



Deelnemerslijst

Datawarehousing & Business Intelligence Summit

24 maart 2026

BEDRIJF	NAAM DEELNEMER		FUNCTIE
APG Asset Management	Joey	Hagelstein	data architect
APG Asset Management	Roy	Krout	it-information-management-consultant
APG Asset Management N.V.	Erwin	Meyer	business-analyst
ASN Bank	Ben	Huizinga	data architect
ASN Bank	Ester	Beckers	data architect
ASN Bank	Philip	Dijkstra	data architect
ASN Bank	Victor	de Hoon	datawarehouse-specialist
Acceldata	Colin	Fernandes	marketing-manager
AmsterdamUMC	Frank	Van de Ruit	bi-specialist
AmsterdamUMC	Riet	Albers	bi-specialist
BDO Holding BV	Bart	van den Bergh	other
Bank Mendes Gans	Jan Willem	Dijkstra	bi-specialist
Belastingdienst	Chang	Pan	information-analyst
Belastingdienst	Francois	Essers	other
Belastingdienst	Harm Willem	Stuut	data architect
Belastingdienst	Hicham	Ghalibe	other
CRV	Candice	Johnson	bi-specialist
Channable	Jeroen	Meijaard	business-analyst
Conclusion Mediaan BV	Heidi	Luyten	organisation-consultant
Conclusion Mediaan BV	Marcel	Pereboom	data architect
Connected Data Group	Timon	Haverkamp	managing-director-owner
Contact Media	Arnoud	van Gemeren	managing-director-owner
DKG services	Denise	de Reus	data architect
DKG services	Jeroen	Luijcks	it-manager
DLL	Bob	Mooren	data engineer
DLL	Willem	Aerts	data engineer
DSW Zorgverzekeraar	Frederik	Spanhoff	data engineer
DSW Zorgverzekeraar	Laurens	Boers	data engineer
Datatune BV	Bart	Buvs	bi-specialist
Datatune BV	John	Gerits	bi-specialist
Datatune BV	Willem	Wijns	bi-specialist
De Belastingdienst	Jorrit	Janssen	data architect
Digital Power	Guus	van	organisation-consultant
Digital Power	Niels	Bosmans	product-manager

Dutch Tax Authority - Belastingdienst	Esmee	Tomaula	business-analyst
Edgeconnex	Bruno	Reis	data architect
Eraneos	Jasper	Stoop	data architect
Eraneos	Saumya	Kapoor	data architect
Flex4Data	Anke	Ruigrok	data architect
Freelance	Serge	Lodens	data architect
Genesee Academy	Remco	Broekmans	managing-director-owner
Goodin	Kira	Sjoberg	organisation-consultant
Growteq B.V.	Arjan	de Winter	data architect
Growteq B.V.	Gertjan	van Luik	data engineer
Growteq B.V.	Robert Jan	Quist	data architect
HAN University of Applied Sciences	Peter	De Belie	data architect
HAN University of Applied Sciences	Tyco	Hendriks	data architect
HAS green academy	Harm	Bodewes	other
Het Consultancy Huis	Frank	Keppel	data architect
Hogeschool Rotterdam	Hassan	Alaoui	teacher-trainer-docent-informatica
ING	Eric	Greuter	other
ING	Jeroen	van Dulleman	other
Iconsultancy	Martijn	Ceelen	bi-specialist
KEMBIT	Math	Huntjens	other-it-related-job-title
KEMBIT	Tim	Lustberg	data architect
Knowit	Jorn Atle	Husbyn	datawarehouse-specialist
Knowit	Kjetil Hoeie	Hansen	datawarehouse-specialist
NVWA	Hedwig	van der Linden	bi-specialist
NVWA	Tessa	van Enter	data engineer
Nederlandse Spoorwegen	Andrea	Huck	data architect
Nederlandse Spoorwegen	Roy	Schenk	data architect
Nippur	Ecem	Okse	data engineer
Nippur	Irini	Wiersma	data engineer
Nippur	Marc	Sanders	data engineer
Ontwar	Pieter-Jan	Dijs	data engineer
Orsted	Wojciech	Lukawski	datawarehouse-specialist
PGGM	Frank	Reterink	data engineer
PW CONSULTING	Marius	Geleijnse	data engineer
RVO	Tamar	Bakker	other
RVO /Ministerie Economische Zaken	Gerard	Moesker	other-it-related-job-title
Rabobank	Bram	Laumans	data engineer
Rabobank	Jeroen	Wolfs	it-manager
SEDIME	Frank	Pieper	data architect
Stater N.V.	Arno	van der Kwast	other
Stater N.V.	Robert	Lautenbach	data architect
Stater NV	Albert	de Ruiter	datawarehouse-specialist
Stedin	Ferry	Brand	other

Stedin	Maurice	Kalis	information-analyst
Suithouse Consulting	Gerco	Soet	general-manager
Suithouse Consulting	Olivier	van den Eertwegh	data engineer
Suithouse Consulting	Peter	Aalvink	bi-specialist
Suithouse Consulting	Rob	Klaassen	bi-specialist
Suithouse Consulting	Rob	de Kruijk	data architect
SureSync BV	Alexander	Al	data engineer
SureSync BV	Kas	De Man	data engineer
TU Delft	Rene	Haverslag	product-manager
TU Delft	Ron	Van Velzen	information-manager
The Future Group	Eric	Fickel	it-information-management-consultant
Tivato	Vanessa	Delsemme	it-information-management-consultant
Topigs Norsvin Research Center BV.	Stefan	Hofman	other-it-related-job-title
University of Vienna	Markus	Janoch	datawarehouse-specialist
University of Vienna	Zvezdana	Milosevic	database-administrator-dba
Vista Veritas Data & AI Services	Erik	Fransen	managing-director-owner
Vlaamse Overheid	Jan	Dhaese	bi-specialist
Vlaamse Overheid	Jorge	Biltris	datawarehouse-specialist
Vlaamse Overheid	Lisa	Vanluchene	bi-specialist
Vlaanderen connect	Raf	Carpentero	data engineer
Vopak	Carina	Appeldoorn-Baud	data architect
Waterschap Noorderzijlvest	Roosmarijn	Sippens Groenewegen	data architect



WHO WE ARE



Werner Schoots
Founder Adept Events



BI-Platform.

- Launched in 2008 as online spin-off from Database Magazine (DB/M)
- Topics: Business Intelligence, Data Warehousing, Analytics, Data Management

News

Job board

Selected Whitepapers

Events

Articles

Blogs

Video interviews

Cases

- We welcome your input: redactie@biplatform.nl



www.biplatform.nl & weekly newsletter



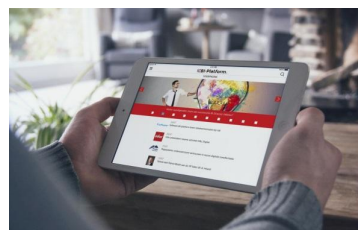
@BIplatform on X.com & YouTube



Download the BI-Platform App



Join our LinkedIn Discussion Group



SEMINARS

All seminars and workshops are organised twice a year

Alec Sharp	Business-oriented Data Modelling Masterclass Working with Business Processes Masterclass Concept Modelling for Business Analysts
Rick van der Lans	Ontwerpen van een Nieuwe Data Architectuur
Mathias Vercauteren	Data Governance Sprint AI Governance, Responsible AI and Data Governance – Connecting the Dots
Chris Bradley / Winfried Etzel	Data Management Fundamentals
Lawrence Corr	Agile Data Warehouse Design & Dimensional Modeling
Christian Gijssels	Generatieve AI in Business Analyse
Juha Korpela	Data Mesh – Modeling Data Products and Domains
Thomas Gijssels	AI Agents in de Praktijk – Van Concept tot Implementatie
Multiple speakers	Data Warehousing & BI Summit – Yearly conference in March/April

IN-HOUSE

All seminars and workshops can be organized in-company.
With local speakers and international speakers!



Please contact Werner Schoots

☎ +31 (0)172 742680

✉ seminars@adeptevents.nl

 **AdeptEvents**

📧 @AdeptEventsNL

📘 Adept Events

🌐 Adept Events

ABOUT THIS EVENT

- Old school: No Streaming – Video Recordings
- Packed Programme 13th Edition with 9 sessions + 3 workshops
- Chairman and moderator: Dennis van Gelder en Tanja Ubert
- Binder with proceedings & PDF files on Downloadpage
- Evaluation forms – end of day
- Video recordings available after the conference
- Free Wifi available: Hotel Utrecht Business | ValkEvent

ABOUT THIS EVENT



Schedule Change

- Session by Eevamaija Virtanen replaced by Ron Tolido

Breaks:

10:15 – 10:30	Coffee/Tea
12:30 – 13:30	Lunch in Foyer (not in Nieuwegracht 1)
15:30 – 15:45	Coffee/Tea/Soda
16:50	Reception!

An Holistic Data Architecture: From Source tot Insight



Rick F. van der Lans
Industry analyst



Copyright © 2026 R20/Consultancy B.V., The Netherlands. All rights reserved. No part of this material may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photographic, or otherwise, without the explicit written permission of the copyright owners.



Rick F. van der Lans



Rick F. van der Lans is a highly-respected independent analyst, consultant, author, and internationally acclaimed lecturer specializing in data warehousing, business intelligence, big data, and database technology. He is managing director of R20/Consultancy BV.

He has presented countless seminars, webinars, and keynotes at industry-leading conferences. Rick helps clients worldwide to design their data warehouse, big data, and business intelligence architectures and solutions and assists them with selecting the right products. He has been influential in introducing the new logical data warehouse architecture worldwide which helps organizations to develop more agile business intelligence systems.

He is the author of several books on computing, including his new *Data Virtualization: Selected Writings* and *Data Virtualization for Business Intelligence Systems*. Some of these books are available in different languages. Books such as the popular *Introduction to SQL* is available in English, Dutch, Italian, Chinese, and German and is sold world wide. He also authored numerous whitepapers for vendors.

In 2018 he was selected the sixth most influential BI analyst worldwide by analytica.com.

You can get in touch with Rick van der Lans via:

Email: rick@r20.nl

Website: www.r20.nl

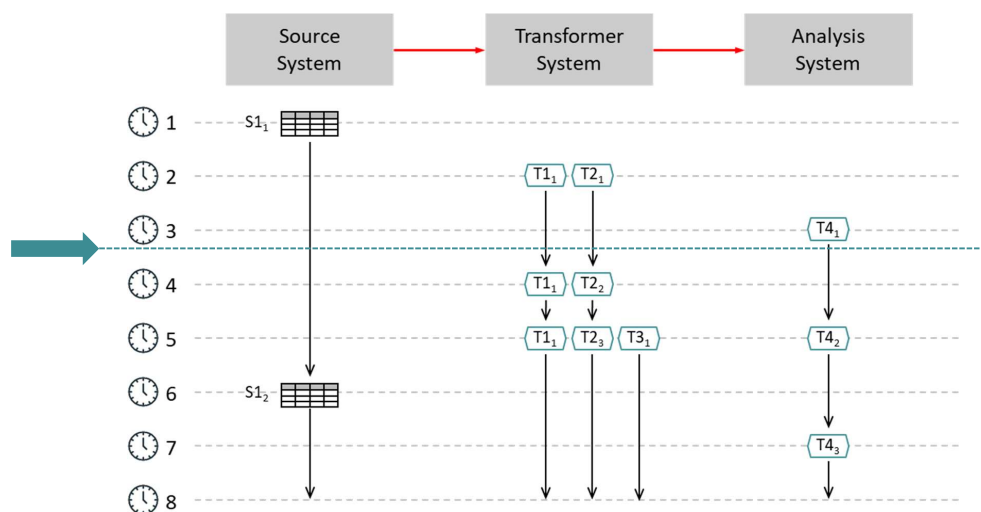
LinkedIn: <http://www.linkedin.com/pub/rick-van-der-lans>



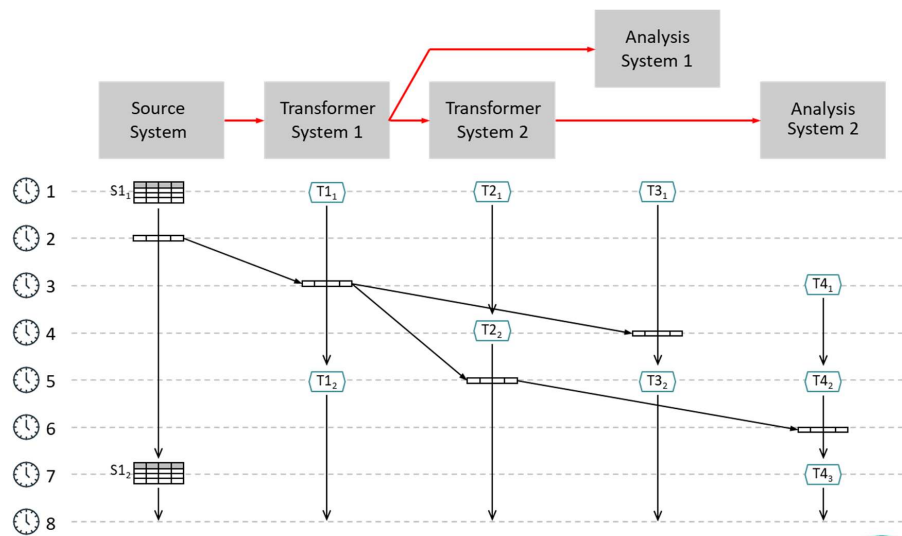
Challenges, Wishes and Requirements

- Accelerating the implementation of *cross-system analysis systems*
- Accelerating the implementation of *cross-organization analysis systems* (open data)
- Simplifying the *synchronization of source systems*
- Increasing the transparency of data processing for *process/decision reconstruction*
- Increasing the *discoverability* and clarity/meaning of data
- Improving *data quality*
- Strengthening *data security* and *data privacy*
- Strengthening *data history*
- Handling unstructured data: video, sound, text, and images
- Simplifying the data processing landscape
- Smarter uses of new technology

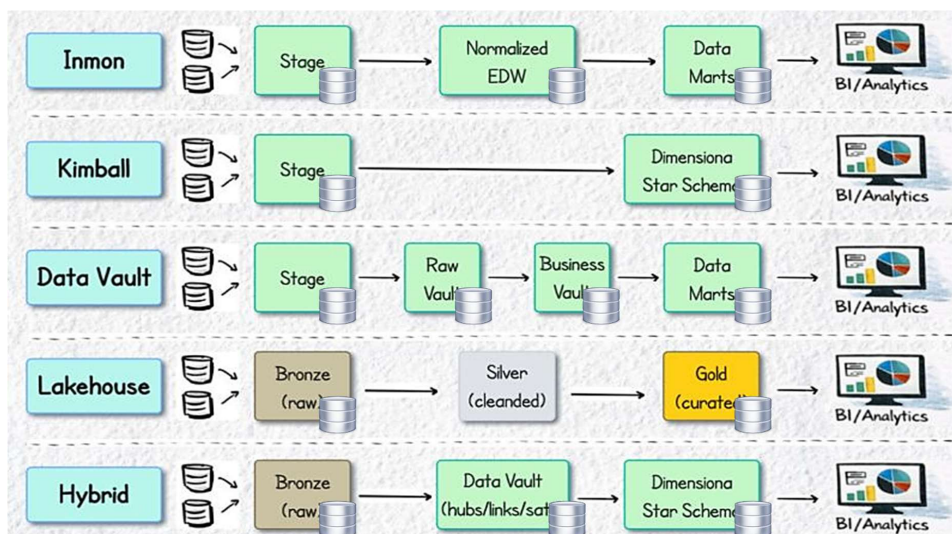
Example of Horizontal Lineage With Versions



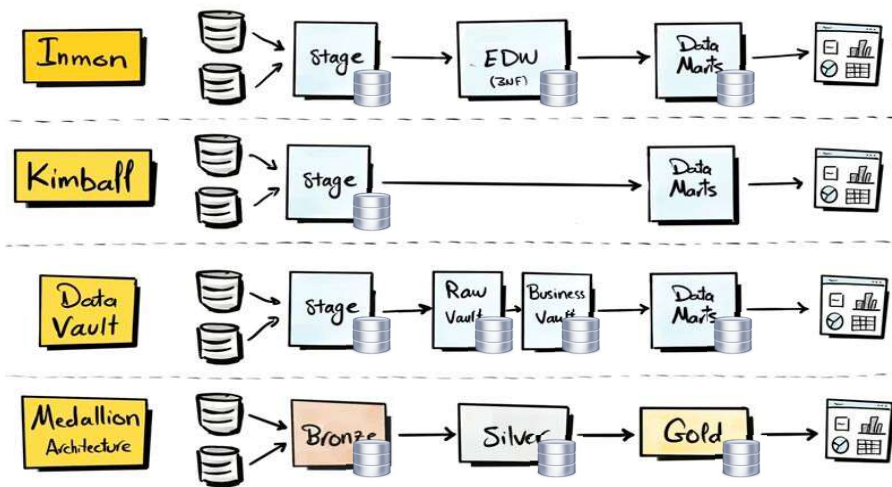
Example of Operational Lineage



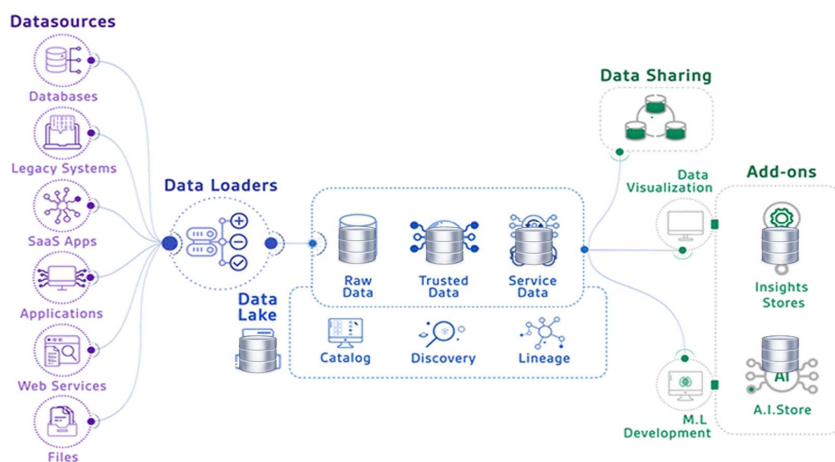
Domain Data Architectures (1)



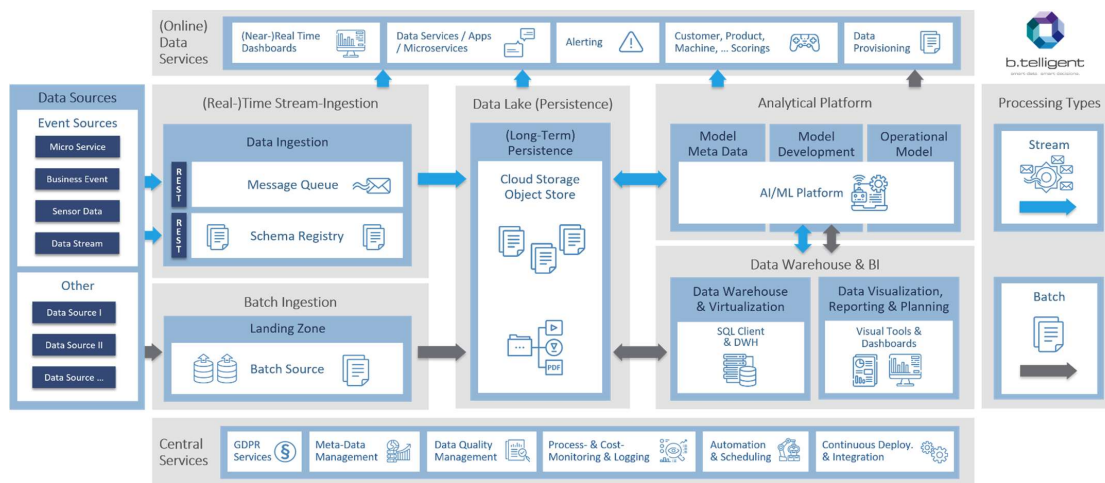
Domain Data Architectures (2)



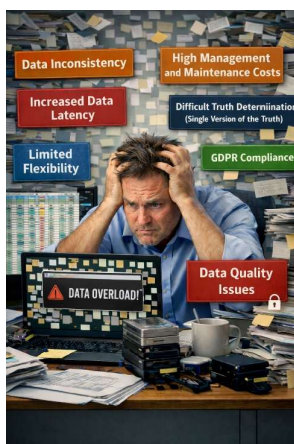
Domain Data Architectures (3)



Data-copy Rich Data Architectures

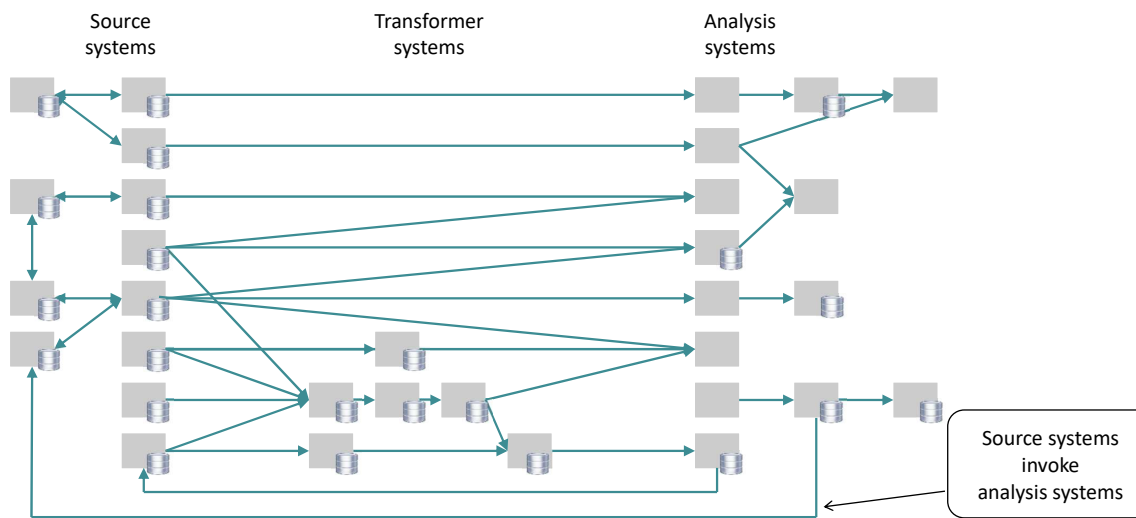


Disadvantages of Data Copying

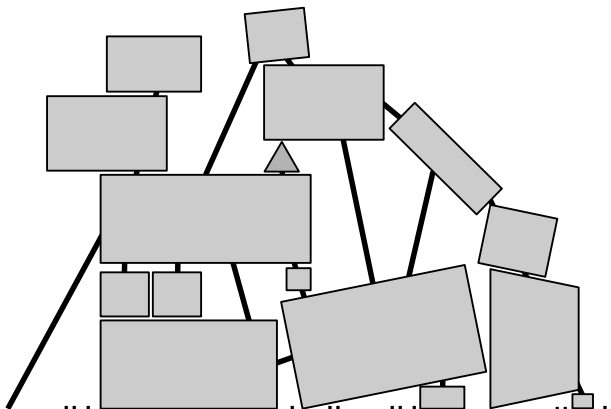


- Data inconsistency
- Difficult determination of the truth ("single version of the truth")
- Limited flexibility and agility
- High management, development, and maintenance costs
- More complex compliance with General Data Protection Regulation (GDPR)
- Complex data synchronization issues
- Data quality challenges
- More complex data security
- Outdated data (increased data latency)
- ...

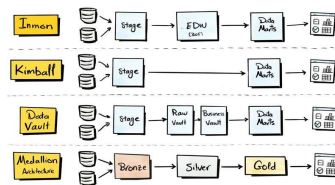
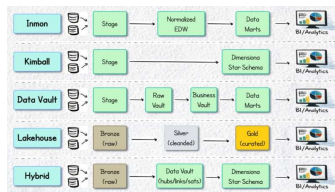
Evolution of the Data Processing Landscape



“Wobbly” Foundation of the Data Processing Landscape



Limitations of Some Data Architectures



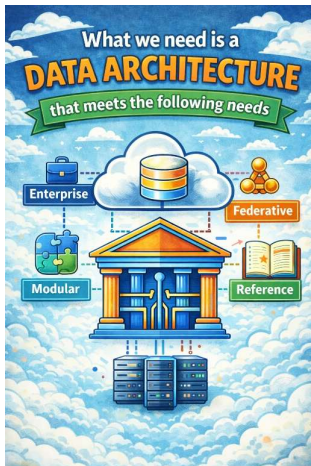
- Source systems are unjustly out of scope
 - Old and inflexible
 - Software packages
 - No IT sovereignty
 - No or limited data history
 - Poor data security
 - Complex data synchronization
- Analysis Systems are unjustly out of scope
 - Many data copies
 - Not under control of IT
 - Analysis models input to source systems
- Many requirements are unjustly out of scope
 - Reconstruction
 - Metadata for everyone
 - Data-sharing

Bring Your (Data)House in Order



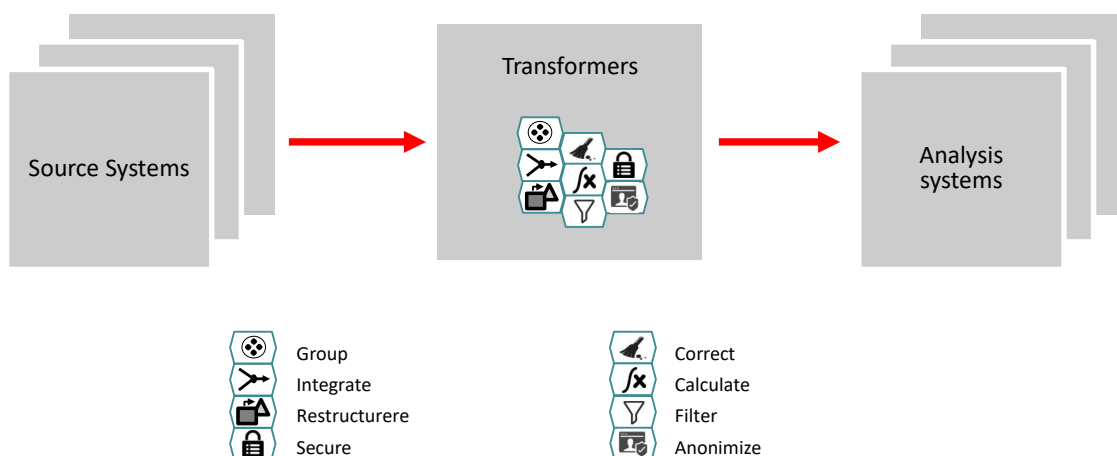
- Data in order
- Metadata in order
- Lineage in order (origin and impact)
- Reconstruction in order
- Data security in order
- Technology in order
- It all starts with *data architecture in order*

What We Need is a ...

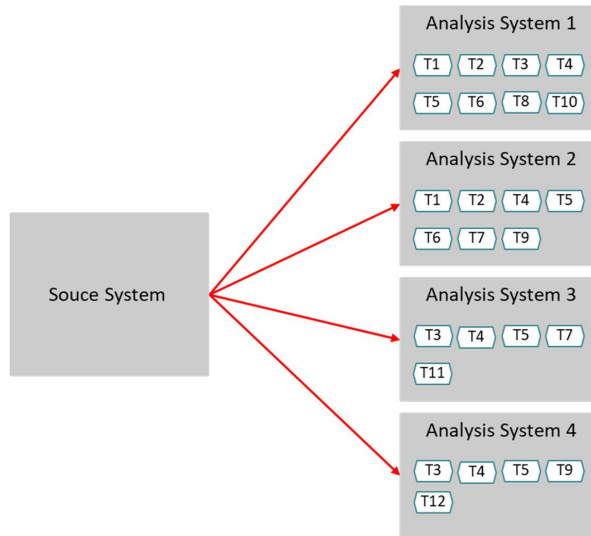


- *Enterprise* data architecture
 - Not a domain data architecture
 - Covers the entire data route
 - Not only focused on analysis and data science (and AI)
- *Modular* data architecture
 - Think big, act small
- *Federative* data architecture
 - Data storage minimization – less data copies
- *Reference* data architecture
 - Generic, reusable

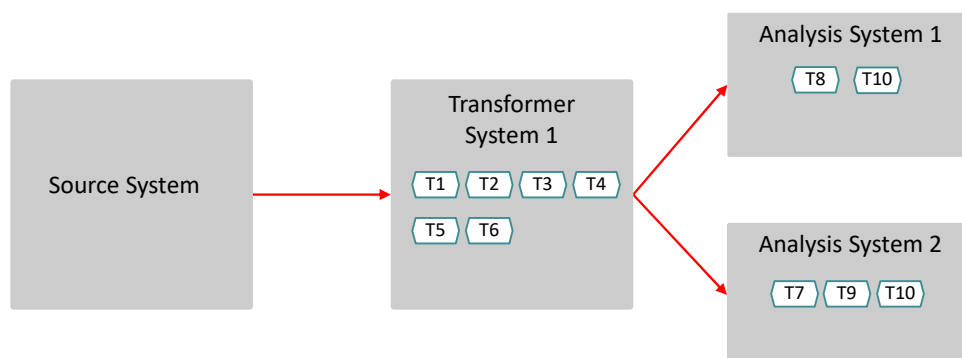
Start with Transformers, Not With Databases



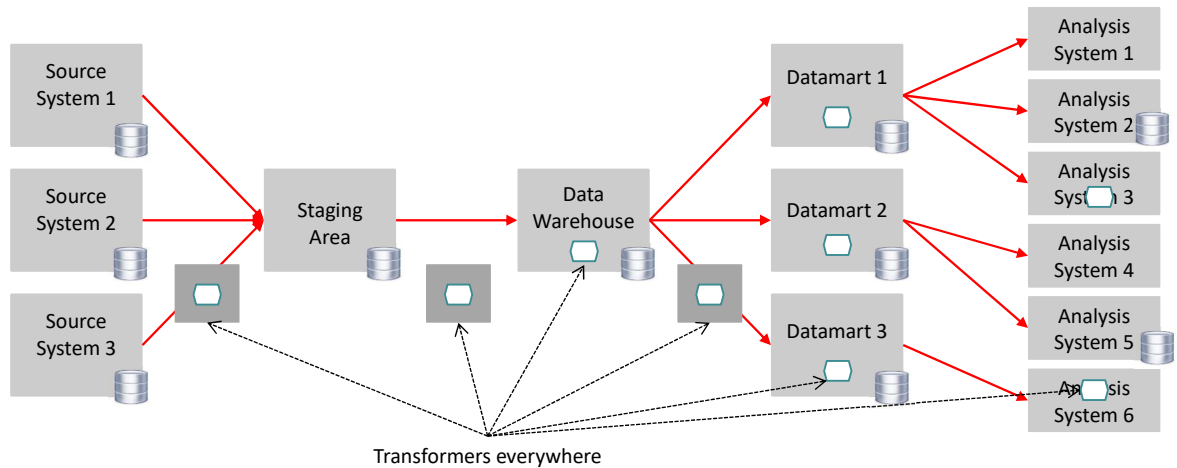
Duplicate Transformers in Analysis Systems



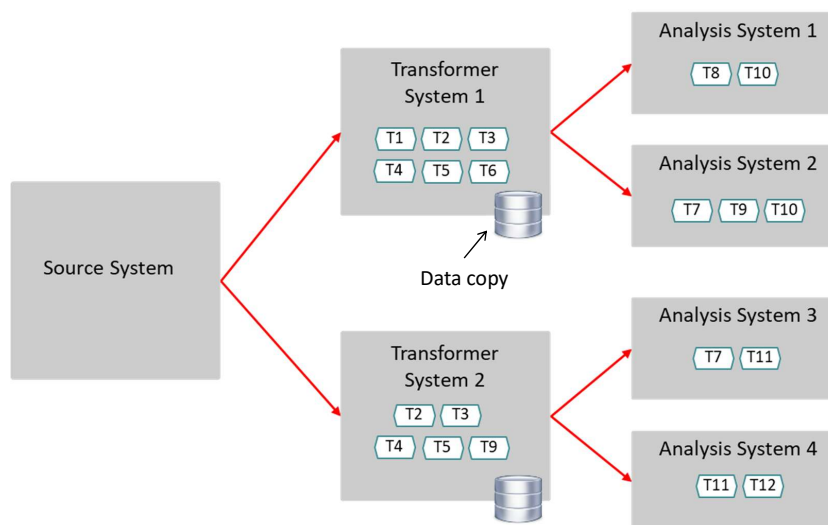
Centralizing Transformers



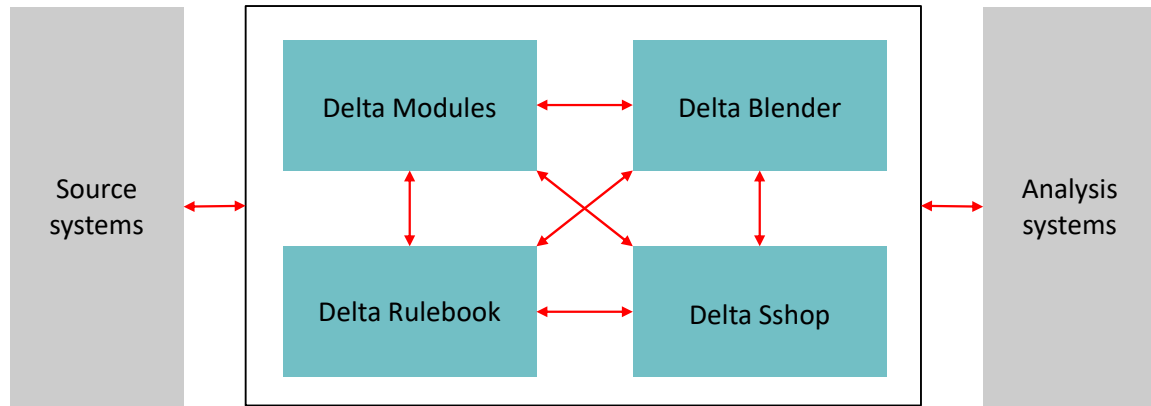
A Traditional Data Warehouse Architecture



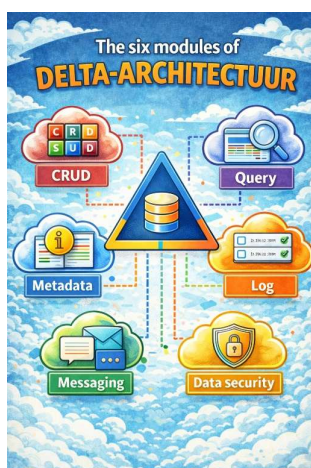
Duplicate Transformers in Transformer Systems



Delta: A Modern, Enterprise Data Architecture

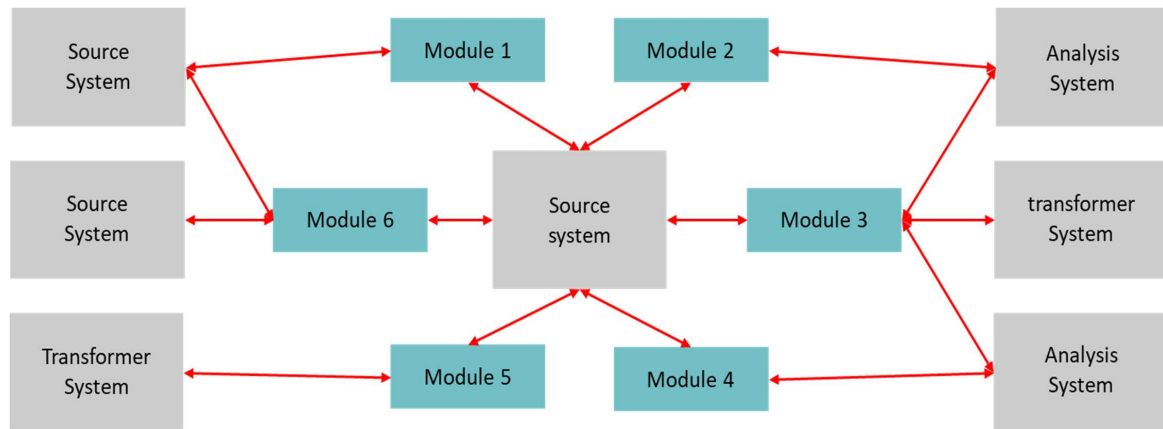


Six Delta Modules

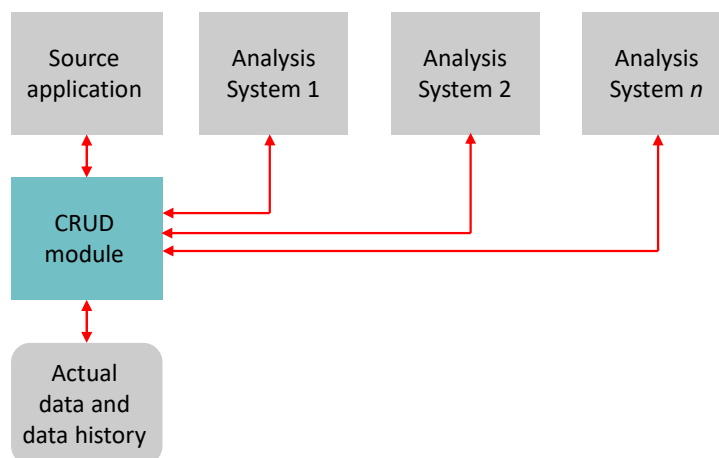


- CRUD
- Query
- Metadata
- Log
- Messaging
- Data security

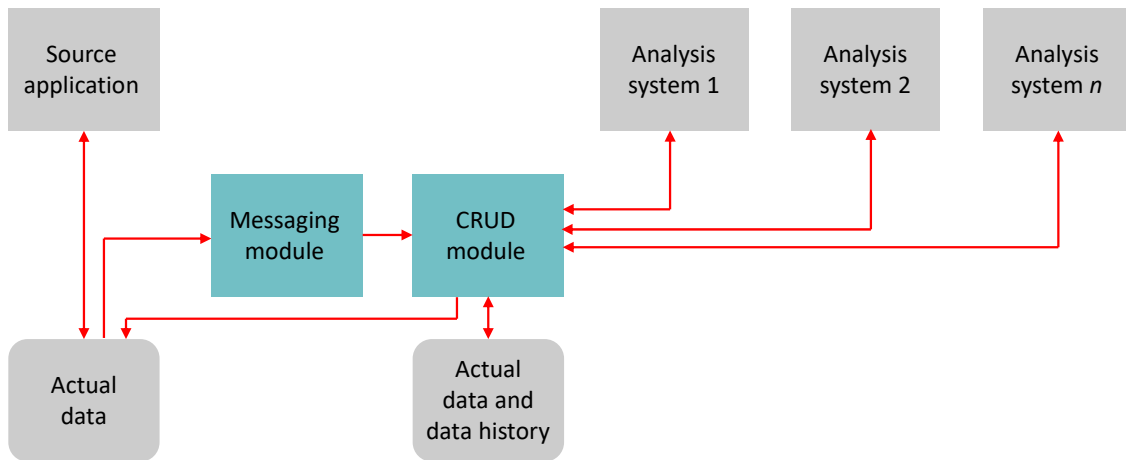
Wrapping the Source Systems - Abstraction



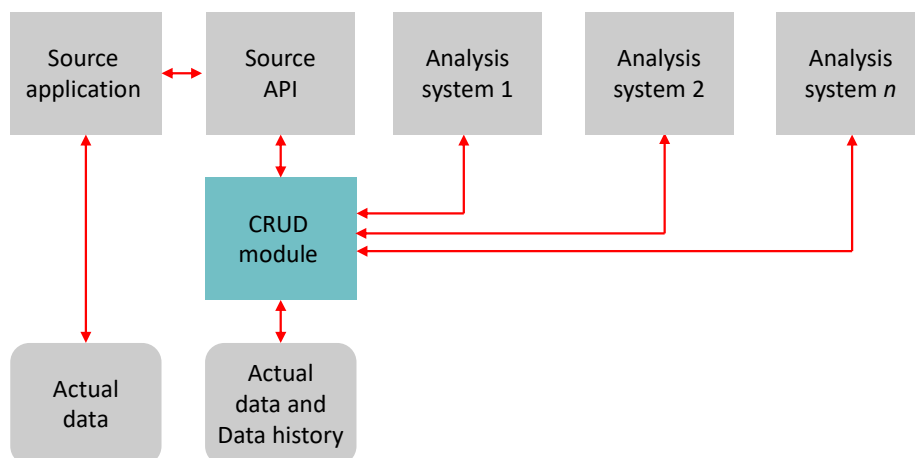
CRUD Module (1) – The Dream



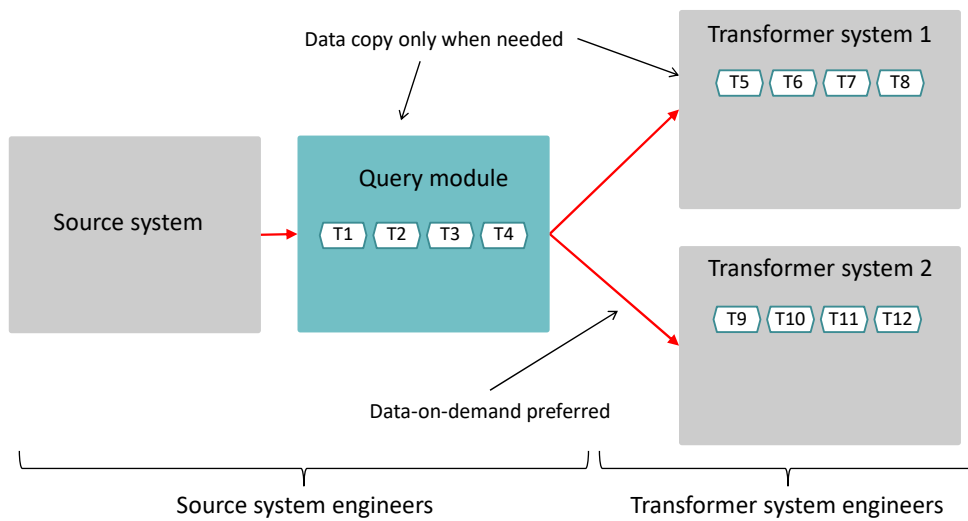
CRUD Module (2) Via Messaging Module



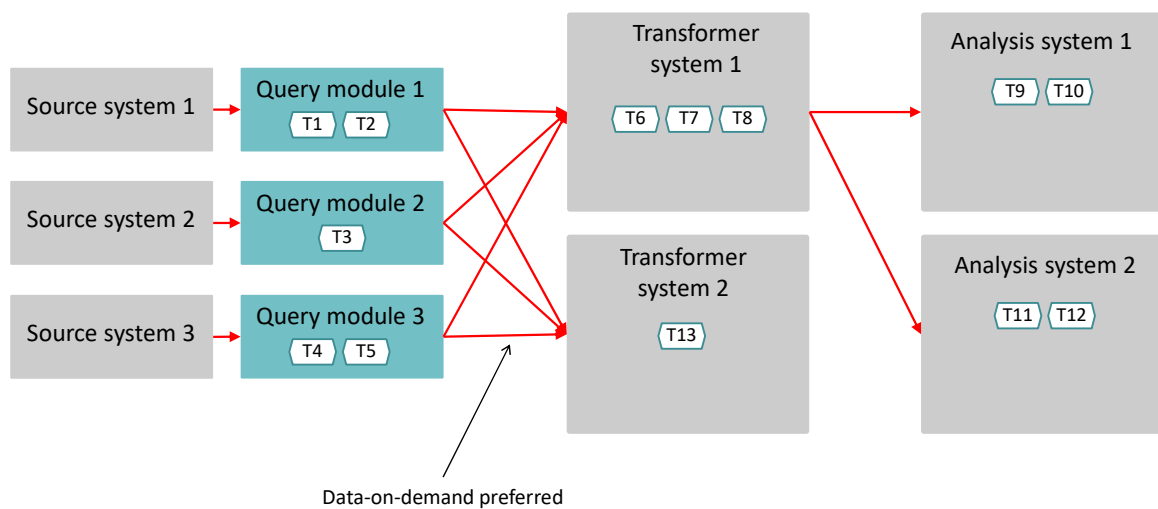
CRUD Module (3) Via API of Source System



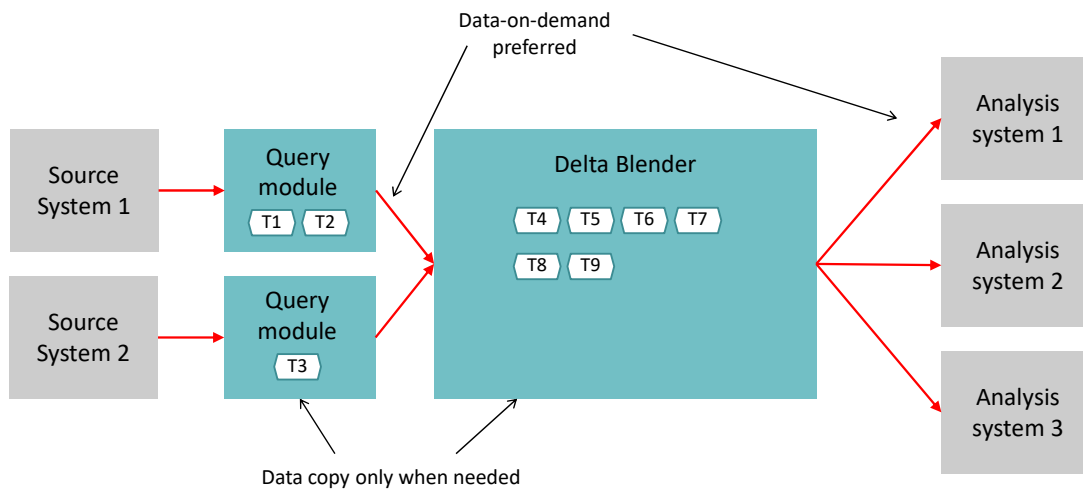
Transformers Closer to the Source Systems



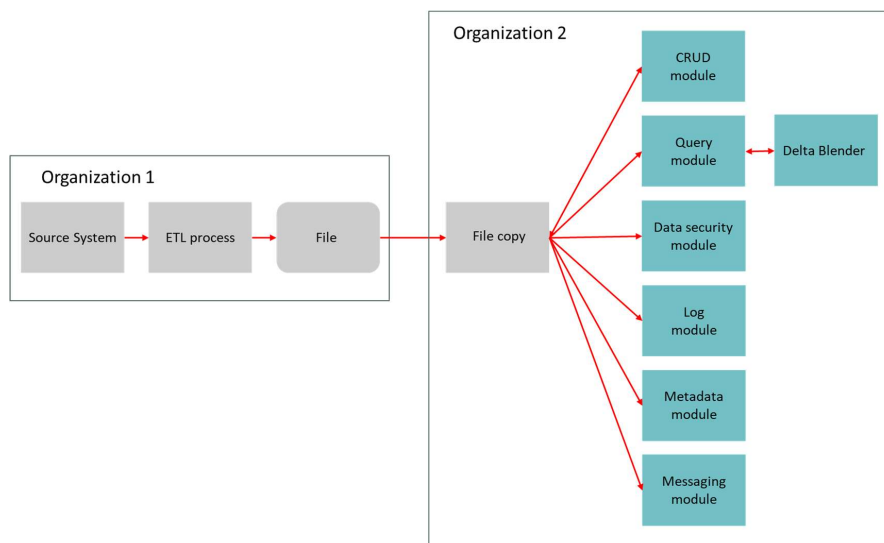
Transforming Data in Delta Modules



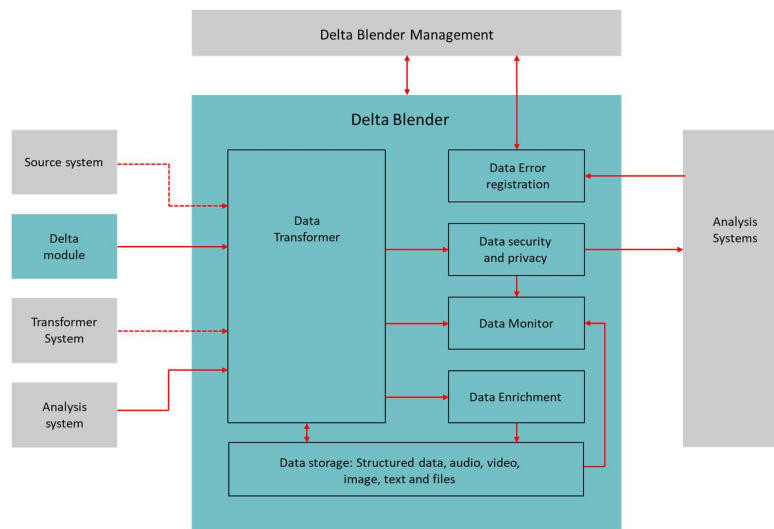
Transforming Data in Delta Blender



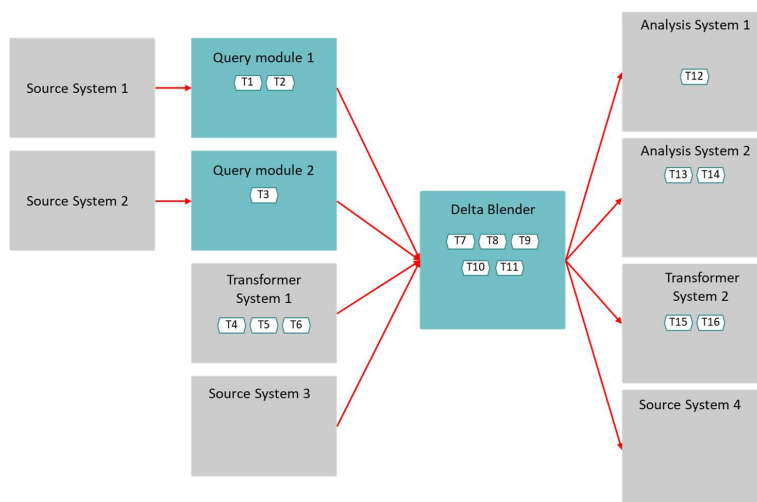
Files Hidden Behind Delta Modules



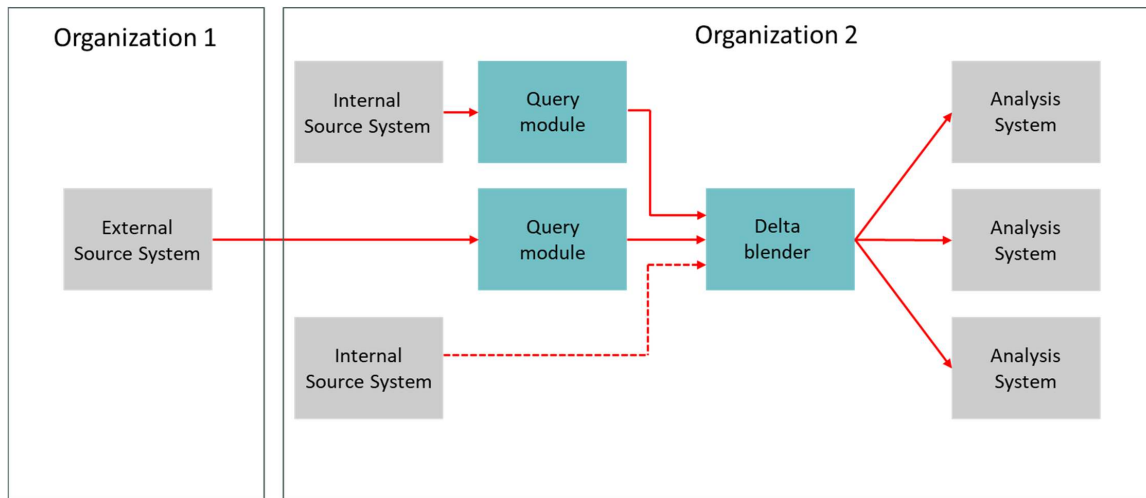
Delta Blender Components



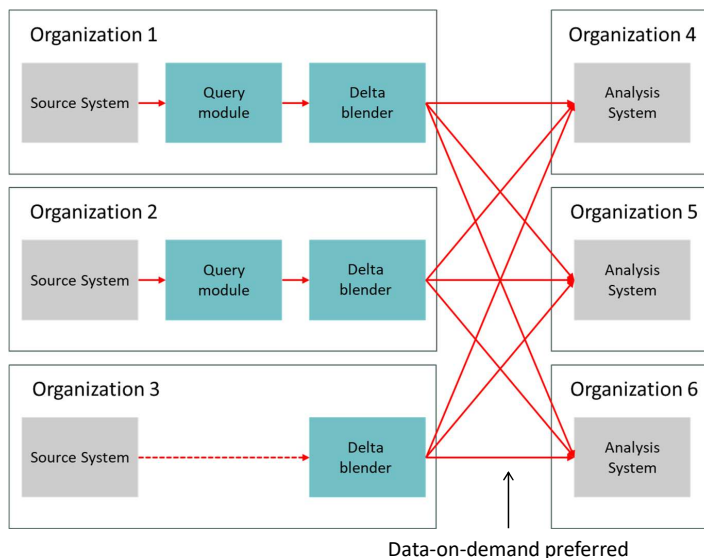
Delta Blender: Transformers



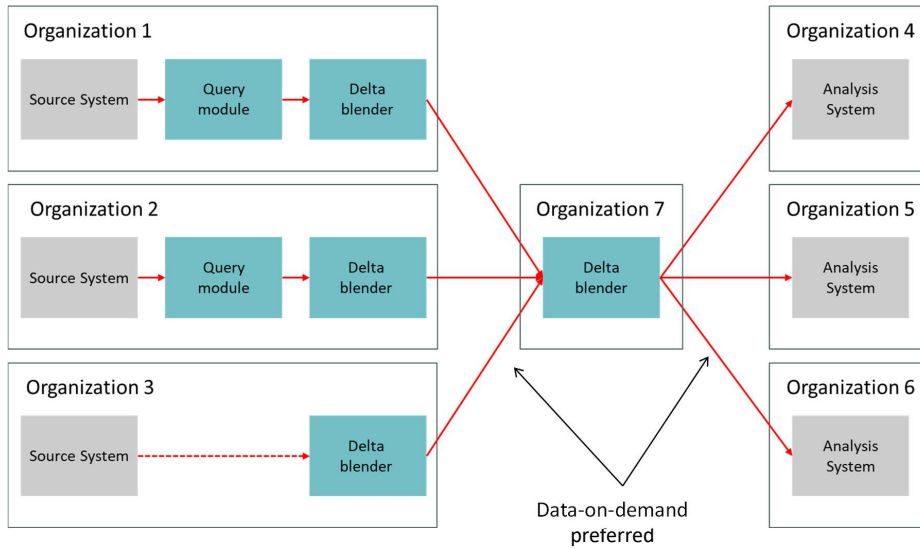
Local Federative Data Architecture



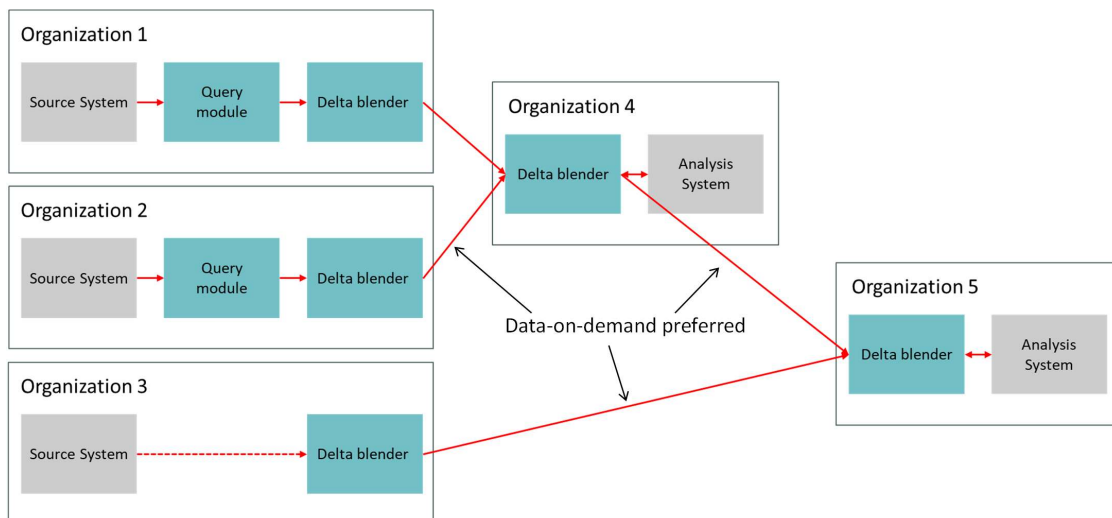
No Federative Data Architecture



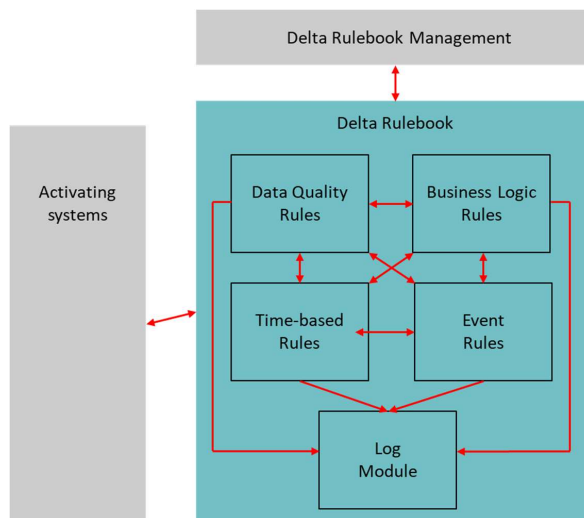
Global Federative Data Architecture



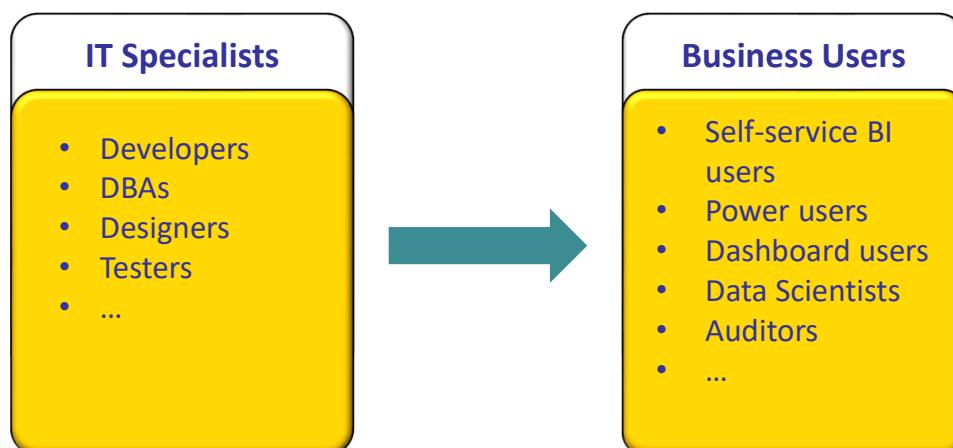
Layered Globale Federative Data Architecture



