Roche Diagnostics Data Mesh and Data Vault Journey

Two systems working in harmony

Paul Rankin
Roche Diagnostics

September 2022

Roche

Table of contents

Roche

Swiss multinational healthcare company that operates under two divisions:
Pharma and Diagnostics

The world's largest biotech
company

125-year history

Revenue: 62.8 billion CHF

Number of employees:
100,920

How the journey started

18 months ago my boss proposed a new approach within Roche to delivering analytics

Could Data Mesh really help Roche?
We needed to find out more about Data Mesh and what changes were required

Hi Zhamak, can you please help us how to implement Data Mesh?

Roche

What is Data Mesh?

What is data mesh?
Data mesh is a sociotechnical paradigm

- Data mesh is a decentralized sociotechnical approach to share, access, and manage analytical data in complex and large-scale environments—within or across organizations.

- Data mesh, at core, is founded in *decentralization* and *distribution of responsibility* to people who are closest to the data in order to support continuous change and scalability.

- Data mesh expects independent teams to own and serve their analytical data. It expects this data to be served as a product, accompanied by behavior that enriches the experience of the data consumer, to discover, trust, and use it for multiple purposes.

Zhamak Dehghani

Director of Emerging Technologies
ThoughtWorks, North America

Principles of Data Mesh

Four simple principles capture what underpins data mesh's logical architecture and operating model

Decentralize the ownership of analytical data to business domains closest to the data

Domain-oriented data is shared as a product directly with data users

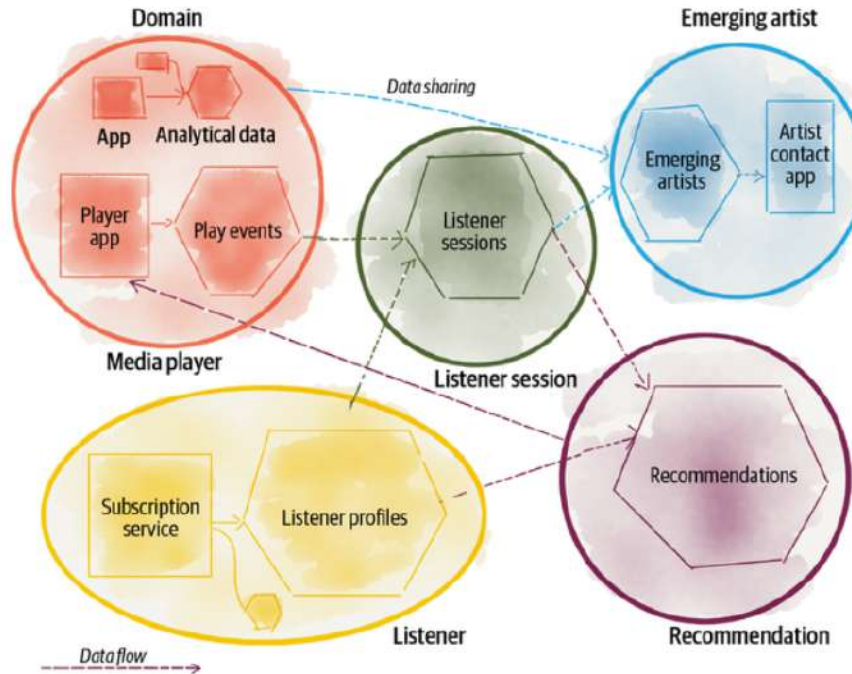Empower domains to build their own pipelines from source to consumption

A data governance operating model based on a federated decision-making structure

These principles are designed to progress us toward the objectives of data mesh: increase value from data at scale, sustain agility as an organization grows, and embrace change in a complex and volatile business context
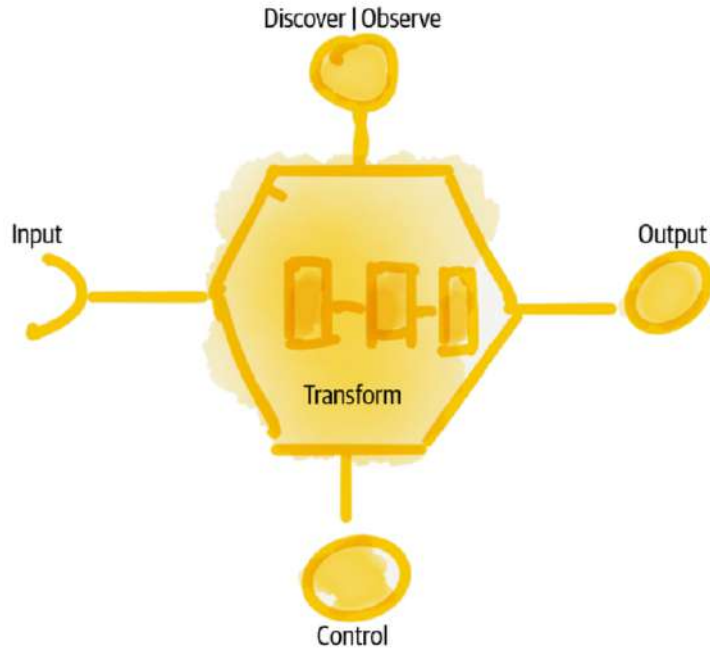
Domain-oriented ownership

Founded in decentralization and distribution of data responsibility to people closest to the data

Data as a product
Applying Product Thinking to Data

Discover | Observe

Input

Output

Transform

Control

Data Product Characteristics

Secure

Discoverable

Valuable (on its own)

Addressable

Interoperable

Understandable

Natively Accessible

Truthful (Trustworthy)

Images source: Data Mesh, O'Reilly 2022

10

Self-serve data platform

Extracting and harvesting domain agnostic infrastructure into a separate data platform

Self-serve Data Platform
addresses:
- Duplication of efforts in each domain
- Increased cost of operation
- Likely large-scale inconsistencies
- Incompatibilities across domains

Images source: Data Mesh, O'Reilly 2022
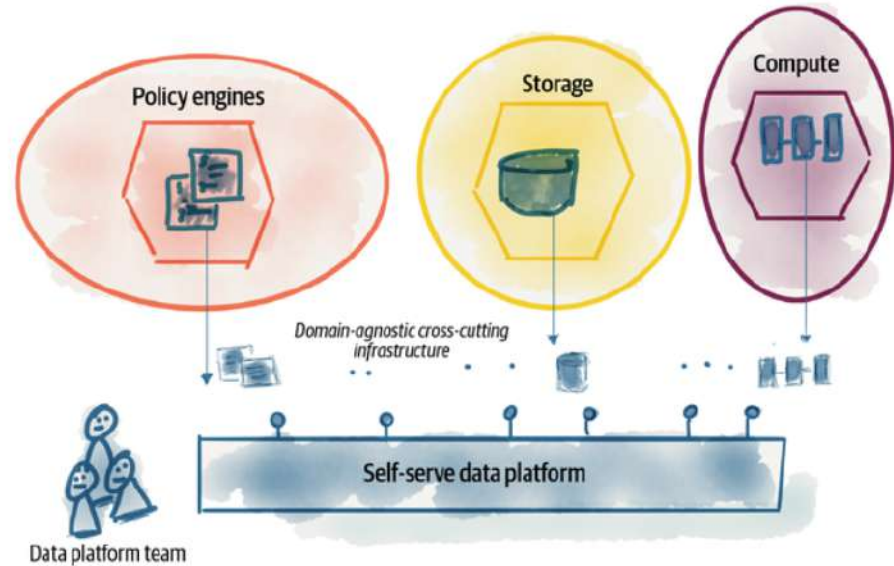
Federated computational governance

