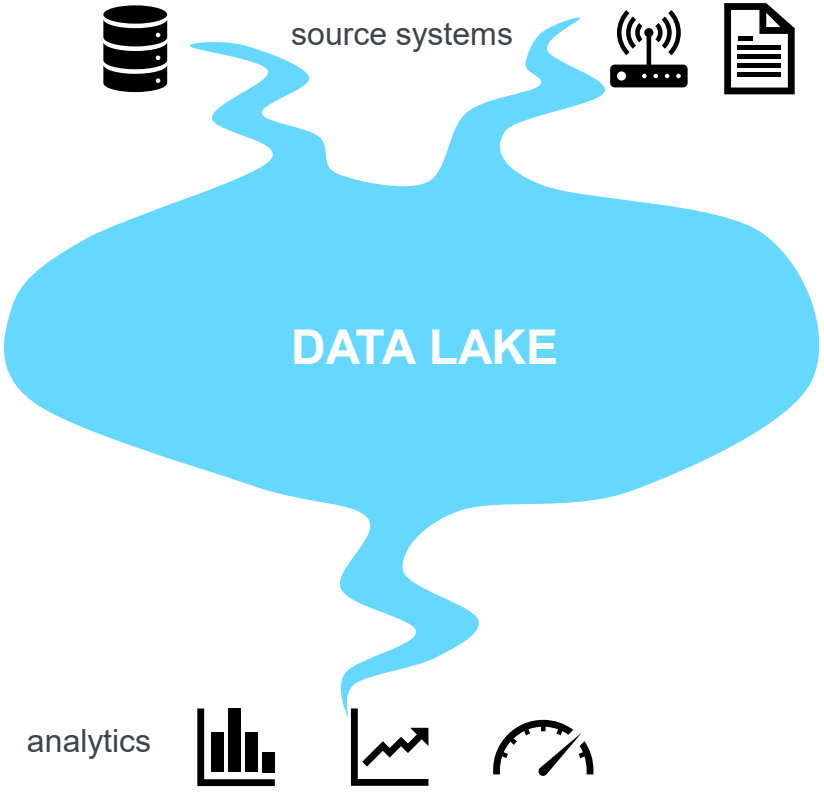
Data Warehouses & Data Lakes



Property	Data Warehouse	Data Lake
Workloads:	Reporting, OLAP	Advanced analytics
Users:	Business users, data analysts	Data scientists
Data Access:	Query language, data export	Direct access on storage
Guarantees:	ACID	Weak
Schema:	On-write	On-read
Data type:	Mainly structured	All types
Addressing:	Relational	Via metadata
Data granularity:	Aggregated	Raw and aggregated
Data Storage:	RDBMS	Polyglot
Flexibility:	Low	High
Mgt. features:	Advanced	Rudimentary

