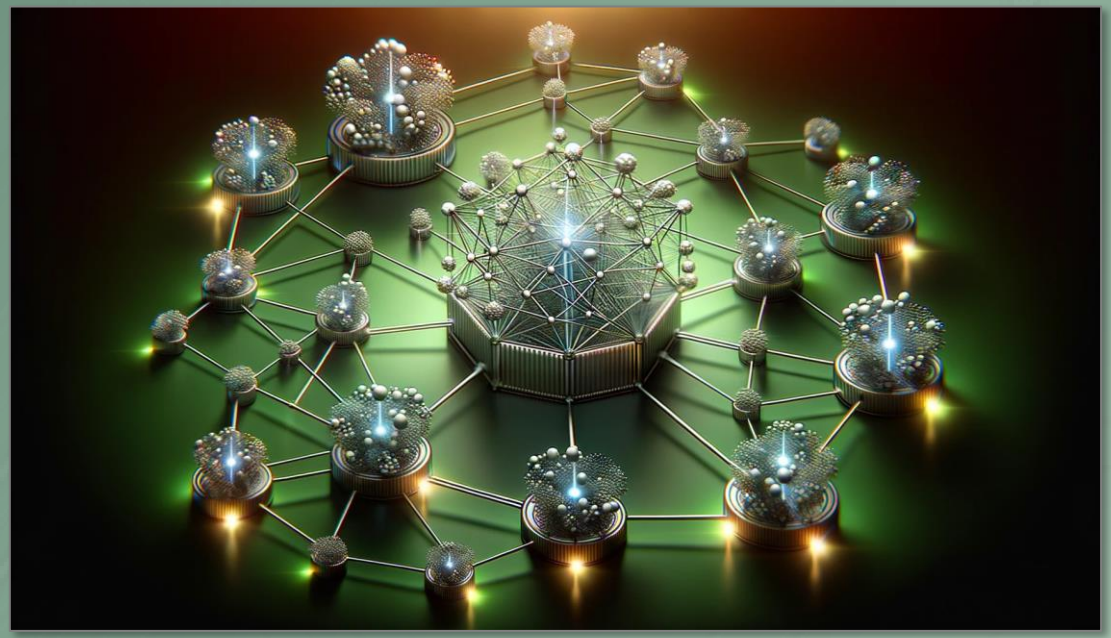
BHT Berliner Hochschule für Technik

Data Mesh



Advanced Software-Engineering

Dr. Harald Stein, Prof. Dr.-Ing. Stefan Edlich

Dez 2023

Agenda

- **What is Data Mesh?**

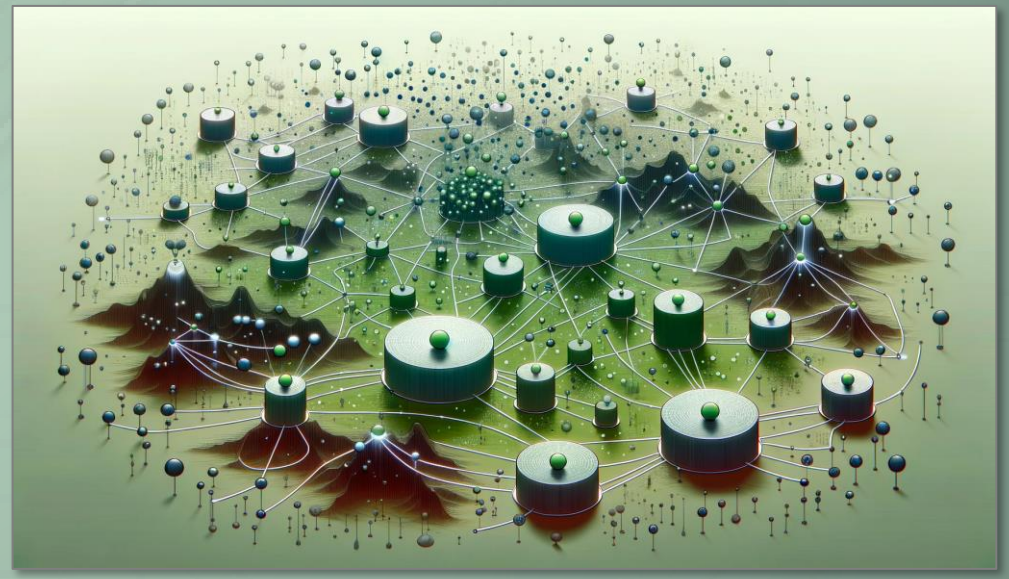
  ...and why you may need it

- **Data Mesh Architecture**
  ... Data Product, Domains, Self-Serve Data Platform, Federated Governance

- **Distinction from Data Warehouses and Data Lakes**

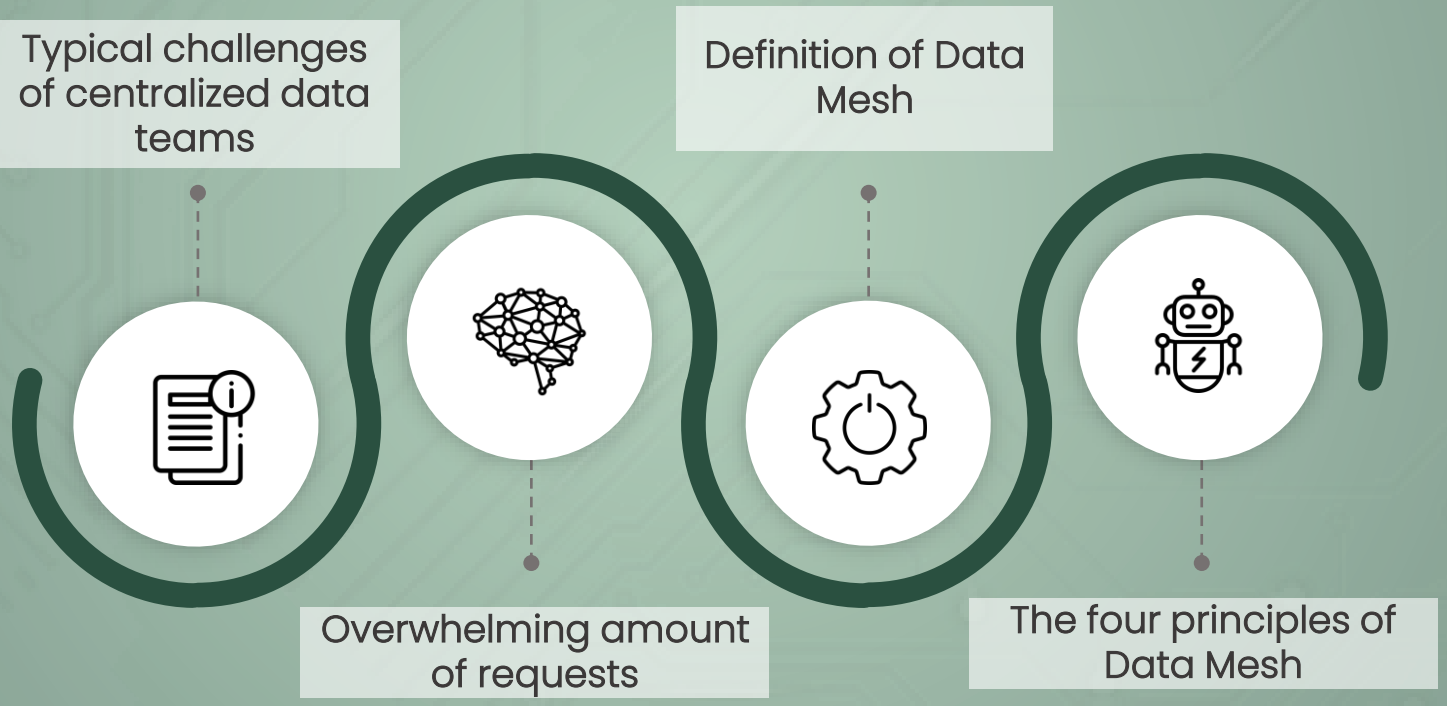
- **Domain Team's Journey**

- **Practical examples**



What is Data Mesh?

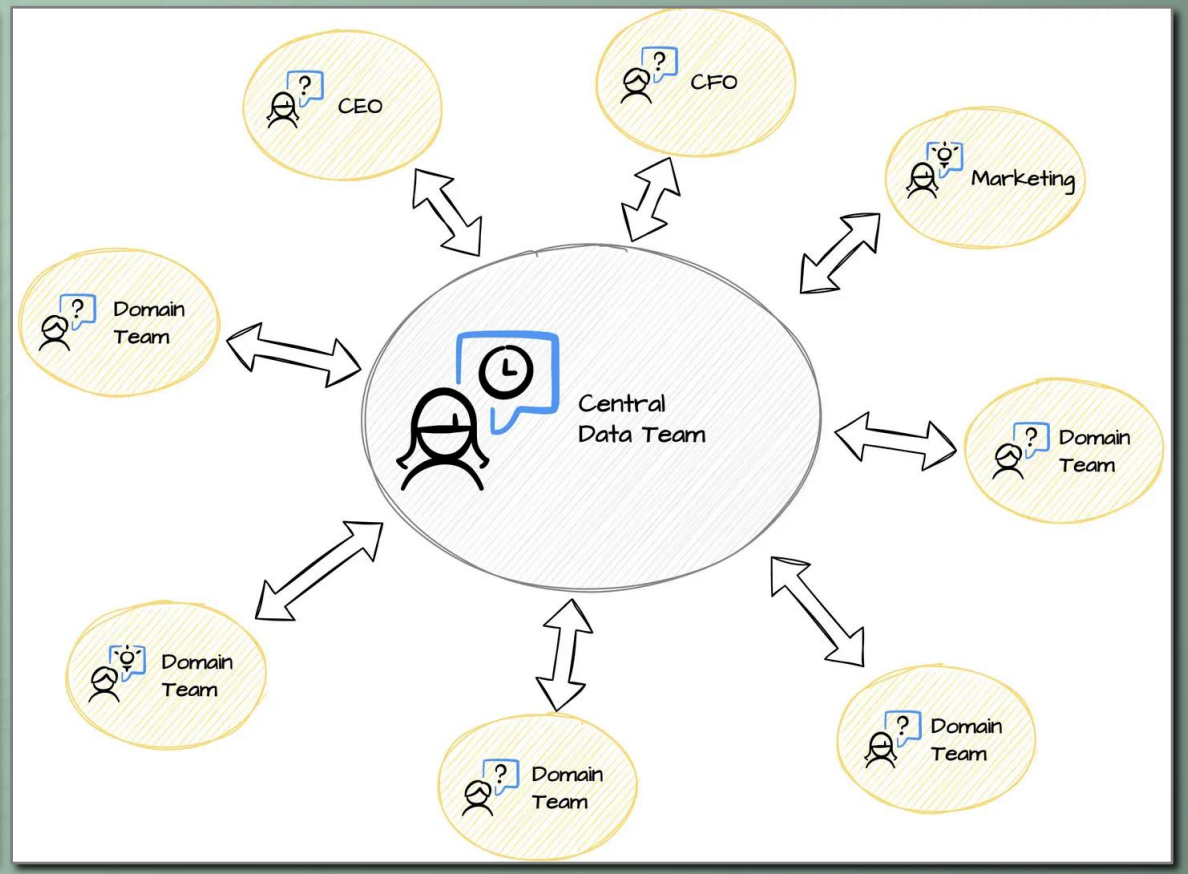
... and why you may need it



Typical challenges of central data team

... occur, when single team is responsible for the entire organization's data needs

- **Bottlenecks**:
  Overwhelmed with numerous requests, causing delays.
- **Limited Domain Knowledge**:
  Potential for misinterpretations due to lack of specific expertise.
- **Scalability Issues**:
  Difficulty handling growing data volumes and requests.
- **Single Point of Failure**:
  Issues with team can disrupt entire organization's data operations.
- **Prioritization Conflicts**:
  Challenges in ranking diverse departmental requests.
- **Dependency**:
  Other departments rely heavily on the centralized team, leading to inefficiencies.
- **Data Silos**:
  Risk of isolated data pools, even within centralized structure.



Overwhelming amount of requests

... for central data team creates bottlenecks, i.e. inefficiency

Predictive Modeling:
"Can you create a model to predict next quarter's revenue?"

Data Extraction
"Can you pull the last year's sales data for region X?"

Ad-hoc Analysis:
"How did our sales perform during the last holiday season compared to the previous year?"

Security Checks:
"Ensure that our data storage is compliant with the latest data protection regulations."

Historical Data Retrieval: "Can you retrieve the product data from five years ago?"

Issue Resolution:
"There's a discrepancy in the financial reports. Can you check?"

Integration Tasks:
"Can we integrate social media metrics into our existing CRM?"

Data Visualization:
"Can you create a heatmap of our product sales by region?"

Tool Implementation:
"Can you set up and integrate a new analytics tool for the sales team?"

Report Generation:
"We need a monthly performance dashboard for the marketing team."

Data Annotation:
"Label the new dataset for our machine learning project."