The definition of data governance can often be understood to mean many things
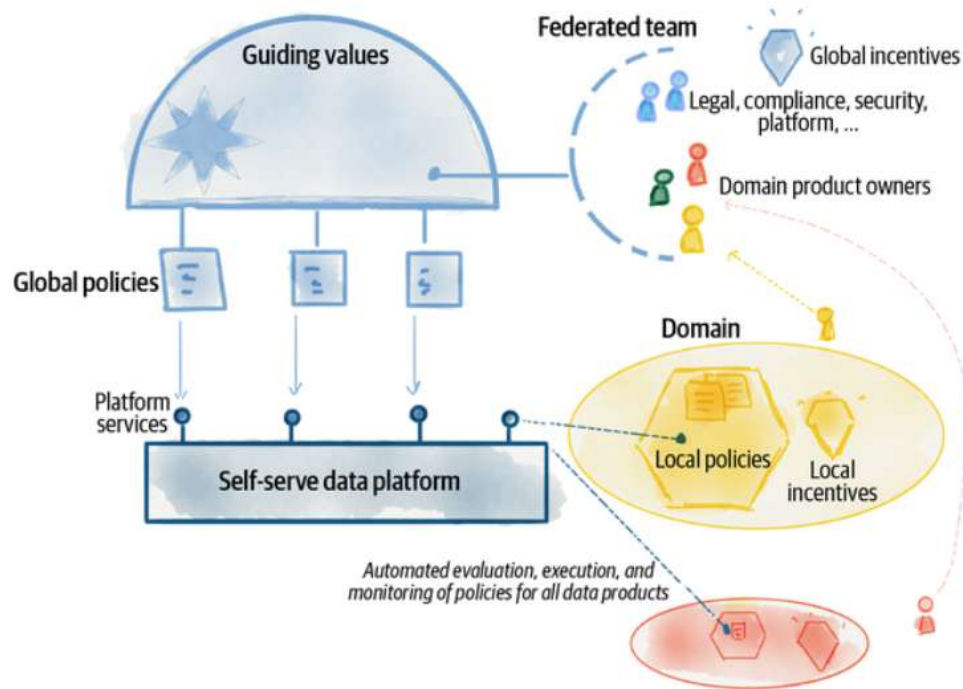
to steer and
guide

to rule with authority

Data Product Teams Responsibility:
Ensure the availability of safe, high quality, consistent, compliant, privacy-respecting, and usable data across an organization with managed risk
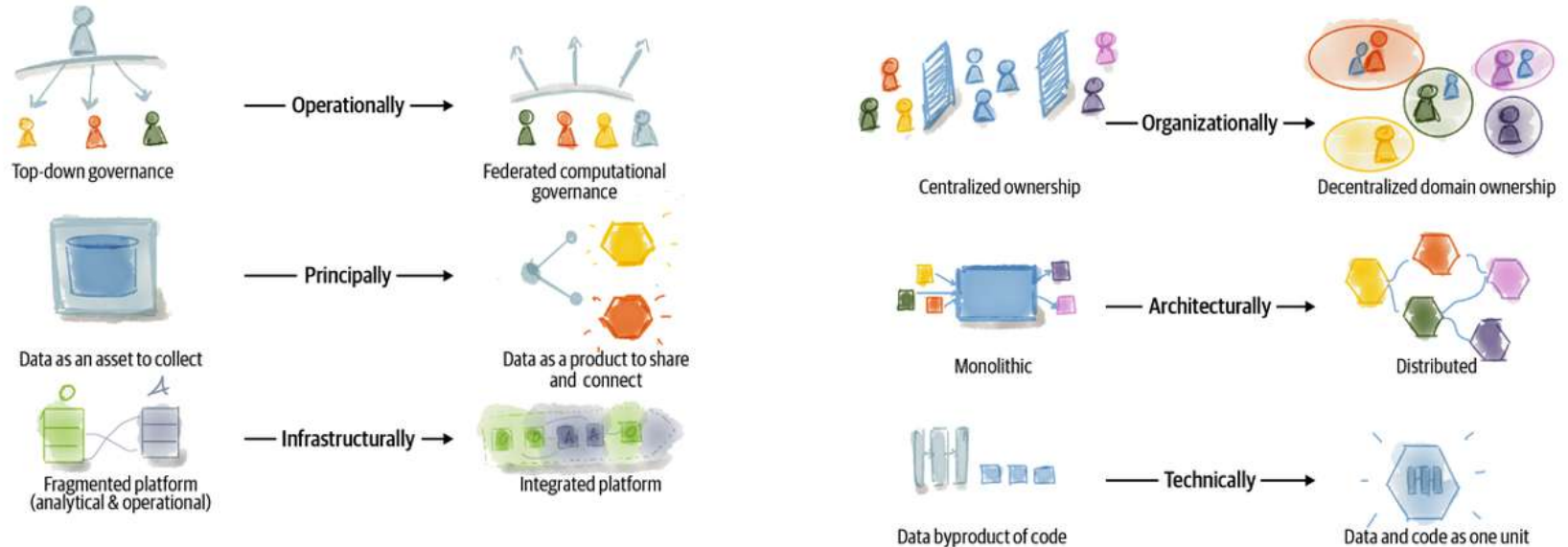
Federated computational governance

Data mesh proposes a governance operating model that benefits from federated decision making

# Multidimensional technical and organizational shifts

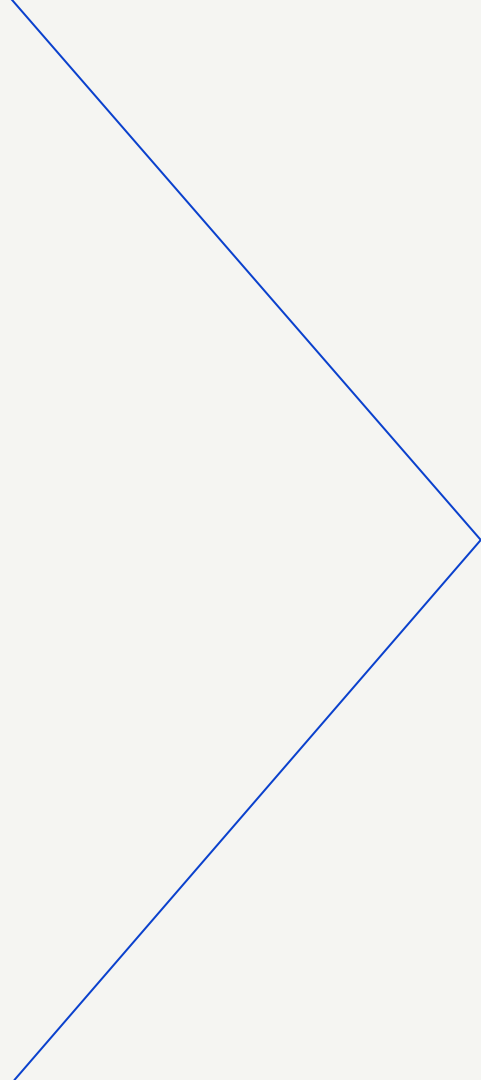Let's summarize the shifts that data mesh introduces compared to past approaches

Top-down governance — Operationally → Federated computational governance

Data as an asset to collect — Principally → Data as a product to share and connect

Fragmented platform (analytical & operational) — Infrastructurally → Integrated platform

Centralized ownership — Organizationally → Decentralized domain ownership

Monolithic — Architecturally → Distributed

Data byproduct of code — Technically → Data and code as one unit

Images source: Data Mesh, O'Reilly 2022

14

Was Roche ready for Data Mesh?

Yes, we believed that Roche was at a level of maturity to benefit from Data Mesh
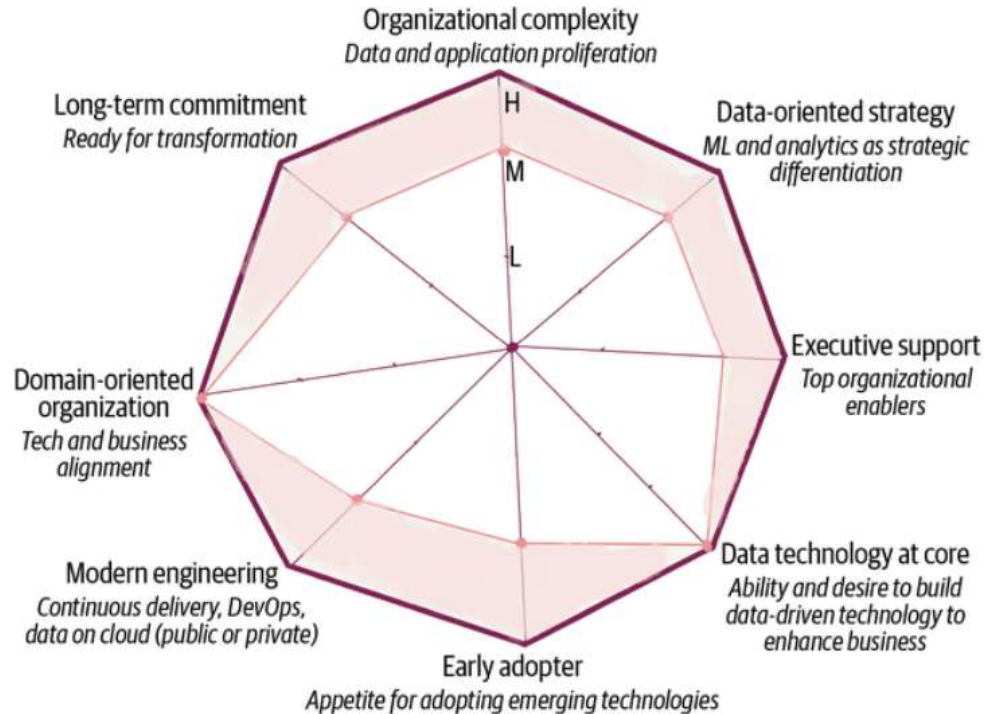
15