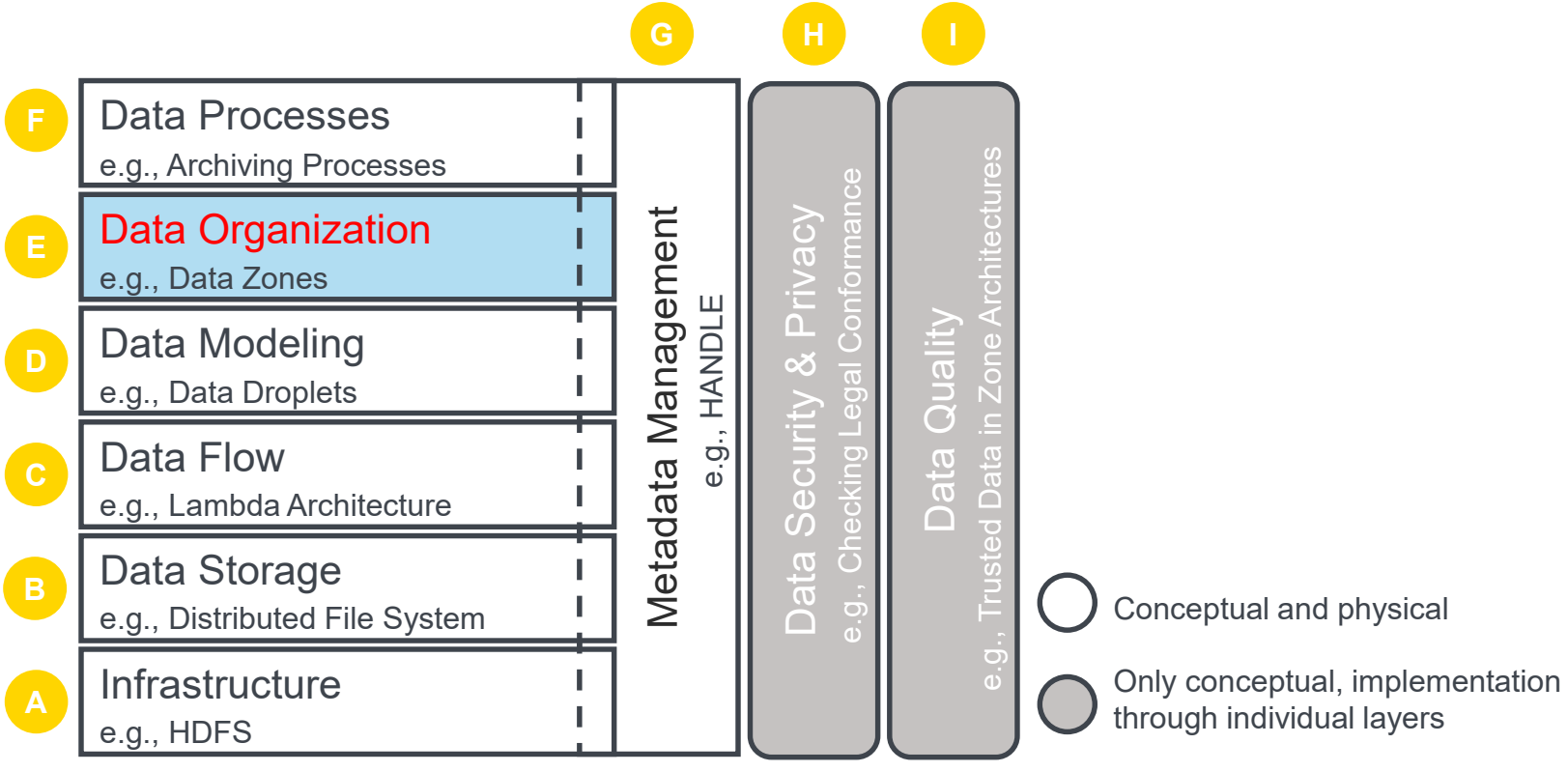
📖 Schneider, J. et al.: First Experiences on the Application of Lakehouses in Industrial Practice. Workshop GvDB, Herdecke 2024.

Data Lake



- Data lake as supplement to data warehouses
  - fexible analytics without predefined use cases
  - heterogeneous data in raw format
- **How to build a data lake?**
  - vague, abstract and inconsistent literature
  - only few best practices

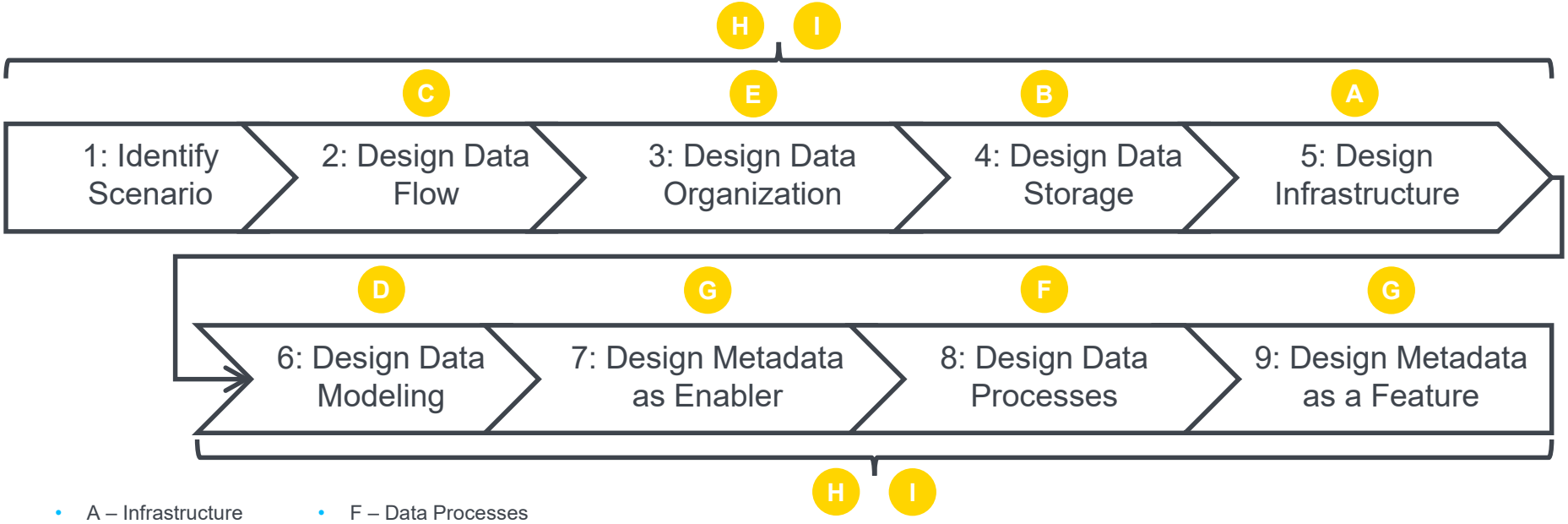
Data Lake Architecture Framework



📖 Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., Mitschang, B.: The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture. In: Proceedings der 19. Fachtagung Datenbanksysteme für Business, Technologie und Web (2021)