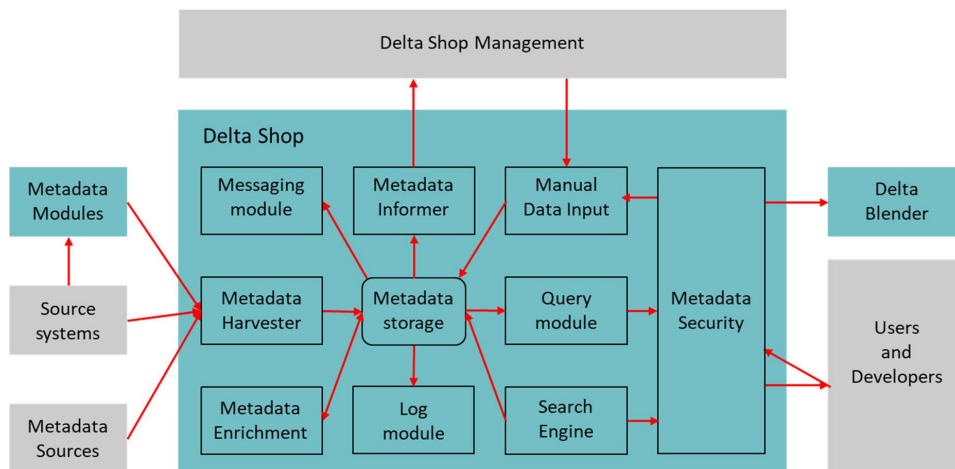
Delta Rulebook Components



Target Audience of Metadata Has Changed



Delta Shop Components



Summary



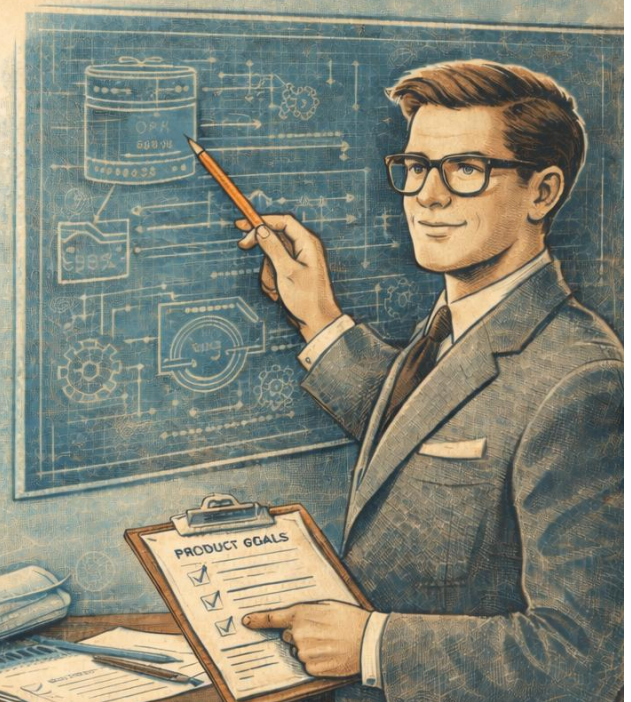
- Unbridled creation of data copies must stop
- Requirement: the (data)house must be in order
- A good (data)house starts with a holistic data architecture
 - Federated data architecture
 - Enterprise data architecture
 - The system is formed by the transformers, not the collection of databases
 - Avoid duplicate transformers
 - Data must be ready-to-use, findable, and insightful
 - Metadata for everyone
 - Operational lineage for reconstruction, transparency, and auditability



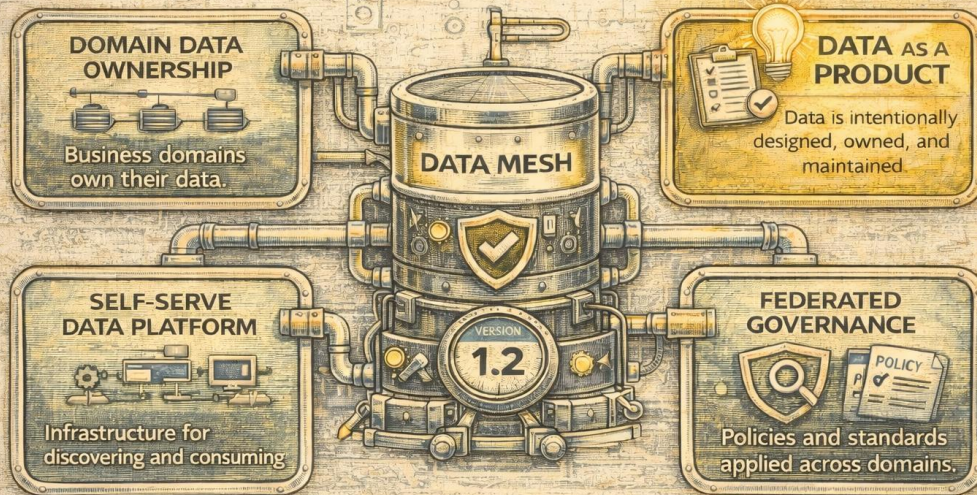
The 5 Lessons of the Open Data Product Specification

Datawarehousing &
Business Intelligence Summit

Ron Tolido
TechstraOrdinary
March 24, Utrecht



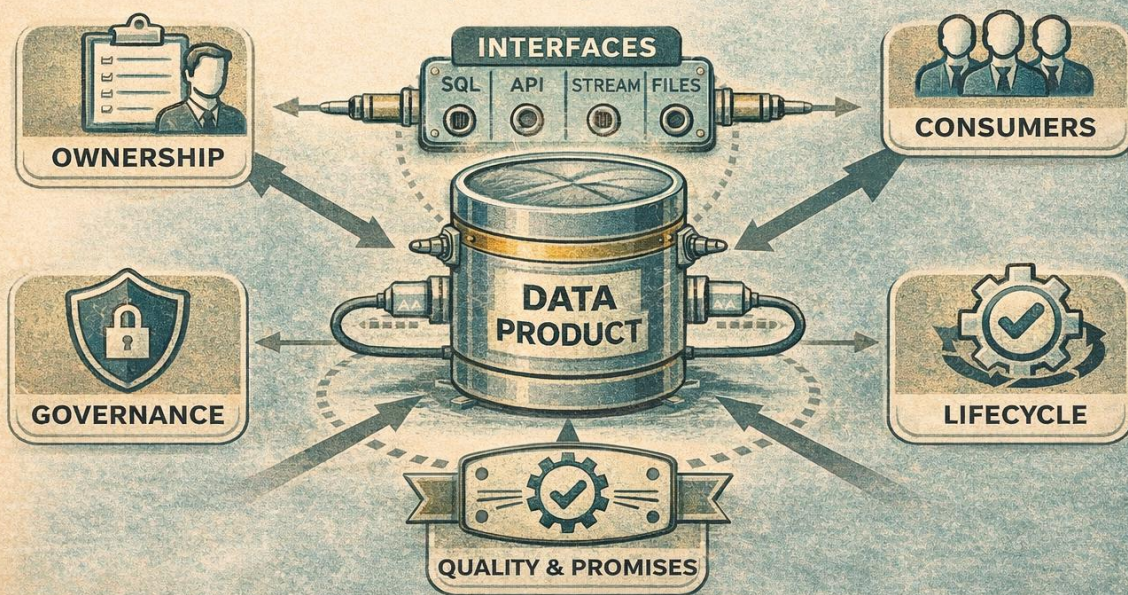
THE FOUR PILLARS OF DATA MESH



Data Mesh gave us the philosophy



A **Data Product** is a reusable, governed data asset that is intentionally designed, owned, and delivered to consumers with clear interfaces, quality guarantees, and lifecycle management.



HOW TO BUILD A PROPER DATA PRODUCT

Lessons from the Open Data Product Specification

1 Purpose Before Plumbing

DATA PRODUCT SPECIFICATION
WHAT:
WHO:
WHY:

A product must clearly state **what** it is, **who** it serves, and **why** it exists.

2 No Owner, No Product

Every product has a named owner responsible for value and success.



3 Interfaces Are Promises

Stable interfaces make consumption predictable and scalable.



4 Trust Is Engineered

Reliability, freshness and accuracy must be explicit guarantees.



5 Governance by Design

Compliance, access and versioning are built into the product itself.

TechstraOrdinary

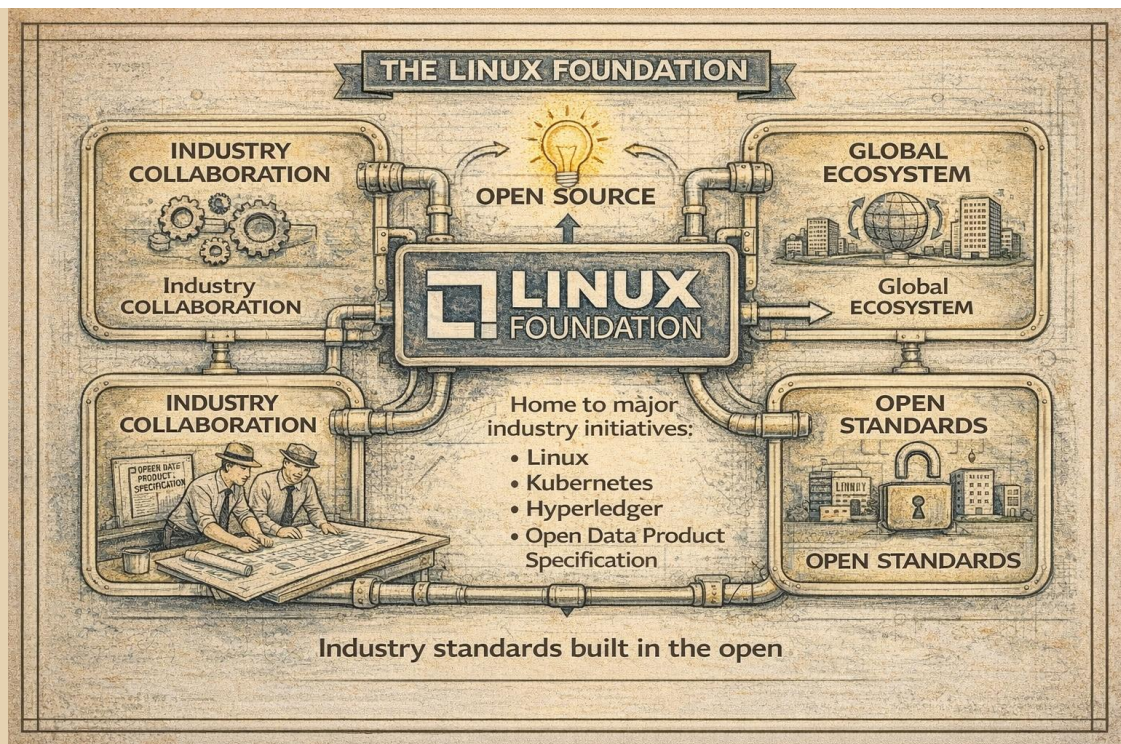


[What is ODPS](#) [Versions](#) [HowTo](#) [Use cases](#) [Testimonials](#) [Development](#)

Open Data Product Specification

Powering Reuse, Revenue, and Rapid Productization

Start Building with ODPS 4.1



ODPS is designed for both open and commercial data products, enabling scalable governance, monetization, and AI-native integration — all through a modular, machine-readable YAML specification. Whether you're offering public sector open data or building monetized data APIs, ODPS provides a standardized structure to define and manage your data products. With support for twelve flexible pricing plan templates, organizations can easily adapt to freemium, subscription, enterprise licensing, or usage-based models.

ODPS 4.1 introduces Product Strategy linking data product KPIs to shared business objectives. Together with in-spec referencing for SLAs, Data Quality, Access, and Pricing Plans, it enables measurable, outcome-driven governance and consistent, reusable metadata across large-scale data ecosystems.

The specification embraces a computational, Everything-as-Code philosophy. Data quality rules, SLAs, pricing enforcement, and contract logic can all be automated and integrated into modern workflows. ODPS supports leading platforms such as Stripe, Checkout.com, PayPal, SodaCL, Monte Carlo, DQOps, and Prometheus — making it compatible with both technical operations and business processes.

Hosted by the Linux Foundation since May 14th, 2024, ODPS is a vendor-neutral, open-source initiative built to reshape how organizations govern, share, and monetize data. Its core mission is to provide a machine-readable, modular, and extensible specification that makes data products interoperable, governable, and AI-ready by design.

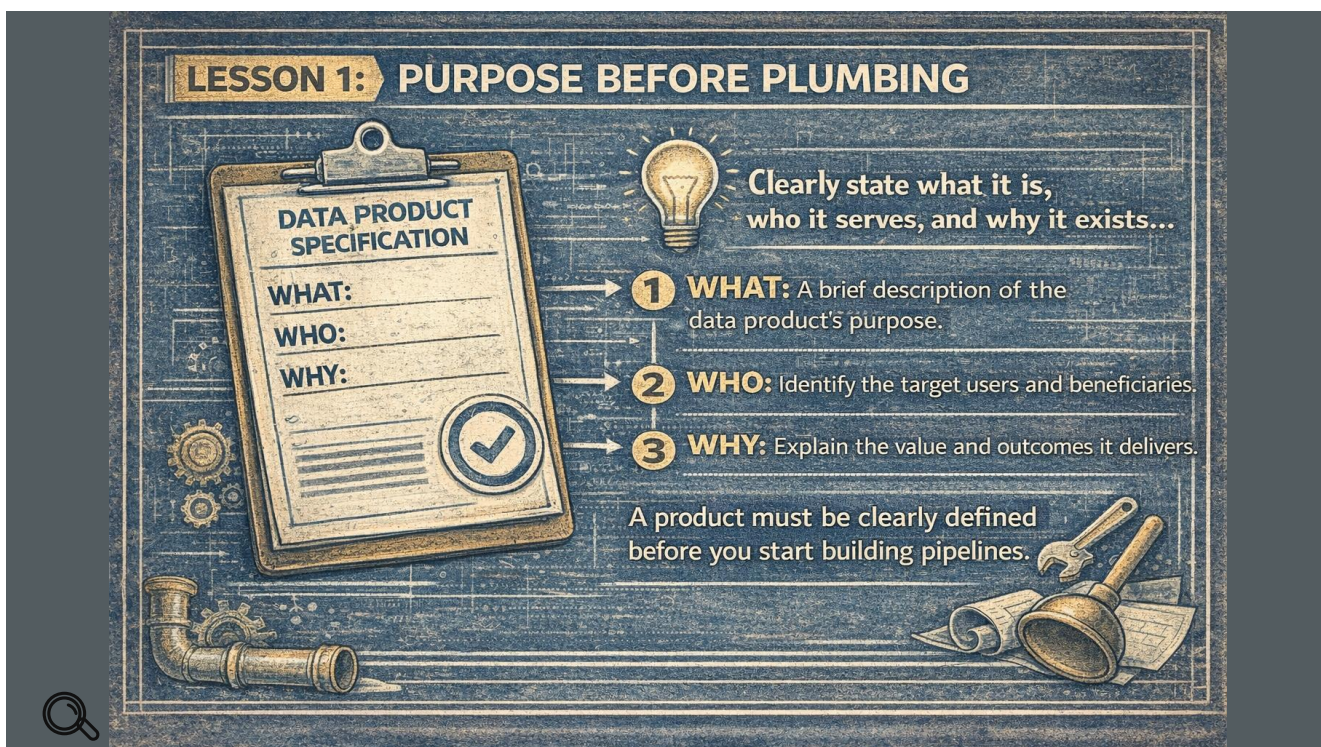
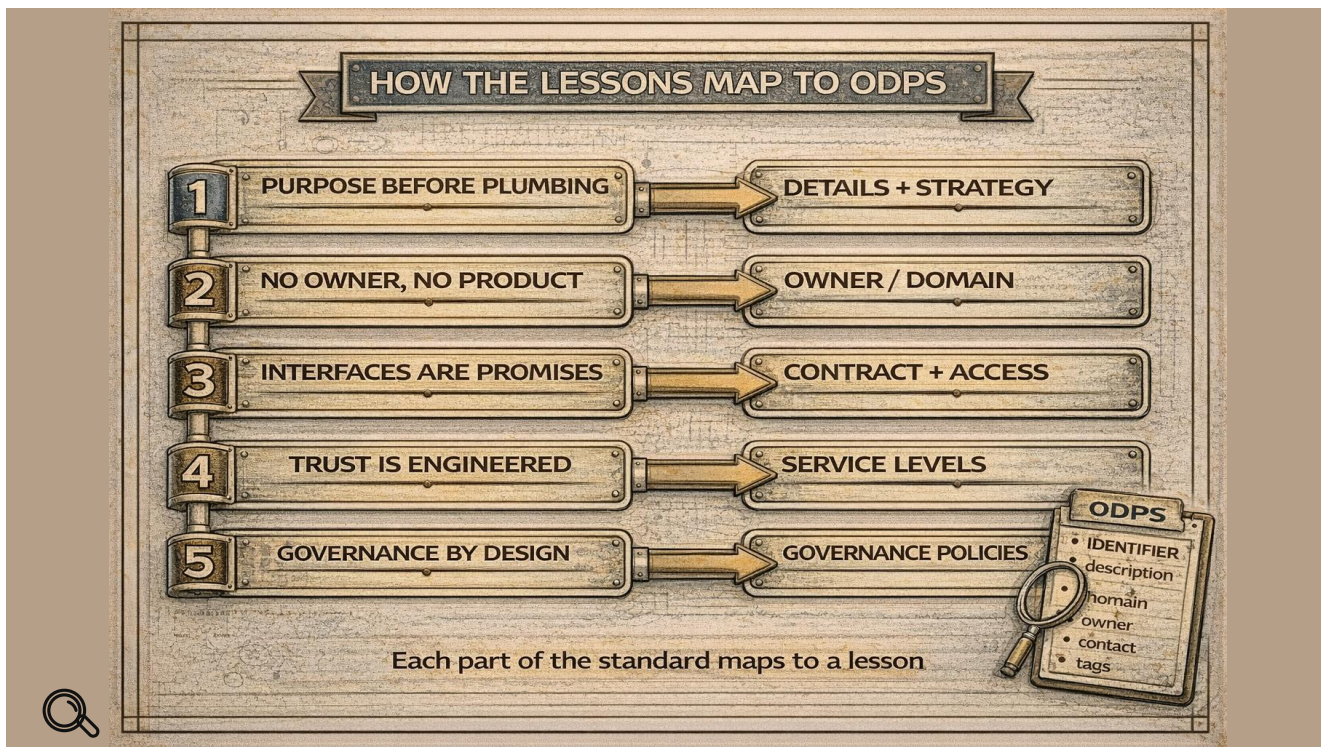


```
schema: 'https://opendataproductions.org/v4.1/schema/odps.yaml'
version: 4.1
product:
  productStrategy:
    objectives:
      - en: Reduce emergency response time
    strategicAlignment:
      - en: Smart City Vision 2030
  contributesToKPI:
    id: bizkpi-city-response-time
    name: City Emergency Response Time
    description: Average minutes from incident to first responder arrival
    unit: minutes
    target: 5
    direction: at_most
    timeframe: "by Q4"
  relatedKPIs:
    - id: bizkpi-traffic-congestion
      name: Traffic Congestion Index
      unit: percentage
      target: -10
      direction: decrease
  productKPIs:
    - id: kpi-detection-coverage
      name: Event Detection Coverage
      description: % of reported incidents captured in real time
      unit: percentage
      target: 95
      direction: at_least
      frequency: hourly
      calculation: detected_events / reported_events

    - id: kpi-time-to-insight
      name: Average Time to Insight
      description: Median time from event occurrence to product update
      unit: seconds
      target: 60
      direction: at_most
      calculation: p50(update_ts - event_ts)
```



```
contract:
  id: 02323M123
  type: ODCS
  contractVersion: 2.2.2
  contractURL: 'https://datamesh-manager.com/urbanltd/dataproducts/9bd530'
details:
  en:
    name: UrbanPulse Events
    productID: urbanpulse-events-001
    valueProposition: >-
      Enable smarter city experiences by providing structured, real-time
      public event data ready for integration into travel apps, tourism
      platforms, and smart city services.
    description: >-
      UrbanPulse Events is a SmartCity data product that aggregates and
      structures public event information – concerts, exhibitions, festivals,
      sports events – making it accessible through APIs and dashboards for
      internal and future external use.
    productSeries: SmartCity Living Data Products
    visibility: internal
    status: production
    productVersion: 0.1.0
    versionNotes: >-
      Initial internal release with basic event metadata structure and shadow
      pricing model implemented.
    issues: >-
      Current limitations include manual ingestion of some event sources and
      partial metadata for smaller events. These will be addressed in the next
      update with automated feeds and metadata enrichment.
    categories:
      - city-events
      - tourism
      - smartcity
    standards:
      - ODPS 4.0
    tags:
      - smartcity
      - events
      - tourism
      - public-data
    brandSlogan: Turning City Buzz into Business Value
    type: dataset
```



Q Search

OPEN DATA PRODUCT SPEC...

Document level structure

Data Product Details

Mandatory attributes and ele...

Optional attributes and eleme...

Data Product Strategy

Data Contract

Data SLA

Data Quality

Data Pricing Plans

Data Licensing

Data Access

Data Holder

Payment Gateways

Specification extensions

Hello world example

Mandatory-only example

Data Product Catalog

ODPS 4.0 → 4.1 Migration G...

Terms used

Editors and contributors



The pattern to implement multilanguage support for data products was adopted from de facto UI translation practices. The attributes inside this element are commonly rendered in the UI for the consumer and providing a simple way to implement that was the driving reasoning. See for example [JSON - Multi Language](#)

name	string	max length 256 chars	REQUIRED The name of the product.
productID	string	max length 256 chars	REQUIRED Product identifier.
visibility	string	one of: private, invitation, organisation, dataspace, public	REQUIRED The publicity level eg who can see this product. Private - just the creator. Invitation - visible only to parties explicitly invited. Organisation - visible to all in your organisation. Dataspace - visible to all existent members of the data space. Public - visible to all publicly.
status	string	one of: announcement, draft, development, testing, acceptance, production, sunset, retired	REQUIRED The status of the product. Lifecycle model discussed in details in here (link).
type	string	one of: raw data, derived data, dataset, reports, analytic view, 3D visualisation, algorithm, decision support, automated decision-making, data-enhanced product, data-driven service, data-enabled performance, bi-directional.	REQUIRED The type of the product. Options are derived from examples and lists found from academic literature.



Q Search

OPEN DATA PRODUCT SPEC...

Document level structure

Data Product Details

Data Product Strategy

Attributes and options

Data Contract

Data SLA

Data Quality

Data Pricing Plans

Data Licensing

Data Access

Data Holder

Payment Gateways

Specification extensions

Hello world example

Mandatory-only example

Data Product Catalog

ODPS 4.0 → 4.1 Migration G...

Terms used

Editors and contributors

License Apache 2.0

Specification home

Linux Foundation



The `productStrategy` object captures the business intent behind a data product and links it to a **single higher-level business KPI** it is primarily accountable for. It defines the objectives, the primary KPI at the business level, and the **product-level KPIs** that measure its direct contribution.

Optionally, it can also include **related KPIs** — secondary or cross-unit measures that:

1. **Capture additional value** — show extra benefits beyond the main KPI, useful for demonstrating broader impact.
2. **Track side effects** — monitor unintended positive or negative impacts on other KPIs.
3. **Highlight cross-department value** — reveal benefits for other business units, strengthening prioritization and funding cases.

Beyond KPIs, `productStrategy` can include:

- **objectives** — natural-language statements describing the business outcomes the product supports.
- **strategicAlignment** — references to corporate initiatives, programs, or policy documents that the product contributes to (e.g., Smart City Vision 2030).

By embedding both primary and related KPI connections directly into the product specification, `productStrategy` ensures that data products remain technically sound, strategically aligned, and transparent in how they deliver value.

Attributes and options

Element	Type	Options	Description
productStrategy	object	–	Top-level block that connects the data product to business goals and KPIs.
objectives	array of objects	language-tagged strings	Business objectives the product supports, written in natural language.
contributesToKPI	object	required	Single higher-level business KPI (from SMART objectives) that this product is accountable for.
id	string	optional	Identifier of the business KPI (use shared IDs for roll-ups).
name	string	required	KPI name.

Example of catalog object usage:

```
schema: https://opendataproducs.org/v4.1/schema/odps.yaml
## JSON schema: https://opendataproducs.org/v4.1/schema/odps.json
version: 4.1
product:
  productStrategy:
    status: Planned
    startDates: 2026-01-12
    endDate: 2026-06-30
    objectives:
      - en: Reduce emergency response time
    strategicAlignment:
      - en: Smart City Vision 2030
    contributesToKPI:
      id: bizkpi-city-response-time
      name: City Emergency Response Time
```

LESSON 2: NO OWNER, NO PRODUCT

1

Every data product has a named owner responsible for value and success.

1

WHAT: A brief description of the data product purpose.

2

WHO: Identify the target users and beneficiaries.

3

WHY: Explain the value and outcomes it delivers.

DATA PRODUCT OWNER

NAME:

EMAIL:

RESPONSIBILITIES:

PRODUCT OWNER

No clear ownership? No data product to deliver.

API

DB

STREAM

ODPS

OPEN DATA PRODUCT SPECIFICATION

Q Search

OPEN DATA PRODUCT SPEC...

Document level structure

Data Product Details

Data Product Strategy

Attributes and options

Data Contract

Data SLA

Data Quality

Data Pricing Plans

Data Licensing

Data Access

Data Holder

Payment Gateways

Specification extensions

Hello world example

Mandatory-only example

Data Product Catalog

ODPS 4.0 → 4.1 Migration G...

Terms used

Editors and contributors

License Apache 2.0

Specification home

Linux Foundation

description	string	optional	Human-readable description.
unit	string	e.g., percentage, minutes, s	Unit of measurement.
target	number/string	-	Target value for the KPI.
direction	enum	increase, decrease, at_least, at_most, equals	Desired direction of movement.
timeframe	string	optional	When the target should be met.
frequency	string	Measurement cadence (e.g., hourly, daily, monthly).	
owner	string	Responsible role/team (optional).	
calculation	string	Human-readable formula (optional).	
productKPIs	array	optional	KPIs measured at product level that influence <code>contributesToKPI</code> . Useful for contribution analysis and governance checks.
relatedKPIs	array	optional	Secondary/cross-unit KPIs to monitor side-effects and additional value (informational; not for prioritization).
strategicAlignment	array	language-tagged strings	Strategic initiatives, policies, or visions the product aligns with.

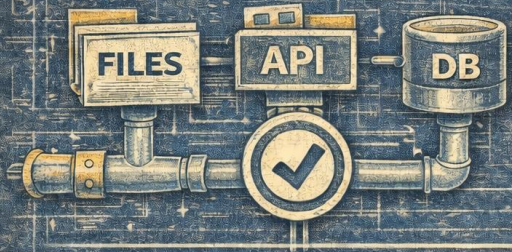
Governance-by-Design checks (Minimum Lovable Gates)

- **Required:** `contributesToKPI.name` present.
- **Recommended:** `≥1 productKPIs` with `unit`, `target`, and `calculation`.
- **Optional:** `relatedKPIs` for secondary/side effects.
- **Traceability:** use shared KPI `ids` to enable cross-product roll-ups against the same business KPI.

```
unit: percentage
target: 95
directions: at_least
frequency: hourly
calculation: detected_events / reported_events

- id: kpi-time-to-insight
  name: Average Time to Insight
  description: Median time from event occurrence to product update
  unit: seconds
  target: 60
  direction: at_most
  calculation: p50(update_ts - event_ts)
```


LESSON 3: INTERFACES ARE PROMISES



SQL | API | Stream | Files
established expectations.

Stable interfaces make consumption
predictable and scalable.

- 1 Clear data contracts.
- 1 Backward-compatible changes.
- 1 Documentation and support.

Interfaces set expectations,
like a promise you must keep or face

BREAKING CHANGES



Q Search

OPEN DATA PRODUCT SPEC...

Document level structure

Data Product Details

Data Product Strategy

Data Contract

Optional attributes and eleme...

Data SLA

Data Quality

Data Pricing Plans

Data Licensing

Data Access

Data Holder

Payment Gateways

Specification extensions

Hello world example

Mandatory-only example

Data Product Catalog

ODPS 4.0 → 4.1 Migration G...

Terms used

Editors and contributors

License Apache 2.0

Specification home

Linux Foundation

The `contract` object in the ODPS schema defines the structure for linking a data product to a formal data contract. It includes an `id` as a unique identifier, a `type` field that specifies the contract standard (either ODCS or DCS), a `contractVersion` indicating the version of the standard, and a `contractURL` that points to an external contract document in a contract management service. Optionally, an inline YAML spec element can be used instead of a URL to embed the contract directly in the product file. These specifications allow data producers to clearly declare governance, usage rights, and technical expectations, supporting interoperability across platforms using standards such as the Open Data Contract Standard (ODCS) or the Data Contract Specification (DCS). More details in the [Knowledge Base](#).

Contract can be defined in 3 ways:

- Use ODPS given attributes to define basic information of the data contract and add URL to details found in contract management system
- Same as above but add contract details as YAML on `spec` element.
- Use `$ref` to define contract object content in a separate file.

Optional attributes and elements

Element name	Type	Options	Description
<code>contract</code>	element	-	Binds together data contract details. You can use both URL and inline (YAML) for data contract content.
<code>id</code>	string	-	UUID of the data contract
<code>type</code>	string	one of: ODCS, DCS	Defines the standard used in data contract. Currently supported options: ODCS and DCS .
<code>contractVersion</code>	string	-	Version of the standard used to define the Data Contract. Type attribute defines the standard/specification. NOTE! This is not the possible iterated version of the data contract itself.
<code>contractURL</code>	URL	Valid URL. See more from RFC 3986 .	URL pointing to data contract in data contract management service or alike. Optionally you can use <code>spec</code> to add data contract details as YAML inline element.

Example of contract object usage:

```
schema: https://opendataproducts.org/v4.0/schema/odps.yaml
version: 4.1
product:
  contract:
    # Optionally se URI of external file, contains all contract details
    $ref: "https://example.org/contracts/default.yaml"
    # Or then define with attributes
    id: 02323M123
    type: ODCS
    contractVersion: 2.2.2
    contractURL: https://demo.datamesh-manager.com/demo834016807886/dataproducts/
    # Or use inline style and add contract under spec element
    spec:
      ....
```




OPEN DATA PRODUCT SPECIFICATION

Q Search

OPEN DATA PRODUCT SPEC...
Document level structure
Data Product Details
Data Product Strategy
Data Contract
Data SLA
Data Quality
Data Pricing Plans
Data Licensing
Data Access
Optional attributes and eleme...
Data Holder
Payment Gateways
Specification extensions
Hello world example
Mandatory-only example
Data Product Catalog
ODPS 4.0 → 4.1 Migration G...
Terms used
Editors and contributors

License Apache 2.0

Specification home

Linux Foundation



The `dataAccess` object defines how users—or machines—can technically access the data product. It allows publishers to describe **multiple, named access methods** tailored to different consumer needs: from simple file downloads and APIs to AI agent integration via protocols like **MCP**.

Each entry under `dataAccess` (such as `default`, `API`, or `Agent`) represents a distinct access interface with its own metadata, authentication requirements, and documentation references. This structure makes it possible to:

- Offer **flexible access modes** for various user personas (analysts, developers, AI agents, etc.)
- Support **multilingual UI presentation** through localized `name` and `description` fields
- Clearly declare **security expectations** using `authenticationMethod`
- Link to both **machine-readable specs** (`specsURL`) and **human-readable guides** (`documentationURL`)
- Promote **reusability** by referencing these interfaces throughout the ODPS YAML using `$ref`

Including an AI agent-specific access interface (`outputPorttype: AI`) supports MCP-based agent interactions, aligning your product with **AI-native data delivery patterns**.

Referencing Examples

For example in your `access` section in Pricing, you can reuse any defined method from `dataAccess` like this: `$ref: '#/Product/dataAccess/default'`

Optional attributes and elements

Element name	Type	Options	Description
dataAccess	object	-	Root-level object containing named access configurations. Each key (e.g., <code>default</code> , <code>API</code> , <code>Agent</code>) defines an access method that can be reused across the ODPS YAML.
\$ref	filepath or valid URL	-	Define the Data Access in external file for reuse purposes, example <code>\$ref: 'https://example.org/dataAccess/all-packages.yaml'</code> See example. This makes it easy to keep related profiles (e.g. <code>default</code> , <code>API</code> , <code>agent</code>) together, apply versioning and validation once, and publish all variants from a single repo or source.



OPEN DATA PRODUCT SPECIFICATION

Q Search

OPEN DATA PRODUCT SPEC...
Document level structure
Data Product Details
Data Product Strategy
Data Contract
Data SLA
Data Quality
Data Pricing Plans
Data Licensing
Data Access
Optional attributes and eleme...
Data Holder
Payment Gateways
Specification extensions
Hello world example
Mandatory-only example
Data Product Catalog
ODPS 4.0 → 4.1 Migration G...
Terms used
Editors and contributors

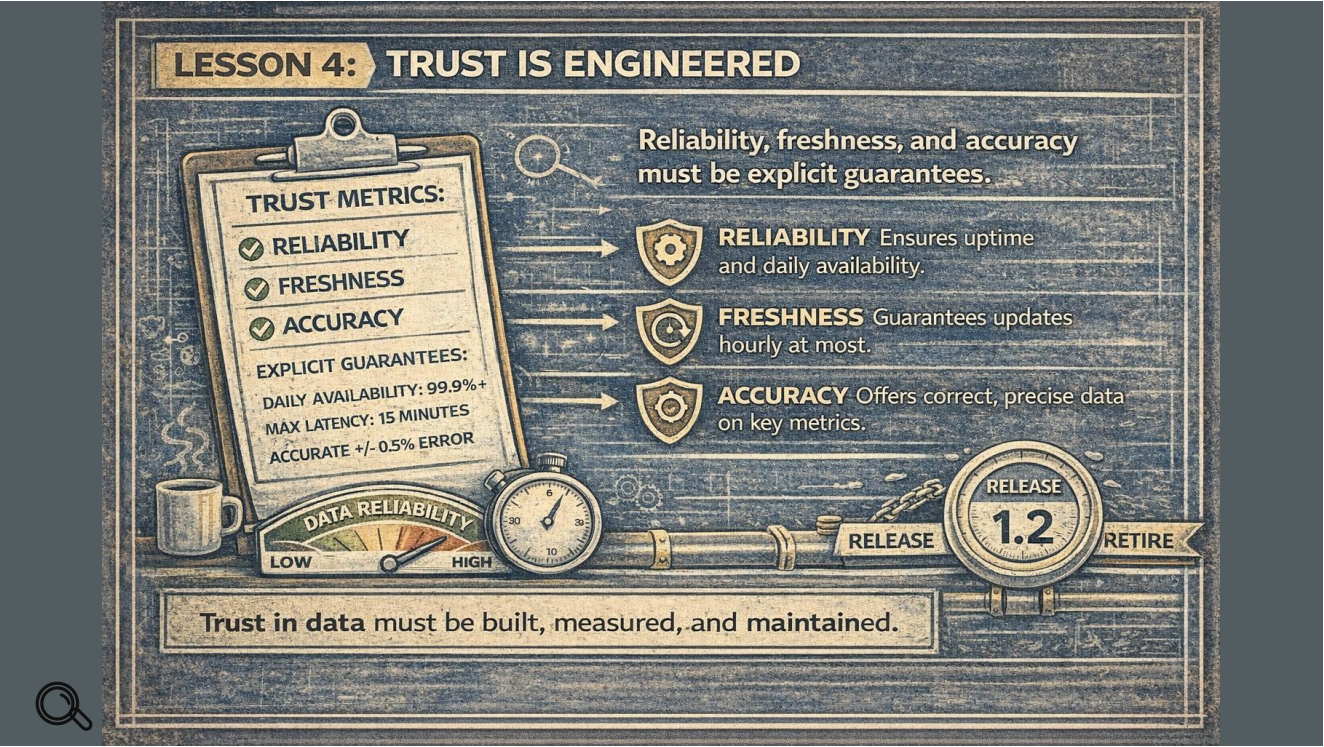
License Apache 2.0

Specification home

Linux Foundation



outputPorttype	string	file, API, SQL, AI, gRPC, sFTP, etc.	Describes the technical method for delivering data (e.g., <code>file</code> for file downloads, <code>API</code> for web services).
format	string	TOON, JSON, XML, CSV, Excel, zip, plain text, GraphQL, MCP	Specifies the data format made available through this access channel.
authenticationMethod	string	OAuth, Token, API key, HTTP Basic, none	Security model required to access the data.
specification	string	OAS, RAML, Slate, MCP	Defines the type of API or protocol specification used to describe access (e.g., OpenAPI, RAML, or a custom protocol like MCP).
specsURL	URL	Valid URL	Points to the machine-readable technical documentation (e.g., OpenAPI YAML).
accessURL	URL	Valid URL	The direct access point to retrieve the data – can be for example an API endpoint or a file link.
documentationURL	URL	Valid URL	A human-readable documentation or guide for access setup, authentication steps, or onboarding.
hashType	string	SHA-1, SHA-2, SHA-256, MD5, etc.	(Optional) Defines hash algorithm used when providing file integrity verification.
checksum	string	any string	(Optional) File hash/checksum value, useful for verifying data integrity after download.





ODPS
OPEN DATA PRODUCT SPECIFICATION

Q Search

OPEN DATA PRODUCT SPEC...

Document level structure

Data Product Details

Data Product Strategy

Data Contract

Data SLA

SLA can be defined with 11 sta...

Optional attributes and eleme...

Data Quality

Data Pricing Plans

Data Licensing

Data Access

Data Holder

Payment Gateways

Specification extensions

Hello world example

Mandatory-only example

Data Product Catalog

ODPS 4.0 → 4.1 Migration G...

SLA can be defined with 11 standardized dimensions with decoupled Everything as Code monitoring

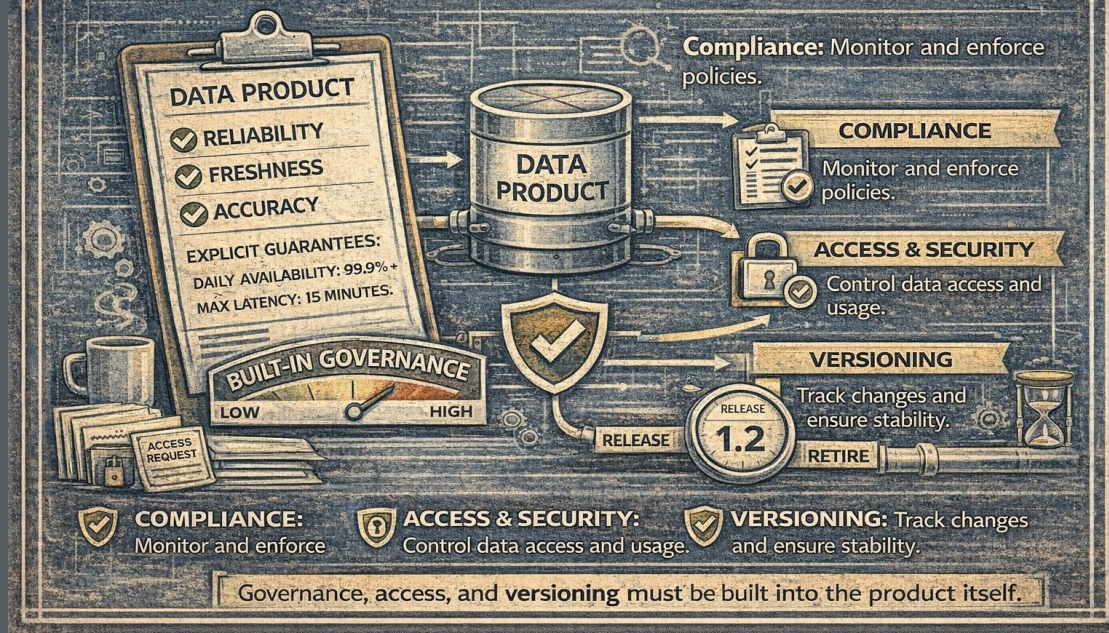
SLA Dimension	Description
latency	minimal amount of time before getting any response.
uptime	Uptime is a measure of system reliability, expressed as the percentage of time a machine, typically a computer, has been working and available. See more https://uptime.is/ .
responseTime	amount of time to process external request.
errorRate	Maximum tolerated errors in data, percentage.
endOfSupport	The date at which your product will not have support anymore.
endOfLife	The date at which your product will not be available anymore. No support, no access.
updateFrequency	how often data is updates.
timeToDetect	How fast can you detect a problem?
timeToNotify	Once you see a problem, how much time do you need to notify your users?
timeToRepair	How long do you need to fix the issue once it is detected?
emailResponseTime	How long do you need to respond to email support requests?

Example of SLA component usage:

```
SLA:
  declarative:
    - dimension: uptime
      displaytitle:
        - en: Uptime
        objective: 99
        unit: percent
    - dimension: responseTime
      objective: 200
      unit: milliseconds
```



LESSON 5: GOVERNANCE BY DESIGN



MINIMUM LOVABLE GOVERNANCE FOR DATA PRODUCTS

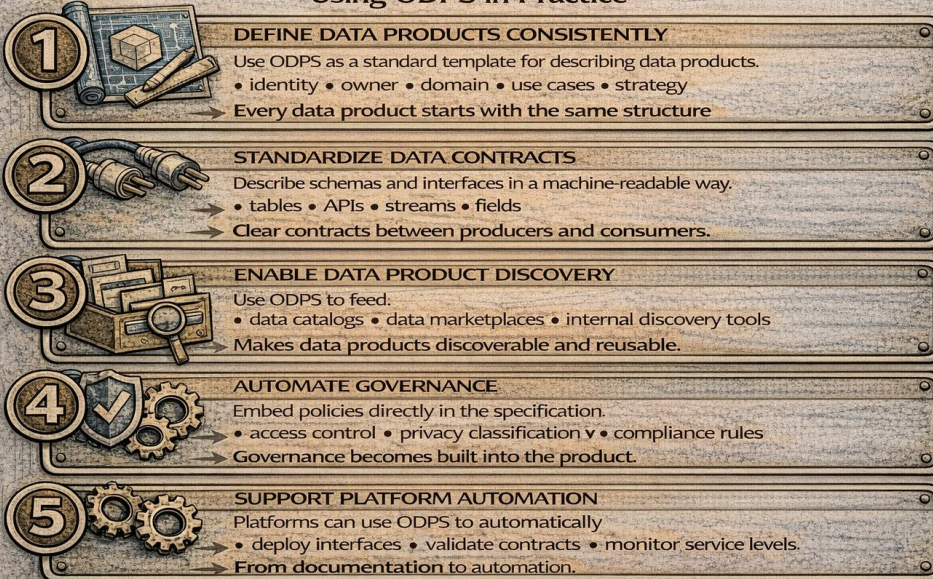
Aspect	Data Product Management (MLG)
Primary Goal	Deliver value and outcomes
Focus	Data as a product with lifecycle, owner, and customers
Approach	Lightweight, embedded governance within each product
Process	Iterative, automated, tied to delivery
Success Measure	Business KPIs, adoption, trust, and impact

Core Principles of MLG

1. **Value First** — Every data product must tie to at least one business KPI. Governance begins with outcomes, not just controls.
2. **Lean but Strong** — Only apply rules that matter. Automate wherever possible. Don't create work that nobody uses.
3. **Governance by Design** — Governance is built into the product itself: metadata, contracts, SLAs, and quality checks travel with the product.
4. **People-Centric** — Governance should be simple, clear, and supportive. Product teams should see it as an enabler, not a blocker.

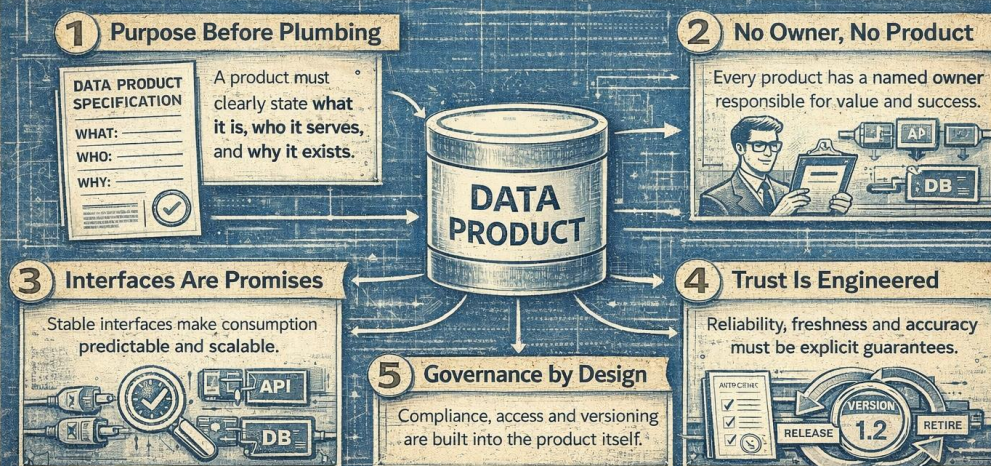
THE ODPS TOOLBOX

Using ODPS in Practice



HOW TO BUILD A PROPER DATA PRODUCT

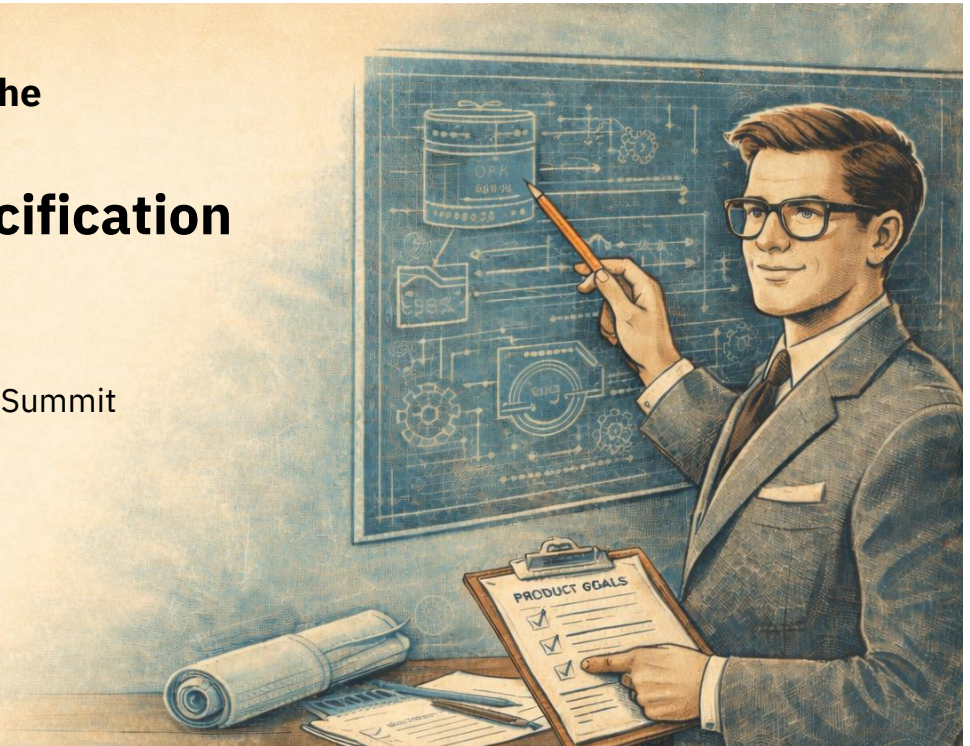
Lessons from the Open Data Product Specification



The 5 Lessons of the **Open Data** **Product Specification**

Datawarehousing &
Business Intelligence Summit

Ron Tolido
TechstraOrdinary
March 24, Utrecht



DAIG

DATA & AI GOVERNANCE

PARTNERS

Data Governance Sprint™

Establish Data Governance in Weeks instead of Months

DW & BI Summit 2026
Utrecht – March 24, 2026

© 2026 Data & AI Governance Partners
Reproduction without permission is prohibited.

Who is **MATHIAS** Vercauteren

Consulting & Advisory Services

- DAMA –DMBOK 3.0 (Project Manager)
- UZA (Hospital)
- MLOZ (Healthcare Insurance)
- Monument Group (Insurance)
- De Lijn (Logistics)
- MPET (Logistics)
- Securex (Professional Services)
- Federal Insurance (Insurance)
- Flemish Government (Governmental Institution)
- Belfius (Financial Services)
- Barry Callebaut (Manufacturing)
- Carrefour (Retail)
- Hilti (Manufacturing)

Research

- President of Data & AI Governance Research Institute (2026, founding phase)
- PhD in Data Governance (AMS, 2025 - 2029)
- Book "Data Governance Sprints" (Technics Publication, est. Q2 2026)
- Data Governance Digest (LinkedIn newsletter)



- PhD in Data Governance
- MSc in Business Economics
- BSc in Sociology

(AMS, 2025 - 2029)
(2012, Ghent University)
(2009, Ghent University)



Educational Services

Training and Coaching Engagements – both in-house and classroom:

- Data Governance
- AI Governance
- DAMA-DMBOK® / CDMP®
- Data Strategy
- Data Quality
- Master Data Management

Speaking Engagements:

- DGIQ/EDW (San Diego, 2026)
- Data Modeling Zone (San Francisco, 2026)
- DGIQ/EDW (Anaheim, 2025)
- Data and AI Conference (London, 2025)
- Data Modeling Zone (Phoenix, 2025)
- DGIQ East (Washington DC, 2024)
- Data and AI Conference (London, 2024)
- DGIQ West (San Diego, 2024)
- Enterprise Data World (Orlando, 2025)
- DG & MDM Conference (London, 2023)
- DGIQ East (Washington DC, 2023)



AGENDA: Welcome to “Data Governance Sprints™”

We'll cover the following topics



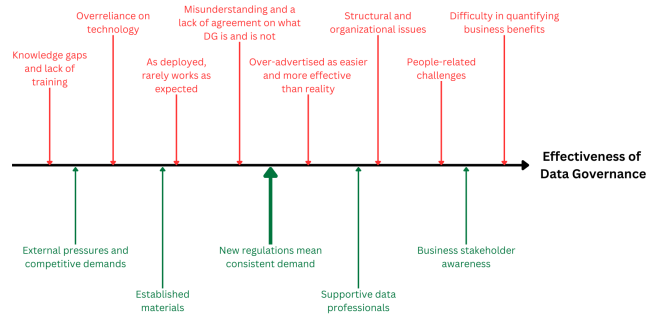
[Back to Agenda](#)

1

Data Governance is **Failing!**

Opinion & Analysis

Data Governance Is Failing – Here's Why



Representative photo by Claudio Schwarz on Unsplash

Author: 

Thomas C. Redman, Anne Marie Smith, John Ladley, Mathias Vercauteren, Malcolm Hawker, Aaron Wilkerson

Published on: Jan 22, 2025, 4:00 pm

Source: Redman, T. C. (2025, January 22). Data governance is failing—Here's why. *CDO Magazine*. <https://www.cdomagazine.tech/opinion-analysis/data-governance-is-failing-heres-why>

MOST Data Governance programs fail!

- Most businesses today are **unfit for data**.
- Data Governance programs are **bolted on, not built in**.
- Policies, standards, glossaries, roles all exist **on paper**.
- The data office is a **supplicant**.
- People experience it as **extra work** added on top of their “day job”, **competing** with “real work”.
- **People side** versus Technology side!
- Growing complexity of **modern data organizations**.
- Where's **the value**? And what about **the communication**?
- Data governance, as practiced, is still an **unproven theory**.
- **Disconnect** between **speed** of business and speed of a data governance program.
- **Traditional** Data Governance strategies fail...time and time again!

What is DATA GOVERNANCE?

The exercise of authority, control, and shared decision-making (planning, monitoring, and enforcement) over the management of data assets. ~ DAMA DMBoK

Rules: The standards and guidelines for data use.

- Purpose: Define how data should be managed to maintain quality, ensure compliance, and protect sensitive information.
- Examples:
 - Data privacy policies
 - Data quality standards
 - Access control protocols

Roles: The people responsible for managing, governing, and overseeing data.

