SummerSoC / 2025-06-17

Data Products and Data Contracts

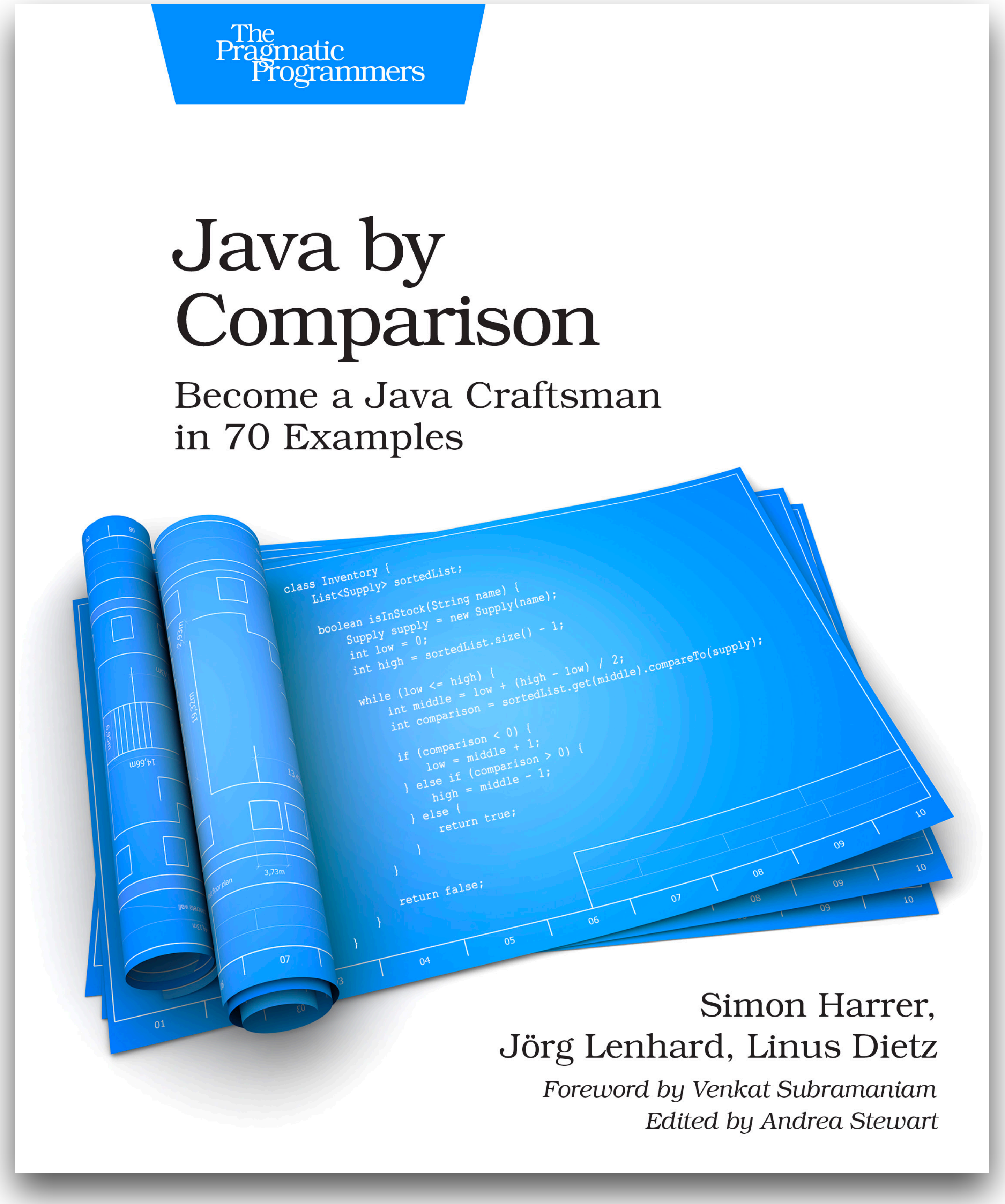
Open Standards & Implementations

Warning: There will be YAML!

DR. SIMON HARRER

/IN/SIMONHARRER

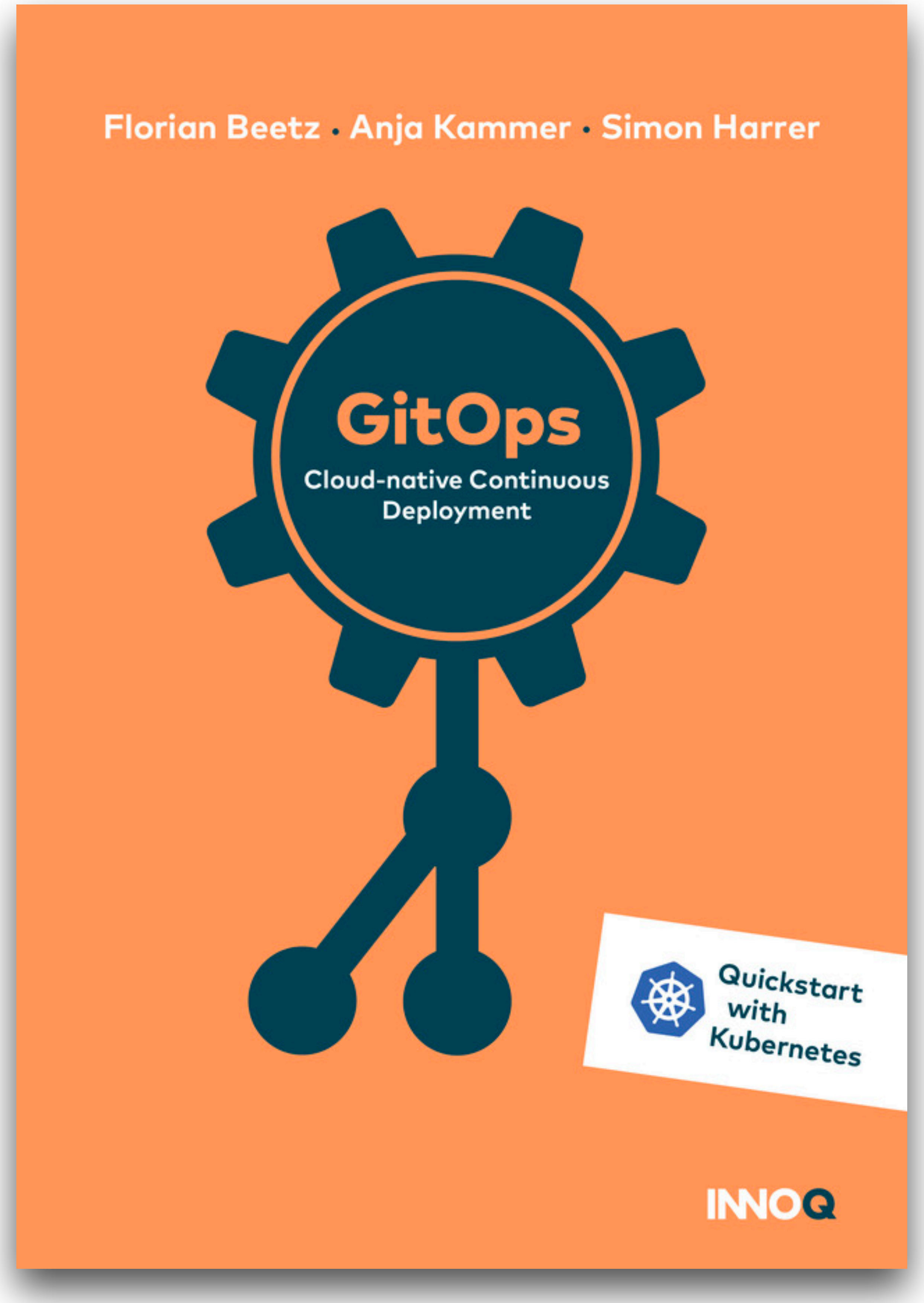
From Software Engineering …


The Pragmatic Programmers
Java by Comparison
Become a Java Craftsman in 70 Examples
Simon Harrer, Jörg Lenhard, Linus Dietz
Foreword by Venkat Subramaniam
Edited by Andrea Stewart

java.by-comparison.com

Jochen Christ
Simon Harrer
Martin Huber
Remote Mob Programming
At home, but not alone
INNOQ
Foreword by Mark Pearl

remotemobprogramming.org

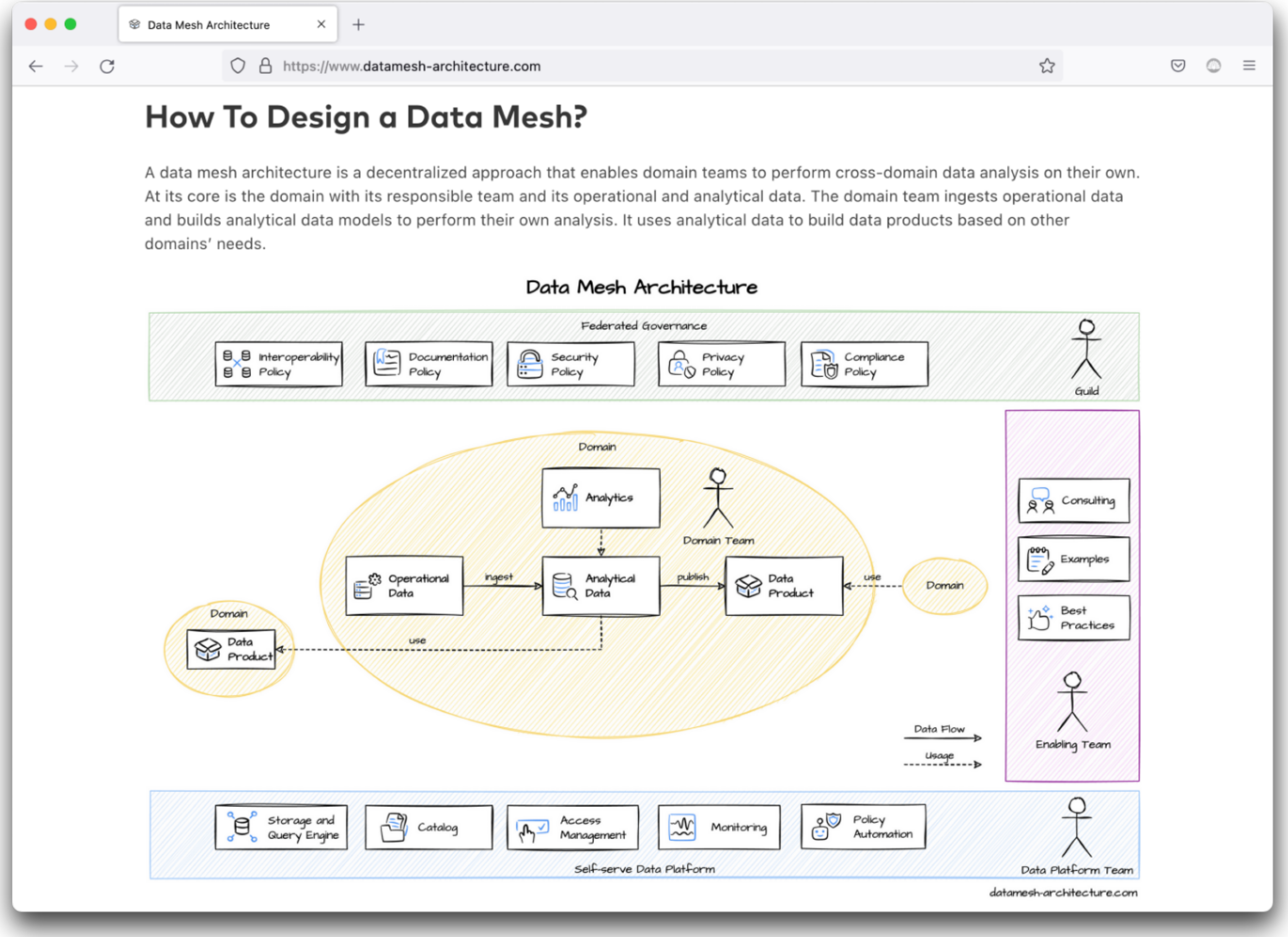

Florian Beetz · Anja Kammer · Simon Harrer
GitOps
Cloud-native Continuous Deployment
Quickstart with Kubernetes
INNOQ