Data Lake Architecture Framework

Methodology

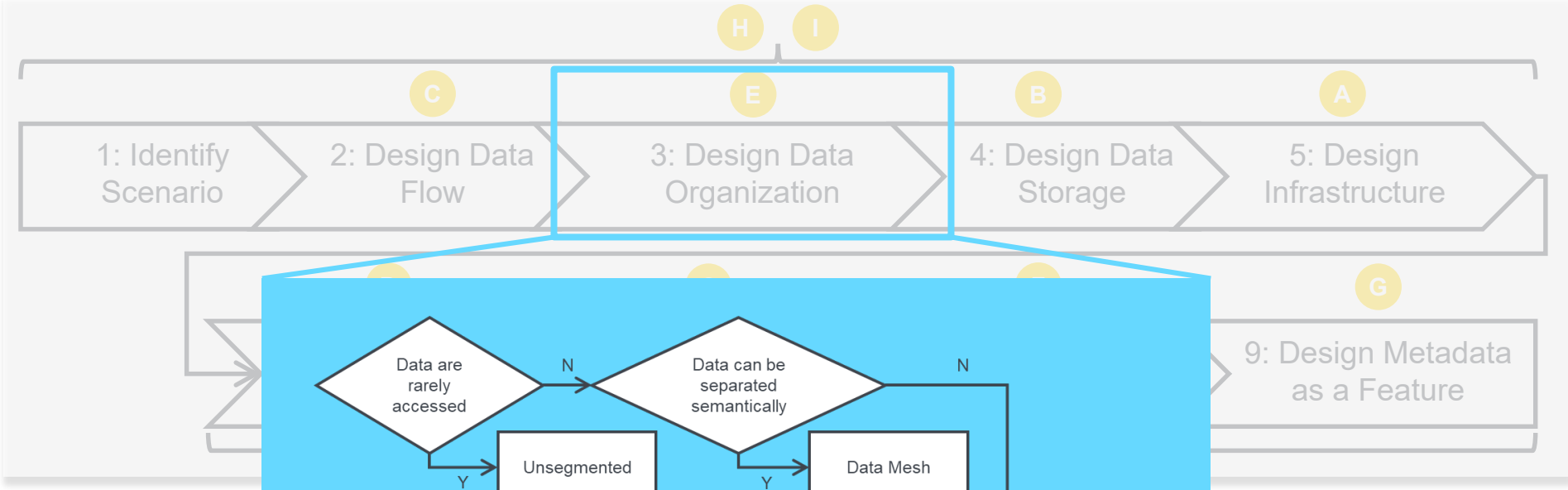


- A – Infrastructure
- B – Data Storage
- C – Data Flow
- D – Data Modeling
- E – Data Organization
- F – Data Processes
- G – Metadata
- H – Data Security & Privacy
- I – Data Quality

📖 Giebler, C., Gröger, C., Hoos, E., Eichler, R., Schwarz, H., Mitschang, B.: The Data Lake Architecture Framework: A Foundation for Building a Comprehensive Data Lake Architecture. In: Proceedings der 19. Fachtagung Datenbanksysteme für Business, Technologie und Web (2021)