Roche

Data Vault and Data Mesh -
Synergies and Conflicts

Let's look at what the authors say

On face value both approaches seem to be trying to solve similar problems

Data Mesh is a new approach in sourcing, managing, and accessing data for analytical use cases at scale
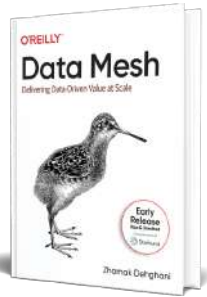
Zhamak Dehghani

Data Vault allows the ability to adapt quickly, model the business accurately, and scale with the business needs

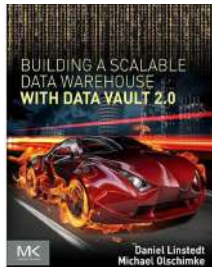***Michael Olschimke & Dan Linstedt***

Let's look at what the authors say
A closer look highlights some obvious conflicts

I often get asked if the existing warehouses, lakes, and lakehouses that the organizations have created can coexist with the mesh. The short answer is transitionally yes, but ultimately no.

Zhamak Dehghani

Data Vault represents a system of Business Intelligence. A number of aspects that relate to the business of designing, implementing and managing a data warehouse.

Dan Linstedt & Michael Olschimke

Compare and contrast
Let's try to break it down and compare each component

Data Vault 2.0

- Data Vault Architecture
- Data Vault Methodology
- Data Vault Modeling
- Data Vault Implementation

"Each of these components plays a key role in the overall success of an enterprise data warehousing project"

Data Mesh

- Domain Ownership
- Data as a Product
- Self-Serve Data Platform
- Federated Computational Governance

"Interplay of the principles are collectively necessary and sufficient to addresses new challenges that may arise"
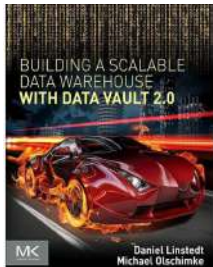
Architecture
Both books have dedicated sections on architecture

The ultimate goal of the architecture is to remove and reduce the overall level of *centralization*. Hence, ultimately a mesh implementation and a *central* data warehouse or lake should not coexist.

Zhamak Dehghani

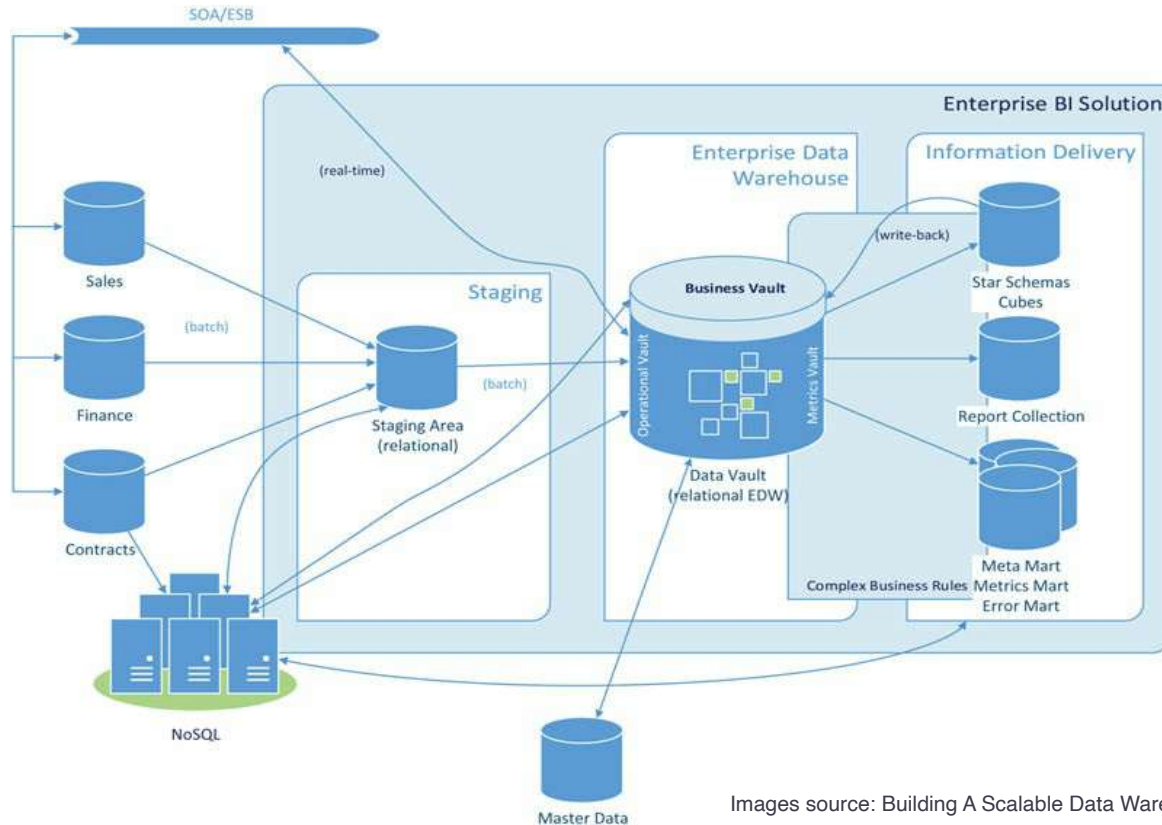The Data Vault 2.0 architecture addresses the extensibility and dimensions of scalability including workload, complexity, availability and security.

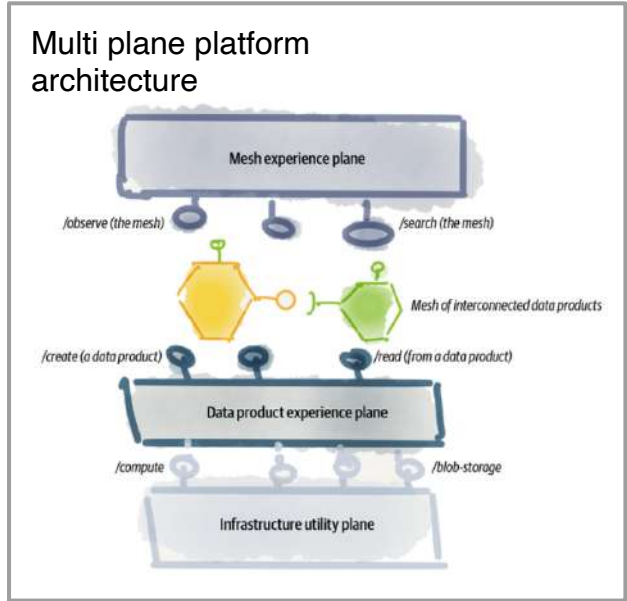***Dan Linstedt & Michael Olschimke***

Architecture
Data Vault 2.0 presents an alternate three layer architecture

Images source: Building A Scalable Data Warehouse, Morgan Kaufmann 2016

Architecture

Data Mesh presents a distributed architecture and multi-plane self-serve data platform
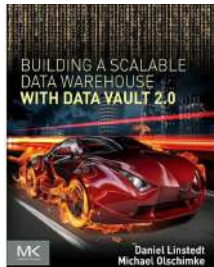