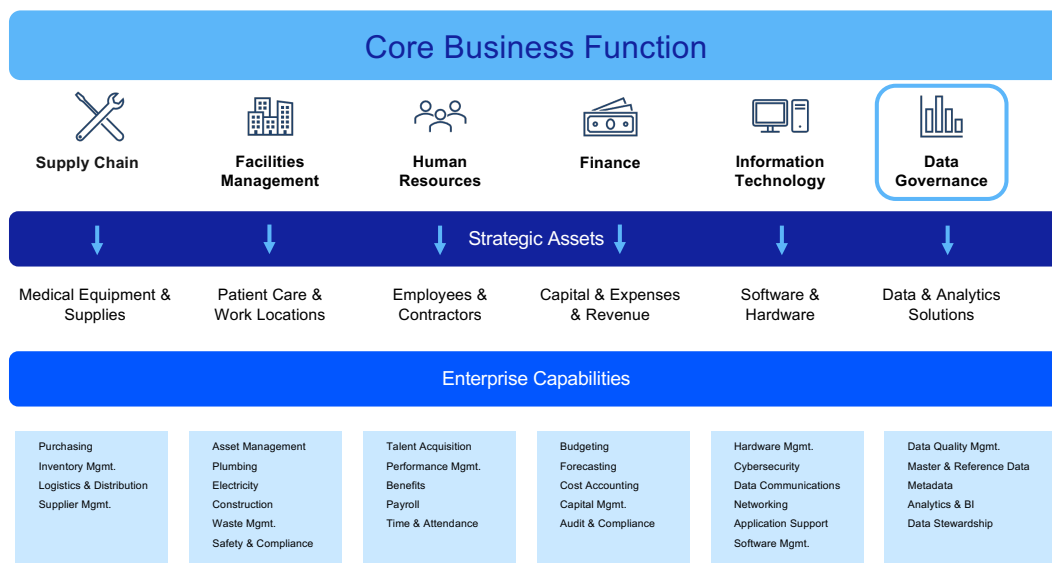
- Purpose: Establish accountability, define responsibilities, and ensure collaboration across the organization.
- Examples:
 - Data Stewards
 - Data Owners
 - Data Custodians

Control: The mechanisms that ensure compliance and enforcement.

- Purpose: Monitor adherence to rules, enforce policies, and mitigate risks through oversight and governance structures.
- Examples:
 - Data audits and compliance reviews
 - Data governance councils or committees
 - Data quality monitoring

Rules, Roles, and Control work together to turn data into a valuable, trusted asset.

Data Governance is MANAGING data as an ASSET



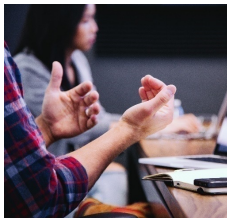
2

Why **Workshops** are Better

© 2026 Data & AI Governance Partners
Reproduction without permission is prohibited.

9

How effective are your **MEETINGS**?



Endless debate
cycles



Biased meetings &
wrong priorities



Frustrated teams &
loss of momentum



Communication
barriers

Why workshops are **BETTER** than meetings?

- ☐ Grab pen and paper
- ☐ I flash some words on the screen
- ☐ Take 20 seconds to write down every word you remember
- ☐ I'll show the words again, how many did you remember?

Cat Fish Tree Flower **Tape**

Key *Door* Head **King**

Hammer **Ball** Square Shoe

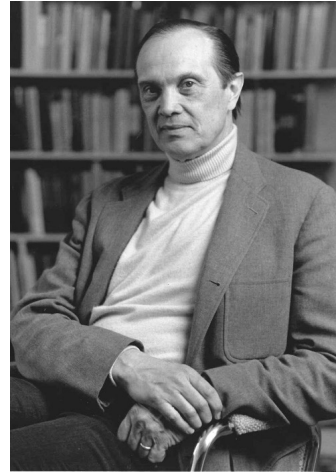
Apple House Box Ball

Milk

The magical number 7 (plus or minus 2)

Some limits on our capacity for processing information.

So how does this play out in a meeting?



George Miller

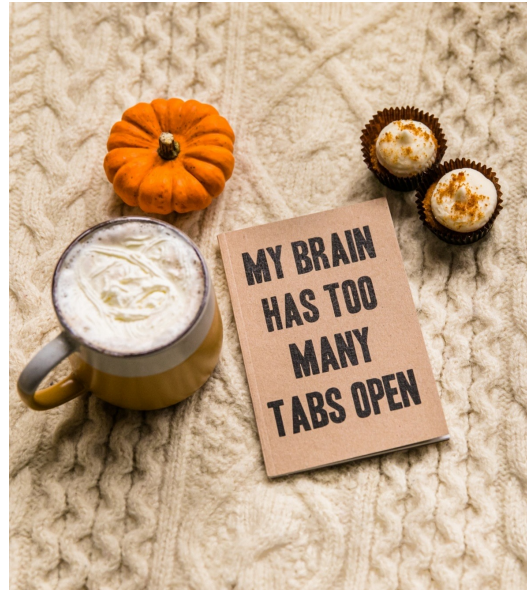
A **TYPICAL** meeting



The **BRAIN** stuff

In meetings, you have a ton of information being produced, but people don't actually remember. Our brains struggle to process this data avalanche, especially when cognitive biases like **recency bias** and **primacy bias** come into play.

Our brains are just not made to process the amount of data that happens in a meeting.



A workshop is a **POWERHOUSE!**

*Workshops allow you to **visualize** meetings and make **decisions** based on the data that is visualized around you.*

3

Breaking Free from Legacy Data Governance

© 2026 Data & AI Governance Partners
Reproduction without permission is prohibited.

17

The **SPEED PROBLEM** in data governance

The gap between rapid business movement and slow data governance implementation, plus the failure of traditional strategies, highlights the need for a more flexible and pragmatic approach.

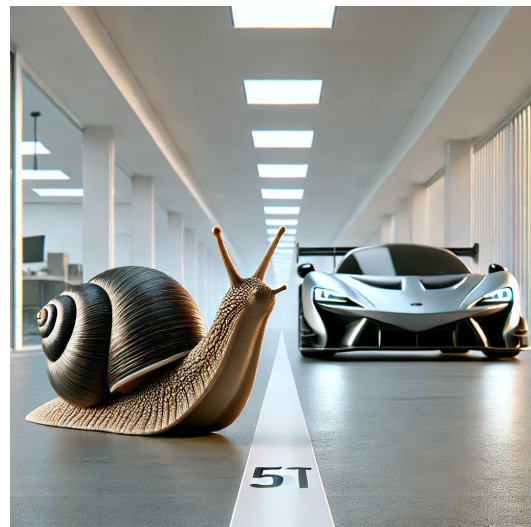


Image Credit: Image generated by DALL-E, OpenAI.

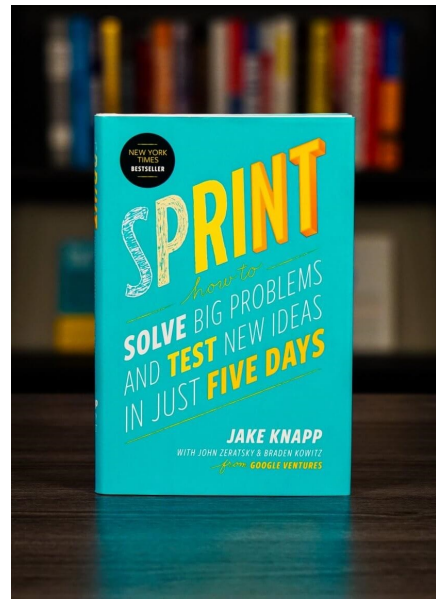
History of the DESIGN SPRINT

Jake Knapp worked at Microsoft and later at Google.

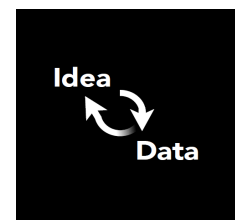
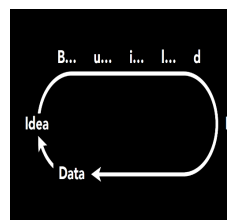
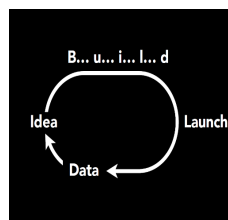
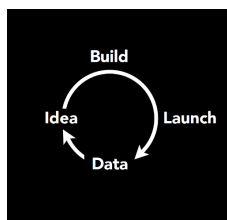
He experienced firsthand how building products would often take months, or even a year, before they could be launched.

Something clicked for Jake in 2009 when he was in Stockholm for **1 week** in January. During that short amount of time, he **designed, prototyped, and tested an idea** - which turned out to be Google Hangout.

That's when he realized that there had to be a better way to build products!

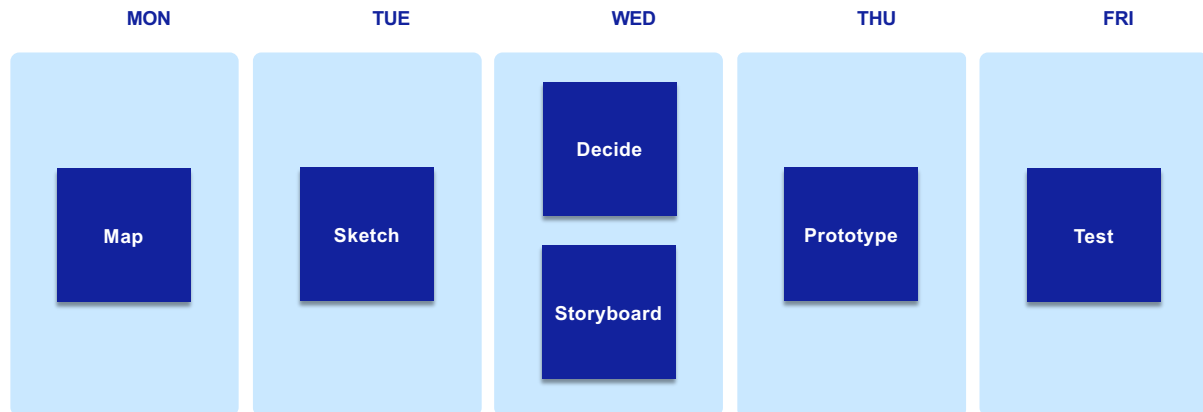


Get data about an idea **FAST!**



The Design Sprint

The **STANDARD** design sprint



DATA GOVERNANCE SPRINT™: The secret to speed

Imagine a five-week journey where you **design**, **build**, and **test** core components of your data governance program. This is what our Data Governance Sprint™ is all about.

It's a structured, **hands-on experience** for data leadership teams, that uses a proven, repeatable approach that transforms your data governance concepts into reality **in weeks, not months**.

From designing a winning, action-oriented data governance strategy that moves your company to managing data as an asset, to creating a business glossary and piloting an operating model with engaged data stewards—this **iterative approach** helps you make tangible progress in no time, while keeping momentum and morale high.

Data Governance **SPRINT**™ to the rescue

1 Establishing Data Governance **CAPABILITIES**



2 Through Incremental Implementation of **USE CASES**



Data Quality



Metadata



Critical Data Elements



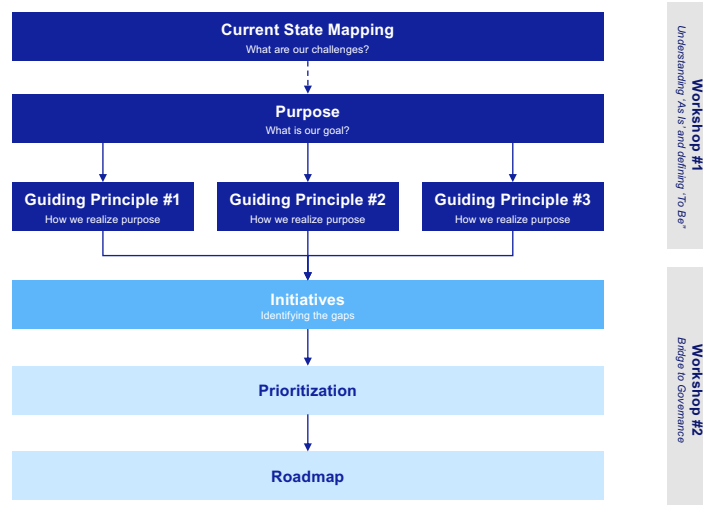
Data Literacy

3 In **WEEKS** Instead of Months

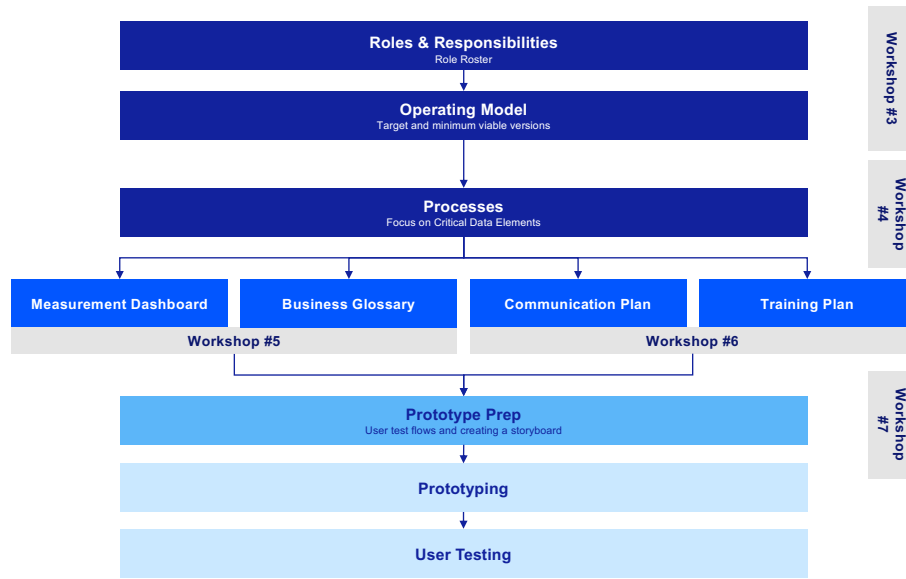


Data Governance SPRINT™

Data Governance **SPRINT**™



Data Governance SPRINT™



4

Data Governance Sprint™: Week by Week

WEEK 1 of 5

Laying the Groundwork | The Fundamentals



WEEK 1 of 5

The Fundamentals | Workshops 1 and 2 of 7

Week 1



The Fundamentals Workshops: Establishing the Foundation for Data Governance

The Fundamentals workshops are the cornerstone of our Data Governance Sprint™, where we take a comprehensive look at how your organization currently manages its data assets and chart a clear path toward your desired future state.

By the end of The Fundamentals workshops, you'll have a thorough understanding of the **"As Is"** state, the **"To Be"** vision, and **the gaps** that need to be bridged. This clarity will provide a strong, actionable foundation for the remainder of the DG Sprint™ and beyond, ensuring that your data governance efforts are aligned with your strategic objectives and ready to drive meaningful progress.

Workshop 1: Thursday Full-Day

- Focus on understanding the current state of data management within your organization.
- Identify existing challenges and examine current practices.
- Define the "To Be" state, our vision for data governance, aligned with business objectives and strategic goals.
- Work through four key exercises: The Challenges Map, Challenges Interview, Purpose, and Guiding Principles.

Workshop 2: Friday Full-Day

- Highlight the gaps between the current state and the future vision established on Thursday.
- Prioritize the actions needed to bridge these gaps.
- Focus on improving the quality and understanding of critical data elements.
- Work through five engaging exercises designed to create a clear roadmap with prioritized use cases.

WEEK 1 of 5

The Fundamentals | Workshop 1 of 7

Exercise 1. The Challenges Map

Exercise 2. Challenges Interview

Exercise 3. Purpose

Exercise 4. Guiding Principles

We start by capturing the data challenges and pain points on a Challenges Map.

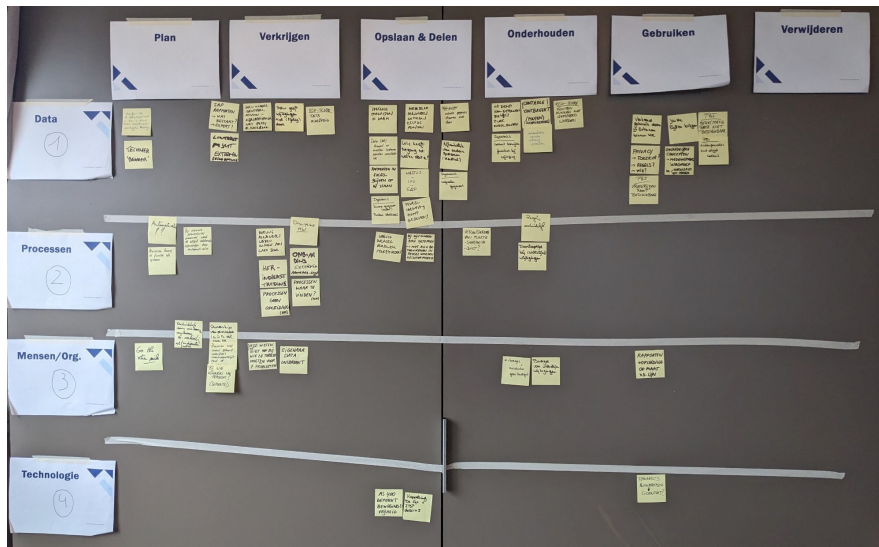
The purpose is to get a **high-level overview** of how the key components (data, processes, people, & technology) interact with the information life cycle within the organization.

It helps to **identify root causes** and **determine improvements** needed to correct existing problems and prevent them from reappearing.

The challenges MAP



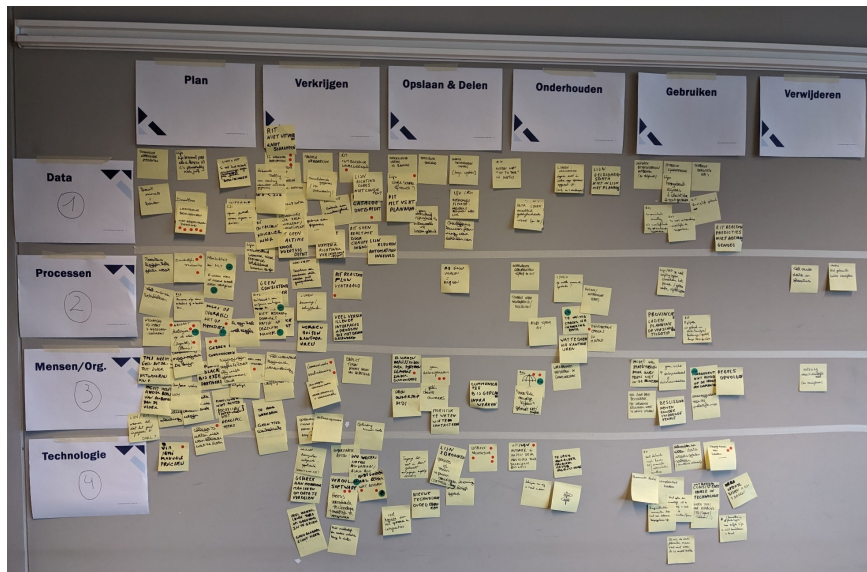
The Challenges Map



The challenges MAP



The Challenges Map



WEEK 1 of 5

The Fundamentals | Workshop 1 of 7

Exercise 1. The Challenges Map

Exercise 2. Challenges Interview

Exercise 3. Purpose

Exercise 4. Guiding Principles

We will try to get a deeper understanding of the challenges and the **broad-impact components** that affect the quality of information.

This is a **structured discussion** where we listen to our participants and collect all the challenges that come up during the interview.

We capture our thoughts as questions in the **'How Might We'** format. The HMW format helps us define our biggest challenge without prescribing a solution.

Next, we narrow down the Challenges Interview (written in the form of "How Might We" questions) and identify **focus areas** in the Challenge Map.

Challenges **INTERVIEW**

Workshop 1

Exercise 2



Let the broad-impact components guide your open-ended questioning

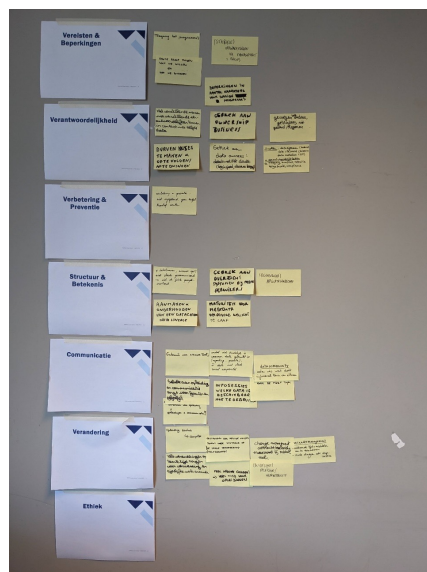
Requirements and Constraints	Business, User, Functional, Technology, Legal, Regulatory, Compliance, Contractual, Industry, Internal Policies, Access, Security, Privacy, Data Protection
Responsibility	Accountability, Authority, Ownership, Governance, Stewardship, Motivation, Reward
Improvement and Prevention	Continuous Improvement, Root Causes, Prevention, Correction, Enhancement, Audit, Controls, Monitoring, Metrics, Targets
Structure and Meaning	Definitions, Relationships, Metadata, Standards, Reference Data, Data Models, Business Rules, Architecture, Semantics, Taxonomies, Ontologies, Hierarchies
Communication	Awareness, Engagement, Outreach, Listening, Feedback, Trust, Confidence, Education, Training, Documentation
Change	Management of Change and Associated Impact, Organizational Change Management, Change Control
Ethics	Individual and Societal Good, Justice, Rights and Freedoms, Truthfulness, Standards of Conduct, Avoiding Harm, Supporting Well-Being

Source: **Broad Impact Components** © 2005, 2020 Danette McGilvray, Granite Falls Consulting, Inc. www.gfalls.com

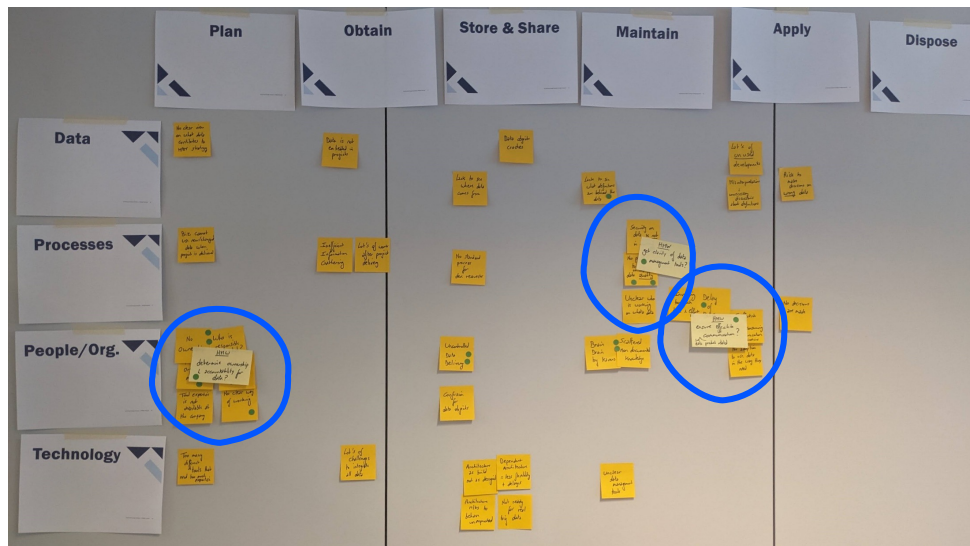
Challenges **INTERVIEW**

Workshop 1

Exercise 2



The challenges MAP



WEEK 1 of 5

The Fundamentals | Workshop 1 of 7

Exercise 1. The Challenges Map

Exercise 2. Challenges Interview

Exercise 3. Purpose

Exercise 4. Guiding Principles

In this exercise, we will focus on identifying the **fundamental reason why** we are implementing data governance within our organization. This isn't just about the technical aspects or the tasks we need to perform; it's about aligning our data governance efforts with the overarching business objectives and goals.

Understanding this purpose will ensure that every data governance initiative we undertake directly contributes to the success and strategic direction of the company.

A well-defined purpose will become **our Vision**, guiding our decisions, helping us prioritize, and making sure that everyone—across all levels—understands the critical role that data governance plays in our collective success.

WEEK 1 of 5

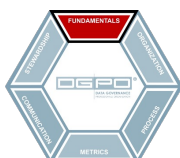
The Fundamentals | Workshop 1 of 7

Exercise 1. The Challenges Map

Exercise 2. Challenges Interview

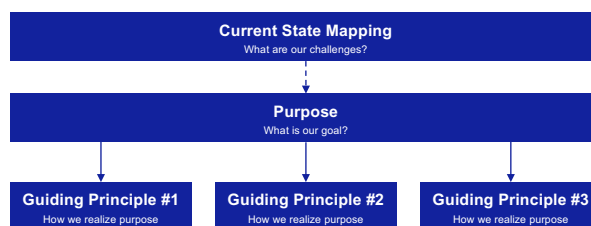
Exercise 3. Purpose

Exercise 4. Guiding Principles



As we build upon the foundation we've established in our previous exercises—assessing the current state of our data landscape, highlighting key use cases and pain points, and defining our vision and purpose—it's now time to finalize our approach by defining **our Guiding Principles**.

These principles are the **fundamental beliefs** that will steer your organization's approach to data. Think of them as your **North Star**, providing clear direction and setting the stage for how data should be managed and treated within your company.

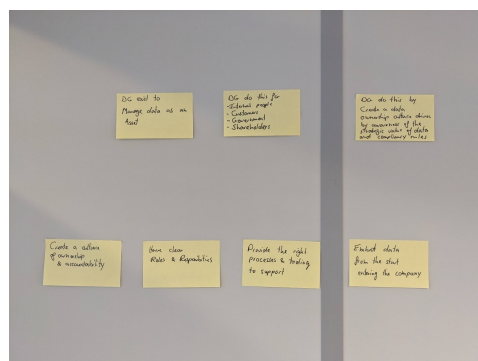


Guiding PRINCIPLES

We realize this by...

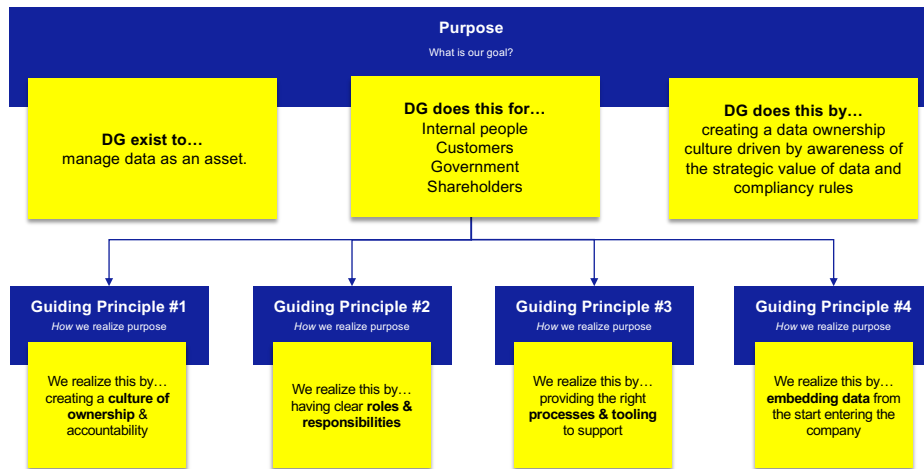
We realize this by...

We realize this by...



WEEK 1 of 5

The Fundamentals | Workshop 1 of 7



WEEK 1 of 5

The Fundamentals | Workshop 2 of 7

Exercise 1. 10 for 10 Brainstorm

Exercise 2. Action Board

Exercise 3. Action Steps

Exercise 4. Roadmap

Exercise 5. CDE Shortlist

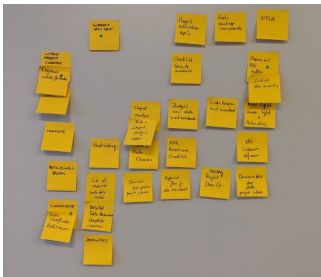
Let's kick off our creative engines with the 10 for 10 Brainstorm! This exercise is all about generating a wide range of ideas to **bridge the gap** between our current state and the future vision for data governance.

In this session, we'll draw inspiration from our Purpose, Guiding Principles, and the focus areas identified in our Challenges Map. The aim is to come up with **as many ideas as possible**—whether they seem practical or far-fetched, all ideas are welcome!

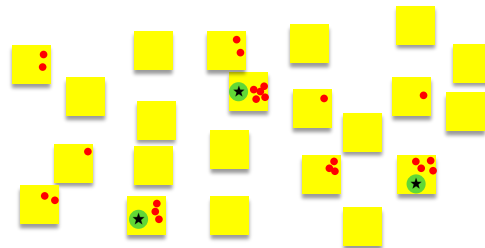
Why is this important? Because the more ideas we generate, the more likely we are to uncover innovative solutions that will propel us forward.

10 for 10 BRAINSTORM

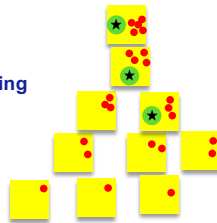
Presenting the ideas



Voting on the ideas



Revisualizing



WEEK 1 of 5

The Fundamentals | Workshop 2 of 7

Exercise 1. 10 for 10 Brainstorm

Exercise 2. Action Board

Exercise 3. Action Steps

Exercise 4. Roadmap

Exercise 5. CDE Shortlist

Welcome to our Action Board exercise, where we'll take all the great ideas generated during the 10 for 10 Brainstorm and **prioritize** them based on their **impact and feasibility**.

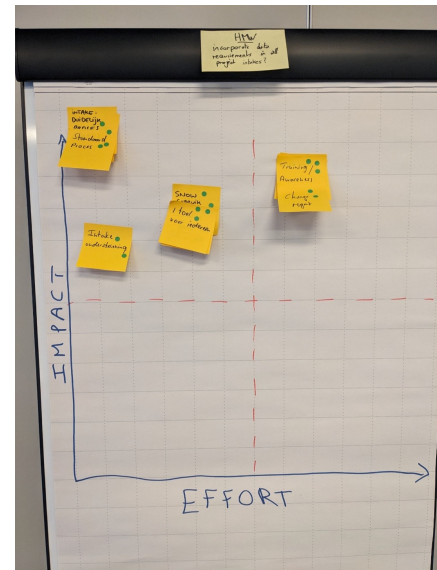
The Action Board is a powerful tool that helps us focus on what really matters—those initiatives that will drive the most significant change with the least amount of effort. By the end of this exercise, we'll have a clear understanding of which ideas we should tackle first and which ones may require more time and resources.

Why is this important? Because prioritization is key to ensuring that our efforts are directed toward the **most valuable and achievable goals**. This process will help us create a realistic and actionable roadmap that aligns with our data governance objectives.

ACTION Board

Workshop 2

Exercise 2



WEEK 1 of 5

The Fundamentals | Workshop 2 of 7

Friday

Exercise 1. 10 for 10 Brainstorm

Exercise 2. Action Board

Exercise 3. Action Steps

Exercise 4. Roadmap

Exercise 5. CDE Shortlist

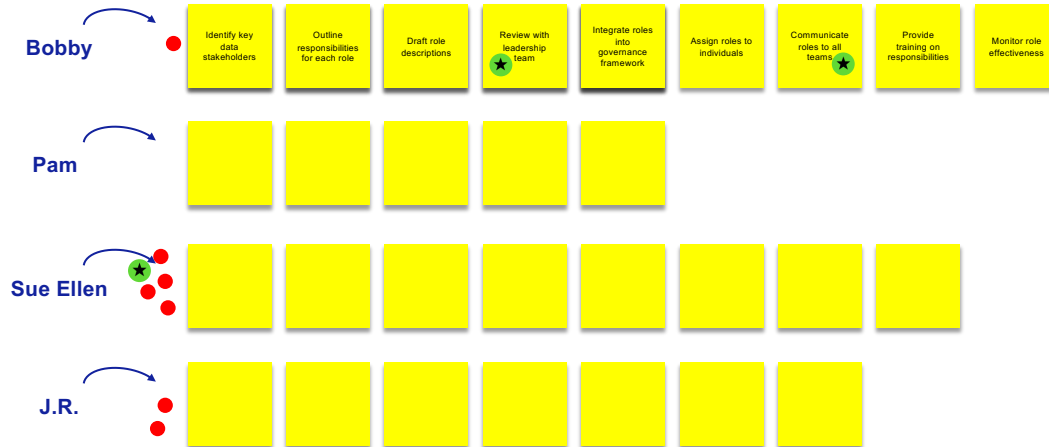
Next up, the Action Steps exercise, where we'll transform our prioritized "quick win" ideas from the Action Board into **actionable steps**.

This exercise is crucial because it **turns concepts into concrete actions**, ensuring that we can effectively implement our ideas and start seeing tangible results.

The focus here is on **high-level, strategic steps** that will guide us in bringing these quick wins to life, moving us closer to achieving our data governance goals.

Why is this important? Because having a clear, actionable plan is what separates good ideas from real, impactful change. By the end of this session, we'll have a series of task flows that outline the steps necessary to implement each of our selected initiatives.

ACTION Steps



WEEK 1 of 5

The Fundamentals | Workshop 2 of 7

Exercise 1. 10 for 10 Brainstorm

Exercise 2. Action Board

Exercise 3. Action Steps

Exercise 4. Roadmap

Exercise 5. CDE Shortlist

Let's hit the road, Jack! In this Roadmap exercise we'll take the prioritized initiatives and their corresponding task flows and strategically place them on a timeline.

This exercise is critical because it transforms our ideas and plans into a **clear, actionable roadmap** that outlines when and how each initiative will be implemented.

By the end of this session, we'll have a comprehensive plan that not only reflects our data governance goals but also aligns with the organization's overall strategy and priorities.

Why is this important? Because having a well-structured roadmap ensures that our efforts are organized, realistic, and geared toward **achieving tangible results** within a defined timeframe. This roadmap will serve as our guide, helping us maintain momentum and focus as we move forward with our Data Governance Sprint.

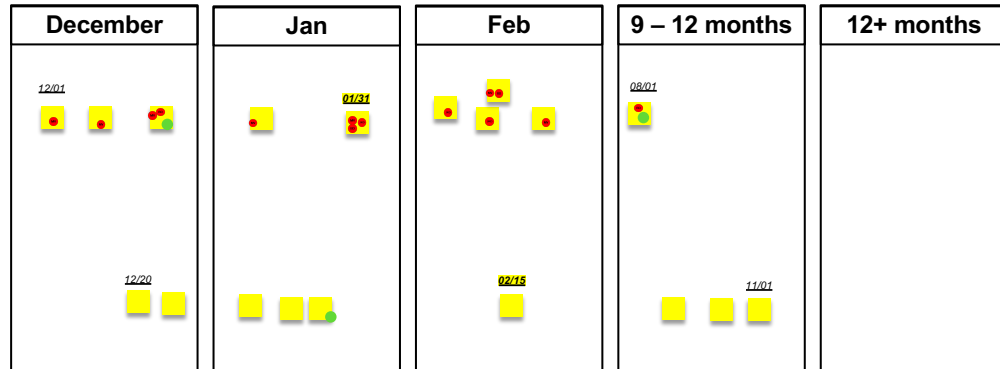
ROADMAP Exercise

Workshop 2

Exercise 4

★
Define roles
for data
accountability

★
Create a
centralized
data catalog



BE

SE

SE

BE

PM

PM

SE

J.R.

SE

J.R.

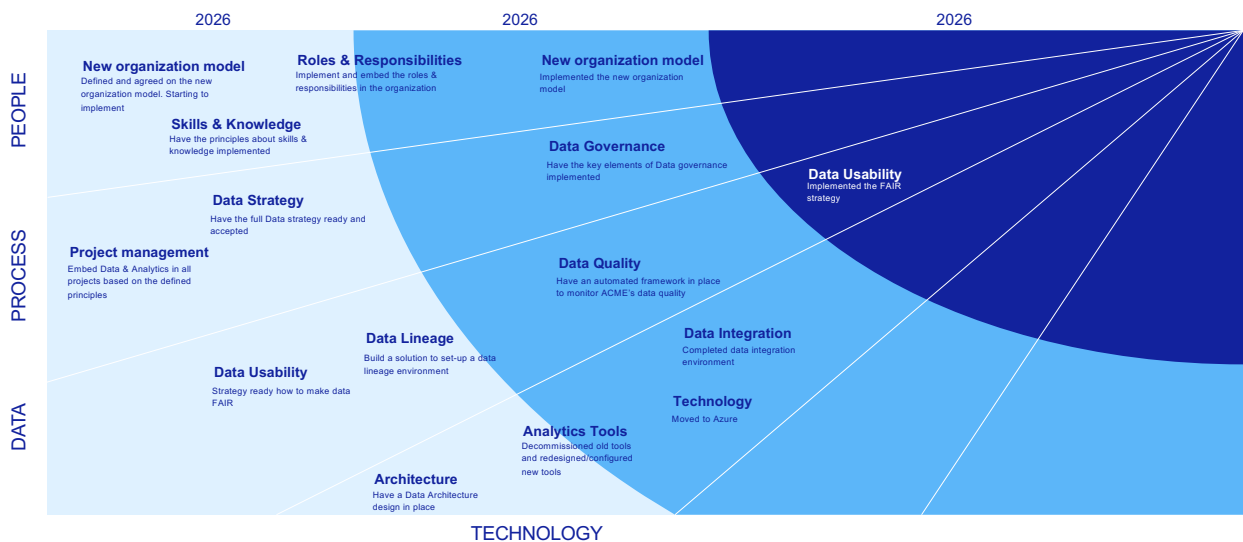
BE

J.R.

WEEK 1 of 5 Workshop 2 of 7 | Roadmap

Workshop 2

Template



WEEK 1 of 5

The Fundamentals | Workshop 2 of 7

Exercise 1. 10 for 10 Brainstorm

Exercise 2. Action Board

Exercise 3. Action Steps

Exercise 4. Roadmap

Exercise 5. CDE Shortlist

Welcome to the final exercise of our workshop, the **Critical Data Element (CDE) Shortlist**. In this session, we'll focus on identifying and prioritizing the most important data elements within our organization.

By narrowing down our focus to the top 10 CDEs, we lay the foundation for effective data governance, enabling us to make fast progress and **demonstrate value quickly**.

These CDEs will later be central to our prototyping phase, where we'll conduct data quality assessments and define these elements in our business glossary. For now, our goal is to identify and select these key data elements that are crucial to our organization's success.

CDE Shortlist



WEEK 2 of 5

The Data Players | The Rules of the Game

MON	TUE	WED	THU	FRI
Documentation Summarizing the outcomes and draft deliverables	Workshop 3 of 7 The Data Players ca. 6h, with the entire team	Documentation Summarizing the outcomes and draft deliverables	Workshop 4 of 7 The Rules of the Game ca. 6h, with the entire team	Documentation Summarizing the outcomes and draft deliverables
Workshop set-up Preparing for the next workshop		Workshop set-up Preparing for the next workshop		Workshop set-up Preparing for the next workshop

WEEK 2 of 5

The Data Players | The Rules of the Game

Week 2

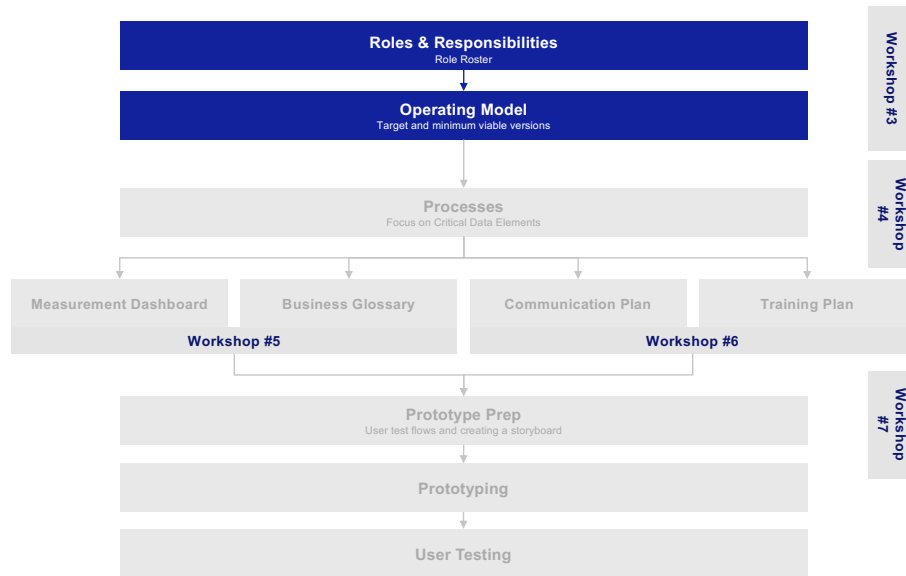


Welcome to Week 2: Building the Data Governance Framework

In Week 2, we transition from strategic visioning to the practical development of your **data governance framework**. This week is dedicated to defining the key roles, responsibilities, and operating models that will drive your data governance efforts, as well as establishing the foundational processes for managing your Critical Data Elements (CDEs).

Through two targeted workshops, we'll develop clear, actionable deliverables that will ensure your data governance structure is both **robust and sustainable**. By the end of this week, you'll have a well-defined framework that aligns with your organization's goals and prepares you for the next steps in your data governance journey.

Data Governance SPRINT™



WEEK 2 of 5

The Data Players | Workshop 3 of 7

Tuesday



Exercise 1. Responsibility Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Exercise 4. Roles Roster

In this exercise, we'll embark on a journey to uncover and define the **key responsibilities** that are essential for achieving our data governance vision.

Rather than starting with predefined roles, we'll focus first on identifying **what needs to be done**—what are the crucial tasks, decisions, and accountabilities that will drive our success?

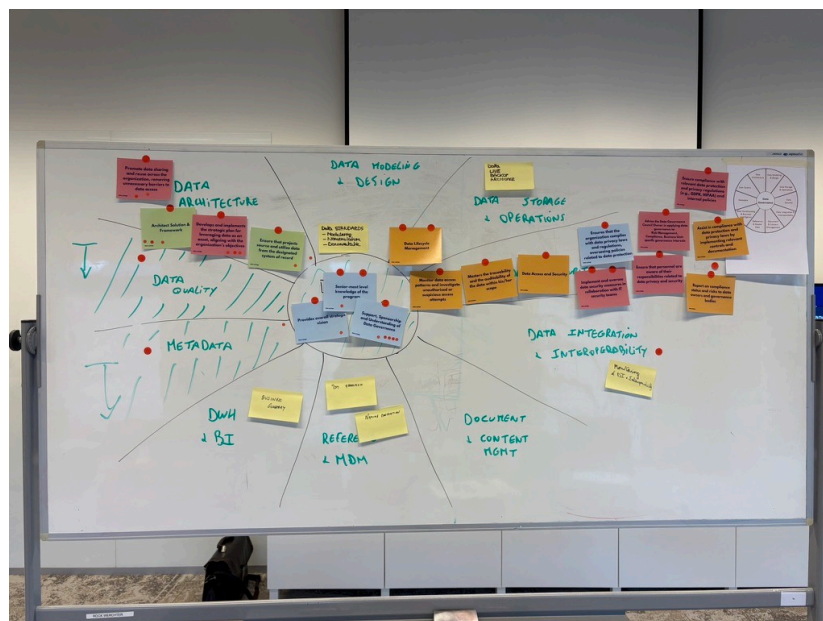
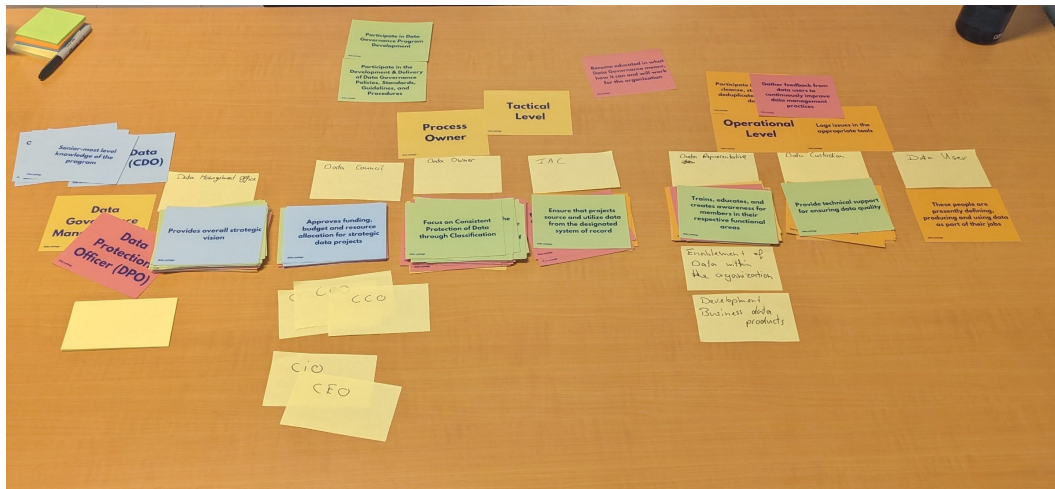
By mapping these responsibilities, we'll then naturally evolve into **defining the roles** that will carry them out. This approach ensures that our roles are rooted in the real needs of our organization, making our data governance framework both practical and impactful.

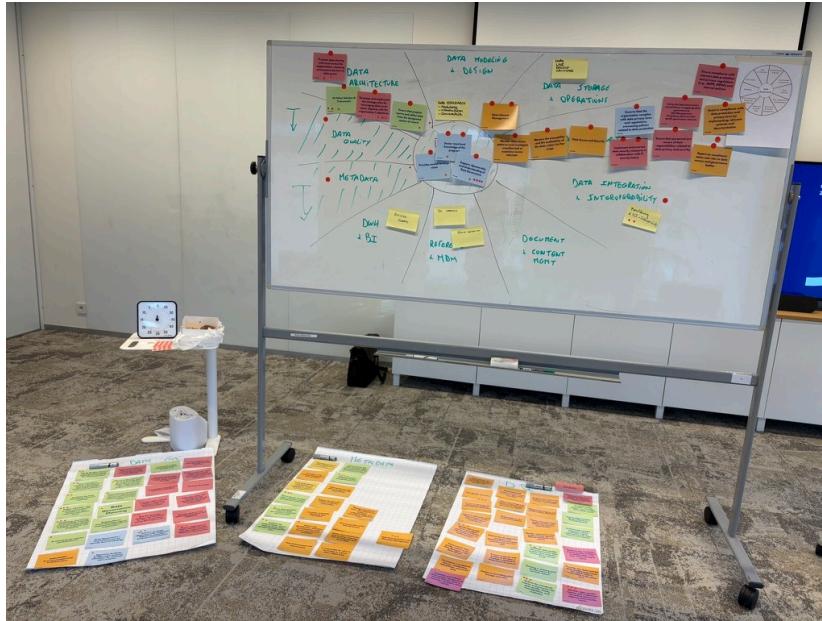
RESPONSIBILITY Mapping

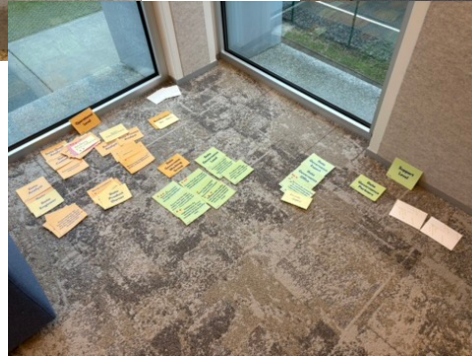
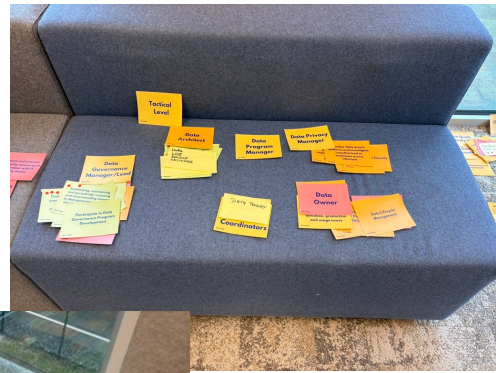
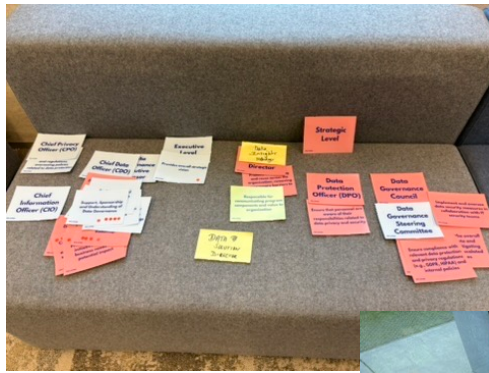
Workshop 3

2

Exercise 1







WEEK 2 of 5

The Data Players | Workshop 3 of 7

Tuesday



Exercise 1. Responsibility Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

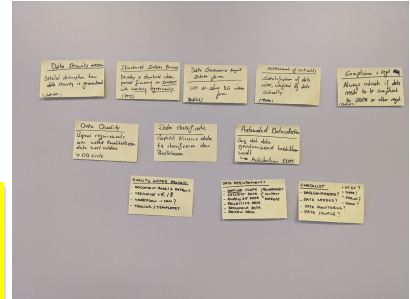
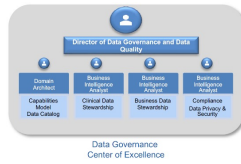
Exercise 4. Roles Roster

Now that we've identified and prioritized the key roles and responsibilities crucial to our data governance framework, it's time to explore how these elements will interact within an **effective operating model**.

This exercise is an inspiration-gathering exercise to stimulate creativity before concept creation and is aimed to **inspire creative and practical ideas** for structuring our data governance operating model.

This fast-paced, idea-generating session is designed to spark innovation by **showcasing different approaches** to data governance structures.

LIGHTNING Demos

HOSPITAL

Big Idea: A Data Governance Center of Excellence

LARGE BANK

Big Idea:
Federated operating model
with support structure

COLLIBRA

Big Idea:
Centralized operating
model with Data citizens

WEEK 2 of 5

The Data Players | Workshop 3 of 7

Exercise 1. Responsibility Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Exercise 4. Roles Roster

Next up, we'll take the insights gained from our Lightning Demos and the roles and responsibilities we've defined to design a set of **operating model concepts** tailored to our organization.

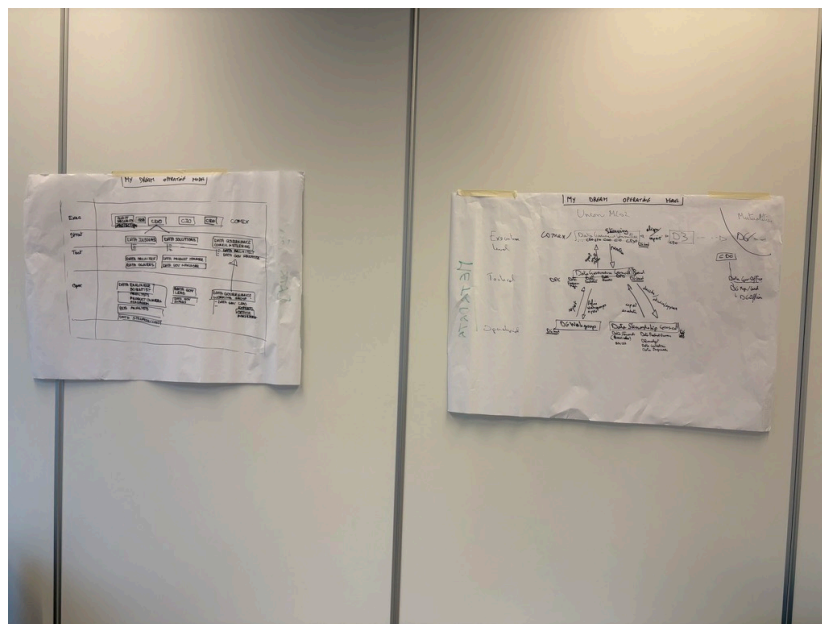
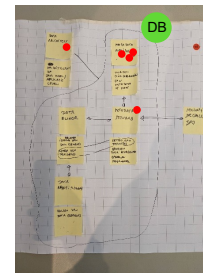
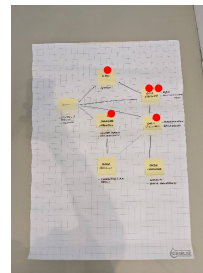
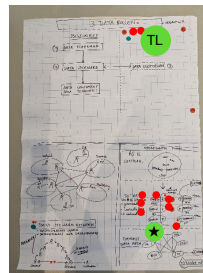
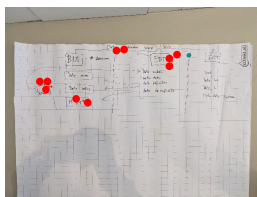
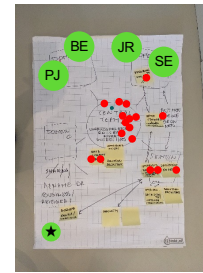
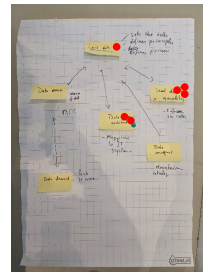
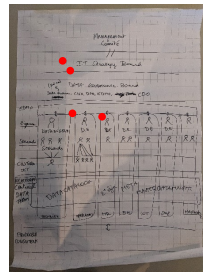
This is where creativity meets strategy—each participant will have the opportunity to envision how our data governance framework could be structured to optimize **communication flows, workflows, and organizational bodies**.

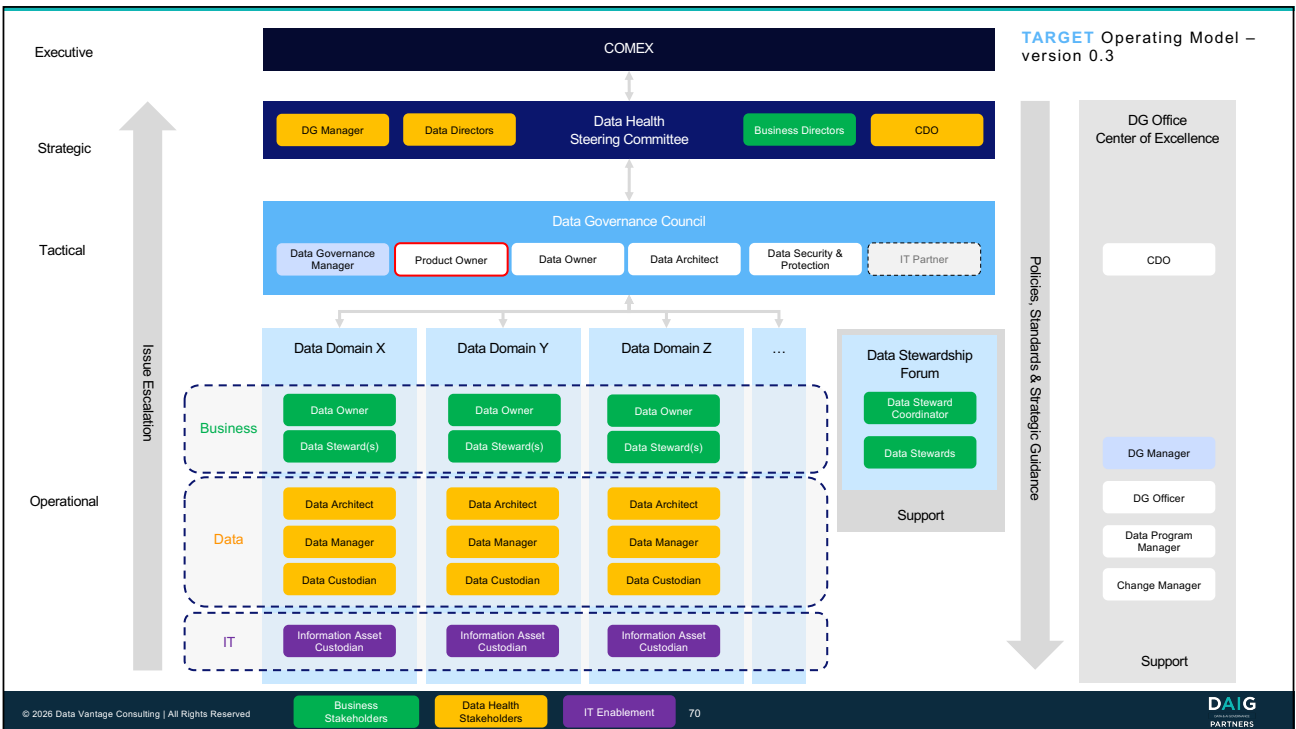
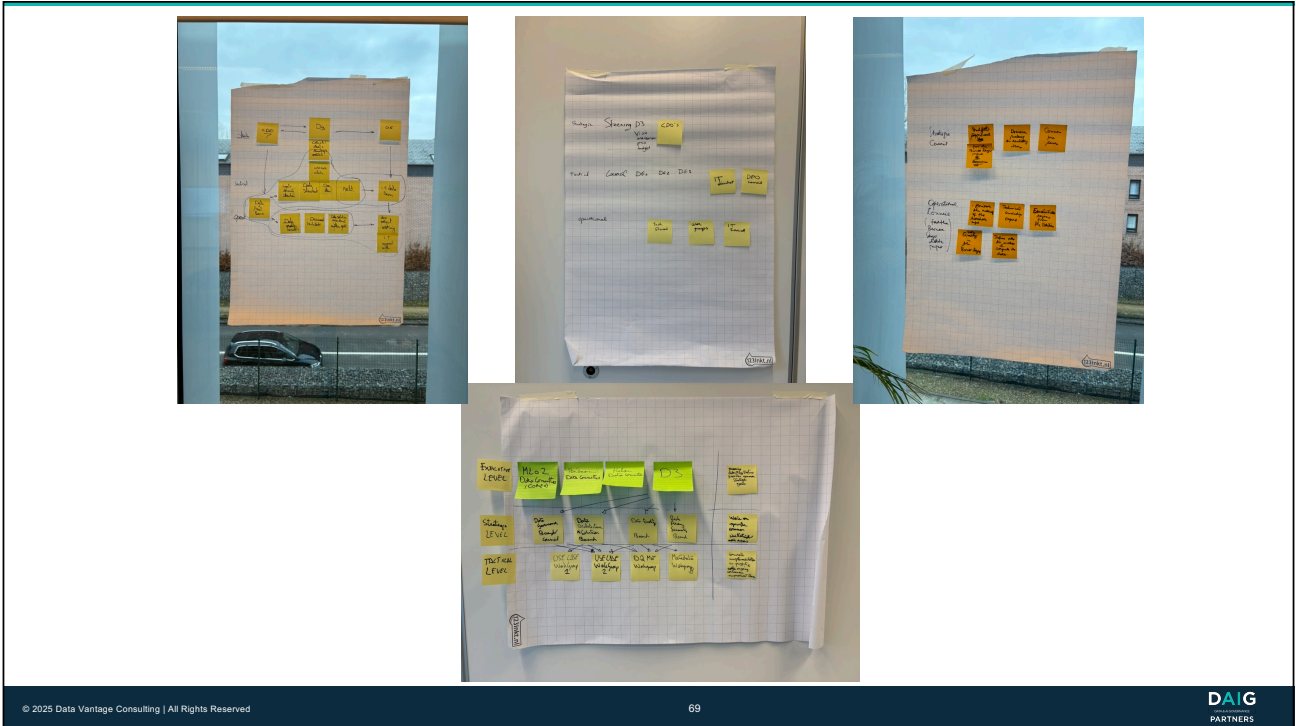
The goal is to create a robust, effective operating model that aligns with our data governance purpose and strategic goals. By the end of this exercise, we'll collaboratively refine these ideas into a **cohesive target operating model** that will guide our future efforts.