Data Lake Architecture Framework

Methodology

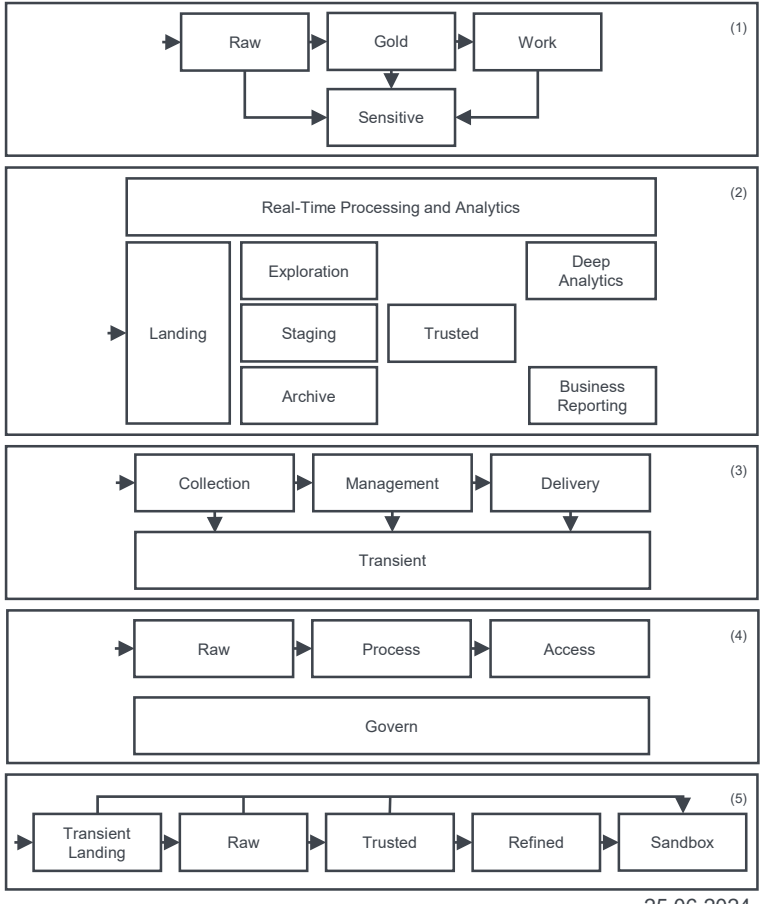


- A – Infrastructure
- B – Data Storage
- C – Data Flow
- D – Data Modeling
- E – Data Organiza

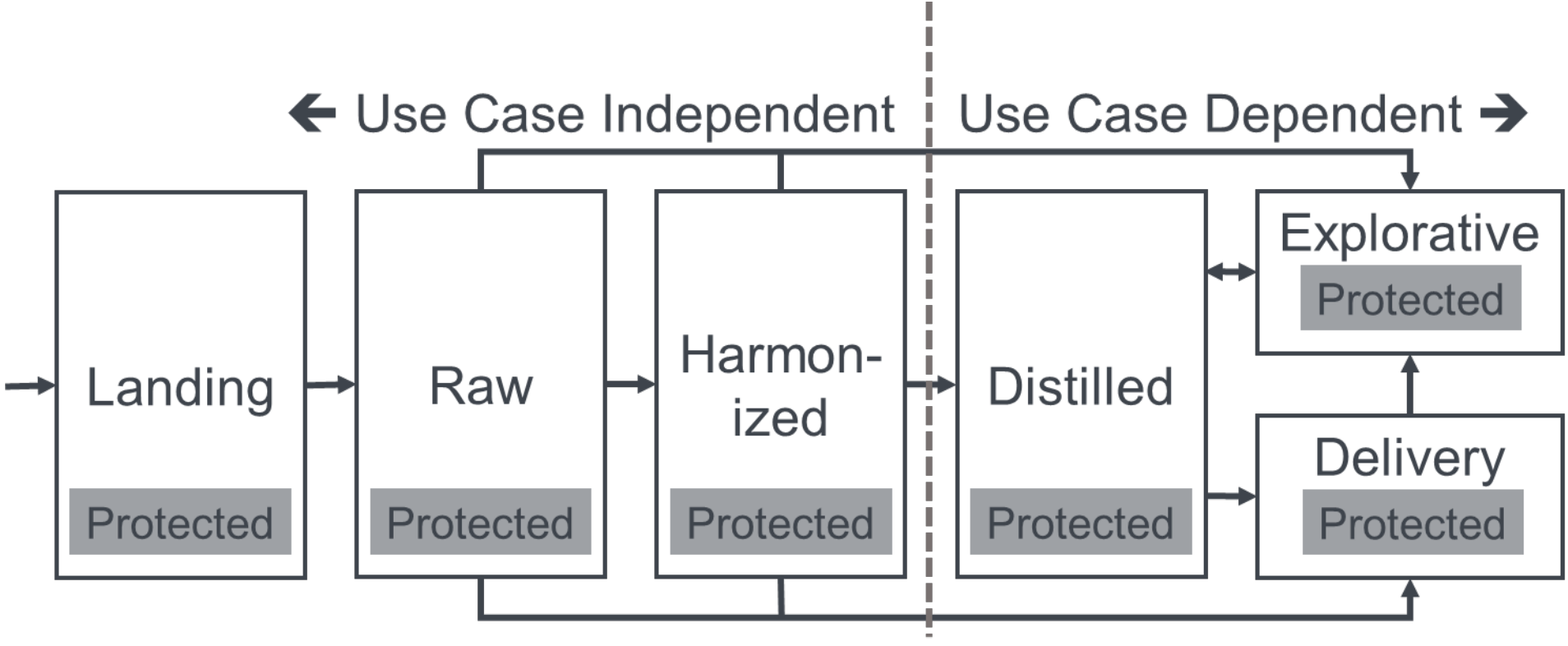
📖 Giebler, C., Grög … A Foundation for Building a Comprehensive Data… …logie und Web (2021)