Multi plane platform architecture
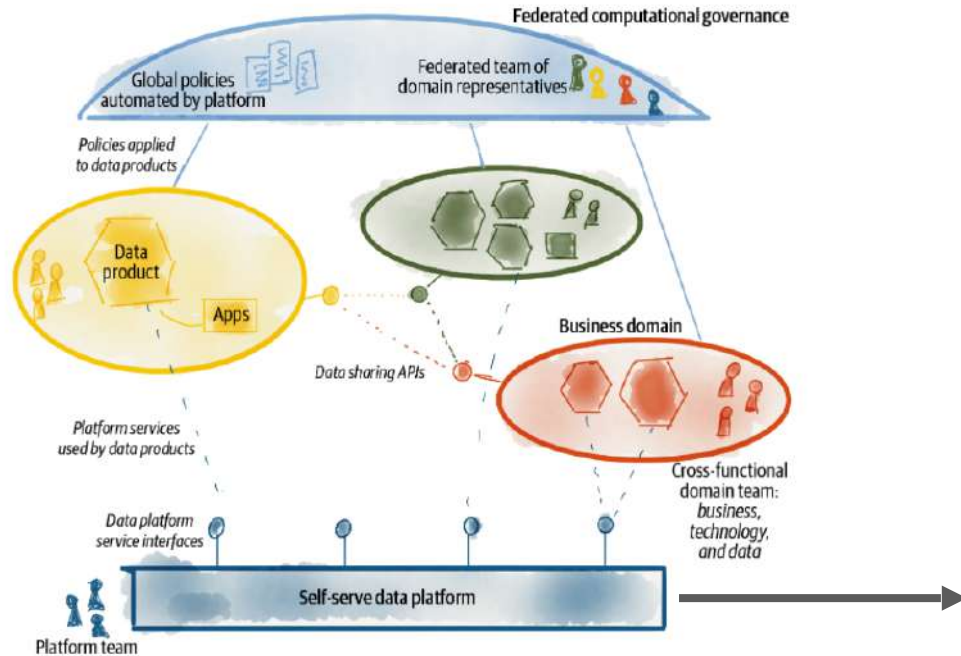
Methodology
There is no mention of methodology in Data Mesh

Data Mesh does not talk about methodology but a change in organization and culture adopting the movement-based change approach.

Zhamak Dehghani
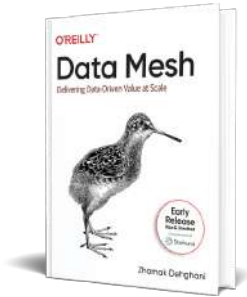
The Data Vault 2.0 Methodology provides best practice for project execution including project planning, project execution, review and improvement.

Dan Linstedt & Michael Olschimke

Modeling
There is no real mention of data modeling in Data Mesh

The transformation code is domain-specific and encapsulates tasks such as a domain's business logic as well as aggregating and modeling data.
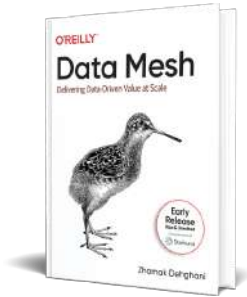
Zhamak Dehghani
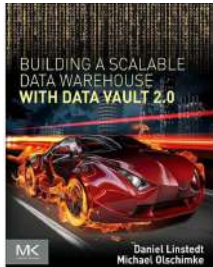
The Data Vault model was indeed invented by Dan Linstedt and considered by most industry experts to be the only truly scalable data modeling technique available today.

Dan Linstedt & Michael Olschimke

Implementation
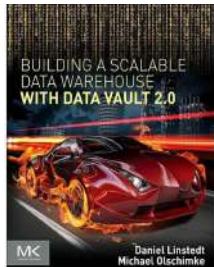Data Vault 2.0 presents best practice implementation

Data Mesh presents a high level Execution framework but does not go into any great detail about an implementation strategy.

Zhamak Dehghani

The Data Vault 2.0 Implementation deep dives into how to make use of pattern based automation to ensure code standards are met, data quality is maintained and pipelines are stable.
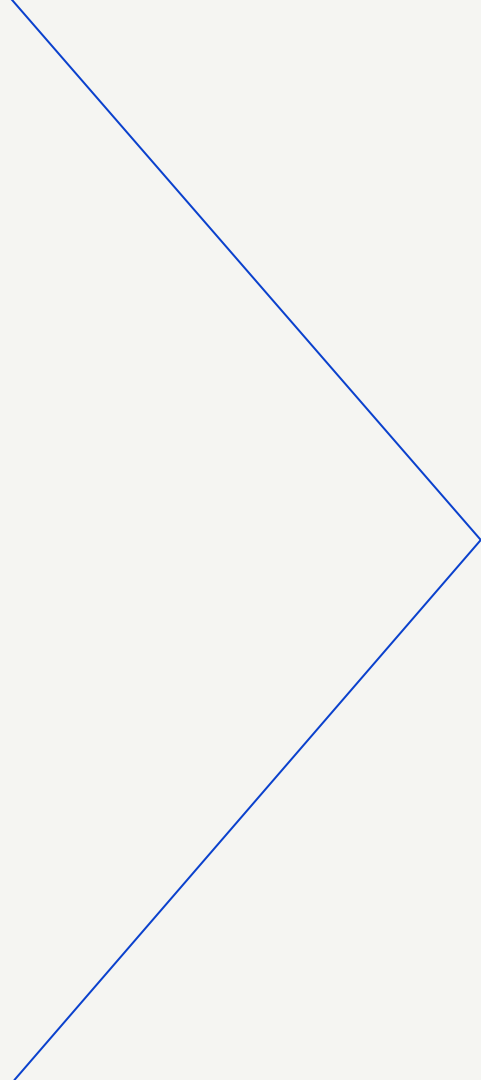
Dan Linstedt & Michael Olschimke

Conclusion
After many months of internal and external discussion

- Both trying to solve similar problems
- You cannot really compare a change in Organizational and culture to a data modeling technique and implementation methodology
- What you can compare are the different logical architectures
- Redefine the ownership and boundaries of an enterprise data warehouse
- Business keys are the only way to ensure interoperability
- Pattern based automation is key to scaling
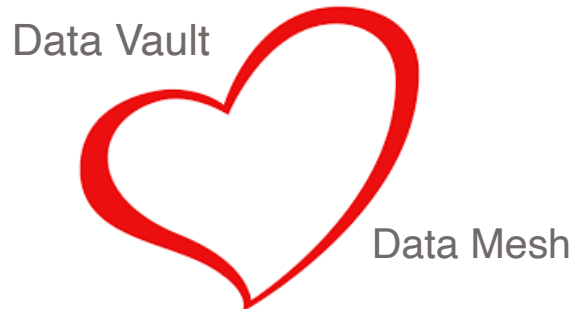
Data Vault

Data Mesh

Data Vault Modeling for Data Mesh

Pre Data Mesh

Analytics was delivered in a traditional IT as producers and Business as consumers way

CENTRAL IT TEAM

DEVELOPMENT

OPERATIONS

DOMAIN/AFFILIATE BI TEAMS

MANAGEMENT REPORTING

DATA SCIENCE TEAM

Pre Data Mesh

We were already following the Data Vault modeling approach with some degree of success

PSA_SALESFORCE
▾ Tables
- ACCOUNT
- CASE
- PRODUCT
- TERRITORY

PSA_SAP
▾ Tables
- SALES
- PRODUCT
- CUSTOMER
- COMPANY

RAW_DV
▾ Tables
- HUB_PRODUCT
- HUB_CUSTOMER
- HUB_COMPANY
- HUB_ACCOUNT
- HUB_CASE
- HUB_TERRITORY
- SAT_PRODUCT_SAP
- SAT_CUSTOMER_SAP
- SAT_CUSTOMER_SFDC
- SAT_CASE
- SAT_TERRITORY
- LSAT_TERRITORY_SALES
- LNK_TERRITORY_SALES
- XLNK_SALES
- LNK_CASE_ACCOUNT
- SAT_PRODUCT_SFDC

BUSINESS_DV
▾ Tables
- SLAT_CUSTOMER
- BSAT_TERRITORY_HIERARCHY
- PIT_TERRITORY_SALES
- LNK_CASE_ACCOUNT
- BRG_SALES
- BRG_CASE_DETAILS