gitops.tech

2021

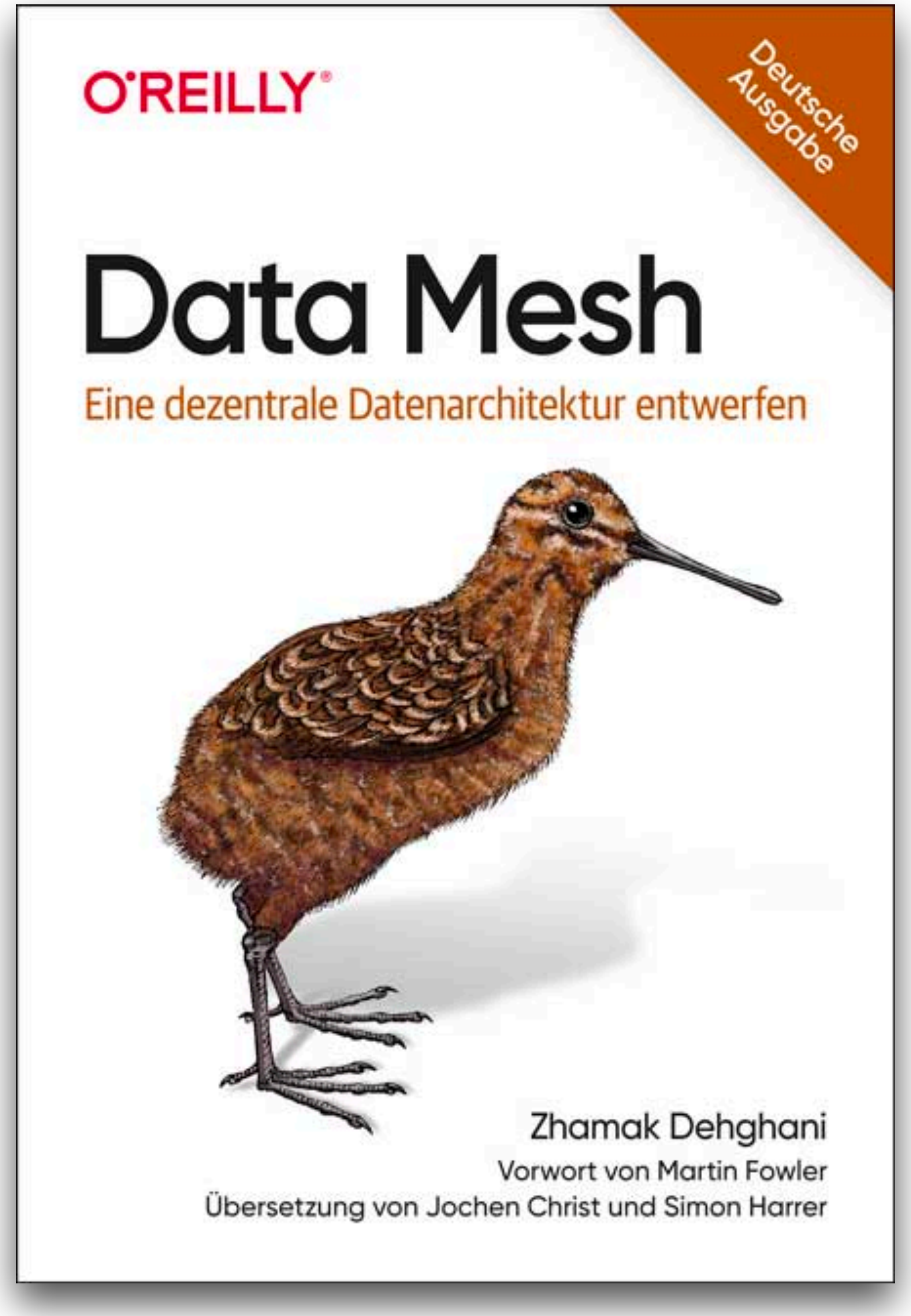
Photo by Krisjanis Mezulis on Unsplash

… to Data Mesh …



How To Design a Data Mesh?

A data mesh architecture is a decentralized approach that enables domain teams to perform cross-domain data analysis on their own. At its core is the domain with its responsible team and its operational and analytical data. The domain team ingests operational data and builds analytical data models to perform their own analysis. It uses analytical data to build data products based on other domains' needs.



O'REILLY®

Deutsche Ausgabe

Data Mesh

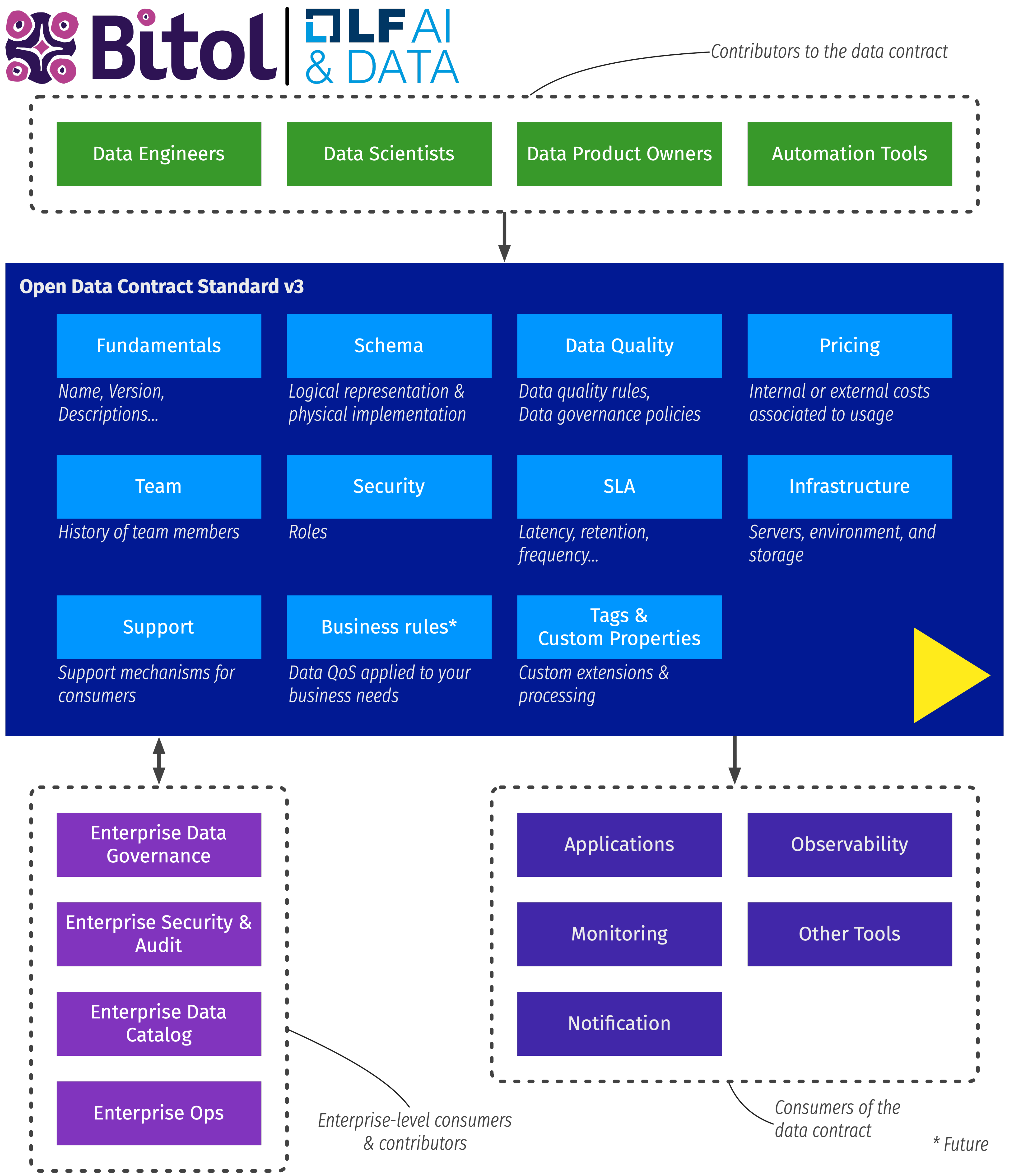
Eine dezentrale Datenarchitektur entwerfen

Zhamak Dehghani
Vorwort von Martin Fowler
Übersetzung von Jochen Christ und Simon Harrer

datamesh-architecture.com

oreilly.de/produkt/data-mesh

... to Open Standards and Tools



Bitol | LF AI & DATA

Contributors to the data contract

Data Engineers | Data Scientists | Data Product Owners | Automation Tools

Open Data Contract Standard v3

- **Fundamentals** – *Name, Version, Descriptions...*
- **Schema** – *Logical representation & physical implementation*
- **Data Quality** – *Data quality rules, Data governance policies*
- **Pricing** – *Internal or external costs associated to usage*
- **Team** – *History of team members*
- **Security** – *Roles*
- **SLA** – *Latency, retention, frequency...*
- **Infrastructure** – *Servers, environment, and storage*
- **Support** – *Support mechanisms for consumers*
- **Business rules*** – *Data QoS applied to your business needs*
- **Tags & Custom Properties** – *Custom extensions & processing*