Zone Architectures for Data Lakes

- Motivation for zone models
  - need to organize raw data and pre-processed data
  - organize data by its characteristics
    - degree of processing, degree of applied governance, …
  - reuse of data integration, data transformation, data models, and more across use cases
- Challenges
  - many proposals for zone architectures
  - no guidance for their implementation



Zone Reference Model



🕮 Giebler, C., Gröger, C., Hoos, E., Schwarz, H., Mitschang, B.: A Zone Reference Model for Enterprise-Grade Data Lake Management.
In: Proceedings of the 24th IEEE Enterprise Computing Conference (2020)

Zone Reference Model

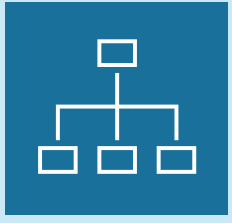
	Landing	Raw	Harmonized	Distilled	Explorative	Delivery
Granularity (Raw – Aggregated)	Raw	Raw	Raw	Aggregated	Any	Any
Schema (Any – Consolidated)	Any	Any	Consolidated	Consolidated, enriched	Any	Any
Syntax (Unchanged – Consolidated)	Basic transformations	Basic transformations	Consolidated	Consolidated	Any	Any
Semantics (Unchanged – Processed)	Mostly unchanged, unless needed for compliance	Mostly unchanged, unless needed for compliance	Mostly unchanged, unless needed for compliance	Complex processing	Any	Any
Properties	Governed, non-historized, non-persistent, protected part, use case independent	Governed, historized, persistent, protected part, use case independent	Governed, historized, persistent, protected part, use case independent	Governed, historized, persistent, protected part, use case dependent	Not governed, non-persistent, protected part, use case dependent	Governed, persistent, protected part, use case dependent
User Groups	Systems, processes	Data scientists, systems, processes	Data scientists, systems, processes	Data scientists, domain experts, systems, processes	Data scientists	Any human users, systems, processes
Modeling Approach	Any	Any	Standardized	Standardized	Any	Any

📖 Giebler, C., Gröger, C., Hoos, E., Schwarz, H., Mitschang, B.: A Zone Reference Model for Enterprise-Grade Data Lake Management. In: Proceedings of the 24th IEEE Enterprise Computing Conference (2020)

Zone Implementation Patterns

Categories of Patterns

Zone Structure Patterns



- How is the zone model represented in the Data Lake?
- How does one know which data belong to which zone?
- How are zones separated from each other?

Zone Storage Patterns

- How do zones refer to the underlying storage?
- What kinds of storage systems are used?
- How do storage systems interact within the zone model?

Zone Data Flow Patterns

- How are streaming data handled in the zone model?
- What zones apply to streaming data?
- How does streaming data interact with batch data?

📖 Giebler, C.; Gröger, C.; Hoos, E.; Schwarz, H.; Mitschang, B.: Implementation Patterns for Zone Architectures in Enterprise-Grade Data Lakes. In: Proceedings 36th International Conference Advanced Information Systems Engineering - CAiSE 2024, Limassol, Cyprus, June 3-7, 2024

Zone Implementation Patterns

Zone Structure

	Separation	Complexity	Latency	Centrality
Structure Through Systems	+	-	-	-
Structure Through Containers	+	+	+	-
Structure Through Metadata	-	+	+	+