DM_SALES
▾ Tables
- DIM_CUSTOMER
- DIM_TERRITORY
- DIM_PRODUCT
- FACT_SALES

DM_SERVICE
▾ Tables
- DIM_CUSTOMER
- DIM_TERRITORY
- DIM_PRODUCT
- FACT_CASE_DETAILS

Ownership: | Central IT Team | Data Product Team1 | Data Product Team2 |

Pre Data Mesh

Pros and Cons of analytics delivery pre data mesh

Pros

- Pattern based automation and configuration
- Highly skilled central data modeling team completely aligned with each other

Cons

- Releases slow
- Backlog huge
- Business demanding quicker development times

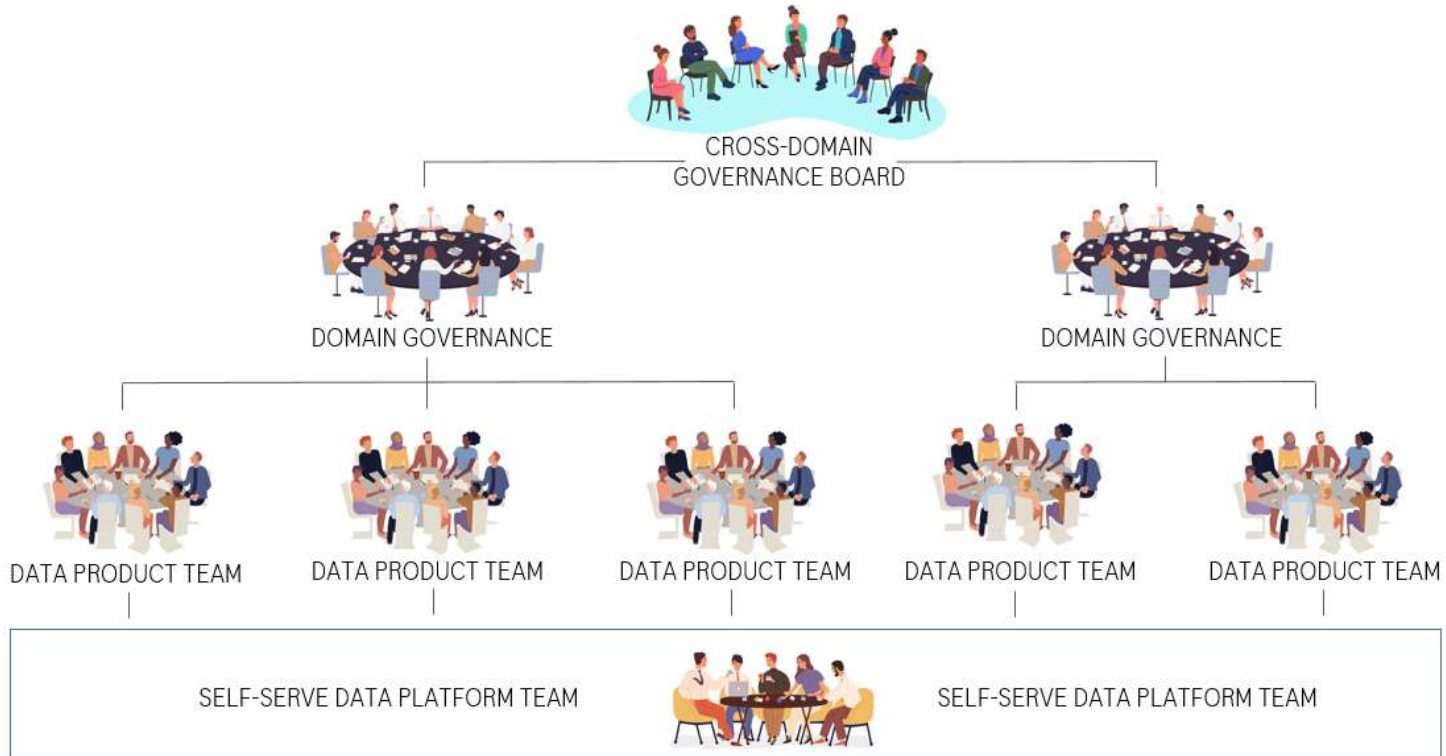Ownership: Central IT Team   Data Product Team1   Data Product Team2

Data Mesh - organizational changes
Domain owned, cross-functional data product teams were introduced

CROSS-DOMAIN
GOVERNANCE BOARD

DOMAIN GOVERNANCE

DOMAIN GOVERNANCE

DATA PRODUCT TEAM

DATA PRODUCT TEAM

DATA PRODUCT TEAM

DATA PRODUCT TEAM

DATA PRODUCT TEAM

SELF-SERVE DATA PLATFORM TEAM

SELF-SERVE DATA PLATFORM TEAM

Data Mesh - ownership changes
Data product teams take ownership of their own pipelines and code

Roche

PSA_SALESFORCE
- ▼ Tables
  - ⊞ ACCOUNT
  - ⊞ CASE
  - ⊞ PRODUCT
  - ⊞ TERRITORY

PSA_SAP
- ▼ Tables
  - ⊞ SALES
  - ⊞ PRODUCT
  - ⊞ CUSTOMER
  - ⊞ COMPANY

RAW_DV
- ▼ Tables
  - ⊞ HUB_PRODUCT
  - ⊞ HUB_CUSTOMER
  - ⊞ HUB_COMPANY
  - ⊞ HUB_ACCOUNT
  - ⊞ HUB_CASE
  - ⊞ HUB_TERRITORY
  - ⊞ SAT_PRODUCT_SAP
  - ⊞ SAT_CUSTOMER_SAP
  - ⊞ SAT_CUSTOMER_SFDC
  - ⊞ SAT_CASE
  - ⊞ SAT_TERRITORY
  - ⊞ LSAT_TERRITORY_SALES
  - ⊞ LNK_TERRITORY_SALES
  - ⊞ XLNK_SALES
  - ⊞ LNK_CASE_ACCOUNT
  - ⊞ SAT_PRODUCT_SFDC

BUSINESS_DV
- ▼ Tables
  - ⊞ SLAT_CUSTOMER
  - ⊞ BSAT_TERRITORY_HIERARCHY
  - ⊞ PIT_TERRITORY_SALES
  - ⊞ LNK_CASE_ACCOUNT
  - ⊞ BRG_SALES
  - ⊞ BRG_CASE_DETAILS

DM_SALES
- ▼ Tables
  - ⊞ DIM_CUSTOMER
  - ⊞ DIM_TERRITORY
  - ⊞ DIM_PRODUCT
  - ⊞ FACT_SALES

DM_SERVICE
- ▼ Tables
  - ⊞ DIM_CUSTOMER
  - ⊞ DIM_TERRITORY
  - ⊞ DIM_PRODUCT
  - ⊞ FACT_CASE_DETAILS

Ownership: | Central IT Team | Data Product Team1 | Data Product Team2 |

32

Data Mesh - ownership changes

Data product teams take ownership of their own pipelines and code

PSA_SALESFORCE
▾ Tables
- ACCOUNT
- CASE
- PRODUCT
- TERRITORY

PSA_SAP
▾ Tables
- SALES
- PRODUCT
- CUSTOMER
- COMPANY

RAW_DV_DPT1
▾ Tables
- HUB_CUSTOMER
- HUB_ACCOUNT
- HUB_CASE
- HUB_TERRITORY
- SAT_CUSTOMER_SFDC
- SAT_CASE
- SAT_TERRITORY
- LSAT_TERRITORY_SALES
- LNK_TERRITORY_SALES
- LNK_CASE_ACCOUNT
- SAT_PRODUCT_SFDC

RAW_DV_DPT2
▾ Tables
- HUB_PRODUCT
- HUB_COMPANY
- SAT_PRODUCT_SAP
- SAT_CUSTOMER_SAP
- XLNK_SALES

BUSINESS_DV_DPT1
▾ Tables
- SLAT_CUSTOMER
- BSAT_TERRITORY_HIERARCHY
- PIT_TERRITORY_SALES
- LNK_CASE_ACCOUNT
- BRG_SALES
- BRG_CASE_DETAILS