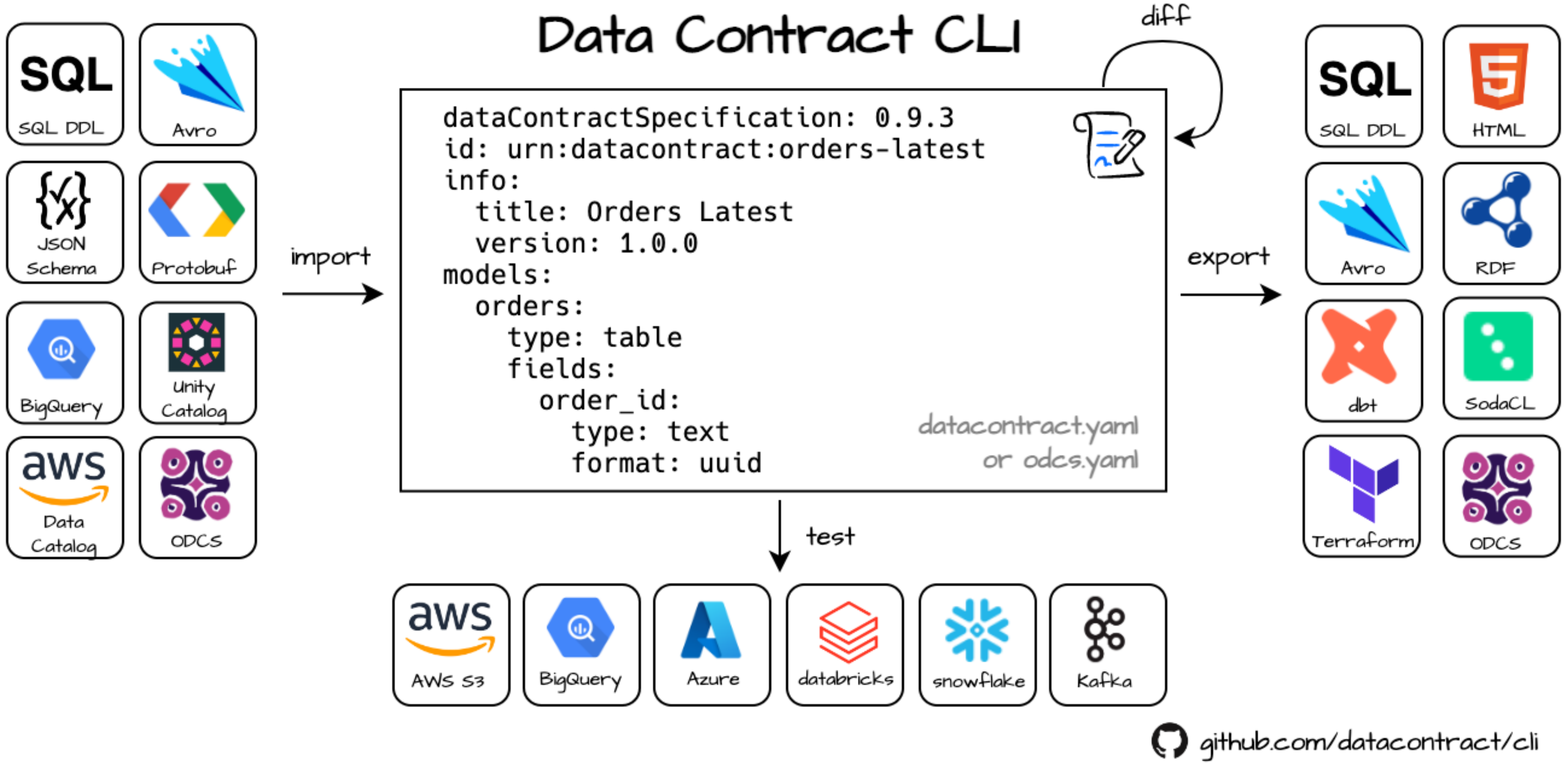
Enterprise Data Governance | Enterprise Security & Audit | Enterprise Data Catalog | Enterprise Ops

Enterprise-level consumers & contributors

Applications | Observability | Monitoring | Other Tools | Notification

Consumers of the data contract

** Future*



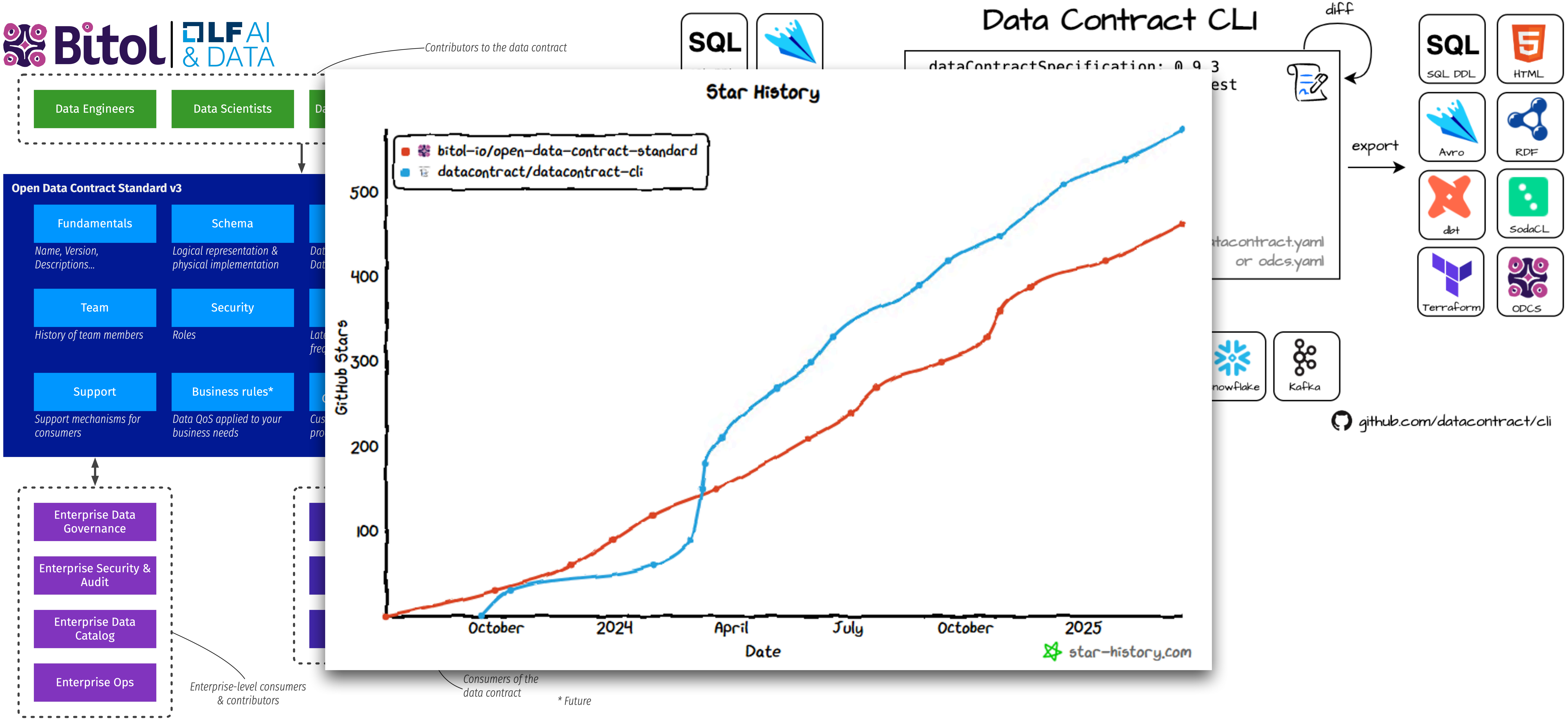
Data Contract CLI

```
dataContractSpecification: 0.9.3
id: urn:datacontract:orders-latest
info:
  title: Orders Latest
  version: 1.0.0
models:
  orders:
    type: table
    fields:
      order_id:
        type: text
        format: uuid
```

datacontract.yaml or odcs.yaml

github.com/datacontract/cli

... to Open Standards and Tools



I’ll talk about data mesh,

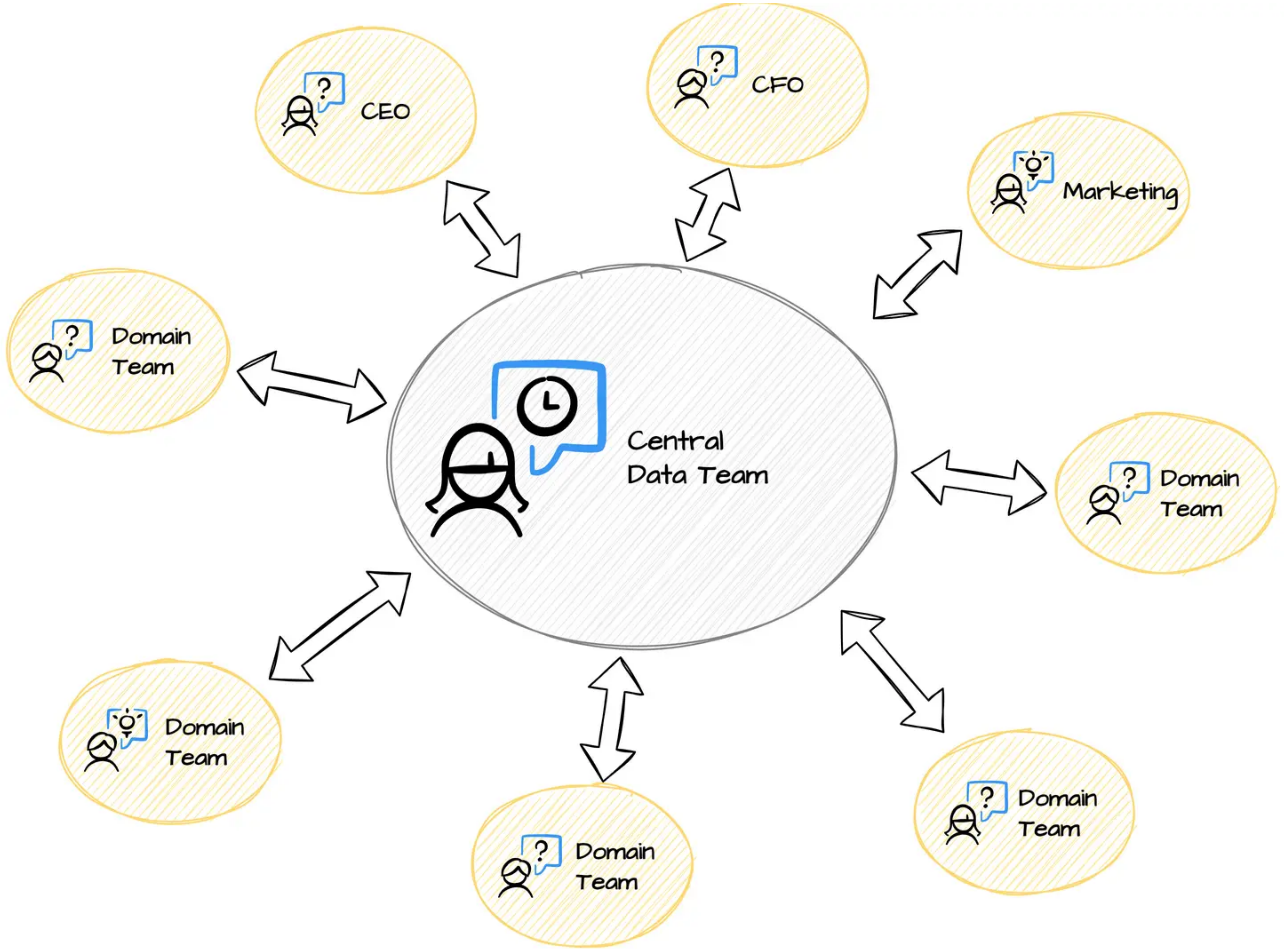
data products,

and data contracts.

And, if time allows, using AI in this space.