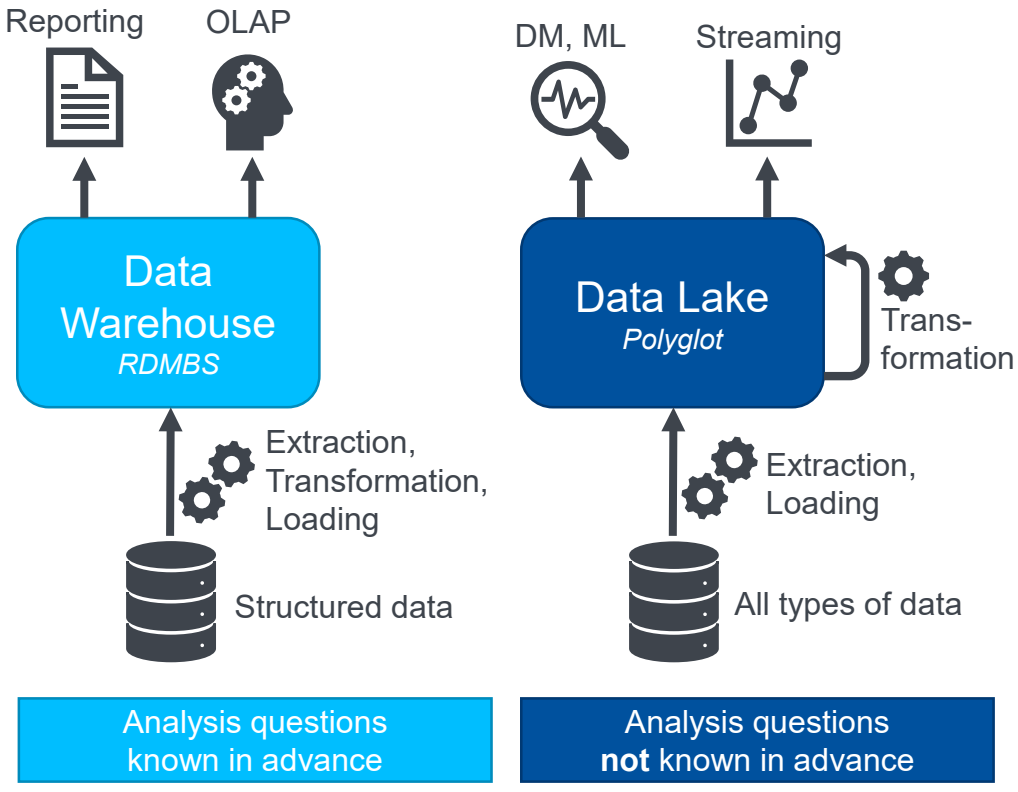
Zone Storage

	Management Functionality	Complexity	Latency	Redundancy
Single Storage	-	+	+	+
Polyglott – Data Oriented	+	o	o	o
Polyglott – Usage Oriented	+	o	o	o
Polyglott – Best Fit	+	-	o	o

Zone Data Flow

	Intermediate Results Available	Complexity	Latency
Streaming Zone	-	+	+
Zone-Based Architecture for Streaming	+	-	-

Data Warehouses & Data Lakes



Property	Data Warehouse	Data Lake
Workloads:	Reporting, OLAP	Advanced analytics
Users:	Business users, data analysts	Data scientists
Data Access:	Query language, data export	Direct access on storage
Guarantees:	ACID	Weak
Schema:	On-write	On-read
Data type:	Mainly structured	All types
Addressing:	Relational	Via metadata
Data granularity:	Aggregated	Raw and aggregated
Data Storage:	RDBMS	Polyglot
Flexibility:	Low	High
Mgt. features:	Advanced	Rudimentary

📖 Schneider, J. et al.: First Experiences on the Application of Lakehouses in Industrial Practice. Workshop GvDB, Herdecke 2024.