Data Migration:
"We're transitioning to a new system. Can you migrate the existing datasets?"

Survey Analysis:
"We've collected customer feedback. Can you analyze the sentiment and summarize it?"

Training Requests:
"The new hires need training on how to use the data tools. Can you arrange a session?"

Data Backups:
"Ensure regular backups of our critical datasets."

Custom Queries:
"Can you write a custom SQL query to extract specific insights from our database?"

Data Enrichment:
"Add demographic details to our customer database."

Real-time Analytics:
"Set up a real-time dashboard for tracking our website's user activity."

Data Cleaning:
"The recent survey data has inconsistencies. Can you clean and standardize it?"

Database Maintenance:
"The database seems slow. Can you optimize it?"

What is Data Mesh?

A modern architectural paradigm that decentralizes data roles, treating data as a product and emphasizing domain-oriented ownership, introduced by Zhamak Dehghani

A Data Mesh

- decentralizes data ownership for scalability
- empowers domain experts
- boosts agility and data quality
- fosters innovation
- and balances team autonomy with federated governance
- making it a resilient and efficient approach to data management.

Sociotechnical approach

to

- share
- access and
- manage

analytical data in complex and large-scale environments - within or across organizations.

Key Points

- Data as a Product: Elevating data to have its own lifecycle and value.
- Domain Ownership: Assigning data responsibilities to domain experts.
- Self-Serve Infrastructure: Empowering teams with tools for autonomous data operations.
- Federated Governance: Balancing autonomy with overarching standards.

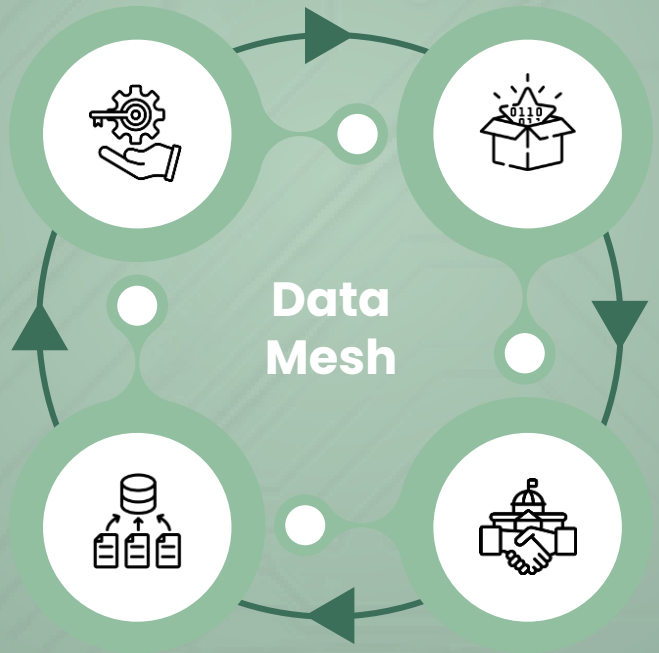
The four principles of Data Mesh

Domain Ownership

- Domain teams own and are responsible for their data.
- Analytical data aligns with team boundaries and system contexts.

Self-serve Data Platform

- Platform thinking applied to data infrastructure.
- Dedicated platform team offers tools for creating and consuming data products.



Data as a Product

- Analytical data seen as a consumable product.
- Domain teams provide high-quality data to other domains, akin to public APIs.

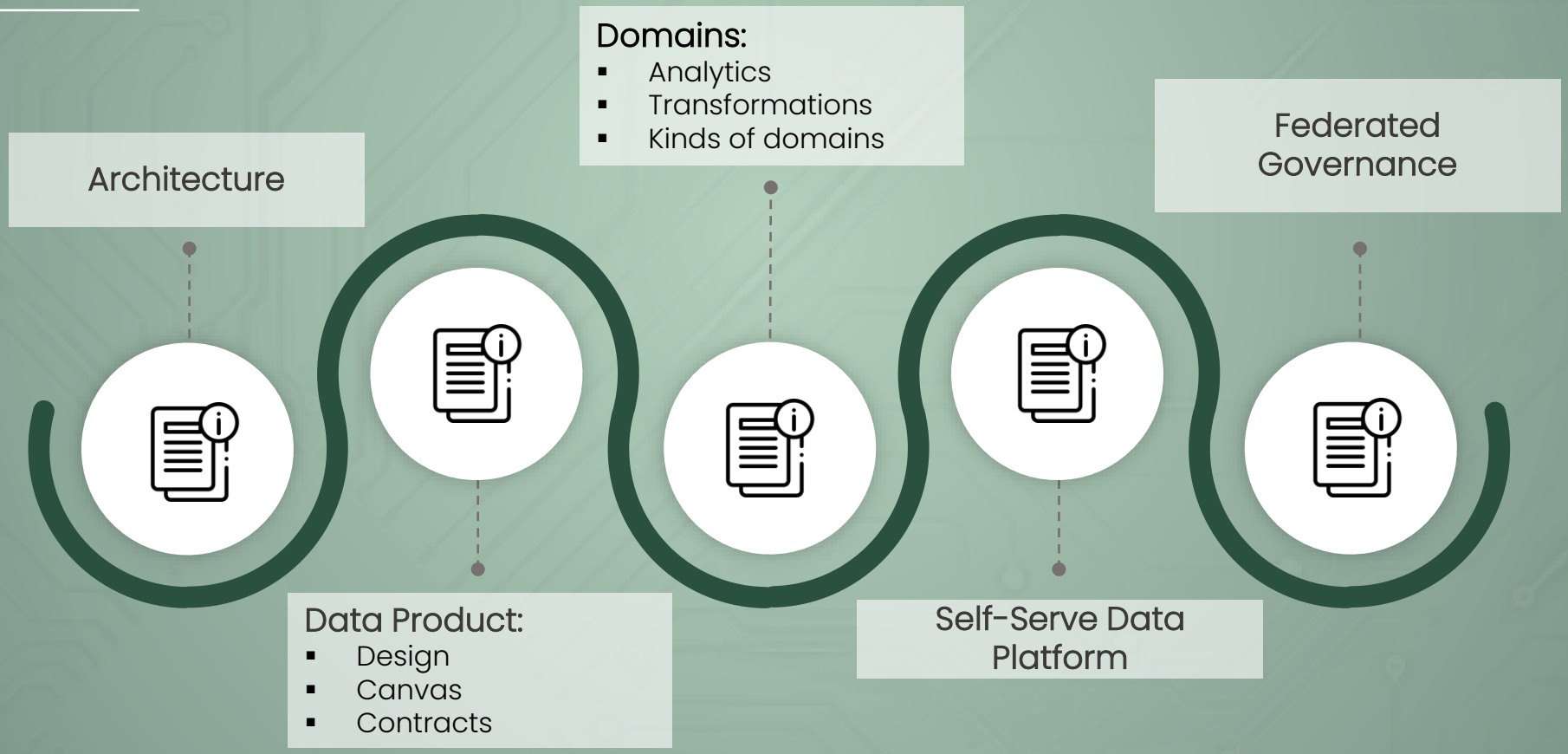
Federated Governance

- Ensures data product interoperability through standardization.
- Aims for a cohesive data ecosystem following organizational and industry rules.

The four principles of Data Mesh

	Domain Ownership	Data as a Product	Self-serve Data Platform	Federated Governance
Strategic Domain-driven Design	Domain Bounded Context	Product Thinking	Domain-agnostic	Context Mapping
Socio-technical Perspective	Domain Teams	Data Product by Domain Team	Data Platform Team	Governance Group
Technology	Operations & Analytical Data	Interoperability Interfaces	Self-serve Data Platform	Policy Automation

Data Mesh Architecture



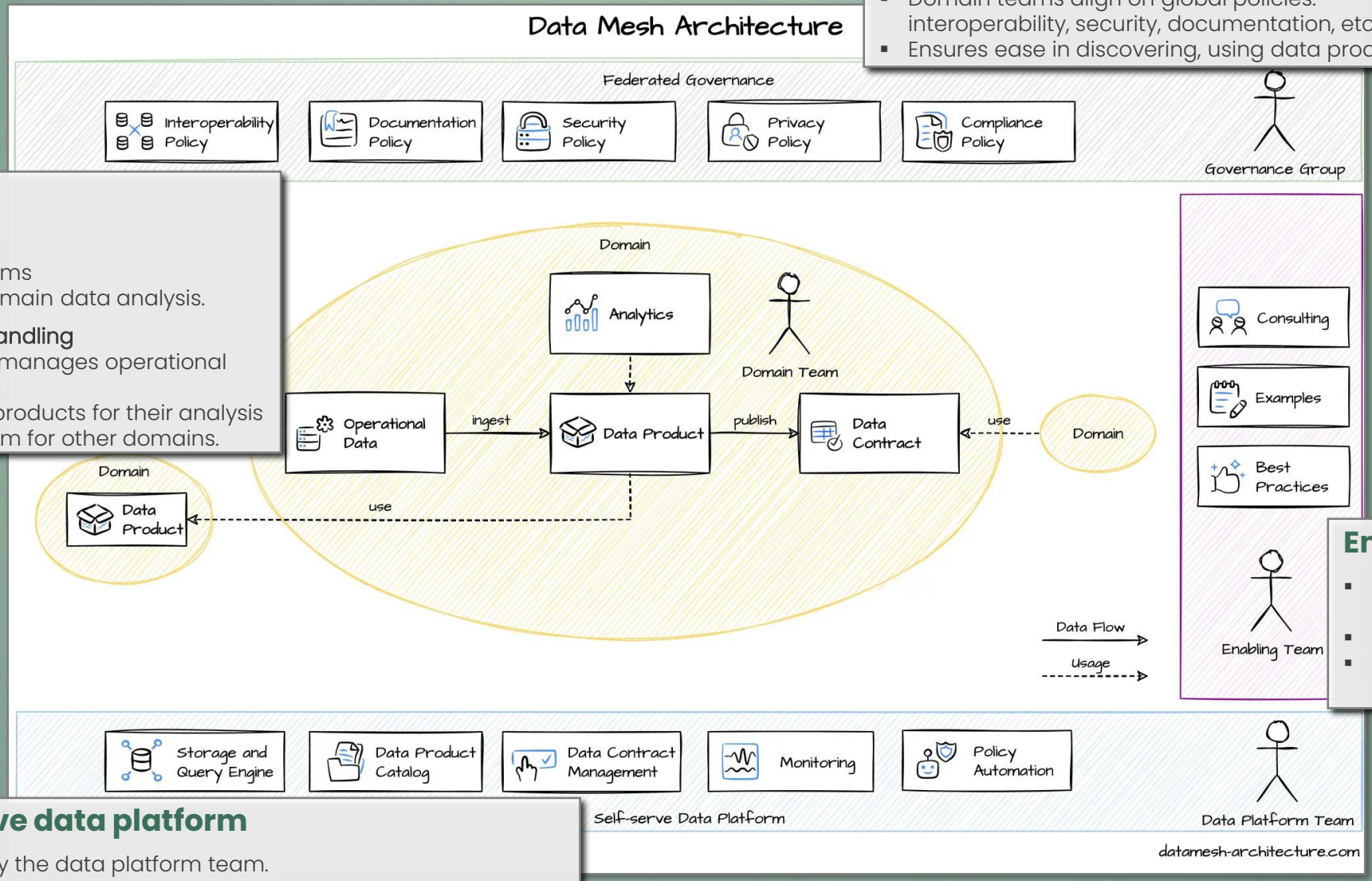
Data Mesh Architecture

Federated Governance

- Domain teams align on global policies: interoperability, security, documentation, etc.
- Ensures ease in discovering, using data products within the mesh.

Domain Teams

- Decentralized Approach
  - Enables domain teams
  - to execute cross-domain data analysis.
- Domain-Centric Data Handling
  - Each domain team manages operational and analytical data.
  - They develop data products for their analysis
  - and can publish them for other domains.



Enabling team

- Guides domain teams on data modeling
- platform usage
- and maintaining interoperable data products.

Self-serve data platform

- Offered by the data platform team.
- Assists domain teams in creating and analyzing data products

Designing Data Products in Data Mesh

Treating data as a tangible product, tailored for specific domain needs.