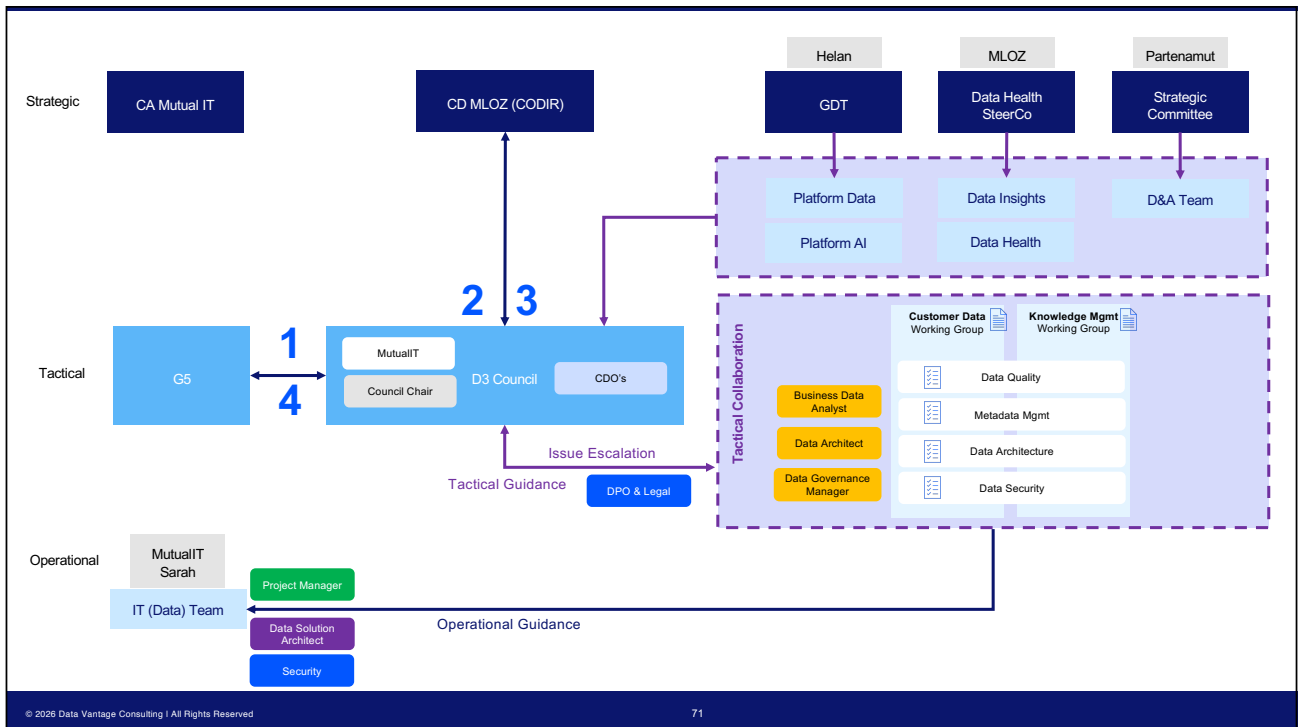
CONCEPT Creation

Workshop 3

Exercise 3







WEEK 2 of 5

The Data Players | Workshop 3 of 7

Tuesday

Exercise 1. Responsibility Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Exercise 4. Roles Roster

The goal of this final exercise – Roles Roster - is to identify the **bare minimum responsibilities** and roles necessary to effectively maintain data quality and document metadata for the selected CDEs.

By downsizing our target roles, responsibilities, and operating model, we will establish the most efficient and sustainable structure that can be **practically implemented** during our prototype and testing phase.

This simplified version will ensure that we are focusing our resources where they matter most, while still upholding the core principles of our data governance strategy.

Week 2

- 1

Monday
- 2

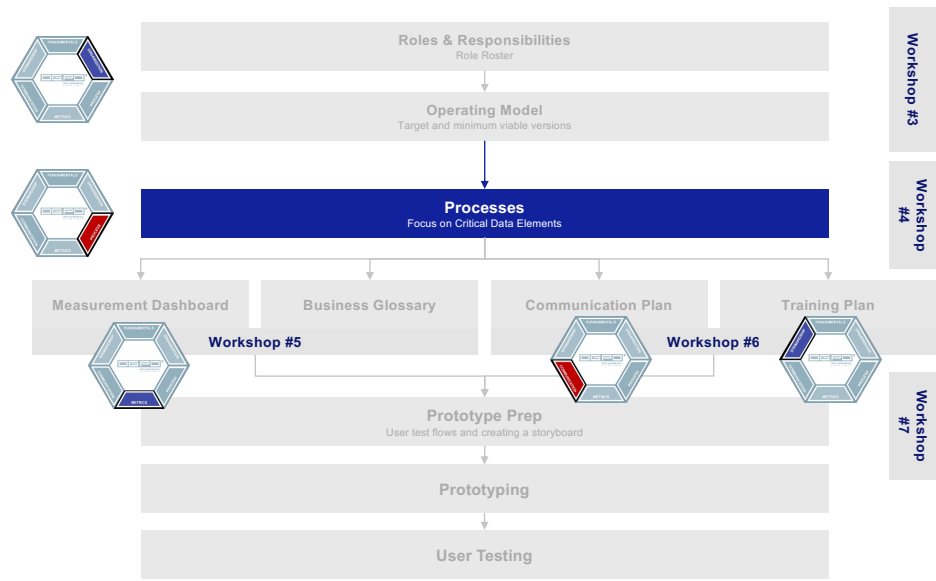
Tuesday
- 3

Wednesday
- 4

Thursday
- 5

Friday

Data Governance SPRINT™



WEEK 2 of 5

The Rules of the Game | Workshop 4 of 7

Round 1: Data Quality Focus

Exercise 1. Applied Users Story Mapping

Exercise 2. The Process Map

Exercise 3. Applied Users Story Mapping

Exercise 4. The Process Map

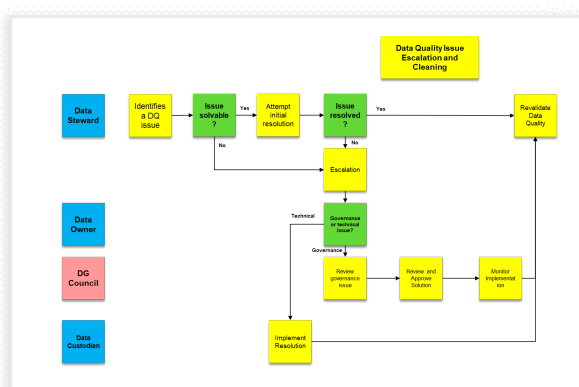
In this exercise, we will dive into the practical aspects of **maintaining and improving the quality** of our Critical Data Elements (CDEs).

By mapping out the story of how data quality is managed within our organization, we'll create a clear, **step-by-step process** that ensures our data remains accurate, reliable, and valuable.

This is where strategy meets action—each of you will bring your insights to the table, mapping out how data quality issues are identified, addressed, and resolved.

By the end of this exercise, we'll have a well-defined process that can be implemented to safeguard the integrity of our most critical data.

Applied Users STORY MAPPING

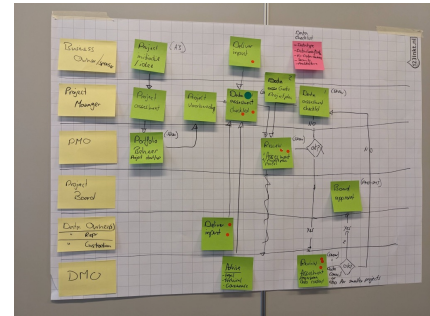
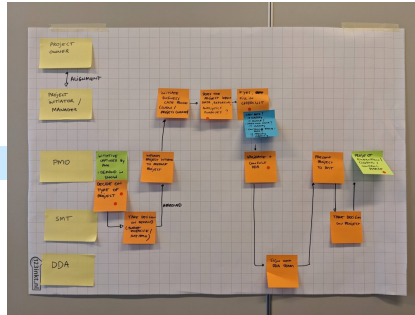


WEEK 2 of 5

Workshop 4 of 7 | Applied Users Story Mapping

Step 2: Create the Story

1. Timing
2. What you'll need
3. Introducing the exercise
4. Flow

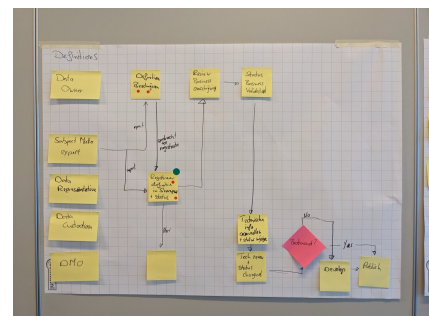
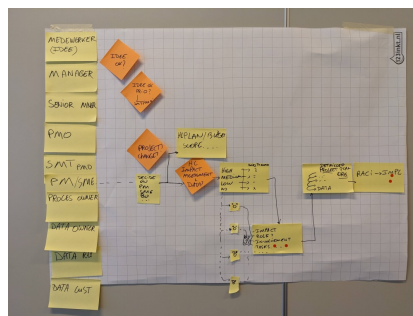


WEEK 2 of 5

Workshop 4 of 7 | Applied Users Story Mapping

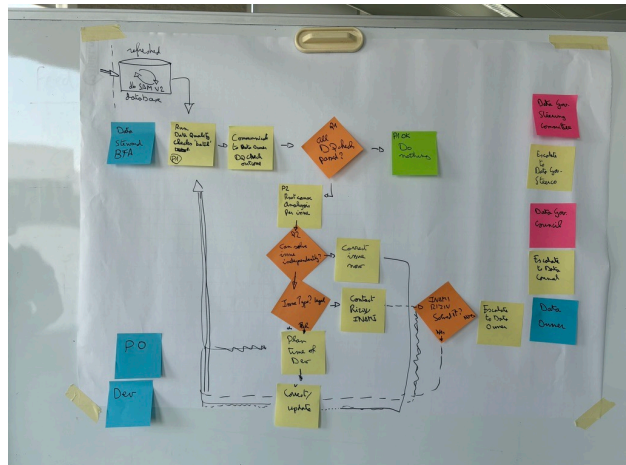
Step 2: Create the Story

1. Timing
2. What you'll need
3. Introducing the exercise
4. Flow



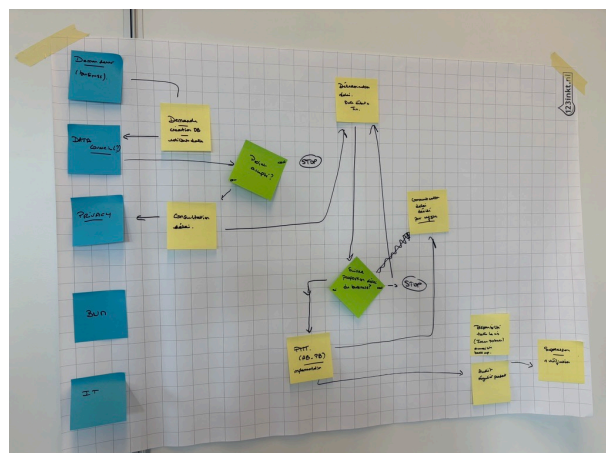
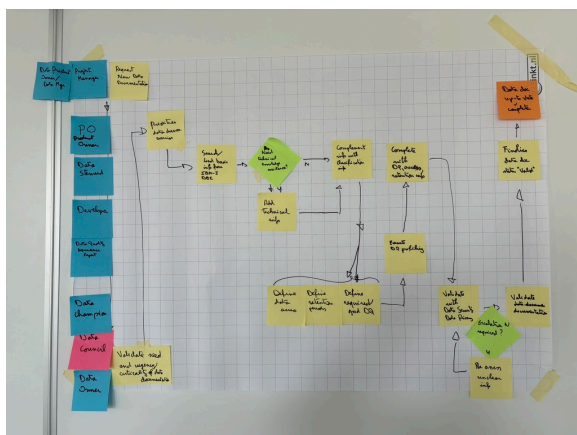
WEEK 2 of 5

Workshop 4 of 7 | Applied Users Story Mapping



WEEK 2 of 5

Workshop 4 of 7 | Applied Users Story Mapping



WEEK 2 of 5

The Rules of the Game | Workshop 4 of 7

Round 1: Data Quality Focus

Exercise 1. Applied Users Story Mapping

Exercise 2. The Process Map

Exercise 3. Applied Users Story Mapping

Exercise 4. The Process Map

After exploring individual concepts in the previous exercise, it's time to bring our ideas together and collaboratively create a **unified process map** for maintaining and improving the data quality of our Critical Data Elements (CDEs).

This exercise will **synthesize the best elements** from our individual efforts into a clear, practical process that can be implemented across the organization.

By working together, we'll ensure that the final process map is comprehensive, effective, and aligned with our data governance objectives.

WEEK 3 of 5

Designing Governance Essentials | Define & Decide

MON

Workshop 5 of 7
Keeping Score &
Language of the Game
ca. 6h,
with the entire team

TUE

Documentation
Summarizing the outcomes
and draft deliverables

Workshop set-up
Preparing for the next
workshop

WED

Workshop 6 of 7
Rooting for the Game &
Training Camp
ca. 6h,
with the entire team

THU

Workshop 7 of 7
Prototype Prep
ca. 4h,
with the entire team

Documentation
Summarizing the outcomes
and draft deliverables

FRI

Prototyping
Design a hi-fidelity
interactive prototype

User Recruiting
Find and schedule users

WEEK 3 of 5

Designing Governance Essentials | Define & Decide

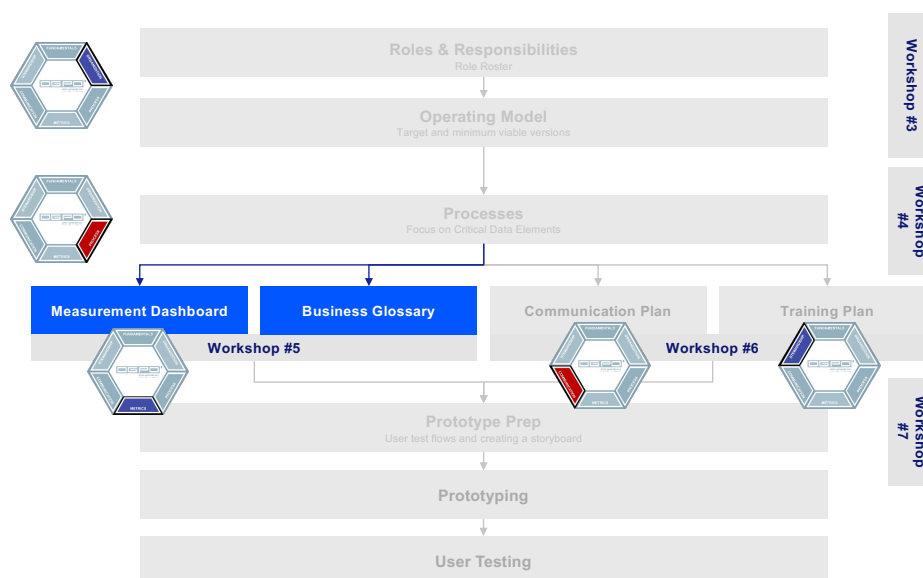
Welcome to Week 3: Designing Governance Essentials

This week marks a critical phase in our Data Governance Sprint™ as we shift our focus to creating and building the **tangible, essential elements** of our data governance program. Our goal is to develop practical tools that are ready for immediate use, starting with a simple yet effective **measurement dashboard** and a **business glossary** template. These tools will provide the foundation for consistent communication and tracking of our data governance efforts.

Recognizing that change management and communication are pivotal to the success of any data governance program, we'll also identify and plan the **communication and training activities** that will ensure all stakeholders are informed, engaged, and prepared to execute their roles.

As we progress, we'll prepare for the **prototyping phase**, where we'll bring together everything we've built and create a **minimum sustainable version** of our data governance program. This version will be put into practice and tested next week. To support this, we will also begin **recruiting our test audience**. The entire week is centered around defining the practical aspects of our program and making crucial decisions that will guide our next steps.

Data Governance SPRINT™



WEEK 3 of 5

Designing Governance Essentials | Workshop 5 of 7

Keeping Score Workshop: Creating a Measurement Dashboard

The first part of Monday's workshop focuses on creating a Measurement Dashboard that will serve as the backbone of our data governance **performance tracking**. We'll work together to identify the most **meaningful metrics** that align with our business goals and governance objectives.

By the end of this session, we'll have a clear, actionable dashboard template that includes both quantitative and qualitative metrics.

The Language of the Game Workshop: Setting up the Business Glossary

In the second part of the day, we shift our focus to establishing a Business Glossary. This glossary will be the cornerstone of **consistent and effective communication** across all data governance efforts.

The template we will design will help define critical terms and concepts, ensuring that everyone in the organization is on the same page.

WEEK 3 of 5

Keeping Score | Workshop 5 of 7

Round 1: Measurement Focus

Exercise 1. Metrics Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Exercise 4. Lightning Demos

Exercise 5. 10 for 10 Brainstorm

Exercise 6. Concept Creation

In this first exercise, we will identify and map out the **key data quality metrics and governance KPIs** that will form the foundation of our data governance dashboard.

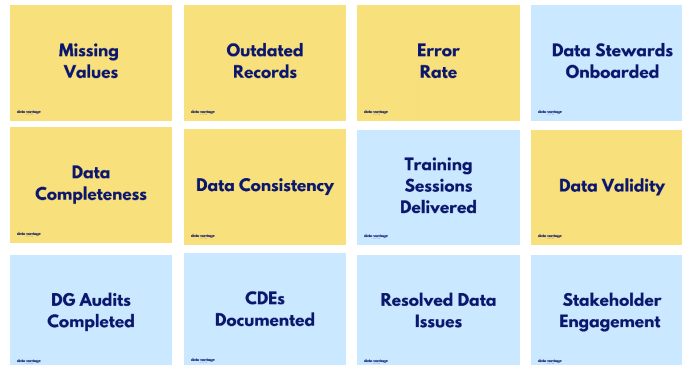
These metrics will allow us to measure not only the health and integrity of our Critical Data Elements (CDEs) but also the efficiency and effectiveness of our overall data governance program.

By selecting the right metrics, we ensure that our data governance initiatives are driven by meaningful, measurable goals. Through collaboration, we will narrow down the most important metrics, enabling us to track progress and drive impactful decision-making

METRICS Mapping

Workshop 5

Exercise 1



WEEK 3 of 5

Keeping Score | Workshop 5 of 7

Monday

Round 1: Measurement Focus

Exercise 1. Metrics Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Exercise 4. Lightning Demos

Exercise 5. 10 for 10 Brainstorm

Exercise 6. Concept Creation

In this exercise, we'll take a moment to step outside our own organization and **gather inspiration** from the world around us.

By looking at how other companies and industries visualize their data quality metrics and governance KPIs, we can uncover fresh ideas and discover innovative ways to design our own dashboard.

These "Lightning Demos" will help spark creativity, ensuring that we **create a dashboard** that is not only functional but also intuitive and engaging for stakeholders. Think of it as a quick dive into the best practices and visualizations that can elevate our data governance efforts.

LIGHTNING Demos



HEAT MAP

Big Idea:
A summary heatmap showing the progress of DG implementation

STEWARDSHIP DASHBOARD

Big Idea:
Steward dashboard
showing metadata metrics

SCORE CARD

Big Idea:
Colored scorecard to show progress

WEEK 3 of 5

Keeping Score | Workshop 5 of 7

Round 1: Measurement Focus

Exercise 1. Metrics Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Exercise 4. Lightning Demos

Exercise 5. 10 for 10 Brainstorm

Exercise 6. Concept Creation

Now that we've drawn inspiration from a range of dashboards through our Lightning Demos, it's time to shift gears and start **creating our own**.

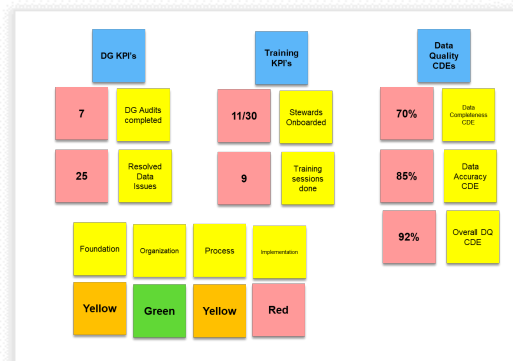
In this Concept Creation exercise, we'll design the foundation of our data metrics dashboard, turning ideas into a tangible, workable structure. This is where creativity meets practicality.

Each participant will **map out their vision** for a dashboard that effectively displays the key metrics we've prioritized. The goal is to create something simple, intuitive, and impactful—something that will help drive data governance forward by keeping the most essential metrics front and center.

CONCEPT Creation

Workshop 5

Exercise 3



WEEK 3 of 5

Workshop 5 of 7 | Lightning Demos

Workshop 5

Exercise 4

Exercise 1. Metrics Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Round 2: Business Glossary Focus

Exercise 4. Lightning Demos

Exercise 5. 10 for 10 Brainstorm

Exercise 6. Concept Creation

- **Reference to the previous Lightning Demos exercise** from Exercise 2 in the "Keeping Score" part of the workshop. This time, participants will focus on gathering examples of Business Glossaries and Data Dictionaries.
- Specifically, they should look at the **essential metadata elements** that other organizations use to document their data assets, such as descriptions, data types, definitions, and governance-related metadata.
- The goal is to **get inspired** by how others organize and structure their business glossaries, ensuring participants can bring back useful ideas for the next exercises.

WEEK 3 of 5

Workshop 5 of 7 | Lightning Demos

Workshop 5

Exercise 4

Exercise 1. Metrics Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

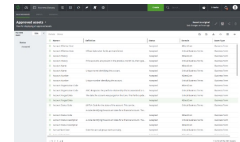
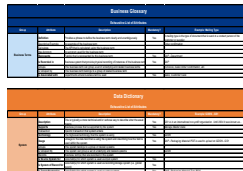
Round 2: Business Glossary Focus

Exercise 4. Lightning Demos

Exercise 5. 10 for 10 Brainstorm

Exercise 6. Concept Creation

- Show some real **examples** before the exercise starts.



EXCEL

Big Idea:
A simple data dictionary
and business glossary in
Excel

SHAREPOINT

Big Idea:
Data dictionary in
SharePoint, including
approval flow

COLLIBRA

Big Idea:
Fancy glossary in a high-
end tool

WEEK 3 of 5

The Language of the Game | Workshop 5 of 7

Monday

Exercise 1. Metrics Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Round 2: Business Glossary Focus

Exercise 4. Lightning Demos

Exercise 5. 10 for 10 Brainstorm

Exercise 6. Concept Creation

The "10 for 10 Brainstorm" is where the creative ideas flow freely, allowing the group to **generate as many ideas** as possible about what metadata elements we should include in the business glossary.

This exercise is critical because the success of the glossary depends on **capturing the right data** about each Critical Data Element (CDE) and business term. We want to think beyond just the typical fields and get creative, but also practical. What are the metadata elements that will bring the most clarity and value to the organization?

From descriptions and business rules to data quality guidelines and possible values, we will aim to identify every potential metadata element that could help ensure consistency and understanding across the organization.

WEEK 3 of 5

Workshop 5 of 7 | Concept Creation

Workshop 5

Exercise 6

Exercise 1. Metrics Mapping

Exercise 2. Lightning Demos

Exercise 3. Concept Creation

Round 2: Business Glossary Focus

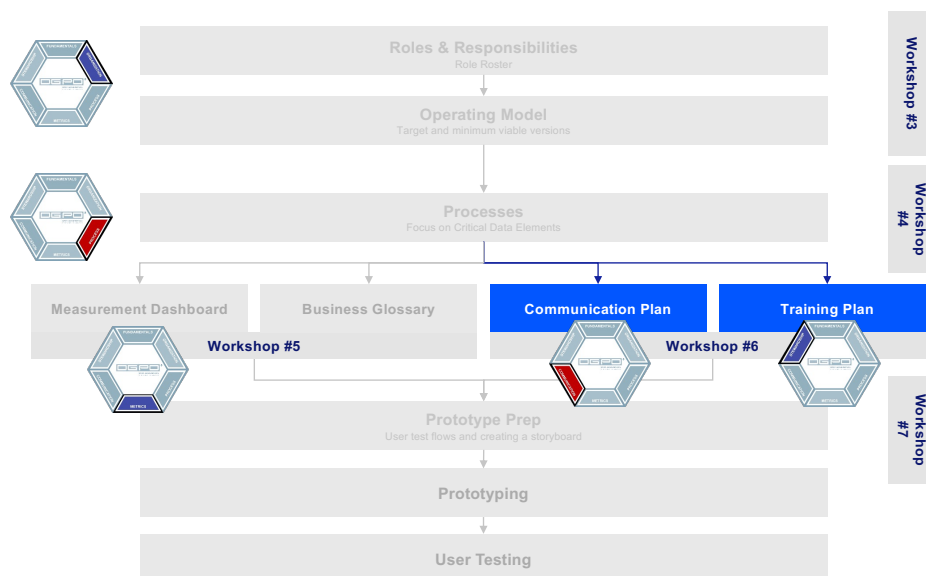
Exercise 4. Lightning Demos

Exercise 5. 10 for 10 Brainstorm

Exercise 6. Concept Creation

Data Element	Definition	Data Type	Owner	Source	Data Quality Rule
Customer ID	A unique identifier assigned to each customer in the system	Number	Sales Director	SAP MDG	Must be unique and not null. No duplicates allowed.
Customer Name	The full legal name of the customer, used for identification and correspondence	Text	Customer Service Manager	SAP MDG	Must be filled, no blank or incomplete names allowed.
Email address	The email address provided by the customer for communications and notifications	Text	Sales Director	Salesforce CRM	Must be a valid email format (e.g., name@domain.c). No duplicates allowed.

Data Governance SPRINT™



WEEK 3 of 5

Designing Governance Essentials | Workshop 6 of 7

Rooting for the Game Workshop: Defining Communication Plan

On Wednesday, we'll focus on creating a structured Communication Plan that ensures that data governance efforts are clearly communicated across the organization and that all stakeholders are properly informed.

Effective communication is critical for creating awareness, securing buy-in, and fostering ongoing engagement from stakeholders. We'll explore multiple strategies, prioritize key activities, and craft a detailed communication roadmap.

Training Camp Workshop: Defining Training Plan

In the second half of the day, we'll shift our attention to developing a Training Plan that equips all data governance roles with the necessary skills and knowledge.

We will brainstorm training activities and key topics, prioritize the most essential ones, and develop a roadmap to deliver targeted training initiatives that align with our governance milestones.

WEEK 3 of 5

Workshop 6 of 7 | Lightning Demos

Round 1 + 2: Same exercises

Exercise 1. Lightning Demos

Exercise 2. 10 for 10 Brainstorm

Exercise 3. Action Board

Exercise 4. Concept Creation

Exercise 5. Roadmap

- Show some **real examples** before the exercise starts.

➤ Morning Session:



WEBSITE

Big Idea:
Data Governance
SharePoint site

ARTICLES

Big Idea:
Quarterly Data Literacy
articles

INFOGRAPHIC

Big Idea:
Engaging infographic
showcasing the DG
Program

WEEK 3 of 5

Workshop 6 of 7 | Lightning Demos

Workshop 6

Exercise 1

Round 1 + 2: Same exercises

Exercise 1. Lightning Demos

Exercise 2. 10 for 10 Brainstorm

Exercise 3. Action Board

Exercise 4. Concept Creation

Exercise 5. Roadmap

- Show some **real examples** before the exercise starts.

➤ Afternoon Session:



Data Governance Explained in 5 Minutes



SKILL: ENTERPRISE ARCHITECTURE	
Level	Description
1	Define the purpose, goals, and intended outcomes of the enterprise architecture.
2	Identify the key stakeholders and their roles in the development of the enterprise architecture.
3	Develop the enterprise architecture framework and standards.
4	Develop the enterprise architecture framework and standards.
5	Develop the enterprise architecture framework and standards.

VIDEOS

Big Idea:
Short, targeted videos that explain data governance

QUICK REFERENCE CARDS

Big Idea:
Printable or digital cheat sheets

SKILLS MATRIX

Big Idea:
Outlining the required competencies and skills for each role

WEEK 3 of 5

Workshop 6 of 7 | 10 for 10 Brainstorm

Workshop 6

Exercise 2

Round 1 + 2: Same exercises

Exercise 1. Lightning Demos

Exercise 2. 10 for 10 Brainstorm

Exercise 3. Action Board

Exercise 4. Concept Creation

Exercise 5. Roadmap

- Reference to the previous 10 for 10 Brainstorm exercise** from Exercise 6 in the "Keeping Score" part of Monday workshop, but this time with a specific focus.

➤ **Morning Session:** focus on generating ideas for communication events, activities, and artifacts that will support the rollout and ongoing promotion of the data governance program.

➤ **Afternoon Session:** focus on brainstorming training methods, events, activities, and artifacts that will empower and educate the key data governance roles, such as data stewards, data owners, and data custodians.

- The goal is to **generate a wide range of creative ideas** that can be further refined and prioritized for the development of effective communication and training strategies.

WEEK 3 of 5

Workshop 6 of 7 | 10 for 10 Brainstorm

Workshop 6

Exercise 3

Round 1 + 2: Same exercises

Exercise 1. Lightning Demos

Exercise 2. 10 for 10 Brainstorm

Exercise 3. Action Board

Exercise 4. Concept Creation

Exercise 5. Roadmap



WEEK 3 of 5

Workshop 6 of 7 | 10 for 10 Brainstorm

Workshop 6

Exercise 4

Round 1 + 2: Same exercises

Exercise 1. Lightning Demos

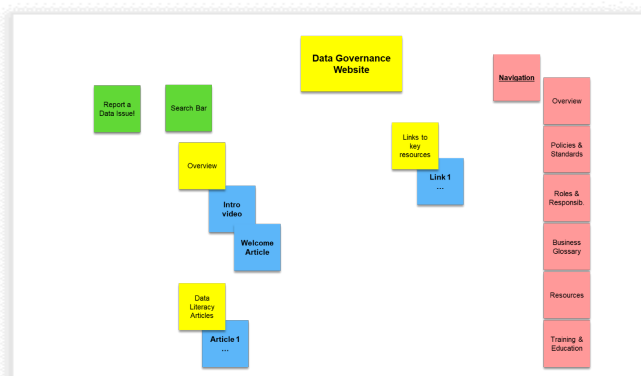
Exercise 2. 10 for 10 Brainstorm

Exercise 3. Action Board

Exercise 4. Concept Creation

Exercise 5. Roadmap

- An example of a **communication concept**:



WEEK 3 of 5

Workshop 6 of 7 | 10 for 10 Brainstorm

Workshop 6

Exercise 4

Round 1 + 2: Same exercises

Exercise 1. Lightning Demos

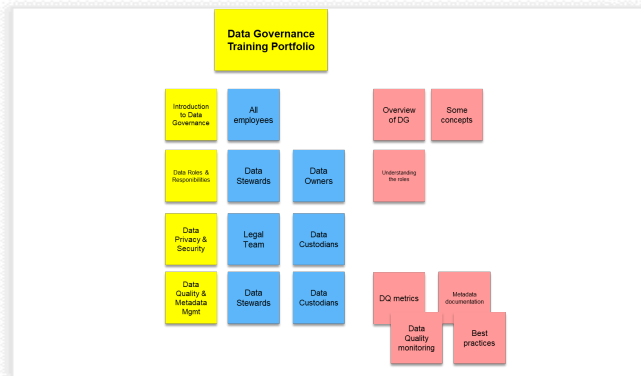
Exercise 2. 10 for 10 Brainstorm

Exercise 3. Action Board

Exercise 4. Concept Creation

Exercise 5. Roadmap

- An example of a **training concept**:



WEEK 3 of 5

Workshop 6 of 7 | 10 for 10 Brainstorm

Workshop 6

Exercise 5

Round 1 + 2: Same exercises

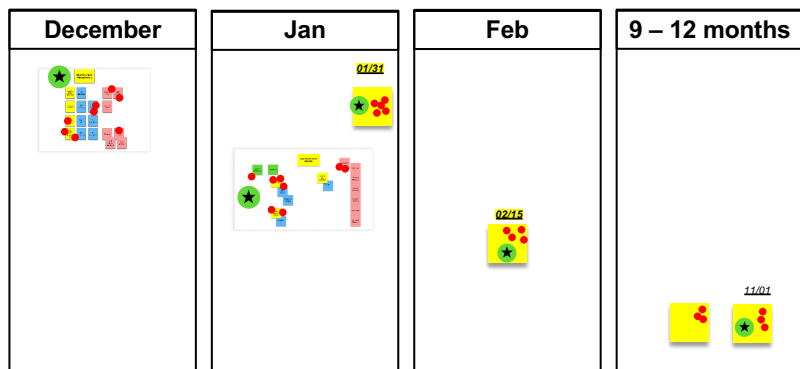
Exercise 1. Lightning Demos

Exercise 2. 10 for 10 Brainstorm

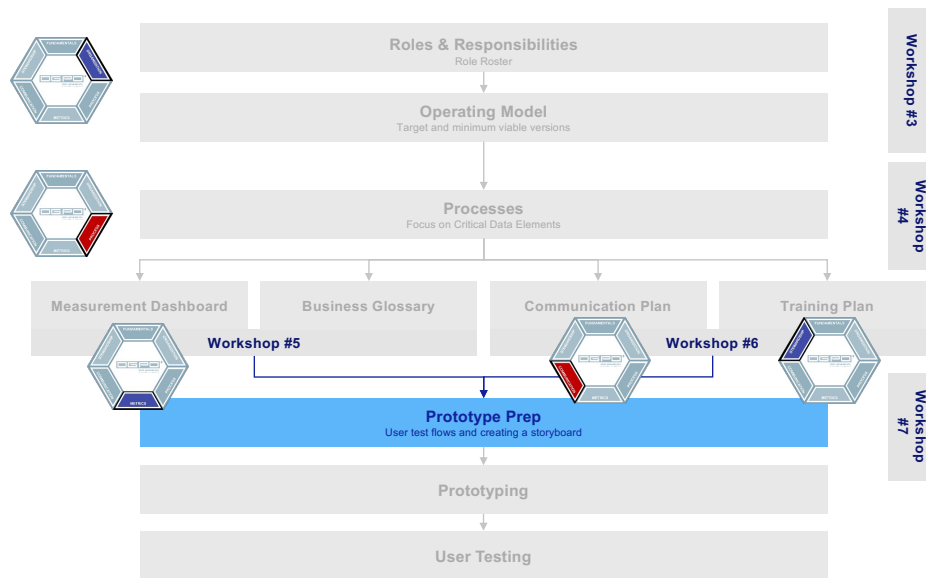
Exercise 3. Action Board

Exercise 4. Concept Creation

Exercise 5. Roadmap



Data Governance SPRINT™



WEEK 3 of 5

Prototype Preparation | Workshop 7 of 7

Thursday



Prototype Prep Workshop: Bringing Everything Together

In this final workshop, we bring everything together and begin **preparing the prototypes** that will be used during the testing phase. The goal is to take the elements we've developed so far—such as the process flows, business glossary, measurements dashboard, training materials, and communication plans—and **downsize them into small, actionable prototypes** focused on the top-voted and easily accessible Critical Data Elements (CDEs).

We aim to create a **minimum sustainable version** of key processes like the CDE data quality and CDE metadata documentation process, ensuring they are pragmatic and easy to implement on a small scale. This workshop will zoom in on the **1 to 3 key responsibilities** of essential data roles (data steward, data custodian, and data owner) to test how these roles function in real-world scenarios.

Through two targeted exercises—Users Test Flow and Storyboard—we will establish a clear sequence for prototyping and testing, making the implementation phase smoother and more actionable.

Workshop 7: Thursday Half-Day

- Preparing the sequence of steps for each prototype, ensuring we have a clear roadmap for testing once we've settled on the concepts and ideas.
- By using a Storyboard, we create a step-by-step guide that outlines the implementation of each prototype in a clear and concise manner, ensuring all aspects of the chosen solution are covered and ready for testing.

WEEK 3 of 5

Prototype Preparation | Workshop 7 of 7

Exercise 1. Users Test Flow

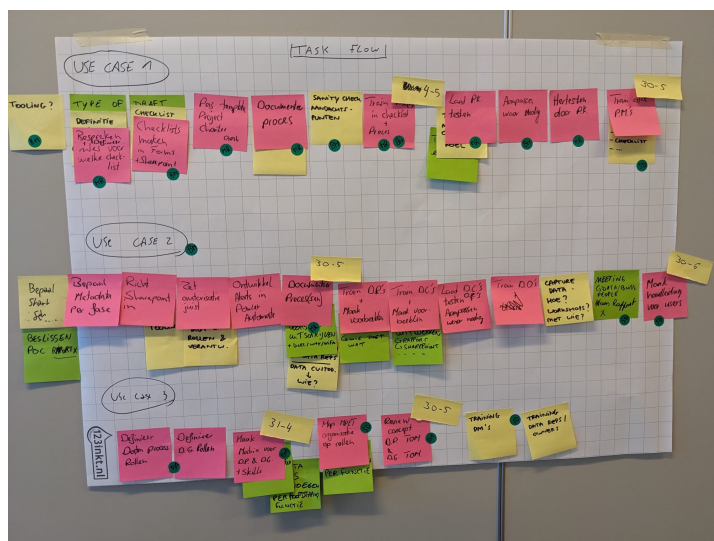
Exercise 2. Storyboard

In this exercise, we'll focus on creating detailed **user flows for the key data governance roles**—data steward, data owner, and data custodian—specifically in the context of managing data quality and metadata. The goal is to map out how these roles will interact with the processes we've defined, ensuring that the workflows are practical and efficient for our upcoming prototype testing phase.

We'll explore how each of these **roles will engage with the Critical Data Elements (CDEs)** and manage their responsibilities. This will help us identify pain points, optimize the flow, and ensure that each role's actions align with the overall governance objectives. The end result will be a clear map of how these roles operate in practice, giving us a solid foundation for testing and validation in the next phase.

WEEK 3 of 5

Workshop 7 of 7 | Users Test Flow



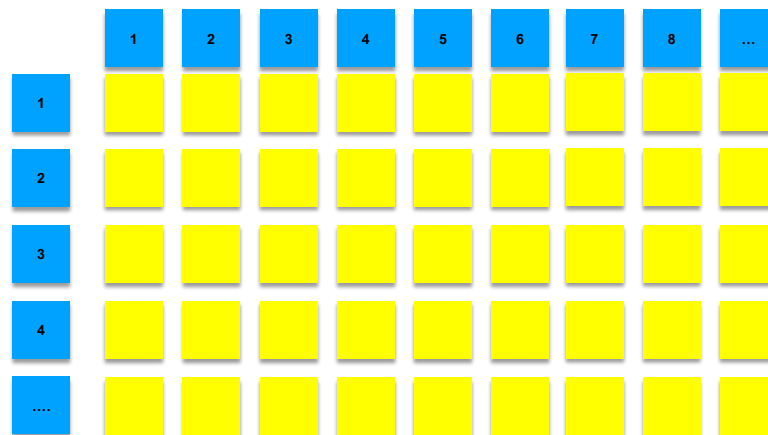
USERS Test Flow

Workshop 7

3

Exercise 1

Guiding axes



WEEK 3 of 5

Prototype Preparation | Workshop 7 of 7

Thursday

4

Exercise 1. Users Test Flow

Exercise 2. Storyboard

In this exercise, we will take the **best user story flow** from the previous exercise and transform it into a **detailed, step-by-step guide**, much like a **storyboard for a movie**.

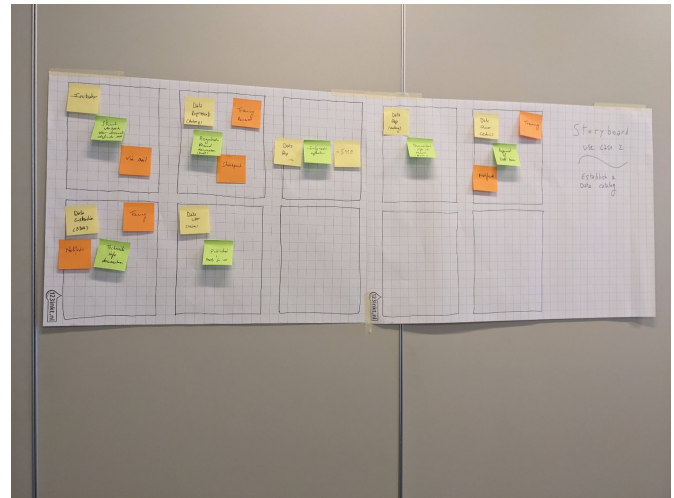
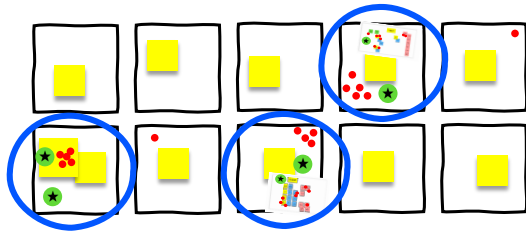
Just as a movie storyboard maps out each scene in sequence, our storyboard will map out each step of the data governance process, focusing on the actions of the **data steward**, **data owner**, and **data custodian**.

The goal is to create a clear, actionable plan that outlines every step necessary for executing the chosen process, including the roles, responsibilities, activities, and key decision points. This storyboard will serve as the blueprint for moving from concept to reality, from prototype to testing.

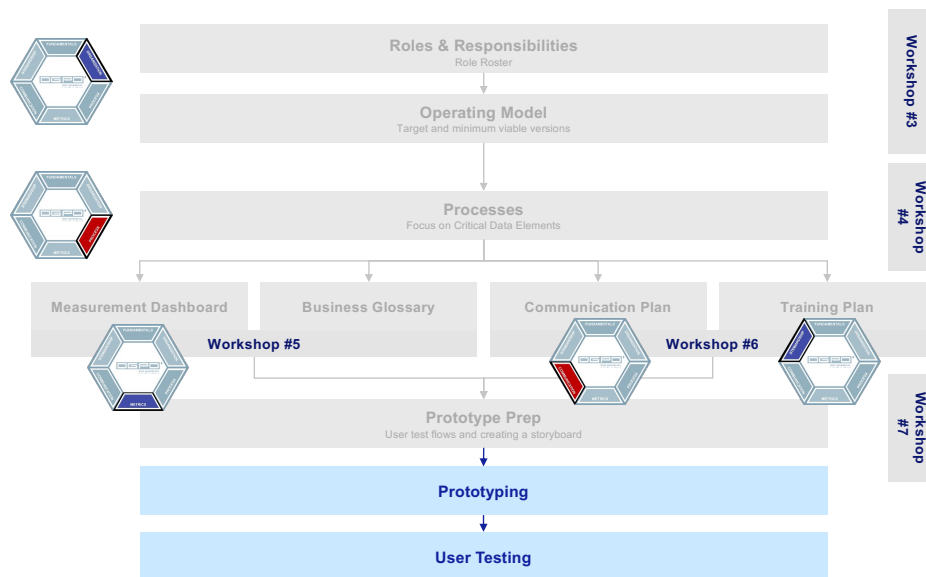
STORYBOARD

Workshop 7

Exercise 2

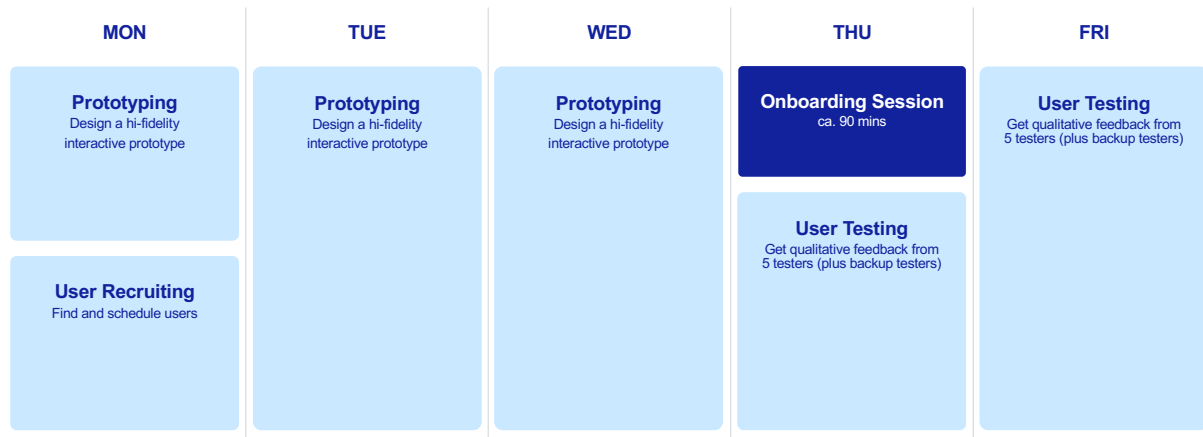


Data Governance SPRINT™



WEEK 4 of 5

Design Sprint | Design & Validate



WEEK 4 of 5

Design Sprint | Design & Validate

Week 4



Welcome to Week 4: From Concept to Reality

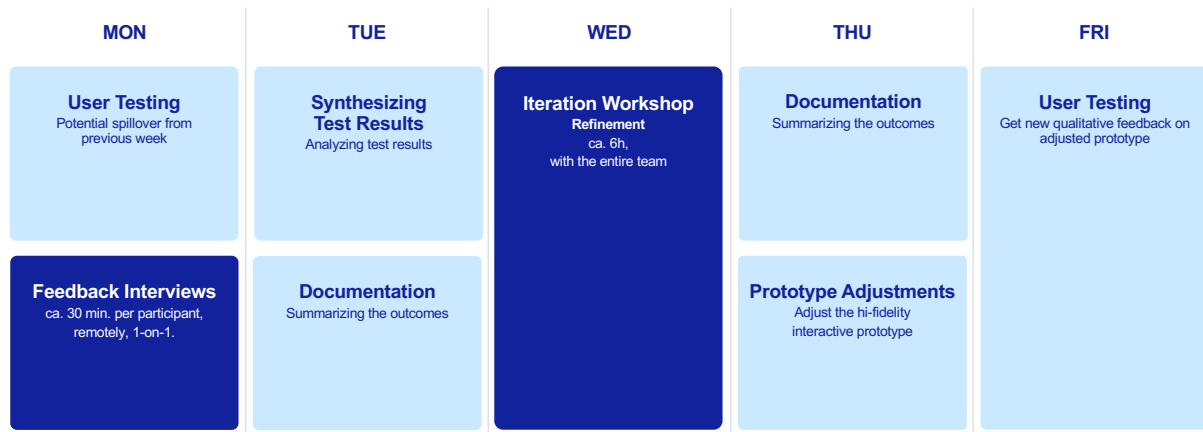
In Week 4, we transition from planning and concept development into the hands-on phase of **prototyping and validation**. This is the time to bring everything we've built over the past few weeks into tangible forms that can be tested.

We'll start by **refining and creating prototypes** of our core data governance tools—such as the Data Governance Website, the business glossary template, and process documentation—ensuring they are ready for **user interaction**.

By the end of Week 4, our goal is to have not only tested the prototypes but also gathered **initial insights** on how they can be improved. The results of these tests will be critical as we move into the final stages of refining and implementing the data governance program.

WEEK 5 of 5

Design Sprint | Refine & Implement



WEEK 5 of 5

Design Sprint | Refine & Implement

Week 5



Welcome to Week 5: The Final Stretch

In Week 5, we enter the critical phase of **refining and implementing** our data governance solutions, putting the finishing touches on everything we've built and tested so far.

This is the culmination of our Data Governance Sprint™—where we turn insights from user feedback into **actionable adjustments**, ensuring our prototypes are not just functional but optimized for **real-world use**.

This week is all about refinement, ensuring that we leave with a strong, actionable version of our prototypes that can be **scaled confidently** and with momentum. It's the final push towards making the Data Governance Sprint a success, ensuring that the groundwork is set for long-term adoption and growth.

WEEK 5 of 5

Design Sprint | Iteration Workshop

Iteration Workshop: Refining for Success

Welcome to the Iteration Workshop, where we take everything we've learned from our user testing and **make it better**. Today is all about getting hands-on with the feedback we've gathered and transforming those insights into improvements for our data governance prototypes.

It's our chance to spot what's working well and **tackle the pain points** head-on. We'll kick off by reviewing what the users told us, and then we'll dive into exercises that let us brainstorm, create, and refine. By the end of the day, we aim to have **sharper, smarter prototypes** that are ready for the next stage. This is the workshop where everything comes together, so let's roll up our sleeves and build something great!

Iteration Workshop: Wednesday Full-Day

- Refined prototypes based on user feedback, ensuring improved tools for data governance.
- A clear understanding of what's working, what's not, and actionable insights to guide improvements.
- Strategic priorities and focus areas identified, setting the direction for the next steps.
- Finalized concepts chosen through team discussions and voting, ready for further development.

WEEK 5 of 5

Design Sprint | Iteration Workshop

Exercise 1. Review Feedback

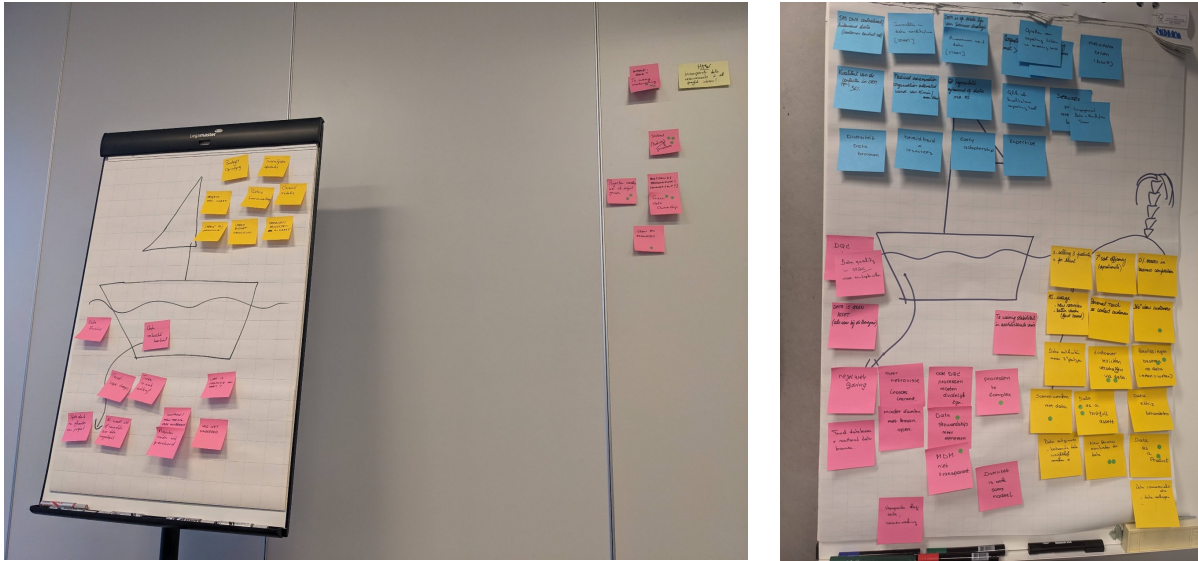
Exercise 2. Sailboat

Exercise 3. Strategic Questions

Exercise 4. Lightning Demos

Exercise 5. Concept Creation

WORKSHOP ACTIVITY: The sailboat exercise



Images © Data Vantage Consulting. Used with permission.

WEEK 6+

LiftOff | Embedding & Scaling Data Governance

Week 6+



Next Steps After the Data Governance Sprint™

As we wrap up the Data Governance Sprint™, it's important to reflect on the solid foundation we've built in just five weeks.

Throughout this DG Sprint™, we have **designed**, **tested**, and **refined** core elements of a data governance framework that can be scaled and embedded within the organization.

The journey, however, doesn't end here—**this is just the beginning**. The next phase is **LiftOff**, where we transition from prototypes to fully operational systems, further embedding governance practices and scaling them across the organization.

WEEK 6+

LiftOff | Embedding & Scaling Data Governance

Week 6+

6

Next Steps After the Data Governance Sprint™

As we move forward with LiftOff, the focus shifts to making data governance a core part of the organization's culture. The prototypes will evolve into **embedded practices and capabilities** that not only improve data quality but also enhance decision-making, transparency, and compliance across the organization.

By continuously implementing the use cases on the roadmap, maturing our prototypes into fully integrated solutions, and scaling the framework using agile methodologies, we will ensure that data governance drives real business value and creates a **sustainable data-driven culture** where you manage your data as an asset.

Moving fast without sacrificing **QUALITY**

1. Adaptability to change

- Agile methods allow data governance initiatives to quickly adapt to evolving business needs, regulatory requirements, and market changes. This flexibility ensures that governance efforts remain relevant and effective without slowing down progress.

2. The power of prototyping

- Prototyping allows organizations to test data governance practices on a small scale, gathering feedback and making adjustments before broader implementation. This minimizes risks and ensures high-quality, practical solutions that can be scaled effectively.

3. Incremental improvements

- Agile emphasizes iterative development, enabling organizations to achieve quick wins by implementing manageable improvements. These incremental steps enhance data quality over time while focusing on high-impact areas.

4. Enhanced collaboration across teams

- Agile fosters cross-functional collaboration, uniting IT, business, and compliance teams. This collaborative approach improves alignment, ensures governance practices meet diverse needs, and promotes collective accountability for data quality.

5

Must See Case Studies

Bringing your data governance framework to life with **PROTOTYPING**



Images © Data Vantage Consulting. Used with permission.

CASE STUDY: Success with small teams and budgets!