BUSINESS_DV_DPT2
▾ Tables
- BRG_SALES

DM_SALES
▾ Tables
- DIM_CUSTOMER
- DIM_TERRITORY
- DIM_PRODUCT
- FACT_SALES

DM_SERVICE
▾ Tables
- DIM_CUSTOMER
- DIM_TERRITORY
- DIM_PRODUCT
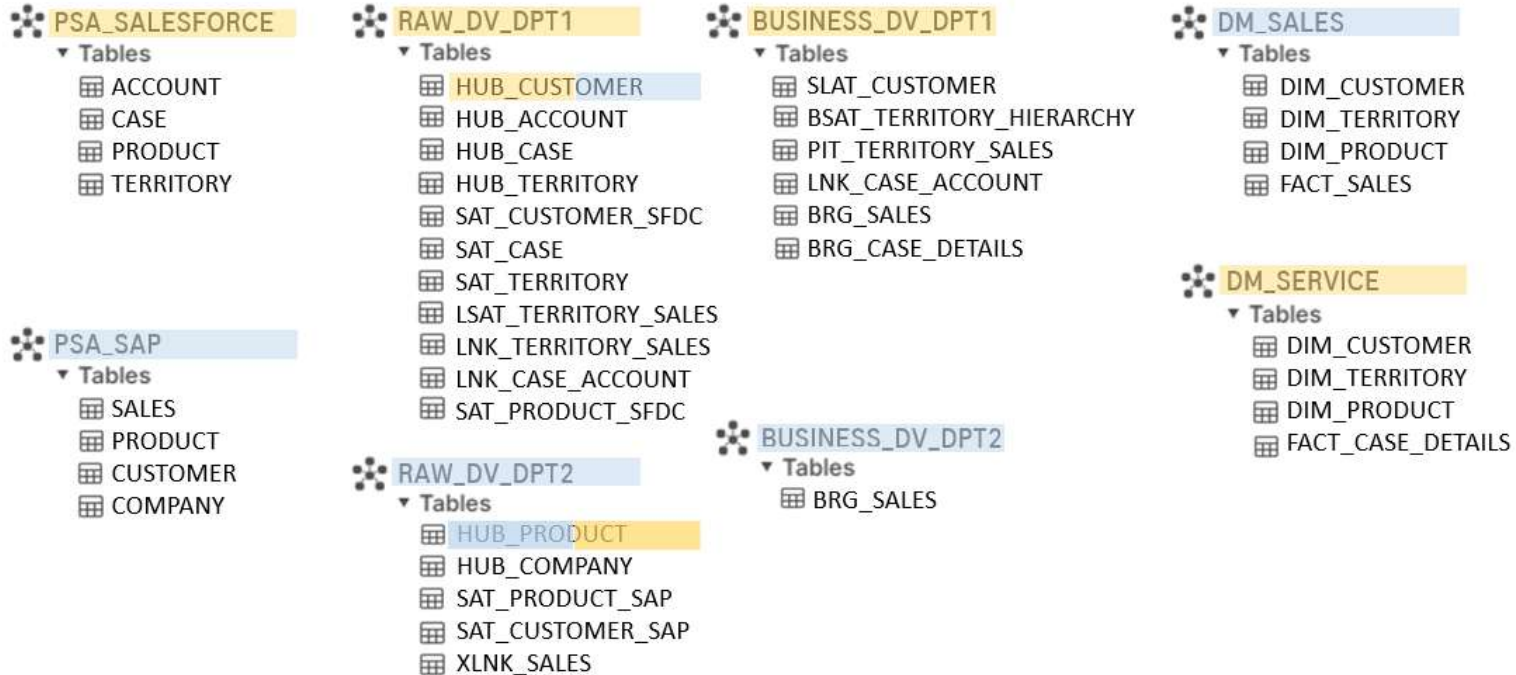- FACT_CASE_DETAILS

Ownership: Central IT Team | Data Product Team1 | Data Product Team2

33

What about data modeling
How should data modeling be addressed when implementing Data Mesh

Pros

- Scaling faster as data product teams bring new developers
- More frequent releases
- Data modeling is closer to those who understand the data
- Reduced IT and business friction
- Easier access to source systems as there is clear business use case

Cons

- Multiple data product teams loading to same HUBs and LNKs breaches data product team boundaries
- Data modeling skills are often difficult to find for every data product team
- Lack of 'how to' implement

Data Mesh - reusable data products
Components of the data vault model considered reusable data products

Roche

PSA_SALESFORCE
- ▼ Tables
  - ACCOUNT
  - CASE
  - PRODUCT
  - TERRITORY

PSA_SAP
- ▼ Tables
  - SALES
  - PRODUCT
  - CUSTOMER
  - COMPANY

RAW_DV_DPT1
- ▼ Tables
  - HUB_CUSTOMER
  - HUB_ACCOUNT
  - HUB_CASE
  - HUB_TERRITORY
  - SAT_CUSTOMER_SFDC
  - SAT_CASE
  - SAT_TERRITORY
  - LSAT_TERRITORY_SALES
  - LNK_TERRITORY_SALES
  - LNK_CASE_ACCOUNT
  - SAT_PRODUCT_SFDC

RAW_DV_DPT2
- ▼ Tables
  - HUB_PRODUCT
  - HUB_COMPANY
  - SAT_PRODUCT_SAP
  - SAT_CUSTOMER_SAP
  - XLNK_SALES

BUSINESS_DV_DPT1
- ▼ Tables
  - SLAT_CUSTOMER
  - BSAT_TERRITORY_HIERARCHY
  - PIT_TERRITORY_SALES
  - LNK_CASE_ACCOUNT
  - BRG_SALES
  - BRG_CASE_DETAILS

BUSINESS_DV_DPT2
- ▼ Tables
  - BRG_SALES

DP_MONTHLY_SALES
- ▼ Tables
  - FISCAL_CALENDAR
  - MTHLY_SALES_DETAIL
  - MTHLY_SALES_BY_PRODUCT
  - MTHLY_TERRITORY_SALES

DP_SERVICE_ANOMALIES
- ▼ Tables
  - SERVICE_ANOMALY_CASES
  - SERVICE_ANOMALY_DETAILS

Ownership: | Central IT Team | Data Product Team1 | Data Product Team2 |

Data Mesh - reusable data products

Components of the data vault model considered reusable data products

Pros

- Fully aligned with data product boundaries
- Promotes accountability within each domain
- Outward scales very quickly
- IT can focus on capabilities and automation