Examples of Data Products

- Database tables/views
- Unstructured files with metadata (e.g., images, videos)
- Data streams from transaction systems
- Change logs (e.g., billing account events)
- Files (CSV, Excel, Parquet)
- REST APIs
- Master Data Management database
- Features for ML models
- Dashboards and visualizations

Aspects of Data Product

A data product is bounded, interoperable, self-aware, discoverable, secure, shareable

Interoperable

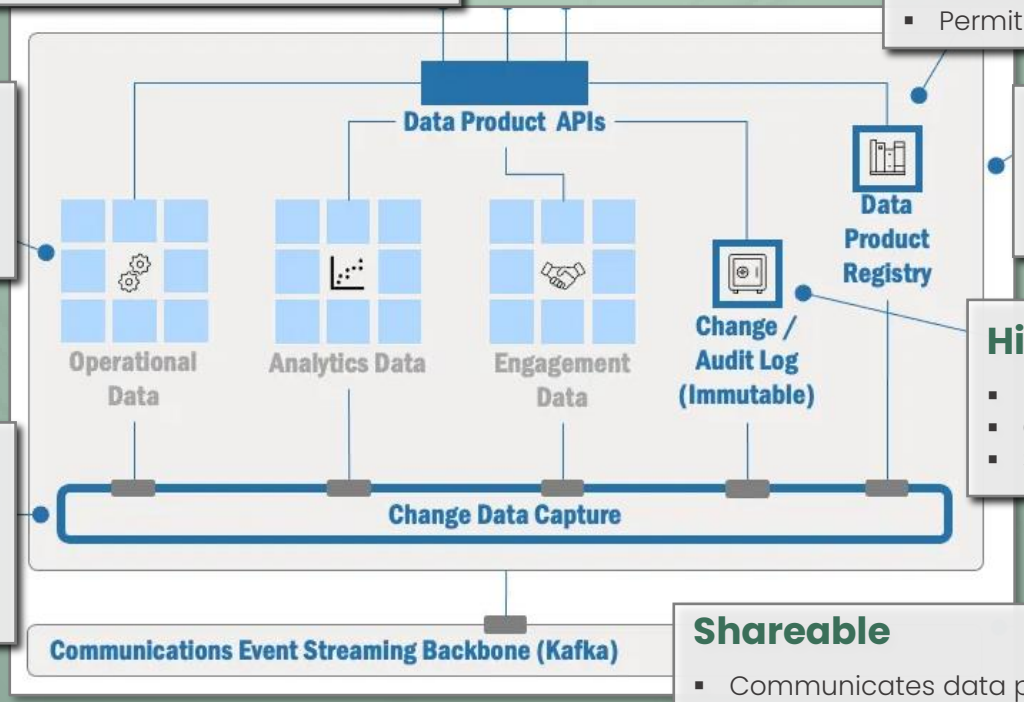
- Interfaces include queries, APIs, pipelines, files, events, etc.
- Interfaces come with formal contracts, e.g., APIs using OpenAPI specifications.
- Additional interfaces, mainly APIs, available for various operations.

Discoverable

- Data product registry serves as "one-stop-shop" for developers, data scientists, and analysts.
- Enables finding, consuming, sharing of data.
- Surfaces metadata related to data product.
- Allows sophisticated user interactions to request data.
- Permits "owners" to create new data products.

Bounded

- Data products store data with defined boundary and owner.
- Primary use case: analytic data.



Secure

- Ensures data is secure both at-rest and in-motion.
- All data products should operate in "Zero-Trust" container/environment..

Historical

- Monitors changes to data product state, exceptions, etc.
- Captures, manages these in immutable log.
- Supports federated governance and security framework.

Self-Aware

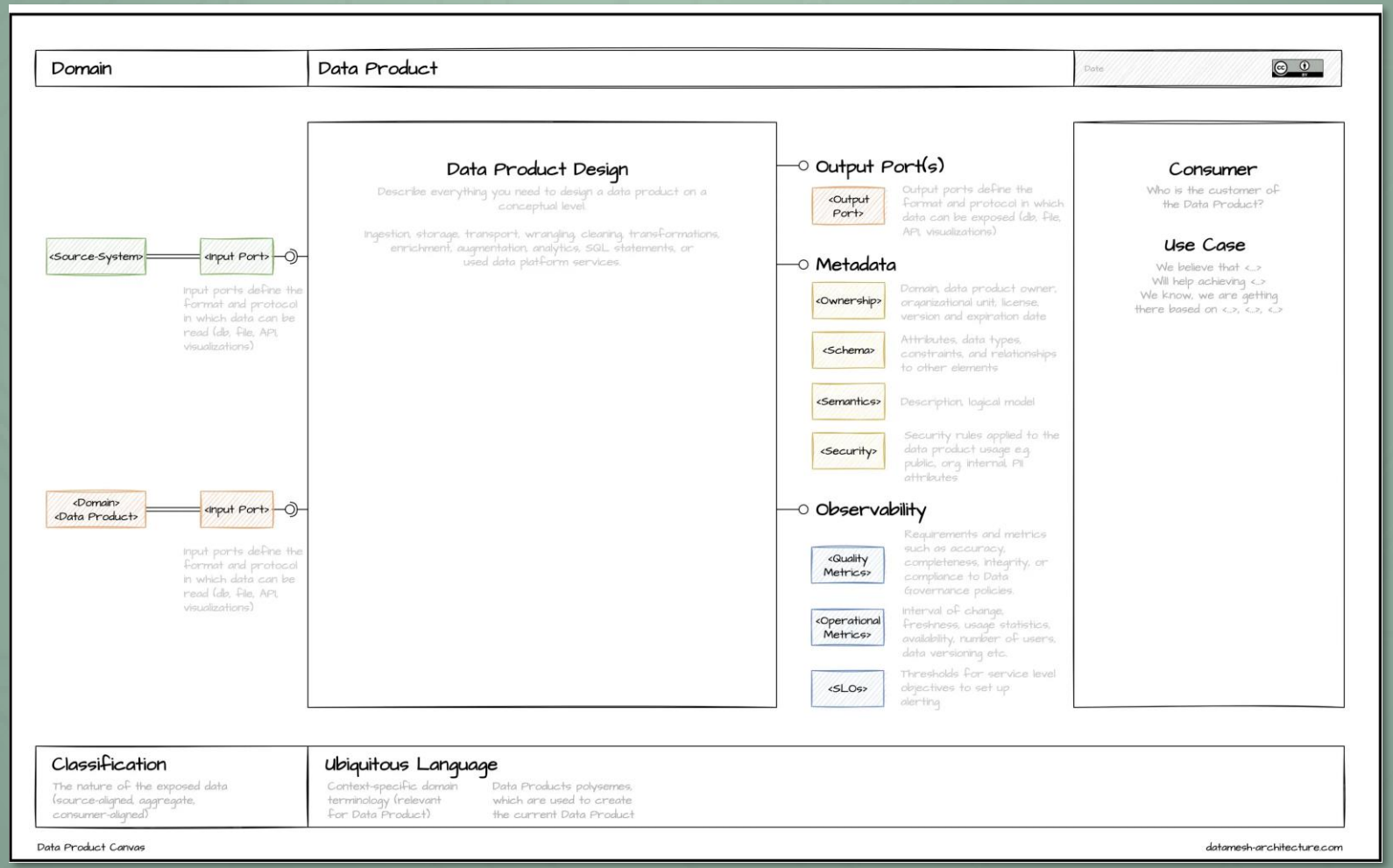
- Captures all changes in data products.
- Changes are distributed as "events".
- Events can be shared within data product, to other data products, parties in the enterprise.

Shareable

- Communicates data product information, events in near real-time.
- Examples of events: data changes, API calls.
- Communication: data product domain ⇔ different data products.
- Extends across the entire organization.
- Uses robust, reliable, resilient infrastructure for sharing.

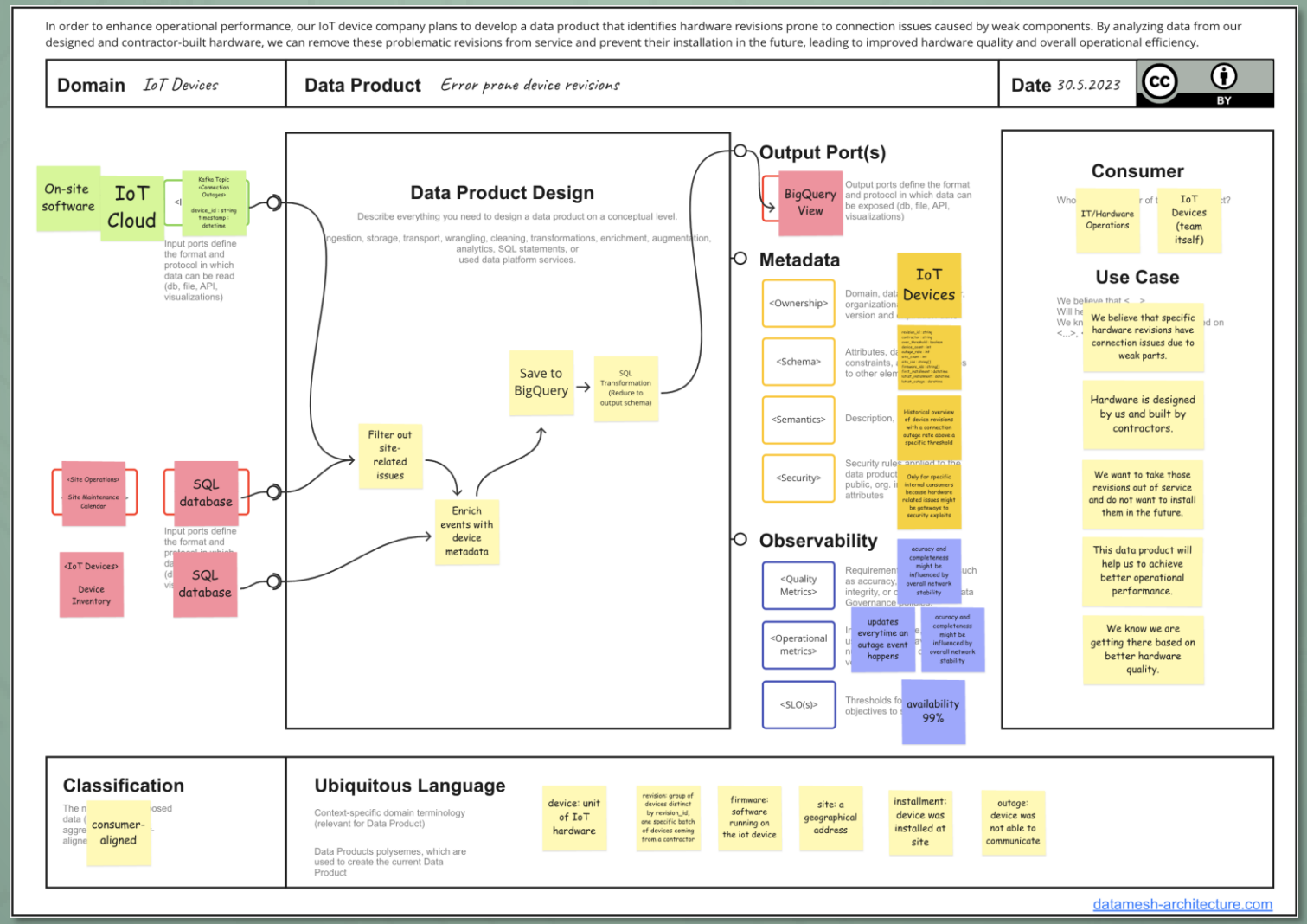
Data Product Canvas

… is a visual framework that guides a team through data product specification



Data Product Canvas example

… is a visual framework that guides a team through data product specification



Data Contracts within Data Mesh

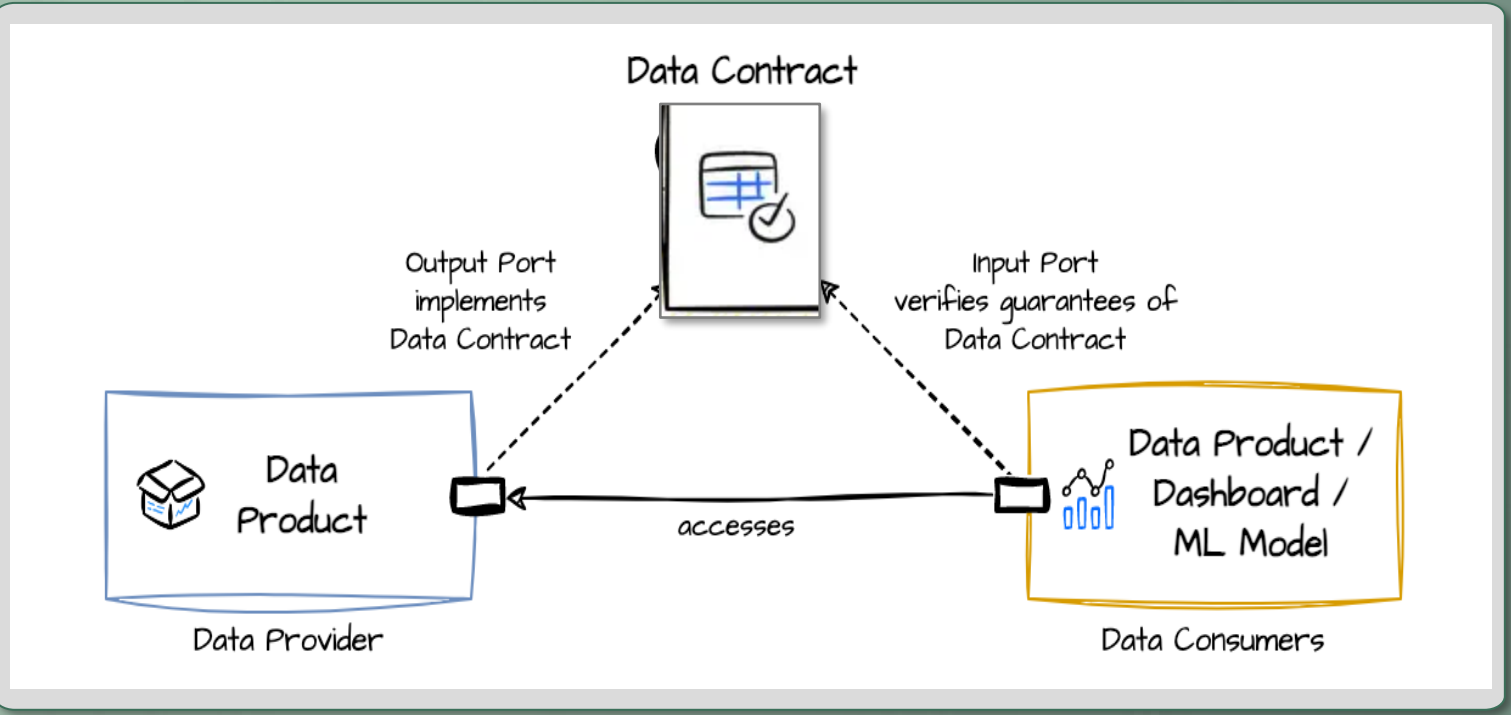
are formalized, explicit agreements detailing structure, quality, frequency of shared datasets. They connect data producers/providers and data consumers

Purpose

- Ensures data consistency, integrity across teams.
- Provides blueprint for how data should be produced and consumed.
- Reduces ambiguity, ensuring data is trustworthy, reliable.