Part 1: Data Mesh

Status Quo


CEO
CFO
Marketing
Domain Team
Central Data Team
Domain Team
Domain Team
Domain Team
Domain Team

The Problem

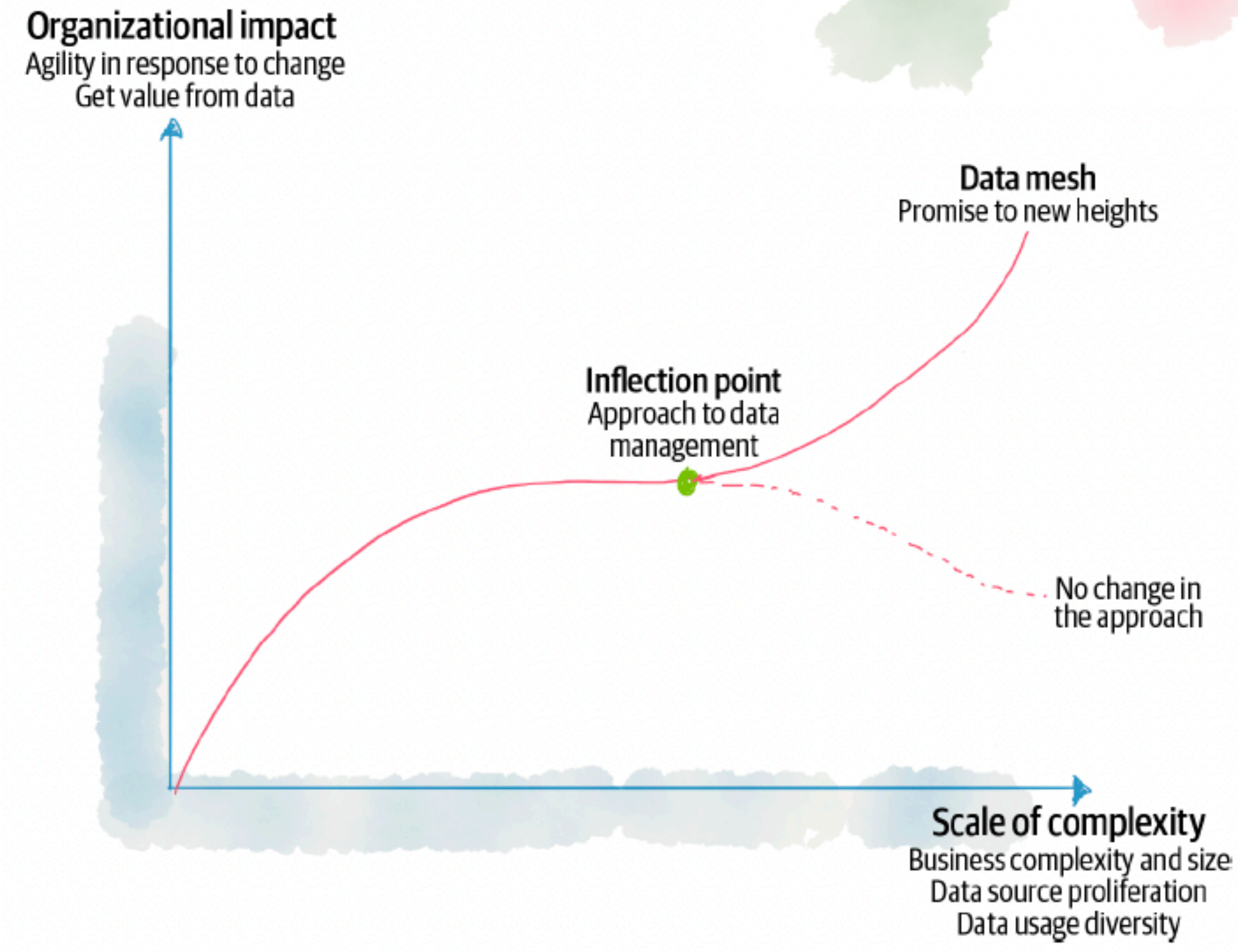
Great expectations of data
Diverse and wide applications of ML and analytics

Great divide of data
Complexity risen from fragmentation of operational and analytical data

Scale
Large scale data source proliferation

Business complexity & volatility
Continuous change and growth of businesses

Discord between data investments and returns
Expensive data solutions lacking impact



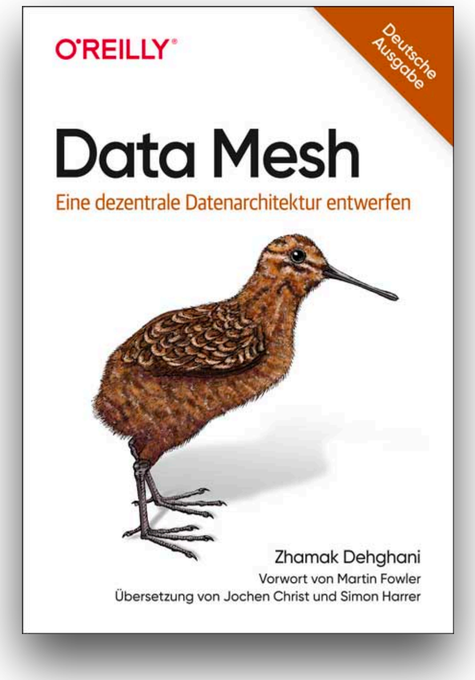
Organizational impact
Agility in response to change
Get value from data

Data mesh
Promise to new heights

Inflection point
Approach to data management

No change in the approach

Scale of complexity
Business complexity and size
Data source proliferation
Data usage diversity



O'REILLY

Deutsche Ausgabe

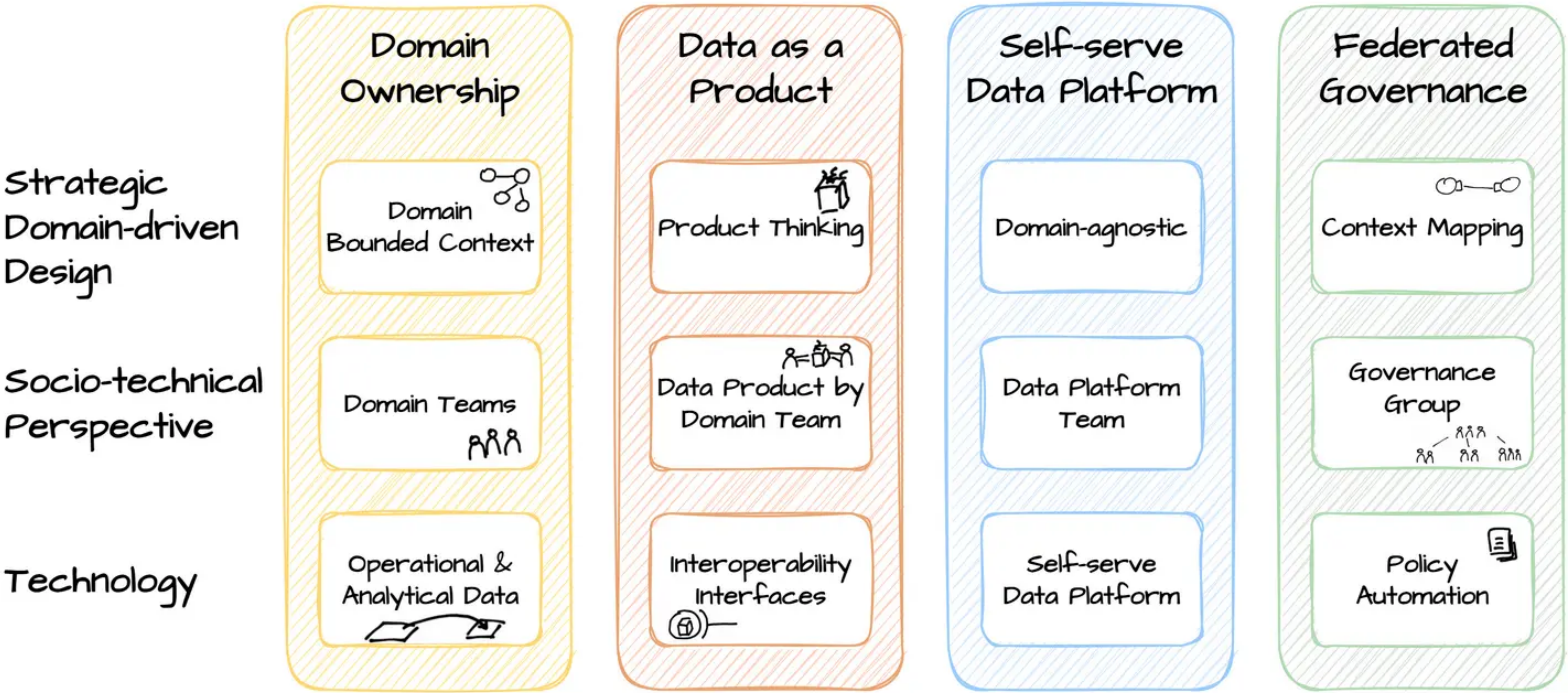
Data Mesh

Eine dezentrale Datenarchitektur entwerfen

Zhamak Dehghani
Vorwort von Martin Fowler
Übersetzung von Jochen Christ und Simon Harrer

The Solution

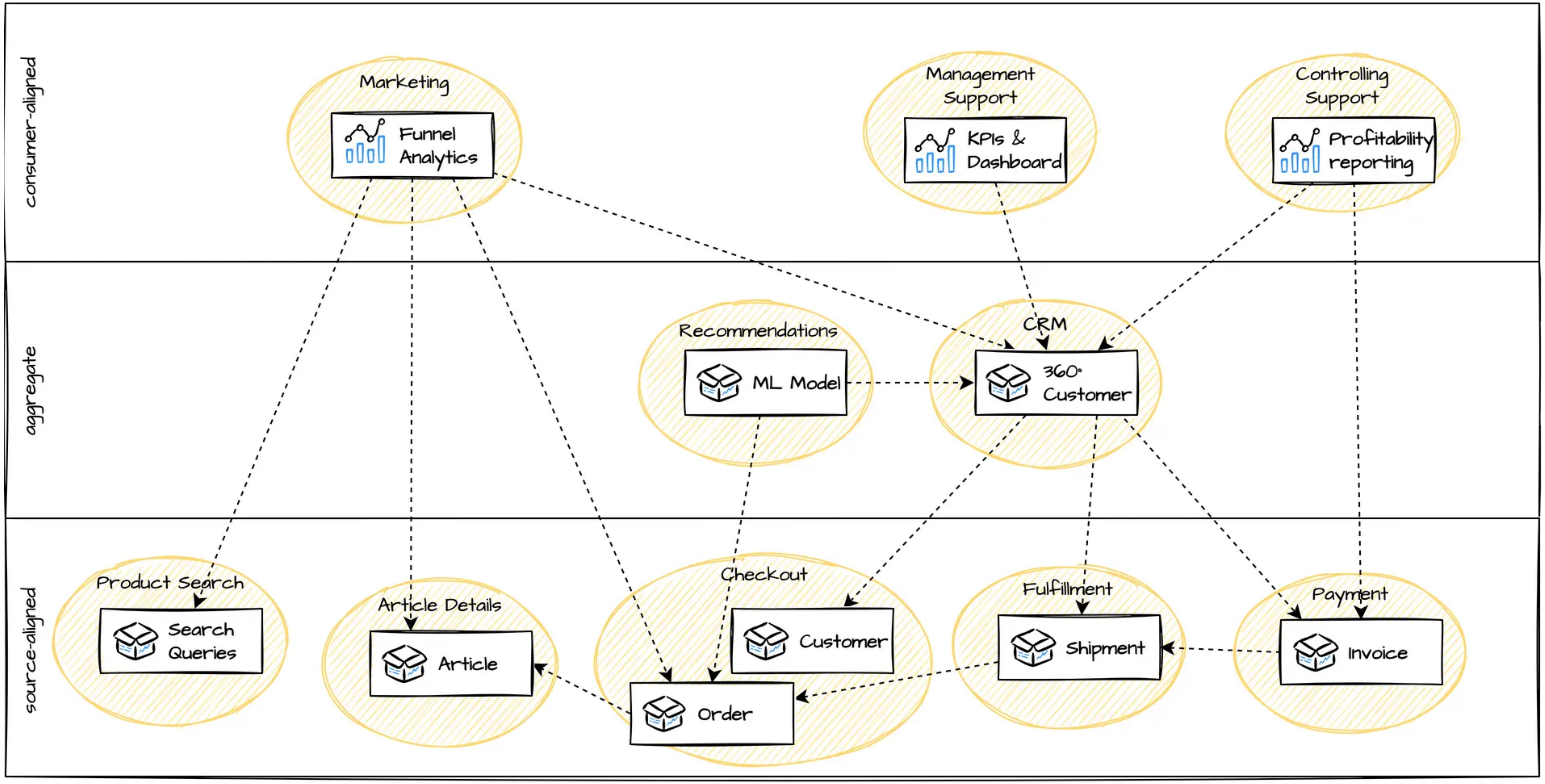
What Is Data Mesh?