- **Team size:** 2 people (1 DG lead, 1 support)
- **Budget:** Stand-up completed within a \$24K budget

Objective

To establish a structured data governance framework tailored to this insurance company, enhancing data quality, supporting compliance, and aligning with the organization's strategic goals.

Key Takeaways

- **Quick wins:** Addressed immediate issues, like inconsistent customer data, using real-world applications of data governance.
- **Strategic alignment:** Aligned data governance with company's key business goals.
- **Scalable foundation:** Established a framework ready for future expansion of data governance.

Approach & Key Deliverables

- 1. Business glossary and data dictionary**
 - Created a central repository (Sharepoint) of key business terms and definitions to ensure consistent understanding across departments.
- 2. Data governance charter and operating model**
 - Developed a charter outlining the goals, roles, and responsibilities within the data governance framework, aligning with the company's corporate strategy.
- 3. Data quality policy and measurement scorecard**
 - Introduced a policy to improve data accuracy and created a scorecard (Excel) to track data quality metrics, ensuring ongoing monitoring and accountability.
- 4. Communication and training plan**
 - Established a plan to raise awareness, train staff, and foster a data-centric culture across the organization.



Area	Current Score	Target Score
Business Glossary	8.5 (20/20)	9.0 (20/20)
Data Governance Charter	8.0 (20/20)	8.5 (20/20)
Data Quality Policy	7.5 (20/20)	8.0 (20/20)
Data Quality Scorecard	7.0 (20/20)	7.5 (20/20)
Communication & Training Plan	6.5 (20/20)	7.0 (20/20)
Overall Score	7.5 (100/100)	8.0 (100/100)

Area	Current Score	Target Score
Business Glossary	8.5 (20/20)	9.0 (20/20)
Data Governance Charter	8.0 (20/20)	8.5 (20/20)
Data Quality Policy	7.5 (20/20)	8.0 (20/20)
Data Quality Scorecard	7.0 (20/20)	7.5 (20/20)
Communication & Training Plan	6.5 (20/20)	7.0 (20/20)
Overall Score	7.5 (100/100)	8.0 (100/100)

CASE STUDY: Data quality initiative for financial institution



Objective

Establish a data quality framework to improve the accuracy of critical Key Performance Indicators (KPIs) in executive reports.

Key Takeaways

- **Enhanced consistency:** The business glossary standardized terminology, reducing misinterpretations and improving interdepartmental communication.
- **Improved decision-making:** Reliable KPIs supported better executive decisions and compliance with regulatory requirements.
- **Proactive data monitoring:** The SAS-based dashboard allowed early detection of data issues, fostering continuous quality improvement.

Approach & Key Deliverables

- 1. Business glossary creation**
 - Established clear definitions and standards for critical data elements, using SAS Business Data Network (BDN), ensuring consistent understanding and use across departments.
- 2. Data quality rules setup**
 - Defined data quality rules for each KPI, aligning them with business objectives to monitor and maintain data integrity.
- 3. Data quality dashboard implementation**
 - Utilized SAS Visual Analytics (VA) to create an initial data quality dashboard, providing real-time insights for proactive data management.

Area	Current Score	Target Score
Business Glossary	8.5 (20/20)	9.0 (20/20)
Data Governance Charter	8.0 (20/20)	8.5 (20/20)
Data Quality Policy	7.5 (20/20)	8.0 (20/20)
Data Quality Scorecard	7.0 (20/20)	7.5 (20/20)
Communication & Training Plan	6.5 (20/20)	7.0 (20/20)
Overall Score	7.5 (100/100)	8.0 (100/100)



Area	Current Score	Target Score
Business Glossary	8.5 (20/20)	9.0 (20/20)
Data Governance Charter	8.0 (20/20)	8.5 (20/20)
Data Quality Policy	7.5 (20/20)	8.0 (20/20)
Data Quality Scorecard	7.0 (20/20)	7.5 (20/20)
Communication & Training Plan	6.5 (20/20)	7.0 (20/20)
Overall Score	7.5 (100/100)	8.0 (100/100)

Area	Current Score	Target Score
Business Glossary	8.5 (20/20)	9.0 (20/20)
Data Governance Charter	8.0 (20/20)	8.5 (20/20)
Data Quality Policy	7.5 (20/20)	8.0 (20/20)
Data Quality Scorecard	7.0 (20/20)	7.5 (20/20)
Communication & Training Plan	6.5 (20/20)	7.0 (20/20)
Overall Score	7.5 (100/100)	8.0 (100/100)

Area	Current Score	Target Score
Business Glossary	8.5 (20/20)	9.0 (20/20)
Data Governance Charter	8.0 (20/20)	8.5 (20/20)
Data Quality Policy	7.5 (20/20)	8.0 (20/20)
Data Quality Scorecard	7.0 (20/20)	7.5 (20/20)
Communication & Training Plan	6.5 (20/20)	7.0 (20/20)
Overall Score	7.5 (100/100)	8.0 (100/100)

CASE STUDY: Foundational activities at public transportation company



Objective

To establish a foundational data governance framework to support a larger data quality initiative.

Key Takeaways

- **Identified key pain points:** Conducted Challenge Map workshops, surfacing nearly 600 data-related issues to target.
- **Strategic alignment:** Defined the purpose of data governance and established guiding principles through a strategy workshop.
- **Building blocks for future growth:** Set the groundwork to enable scalable and sustainable data quality improvements.

Approach & Key Deliverables

1. Challenge Map workshops

- Conducted 7 workshops across departments to identify data pain points, resulting in a comprehensive list of 585 issues.
- Provided a clear view of the organization's current data challenges, creating a prioritized action list.

2. Prioritized issue list and use cases

- Enabled the organization to prioritize critical data governance initiatives, focusing on use cases that would deliver quick wins and long-term value.

3. Data governance strategy workshop

- Facilitated a strategy session to define the overarching purpose of data governance and establish guiding principles.
- Created a unifying vision for data governance that aligns with organizational goals and sets a clear direction.

Issue ID	Issue Description	Priority	Status
1	Data quality issues in customer database	High	In Progress
2	Inconsistent data formats across systems	Medium	Not Started
3	Lack of data ownership and accountability	Low	Planned

Issue ID	Issue Description	Priority	Status
1	Data quality issues in customer database	High	In Progress
2	Inconsistent data formats across systems	Medium	Not Started
3	Lack of data ownership and accountability	Low	Planned

Issue ID	Issue Description	Priority	Status
1	Data quality issues in customer database	High	In Progress
2	Inconsistent data formats across systems	Medium	Not Started
3	Lack of data ownership and accountability	Low	Planned

Issue ID	Issue Description	Priority	Status
1	Data quality issues in customer database	High	In Progress
2	Inconsistent data formats across systems	Medium	Not Started
3	Lack of data ownership and accountability	Low	Planned

Issue ID	Issue Description	Priority	Status
1	Data quality issues in customer database	High	In Progress
2	Inconsistent data formats across systems	Medium	Not Started
3	Lack of data ownership and accountability	Low	Planned

CASE STUDY: Enhancing data literacy at a large container terminal



Objective

Boost data literacy and raise awareness of data governance principles across the organization through creative engagement.

Key Takeaways

- **Increased engagement:** The cartoon and storytelling approach made data governance fun and engaging, boosting interest and retention.
- **Improved data literacy:** Simplified concepts in the comic helped employees understand data governance basics, enhancing data literacy.
- **Foundation for culture change:** Container Chris is now a recognizable symbol, strengthening the company's commitment to data governance and supporting a data-driven culture.

Approach & Key Deliverables

1. Creation of "Container Chris" character

- Developed a relatable cartoon figure, "Container Chris," who personifies data governance concepts, making them more accessible and engaging for employees.

2. Comic story development

- Designed an easy-to-understand comic story featuring Container Chris. The storyline explains the importance of data governance and covers essential topics, such as data quality, security, and compliance.

3. Lunch & Learn sessions

- Hosted interactive "Lunch and Learn" sessions to introduce Container Chris to employees and dive deeper into the core data governance principles, fostering open discussions and increasing engagement.



What success looks like: **KEY DELIVERABLES**

Foundational Deliverables

- **Business glossary:** A shared vocabulary for key terms and data definitions.
- **Data governance charter:** A formal document outlining the purpose, principles, and framework.
- **Roles & responsibilities matrix:** Clearly defined accountability for Data Owners, Stewards, and Custodians.

Communication Deliverables

- **Awareness materials:** Posters, website, newsletters, or infographics to promote data governance.
- **Training materials:** Structured guides, videos, and presentations to educate teams.
- **Stakeholder engagement plan:** Strategies for engaging and maintaining stakeholder buy-in.

Process Deliverables

- **Data quality dashboard:** Visual representation of data quality metrics and trends.
- **Critical data elements (CDEs) documentation:** Defined and prioritized CDEs with data quality rules.
- **Workflow and process maps:** Clear diagrams of key governance processes like issue resolution or metadata management.

Success Tracking Deliverables

- **Metrics scorecard:** A summary of data governance KPIs to monitor progress.
- **Roadmap:** A phased plan of initiatives, milestones, and timelines.

Key **SUCCESS** Factors! ...but also **PITFALLS**

- **Start with the problem and assess the current state**
 - ❑ Data Governance must be problem-focused—address specific business problems that need solving. But don't boil the ocean!
 - ❑ Evaluate your existing data landscape to understand the current state before initiating efforts.
- **Focus on critical data and deliver measurable benefits**
 - ❑ Start with governing Critical Data Elements (CDEs) to create the greatest impact.
 - ❑ Ensure the focus is on improving data quality, accessibility, security, privacy, and retention.
- **Deliver early wins and demonstrate value**
 - ❑ Deliver measurable benefits early on to build momentum and maintain stakeholder support.
 - ❑ Data Governance must demonstrate real and ongoing value to the organization, so link it to the business strategy.
- **Secure commitment and communicate effectively**
 - ❑ Gain ongoing support both from the top (leadership) and the bottom (operational level).
 - ❑ Keep stakeholders informed and engaged—communicate the value of Data Governance consistently.
- **Prioritize process over tools**
 - ❑ Focus on establishing strong processes for Data Governance rather than relying on technology alone.
- **Define control measurements and evolve into business as usual**
 - ❑ Set clear metrics to track progress and success.
 - ❑ Treat Data Governance as a continuous process that evolves into business as usual, not a one-off initiative.



Questions? Feel free to contact me

mathias@daigpartners.com
+32 468 258 947
<https://daigpartners.com>

© 2026 Data & AI Governance Partners
Reproduction without permission is prohibited.

Beyond Hive: The Open Table Format Revolution

Erasmus Data Collaboratory | House of AI

THE LEGACY BOTTLENECK (The "Data Swamp")

Rigidity and Unsafe Writes

Legacy Hive tables lack ACID compliance, meaning failed writes can lead to silent data corruption.



The "Rewrite" Nightmare

Simple partition changes (e.g., Month to Day) require expensive, full table rewrites in legacy systems.



Performance Latency

Slow directory-based scans create massive overhead as data volumes grow into the petabyte range.



THE MODERN SOLUTION (The "Lakehouse Vision")

Scale Meets Governance

Lakehouses combine low-cost cloud storage with the ACID transactions and governance of a database.



The Metadata "Secret Sauce"

Formats like Iceberg and Delta Lake use a metadata layer to enable features like Time Travel.



Multi-Engine Freedom

Open formats allow the same data to be accessed by Spark, Trino, Flink, and DuckDB simultaneously.

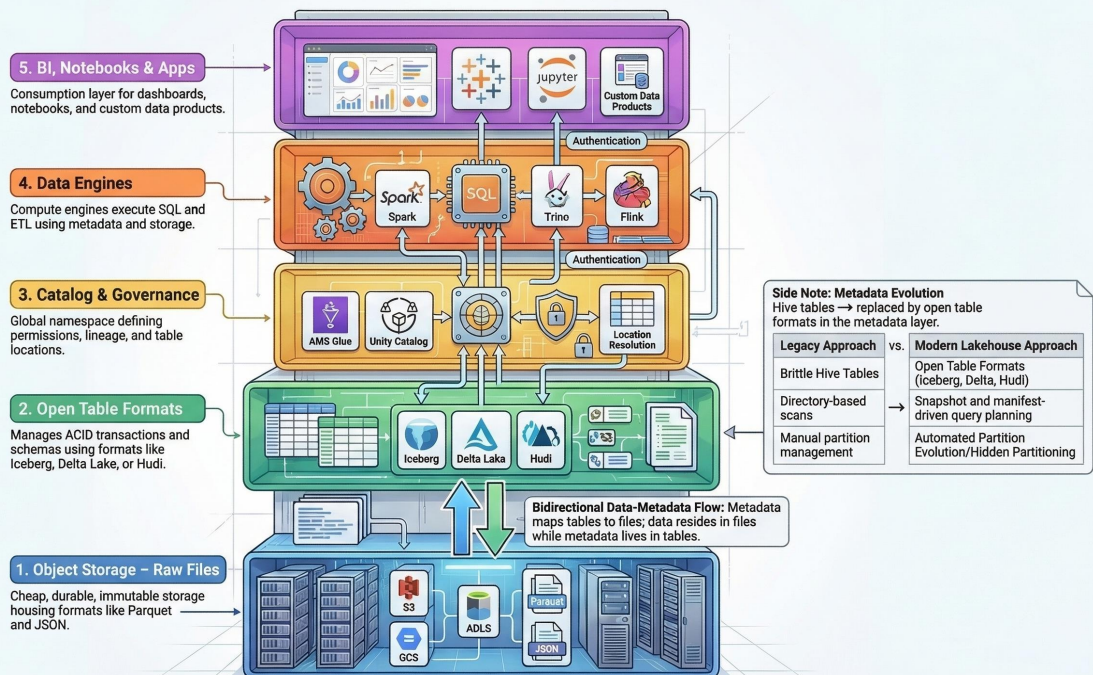


Jos van Dongen

Director Erasmus Data Collaboratory | House of AI

Metadata-Enabled, High-Performance Data Architecture with Open Table Formats. | Reliable. Scalable. Open.

Modern Data Lakehouse Architecture (Open Table Formats)



Lakehouse at Scale: Industry Giants & Open Format Metrics

Enterprises are shifting away from legacy Hive-based data lakes toward open table formats like Apache Iceberg and Delta Lake. This transition enables warehouse-like reliability (ACID transactions, schema evolution) at a petabyte scale while drastically reducing costs and latency.



Apache Iceberg Implementations

Apple: Billions of events per day

Billions

Manages petabyte-scale tables across all divisions for real-time streaming and ETL workloads.



Pinterest: 24-hour latency cut to 15 minutes

24h to 15m

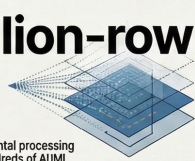
Consolidated event logging into Iceberg, saving \$1.5MM in annual infrastructure costs.



Netflix: Trillion-row data scale

Trillion-row

Uses incremental processing to power hundreds of AUML applications and global content insights.



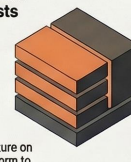
Comparing high-volume event processing across Iceberg adopters

Company	Daily Metric	Key Outcome
Adobe	32 Billion Events	Eliminated data corruption via schema enforcement
Skyscanner	30 Billion Events	Simplified analytical infrastructure for AI governance

Delta Lake & Data Intelligence

Block: 12x reduction in computing costs

12x



Standardized infrastructure on a Data Intelligence Platform to accelerate GenAI onboarding.

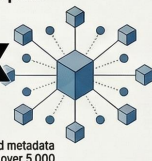
Itaú: 70% faster data availability

70%

Processes 650,000+ daily events with transactional data reaching the mesh in under 10 minutes.

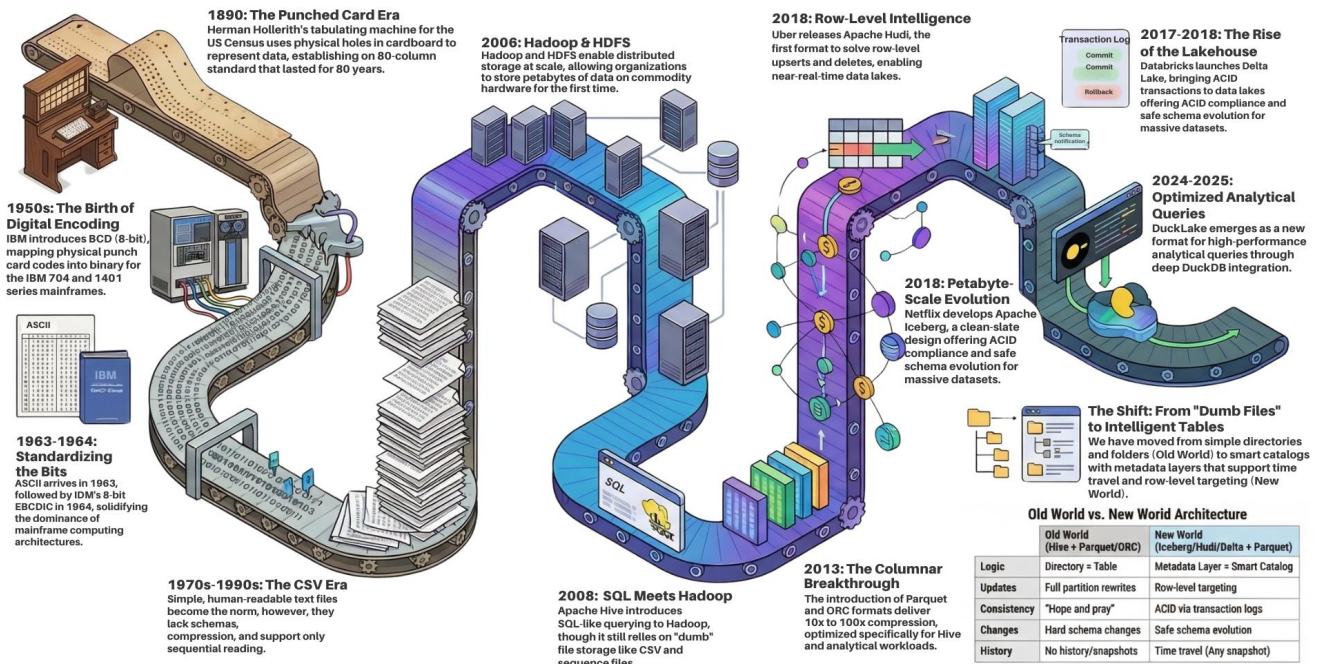
Unilever: 10x increase in development speed

10x



Utilizes a centralized metadata framework to serve over 5,000 internal users.

The Great Data Conveyor: 134 Years from Punched Cards to Intelligent Tables



The 'Why' and 'How' of Open Table Formats: From Raw Files to Reliable Lakehouses

THE 'WHY'—FIXING THE BRITTLE DATA LAKE

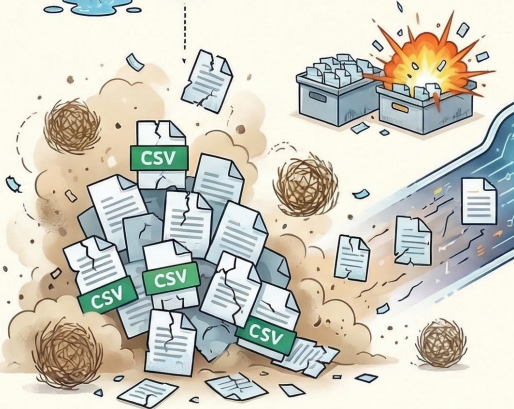
The "Wild West" of Raw Files

Raw files lack transactions, causing corrupted data when multiple jobs write simultaneously.



High Maintenance & "Small-File Explosions"

Changing schemas or deleting single rows requires rewriting entire partitions, causing performance bottlenecks.



THE BRAINS OF A DATABASE

Table formats turn piles of files into reliable tables using metadata and protocols.

THE 'HOW'—CRITICAL LAKEHOUSE CAPABILITIES

ACID Transactions & Schema Evolution

Ensures concurrent write reliability and allows renaming fields without breaking downstream queries.



Time Travel & Efficient Querying

Enables auditing historical data versions and uses metadata to prune unnecessary file scans.



Multi-Engine Interoperability

Allows the same table to be accessed by Spark, Flink, Trino, and DuckDB.



Table Format Comparison: At a Glance

Format	Primary Optimization	Key Strength
Iceberg	Large-scale Analytics	Engine-agnostic and wide industry adoption.
Delta Lake	Batch & Streaming	Deep integration with Spark and Databricks.
Hudi	Upserts & CDC	Built for near-real-time ingestion and incremental processing.
Paimon	Streaming-first	LSM-tree design for high-velocity IoT and CDC updates.
DuckLake	Metadata Simplicity	Stores metadata in a SQL database for fast commits.

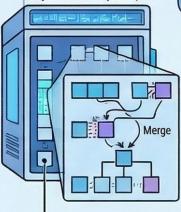
The Three Waves of Open Table Format Innovation: From Hive to Modern Lakehouses

Wave 1 – The Incremental Pioneers (2016–2018)

Fixing Hive with Transactionality

Addressed Hive's inability to handle updates/deletes and rewrites for small changes.

Apache Hudi (2016)



Delta Lake (2017)



Philosophy of "Fixing Hive":

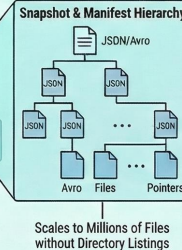
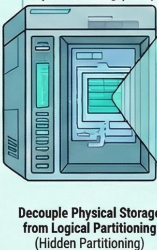
Add transactionality to existing batch-centric infrastructures.

Wave 2 – The Universal Standard (2018–2022)

Engine Neutrality & Metadata Scalability

Solved Wave 1 engine lock-in and metadata bottlenecks at petabyte scale.

Apache Iceberg (2018)



Philosophy of "Outlast Any Engine":

Vendor-neutral standard for AWS, Google, Snowflake, Databricks.

Wave 3 – Streaming Disruptors & Simplicity (2022–Present)

Rethinking Core Assumptions

Addresses Wave 2's latency struggles and high metadata complexity for smaller teams.

Technical Spotlight

Apache Paimon: The Streaming Disruptor



DuckLake: The Simplicity Advocate



Philosophy of Radical Efficiency:

Prioritizes sub-minute freshness and ease of use over batch-centricity.

Positioning & Decision Guide

Wave	Format	Best For	Key Advantage
1	Hudi	CDC, AWS Stacks, Audit Trails	Proven incremental pipelines
1	Delta Lake	Databricks/Spark Ecosystems	Deep Spark integration
2	Iceberg	Interoperability, Large Batch	Engine-agnostic, safe bet
3	Paimon	Sub-minute freshness, Flink	LSM-tree for high-velocity streaming
3	DuckLake	Greenfield, Dev Velocity	Radical simplicity via SQL metadata

2016–2018 (Wave 1)

2018–2022 (Wave 2)

2022–Present (Wave 3)

Navigating the Lakehouse: A Guide to Open Table Formats

THE FORMAT COMPARISON MATRIX

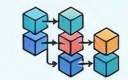
Metadata Architecture Differences

Formats range from file-based manifests to sequential logs and relational databases.

Row-Level Strategy: COW vs. MOR

Formats use Copy-on-Write (COW) or Merge-on-Read (MOR) to balance write speed against query performance.

Apache Iceberg Delta Lake Apache Hudi Apache Paimon DuckLake



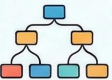
Snapshot/Manifests



Transaction Log



Commit Timeline



LSM-Tree Design



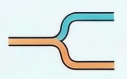
SQL Database



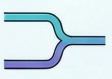
COW + Delete Files



COW + Deletion Vectors



COW or MOR



Native MOR



COW (Fast Commits)

Engine-Agnostic
(Trino, Snowflake)

Spark & Databricks

AWS AWS & CDC Ingestion

Flink & Streaming

DuckDB & MotherDuck

STRATEGIC SELECTION GUIDE

Choose Iceberg for Universal Neutrality

Best for organizations requiring the widest vendor support and cross-engine interoperability.



Choose Delta Lake for Databricks Centricity

Ideal for teams heavily invested in Spark and Databricks' optimized performance features.



Choose Paimon or Hudi for Real-Time Needs

Best for high-velocity streaming, CDC ingestion, and sub-minute data freshness.



ANATOMY OF THE UNIVERSAL STANDARD: INSIDE APACHE ICEBERG

CORE IDENTITY & ORIGIN

HEADLINE: Born at Netflix (2018)
SUPPORTING DETAIL: Created to fix Hive's scalability issues and donated to the Apache Software Foundation in 2019.

HEADLINE: Engine-Agnostic Philosophy
SUPPORTING DETAIL: Designed with an architecture intended to outlast any single compute platform or query engine.

HEADLINE: Petabyte-Scale Maturity
SUPPORTING DETAIL: Production proven across hundreds of major organizations including Apple, Adobe, and LinkedIn.

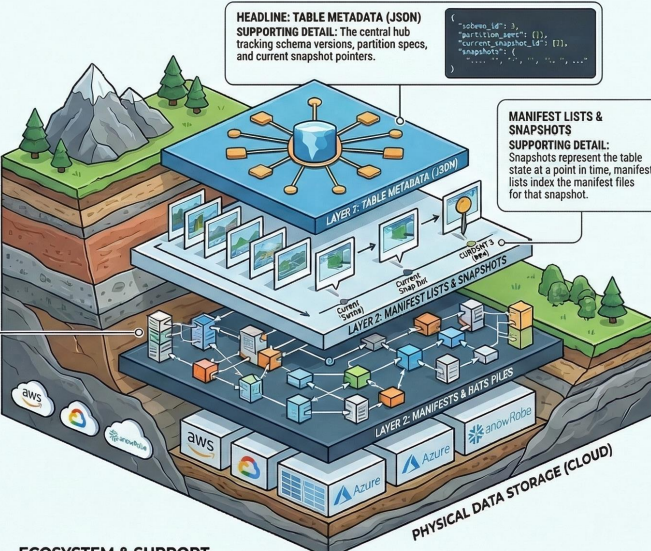
HEADLINE: Manifests & Data Files
SUPPORTING DETAIL: Manifests track individual data files (Parquet/ORC) and provide column-level stats for efficient pruning.

TRADE-OFFS & OPERATIONAL REALITIES

HEADLINE: Batch-Optimised Design
SUPPORTING DETAIL: Not ideal for sub-minute streaming latency requirements where formats like Paimon may excel.

HEADLINE: Operational Maintenance
SUPPORTING DETAIL: Requires active orchestration for snapshot expiration and "orphan file" cleanup to manage storage costs.

HEADLINE: Metadatas Complexity
SUPPORTING DETAIL: The three layer hierarchy offers high performance but requires tuning and deeper understanding for troubleshooting.



ECOSYSTEM & SUPPORT

Spark
HEADLINE: Broadest Engine Support
SUPPORTING DETAIL: Native, first class integration with Spark, Flink, Trino, Presto, Athena, Snowflake, and BigQuery.

Zero Vendor Lock-in
HEADLINE: Zero Vendor Lock-in
SUPPORTING DETAIL: Adopted as a native format by AWS, Google Cloud, and Snowflake, ensuring long-term data portability.

Governance & Tooling
HEADLINE: Governance & Tooling
SUPPORTING DETAIL: Supported by robust catalog services like Apache Polaris, Nessie, and AWS Glue.

EVOLUTIONARY STRENGTHS

HEADLINE: Comprehensive Schema Evolution
SUPPORTING DETAIL: Add, drop, rename, or reorder columns using unique IDs—never breaking downstream queries.

HEADLINE: Hidden Partitioning
SUPPORTING DETAIL: Decouples physical layout from logical auctev, users don't need to know the partition column to query efficiently.

HEADLINE: True Partition Evolution
SUPPORTING DETAIL: Change partitioning schemes (e.g., Daily to Hourly) without rewriting existing historical data.

HEADLINE: Point-in-Time "Time Travel"
SUPPORTING DETAIL: Snapshot-based queries allow for historical analysis, auditing, and easy data rollbacks.

DECISION MATRIX: WHEN (AND WHEN NOT) TO CHOOSE

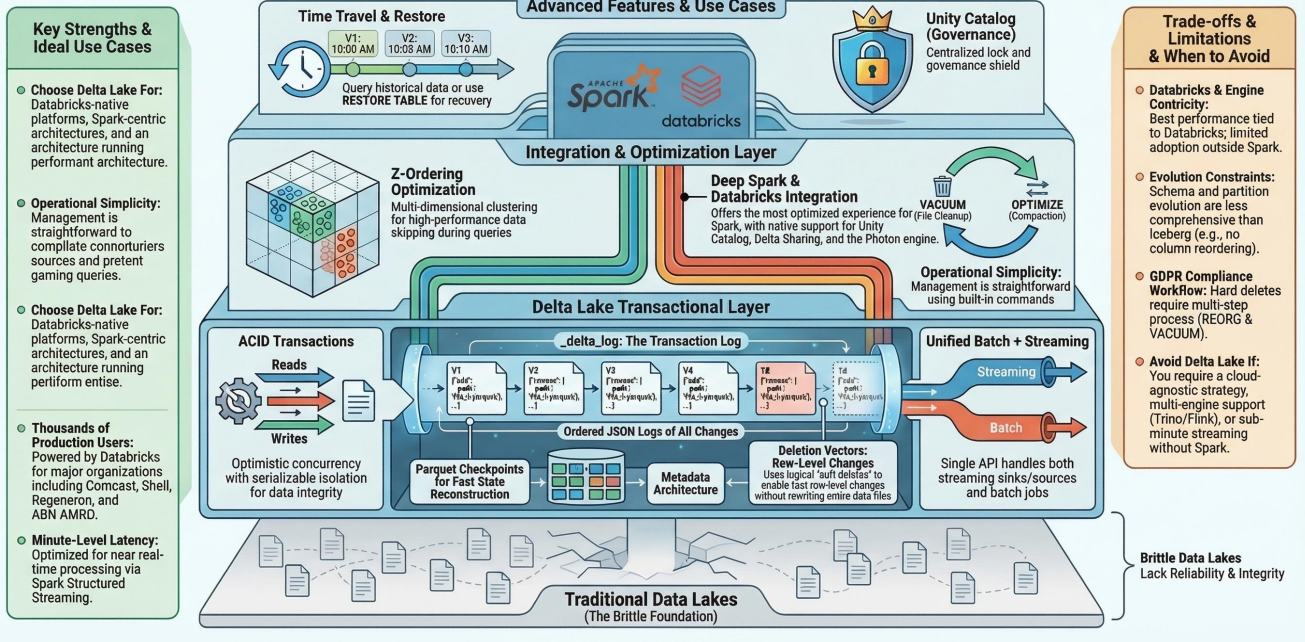
- IDEAL:** Multi-Engine Data Platforms (→)
SUPPORTING DETAIL: Best when Spark, Trino, and Flink must all interact with the same tables simultaneously.
- IDEAL:** Long-Term Modernization (→)
SUPPORTING DETAIL: Perfect for organizations requiring decades of backward compatibility and future-proofing.
- AVOID:** Streaming-First Needs (x)
SUPPORTING DETAIL: Avoid if your architecture requires sub-minute latency for real-time CDC.
- AVOID:** Simple/Small Use Cases (x)
SUPPORTING DETAIL: Overhead may exceed benefits for simple projects; consider lighter options like DuckLake.

BY THE NUMBERS

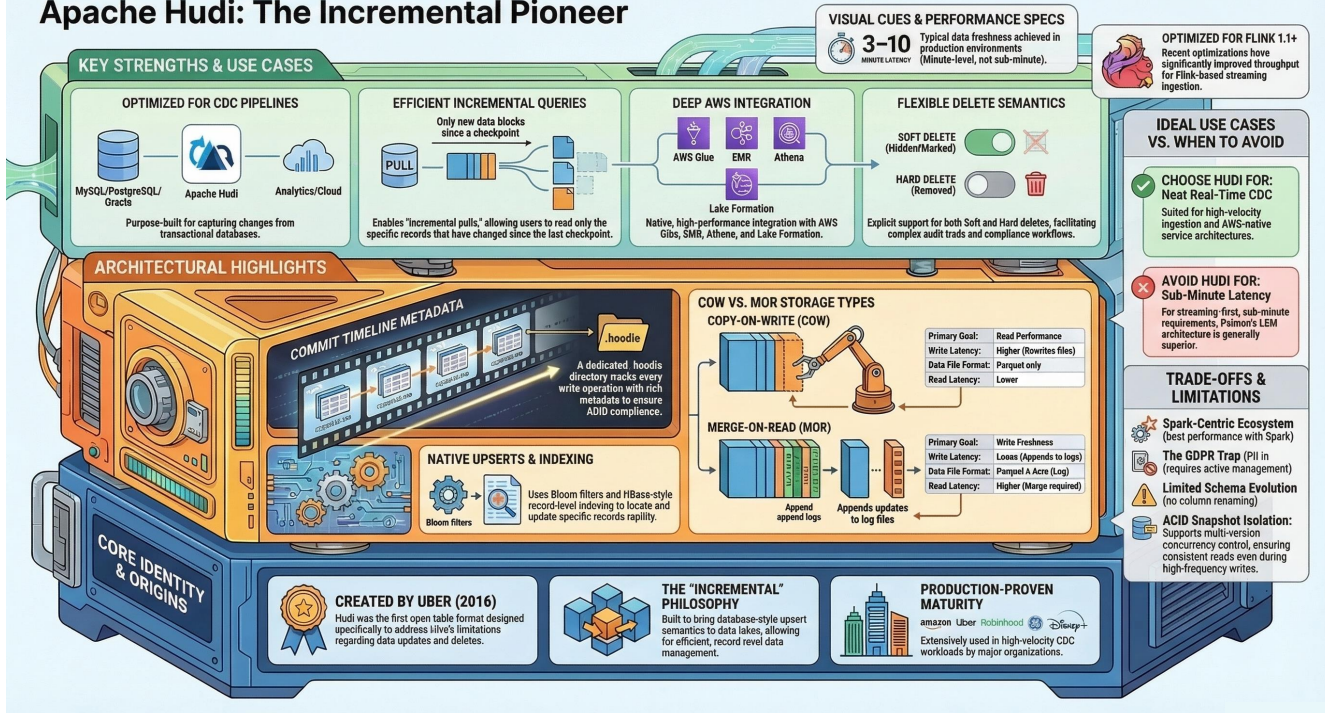
Metric	Value
Supported Engines	10+ Native Integrations
Cloud Support	AWS, GCP, Azure, Snowflake
Community Size	Largest contributor base in format category
Production Scale	Petabytes per organization

Delta Lake: The Spark-Native Transactional Layer

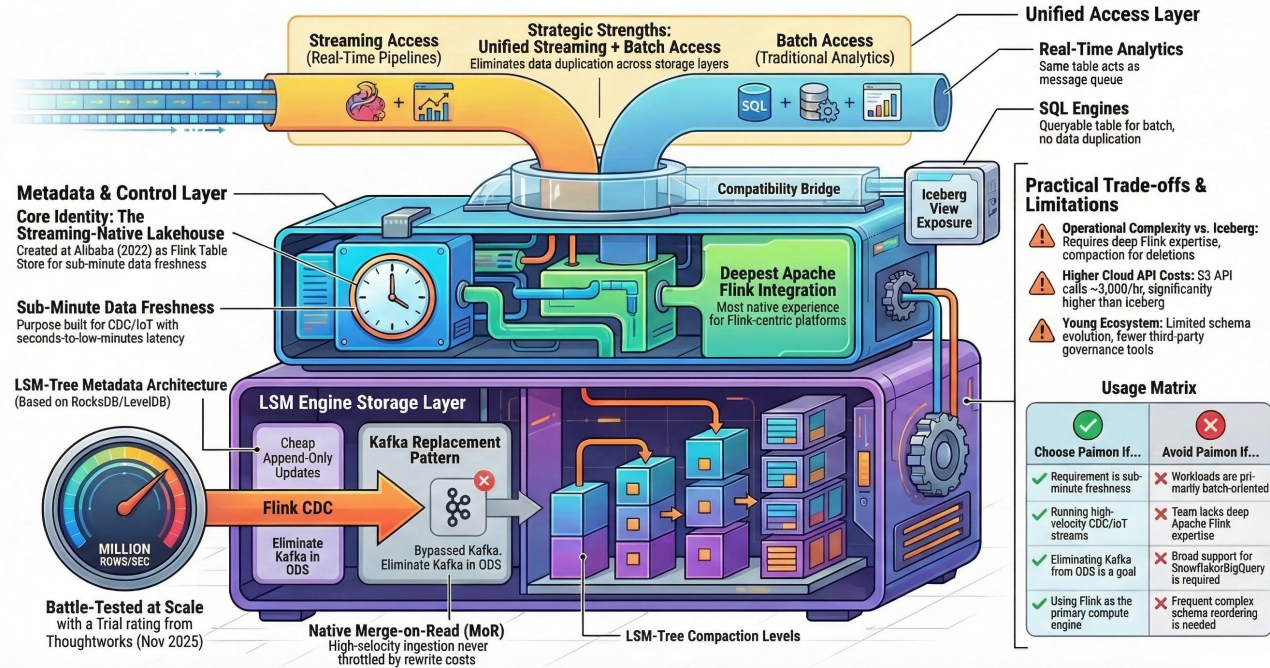
ACID Transactions Meet Data Lakes in the Databricks Ecosystem



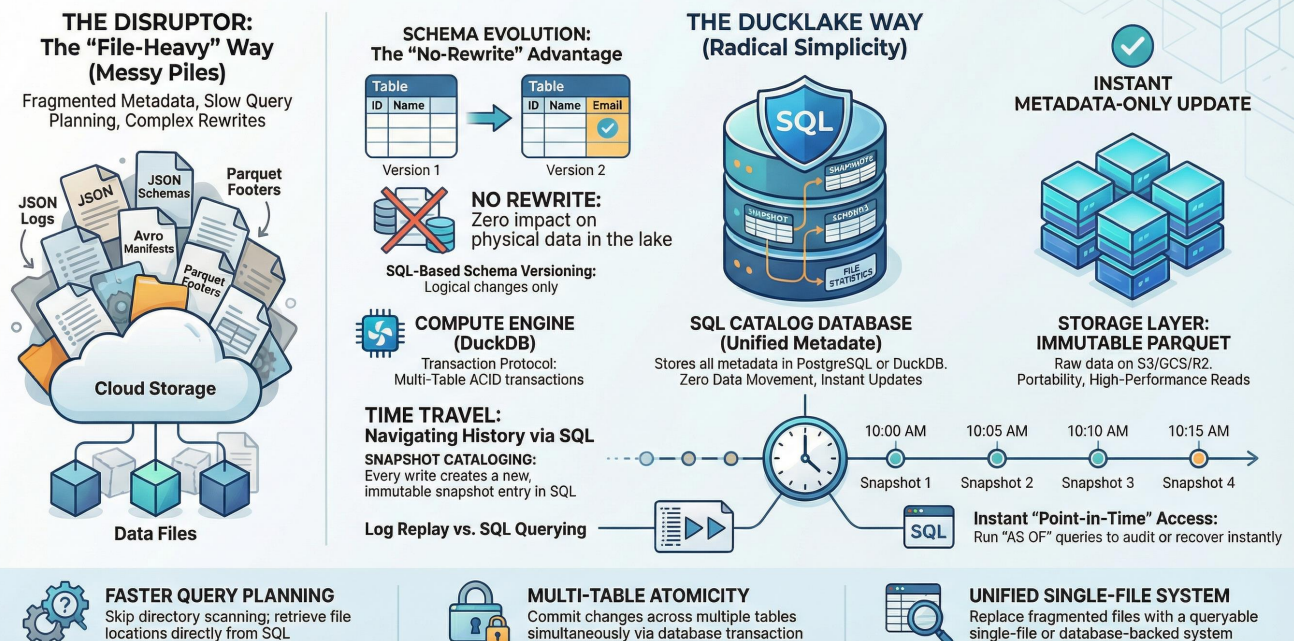
Apache Hudi: The Incremental Pioneer



Apache Paimon: The Streaming Disruptor



DuckLake: Radical Simplicity via SQL-Native Metadata



Copy-on-Write vs Merge-on-Read: How Table Formats Handle Updates

Exploring Data Lake Table Format Strategies for Immutable Storage

Copy-on-Write (CoW) Analogy

Think of it as reprinting an entire book every single time you need to fix a single typo in one chapter.



Merge-on-Read (MoR) Analogy

Think of it as keeping a stack of sticky notes with corrections inside the original book, only reprinting the book occasionally.



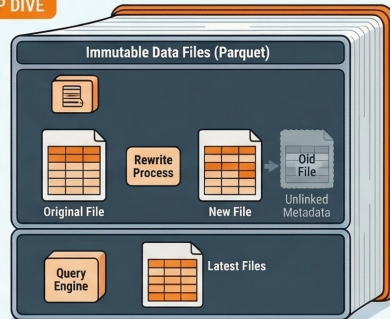
COPY-ON-WRITE (CoW) DEEP DIVE

The Write Path: Full Rewrite

When a row is **UPDATED** or **DELETED**, the system identifies the file containing that row and rewrites the entire file with the changes applied, the old file is then unlinked in the metadata.

Update/Delete Request

Changes



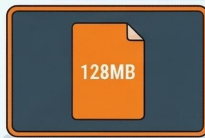
The Read Path: Direct Scan

Reads are simple and extremely fast because the engine only needs to scan the latest set of fully-formed Parquet files.

Read Request

Trade-off: High Write Cost

Writing is expensive due to high "write amplification"—even a single row change requires rewriting a large (e.g., 128MB) data file.



DECISION GUIDE

When to Choose Copy-on-Write: Select CoW when reads dominate your workload, updates are rare/inrequent, or you are running heavy batch analytics.

COMPARISON AND IMPLEMENTATION

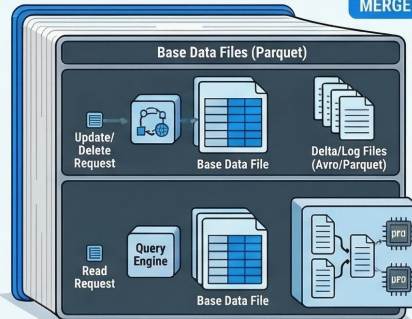
Aspect	Copy-on-Write (CoW)	Merge-on-Read (MoR)
Read Path	Simple, Direct, Fast	Complex (Merge required)
Write Cost	High (Heavy Amplification)	Low (Append-friendly)
Best For	Read-heavy batch analytics	Streaming, CDC, IoT Ingestion
Ops Effort	Snapshot expiration	Compaction scheduling
Format Mapping	Iceberg uses CoW by default (MoR in v2/v3); Delta Lake uses CoW with "Deletion Vectors"; Nodi allows users to choose per table; Paimon is natively MoR using an LSM-tree design.	

The Key Takeaway: "Both strategies are valid—the right choice depends on your read/write mix, latency requirements, and operational complexity tolerance."

MERGE-ON-READ (MoR) DEEP DIVE

The Write Path: Delta Appends

Updates and deletes are written as small, separate "delta" or "log" files (often Avro or specialized Parquet), leaving the original base data files untouched.



The Read Path: On-the-Fly Merging

The engine must read the base file **AND** the delta files, merging them in memory to provide the correct result, which adds CPU overhead.

Compaction (The Cleanup)

Periodic background processes eventually merge the small delta files back into new base files to keep read performance from degrading.



When to Choose Merge-on-Read: Select MoR for frequent updates, low-latency streaming ingestion (CDC), or when write latency is your primary bottleneck.

Snapshot in Time: How Apache Iceberg Enables Data Time Travel

The Table Metadata File (JSON)

The central "source of truth" that tracks schema versions, partition specs, and—most importantly—the list of all valid snapshots.



Section 1: The Brains—The Metadata Hierarchy

Manifest Lists & Manifests (Avro)

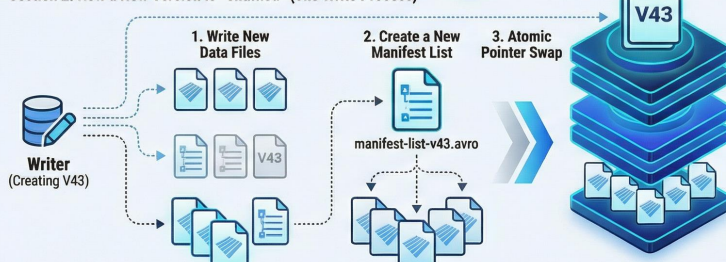
These act as an index, the Manifest List points to individual Manifests, which in turn point to the actual Parquet data files and provide column-level stats.

Total Immutability

Every data file (Parquet) and metadata file (Avro) is immutable, once written, they are never modified, only replaced by newer pointers in the JSON file.



Section 2: How a New Version is "Shuffled" (The Write Process)



Historical State



Reader (Querying the Past)

Section 3: Time Travel—Querying the Past

Jumping Back in Time

Compute engine from the current engine:

`SELECT * FROM table VERSION AS OF 42`

The compute engine ignores the current V43 pointer, and directly following the select V43 pointer.

Snapshot Isolation in Action

While a "Writer" is creating Snapshot V43, a "Reader" can still query Snapshot V42 simultaneously without seeing partial or corrupted data.

Zero Replay Required

Iceberg provides instant access to any historical state by simply reading that snapshot's manifest list.

Section 4: Why Time Travel Matters



Debugging & Recovery

If a pipeline accidentally overwrites emails with NULLs, engineers can query the last "good" snapshot to restore data.



Regulatory Compliance & Auditing

Organizations can prove exactly what data was used for a specific report or model on a specific date.



Safe Machine Learning

Data scientists can "pin" a model to a specific snapshot to ensure experiments are reproducible even as the underlying table continues to update.

GDPR & The Hard Truth of Hard Deletes: From 'Row Deleted' to 'Bits Gone from Disk'

GDPR: The Great Format Equalizer



Every format must answer:
How to go from "deleted" row to physical bits gone?

The Performance vs. Physicality Trade-off

Initial Logical Delete (Fast)



Modern optimizations
(e.g., deletion vectors)
are fast.

Physical Hard Delete (Expensive & Slow)



Subsequent steps to
remove bits are time-
consuming & costly.

The 4-Stage Compliance Lifecycle

Stage 1: Logical Delete (Fast)



Metadata-only
operation. Enables time
travel/snapshots, hides
data from active queries.

Stage 2: Compaction & Rewrite (Heavy Compute)



Merges delete metadata
& physically rewrites
files into new versions
without deleted records.

Stage 3: Retention Expiration (Metadata Cleanup)



Drops old snapshots &
historical versions
pointing to sensitive
data files.

Stage 4: Physical Garbage Collection

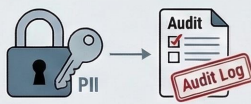


Actual removal of
abandoned, "orphan"
files from object storage.
Bits truly gone.

Format-Specific Workflow Comparison

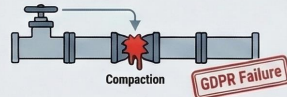
Format	Deletion Mechanism	Workflow Summary	Risk Level
Iceberg v3	Deletion Vectors	Vectors, Compaction, Snapshot Expiration, Orphan Cleanup	Moderate
Delta Lake	Deletion Vectors	Vectors, REORG, VACUUM operations, Physical Removal	Moderate
Apache Hudi	Soft vs. Hard Semantics	Choose Soft (Audit) or Hard (Immediate), Compaction, Cleaning	HIGH
Apache Paimon	LSM-tree + Vectors	Append to LSM, Mandatory Compaction, Dual Expiration, GC	HIGH
DuckLake	Delete Files (v2 Pattern)	Delete Files, Compaction, Retention Expiration	HIGH

Apache Hudi: The PII Key Trap



If Primary Keys contain PII, Hudi's "soft delete"
audit trail becomes a legal liability under GDPR.

Apache Paimon: The Compaction Dependency



Paimon's Dual Retention Configs: Compliance requires
independent management of both snapshot and
changelog retention settings to ensure data is fully purged.

DuckLake: Governance Immature



Lacks mature automated governance tooling,
requiring custom, high-risk compliance pipelines.

The Hidden Costs of Compliance & Key Takeaway

SLA Enforcement Gap & Metadata Housekeeping



No "free" compliance: requires active monitoring &
pipelines. Orchestrated cleanup of manifest files,
transaction logs, catalogs needed.

Operationalize the Full Pipeline

"Choose your format based on your team's ability
to operationalize the full hard-delete pipeline, not
just the initial soft delete performance."

The Lakehouse Brain: Why Data Catalogs are Essential

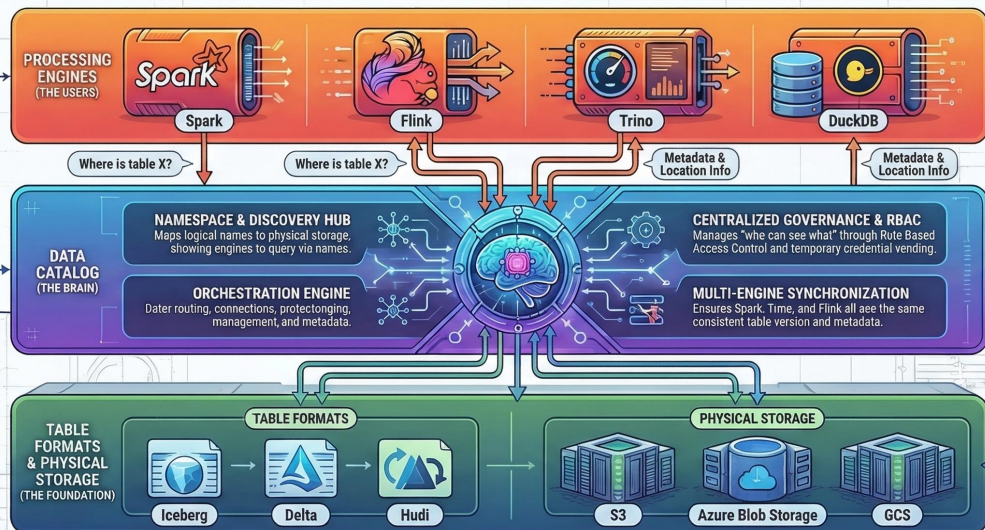
WITHOUT A CATALOG: THE CHAOS SCENARIO



MANUAL PATH MANAGEMENT & BLINDNESS
Engineers hardcode paths,
making new tables "invisible"

METADATA & SCHEMA DRIFT
Engines struggle to agree on
column types and partition specs

GOVERNANCE & COORDINATION FAILURES
Security gaps and conflicting
writes between different engines



Key Takeaway:
Without a catalog, you have a folder of PDFs.
With a catalog, you have a library with a
functional card catalog.

The Three Waves of Data Catalog Evolution: From Metadata to Versioning

“Just as table formats evolved in three waves, so did catalogs—from centralized databases to git-like workflows for data.”

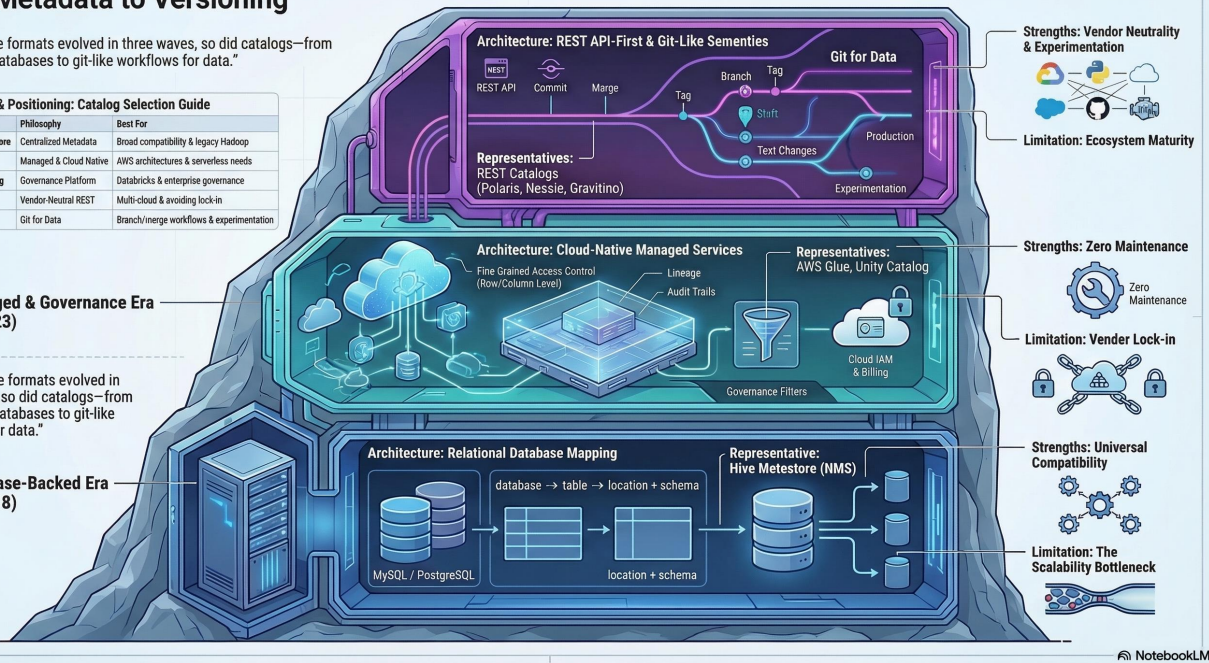
Comparison & Positioning: Catalog Selection Guide			
Wave	Catalog	Philosophy	Best For
1	Hive Metastore	Centralized Metadata	Broad compatibility & legacy Hadoop
2	AWS Glue	Managed & Cloud Native	AWS architectures & serverless needs
	Unity Catalog	Governance Platform	Databricks & enterprise governance
	Polaris	Vendor-Neutral REST	Multi-cloud & avoiding lock-in
3	Nessie	Git for Data	Branch/merge workflows & experimentation

Wave 2: The Managed & Governance Era (2018–2023)

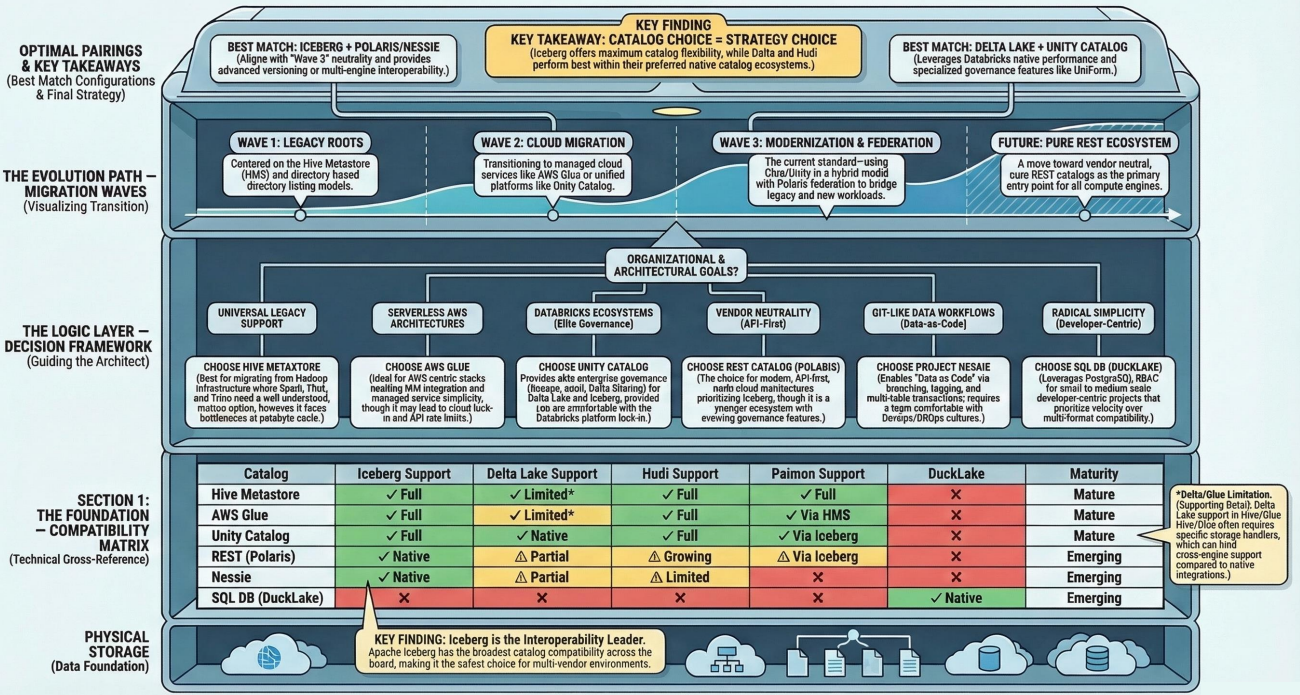
“Just as table formats evolved in three waves, so did catalogs—from centralized databases to git-like workflows for data.”