Implementation

- Key points, structure and use:
  Defines data structure, format, semantics, quality, use terms.
- Implementation:
  Realized through a data product's output port or related technologies.
- Format:
  - Specified in YAML
  - Neutral to data platforms and supports standard schema formats (e.g., dbt models, JSON Schema)
  - and quality checks (e.g., SodaCL, SQL queries).



Data Contracts within Data Mesh

… define fundamentals, data structure, data quality, Service-level agreements, stakeholders, security, pricing, etc. They ensure guarantees about data provisions and usage expectations.

Collaboration

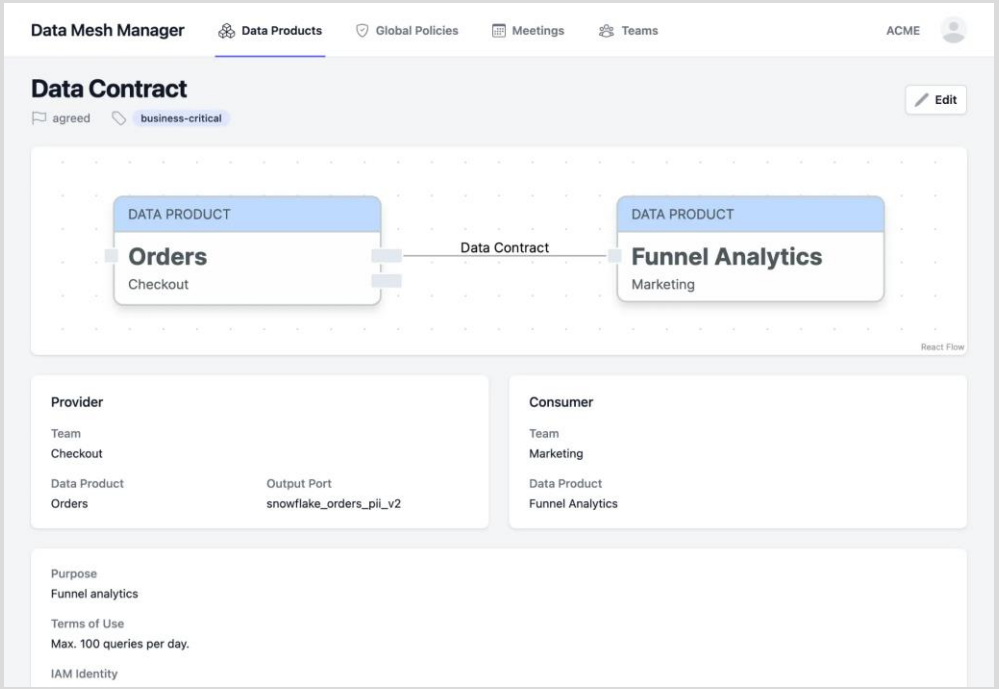
- Facilitates discussions on semantics, quality, and requirements prior to implementation.
- Establishes agreed-upon usage terms and billing policies.

Access Management

- Supported by Data Mesh Manager for end-to-end data access life-cycle.
- Enables consumers to request data product access based on contract terms.
- Data product owners can approve or reject access; event-driven API aids in automating this.

Testing & Compliance

- Ensure data products align with contracts during CI/CD pipeline stages.
- Validates upstream data quality as per contract terms.
- Integrated compatibility with Data Contract CLI tool in Data Mesh Manager.



https://www.datamesh-manager.com/

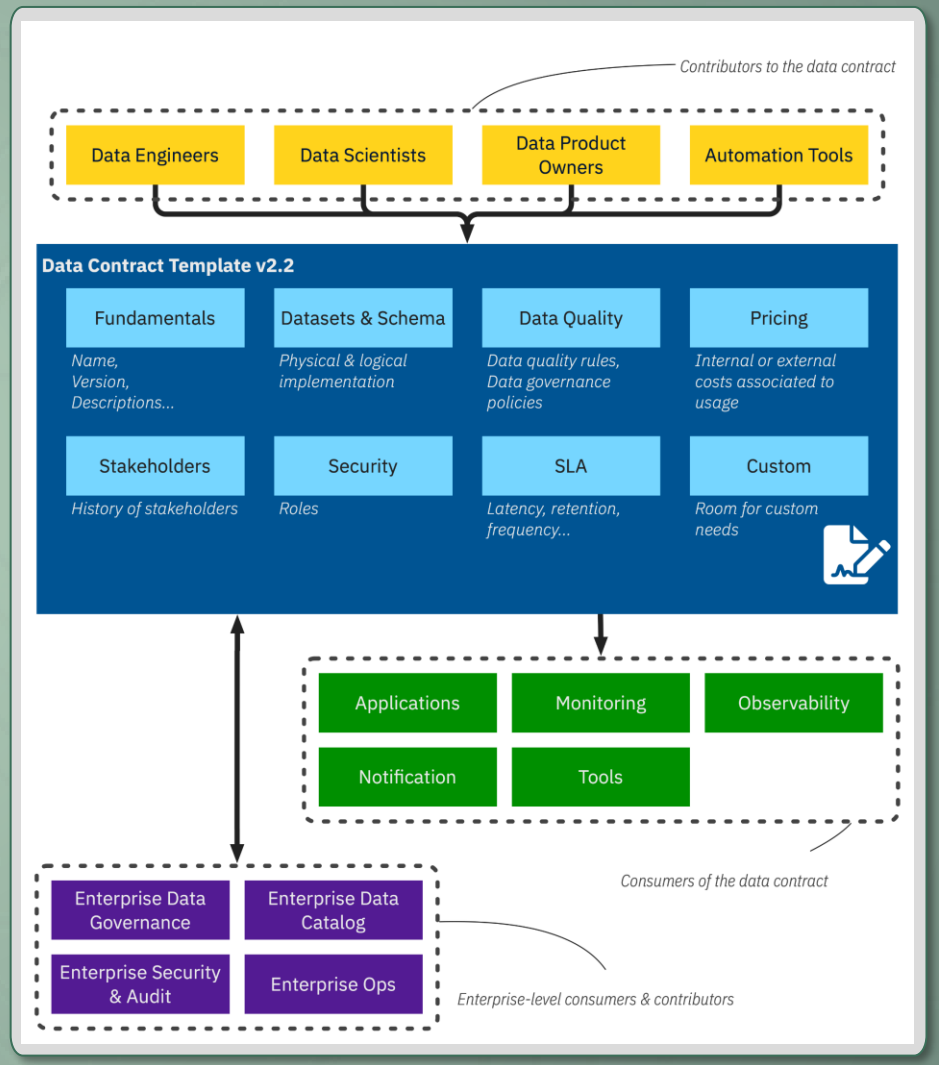
Real-World Application of Data Contracts

Large retail firm gathers data on customer purchases, demographics, etc.

- Data is utilized:
  for enhancing marketing, product development customer service.
- Shared with third-party entities:
  analytics firms, marketing agencies, and cloud providers.
- Vendors offer services
  like customer segmentation, targeted ads, data processing.

Data contracts

- … with vendors ensure compliant and responsible data usage.
- … show in detail data usage, sharing rules, data privacy/security measures.
- Crafting and implementing robust data contracts protects data and ensures responsible usage.



Open Data Contract Standard, Apache 2.0

Kinds of domains

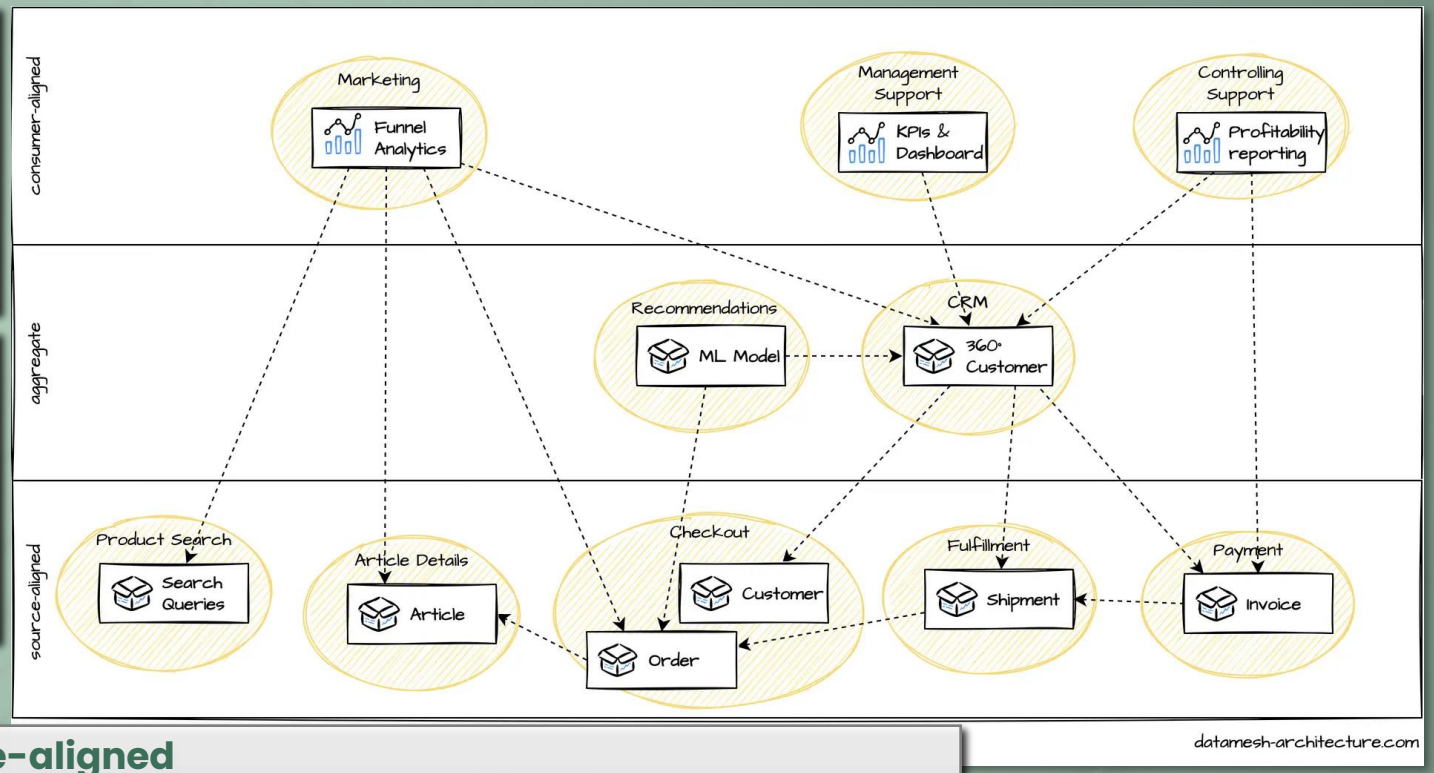
Domains can be classified by data characteristics and data product usage.

Consumer-aligned

- Tailored data for business departments across entire value stream.
- Used by management for detailed reports and KPIs.
- Marketing uses tools like Google Analytics for funnel, web analysis.
- Focuses on specific department needs.
- Business-IT integration grows, leading rising potential for data analytics

Aggregate

- Focus on delivering data products aggregated from various domains.
- Examples:
  - 360° customer view: Combining account data, orders, shipments, invoices, and more.
  - ML teams: Using data from checkout and the 360° view to train recommendation models.