	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	Operational & Analytical Data	Interoperability Interfaces	Self-serve Data Platform	Policy Automation

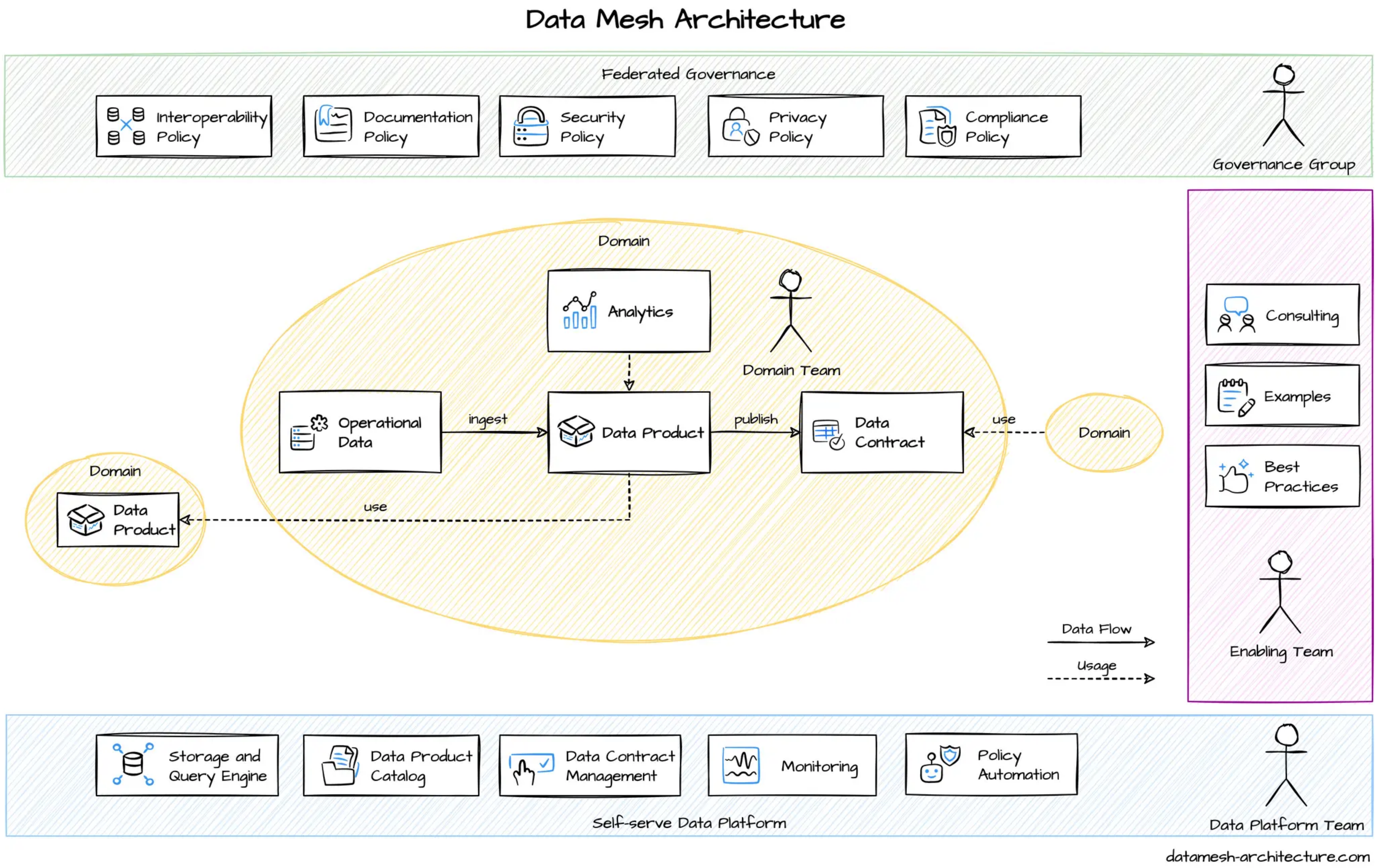
datamesh-architecture.com

I brought a Data Mesh with me



datamesh-architecture.com

The Architecture Perspective



In summary: the vast majority of larger companies try moving towards the principles of data mesh.

But be aware of the name data mesh.

Data Mesh
decentralized data architecture with domain ownership, data as a product, self-serve data platform, and federated computational governance
imgflip.com

Part 2: Data Products

What is a data product?

My own practical definition:

A data product is a logical unit

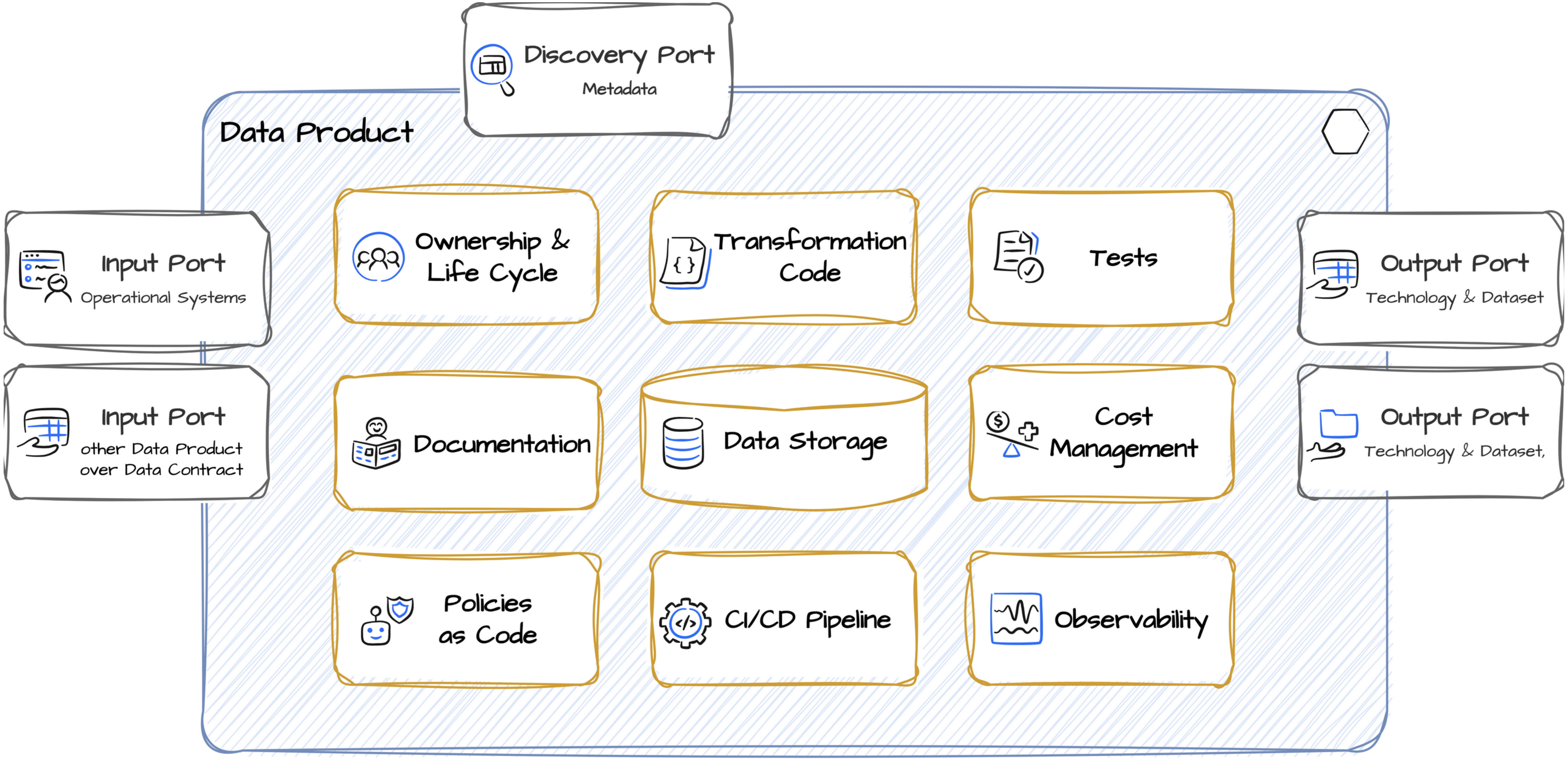
that contains all components to process domain data

and provide data sets via output ports.