Wave 1: The Database-Backed Era (2010–2018)

“Wave 1 gave us metadata. Wave 2 gave us governance. Wave 3 is giving us versioning and neutrality.”
Select your catalog based on your specific governance needs and your tolerance for vendor lock-in.



The Data Lakehouse Catalog Compatibility Matrix & Decision Framework (2025-2026)



DuckLake: What if a SQL Database IS the Catalog?

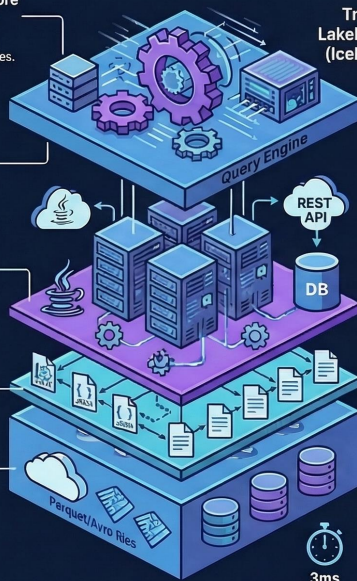
Era 1: The Hive Metastore
Early metadata stored in relational databases struggled with filtering bottlenecks at petabyte scales.

Era 2: File-Based Table Formats
Metadata moved to object storage (JSON/Avro) for infinite scale, but introduced "file-system dancing" and high latency.

Catalog Server (Java/REST/DB)
Hive Metastore Era: MySQL/Postgres for paths, file-listing outliners

Table Format (JSON Manifests)
JSON/Avro

Object Storage
Parquet/Avro files



Traditional Lakehouse Stack (Iceberg/Delta)

The DuckLake Stack

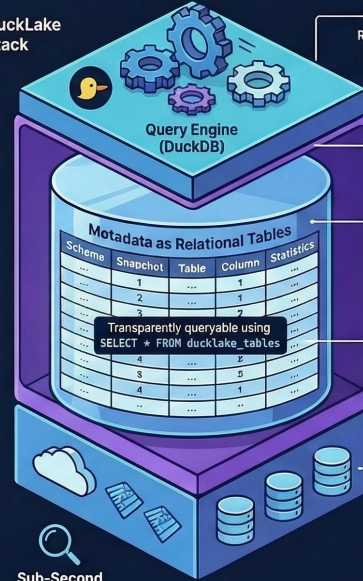
Era 3: The DuckLake "Radical Simplicity"
Reintroducing the SQL database as the catalog, handling metadata as relational tables without the "file zoo."

SQL Database (Catalog + Metadata Merged)

Core Relational Components
Manages snapshots (state), schemas, tables (inception-style), columns, and file-level statistics

Data Inlining
Small appends are buffered directly in the SQL database, then flushed to Parquet later to avoid small files problem

Object Storage (Parquet)



Collapses two layers into one.

The DuckLake Question

If Iceberg + REST catalogs are the "Linux of data lakes," DuckLake asks: What if PostgreSQL IS the catalog?

PERFORMANCE BENCHMARKS

3ms Metadata Inserts
Adding a new snapshot to a petabyte-scale dataset takes just 3 milliseconds in DuckLake.

4x Faster Query Planning
DuckLake outpaced Iceberg by 4x due to fewer network round trips and no Java/REST overhead.

Sub-Second Pruning
Selecting a specific day or month of data from a petabyte-scale lake remains sub-second via SQL indexing.

KEY TRADE-OFFS: SIMPLICITY VS. SCALE

DuckLake offers radical simplicity and multi-table ACID transactions, but introduces a database dependency and is currently DuckDB-centric.

DuckLake: From Browser to Production in 5 Steps

Step 0: Browser Preview (DuckDB Wasm)

Start with SQL directly in the browser window with no installation required, allowing for immediate data exploration.

Step 1: In-Process Analytics (Python + DuckDB)

Move to local development to query CSV and Parquet files with zero configuration using DuckDB as a universal data wrangling tool.

Step 2: The Lakehouse Layer (DuckLake)

Attach the DuckLake extension to add ACID transaction, snapshots, and time travel capabilities to your local files.

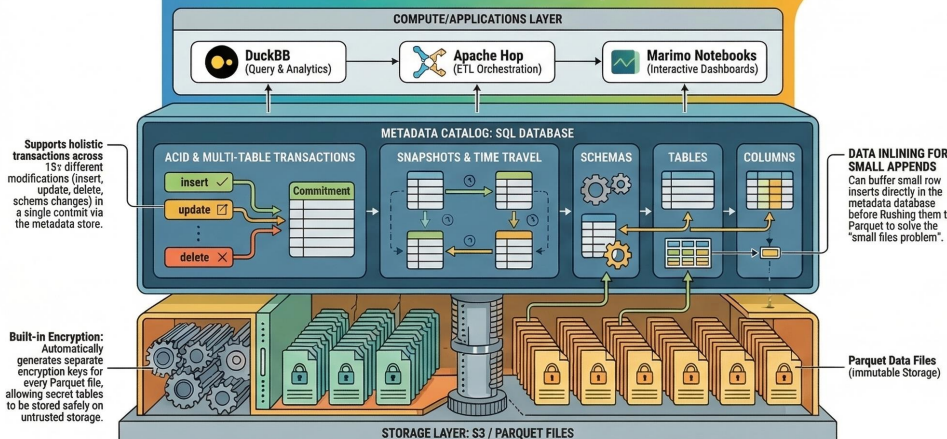
Step 3: Visual ETL (Apache Hop)

Implement drag-and-drop orchestration to build and manage data pipelines that feed directly into the DuckLake format.

Step 4: Reactive Analysis (Marimo Notebooks)

Create interactive dashboards and reactive graphs to visualize and analyze lakehouse data in real time.

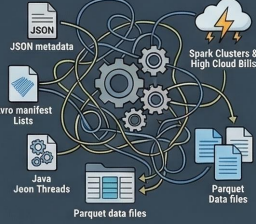
Complete Production Stack (Step 5)



Architectural Contrast: Complexity vs. Simplicity

The Traditional "File Zoo" Stack

Relies on a 528 complexity of Java threads, REST catalogs, JSON manifests, Avro manifest lists, and Parquet data files, often requiring Spark clusters and high cloud bills.



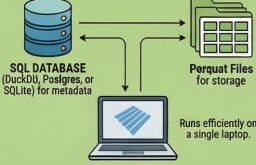
4x

KEY FINDING: 4x Performance Advantage

Benchmarks show DuckLake is up to four times faster for updates than Iceberg because it eliminates multiple non-predictable network round trips to metadata files.

The DuckLake Lean Architecture

Replaces the "zoo" with a simple SQL database for metadata and Parquet for storage, running efficiently on a single laptop.



NotebookLM

From Browser to Lakehouse: The DuckLake Architecture

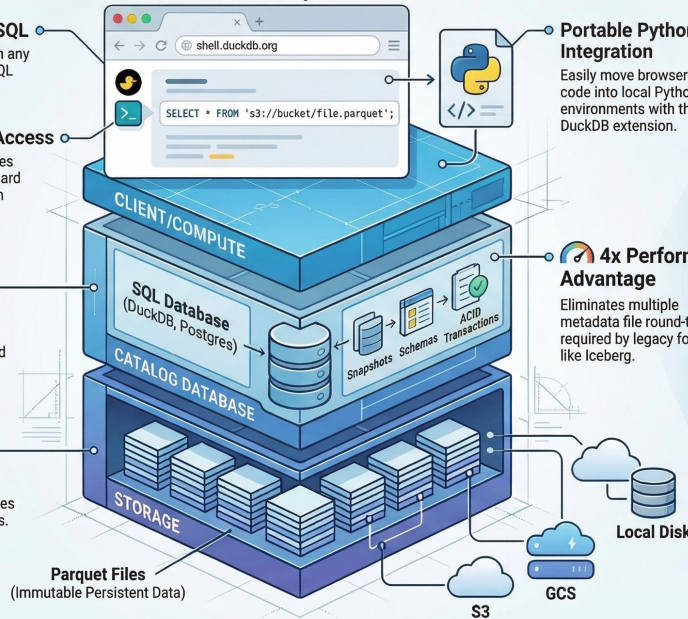
Step 0: The 30-Second Browser Preview

Zero-Installation SQL
Open shell.duckdb.org in any browser to run instant SQL without any setup.

Direct Data Lake Access
Query remote Parquet files immediately using standard SELECT statements from the web.

The Metadata Catalog Layer
Uses a standard SQL database to manage snapshots, schemas, and ACID transactions.

The Scalable Storage Layer
Persistent data is stored as immutable Parquet files on S3, GCS, or local disks.



Step 1: The DuckLake "Database + Files" Architecture

Portable Python Integration
Easily move browser-tested code into local Python environments with the DuckDB extension.

4x Performance Advantage
Eliminates multiple metadata file round-trips required by legacy formats like Iceberg.

ARCHITECTURAL COMPLEXITY COMPARISON

LEGACY TABLE FORMATS
(e.g., Iceberg)



Metadata Storage:
Nested JSON/Avro File "Zoo"



Tech Stack:
Java, Spark, REST Catalog



Update Speed:
High Latency
(Multiple Writes)

DUCKLAKE



Metadata Storage:
Single SQL Database



Tech Stack:
DuckDB + SQL



Update Speed:
3-5ms
(Sub-second planning)

NotebookLM

Step 2: Add DuckLake – The Lakehouse Layer

```
1 import duckdb
2 con = duckdb.connect()
3
4 # Install and load DuckLake extension
5 con.sql("INSTALL ducklake")
6 con.sql("LOAD ducklake")
7
8 # Attach a DuckLake catalog
9 con.sql("""
10 ATTACH 'ducklak:metadata.ducklake' AS my_lake
11 (DATA_PATH 'data_files')
12 """)
13
14 # Switch to the lakehouse
15 con.sql("USE my_lake")
16
17
18 # Creating and Populating Tables
19 # Create a lakehouse-managed table
20 con.sql("CREATE TABLE sales (id INT, amount DECIMAL, product VARCHAR)")
21
22 # Insert data with ACID guarantees
23 con.sql("INSERT INTO sales VALUES (1, 100.50, 'Widget')")
24 con.sql("INSERT INTO sales VALUES (2, 250.75, 'Gadget')")
25
26
27 # Advanced Operations (Time Travel)
28 # Query the lakehouse table
29 con.sql("SELECT * FROM sales").show()
30
31 # Time travel: query previous snapshot
32 con.sql("SELECT * FROM sales AS OF SYSTEM TIME '-5 minutes').show()")
```

Install and Initialize
With just three lines of SQL, DuckLake is installed, loaded, and attached as a catalog to manage your data files.

Instant Time Travel
Because DuckLake tracks every snapshot in its metadata catalog, users can query the database as of a specific system time or version.

THE COMMAND CENTER (INTERFACE)

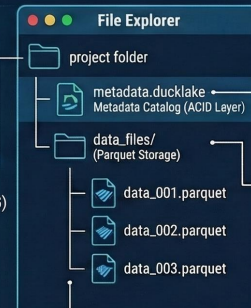
THE LAKEHOUSE ENGINE (METADATA/CATALOG)

THE DEEP STORAGE (FOUNDATION)

Before/After Comparison

Loose Files

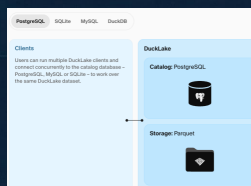
Structured Lakehouse



Catalog-Driven Management:
The metadata layer stores snapshots, schemas, and statistics, acting as the "brain" that coordinates how compute nodes interact with storage.

Parquet-Based Persistence:
Row data is stored as immutable Parquet files in the "data_files" directory, ensuring high-performance compression and portability.

ACID-Guaranteed Transactions:
DuckLake uses a SQL metadata database to ensure that every insert or update is atomic and consistent, preventing data corruption during concurrent access.

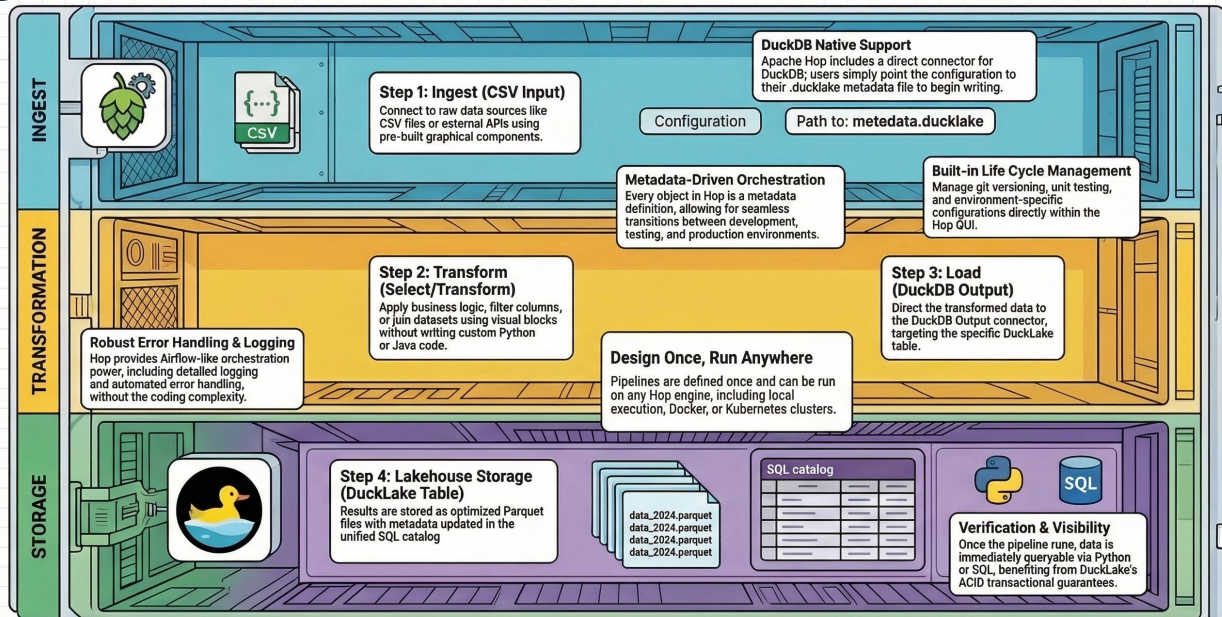


NotebookLM



Orchestrate Pipelines: Drag-Drop ETL to Your Lakehouse

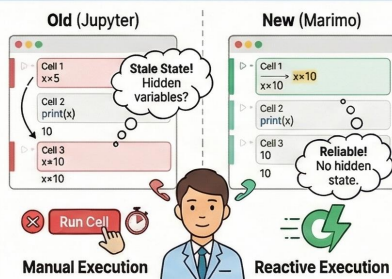
Apache Hop & DuckLake: Build Repeatable Pipelines Without Python ETL Code



Step 4: Marimo Notebooks – Reactive Analysis

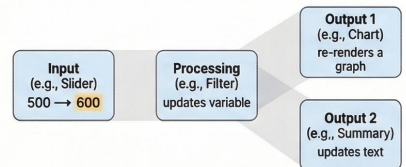


1. The Problem with Traditional Notebooks



2. The Marimo Reactive Solution

Dataflow Graph Enforcement: Change an input, all dependent cells re-run automatically based on dependency.



No More Hidden State: Deleting a cell removes its variables, keeping the environment clean and reproducible.

3. Interactive UI Components

```
import marimo as mo
import duckdb

con = duckdb.connect("metedata.ducklake")
df = con.sql("SELECT * FROM sales").df()

amount_filter = mo.ui.slider(0, 1000, value=500)
filtered = df[df["amount"] > amount_filter.value]

mo.md(f"Sales > ${amount_filter.value}: {len(filtered)} rows")
```

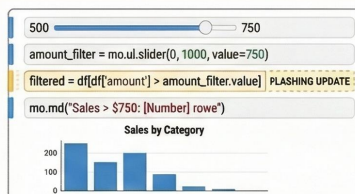
mo.ui 500

Category

Update

Notebooks as Functional Apps: Use `'mo.ui'` (sliders, dropdowns, buttons) for reactive inputs that trigger downstream logic.

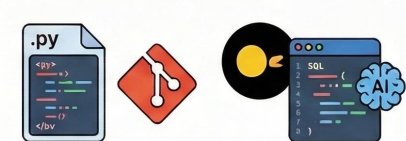
4. Real-Time Data Analysis



Reactive Data Filtering: Variables update instantly when UI inputs change.

Instant Visual Feedback: Charts and summaries re-render without clicking 'Run'.

5. Built for Developers



Pure Python & Git-Friendly

Marimo notebooks are standard `.py` files for easy version control, reviews, and testing (e.g., PyTest).

SQL & AI Integration

Built-in support for DuckDB, SQL cells, and AI-native features like autocomplete and inline edits.

6. Why Reactive Analysis Matters

Fixing the Broken Default: Moves analysis away from fragile, stateful messes toward reliable, shareable apps.



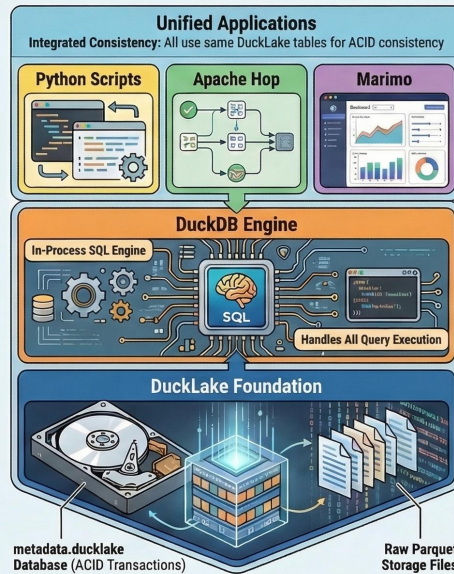
Deploy Anywhere: Export as WebAssembly HTML or serve as a web app using Marimo CLI, working the same locally and deployed.

Production-Ready Lakehouse on a Laptop: The \$0/Month Modern Data Stack

Visual Demo: The Integrated Workflow



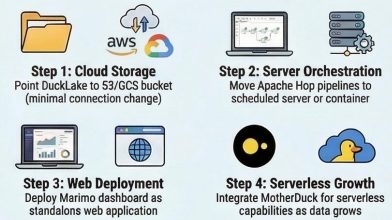
Visualizing the physical storage layer
Visualizing the consumption layer



The Revelation: Cost vs. Complexity

Component	Traditional Enterprise Stack	Local "Laptop" Stack
Messaging/Ingestion	Kafka Cluster (3+ Nodes)	Apache Hop (Java App)
Compute Engine	Spark Cluster (Driver + Executors)	DuckDB (Single Process)
Orchestration	Airflow (Scheduler + Workers + DB)	Apache Hop
Workspace/Storage	Databricks (\$\$\$\$) + Cloud Storage	DuckLake (Local Files)
Dashboarding	Tableau/PowerBI (Licenses)	Marimo (Open Source Python)
Total Cost	\$5,000-\$30,000/Month + DevOps	\$0/Month + Runs on Laptop

The Scaling Path



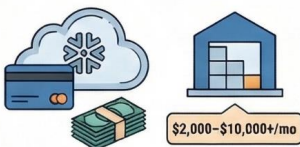
"Don't Start with Databricks and Work Backwards."

Start Simple, Scale Incrementally. The future of data engineering is eliminating complexity. Save expensive infrastructure for proven needs.

From Bill Shock to Blazing Fast: Definite's Migration to the Duck Stack

PART 1: STEP-BY-STEP: THE PATH TO THE DUCK STACK

Step 1: Identify the "Bill Shock"



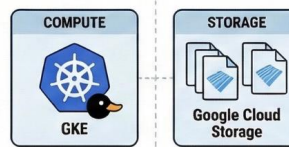
Traditional cloud warehouses can cost mid-sized companies thousands monthly due to complex credit models and storage markups.

Step 2: Choose the Pragmatic Engine



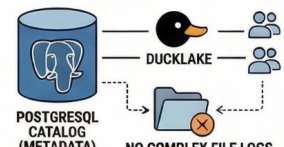
Definite bypassed Postgres and ClickHouse in favor of DuckDB for its embeddable simplicity and Postgres-compatible SQL.

Step 3: Decouple Compute and Storage



DuckDB runs on Kubernetes (GKE) while data is stored as Parquet files on Google Cloud Storage for elastic, low-cost scaling.

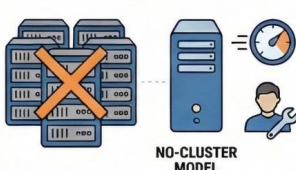
Step 4: Implement DuckLake for ACID



To handle concurrent writers and metadata, DuckLake stores table metadata in a standard PostgreSQL database.

PART 2: THE PERFORMANCE & VELOCITY PAYOFF

Step 5: Eliminate Cluster Overhead



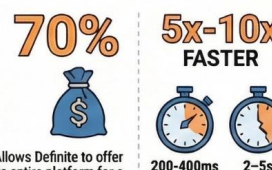
Moving to a "no-cluster" model removes cold-start latencies, rebalancing issues, and need for dedicated infrastructure teams.

The Economic Revolution

MONTHLY COST COMPARISON		
	SHOWFLAKE + LOGKER + FIVETRAN	DEFINITE (DUCKDD/DUCKLAKE)
Compute:	\$2k - \$5k	Included
Storage (1TB):	\$40 - \$80	\$20 - \$40
BI & Connectors:	\$1.5k - \$5k	Included
Total Monthly:	\$3,500 - \$10,000+	\$250 - \$500

Total monthly costs for a mid-market stack plummeted from nearly \$10,000 to under \$500.

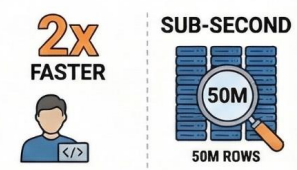
70% Cost Reduction & 5x-10x Faster Queries



Allows Definite to offer its entire platform for a fraction of the price.

Typical queries dropped from 2-5 seconds to 200-400ms by eliminating network hops.

2x Engineering Velocity & Sub-Second Ad-hoc Exploration



Developers run the full stack locally, shortening the development lifecycle.

Filtering and grouping 50M rows happens in under 1 second for instant feedback

BeyondHive: The Anatomy of the Modern Lakehouse

THE THREE WAVES OF EVOLUTION

Wave 1: 2016-2018

Wave 1: The Incremental Pioneers

Led by Hudi & Delta Lake. Introduced ACID transactions and CDC for data lakes, proving cloud storage could be mutable.



Wave 2: 2018-2022

Wave 2: The Universal Standard

Apache Iceberg prioritized engine neutrality and comprehensive scheme evolution, making interoperability the new priority.



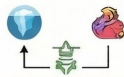
Wave 3: 2022-Present

Wave 3: Streaming & Simplicity

Newcomers like Apache Paimon (streaming first) and DuckLake (SQL-native) rethink fundamentals to prioritize sub-minute freshness and developer velocity.



FORMAT FACE-OFF & DECISION MATRIX

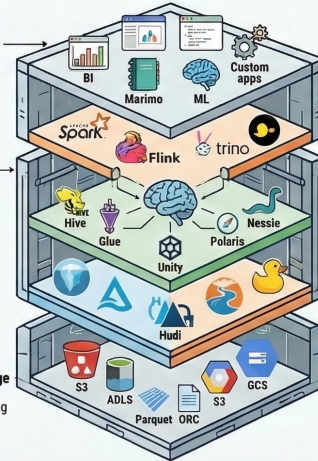


- For Neutrality**
Best for multi-engine platforms and long-term vendor neutrality.
 Apache Iceberg
- For Spark Ecosystems**
Optimized for Databricks and Spark-native environments.
 Delta Lake
- For CDC Specialization**
The preferred choice for Change Data Capture-heavy architectures on AWS
 Apache Hudi
- For Streaming Freshness**
Designed for Flink-heavy workloads requiring sub-minute data freshness
 Apache Paimon
- For Radical Simplicity**
Chosen for high developer velocity where a SQL database serves as the catalog
 DuckLake

LAYER 5: User Interface & Tools
The top-level access points including BI tools, Marimo notebooks, ML applications, and custom apps.

LAYER 3: The Catalog
The 'brain' of the stack (Hive, Glue, Unity, Polaris, Nessie); it manages governance and location discovery.

LAYER 1: Persistent Object Storage
The foundation consisting of S3, ADLS, or GCS, storing data in open files like Parquet or ORC.



LAYER 4: Compute Engines
The processing power (Spark, Flink, Trino, DuckDB) that interacts with the table formats to query data.

LAYER 2: Open Table Format
The metadata layer (Iceberg, Delta, Hudi, Paimon, DuckLake) that defines how files are organized and managed.

THE HIDDEN COSTS REALITY CHECK

The 4-Step GDPR Deletion Workflow:



Write Strategy Trade-offs:

Copy-on-Write (CoW) provides fast reads but expensive writes, Merge-on-Read (MoR) allows fast writes but complex, slower reads.

Operational Overheads:

Costs scale with metadata complexity (S3 API bills) and team expertise (e.g., Flink for Paimon).

THE PRACTICAL PATH FORWARD

The 'Start Simple' Progression:

Browser (DuckDB Wasm) -> Python + DuckDB -> Add DuckLake -> Apache Hop pipelines -> Marimo notebooks

“Start simple. Prove value. Scale when needed.”

(You don't need expensive proprietary platforms to build a functional lakehouse).

Questions?

Modeling the Mesh

Understanding data products and domains

Juha Korpela – Datakor Consulting

About me

Juha Korpela

- Based in Helsinki, Finland
- 15+ years in Data Management in a variety of industries with diverse expertise:
 - Data Products
 - Data Governance
 - Data Modeling
 - Data Engineering
 - Data Strategy
- Entrepreneur & Consultant @ [Datakor Consulting](#)
- Founder & Chairman @ [Helsinki Data Week](#)
- Podcast host @ [Helsinki Data Mafia](#)
- Speaker: Data Day Texas, IRM UK Enterprise Data, Data Mesh Live...
- ex-CPO @ Ellie Technologies Inc.
- ex-Head of Data Platform @ UPM-Kymmene Oyj



Follow me on LinkedIn:
linkedin.com/in/jkorpela



Agenda

- Data Mesh basics
- Why context matters – semantic interoperability
- Data modeling as the key to context
- Modeling at two levels – data products and domains
- The Big Picture of semantics across the enterprise

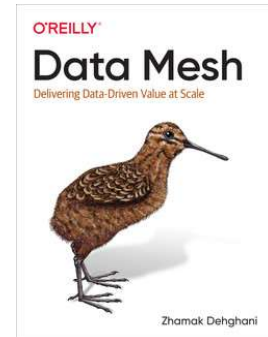


Data Mesh basics

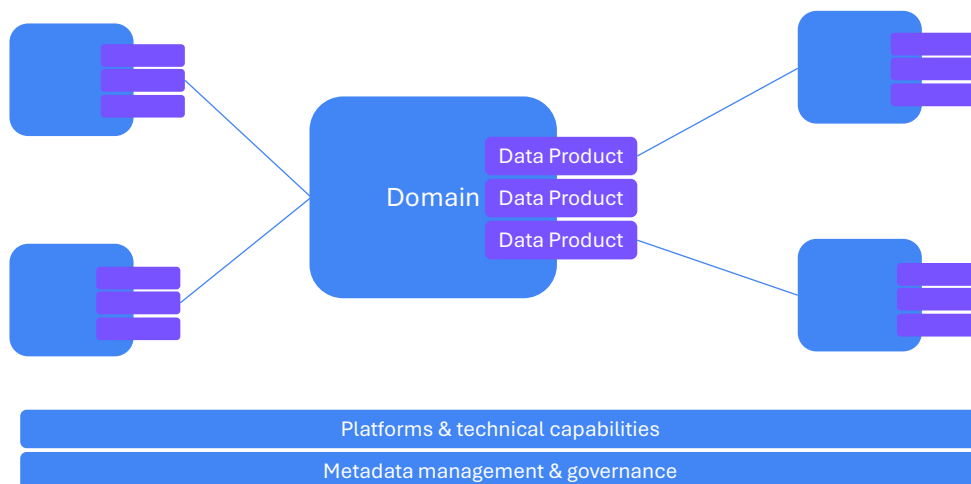
The paradigm, its principles, and how it works in reality

Data Mesh: background

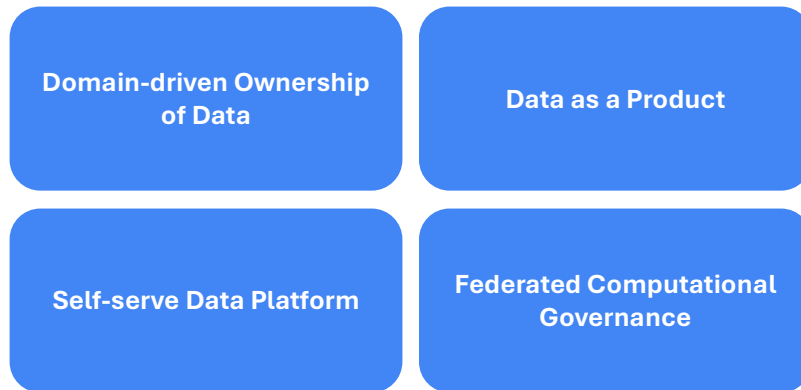
- A framework developed by **Zhamak Dehghani**
- **Domain-oriented**, decentralized model
- Domain teams are responsible for all the data within a domain & distributing it outside the domain
- All data distribution & consumption done through **data products**
- Domains and their data products form a **mesh**
- The Data Mesh is **supported by shared capabilities**: technology platforms + governance model



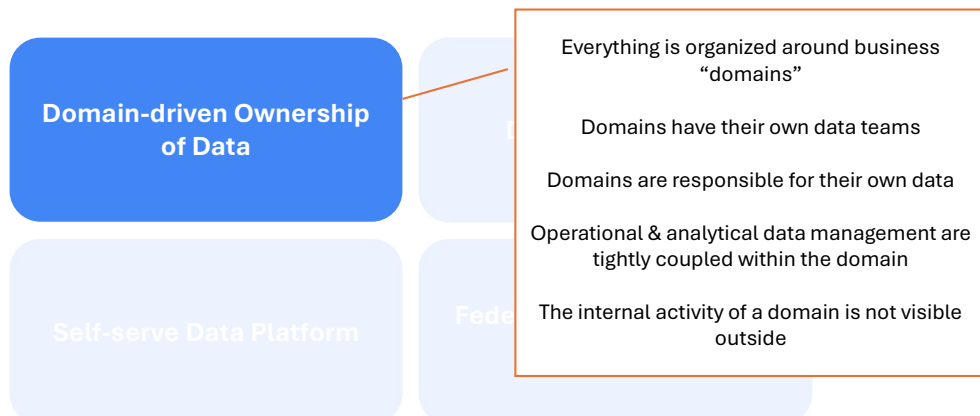
Data Mesh in a nutshell



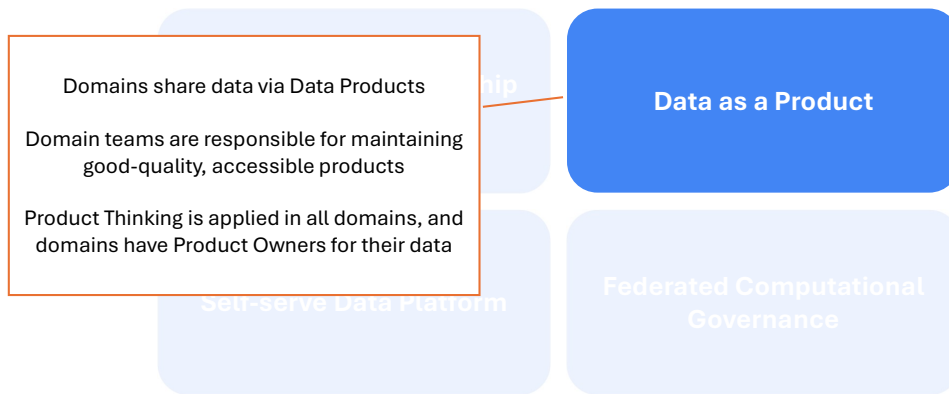
Four principles of Data Mesh



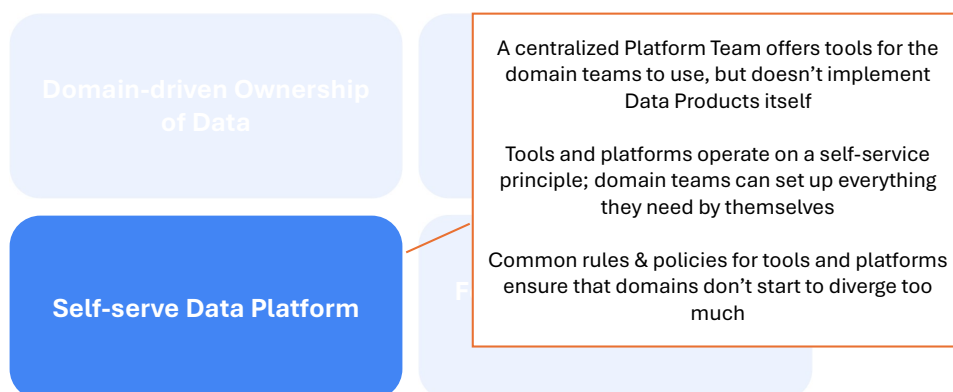
Four principles of Data Mesh



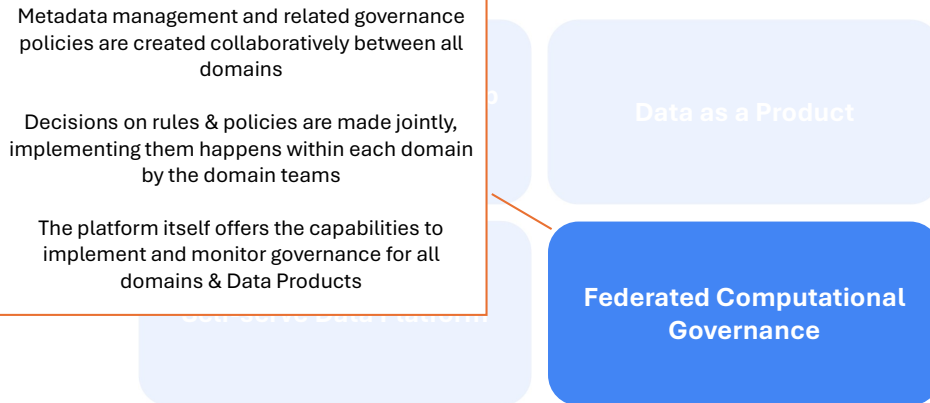
Four principles of Data Mesh



Four principles of Data Mesh



Four principles of Data Mesh



Data Mesh in reality

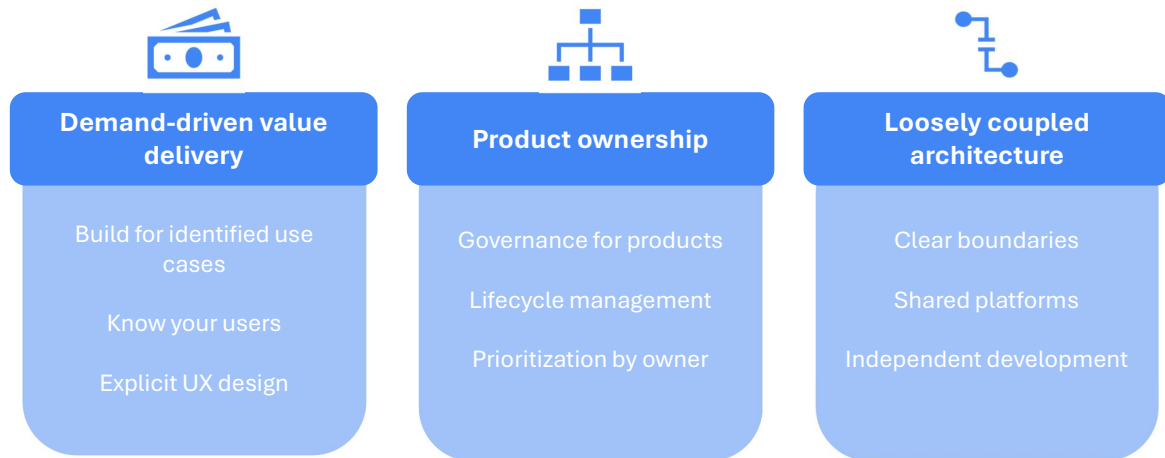
- 100% pure Data Mesh implementation is **rare**
- Significant technology effort
- **Massive** organizational effort
- Benefits from **scale** - example organizations: PayPal, Spotify, Roche...

HOWEVER...

- Real-life benefits in **shifting data responsibilities closer to value-creation**
- **“Data as a Product”** –**thinking** valuable for everyone
- Separating **Platform teams** (building capabilities) and **Domain teams** (building solutions) clarifies roles and boundaries

➤ ***If you want to do decentralized data products, you will end up scaling to Data Mesh!***

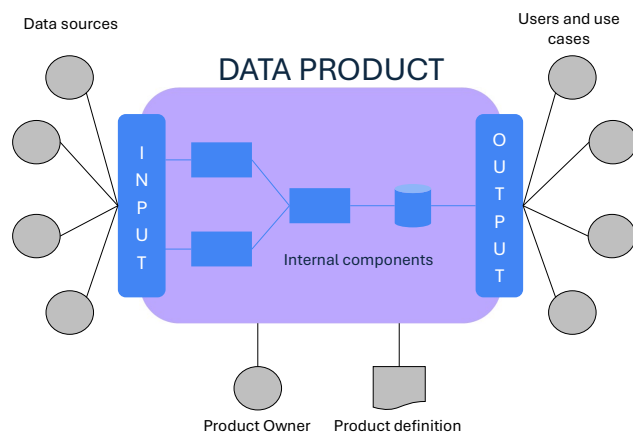
Product thinking: from projects to products



What is a “Data Product”? My definition!

A “packaged” set of data and components, that has:

- Clear boundaries
- One owner
- A lifecycle
- Identified users and benefits
- Input and output interfaces
- Understandable definition
- “Data and code”!



Why context matters

Semantic interoperability of Data Products & Domains

Data



"Productized" data



Product definition

- Contents of the product: raw materials, allergens, production method...
- Instructions & policies: what can be done with the product and how (microwave & eat!)
- Warnings & limitations: how the product is NOT to be used
- Technical information: size, structure, packaging material, recycling...
- Producer and their contact details

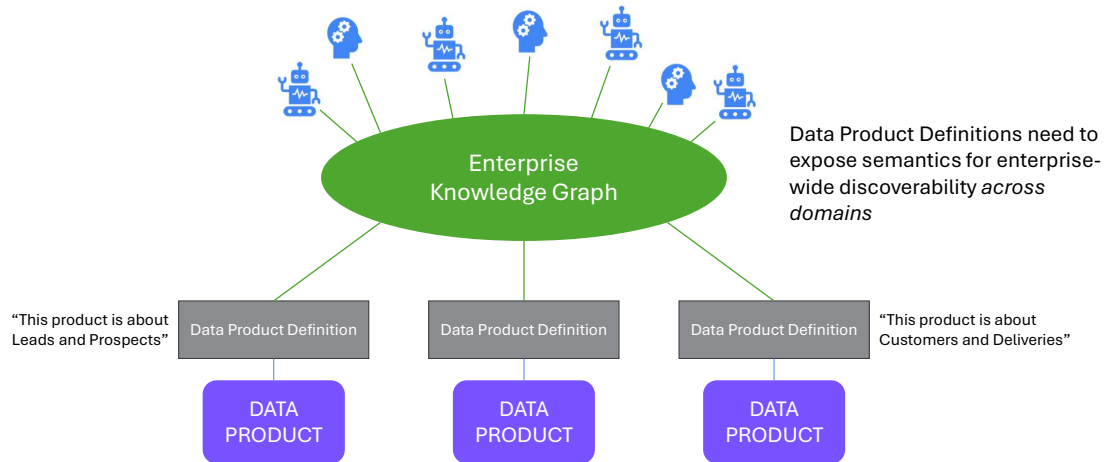


Data Product metadata

What information should we have in the “wrapper” for a Data Product?

- **Contents**: semantic description of the data
- **Instructions & policies**: what are the intended use cases, how to get access
- **Warnings & limitations**: infosec, legal limitations (e.g. GDPR), deployment status (in production or just for testing?)
- **Technical information**: database objects, files, pipelines, authentication methods
- **Producer**: who owns the Data Product, who to ask for help, who is responsible for operations

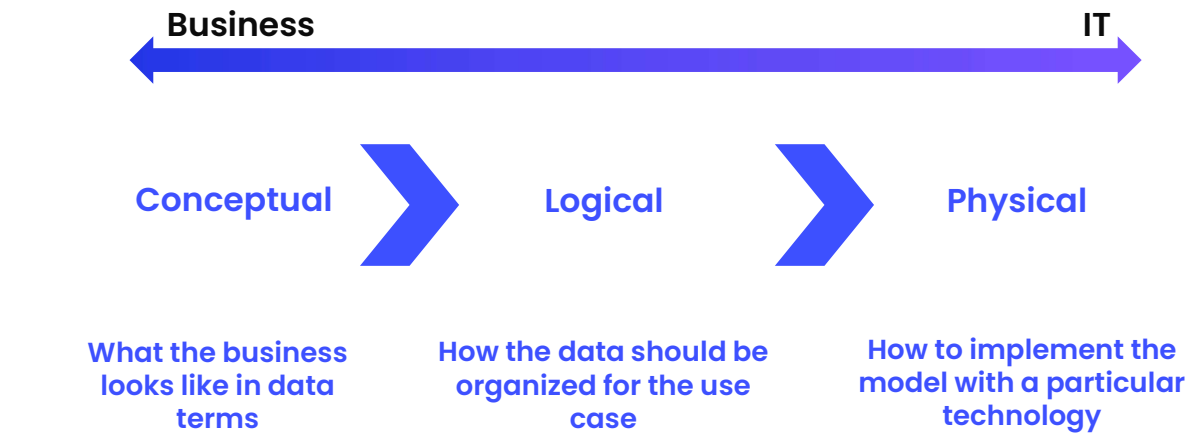
The need for semantics



Data modeling as the key to context

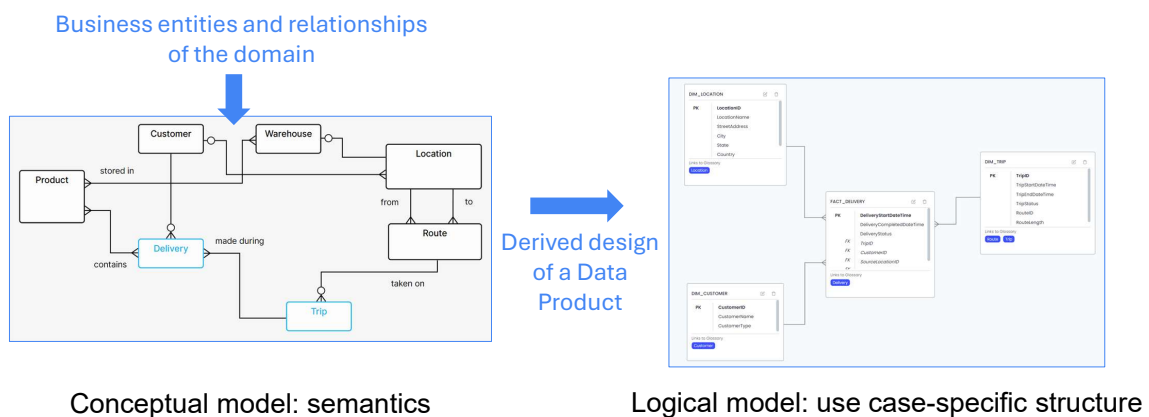
Capturing semantics & structure

Three levels of data modeling



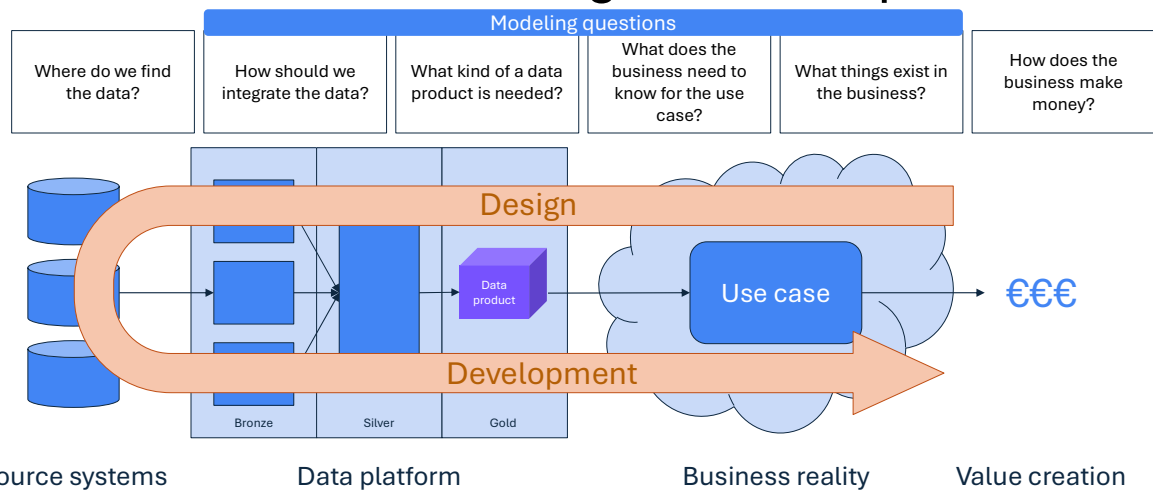
DATAKOR
CONSULTING

Capturing context with data models

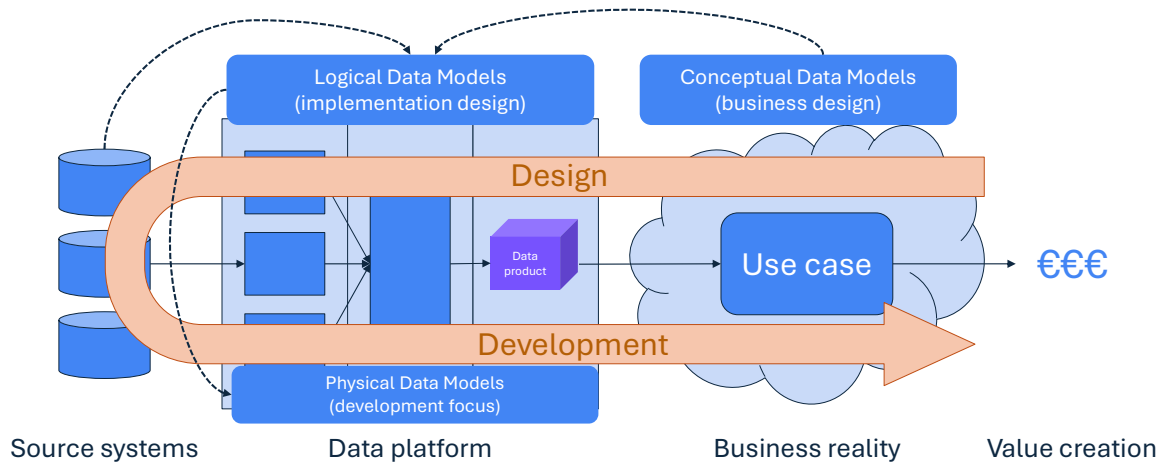


DATAKOR
CONSULTING

Business-driven data design & development



How modeling happens during the process



Modeling Domains & Data Products

Working on two levels

Two problems to solve at the same time

“SMALL PROBLEM”

How does an **individual Data Product** deliver as valuable results as possible, as efficiently as possible?



“BIG PROBLEM”

How does the **whole organization** ensure that ALL Data Products can **CONSISTENTLY** deliver valuable results efficiently?

Different approaches needed

“SMALL PROBLEM”

Focus:

- Understanding the use case
- Development velocity
- Technical optimization
- Minimal documentation overhead

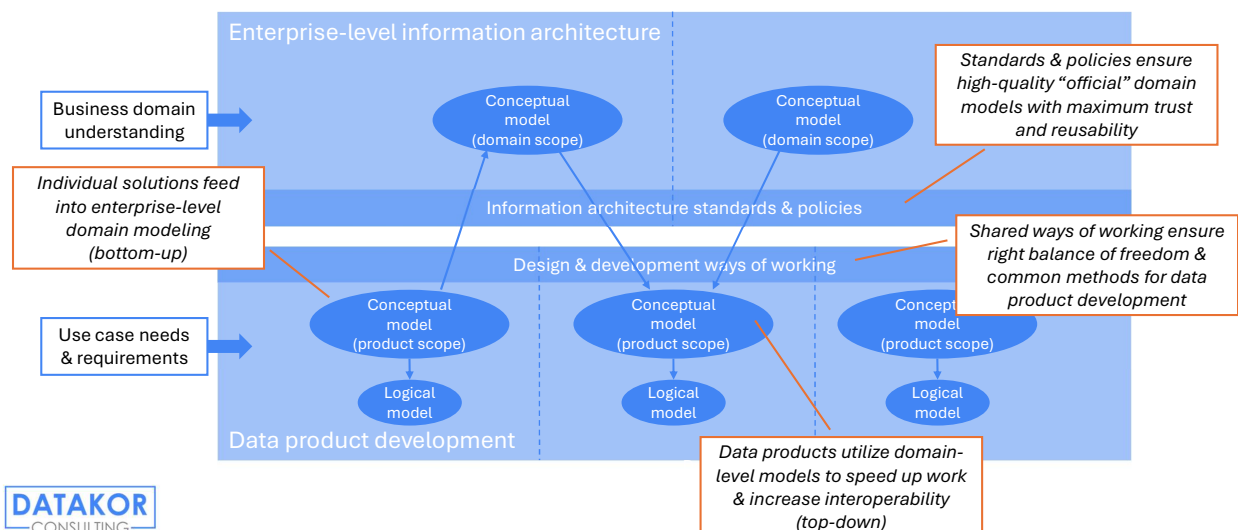


“BIG PROBLEM”

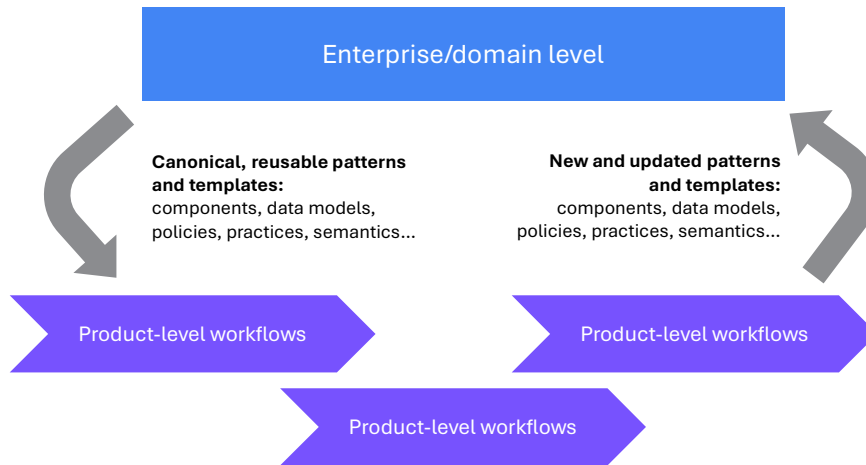
Focus:

- Reusable components & templates
- Shared ways of working
- Avoiding reinventing the wheel
- Shared semantics for interoperability

Organizing data modeling on two levels



Key success factor: feedback loops

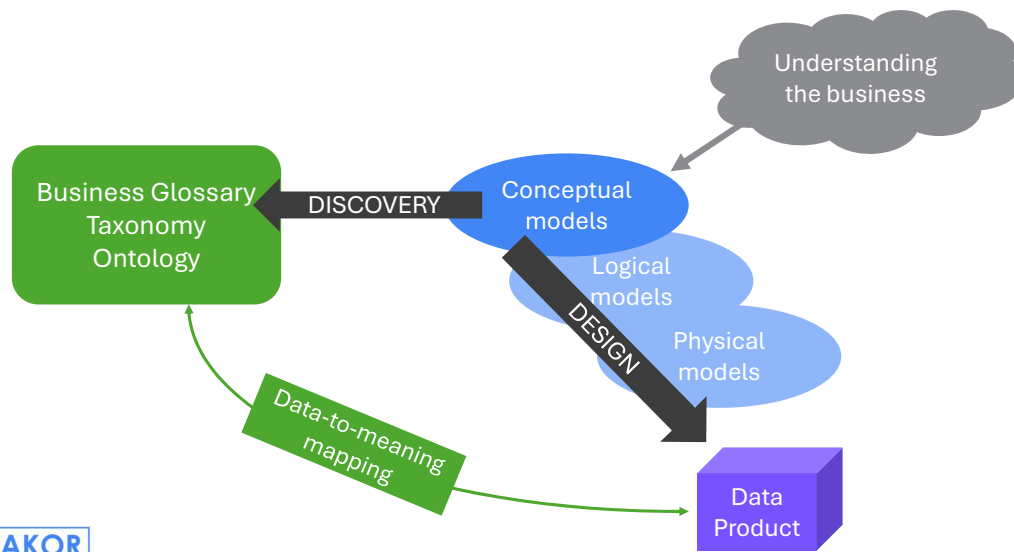


DATAKOR
CONSULTING

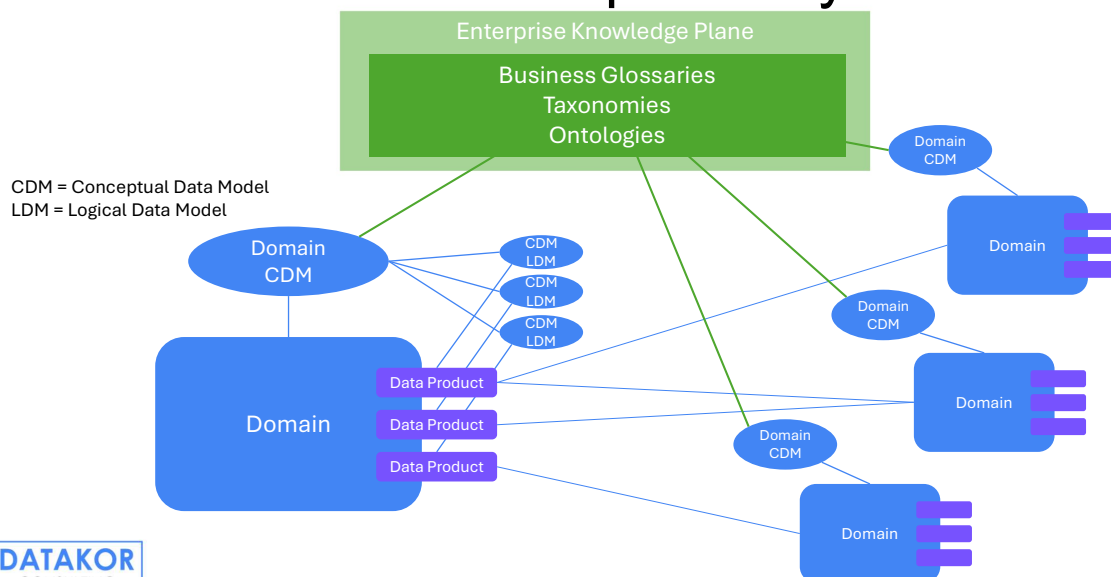
The Big Picture

Managing semantics across the enterprise

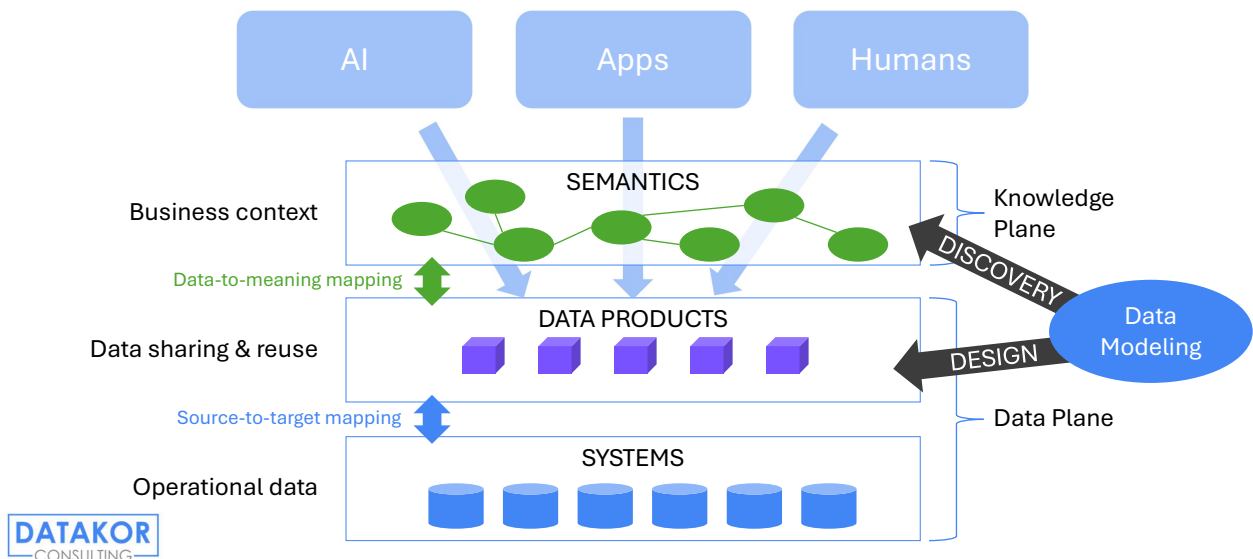
Design & discovery paths



Cross-domain interoperability



Goal: context-aware data utilization



Conclusions

Build interoperability with shared semantics

- Federated model requires common rules & shared understanding
- Data Products must be self-describing – metadata “wrapper”
- Capture semantics with Conceptual models, design product structures with Logical models
- Solve for both product-level delivery and enterprise-level consistency at the same time
- Enable context-aware data use by building a Knowledge Plane of shared semantics



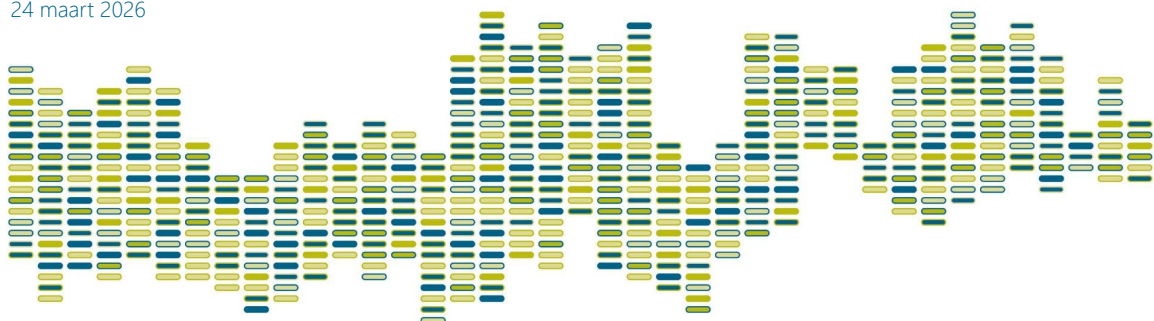
More thoughts on Substack!