Source-aligned

- Dividing online shop into domains (e.g., product search, checkout, payment).
- Domains publish data as products for others to access.
- Engineers analyze their own data to enhance operations.
- Uses neighboring domain data for insights.
- Data is closely tied to domain events and entities.

Analytics by domain teams

To gain insights, domain teams query, process, and aggregate analytical data together with relevant data products from other domains.

Analytical Queries

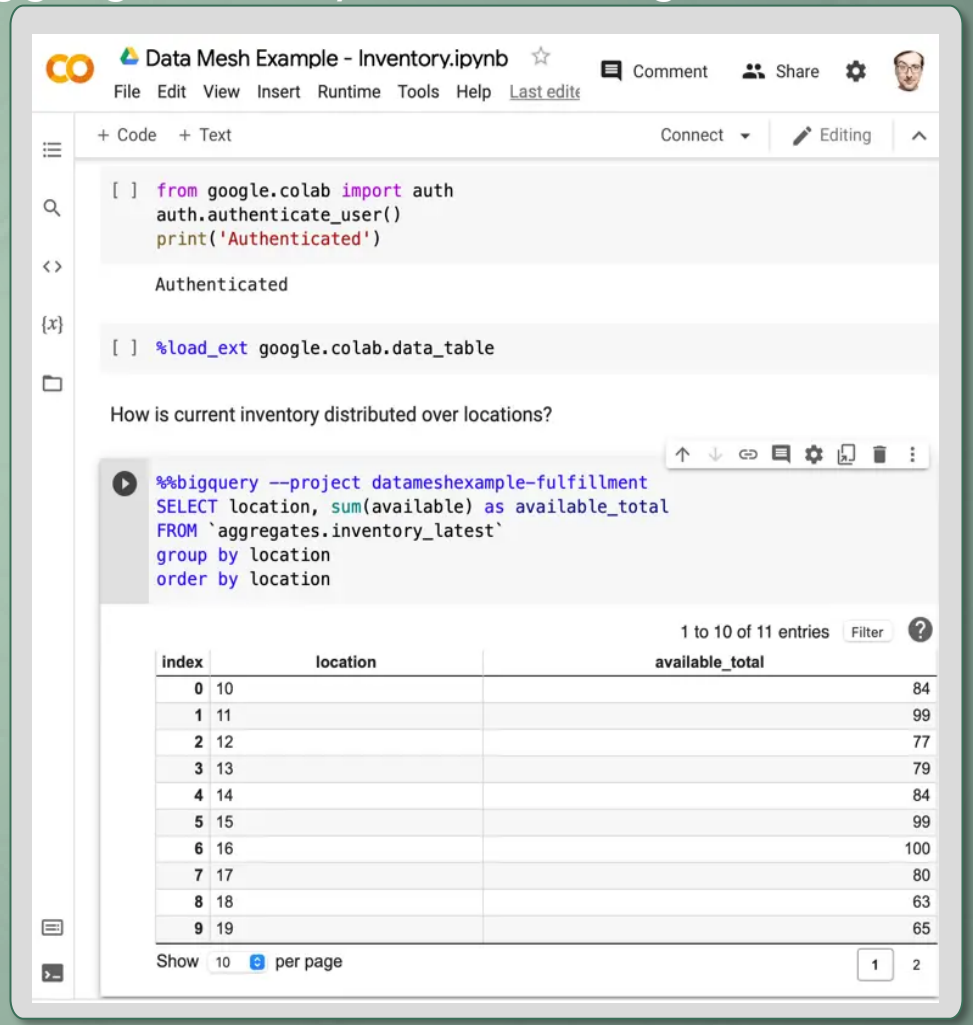
- Domain teams integrate analytical data from various sources.
- Foundation: SQL for in-depth data investigations.
  - Efficient joins for large datasets.
  - Use of aggregations and window functions.
  - Exploratory tools: Jupyter Notebooks, Colab

Data Visualization

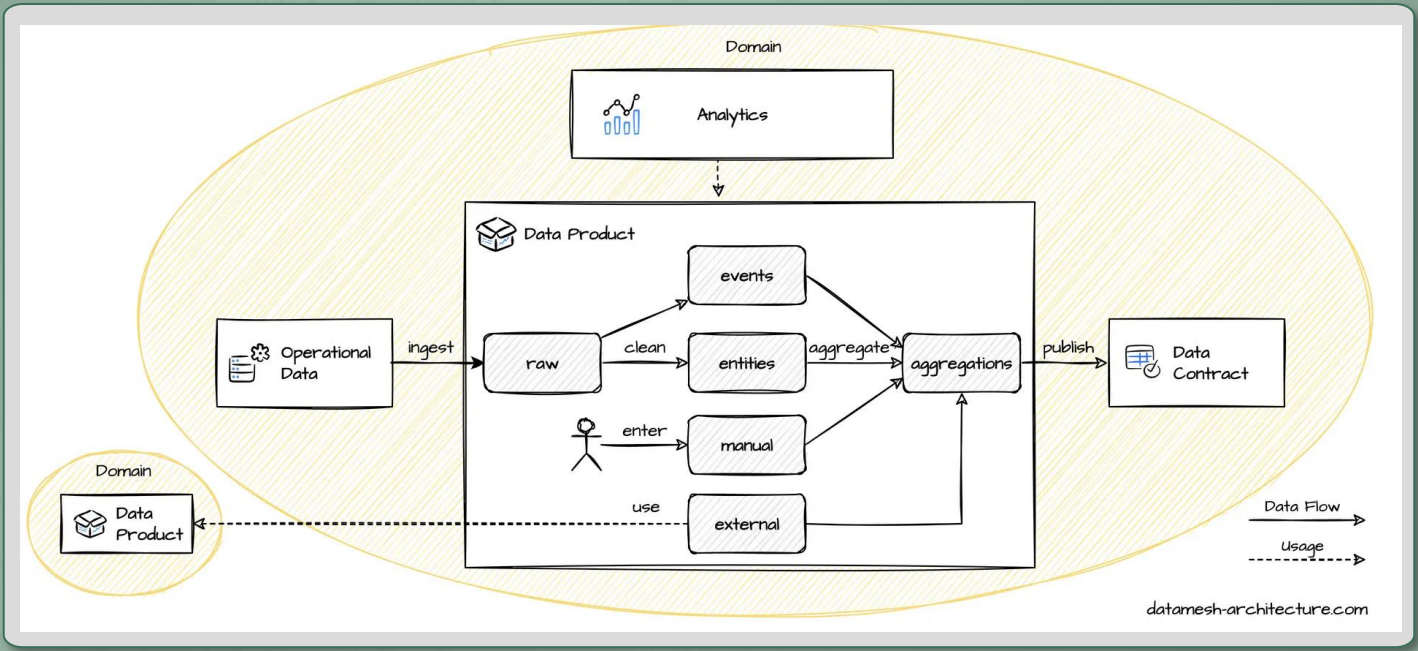
- Visual aids for easier data comprehension.
- Tools for charts, KPIs, dashboards, and reports, like Tableau

Advanced Analytics

- Application of data science and machine learning for deeper insights.
- Correlation, predictions, etc.
- Examples: scikit-learn, PyTorch, TensorFlow.



Data Transformations within domain



Operational data: is typically ingested as raw, unstructured data.

Preprocessing: structures raw data into events and entities

- **Events**: small, immutable, domain-specific (e.g., OrderPurchased, ShipmentDelivered).
- **Entities**:
  represent business objects (e.g., articles) with evolving states. They often shown as snapshot histories; latest snapshot indicates current state.

- **Manual data entries**:
  are common, such as emailed CSV files or text business codes.
- **External data**: data from other teams classified as
- **Aggregations**: merge data for analytical insights.
- Domain data can be shared with other teams using data contract.
- **Data contracts**:
  stable views, consistent despite changes in underlying data models.

Self-Serve Data Platform

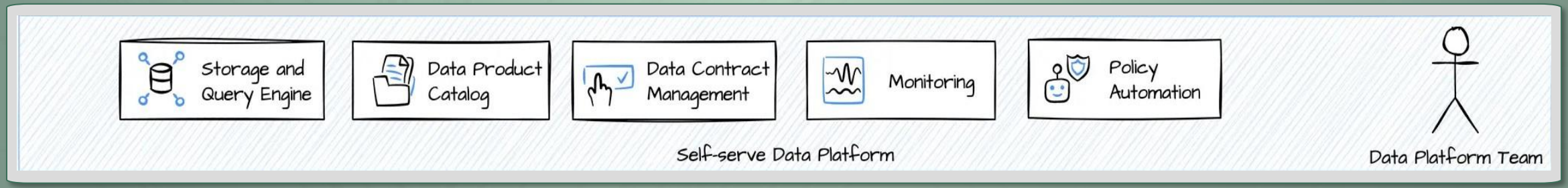
... may vary for each organization

Storage and Query Engine

- Analytical Capabilities
- Enables domain teams to build analytical data model.
- Ingest, store, query, and visualize data.
- Each team gets its isolated area.

Policy Automation

- Automated enforcement of global policies (e.g., standardized metadata, automatic PII removal)



Data Product Catalog

- Create, monitor, discover, access data products.
- Publish and allow other teams to discover data products.
- Implement data catalog via wikis, git, or cloud solutions (e.g., Select Star, AWS Glue Data Catalog).

Efficient Data Product Combination

- Quick cross-domain join operations.
- Shared platforms like Google BigQuery with Google Data Catalog.
- Decentralized options with distributed query engines (e.g., Presto).

Federated Governance

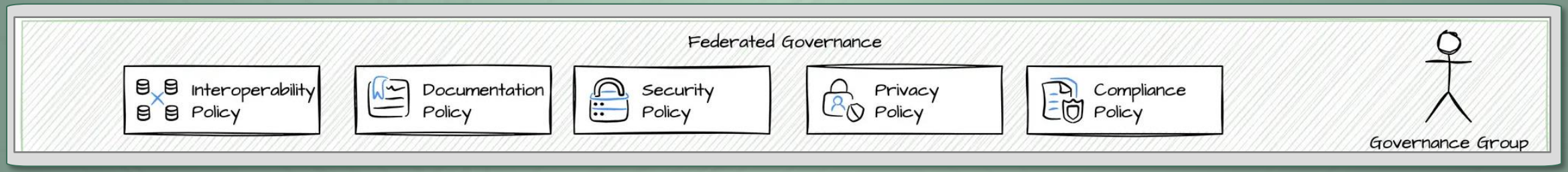
Sets global policies or rules of play are set. Members: representatives from all domain teams.

Policy: Interoperability

- Consistent usage of data products by domain teams.
- Example: CSV on AWS S3 managed by domain teams.

Policy: Access & Security

- Uniform and secure access methods.
- Example: Role-based access via AWS IAM.



Policy: Documentation & Discovery

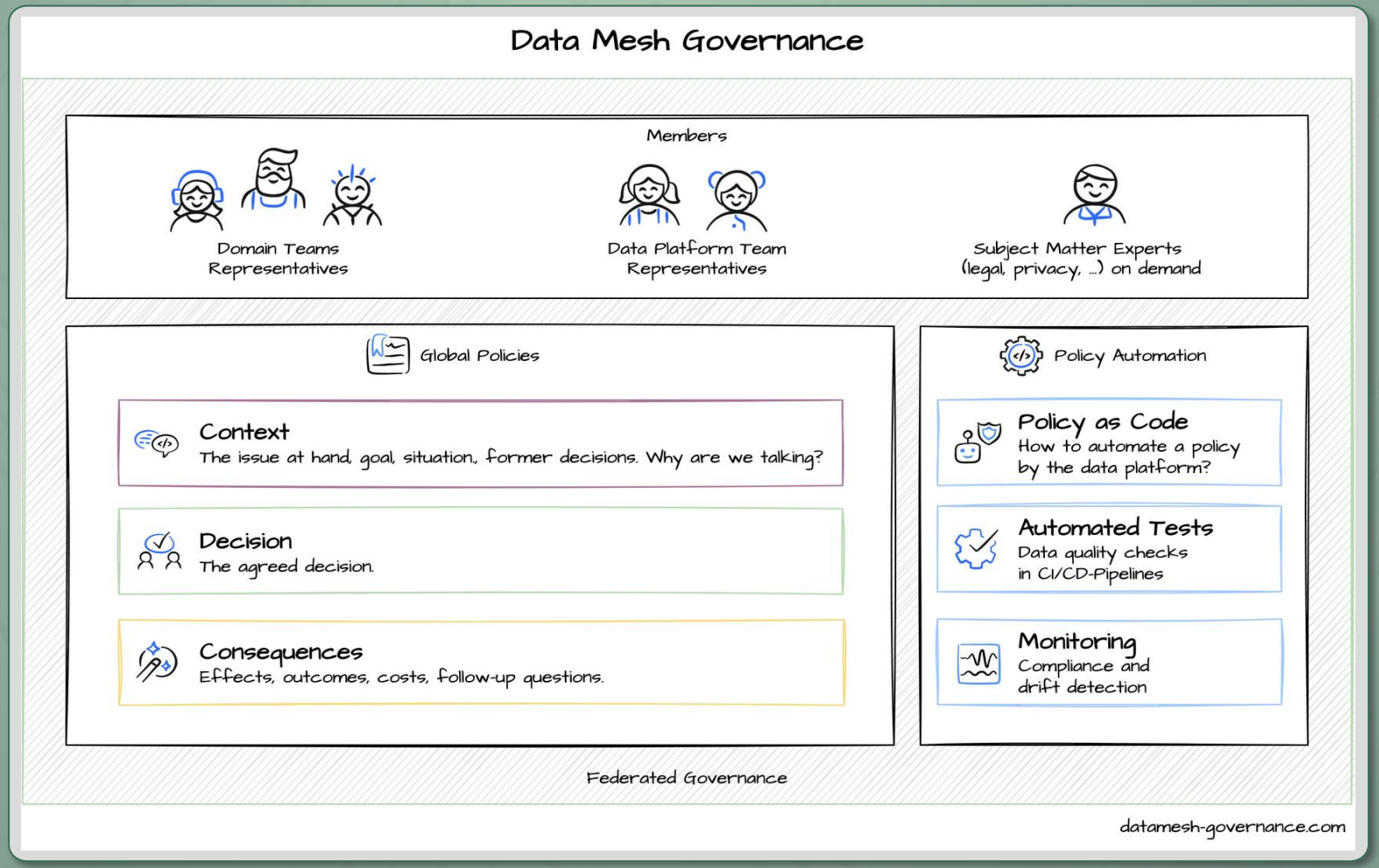
- Required documentation for data product understanding.
- Example: Wiki with metadata - owner, location URL, field descriptions.