(The view of a software architect)

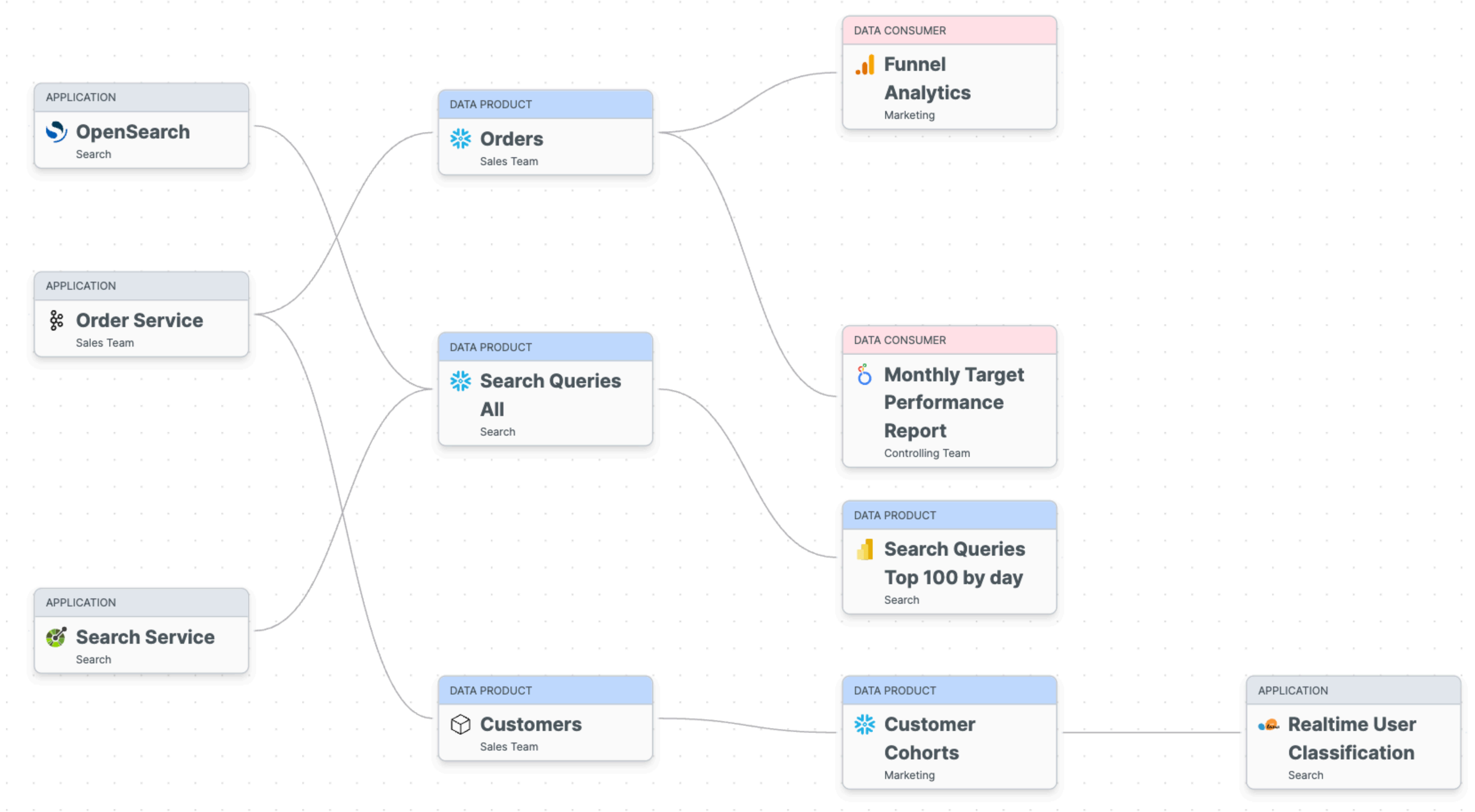
https://www.datamesh-manager.com/learn/what-is-a-data-product

I see them as an architectural unit

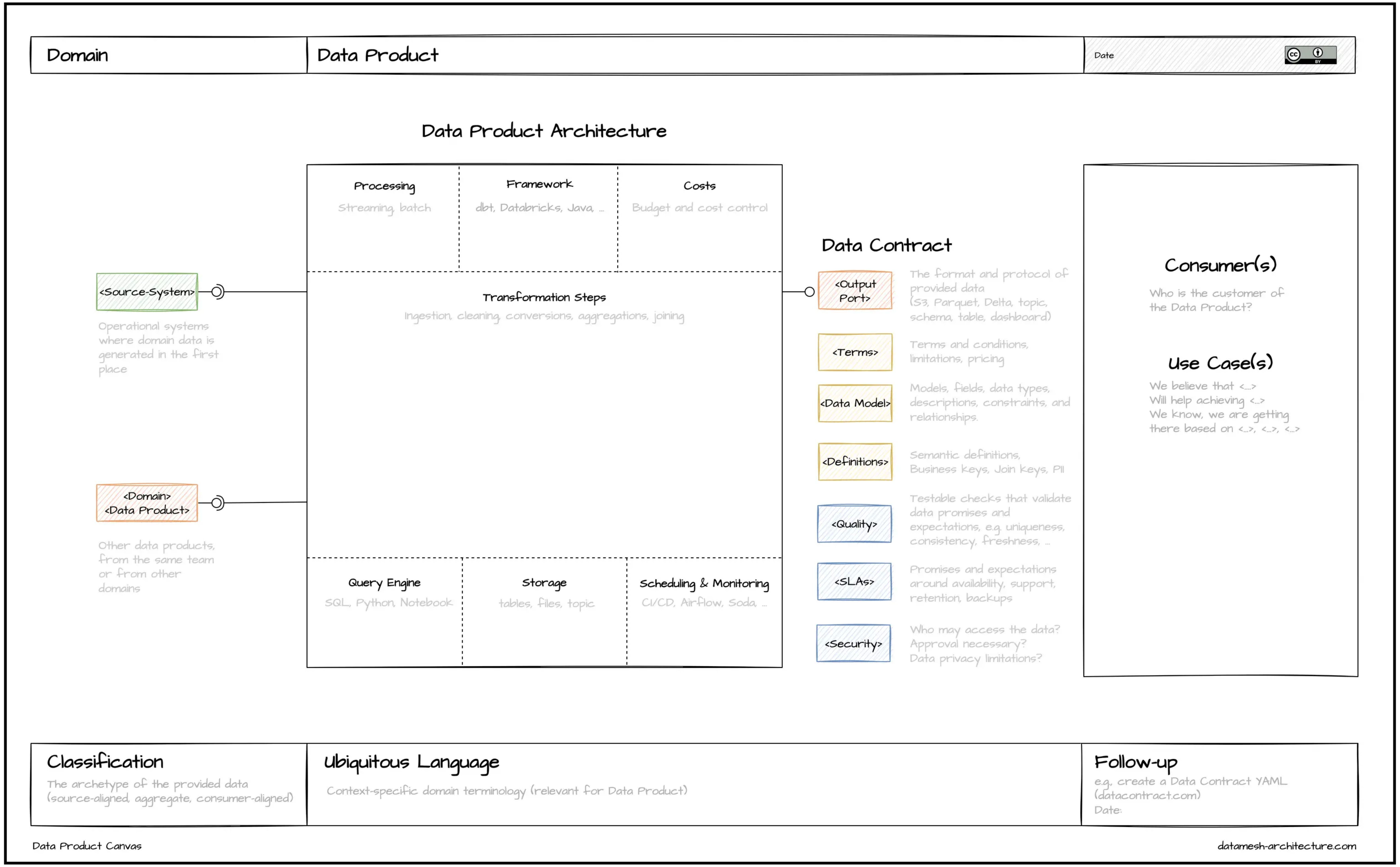


datamesh-architecture.com

That can form a large graph / mesh



The Birth-Certificate



Domain

Data Product

Date

Data Product Architecture

Processing
Streaming, batch

Framework
dbt, Databricks, Java, …

Costs
Budget and cost control

Transformation Steps
Ingestion, cleaning, conversions, aggregations, joining

Query Engine
SQL, Python, Notebook

Storage
tables, files, topic

Scheduling & Monitoring
CI/CD, Airflow, Soda, …

<Source-System>
Operational systems where domain data is generated in the first place

<Domain> <Data Product>
Other data products, from the same team or from other domains

Data Contract

<Output Port> The format and protocol of provided data (S3, Parquet, Delta, topic, schema, table, dashboard)

<Terms> Terms and conditions, limitations, pricing

<Data Model> Models, fields, data types, descriptions, constraints, and relationships

<Definitions> Semantic definitions, Business keys, Join keys, PII

<Quality> Testable checks that validate data promises and expectations, e.g. uniqueness, consistency, freshness, …

<SLAs> Promises and expectations around availability, support, retention, backups

<Security> Who may access the data? Approval necessary? Data privacy limitations?

Consumer(s)
Who is the customer of the Data Product?

Use Case(s)
We believe that <…>
Will help achieving <…>
We know, we are getting there based on <…>, <…>, <…>

Classification
The archetype of the provided data (source-aligned, aggregate, consumer-aligned)

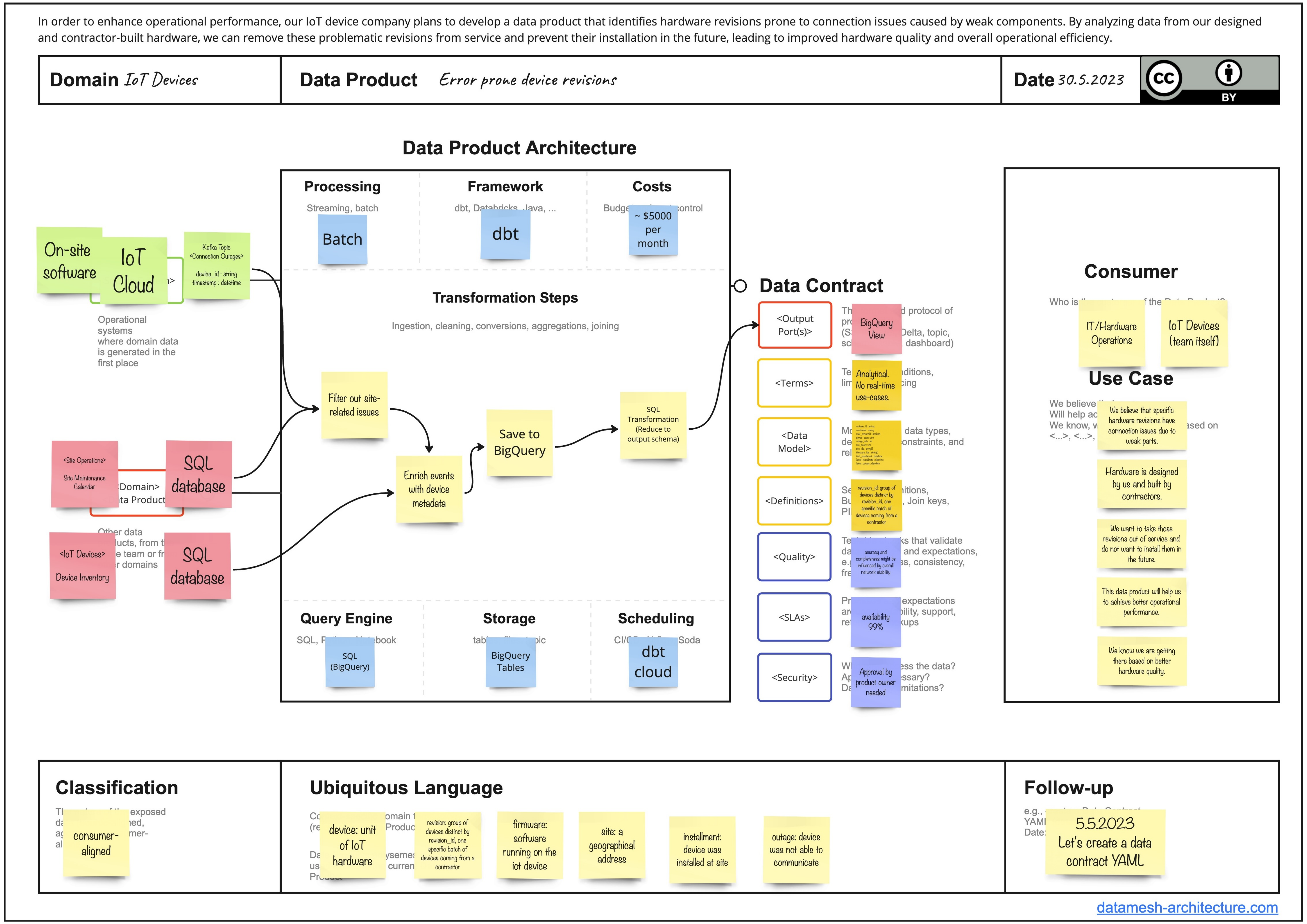
Ubiquitous Language
Context-specific domain terminology (relevant for Data Product)

Follow-up
e.g. create a Data Contract YAML (datacontract.com)
Date:

Data Product Canvas

datamesh-architecture.com

The Birth-Certificate



In order to enhance operational performance, our IoT device company plans to develop a data product that identifies hardware revisions prone to connection issues caused by weak components. By analyzing data from our designed and contractor-built hardware, we can remove these problematic revisions from service and prevent their installation in the future, leading to improved hardware quality and overall operational efficiency.