Data Warehouses & Data Lakes

However, modern enterprises require …

Capabilites of **Data Warehouses**

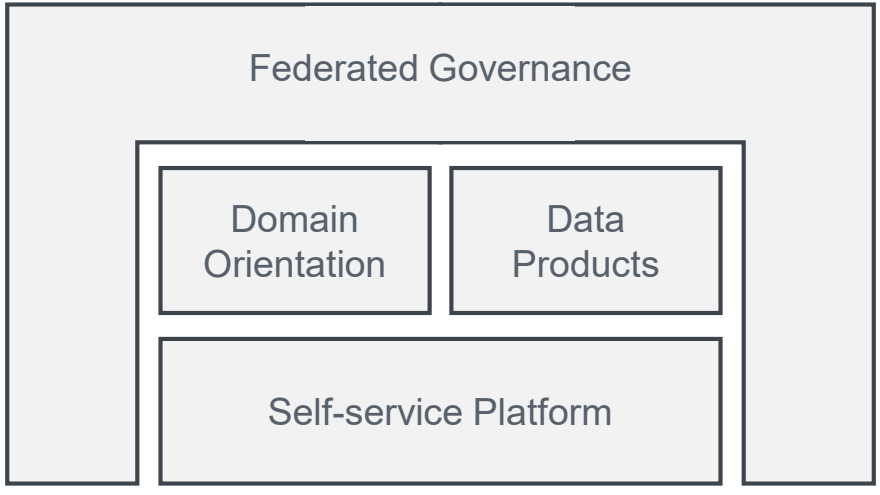
Capabilites of **Data Lakes**

„data lakehouse“

Data Mesh

Motivation

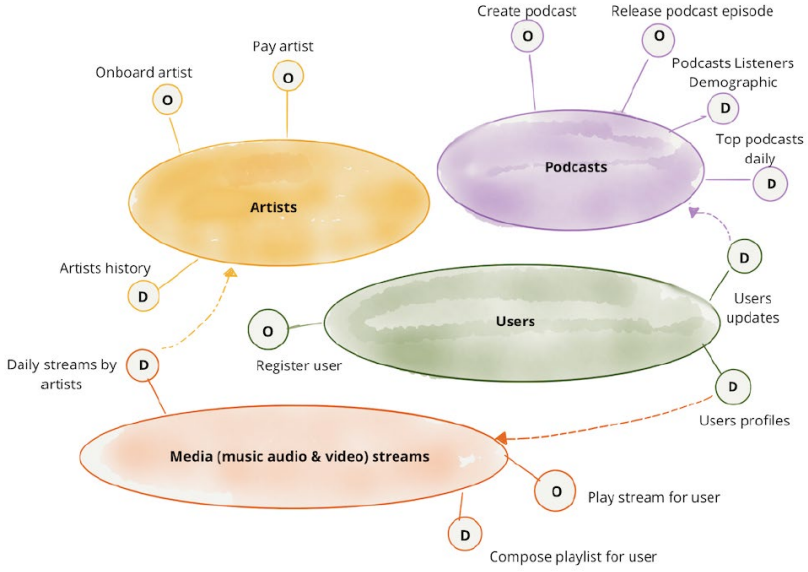
- Data platforms and how they should scale
  - volume of data
  - volume/complexity of data processing
  - changes in the data landscape
  - proliferation of sources of data
  - diversity of data use cases and users
  - speed of response to change
- Goal: support continuous change and scalability



Data Mesh

Principle I: Domain Orientation

- Decompose and decentralize the components of the data ecosystem
- Domains own operational IT systems, analytical IT systems and their data
- Domains provide endpoints for
  - analytical data
  - operational capabilities
- Dependencies between domains
- Decomposition approaches
  - organizational units, business functions, source oriented, consumer oriented

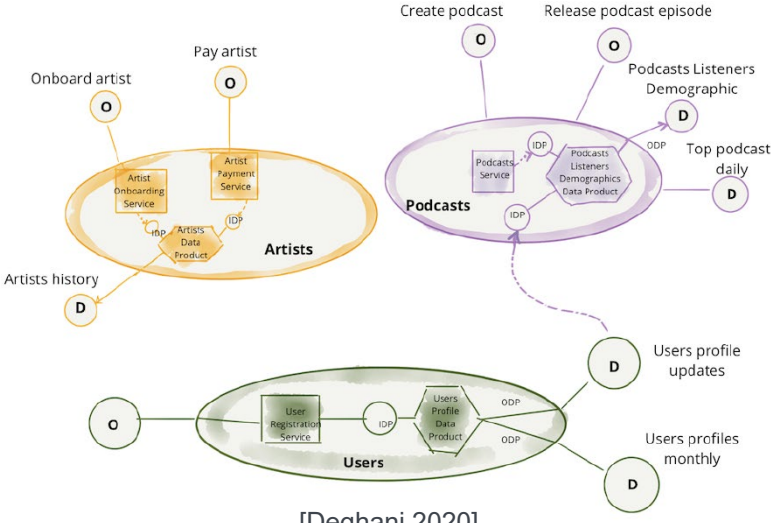


📖 Dehghani (2020): Data Mesh Principles and Logical Architecture, https://martinfowler.com/articles/data-mesh-principles.html

Data Mesh

Principle II: Data Products

- Components of a data product: data and metadata, code, infrastructure
- Data and metadata
  - data served as graph, batch file, relational table, … (depending on domain)
  - metadata e.g. on sytax and semantics, data quality, access contol, …
- Code comprises
  - code for data pipelines
  - code for data and metadata access
  - code for enforcing properties
- Infrastructure to build, deploy and run the code



[Deghani 2020]