Policy: Global Privacy & Compliance

- Protecting PII.
- Meeting industry-specific legal requirements.

Data mesh governance group

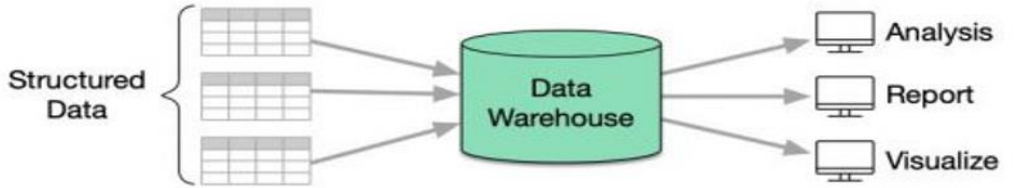
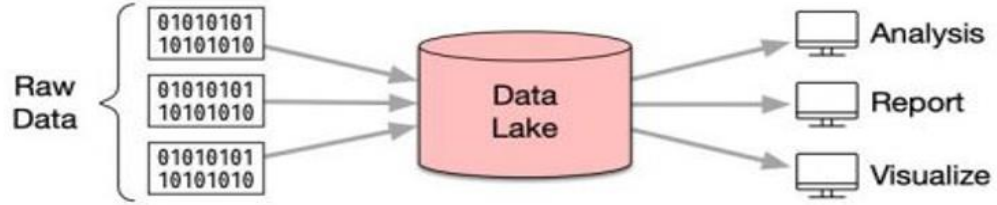
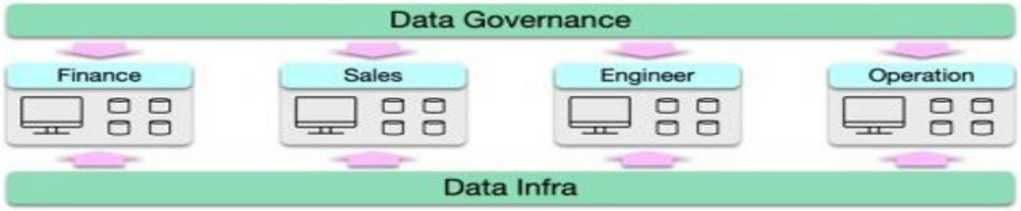
... consists of representatives from domain teams, data platform team with temporary support by subject-matter experts, to address special issues, e.g. legal, compliance, security.



Governance: Impact of Data Mesh style

Governance aspect: Pre Data Mesh	→	Governance aspect: Data Mesh
Centralized team		Federated team
Responsible for data quality		Responsible for defining how to model what constitutes quality
Responsible for data security		Responsible for defining aspects of data security i.e. data sensitivity levels for platform to build in, monitor automatically
Centralized custodianship of data		Federated custodianship of data by domains
Aiming for a well defined static structure of data		Aiming for enabling effective mesh operation embracing a continuously changing and a dynamic topology of the mesh
Governance team is independent from domains		Governance team is made of domains representatives
Centralized technology used by monolithic lake/warehouse		Self-serve platform technologies used by each domain
Responsible for complying with regulation		Responsible for defining the regulation requirements for the platform to build in and monitor automatically

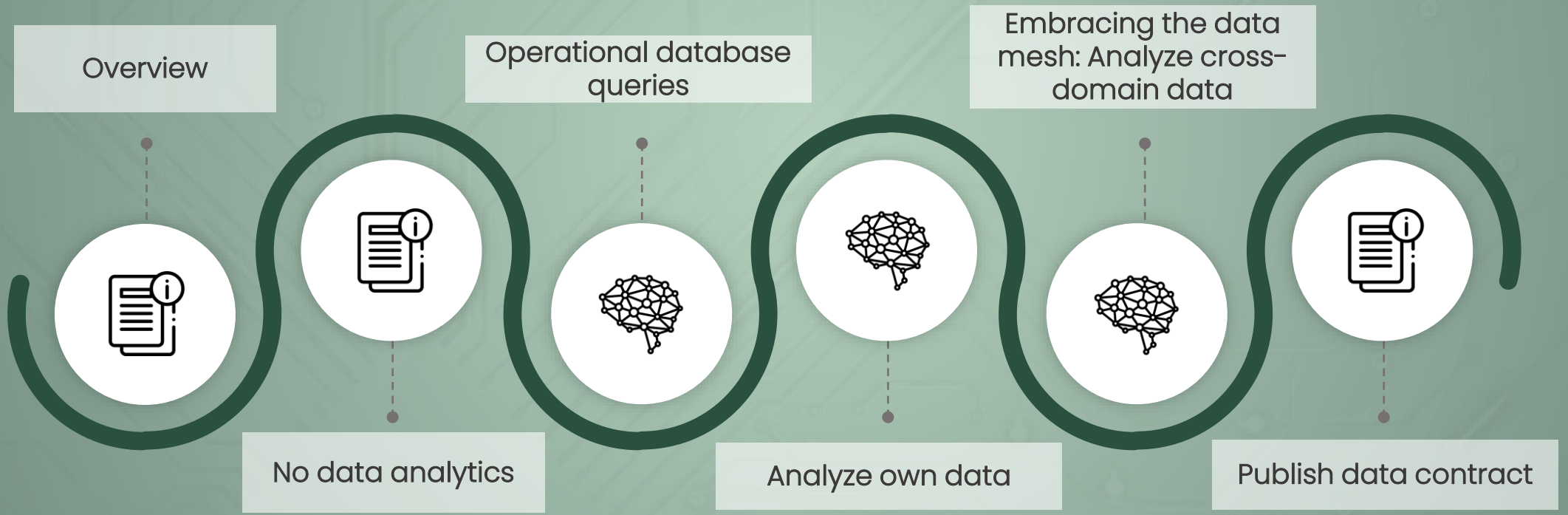
Distinction of Data Mesh from Data Warehouses, Data Lakes

	Explanation	Illustration
Data Warehouse	▪ Large, structured ▪ repository of integrated data ▪ from various sources ▪ used for complex querying ▪ and historical analysis	
Data Lake	▪ Focused, department-specific subset of data warehouse ▪ providing quick data retrieval and analysis	
Data Mesh	▪ Architectural and organizational approach ▪ where data ownership and delivery are decentralized ▪ across domain-specific, cross-functional teams	

Focus

Domain Team's Journey

...involves the organization, summarization, and visualization of data. It provides simple summaries about the sample and the measures.



Overview: Domain Team's Journey

Individualized Journey

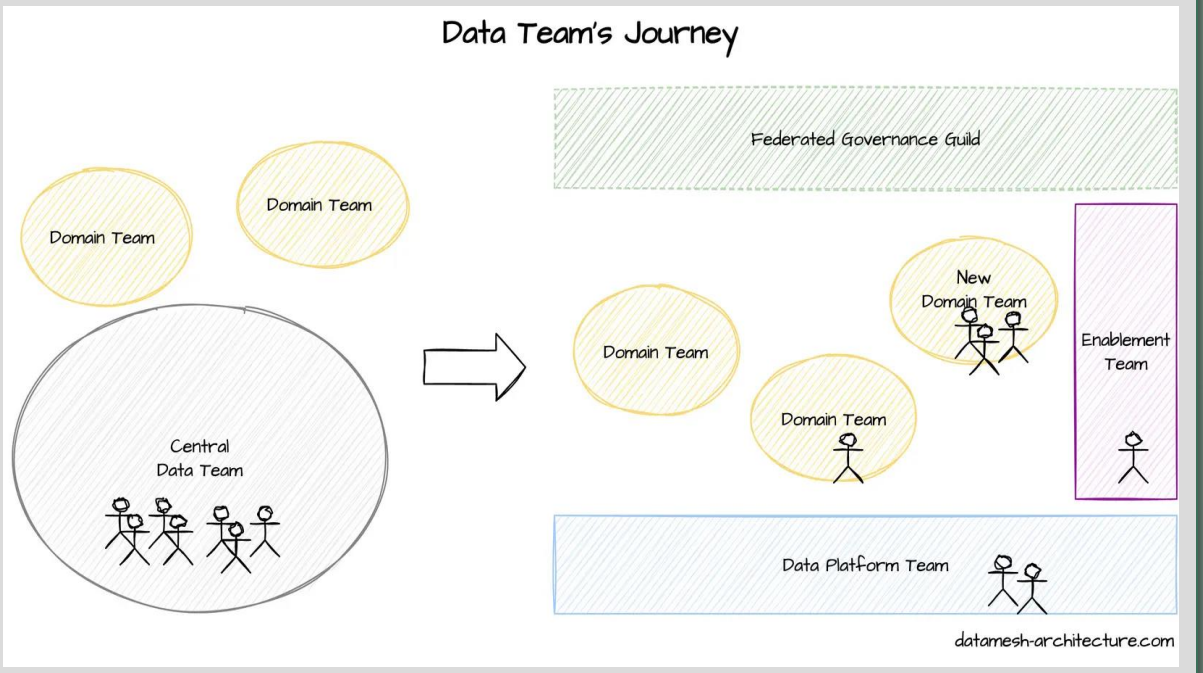
- Domain teams embark at their convenience.
- Flexibility in pace and readiness.

Momentum of Data-Driven Insights

- Immediate advantages from initial data-driven decisions.
- Catalyzing the usage of richer data for profound insights.
- Shared data products amplifying the data mesh ecosystem.

Key Success Factors

- Unified Vision:
  Clear data mesh objective from leadership to align efforts.
- Supportive Infrastructure:
  Intuitive self-serve data platform fostering analytical learning.
- Trust & Autonomy:
  Encouraging teams to chart their journey at their comfort.



No data analytics

Level 0: The starting point

Domain Responsibility

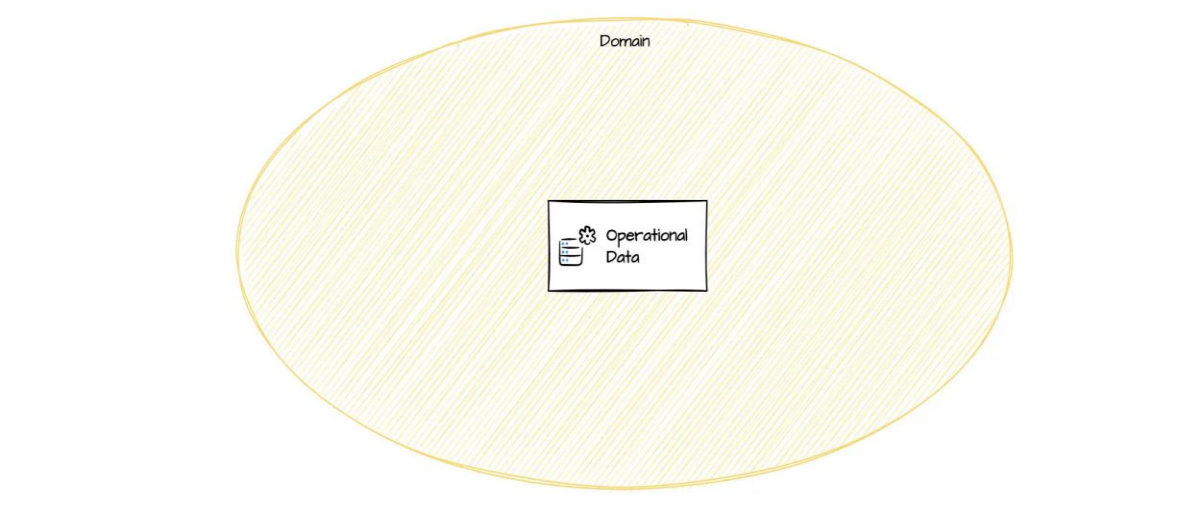
- Team oversees a specific domain.
- Constructs and manages self-contained systems.
- Inclusive of required infrastructure.