Domain *IoT Devices* | **Data Product** *Error prone device revisions* | **Date** *30.5.2023*

CC BY

Data Product Architecture

Processing – Streaming, batch – Batch

Framework – dbt, Databricks, Java, ... – dbt

Costs – Budge... control – ~ $5000 per month

Transformation Steps – Ingestion, cleaning, conversions, aggregations, joining

On-site software
IoT Cloud
...n>
Kafka Topic <Connection Outages> device_id : string timestamp : datetime
Operational systems where domain data is generated in the first place
<Site Operations> Site Maintenance Calendar
<Domain> ...ta Product
SQL database
Other data ...ducts, from t... ...e team or fr... ...r domains
<IoT Devices> Device Inventory
SQL database
Filter out site-related issues
Enrich events with device metadata
Save to BigQuery
SQL Transformation (Reduce to output schema)

Query Engine – SQL, P... ...book – SQL (BigQuery)

Storage – tab... ...pic – BigQuery Tables

Scheduling – CI/C... Soda – dbt cloud

Data Contract

<Output Port(s)>	Th... ...d protocol of pr... (S... ...Delta, topic, sc... , dashboard)	BigQuery View
<Terms>	Te... ...nditions, lim... ...cing	Analytical. No real-time use-cases.
<Data Model>	Mo... data types, de... ...onstraints, and rel...	revision_id : string contractor : string over_threshold : boolean device_count : int outage_rate : int site_count : int site_ids : string[] firmware_ids : string[] first_installment : datetime latest_installment : datetime latest_outage : datetime
<Definitions>	Se... ...itions, Bu... , Join keys, PI...	revision_id: group of devices distinct by revision_id, one specific batch of devices coming from a contractor
<Quality>	Te... ...ks that validate da... and expectations, e.g... ...ss, consistency, fre...	acuracy and completeness might be influenced by overall network stability
<SLAs>	Pr... expectations ar... ...bility, support, re... ...kups	availability 99%
<Security>	Wh... ...ess the data? Ap... ...essary? Da... ...mitations?	Approval by product owner needed

Consumer

Who is ... f the ...

- IT/Hardware Operations
- IoT Devices (team itself)

Use Case

We believe ... Will help ac... We know, w... ...ased on <...>, <...>,

- We believe that specific hardware revisions have connection issues due to weak parts.
- Hardware is designed by us and built by contractors.
- We want to take those revisions out of service and do not want to install them in the future.
- This data product will help us to achieve better operational performance.
- We know we are getting there based on better hardware quality.

Classification

Th... exposed da... ...ned, ag... ...mer- al...

consumer-aligned

Ubiquitous Language

Co... ...omain (re... Produc... Da... ...ysemes us... curren... Pro...

- device: unit of IoT hardware
- revision: group of devices distinct by revision_id, one specific batch of devices coming from a contractor
- firmware: software running on the iot device
- site: a geographical address
- installment: device was installed at site
- outage: device was not able to communicate

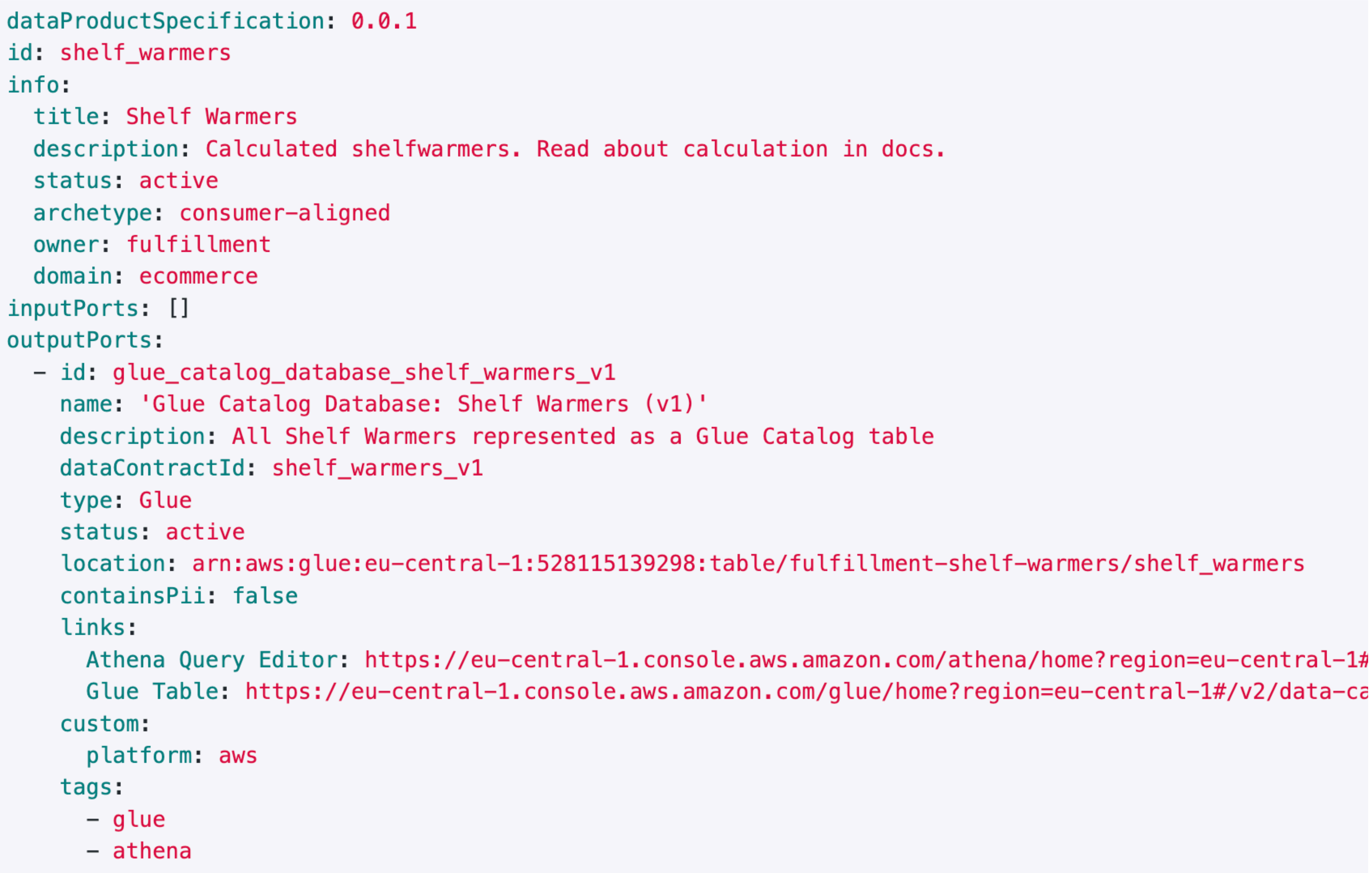
Follow-up

e.g., ... YAML Date:

5.5.2023 Let's create a data contract YAML

datamesh-architecture.com

Specified in YAML



```yaml
dataProductSpecification: 0.0.1
id: shelf_warmers
info:
  title: Shelf Warmers
  description: Calculated shelfwarmers. Read about calculation in docs.
  status: active
  archetype: consumer-aligned
  owner: fulfillment
  domain: ecommerce
inputPorts: []
outputPorts:
  - id: glue_catalog_database_shelf_warmers_v1
    name: 'Glue Catalog Database: Shelf Warmers (v1)'
    description: All Shelf Warmers represented as a Glue Catalog table
    dataContractId: shelf_warmers_v1
    type: Glue
    status: active
    location: arn:aws:glue:eu-central-1:528115139298:table/fulfillment-shelf-warmers/shelf_warmers
    containsPii: false
    links:
      Athena Query Editor: https://eu-central-1.console.aws.amazon.com/athena/home?region=eu-central-1#
      Glue Table: https://eu-central-1.console.aws.amazon.com/glue/home?region=eu-central-1#/v2/data-ca
    custom:
      platform: aws
    tags:
      - glue
      - athena
```

Data Product Specification
(https://dataproduct-specification.com/)

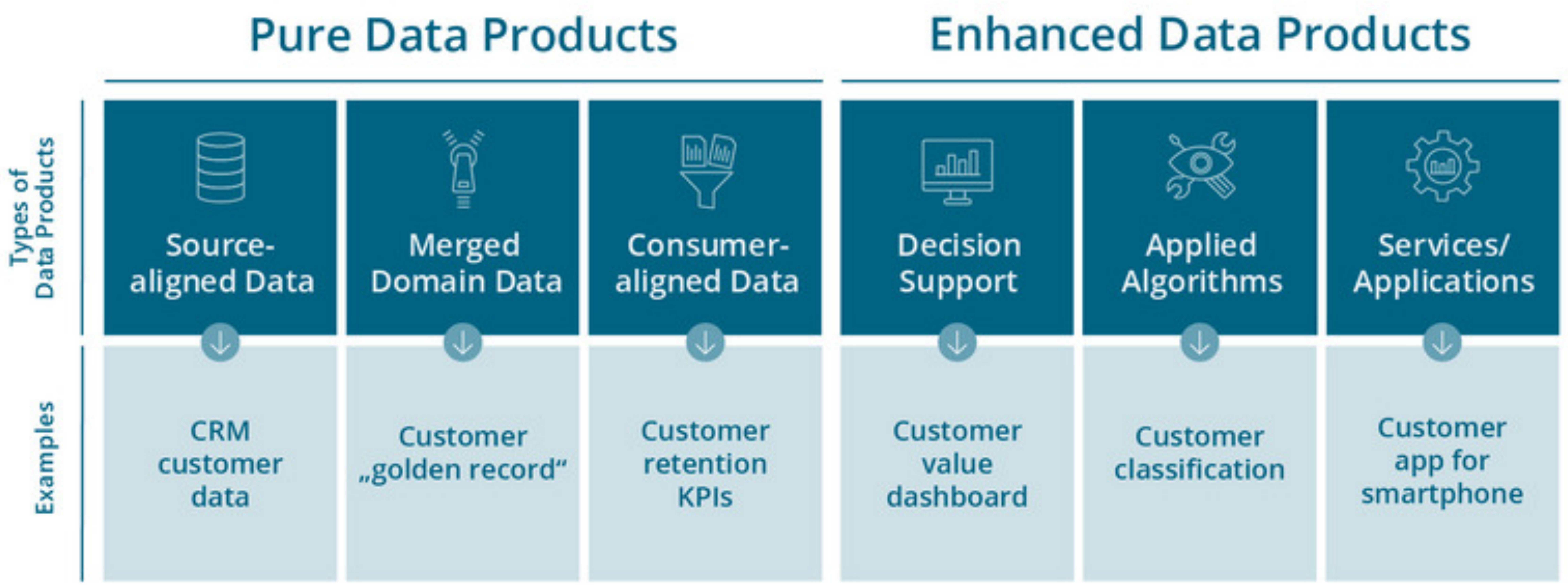
Created by INNOQ

Implementations

- Very different ways
- Depends highly on the data platform the company is using
- Typical: Group all code and YAMLs in a git repo per data product
- Examples:
  - Git repository with dbt that is scheduled in the CI/CD pipeline and runs queries in snowflake
  - Java Application sourcing data from a REST-API and pushing it on an AWS S3 bucket
  - Databricks Asset Bundle with pipelines written in Python
  - ...