- Additional role in domains: data product developer

Data Mesh

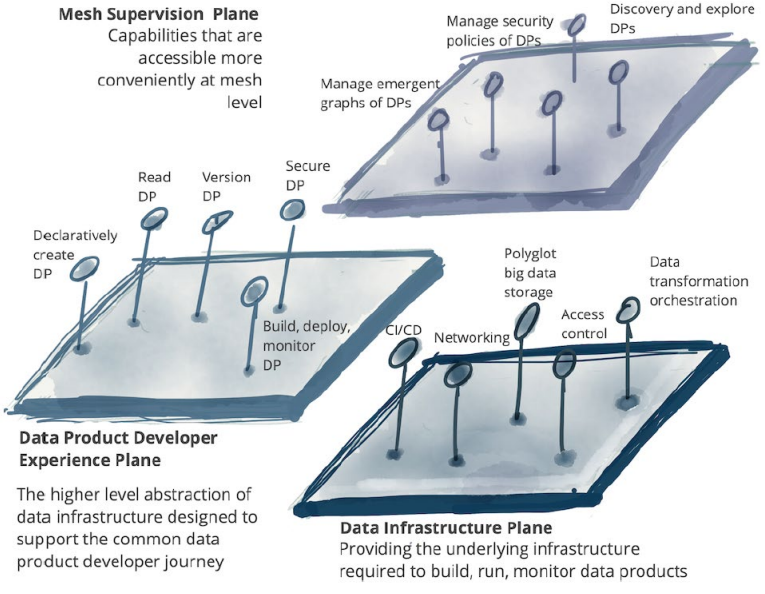
DAUTNIVS capabilities of data products

- Components of a data product: data and metadate, code, infrastructure
- Data and metadata
  - data served as graph, batch file, relational table, … (depending on domain)
  - metadata e.g. on sytax and semantics, data quality, access contol, …
- Code comprises
  - code for data pipelines
  - code for data and metadata access
  - code for enforcing properties
- Infrastructure to build, deploy and run the code

Data Mesh

Principle III: Self-Service Platform

- Should support domain data product developers in creating, maintaining and running data products
- Domains are less relying on central IT
  → supports domain autonomy
- Groups of related capabilities based on profile of users (planes)
  - data infrastructure
  - data product developer experience
  - data mesh supervision

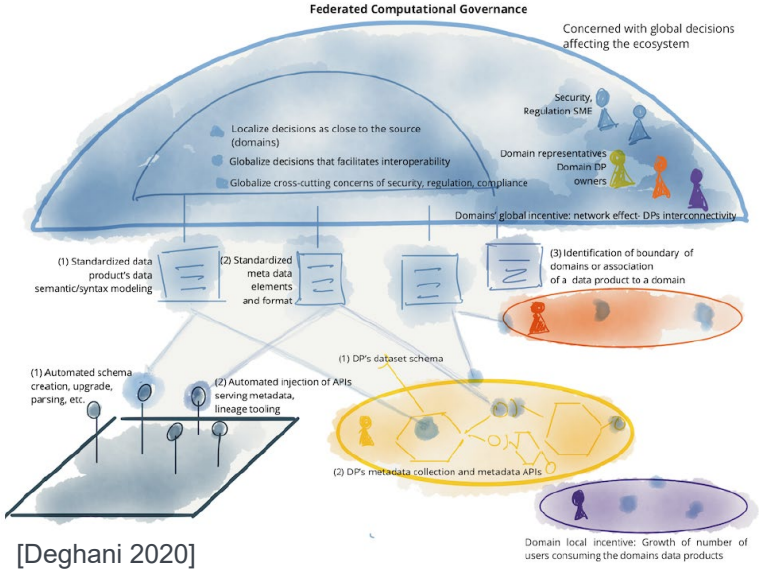


[Deghani 2020]

Data Mesh

Principle IV: Federated Governance

- Independent data products need to interoperate
- Governance model needs to support
  - decentralization and domain self-sovereignty
  - interoperability through global standardization
  - dynamic topology
  - automated execution of decisions
- Decision model has to consider
  - autonomy of domain data product owners and data platform product owners
  - set of global rules applied to all data products and their interfaces
    → ensure a healthy and interoperable ecosystem



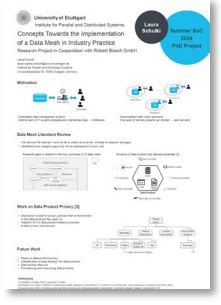
[Deghani 2020]

Overview

- Data platforms – our understanding
- Data Warehouse → Data Lake → Data Lakehouse → Data Mesh
- Data Lakes
  - Architecture framework
  - Zone reference model
  - Zone implementation patterns

- Data Lakehouse (Jan)

- Data Mesh (Laura)



University of Stuttgart
Institute for Parallel and Distributed Systems

Thank you!

Prof. Dr. Holger Schwarz

e-mail Holger.Schwarz@ipvs.uni-stuttgart.de
phone +49 (0) 711 685-88-424
www.ipvs.uni-stuttgart.de/institute/team/Schwarz

University of Stuttgart
IPVS / AS
Universitätsstraße 38
70569 Stuttgart, Germany