Intensive System Development

- Significant effort invested in system creation.
- Primary focus on delivery excellence.
- Operational systems now produce domain data.

Absence of Data Analytics

- Relevance of data analytics not realized.



Operational database queries

Level 1: Balancing operational and analytical needs without compromising system performance

Analytical Needs in Production

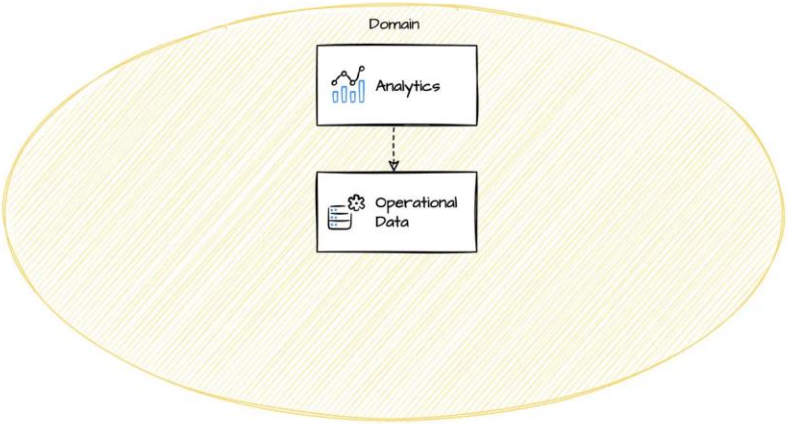
- Investigating incidents.
- Analyzing customer impact.
- Stakeholder data inquiries

Consequences of Analytical Queries on Production Database

- Increased system load.
- Potential disruption to primary operations.

Mitigation Strategies & Drawbacks

- Modify production database for better support by additional indices, etc.
- Drawbacks:
  - Analytical queries remain slow.
  - Cumbersome query formulation.



Analyze own data

Level 2: Utilizing self-serve data platform empowers teams to make informed decisions efficiently.

Problem

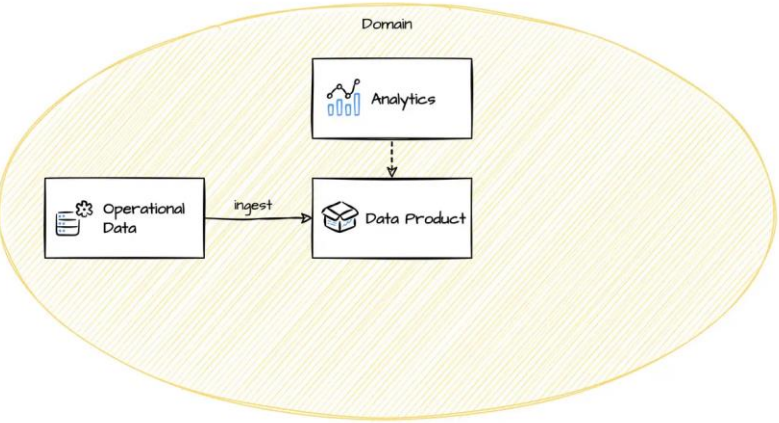
- Slow, hard-to-write analytical queries.

Solution: Self-Serve Data Platform

- Introduction to platforms like Google BigQuery.
- Ingest data via sources like Kafka.
- Creation of first data product.

Benefits

- Maintain fast and easy-to-manage queries.
- No alteration of operational database schemas.
- Gain proficiency in:
  - Data structuring, preprocessing, and cleaning.
  - Data analytics and visualization.
  - Note: Mostly through SQL, a familiar language.



Embracing the Data Mesh

Level 3: Cross-domain data analysis strengthens the collective intelligence of an organization

Beyond Solo Analysis

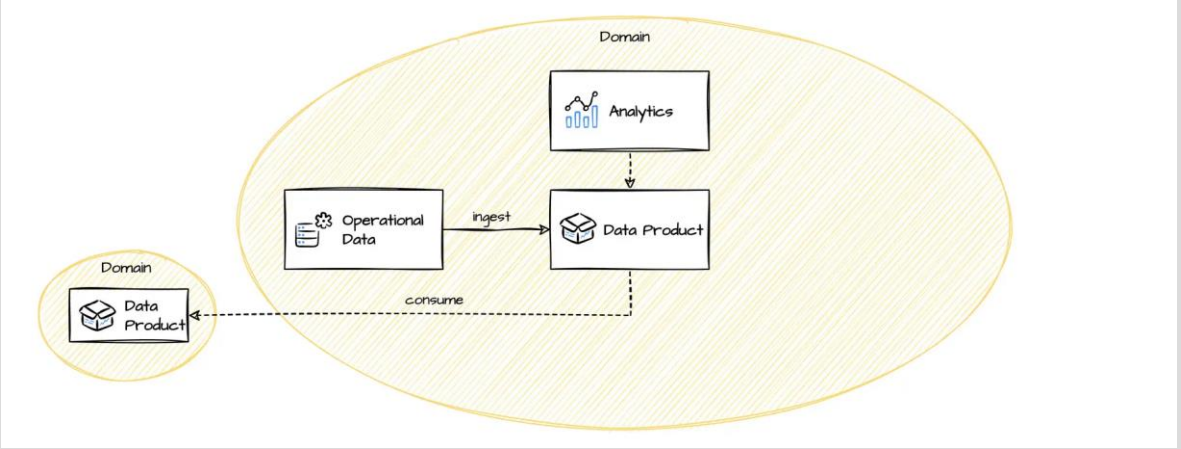
- Combining own domain data with others reveals deeper insights.
- Example use-cases:
  - A/B tests to measure UI change effects on conversion rates.
  - Fraud detection models using purchase history & click stream data.

Dual Role in the Data Mesh

- Transition from consuming data products to publishing them.
- Generating value for other teams, but also taking on increased responsibilities.

Engaging in Governance

- Active participation and contribution in the federated governance group is crucial for alignment.



Publishing data contract

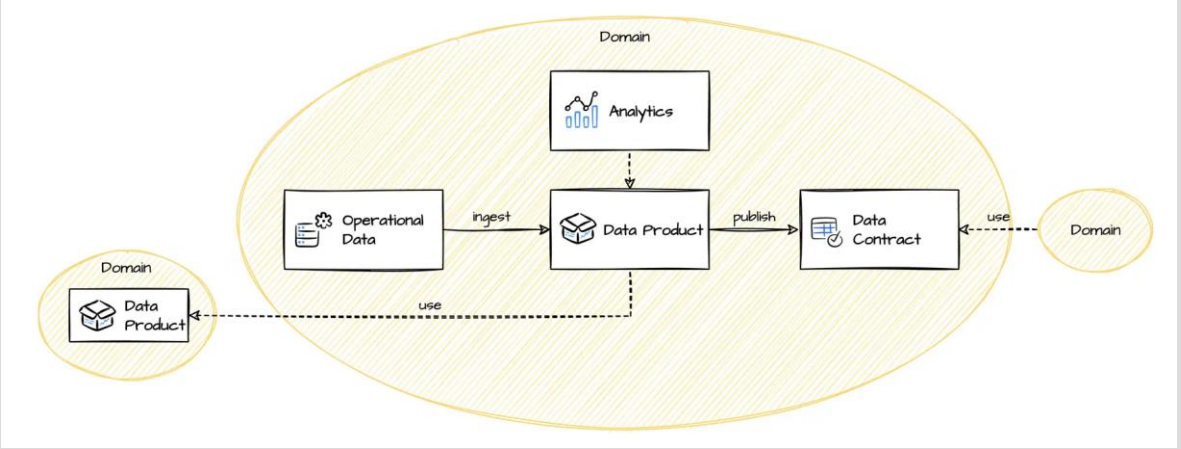
Level 4: The Journey from Consumer to Publisher in Data Mesh

Sharing Data with Others

- Responding to needs of other teams by publishing data contracts.
- Examples:
  Sharing data on confirmed, rejected, aborted orders, etc.

Solution: Self-Serve Data Platform

- Introduction to platforms like Google BigQuery.
- Ingest data via sources like Kafka.
- Creation of first data product.

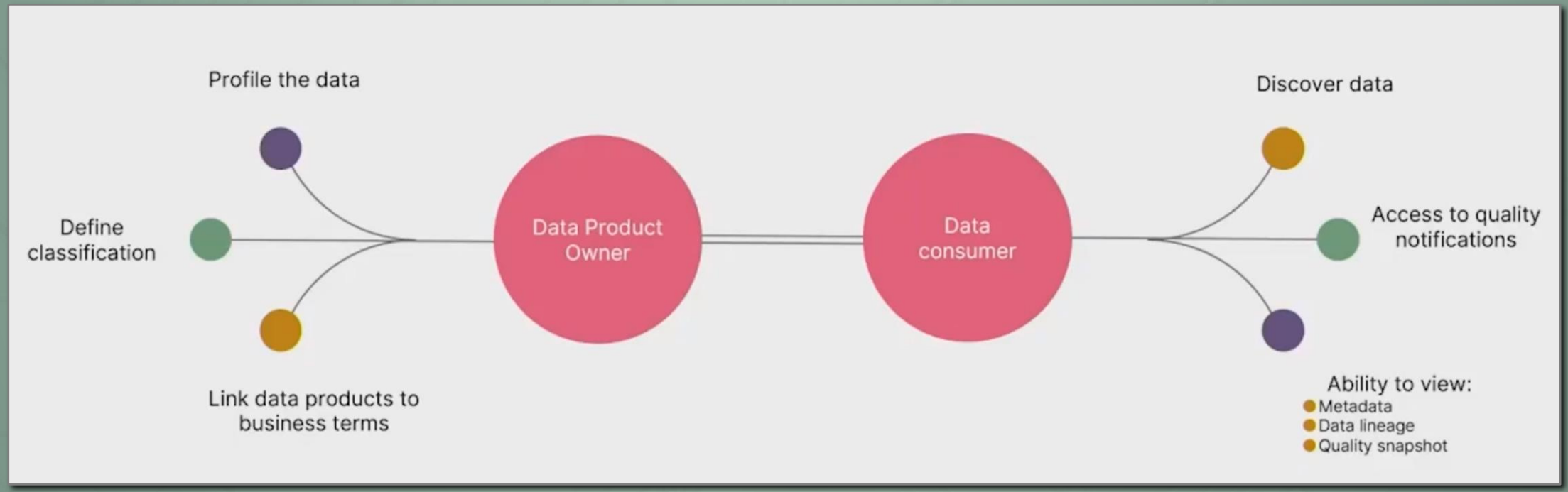


Practical examples

...involves the organization, summarization, and visualization of data. It provides simple summaries about the sample and the measures.