Data administratie

Onmisbare basis voor effectief data management

Wouter van Aerle
24 maart 2026



deltiq
DATA MANAGEMENT

Waar dit over gaat

Oude wijn in nieuwe zakken



© 2026 Deltiq Group B.V. Alle rechten voorbehouden

2 **deltiq**
DATA MANAGEMENT

De Why...

Huib Modderkolk
@huibmodderkolk

De datadietstal via GGD-sy-
zoals de reactie zojuist in d
Jonge. Een aantal punten d

- GGD heeft zelf nog geen
lek
- Ook niet welke systemen
CoronIT en HPZone

Translate Tweet

4:00 PM · Jan 26, 2021 · Twitter Web

Bestuur

Markt en Overheid | Partner nieuws

Datagedreven werken prioriteit bij twee derde van financiële instellingen

Finance teams verliezen kostbare tijd aan het corrigeren van data.



Bestel Shutterstock

26 februari 2024 · Datagedreven werken, Financieel

Volgens 68 procent van de Nederlandse finance professionals speelt datagedreven werken een belangrijke rol in hun organisatie. Dat blijkt uit onderzoek van IT-dienstverlener en specialist in gegevensuitwisseling SureSync, onder 224 medewerkers van financiële instellingen. Toch blijkt het niet eenvoudig om data succesvol in te zetten. De grootste uitdagingen zijn datakwaliteit (28%), versnipperde systemen (21%) en het opzetten van gestandaardiseerde methodes (14%).

Datakwaliteit blijkt bottleneck

FOLLOW THE MONEY

ARTIKELN DOSSIERS AUTEURS PODCAST WOUTER

DIGITALE ZAKEN

Politie gebruikt on

In het nieuws 19

Invloedrijk algoritme reclassering doet al jaren foute voorspellingen

De reclassering gebeurt sinds 2008 met gebruik van algoritmen die het risico op recidive in te schatten.



Van de geschiedenis naar het heden: een algoritme dat de toekomst voorspelt. Het algoritme dat de reclassering doet, doet het al jaren. Het algoritme dat de reclassering doet, doet het al jaren. Het algoritme dat de reclassering doet, doet het al jaren.

[...] 'ernstige psychische problematiek'. Gegevens daarover ontbraken in de Nederlandse dataset [...] terwijl [...] juist een belangrijke voorspeller is"

en who...

RTL Nieuws

Gabriels leven werd een hel, door een foutje van de overheid: 'Ik ben kapot gemaakt'

Roel Schreinemachers · 31 maart 2024 · Aangepast: 2 april 2024

Een klein foutje van de overheid maakt het leven van Gabriel Malki al jaren tot een hel. Hij stond in het systeem als verdachte van een delict met harddrugs en heeft al jaren een hele reeks overheidsinstanties achter zich aan. Onterecht, maar het valt niet mee zijn naam te zuiveren. Toch lijkt dat nu te lukken.



Gabriel Malki.

Het is een taai gevecht dat nu al vier jaar duurt. "Het heeft heel veel pijn en moeite gekost." Maar Gabriel is vastbesloten om niet op te geven. Eindelijk begint zijn vasthoudendheid nu zijn vruchten af te werpen en gloort er licht aan het einde van de tunnel.

en when...

Motie #21

- 6 weken
- om uit te zoeken waar bepaald soort data wordt gebruikt.

- Kamerdebat van **19 januari 2021** over het rapport 'Ongekend Onrecht' van de Parlementaire Onderzoekingscommissie Kinderopvangtoeslag
- Motie Marijnissen (SP) c.s. om '**voor 1 maart 2021** een voorstel te doen hoe overal binnen overheidsinstellingen vervuilde data, risicomodellen en **het gebruik van nationaliteit** worden opgeruimd'.
- Aangevuld met verzoek Klaver (GroenLinks) c.s. om hierbij niet alleen te kijken naar nationaliteit, maar ook naar **etniciteit** en **geboorteplaats**

© 2026 Deltiq Group B.V. Alle rechten voorbehouden



Home > Documenten >

Kamerbrief over rapportage uitvoering motie 21

Kamerstuk | 06-07-2022 **18 maanden!**

Minister Yeşilgöz-Zegerius (JenV) stuurt de Tweede Kamer de rapportage over de uitvoering van motie 21 van de Tweede Kamerleden Marijnissen (SP) en Klaver (GroenLinks). De motie richtte zich op afkomstgerelateerde indicatoren in risicomodellen bij processen binnen het ministerie van Justitie en Veiligheid (JenV).

➔ [Kamerbrief over rapportage uitvoering motie 21](#) (PDF | 4 pagina's | 178 kB)

Bijlagen

➤ [Rapport uitvoering motie 21 door het ministerie van Justitie en Veiligheid](#)

Het rapport bevat een inventarisatie en beoordeling van het gebruik van afkomstgerelateerde indicatoren, in het bijzonder in ...

Rapport | 06-07-2022

© 2026 Deltiq Group B.V. Alle rechten voorbehouden

6



Gemene deler

Om data te (kunnen) managen
moet je om te beginnen weten om
welke data het gaat
en is inzicht vereist in wat de
eigenschappen van die data zijn.

Dit is geen technologieprobleem – dit is een administratieprobleem



© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Hoe nu wordt geadministreerd...



Datamodellen



Business glossary &
data definities



Microsoft Purview
Fysieke tabelstructuren



Data eigenaren &
data stewards



Bedrijfsobjecten-
model (BOM)



Gegevensleverings
overeenkomsten
(GLO's)



SharePoint
Datakwaliteitsissue-log

servicenow
AVG Verwerkingsregister



© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Wat initieel opvalt

Operationele knelpunten

Er valt *veel* te administreren

- Definities
- Rollen en verantwoordelijkheden
- Afspraken
- Data modellen
-

Meerdere 'systems-of-records'

- *Veelheid* aan *verschillende, op zichzelf staande* registratieve oplossingen
- *Formele* en *informele* systemen

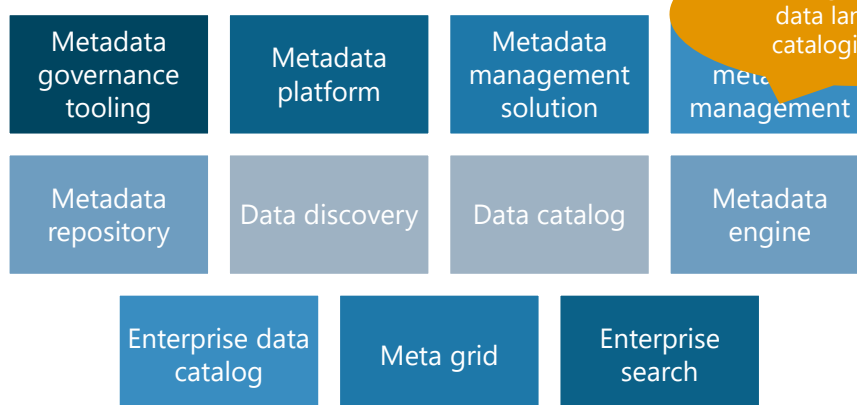
Registratie vaak achteraf

- Metadata corvee...

© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Maar we hebben toch metadata management?

Klopt, en dit is hoe de markt daarop inspeelt



© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Hoe dit soort technologie wordt toegepast...

is vaak niet voldoende

Technologisch gedreven

- Focus op vervangen losse oplossingen
- Veel tijd en aandacht gaat uit naar technische koppelingen

Te instrumenteel

- Vullen van de repository / data catalog staat centraal
- Niet het beheersen van concrete en specifieke data management-vraagstukken

"Build it and they will come"-misvatting

- Verwachting dat als de metadata er eenmaal is, iedereen het vanzelf gaat gebruiken
- Data Market-place anyone?

Beperkte implementatiescope

- Vaak primair in een analytics / BI / data science omgeving maar minder ter ondersteuning van primaire processen
- (meta)data uit secundaire systemen (data lake, data platform, data warehouse) maar minder uit operationele (bron)systemen

Dit is geen technologieprobleem

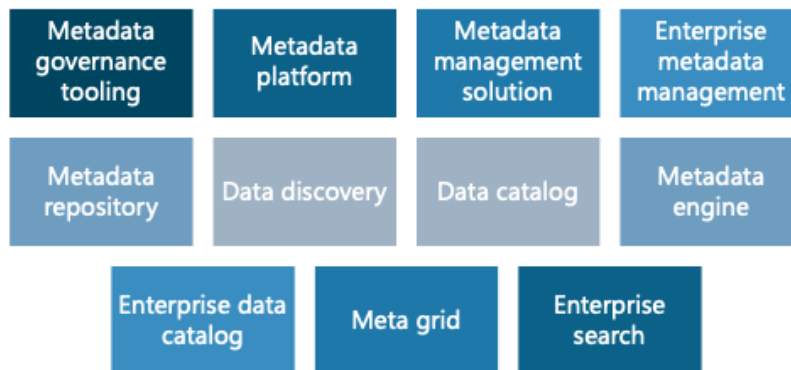
Maar een administratieprobleem

1. Door de **typering** van dit soort problematiek **als metadata management** komt bewust of onbewust de **nadruk** (teveel) te liggen op de **metadata zelf** en de **tools** die deze metadata beheren.
2. Dit leidt af van de **data management-vraagstukken** die we *met* deze metadata willen **beheersen**.
3. En de **aanvullende randvoorwaarden** die nodig zijn om dit mogelijk te maken.

Verbeterde benadering is nodig: *oude wijn in nieuwe zakken*

Data administratie als concept

Taal doet er toe!



In begrijpelijk Nederlands: **Data Administratie**
of **Gegevensboekhouding**

© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Data management als bedrijfsfunctie

Data als bedrijfsmiddel

"Wat SAP of PeopleSoft Financials is voor een CFO, is een Data Administratie voor de CDO."

Bedrijfsfunctie	Finance	Human Resources (HR)	Data Management
Eindverantwoordelijk	CFO	HR Directeur Directeur PZ	CDO
Plaats in de organisatie (afdeling, bedrijfsonderdeel)	Financiën Finance & Control Bedrijfsvoering	HR Personeelszaken Bedrijfsvoering	CDO Office
Kaderstelling	Accounting Manual P&C cyclus Budgetteringscyclus	HR Charter Functiebouwwerk Loon / salarishuis	Gegevensbeleid
Administratie	Financiële administratie Boekhouding	Personeelsadministratie	Data administratie Gegevensboekhouding
Centrale ondersteuning	Debiteuren / crediteuren Treasury	HR-advies Werving & selectie	Beleidsadvies Beheer data administratie
Decentrale uitvoering	Budgetverantwoordelijkheid Controlling	P-gesprekken Formatie	Data steward Gegevensverantwoordelijkheid

© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Essentie van boekhouden

***Boekhouding** is een systematische, veelal dagelijkse, en chronologische neerslag van feiten en gedragingen met een financiële impact waarmee ondernemingen, personen en (overheids)instellingen worden geconfronteerd, uitgedrukt in cijfers en meestal bijgehouden in een journaal.*

Metadata

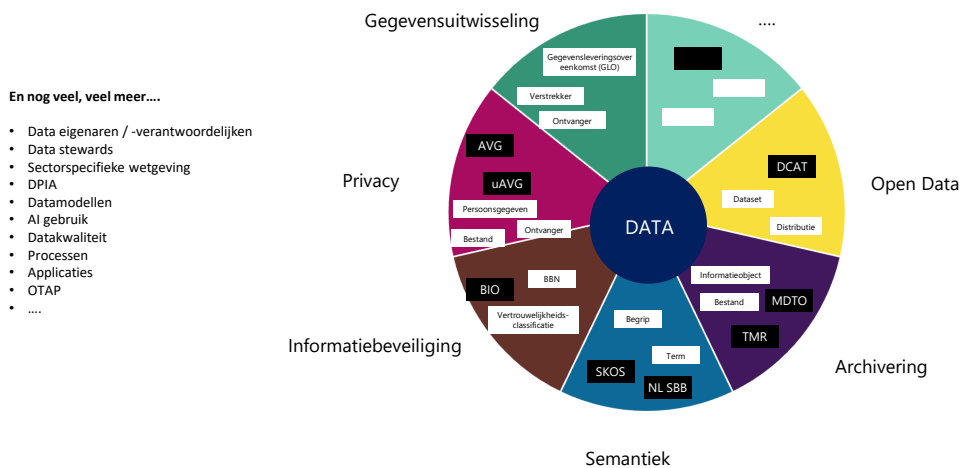
- Systematische neerslag
- Van feiten, gebeurtenissen en eigenschappen met betrekking tot data
- Die binnen een onderneming worden verwerkt en/of gebruikt
- Uitgedrukt in kwalitatieve of kwantitatieve zin
- En vastgelegd in een daartoe ingerichte administratie.

Data administratie

a.k.a. Gegevensboekhouding

is de *voorziening*
waarin de metadata integraal en consistent ligt opgeslagen
en toegankelijk is voor eenieder
die iets over de gegevenshuishouding wil weten
of vastleggen ten behoeve van beheersing of verantwoording.

Feiten, gebeurtenissen en eigenschappen



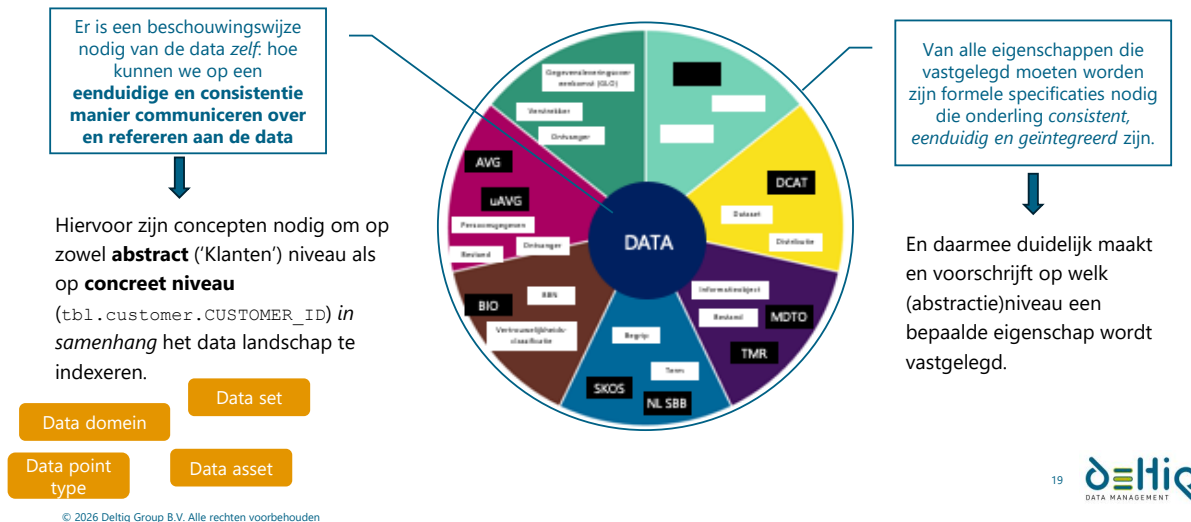
© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Operationele knelpunten

Nader geduid

1. **Feiten, gebeurtenissen en eigenschappen:** begrippenkaders sluiten niet altijd op elkaar aan, kennen verschillende niveaus van abstractie of ontbreken.
2. **Systematische neerslag:** ieder concern (Privacy, Informatiebeveiliging, Archivering etc.) kent zijn eigen systematiek en eigen implementatie. Vaak in de vorm van achteraf vastleggen in plaats van tijdens het ontstaan van de feiten zelf.
3. **Ingerichte administratie:** er bestaat geen integraal, consistent system-of-record. In plaats daarvan losse oplossingen met repositories, Excels, wikipagina's, Sharepoint-lijsten etc.

1. Feiten, gebeurtenissen en eigenschappen



Metamodel

'Rekeningschema' van de data administratie

De formele specificatie van de feiten, gebeurtenissen en eigenschappen van data (metadata).

Formele specificatie zorgt voor:

- Precieze definities
- Wegnemen van ambiguïteit
- Expliciet maken van eigenschappen

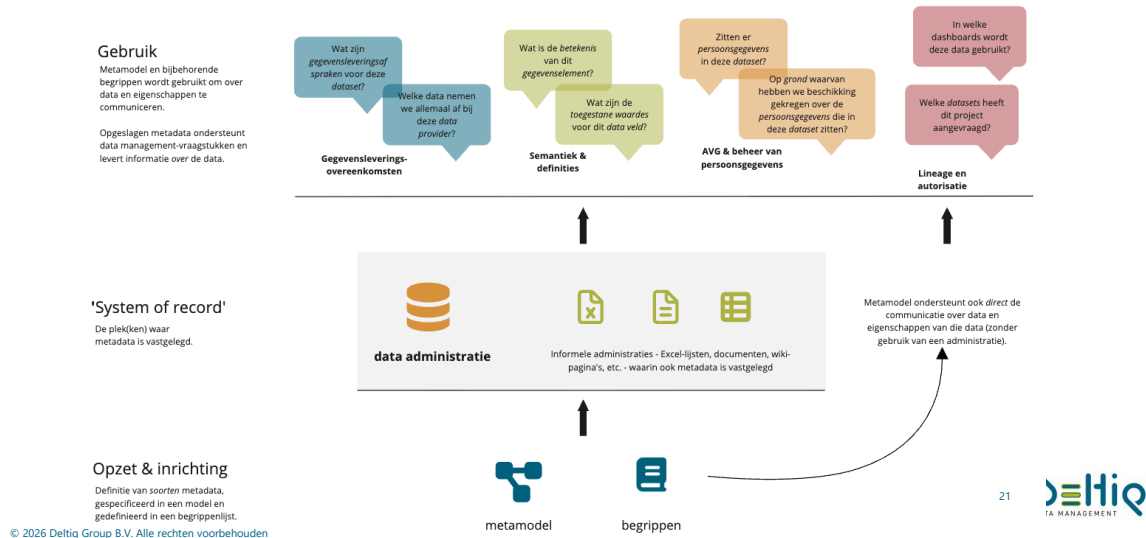
Resulteert in:

- Eenheid van taal
- Eenduidig en consistent referentiekader
- Ontwerp van data administratie

Reikwijdte metamodel verder dan alleen administreren

Taal doet ertoe!

Vormt de taal hoe we over data management communiceren



Metamodel is niet 'gratis'

Vraagt om structureel en systematisch beheer

- Bestaande standaarden niet altijd 1:1 implementeerbaar → vertaling en inpassing nodig.
- COTS metamodelen hebben ook beheer nodig...
 - Definities van de gebruikte concepten vaak niet heel sterk of formeel gedefinieerd
 - Beheermogelijkheden om het metamodel aan te passen / uit te breiden: "even een custom attribute toevoegen!"

2. Systematische neerslag

Hoe dit wel kan (en moet) werken

- **Bedenk:** veel metadata wordt al verzameld, alleen niet systematisch en gestructureerd vastgelegd.
- Koppel het administreren van metadata direct aan (repeterende) werkzaamheden waarin deze metadata ontstaat, bijvoorbeeld:
 - Uitvoeren van een DPIA
 - Dev/Ops sprint om nieuwe functionaliteit te ontwikkelen → neem mee in DoD
 - Afsluiten van een contract met een externe data provider
 - Incidentafhandeling & gebruikersvragen
- Organiseer administratieve / registratieve activiteiten expliciet
- Spreek af wie wat (niet) wanneer mag / moet administreren

Boekhoudregels

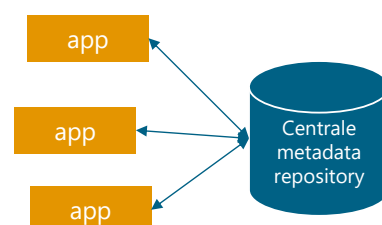
Organisatie & processen

23 **deltiq**
DATA MANAGEMENT

© 2026 Deltiq Group B.V. Alle rechten voorbehouden

3. Ingerichte administratie

Voorziening moet integrale, consistente vastlegging faciliteren



- Meerdere applicaties die metadata vastleggen, beheren en uitwisselen
- Standaard interfaces / koppelvlakken / API's omdat metamodel is gestandaardiseerd

24 **deltiq**
DATA MANAGEMENT

© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Handreikingen voor implementatie

Golden rules

en gebruik AI ☺☺

1. Administreer (alleen dat) wat je wil beheersen

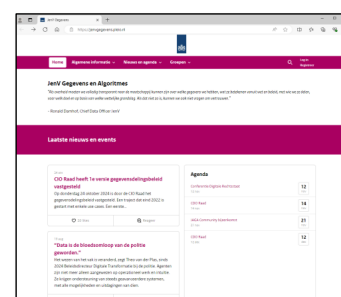
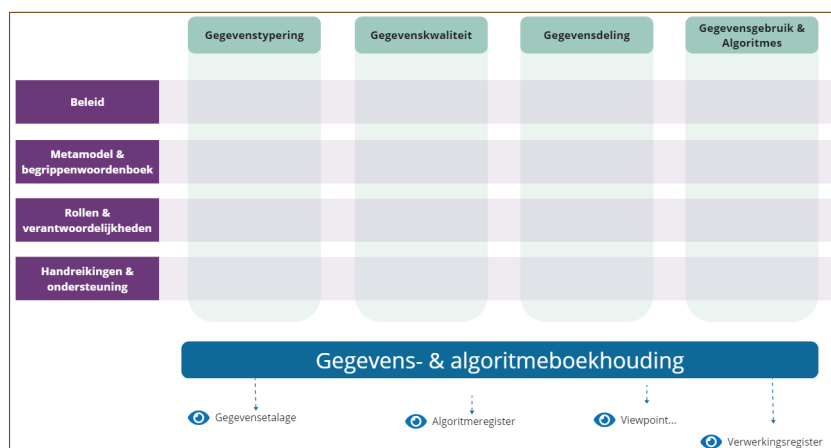
2. Zo abstract als het kan, zo gedetailleerd als het moet

3. Sluit zo naadloos aan bij hoe nu al gewerkt wordt.

4. Wees (technisch) creatief!

Concept in de praktijk

JenV Afsprakenstelsel Gegevens & Algoritmes (JAGA)



jenvgegevens.pleio.nl

Gegevensboekhouding JenV

Finalist Overheidsinnovatie 2025



"Hoe deze jury naar gegevensbeheer kijkt, is fundamenteel veranderd. Dit is een echte gegevensbeheer-revolutie! Een achterstand van ruim 40 jaar wordt hiermee ingelopen. De jury moedigt de collega's van de Gegevensboekhouding aan om met deze innovatie verder te gaan. De Gegevensboekhouding gaat verder dan een technische innovatie, het is een systeeminnovatie. Een verandering van hoe ambtenaren denken en werken. Als deze innovatie grootschalige toepassing krijgt, kan ze een echte systeemdoorbraak creëren."

Impressie

The screenshots illustrate the user interface of the Gegevensboekhouding JenV system. The leftmost screenshot shows a navigation menu with options like 'Home', 'Middeladministratie', and 'Typeringadministratie', along with a search bar and a list of data delivery protocols. The middle screenshot displays a detailed view of a specific record, including a title, description, and a list of associated data types. The rightmost screenshot shows a search results page with a table of records and a detailed view of a specific record, including a list of associated data types.

Vooruitzicht

Ontwikkelingen binnen de overheid

Rijksbrede Gegevensboekhoudstandaard

gegrondvest door een Wet Gegevensboekhouding...

á la Comptabiliteitswet | zie ook: <https://achterkantvandeoverheid.nl/avdo>

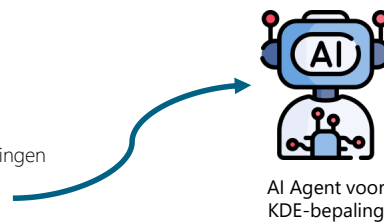
© 2026 Deltiq Group B.V. Alle rechten voorbehouden

29  DATA MANAGEMENT

Vooruitzicht

Mogelijkheden van AI

- **Backlog** in het gevuld krijgen van de data administratie
- **AI** is een **versneller** om content te creëren en administratieve taken uit handen te nemen:
 - Indexeren van gegevenslandschap
 - Classificeren van data
 - Afleiden van definities
 - Identificeren van persoonsgegevens
 - Valideren van wettelijke grondslagen voor gegevensverwerkingen
 - Bepalen van Kritieke Data Elementen (KDE's)
 -
- **MAAR** alle voorgaande randvoorwaarden blijven van kracht!
- En toegang tot eigen, interne content – inclusief data landschap – is nodig.



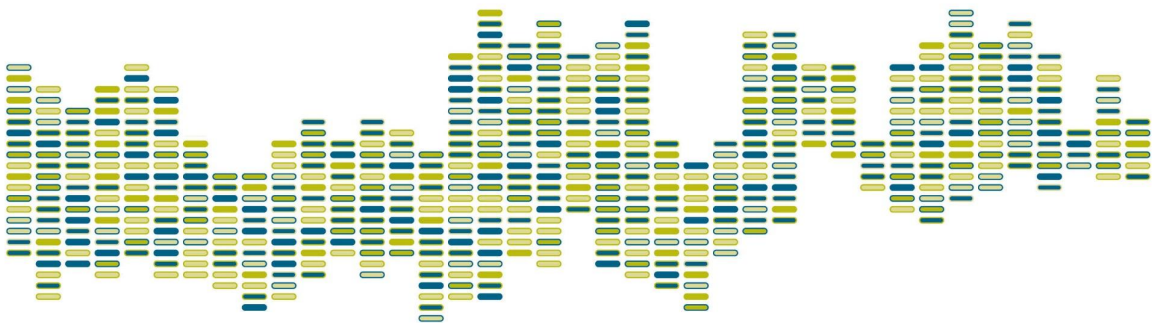
© 2026 Deltiq Group B.V. Alle rechten voorbehouden

30  DATA MANAGEMENT

DANK!

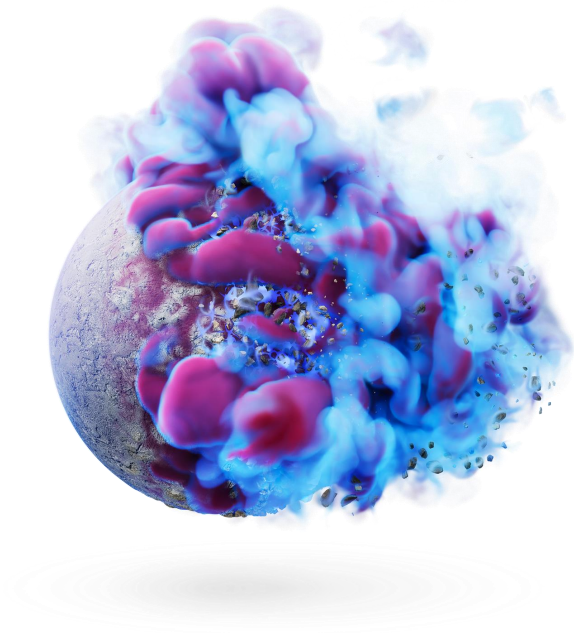
© 2026 Deltiq Group B.V. Alle rechten voorbehouden

Deltiq
Oud Aa 4b
3621 LB Breukelen
Nederland
www.deltiq.com
info@deltiq.com



New Times: From Data for AI to AI for Data

Rutger Rienks



1

 **Forbes**

INNOVATION > AI

Databricks Says AI Agents Now Build 80% Of Enterprise Databases

By **Victor Dey**, Contributor. © Victor Dey is an analyst and writer covering AI a... 

[Follow Author](#)

Published Feb 04, 2026, 10:22am EST, Updated Feb 04, 2026, 10:40pm EST

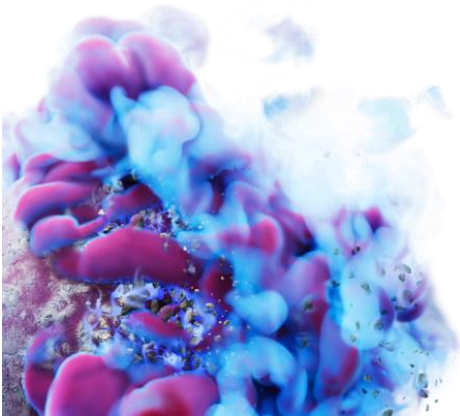


LOADING VIDEO PLAYER...

2

Contents

This presentation outlines the transition from preparing data for AI to using AI to manage data.



01 Market trends and needs

04 Interactive notebook examples

02 Gen- and Agentic AI

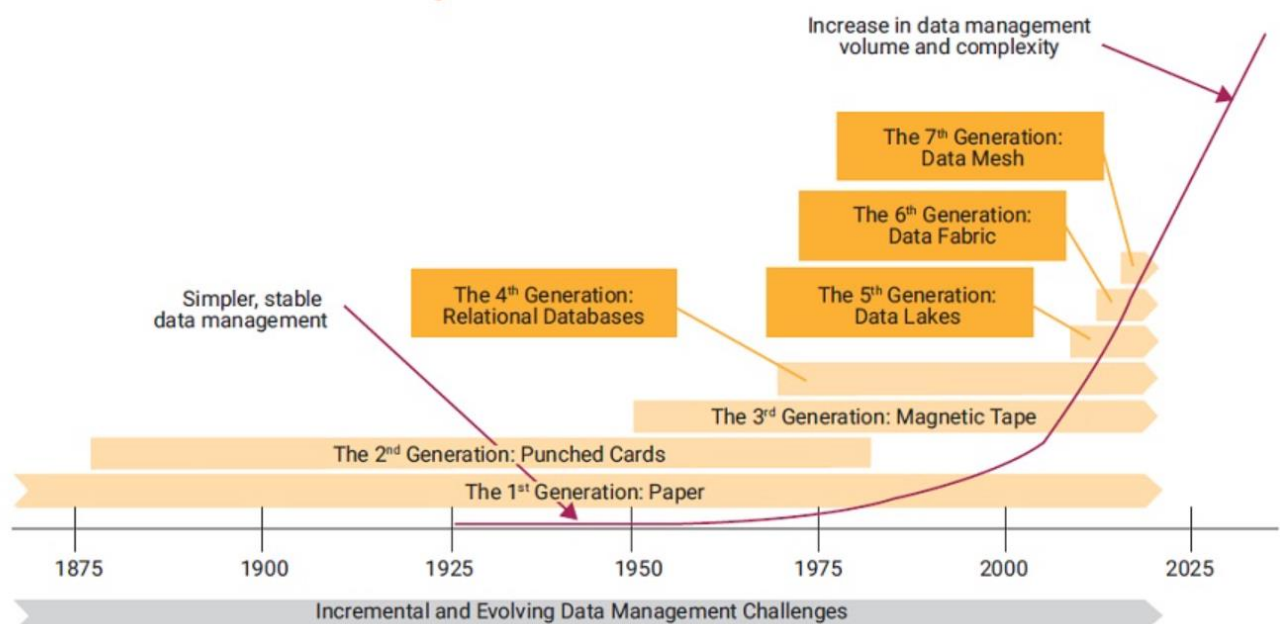
05 Tool overview

03 Five data-management tasks

06 Take home message

Copyright © Deloitte Consulting S.r.l. S.B. All rights reserved.

3



Data Cataloguing, Lineage, and Governance Challenges



Data volume and complexity is increasing at a pace that exceeds data governance resource bandwidth to support and govern

On-demand business analytics and decision-making needs and appetite for high-quality data are increasing

New technologies and capabilities, such as ML, AI, and GenAI, are generating more data while demanding more high-quality data

Data stewards who report to business functions are typically not in full-time roles and prioritize business activities over steward activities.

Data governance organizations are struggling to scale to demand and typically do not have the budget to increase headcount



Organizations should fully embrace automation to scale to demand and expedite Data Cataloguing, Lineage, and governance capabilities

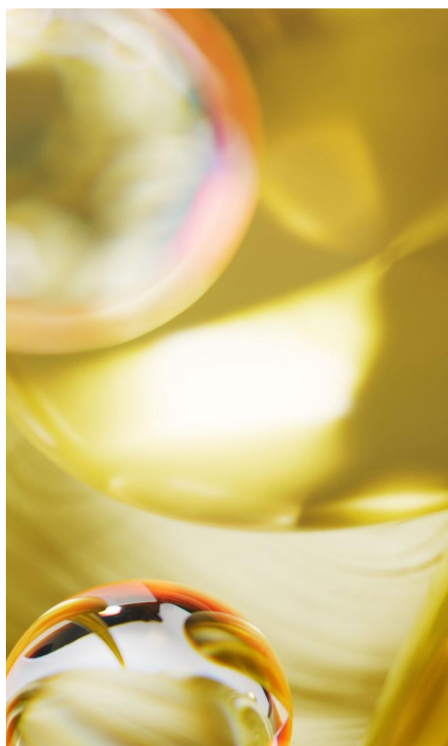


Organization should move towards centralized self-service data & analytics



Organization should build skills in AI/GenAI to integrate these technologies and capabilities into their data governance program

5



GENERATIVE AI

Traditional Generative AI and Large Language Models (LLMs) are **primarily focused on generating and interpreting content**. They excel in tasks such as writing texts, summarizing documents, translating languages, and answering questions based on the context provided. These models respond to specific prompts and perform a single, defined task. Their functionality is reactive; They wait for an input from the user and then produce an output.

AGENTIC AI

An AI agent can **act autonomously to achieve a goal**. This includes creating plans, calling external tools (such as APIs or databases), performing multi-step tasks, and learning from the results. Characteristic of Agentic AI is the use of a memory to retain context over longer interactions and the ability to make decisions independently. Instead of just responding, an agent can proactively initiate a series of actions to solve a complex problem, such as fully managing a data quality process.

6

Where does AI intersect with Data Management

The integration of AI and Gen AI into data cataloging, lineage, and governance processes offers numerous advantages, including improved accuracy, efficiency, transparency, and compliance. By leveraging these technologies, organizations can enhance their data management practices, making it easier to discover, understand, and govern their data assets effectively.



Data Governance

Includes data classification and data security; supports scalability, data governance, and access control, data encryption, and continuous monitoring, making data more manageable, secure, and useful



Data Quality

Ensures data accuracy, consistency, and reliability, which are crucial for effective decision-making and operational efficiency.



Data Lineage

Provides a comprehensive and accurate view of data flows and transformations, enabling better decision-making and risk management.



Metadata Management

Allows organizations to achieve more efficient, accurate, and scalable metadata management, ultimately leading to better data utilization and decision-making.



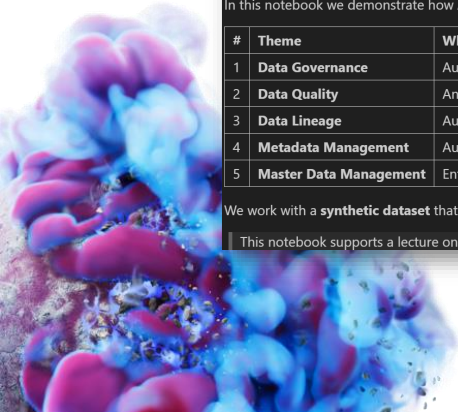
Master Data Management

Supports data enrichment, identifies duplicates, enforces policies, increasing operational efficiency and ensuring accurate and consistent master data

7

Let's make it Interactive!

Download the notebooks to follow along the presentation



AI for Data Management - Interactive Demo

From theory to practice: how AI strengthens the five pillars of data management

In this notebook we demonstrate how **Artificial Intelligence** can be applied to the five core themes of data management:

#	Theme	What do we show?
1	Data Governance	Automatic classification of sensitive data, PII detection, policy suggestions
2	Data Quality	Anomaly detection, fuzzy duplicate detection, automated quality scores
3	Data Lineage	Automatically track and visualize data transformations
4	Metadata Management	Auto-profiling, smart tagging, automatically generated data catalog
5	Master Data Management	Entity resolution, golden record creation, automated standardization

We work with a **synthetic dataset** that is intentionally messy - just like real-world data. Then we show how AI helps bring order.

This notebook supports a lecture on AI and Data Management.

<https://github.com/rutgerrienks/Datamanagement>

8



Data Governance

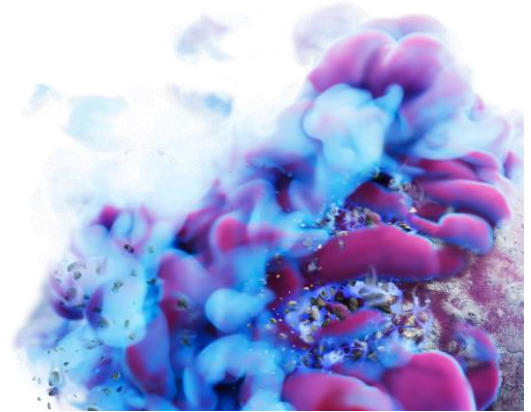
The framework for managing data assets within an organization to ensure data quality, security, integrity, and proper classification, thereby enabling effective data usage, compliance, and protection against unauthorized access.

Key Data Governance Capabilities:

-  **Data Classification Standards:** Tag data with labels such as Public, Confidential, and Personal Information.
-  **Access Control:** Ensuring users have data on a need-to-know basis and verifying identity and permissions before granting access
-  **Data Stewardship:** Responsibly managing and overseeing an organization's data assets to ensure data quality, integrity, security, and compliance

How AI/GenAI Enhances Data Governance:

-  **Automates Data Classification:** AI/GenAI can classify data and apply labels based on predefined criteria, reducing the time and effort required for manual classification and monitoring and classifying new data in real time
-  **Assists Data Stewards:** AI/GenAI can automate data management workflows and processes and enable proactive monitoring and compliance through advanced analytics



Data Lineage

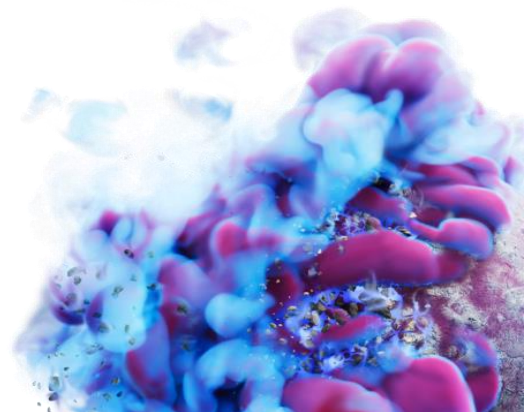
The lifecycle of data, encompassing its origins, movements, transformations, and ultimate destination. It provides a detailed map of how data flows through an organization, from its initial creation or acquisition to its final use

Key Data Lineage Capabilities:

-  **Data Lineage:** Ability to trace data from its origin to its destruction
-  **Data Compliance & Anomaly Detection:** Ensuring data meets regulatory compliance at each step of the data lifecycle, with AI-driven anomaly detection to mitigate risk and operational disruption.

How AI/GenAI Enhances Data Lineage:

-  **Contextual Data Lineage Visualization:** AI-powered tools offer customizable, interactive visualizations of data lineage, enhancing understanding and decision-making by illustrating data flows and transformations across the organization.
-  **Advanced Anomaly Detection:** AI algorithms can detect anomalies and outliers in data lineage, such as unexpected data transformations or unauthorized data access.





Metadata Management

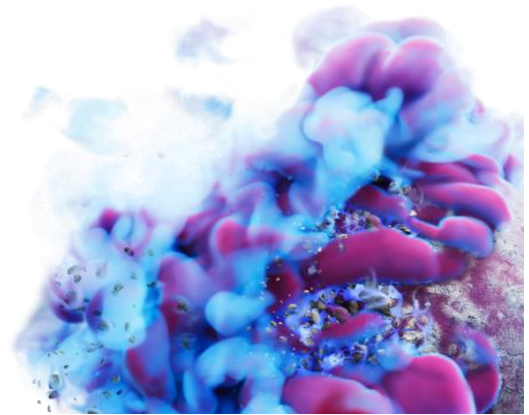
The capability to proactively manage the diverse data standards of the enterprise. It involves creating, controlling, enhancing, and maintaining consistency, quality, and usability of metadata

Key Metadata Management Capabilities:

-  **Data Discovery:** Facilitates the business' ability to quickly find data assets based on metadata attributes
-  **Data Collaboration:** Improves communication and understanding by providing context about data and establishing standards across data areas
-  **Impact Analysis:** Asses the potential effects of changes in data definitions or structures on downstream processes and systems

How AI/GenAI Enhances Metadata Management:

-  **Ease of Search:** AI supports search capabilities using natural language processing (NLP), giving users a simple way to locate data assets
-  **Predictive Impact Analysis:** AI can predict the effects of changes in metadata on downstream processes by analyzing historical data and identifying patterns



Data Quality

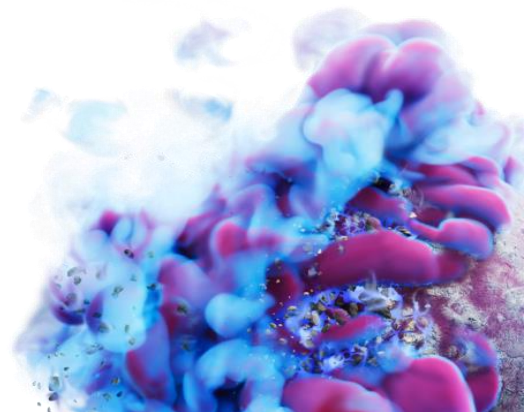
Data Quality (DQ) refers both to the characteristics of high-quality data and to the processes used to measure or improve the quality of data

Key Data Quality Capabilities:

-  **Data Profiling:** Identify potential data quality issues and develop solutions for remediation
-  **Data Cleansing:** Correct or remove inaccurate records from a dataset
-  **Data Quality Tracking:** Track the success and opportunities with metrics reflecting the progress of data quality efforts

How AI/GenAI Enhances Data Quality:

-  **Automate Data Cleansing and Data Preparation:** AI/GenAI can identify and correct errors in datasets. This could include duplicate detection/removal, error correction, and general standardization
-  **Data Quality Rule Creation:** Generative AI can help users write DQ rules so organizations can deploy and monitor data in a faster timeframe





Master Data Management

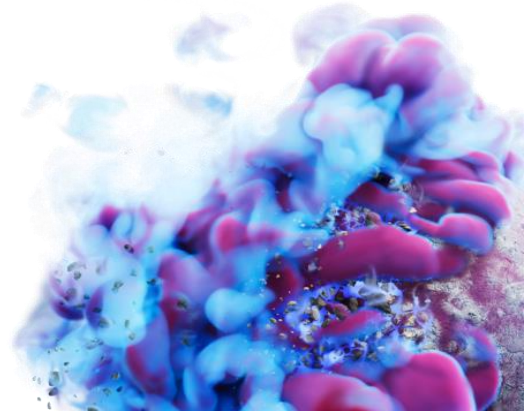
Master Data Management (MDM) is the comprehensive process of managing an organization's critical data to ensure consistency, accuracy, and accountability across the enterprise and creating a single source of truth.

Key Master Data Management Capabilities:

-  **Data Quality Standardization:** MDM facilitates the creation of a centralized and consistent definition of attributes, ensuring all data elements adhere to a consistent set of standards and dimensions (parsing, match/merge, profiling, enrichment)
-  **Master Data Lineage:** Ability to trace and map the origins, movement, transformations and usages of Master Data

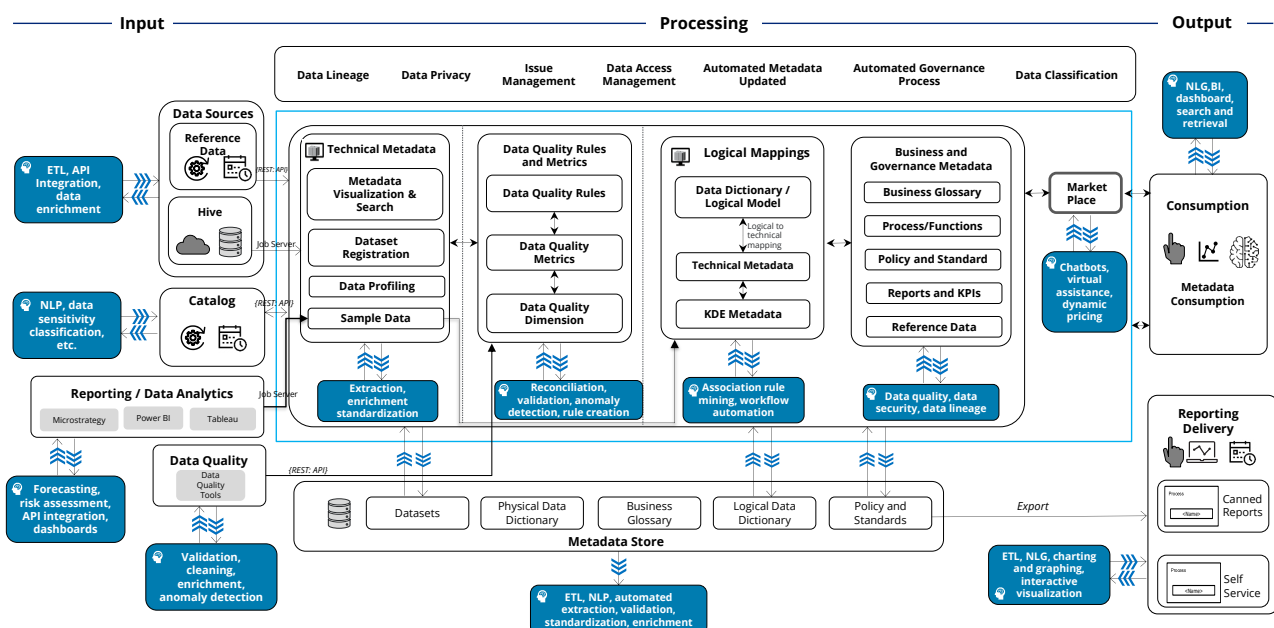
How AI/GenAI Enhances Master Data Management:

-  **Automated DQ Operations:** Built in AI can detect and remove duplicates, augment and enrich hierarchies with third-party data, check for consistency and correct missing data to prevent manual intervention.
-  **GenAI Powered Data Operations:** Gen/AI capabilities can help automate data tasks and create semantic drill downs of data for exploration and analytics



AI Agent Technical Conceptual Architecture

Indicates where the AI agent interacts with a specific part of the architecture.



Tool Overview

Different data governance tools have similar overall functions but offer various capabilities and features.

The commonly used tools listed below offer a comprehensive and collaborative platform for data governance. They help automate tasks such as data lineage, data cataloging, and data discovery. They are ideal and popular choices for organizations looking for a complete data management suite that ensures strong data quality, effective metadata management, and scalability.



Collibra's data governance platform offers automated data cataloging, end-to-end data lineage, robust data quality management, and policy enforcement. It supports role-based access, stewardship workflows, and provides collaboration tools and a centralized business glossary. With customizable dashboards and integration capabilities, Collibra ensures data accuracy, compliance, and efficient decision-making across the organization.



Alation's data governance capabilities include automated data cataloging, data lineage tracking, and data stewardship workflows. It provides robust metadata management, policy enforcement, and role-based access control. Alation also offers collaboration tools, a business glossary, and data quality management features to ensure data accuracy, compliance, and effective decision-making.



Informatica's data governance capabilities include automated data discovery, comprehensive metadata management, and robust data quality tools. It offers end-to-end data lineage, policy enforcement, and role-based access control. The platform supports collaboration through a business glossary and stewardship workflows, ensuring data accuracy, compliance, and effective decision-making across the organization.

15

Tool Overview

●●●/●●○/●○○ = High / Medium / Low Maturity

Use Case	Collibra	Alation	Informatica	AI Capabilities for Each Tool
Metadata Management	●●○	●●○	●●●	Collibra: Supports automated data cataloging, automated asset curation, and NLP search capabilities. Alation: Uses Behavior I/O, an AI-powered engine to automatically identify meaningful patterns about how data is being used, then giving users recommendations. Informatica: Uses active metadata management to create predictive recommendations, supports automated data cataloging and NLP search.
Data Quality	●●○	●○○	●●●	Collibra: Has AI/GenAI-generated AdaptiveRules, supporting automated DQ tracking, and GenAI supported DQ rule creation using NLP. Alation: Integrates Snowflake's Data Quality Metric Functions (DMFs) with its data quality framework, creating high-quality data for AI model development. Informatica: Has CLAIRE AI copilot, which supports automated data profiling, quality analysis, and can generate data quality rules.
Data Lineage	●●●	●●●	●●●	Collibra: Has intelligent filtering through lineage diagrams and automated end-to-end data lineage tracking with Collibra Lineage. Alation: Leverages machine learning algorithms analyze user behavior, data usage patterns, and metadata to identify relevant data assets. Informatica: Supports automated end-to-end data lineage tracking, including automated scanning for data transformation details.
Data Classification	●●●	●●●	●●●	Collibra: Can automatically classify data with industry-specific context Alation: Supports automated mapping and categorization of data, including tagging Informatica: CLAIRE supports intelligent similarity detection and domain discovery with tags, streamlining classification
Data Security	●●○	●●○	●●●	Collibra: Automatically discovers and classifies sensitive data across existing organizations' AI systems, allowing owners to manage access controls Alation: Alation AI and data governance framework flags non-compliant datasets and ensures AI models are built on reliable, governed data Informatica: CLAIRE has data privacy solutions that enable sensitive data tracking, detect anomalies, and simulate risk
Master Data Management			●●●	Collibra: N/A Alation: N/A Informatica: Automatically identifies master data, traces source lineage, and finds mappings between attributes in data models

The three main conclusions

KEY MESSAGES SUMMARIZED

The focus is shifting from manual data preparation to leveraging AI for autonomous data management

Agentic AI significantly increases efficiency by deploying autonomous agents to enrich, validate, and integrate data.

Strong governance and Human-in-the-Loop controls are essential to manage risk and ensure the reliability of the AI.



Building Sovereignty A Two-Year Journey to a Scalable Data & AI Platform

2026, March

Antoine Stelma – Co-founder @ Connected Data Group

● WE LIVE DATA

CONNECTED
DATA GROUP

AGENDA

- Introduction
- The Vision & The Why
- The Infrastructure Challenges
- The Blueprint & Solution Components
- Lessons Learned & Conclusion

● WE LIVE DATA

CONNECTED
DATA GROUP

ANTOINE STELMA

•**20+ Years of Data Mastery:** A seasoned Data Architect and trainer with over two decades of experience in Data Warehousing and Business Intelligence.

•**Co-Founder & Visionary:** Built Connected Data Group alongside Erik Fransen to provide organizations with smarter, sovereign Data & Analytic architectures.

•**Specialist in Scalability:** Deep expertise in Data Vault Modeling, Data Virtualization, and the blueprinting of modern, high-performance data platforms.

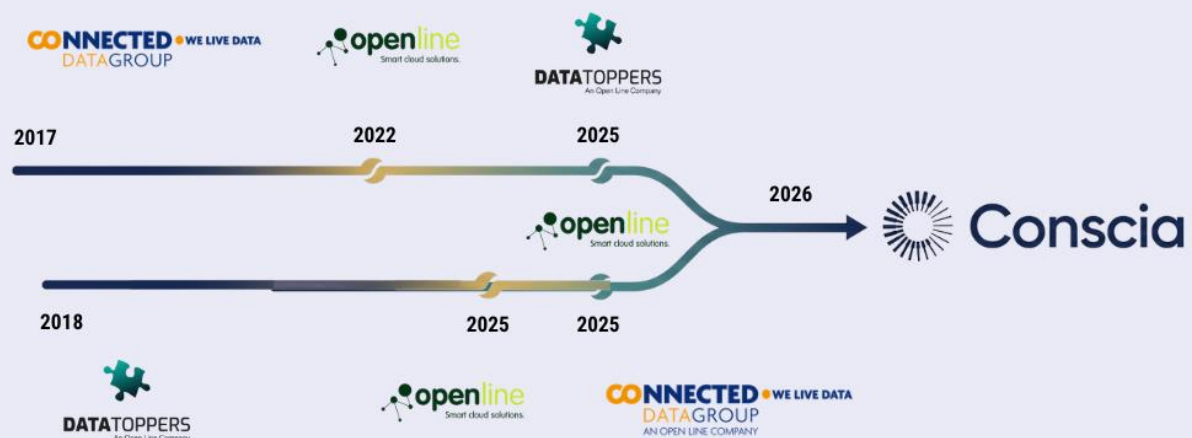
•**Global Educator:** Translating complex theory into practice as an instructor for the Connected Data Academy.

Real-World Grounding: Passionate about skipping the "vendor fluff" to share best practices rooted in extensive research and daily, field-tested use cases.