Cons

- Reluctance to own data products
- Lack of 'how to' implement
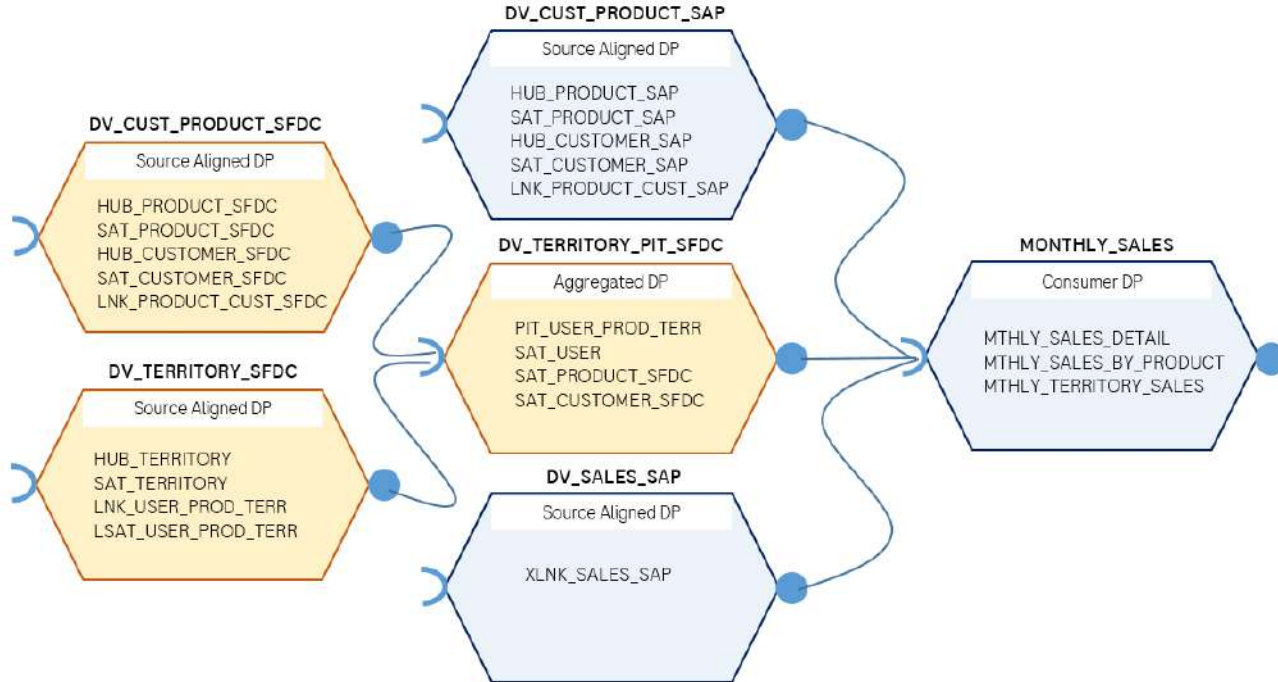- Costs initially seem higher for domains
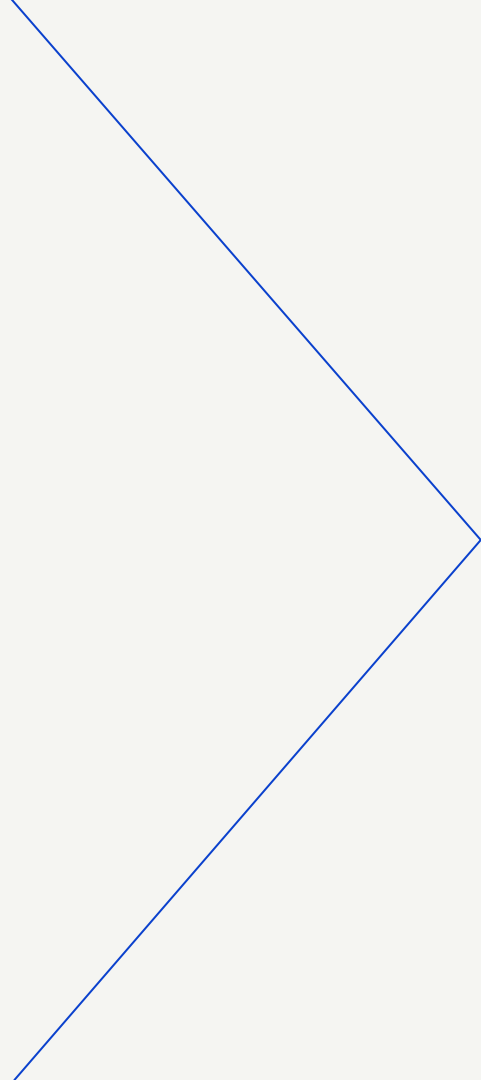
Ownership: Central IT Team   Data Product Team1   Data Product Team2

Data Mesh - reusable data products
Components of the data vault model considered reusable data products

Lessons Learned

Success
Decentralization brings with it many challenges
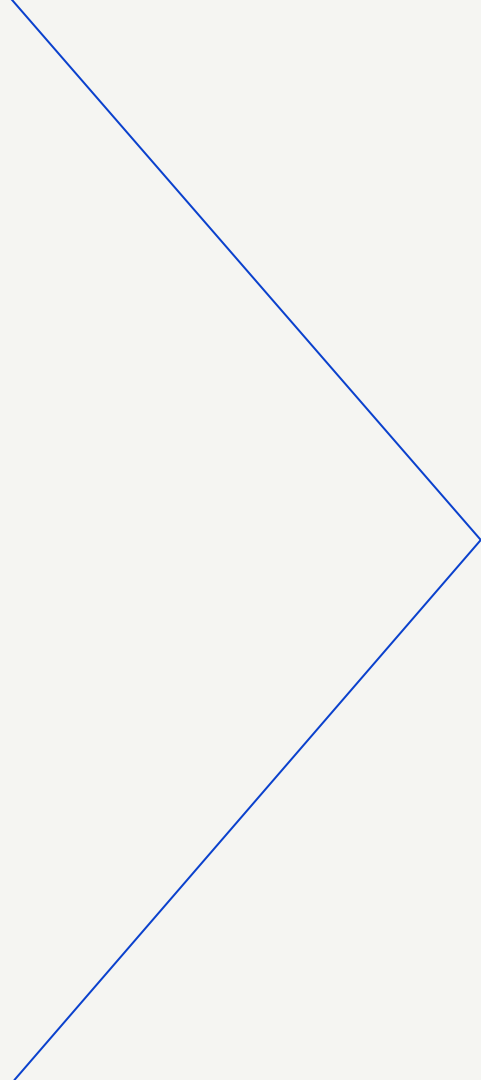
Successes

- Improved data quality
- Increased accountability
- Improved agility and speed of delivery
- Reduced data silos
- Platform team can focus on delivering capabilities, standards and accelerators
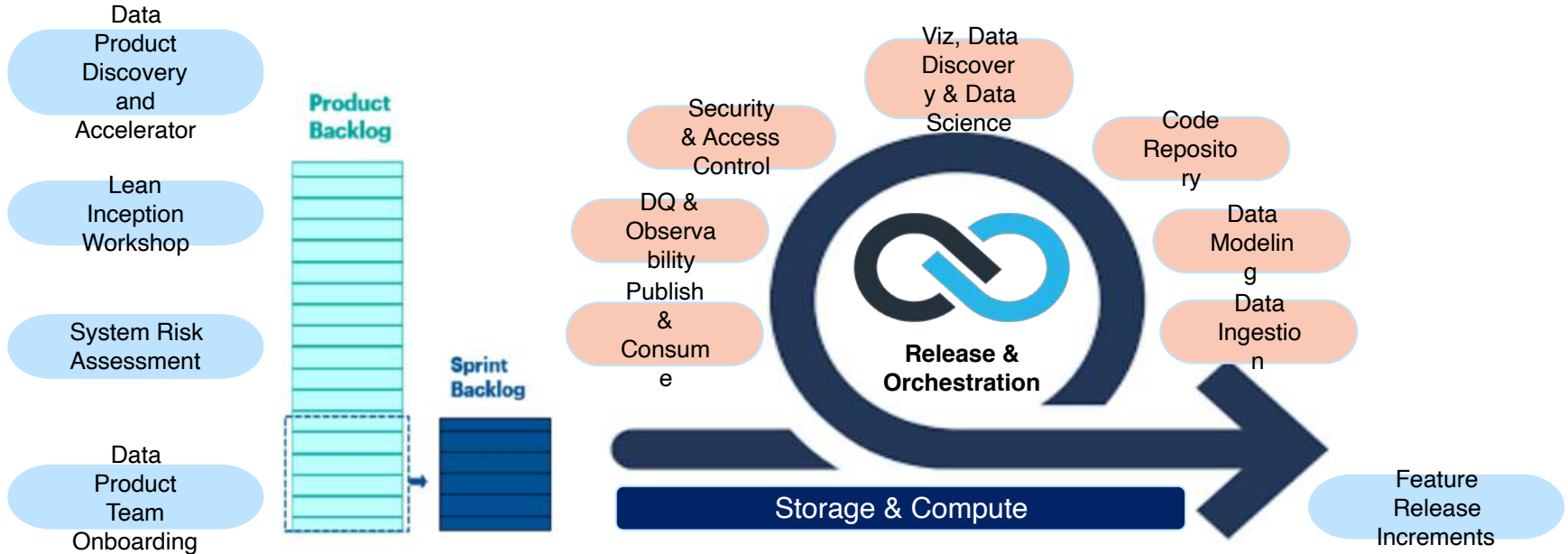
Challenges

- Lacks 'how to'
- Little desire to own Data Products
- Governance of shared data is still sketchy
- Getting good data engineers and data modelers in the data product teams