Typical Data pipeline

Enables domain teams to publish transparent, trustworthy data for customers, e.g. to facilitate onboarding of new partners. Customers are not only retail customers, but also other team within the company in B2B environment

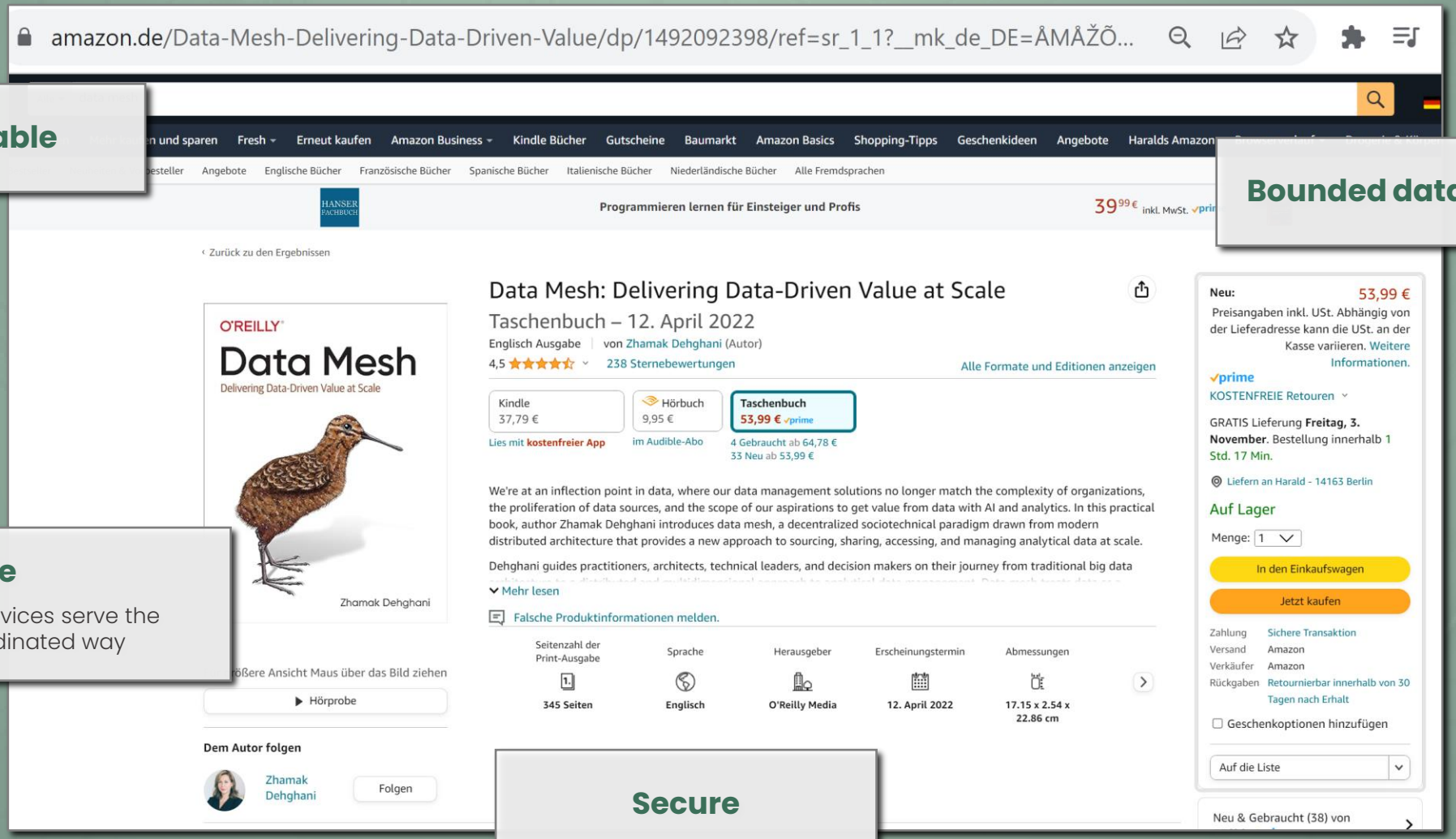


Data Product example: Amazon book store

Provision of online store information for book sales should fulfill properties in sense of data mesh

Discoverable

Bounded data scope



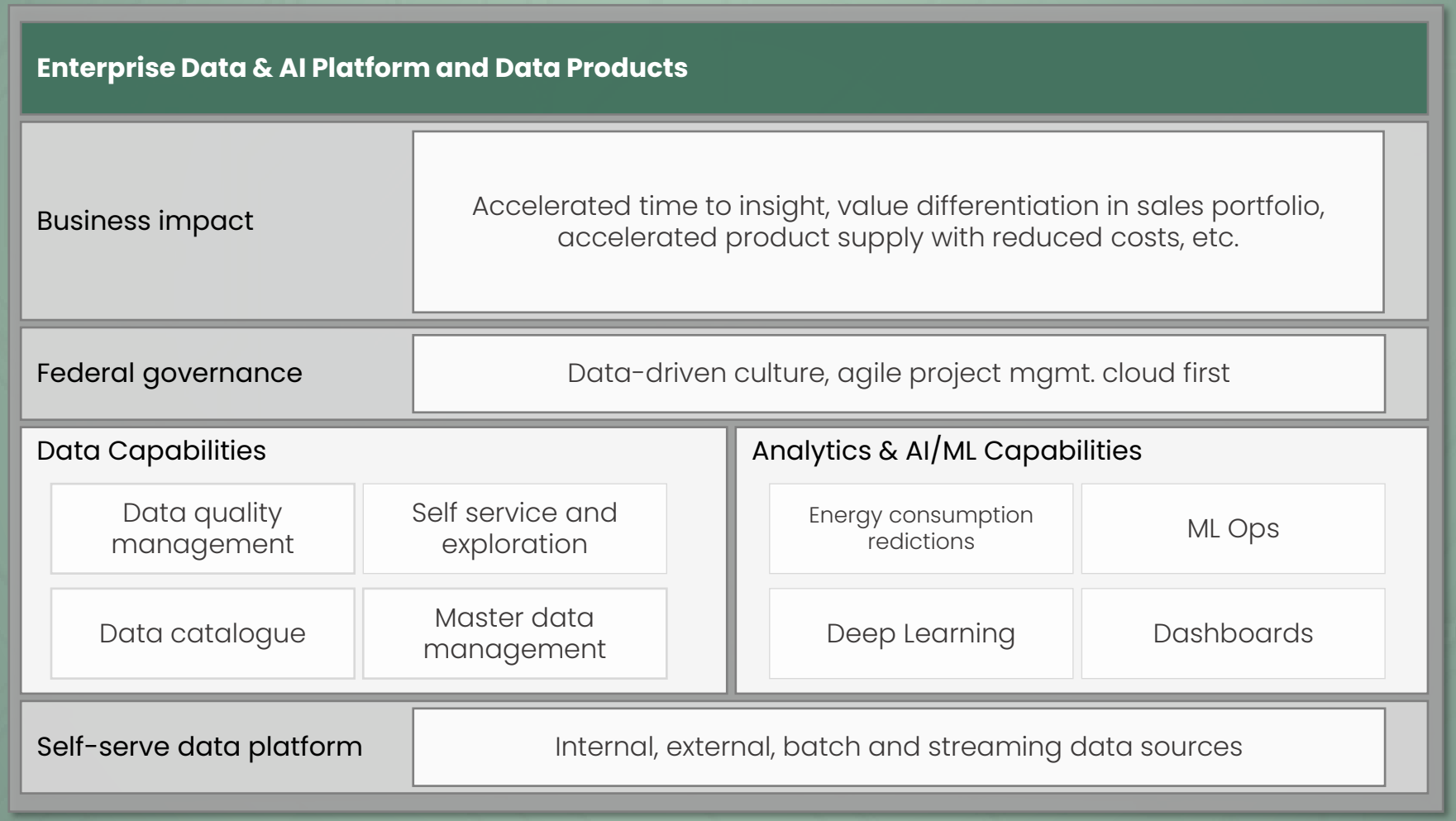
Interoperable

- Lots of microservices serve the surface in coordinated way

Secure

Enterprise Data & AI Platform

Provision of online store information for book sales should fulfill properties in sense of data mesh



Links

i.e. sources for self-learning

	Title	Link
Data Mesh Overview	Data Mesh: Delivering Data-Driven Value at Scale, Zhamak Dehghani	https://www.amazon.de/Data-Mesh-Delivering-Data-Driven-Value/dp/1492092398/ref=sr_1_1?__mk_de_DE=%C3%85M%C3%85%C5%BD%C3%95%C3%91&crid=17AXZI48S51TV&keywords=data+mesh&qid=1698773260&sprefix=data+mesh%2Caps%2C102&sr=8-1
	Data Mesh Architecture - Data Mesh From an Engineering Perspective	https://www.datamesh-architecture.com/
	Data Mesh Governance by Example	https://www.datamesh-governance.com/
	Bring data teams together	https://www.datamesh-manager.com/
	What Is a Data Mesh?	https://www.dremio.com/resources/guides/what-is-a-data-mesh/
	The Definition of Data Mesh: What Is It and Why Do I Need One?	https://www.atscale.com/blog/data-mesh-definition/
	Data Mesh 101	https://www.slideshare.net/ChrisFord803185/data-mesh-101
	Book: Data Mesh in Action - Majchrzak, 2022	https://www.manning.com/books/data-mesh-in-action

Links

i.e. sources for self-learning

	Title	Link
Data Product	The Anatomy of a Data Product	https://towardsdatascience.com/the-anatomy-of-a-data-product-d3140f068311
	Data Contract Specification	https://datacontract.com/
	Driving Data Quality with Data Contracts: A comprehensive guide to building reliable, trusted, and effective data platforms	https://www.amazon.com/dp/B0C37FPH3D
	Open Data Contract Standard	https://github.com/bitol-io/open-data-contract-standard
	Data Contracts: A Bridge Connecting Two Worlds	https://medium.com/@atanas.iliev.ai/data-contracts-a-bridge-connecting-two-worlds-404eff1d970d

	Title	Link
Domain-oriented Ownership	Why Your Data Mesh Needs Domain-oriented Ownership: 7 Reasons to Adopt this Principle	https://www.linkedin.com/pulse/why-your-data-mesh-needs-domain-oriented-ownership-7-reasons-deepak/

	Title	Link
Data governance	Decentralized Data Governance as Part of a Data Mesh Platform: Concepts and Approaches	https://www.researchgate.net/publication/371399084_Decentralized_Data_Governance_as_Part_of_a_Data_Mesh_Platform_Concepts_and_Approaches
	Webinar: Five Things to Consider About Data Mesh and Data Governance	https://www.dataversity.net/webinar-five-things-to-consider-about-data-mesh-and-data-governance-2/

ChatGPT/Dall-E3 Prompts

Sophisticated portrayal of a data mesh platform, capturing the core of decentralized data ownership. Various nodes, each of a different size, are tied together by shimmering links, emphasizing the continuous data movement. Next to these nodes, distinct team clusters can be observed, each glowing in a gradient from a lively green to a muted gray, highlighting the decentralized data teams' involvement in the data management process.

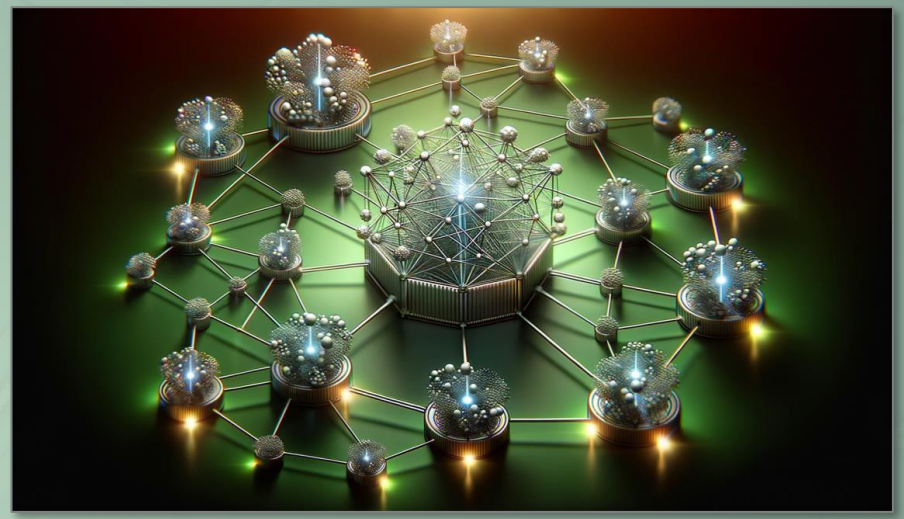
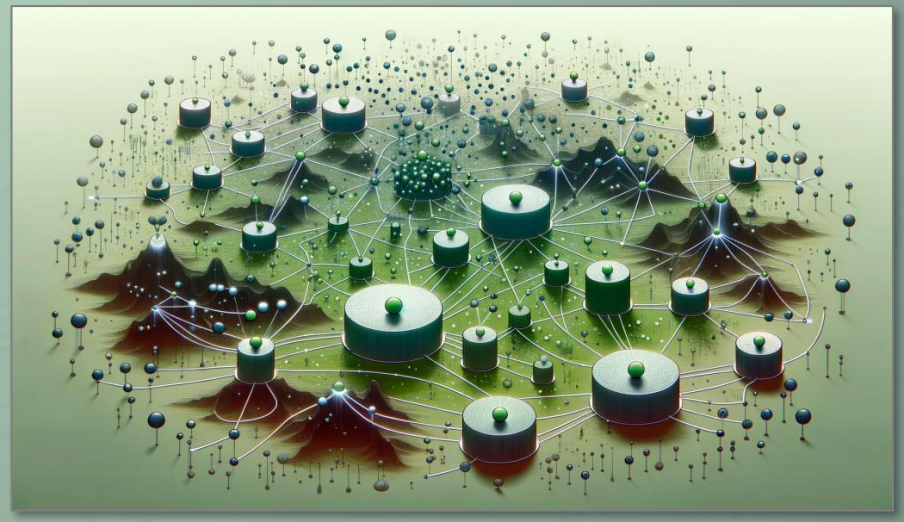


Illustration of a sprawling data mesh landscape where nodes, representing data repositories, are connected by intricate links. Scattered around these nodes are clusters symbolizing decentralized data teams. The color of the nodes, links, and team clusters blend from a rich green to a subdued gray, showcasing the collaborative yet independent nature of these teams within the data ecosystem.



About me

Dr. Harald Stein

- Data Scientist ~ 7 years experience
- Algotrader ~ 4 years experience
- Ph.D. in Economics, Game Theory

- LinkedIn: https://www.linkedin.com/in/harald-stein-phd-1648b51a
- ResearchGate: https://www.researchgate.net/profile/Harald-Stein