● WE LIVE DATA

CONNECTED
DATA GROUP



● WE LIVE DATA

CONNECTED
DATA GROUP



CONNECTED
DATA GROUP
● WE LIVE DATA

CONNECTED
DATA CLOUD
● WE EMPOWER DATA

CONNECTED
DATA ACADEMY
● WE TEACH DATA

● WE LIVE DATA

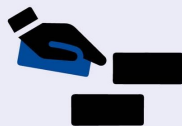
CONNECTED
DATA GROUP

WHAT IS OUR RECIPE

TOGETHER WE CREATE
VALUE FROM DATA



GRIP
ORGANIZING DATA



ANALYTICS PLATFORM
SOLUTION AS A SERVICE



INSIGHTS
ADDING VALUE



KNOWLEDGE
CONNECTED DATA ACADEMY

● WE LIVE DATA

CONNECTED
DATA GROUP

The Vision & The Why

● WE LIVE DATA

CONNECTED
DATA GROUP

The Ambitious Question

Why should we build a truly scalable sovereign platform?

- Escape from Vendor/Cloud lock-in
- Compliance as a Competitive Advantage
- Standardize with best-of-breed solutions
- Control of cost
- AI readiness & Intellectual Property Protection
- Agility through Hybrid Flexibility

● WE LIVE DATA

CONNECTED
DATA GROUP

A definition of sovereignty

True Sovereignty is the **strategic capability** of an organization to maintain **absolute authority and self-determination** over its **digital landscape** while navigating complex regulatory and technical environments.

Three core ownership pillars:

- Ownership of Logic
- Ownership of Data
- Ownership of Solutions

The Market Gap

Why existing off-the-shelf solutions weren't enough.

- The "Black Box" Logic Trap
- Mandatory Vendor Lock-in (The "Cloud Hotel California")
 - Most modern SaaS data platforms are "hotel" architectures:

 'You can check out every time you like. But you can never leave'

- The Trap of Unpredictable Cost

The vision

“Create a scalable and flexible Data & AI platform that is Cloud Agnostic, based on best of breed solutions and provides full control over data, logic and solutions.”

“In our vision, data, logic and AI models are transparent and move with the business, not the vendor.”

“Data Governance & Data Management are continuing processes and not technical solutions. How can we support the process?”

● WE LIVE DATA

CONNECTED
DATA GROUP

The Infrastructure Challenges

● WE LIVE DATA

CONNECTED
DATA GROUP

Challenge 1: Scalability

Kubernetes: the backbone of the operation.

(non-technical explanation)

- Each application is a container
- Data & AI Platform is the cargo
- Each customer has its own private and secured dock
- Kubernetes is the harbour master
- Customer owns the containers and the private dock

CONNECTED
DATA GROUP



Challenge 2: Security

Moving beyond the perimeter
—security in a sovereign environment.

(Just some of the features)

- Private Endpoint Connectivity only
- Context-Aware Access Rules per customer
- Advanced Kubernetes Protection
- Security Matrix for users and usage
- Our Security Operation Centre

Security by design is NOT only the platform but starts with the governance of the data, logic and applications

CONNECTED
DATA GROUP



Challenge 3: Governance & Compliance

How to bake Governance & Compliance into the platform's DNA.

- Automated Data Lineage
- Metadata-Driven Guardrails
- Immutable audit trails
- Standardized Data modelling
- Proactive Risk Monitoring
- Lifecycle Management
- Purpose Limitation
- Data Minimisation
- Track every data movement, 100% transparent
- Using a Datacatalog to enforce policies
- Log every action taken by users and processes
- Data Vault as our core standard
- Security Operation Centre
- How long may we use the data?
- When it's allowed to use this data?
- Do we really need this data here also?

Challenge 4: The Cloud Dilemma

Is there a Cloud Dilemma or should we navigate better?

- Our Data & AI platform is entirely cloud agnostic.
- Public Cloud for Scalability is an option
- Private Cloud for sovereign data for sensitive data and core business logic
- Hybrid as best of both worlds is always possible

But remember the Cloud hotel California:
It's only a good hotel if you check out and really leave.

Challenge 5: The Tooling Balance



The tug-of-war between Open-Source & Closed-Source application

● WE LIVE DATA

CONNECTED
DATA GROUP

Challenge 5: The Tooling Balance

- **Closed Source**
 - Do we need support from the supplier?
 - Exit scenario always results in readable code, logic and data for fast migrations
 - Operational Efficiency versus the cost?
- **Open Source**
 - Can we support the application?
 - Exit scenario always results in readable code, logic and data for fast migrations
 - Is there an active community that we can support?

● WE LIVE DATA

CONNECTED
DATA GROUP

Challenge 6: Delivery Models

Software as a Service versus Product as a Service

SAAS

Customer uses the Data and AI platform as a service.

PAAS

More customers are now using the dataproducts that are generated with our platform.

We provide housing corporations with **VERA-based sovereign data warehouses** as a fully managed service. This includes 150 KPIs and 7 CORA reports to satisfy **mandatory government reporting**, alongside daily data loading, proactive maintenance, and regular updates.

Both models use the same data and ai platform

● WE LIVE DATA

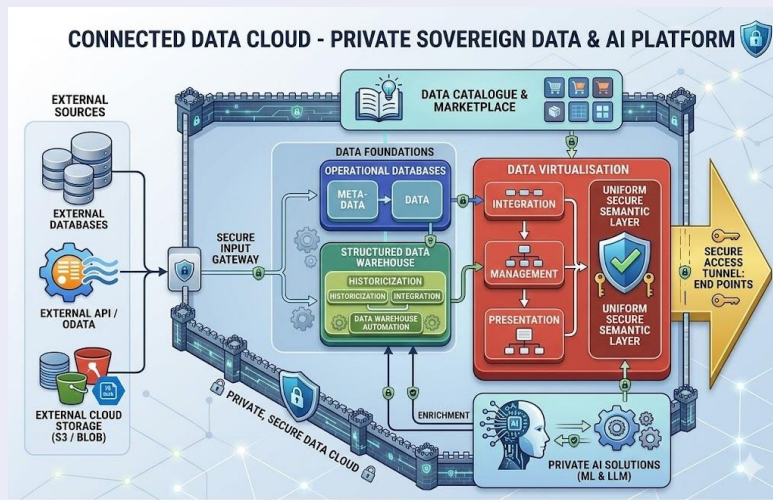
CONNECTED
DATA GROUP

The Blueprint & Solution Components

● WE LIVE DATA

CONNECTED
DATA GROUP

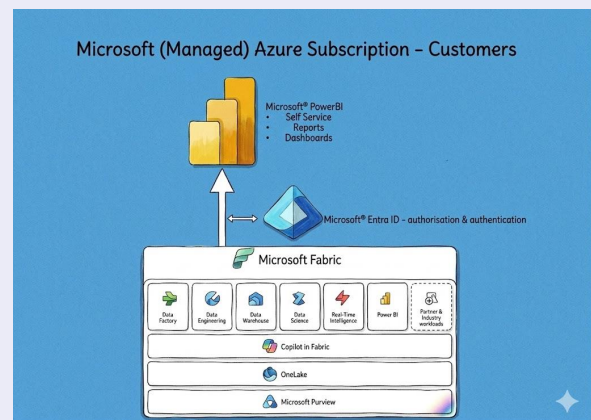
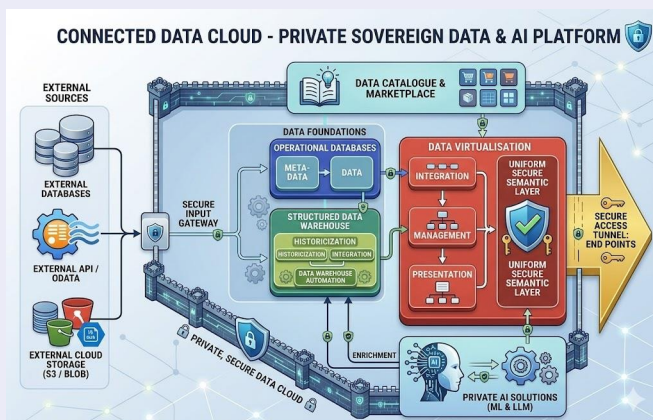
The Big Picture - Sovereign



WE LIVE DATA

CONNECTED
DATA GROUP

The Big Picture - Hybrid



An architectural diagram of the resulting platform hybride with MS Fabric.

WE LIVE DATA

CONNECTED
DATA GROUP

Component 1: Data Warehouse Automation

Our demands

- Automation Excellence
- Native Elasticity
- Continuous Quality
- Progressive Ecosystem
- Performance Driven ROI
- Deep Interoperability
- Standardized Modelling
- Commercial Safety
- Proven track records in automating Data Vaults
- Able to run as a scalable container
- Delivery of code and data products
- Active supplier of community to support our demands
- License cost should match the business case for saving
- Integrate with any source
- Standardization to enhance auditability and historical integrity
- For closed-source a comprehensive SAAS license must be available

All products create with the automation solution must be 100% transparent.

● WE LIVE DATA

CONNECTED
DATA GROUP

Component 1: Data Warehouse Automation

Our choice

With an impressive background in Swiss banking and a strong Data Vault track record, we choose Datavault Builder as our strategic partner.



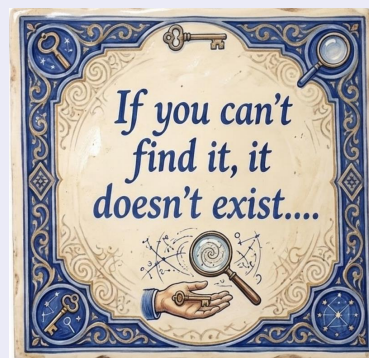
● WE LIVE DATA

CONNECTED
DATA GROUP

Component 2: The Data Catalog

Our demands

- Align with enterprise Data Governance standards
- Automate the ingestion of metadata
- Enable data quality observability
- Provide comprehensive support for data domains
- Maintain a centralized metadata glossary
- Implement automated data tagging by subject area
- Support fine-grained, role-based access control
- Streamline workflows for resolving data quality issues



● WE LIVE DATA

CONNECTED
DATA GROUP

Component 2: The Data Catalog

Our choice

Open Metadata (Open Source)

- Active Community Support
- Comprehensive Data Management
 - Seamlessly handle Discovery, Lineage, Observability, Quality, Collaboration, and Governance.
- Centralized Metadata
 - All data is stored in a Postgres database, ready for Reporting and AI integration.



● WE LIVE DATA

CONNECTED
DATA GROUP

Component 3: Data Virtualization

The "Magic Trick"—accessing data without moving it.

Our demands:

- Zero-Movement Access: Enable real-time data access without costly migrations
- Minimized Redundancy: Eliminate data duplication to streamline storage
- Unified Logical View: Speed up exploration and time-to-market via a single interface
- Virtual Agility: Accelerate development using flexible virtual data models
- Direct Governance: Centralize security by applying policies directly at the data source
- AI Acceleration: Securely fast-track AI use cases through unified data exploration

● WE LIVE DATA

CONNECTED
DATA GROUP

Component 3: Data Virtualization

Our choice

Denodo Data Virtualization



- Dedicated supplier
- Fully audible event logs
- Data Marketplace for direct integration with generative AI
- Proven track records in the Virtual Data Management market
- Transparent backend with readable code for exit scenario's
- Native scaling via container technology
- Excellent support in the Netherlands including a Dutch User Group

● WE LIVE DATA

CONNECTED
DATA GROUP

Component 4: Databases and the Data Lake

The foundation of the storage layer.

Our demands:

- Standardize on PostgreSQL for all organizational database management
- Utilize JSONB for high-performance object storage in a relational model
- Leverage native geospatial support for location-based analysis
- Ensure ACID compliance for data integrity across critical workloads
- Optimize performance through advanced indexing and execution plans
- Secure sensitive data with Row-Level Security (RLS) and enterprise controls
- Reliable Metadata Store: Balance high performance with robust backend stability
- Containerize deployments to increase infrastructure agility

● WE LIVE DATA

CONNECTED
DATA GROUP

Component 4: Databases and the Data Lake

Our choice

Postgres and Single Store

SingleStore: Real-Time Performance

- High-Speed Analytics for real-time analytical requirements
- Hybrid Data Models for SQL and NoSQL functionalities
- In-Memory Processing: Utilizes in-memory capabilities for low-latency operations
- Flexible Storage Architecture: Supports both row-based and column-based storage formats
- Multimodal Capabilities: A single database solution for time series, full-text search, relational, and spatial data



SingleStore

● WE LIVE DATA

CONNECTED
DATA GROUP

Component 5: AI solutions

Bring the models to the data, not vice versa

Our demands:

- Bring the AI to the Data:
 - Developing private ML, Generative & Agentic applications
- Dutch Sovereignty:
 - All solutions are hosted and secured within our local Dutch datacentres
- Security First:
 - A dedicated Data & AI platform built for safety and customer trust

Component 5: AI solutions

Our journey sofar

- Empowering Research by supporting our customers Data Science team with our solutions
- Specialized Knowledge Systems: Developing a high-performance RAG application for complex document knowledge retrieval
- Multi-Model Versatility: Leveraging the Ollama framework to deploy and manage a diverse range of Large Language Models (LLMs) and Vision Language Models (VLMs)
- Cutting-Edge Integration: Engineering Model Context Protocols to bridge Generative AI with Data Vault architectures
- Continuous Innovation: Cultivating a culture of daily learning and technical growth

Lessons Learned & Conclusion

● WE LIVE DATA

CONNECTED
DATA GROUP

Lessons Learned

What we got right - The "Aha!" moments that validated the 2-year journey

- **Strategic Partnerships:** Identified sponsors who provided both organizational support and the expertise needed to transition our platform into a managed service
- **Conceptual Alignment:** Assembled a multi-disciplinary expert team to ensure technical capabilities are always aligned with core business functionalities
- **Methodology as a Product:** While technical specs are vital, our greatest success is embedding our methodology and best practices directly into the platform
- **Outcome-Driven Focus:** Prioritizing the "human" side of the platform—methodology and best practices—as the ultimate driver of customer success

● WE LIVE DATA

CONNECTED
DATA GROUP

Lessons Learned

What We Got Wrong. Honest reflections on the hurdles that slowed the team down

- **The Scalability Gap:** We waited too long to formalize knowledge transfer, which created a bottleneck for our Managed Service transition
- **The Evolution:** Moving from a stand-alone **solution** to a full **ecosystem** proved to be a more complex journey than anticipated
- **The 100% Rule:** While we aimed for perfection, our customers were actually willing to partner with us and grow alongside the product
- **Partnership vs. Product:** Real-world feedback during the "journey" was more valuable than a delayed, perfect launch

Conclusions

3 Key Takeaways

- **Sovereignty is Strategy** True digital sovereignty isn't a buzzword; it's the strategic capability to maintain **absolute authority** over your digital landscape
- **Flexibility Over "Black Boxes"** Avoid being locked into one cloud. Use a **cloud-agnostic blueprint** that blends Public Cloud scalability with Private Cloud security
- **Methodology Drives Success** Value comes from the platform's **governance and standardization**, not just the technical specs or the data itself

Question & Answers

● WE LIVE DATA

CONNECTED
DATA GROUP

CONNECTED
DATA GROUP
● WE LIVE DATA

Concept Modelling with Normal People – Five Key Lessons From 45 Years of Modelling

45 in 55

Adept Events

Data Warehousing & Business Intelligence Summit

24 March 2026 in Utrecht NL

Alec Sharp

Senior Consultant

Clariteq Systems Consulting Ltd.

West Vancouver, BC, Canada

asharp@clariteq.com

www.clariteq.com

© Copyright Alec Sharp / Clariteq 2026

AdeptEvents

DATA WAREHOUSING &
BUSINESS INTELLIGENCE SUMMIT

CLARITEQ

1

Alec's background...

Alec Sharp, Clariteq Systems Consulting – asharp@clariteq.com

- 45 years global experience as an independent consultant:

- Data Modelling and Management** My roots!

- Business Process Modelling & Business Process Change – discover, scope, analyse, and design/redesign processes
- Application Requirements Specification
- Facilitation & Organisational Change
- Project Recovery

Process Business Process Modelling

Application

Use Case Modelling

Service Specification

Data

Concept Modelling

- Awarded DAMA's global Professional Achievement Award for contributions to "human-friendly" data modelling
- Author of "Workflow Modeling"
 - best-selling book on process modelling & improvement
 - second edition – a complete re-write

Check out the nice reviews
on Amazon - <http://amzn.to/dHun1o>



2

Clariteq – small, husband & wife company, global clients

ABB (ASEA Brown Boveri)
Aflac
American Honda
AMP (Australia Mutual Provident)
BackOffice Associates
Bank of Finland
Bellrock
Booking.com
Brisbane City Council (Australia)
Canadian Natural Resources Ltd.
City of Seattle
Civica UK
Clearwater Paper
Corvias
Dell
DHL Express
Dutch National Bank
Elisa
Ericsson
Essity
Eurojust (European Justice Comm.)
European Central Bank
Fortum
Gofore
Helse Vest - Norway
HM Land Registry - UK
Home Depot
Idaho Transportation Dept.
Intel
ISO New England

ING Bank
JP Morgan
Kal Tire
KONE
LGM Financial Services
Liberty Mutual
Livestock Improvement Corp.
MacDonald Dettwiler
Manitoba Public Insurance
Marathon Pipe Line
Microsoft
Ministry of Defence - UK
Ministry of Defence - NL
Ministry of the Interior - Slovakia
MTS Allstream
Nexen
Novo Nordisk
Nusenda Credit Union
OP Bank
Partner Reinsurance
Ritchie Brothers
Phillip Morris
Roche Diagnostics/Pharmaceuticals
Salt River Project
Saudi Aramco
Serco
Shell
Sparta Consulting
State Street Bank
SunGard

SVB (NL)
Synchro
Sysdoc
Talent Base
Teck
The MUSIC Group
The Seattle Times
UK Government
University Med Ctr Groningen
YIT(FI)
Washington Gas & Light

– Higher Education –
Carnegie Mellon University
Cornell University
Douglas College
Gonzaga University
Humboldt State University
The Jackson Laboratory
The Ohio State University
Portland State University
Salt Lake Community College
Southern NH University
University of Arkansas
University of British Columbia
University of the Fraser Valley
University of Maryland
University of Utah
University of Washington
Utah Valley University



© Clariteq 2023

www.clariteq.com

What we'll cover...

★ Five Key Points

1. Start with a nice conversation
2. Understand the "what" vs. "who and how" distinction
3. Patience is a virtue (so are Simplicity and Consistency)
4. Be fearless yet vulnerable - you don't know it all
5. Graphical design principles matter
6. and along the way...
we can use Concept Models for so much more

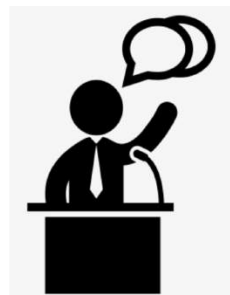
The presentation slides provide a structure, but mostly I plan to *tell stories*.

1 – Start with a conversation

Don't begin with a lecture on data modelling

"Before we begin our data modelling session, let's go over some key points about data modelling. First, *an Entity is any uniquely identifiable person, place, thing, event, concept, or organisation of interest to the enterprise about which facts may be recorded.* Any questions? I didn't think so..."

"Before I begin my speech, let's cover a few of the basic rules of grammar. *A noun is any...*"



If you **avoid** starting with the theory and practice...

Modelling sessions go better

Allows use of concept modelling in non-typical situations

If you can, don't even call it "data modelling"



5

Painful but useful learning experience

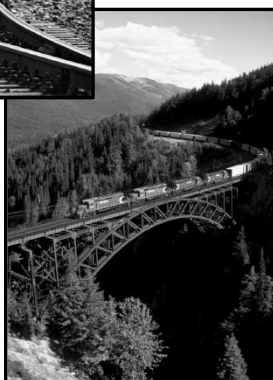
The assignment –
facilitating a Concept Modelling
session for a railway's
Track & Structures group



I began by explaining
data modelling...
"An entity is a uniquely
identifiable person, place,
thing, event, ..."

Bad idea!!!

"I can't stand you IT guys!"

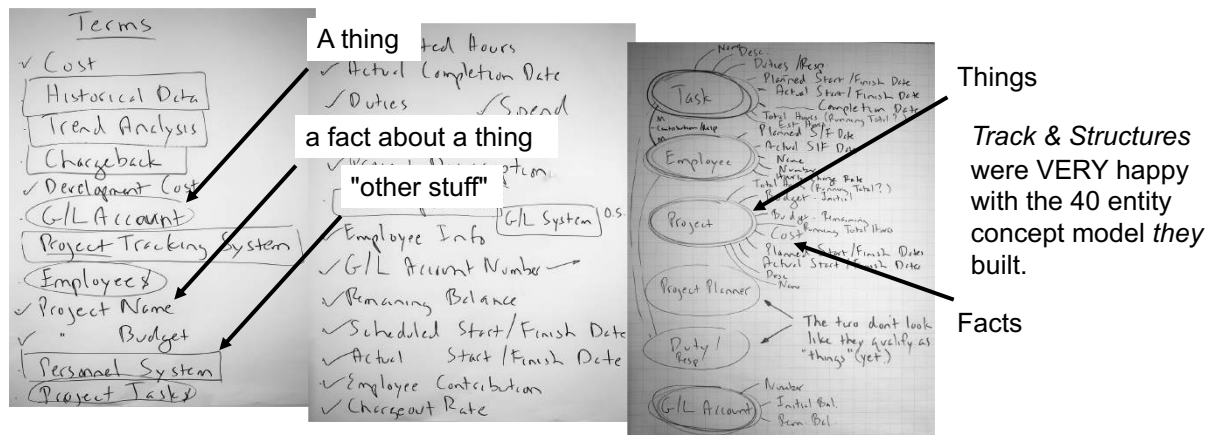


6

It all begins with language

"Why don't you learn *our* language?" "Fair point!"

- Brainstormed over 200 terms –
Track, Structure, Line, Siding, Mileboard, Segment, Sector, Route, ...
- Oh-oh... "Now what?" An idea!
- Is this "a thing, a fact about a thing, or other stuff?"
- Here's a Project Management example...



7

Or, start with one-on-one interviews

- 1) Interview business representatives about their area:
mandate and activities, goals and objectives, issues and opportunities,
needs and wants, likes and dislikes, etc....
Nod sympathetically, but ignore it all (almost!)
Instead, capture "terms" – anything that goes by a name.
- 2) Later, write each term on a large Post-it
- 3) In a facilitated session, participants sort terms into categories:
 - Things (entities, but don't use the term... yet)
 - Facts about things (add new "thing" if it's not there already)
 - "Other stuff"

As needed, introduce criteria to be a "thing" (an entity)

"Other stuff" includes:

- Metrics
- Organisations, departments, jobs, roles, ...
- Processes, functions, activities, tasks, ...
- Systems, tools, equipment, mechanisms, ...
- Reports, forms, screens, queries, ...
- Other – too vague, only one instance, a "fact of life," not a thing we track, etc.

8

Or, start with an email (or other) survey

- *Virtual* first, *in-person* later
- Via email, we gave a "homework assignment"
 - Please spend ~10 minutes listing terms you use daily.
 - Please identify information you need but can't get, or don't trust it
 - No right or wrong – goal is familiarity with your terminology
- The actual text of our email...

Before the session, it would be very helpful if everyone could do two things:

- Spend up to 10 minutes or so listing any terms that come to mind that you use on a frequent basis. Each item in your list could be the name of some thing you need to track, a fact about a thing, a spreadsheet, a report, a metric, a system, a database, or anything else that comes to mind. I'm hoping everyone can list thirty or forty things. There is no "right or wrong" – this helps me learn the language and provides clues to what the most critical terms might be.
- Think of one to three examples of information you'd like to be able to get, but either you can't, or you're not sure how accurate it is. For instance, at a US university last week, a Vice-Provost said she would like to know "How many non-resident, tenure-track Faculty do we have." Of course, this means agreeing what is meant by "Faculty," "tenure-track," and "non-resident." (I've done a LOT of work in higher education, and can promise you there is not agreement on what those terms mean.)

That's the whole point of our sessions next week. :-)

9

Some had little to say, others had lots

- What do we consider a Member in reporting?
- What do we consider an Account in reporting?
- How do we determine how many members we have? Is it based on number of individual memberships or unique SSN?

General Terms

- Member
- Membership
- Individual
- Account
- Services
- Co-borrower
- Primary
- Joint
- Pay off
- FIS (Clientlink)
- Co – op (Springboard)
- DMI
- Past due
- Total due

Member Solutions Terms

- Carmpro (collections system)
- ARM (recovery system in Carmpro)
- Promise to Pay (PTP)
- Payment by interval i.e by hour or day
- Call by interval i.e. by hour or day
- Balances saved by interval i.e by hour or day
- Charge off
- Repossession
- Penny loan
- Workout loan
- Fixed Payment Plan
- Loan Extension
- Bankruptcy by chapter 7,11,13
- Delinquency
- Delinquency rate
- 60 day + delinquency rate
- Charge off rate
- Net Charge off rate
- Recovery
- Recovery rate
- Forced closed
- Lexis nexis
- Credit bureau /credit report
- Skip tracing
- Net flow rate
- Leading edge rate
- Roll rate

All were useful

Terms:

- Household income
- Member growth
- Loan originations per member
- Average relationship balance
- Average relationship account
- Fee income per member
- Net income per member
- Marketing expense per member
- Member trends
- Market segmentation
- Mobile/Online banking penetration
- Member investment products
- Member retirement products
- Lending market overview (credit card, mortgage, auto, home equity)
- Benchmarking
- Payment information (how cards are used)

I went through all the "homework" and selected ~40 terms that qualified as "things" (or entities, or business objects, or classes, or...)

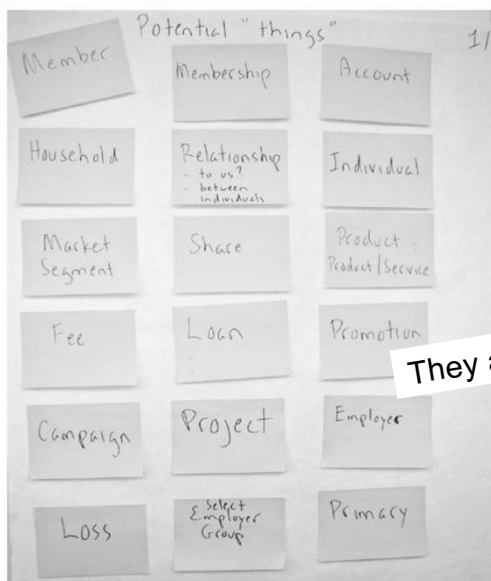
Examples:

- Product/Service adoption based on campaigns
- Track acceptance rates on promotions
- Tracking the new member sales path (similar to what we are doing with the organic growth project)

11

At the first session - "Here are the 'things' you mentioned"

More than enough to work with – here are 30:

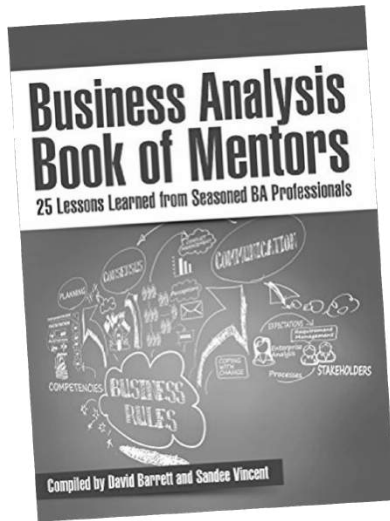


They added more...

"Wow – you actually *did* something with our homework!"

2 — Use the distinction between the "what" and the "who & how"

My chapter in the "BA Book of Mentors"



The premise of the book:

- 25 experienced BAs from around the world would each write a chapter on "The Most Important Lesson I Learned in my BA Career."
- I knew mine instantly – separate the "what" from the "who, how, and why"
- In other words, separate the "essence" from the "accident"

13

Essence and Accident?

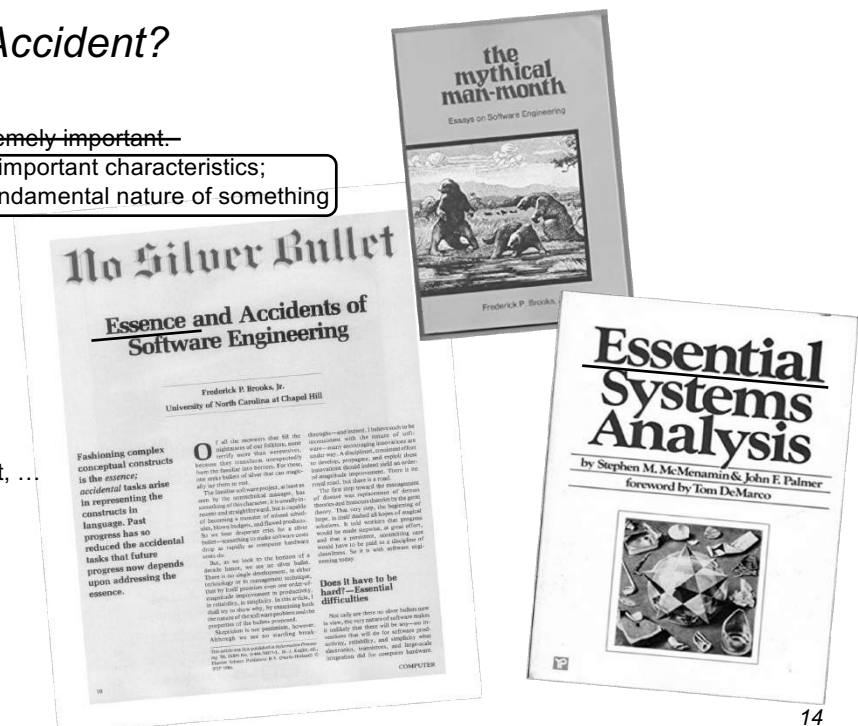
Essential:

- 1) ~~absolutely necessary; extremely important.~~
- 2) something's basic or most important characteristics; the intrinsic, inherent, or fundamental nature of something

Cup:

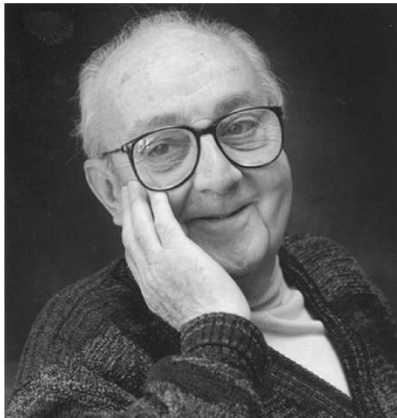
The *essential* characteristics:
a round, handheld container
for drinking from.
What it is.

The *accidental* characteristics:
ceramic vs. bamboo, handle or not, ...
*how it is designed or made and
who it will be used by*



14

"All models are wrong, but some are useful."



George E. P. Box
1919–2013

Two especially useful models

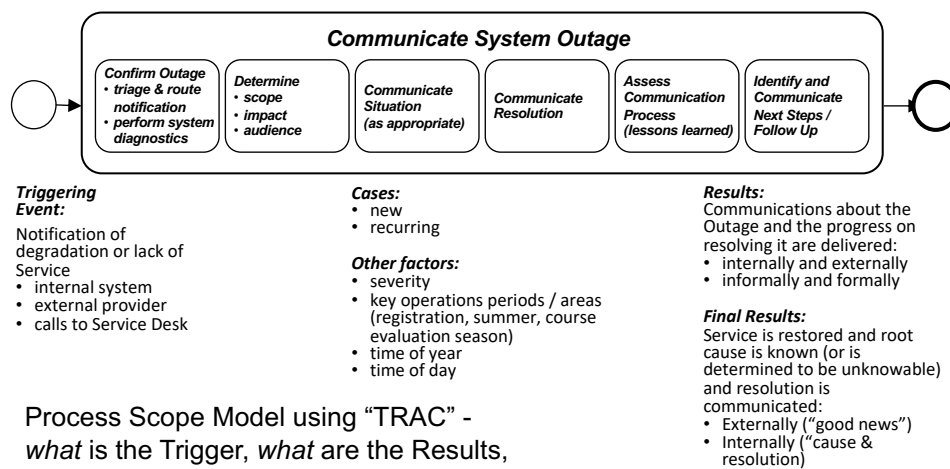
- Business Process Scope Model
- Business Concept Model
(a.k.a Conceptual Data Model)

Both are "essential" –

they show the essence – the "what" –
of a subject with no reference to who,
how, why, etc.

15

"What" - a Process Scope Model

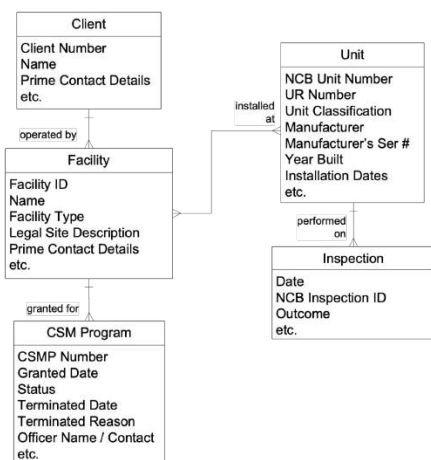


Process Scope Model using "TRAC" -
what is the Trigger, *what* are the Results,
what are the main Activities
(7 ± 2 milestones, phases, or subprocesses,) and *what* are the main cases or variations?

Why 7 ± 2?

16

"What" – a Concept Model



A description of a business in terms of

- **things** it needs to maintain records of – **entities**
- **facts about those things** – **relationships & attributes**
- **policies & rules** – **definitions, constraints, and assertions** governing those things and facts

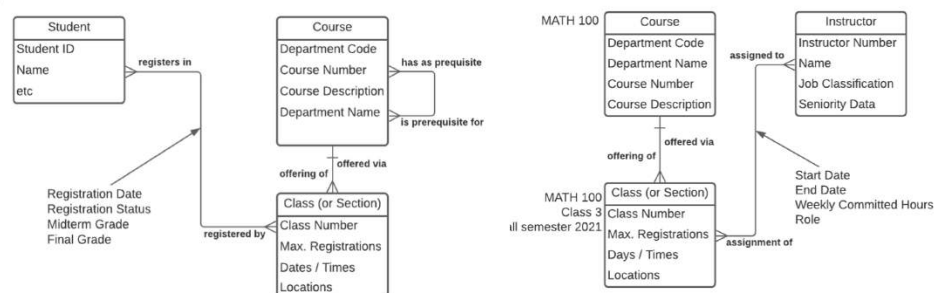
17

Use a Concept Model to demonstrate what's possible

When Concept Modelling in a university environment, everyone wants their favourite report (or artifact) show up on the data model.

We need to be ready to show a near-infinite number of reports etc. can be produced from one essential model:

- Class Roster
- Student Transcript
- Instructor Assignments
- ...



18

3 – *Patience is a virtue*

Two examples I'll illustrate with stories:

1. Don't rush people into agreeing on an entity definition (the heart of concept modelling.)
First, establish there are legitimate differences of opinion
2. Don't rush to generalise or create supertypes.
First, help everyone see the similarities.

Recent student comments -

"I learned modelling should proceed at the pace the group is comfortable with"

and

"Move faster by slowing down"

19

Entity definition - start by opening minds

"Can anyone think of examples that might surprise someone else – that is, anomalies or potential sources of confusion."

E.g., how could we legitimately have different ideas what "Employee" means?

Employee

Project

Account

Task

20

Starting an Entity definition

*"Can anyone think of examples that might surprise someone else – that is, anomalies or potential sources of confusion."
E.g., how could we legitimately have different ideas what "Employee" means?*

F/T vs. P/T?

Only IS Department?

Include management,
or only individual contributors?

Still in recruitment (an applicant)?

Onboarded? on probation? active? retirees?

Include contractors, student interns, vendor staff, etc.?

Volunteers?

A type of worker (DBA or tester) or a specific person?

A robotic, automated, or AI agent?

Employee

Project

Account

Task

21

Starting an Entity definition

*"Can anyone think of examples that might surprise someone else – that is, anomalies or potential sources of confusion."
E.g., how could we legitimately have different ideas what "Employee" means?*

F/T vs. P/T?

– Both

Only IS Department?

– No

Include management,
or only individual contributors?

– Yes, everyone

Still in recruitment (an applicant)?

– No

Onboarded? on probation? active? retirees?

– Yes, all

Include contractors, student interns, vendor staff, etc.?

– Yes, all

Volunteers?

– Yes

A type of worker (DBA or tester) or a specific person?

– No, only a specific person

A robotic, automated, or AI agent?

– No, only a real person

Employee

Project

Account

Task

22

Defining the Entity "Employee" – "Worker"

Definition format:

1. A description of which real-world things are within in scope, and any specific inclusions ("This *includes...*" or "This *is...*")
2. Illustrate with examples – 5 to 10 sample instances or types
3. Interesting points – anomalies, synonyms, common points of confusion, etc.
May include specific exclusions ("This *excludes...*" or "This *is not...*")

Worker (renamed from Employee):

A *Worker* is a person, whether or not directly employed by *the company*, but with some sort of employment contract or arrangement, who has been or may be assigned to a Project.

Worker includes:

- Full or Part-time Employees who have been onboarded, including Probation, Active, Seconded, Suspended, Retired...
- Contractors
- Consultants
- Student Interns
- Vendor Staff Persons
- Company Owners and Managers

Key points:

- "Worker" was chosen as the entity name because it is more generalised than "Employee."
- A Worker may not necessarily be billable on a Project, e.g., a non-chargeable Subject Matter Expert or Volunteer
- Worker excludes:
 - Job Roles, e.g., DBA or Technical Writer
 - Robotic, Automated, or AI Agents (this might change)

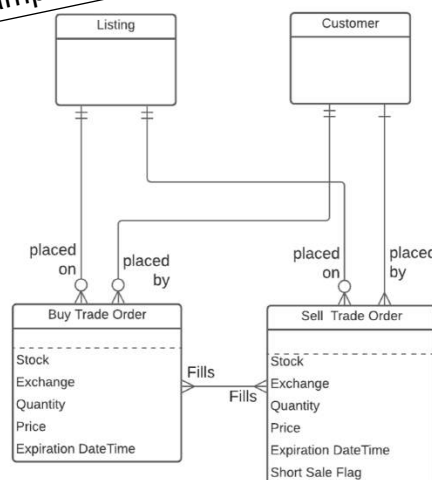
23

A real example - specifics first, generalisation later

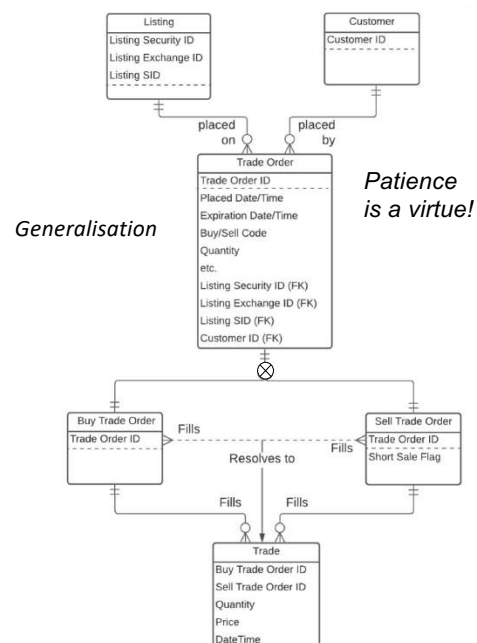
Delaying supertyping/generalisation - a real example

Client:
"Wow! Buy and Sell Trade Orders look very similar!"

Me:
"Omigosh!
That's amazing!"



Specifics



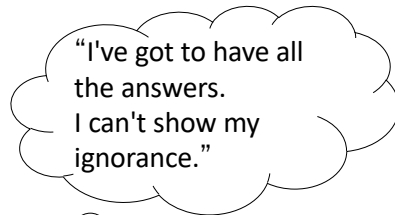
Generalisation

Patience is a virtue!

4) The Principle of Constructive Ignorance

Never be
afraid to ask
“dumb”
questions...

- Myth -



Just one more question, ma'am.
Nothing too important...

Could we go over this just once
more to be sure I've got it right?

There's one thing I'm not clear
on...

- Reality -

- You're not paid to **know** - you're paid to **ask**
- **Someone** will be glad you did
- The number of different answers will surprise **everyone**



Lieutenant Columbo takes up Data Modelling

25

5) Graphical design principles matter

It's learnable! e.g., Gestalt principles (study them on your own)

- how humans instinctively organize individual visual elements
- groups, patterns, and unified wholes vs. separate parts.
- key principles:

- proximity
- similarity
- closure
- continuity
- symmetry



Good Figure

Objects grouped together tend to be perceived as a single figure. Tendency to simplify.



Similarity

Objects tend to be grouped together if they are similar.



Closure

Visual connection or continuity between sets of elements which do not actually touch each other in a composition.



Proximity

Objects tend to be grouped together if they are close to each other.



Continuation

When there is an intersection between two or more objects, people tend to perceive each object as a single uninterrupted object.



Symmetry

The object tend to be perceived as symmetrical shapes that form around their center.

26

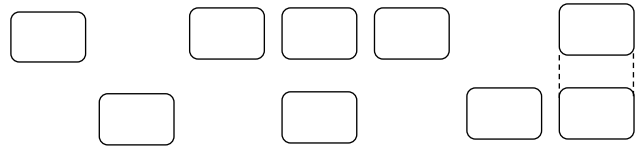
Cognitive psychology of diagramming - a process example

What do people first perceive on a diagram?

1. relative size

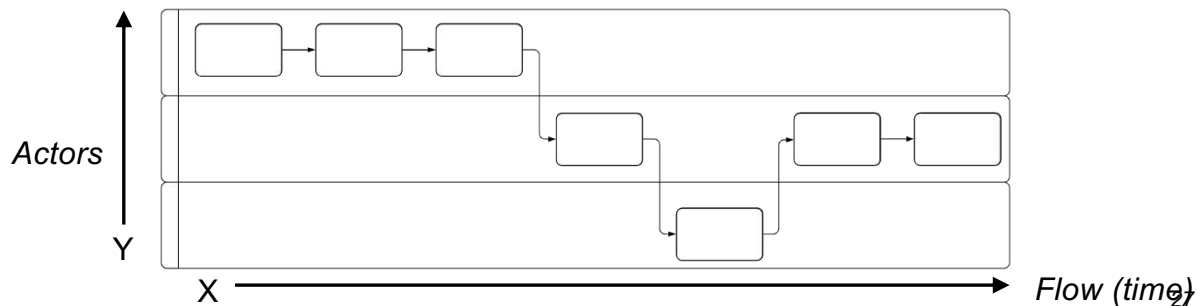
an
important
step

a less
important
step

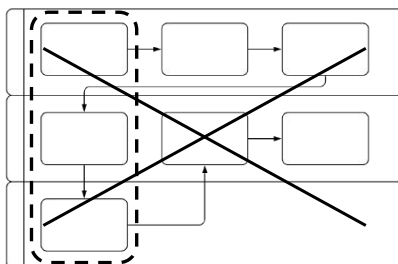


Make all the steps the same size,
unless you're trying to make a point

2. relative X-Y position



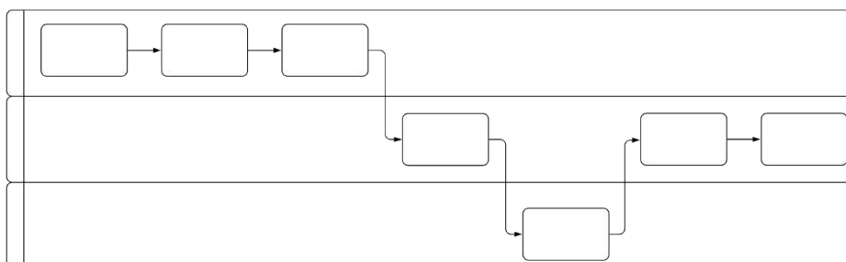
Don't conceal sequence and dependency



Steps perceived as happening in *parallel*,
even though flow lines indicate *sequential*.

Critical in analysing a process:

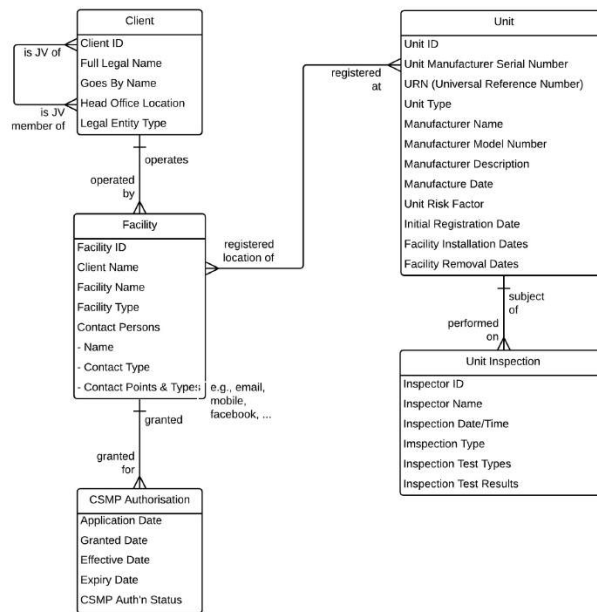
- sequential vs. parallel
- dependent vs. independent



A simple guideline:
flow lines *only* leave
the right edge
and *only* enter
the left edge – never
the top or bottom.

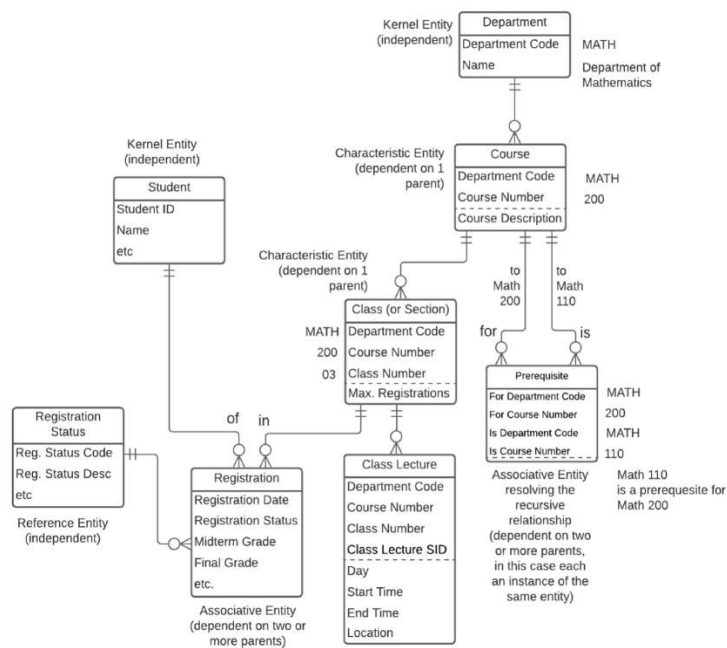
And for data...?

For data models, always draw top-down by dependency



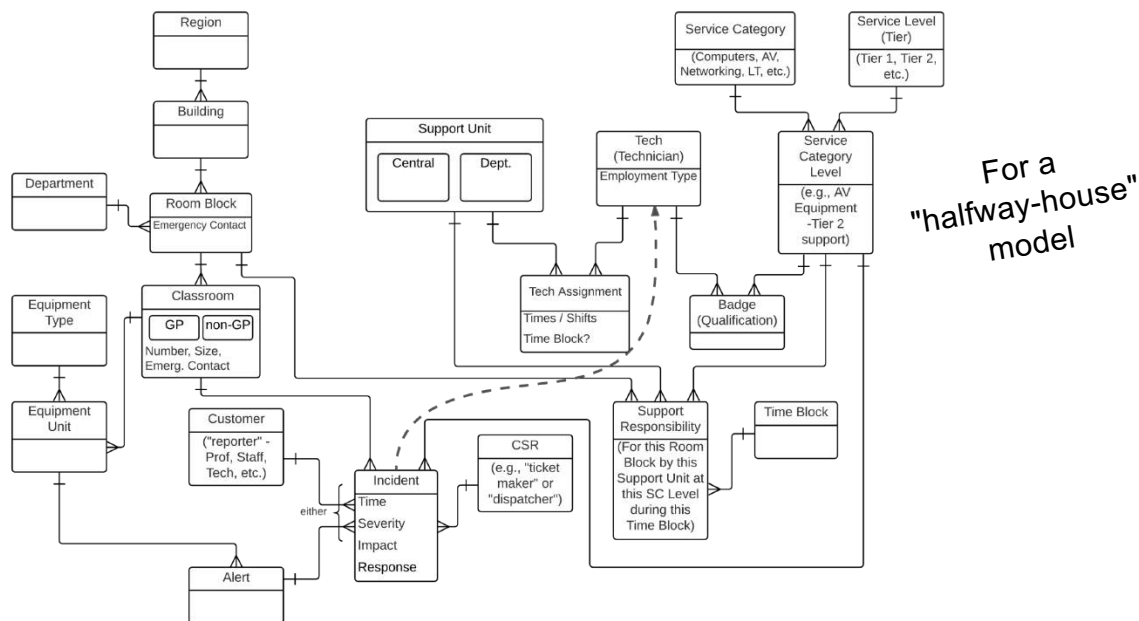
For a Concept Model

Top-down by dependency



For a Logical Model

Top-down by dependency



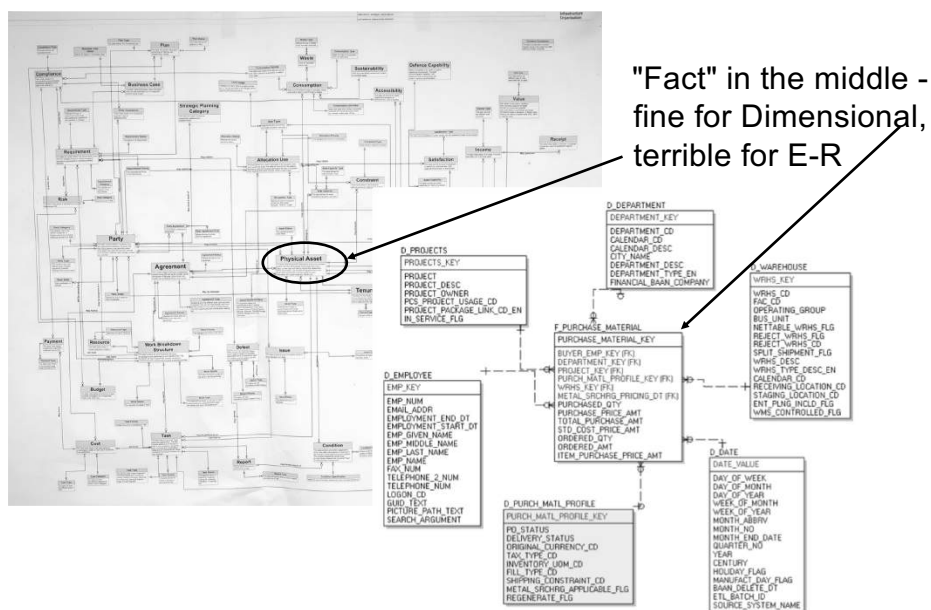
31

What works for Dimensional Models doesn't for E-R Models

A common error –
"the most important
entity should go in the
centre of the diagram."

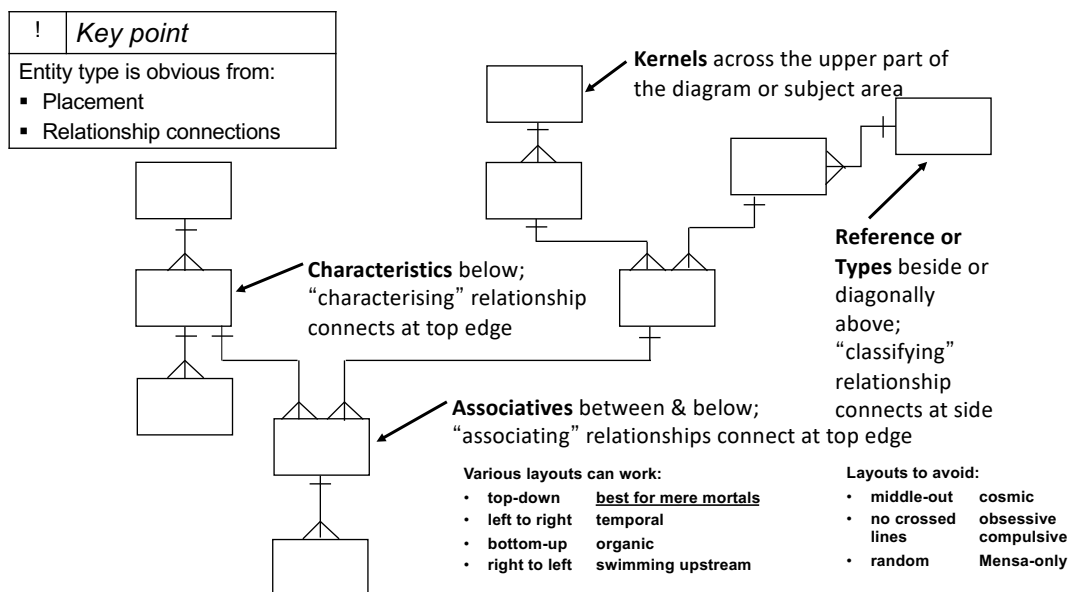
An excellent model
structurally, but very
difficult to follow –
no sense of direction.

Concept Models / E-R
Models should be
drawn top-down by
dependency.



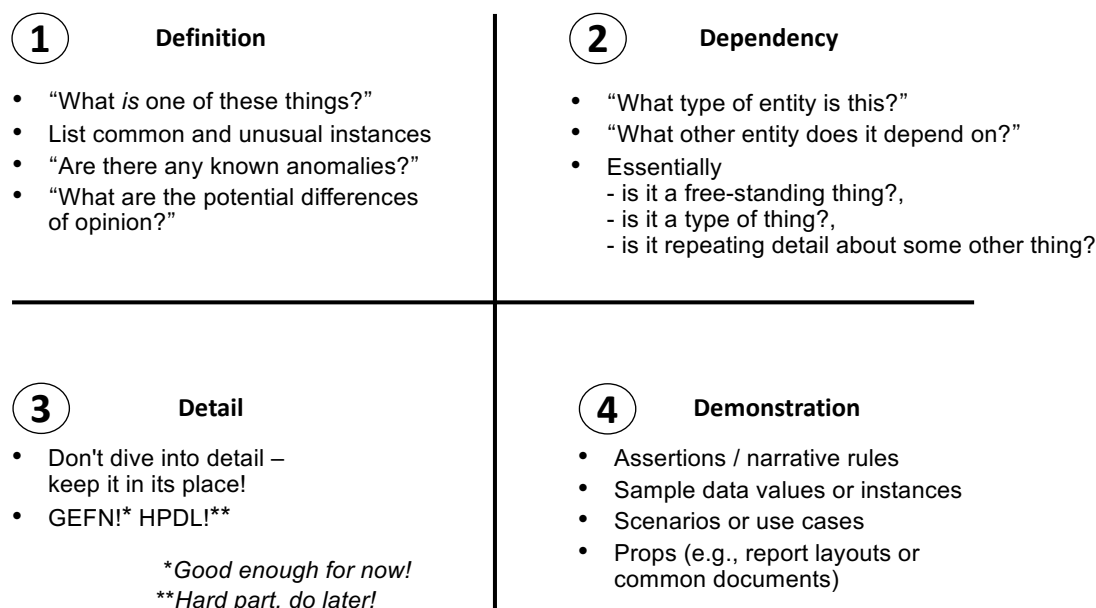
32

Graphic guidelines – the “no dead crows” principle



33

In summary – don't forget the four Ds of Concept Modelling



34

What we covered

★	<i>Five Key Points</i>
1. Start with a nice conversation 2. Understand the "what" vs. "who and how" distinction 3. Patience is a virtue (so are Simplicity and Consistency) 4. Be fearless yet vulnerable - you don't know it all 5. Graphical design principles matter	

*We hope you got some
good ideas from this session!*

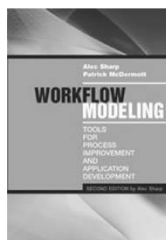
35

Thank you – stay in touch!



Alec Sharp
Clariteq Systems Consulting Ltd.
West Vancouver, BC, Canada

- asharp@clariteq.com
- ig: [@alecsharp01](#)
- www.clariteq.com
- <http://amzn.to/dHun1o>



And most of all, if you have questions or comments...
don't be shy – send me a note!

36