Roche Data Mesh Platform

Roche Data Mesh Platform
The Data Product Lifecycle

Data Product Discovery and Accelerator

Lean Inception Workshop

System Risk Assessment

Data Product Team Onboarding

Product Backlog

Sprint Backlog

Security & Access Control

DQ & Observability

Publish & Consume

Viz, Data Discovery & Data Science

Code Repository

Data Modeling

Data Ingestion

Release & Orchestration

Storage & Compute

Feature Release Increments

Data Mesh <u>Platform</u> Services – Examples

Designing data mesh platform to implement data mesh services

RDDM Self-Service Data Platform
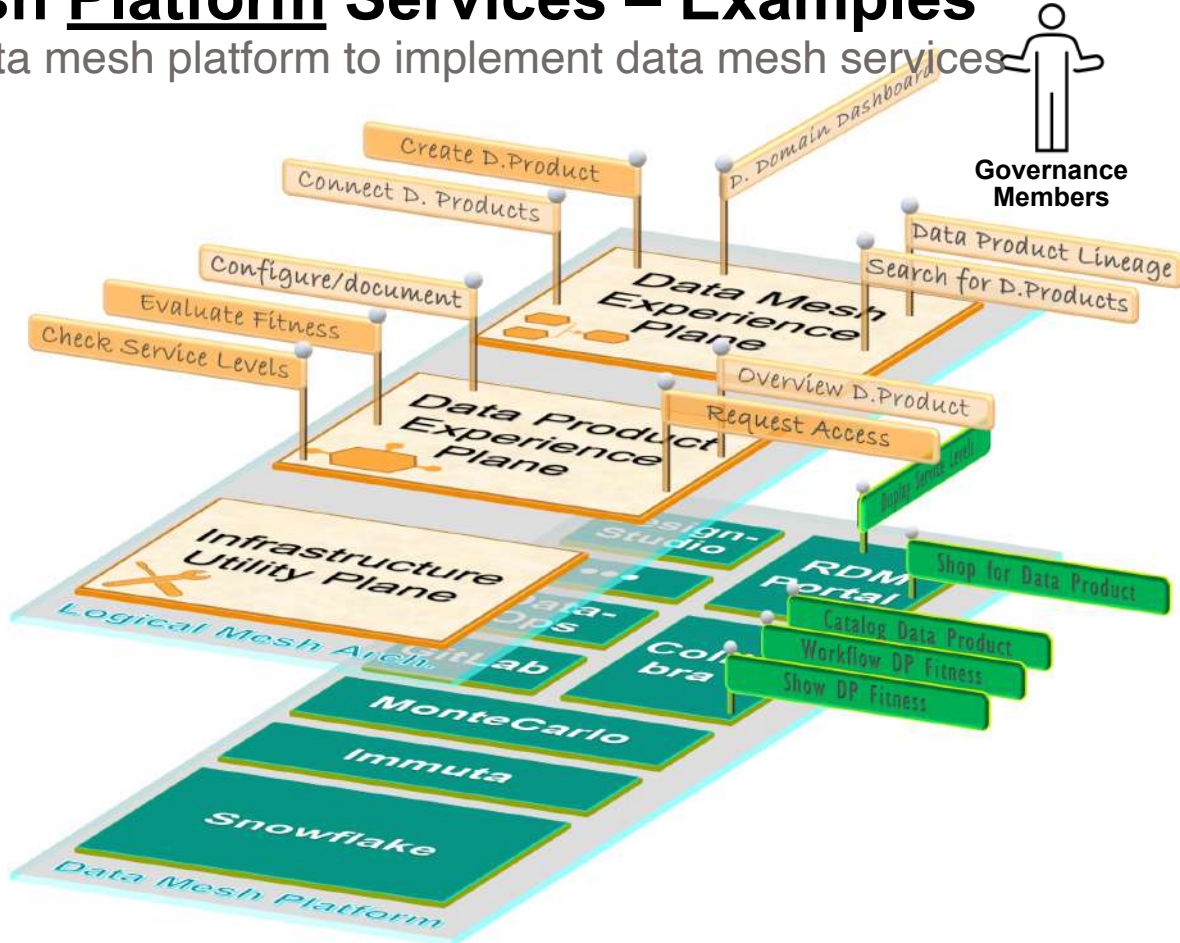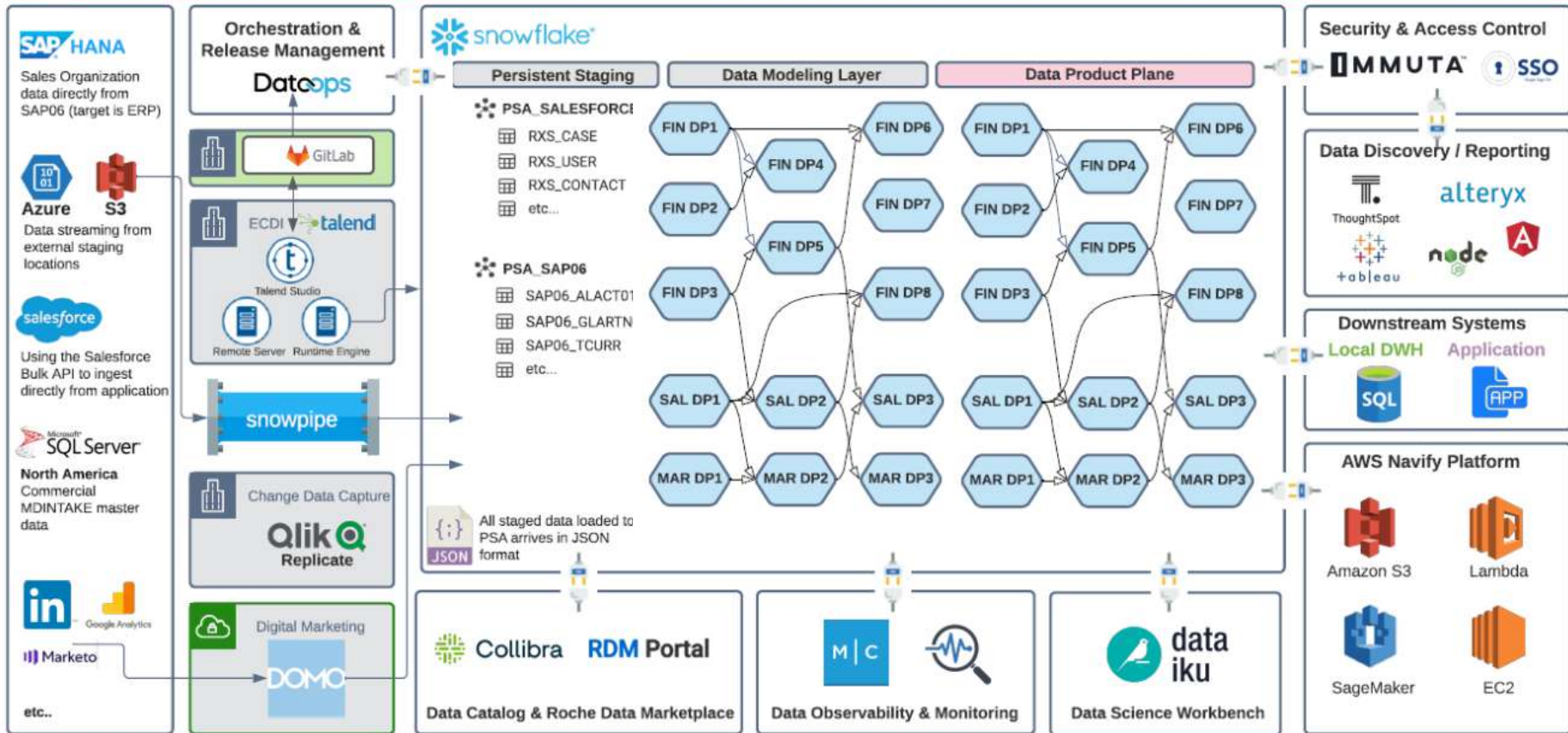
Platform a glance
Roche mesh experience at glance

FEB 2021
Platform build started

6-8 weeks
Average MVP time

#40
Data product teams working

#2680
Users, developers onboarded

180 TB
Storage consumed so far

#50 data products
Published following FAIR principles

120 releases
last month, vs 1 release in 3 months

15+
Tools & technologies integrated in the ecosystem to enabled various capabilities

40+ million
Savings in inventory reduction, cost avoidance and resource optimization

Key Learnings
We needed to find out more about Data Mesh and what changes were required

- Company must be willing to adopt organizational change
- Helps if the company already has a decentralized approach to BI and Analytics
- Company must be at a certain level of data driven maturity
- Must be use-case driven
- Encourage a product mindset
- Good data engineers are key to success
- Agility is key – do not be frightened to change direction, refactor or even u-turn
- There is no book that will tell you exactly how to implement or migrate
- Use the tools and vendors to help you get ahead.
- Using natural business keys in your modeling increases the success for interoperability
- Enjoy the journey