All fine? Sadly, no.

Some call them "Pure Data Products"



	Pure Data Products			Enhanced Data Products		
Types of Data Products	Source-aligned Data	Merged Domain Data	Consumer-aligned Data	Decision Support	Applied Algorithms	Services/ Applications
Examples	CRM customer data	Customer „golden record“	Customer retention KPIs	Customer value dashboard	Customer classification	Customer app for smartphone

©BARC | barc.com

If you ask 5 different people,

you get 6 different answers…

Major Differences

deployment unit (like a Docker container)

logical unit (like a git repository)

data set + quality (like a database table with metadata on guarantees)

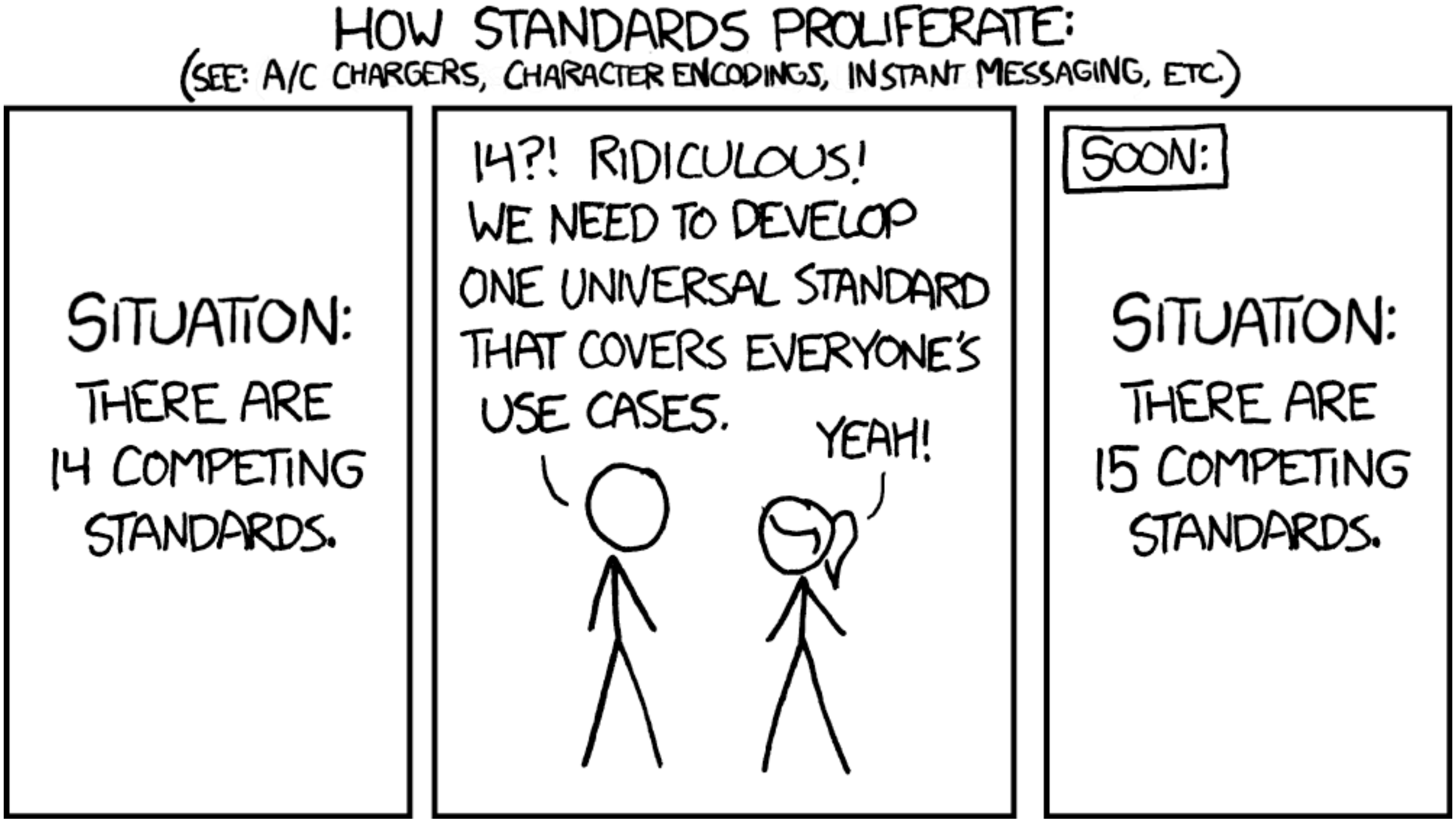
anything that heavily consumes data (like a report or application)

And we haven't talked about any details, like access-request-workflows and breaking-change-processes...

And many formats as well

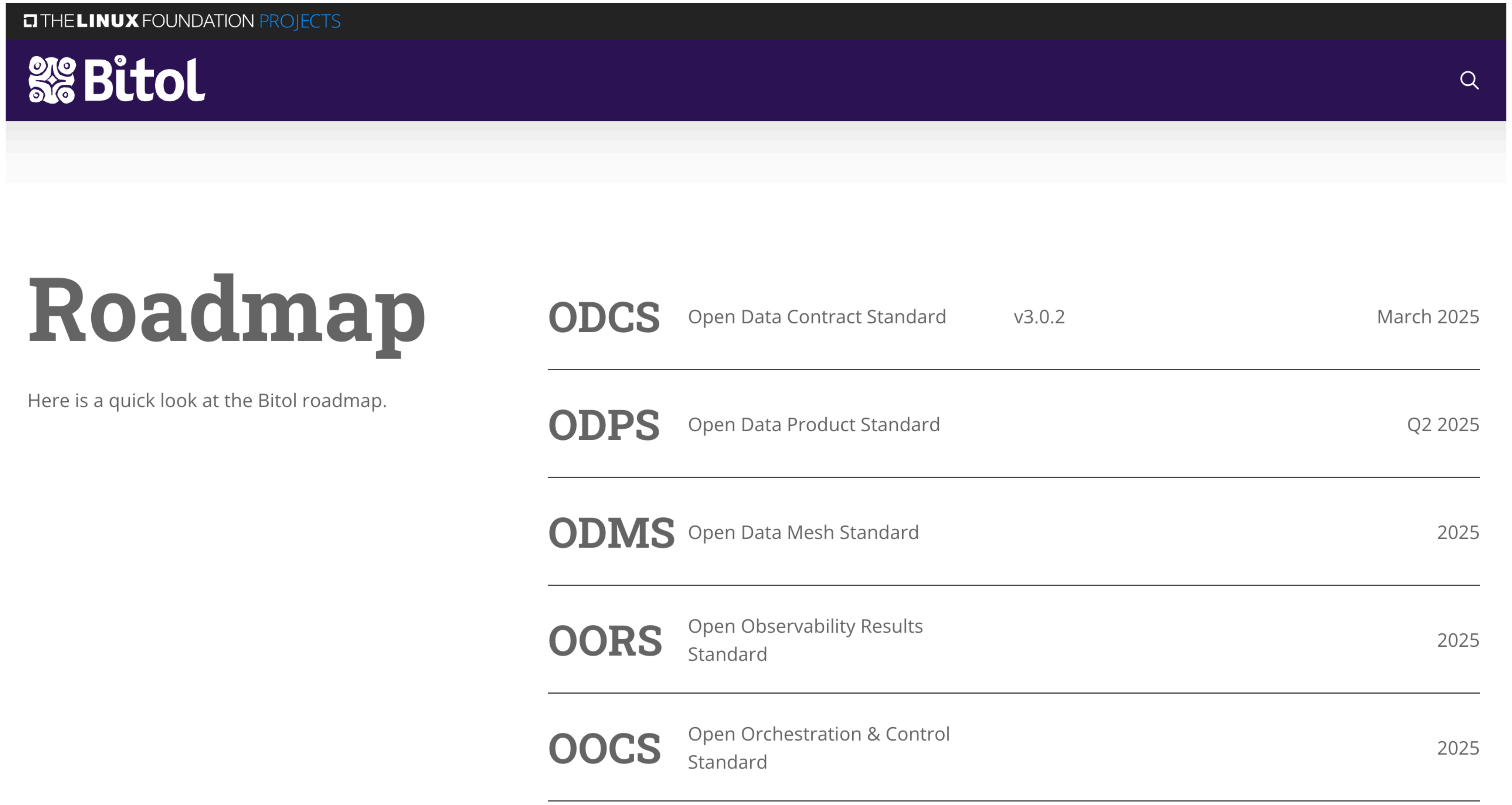
- From vendors and standard bodies that are similar: Data Products Ontology (DPROD from OMG), Data Product Descriptor Specification (DPDS from Quantyca), Data Product Specification (DPS from witboost), Data Product Specification (DPS from INNOQ), Open Data Product Specification (ODPS from LF) …
- A popular one, the Open Data Product Specification, follows a different definition of what a data product is (data set + guarantees) … which makes everything even more confusing. (https://opendataproducts.org/)
- There is no clear leader yet

Our Answer


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

The Open Data Product Standard

- At the Linux Foundation, Data & AI, as part of Bitol
- Bitol: standards around data products, data contracts, data mesh, …
- Status: Work in Progress, release overdue ;-)



Roadmap

Here is a quick look at the Bitol roadmap.

ODCS	Open Data Contract Standard	v3.0.2	March 2025
ODPS	Open Data Product Standard		Q2 2025
ODMS	Open Data Mesh Standard		2025
OORS	Open Observability Results Standard		2025
OOCS	Open Orchestration & Control Standard		2025

In summary: data product is a catchy term

Everybody wants to use it for their own agenda

And now you have to clarify the term before every conversation…

Part 3: Data Contracts

APIs

As every presentation needs an image generated with AI, here you go.

APIs

- REST-API specified with OpenAPI
- Messages and Events specified with AsyncAPI
- What about sharing large datasets? How to specify these APIs?
  - Examples: JSON on AWS S3, SQL tables on BigQuery, Iceberg files on Azure One Lake, SQL views on Snowflake, Delta Live Tables on Databricks, CSV on sftp

Existing Specs are Lacking

- data structure (string with length 5 is implemented with VARCHAR(5))
- data quality on columns (column is nullable, but only 3% null values max)
- terms and conditions (can I use the data for my use case?)
  - data classifications and PII anonymization
- Service-level agreements (freshness, latency, retention, ...)
- Semantics (what the column really is about)

So we need something that fills this gap.

Terminology

A data contract

is not a contract as a mutual agreement,

it is rather an offer to potential consumers.

Cardinality

A data contract is a one-to-many relationship.

One data provider to many data consumers.

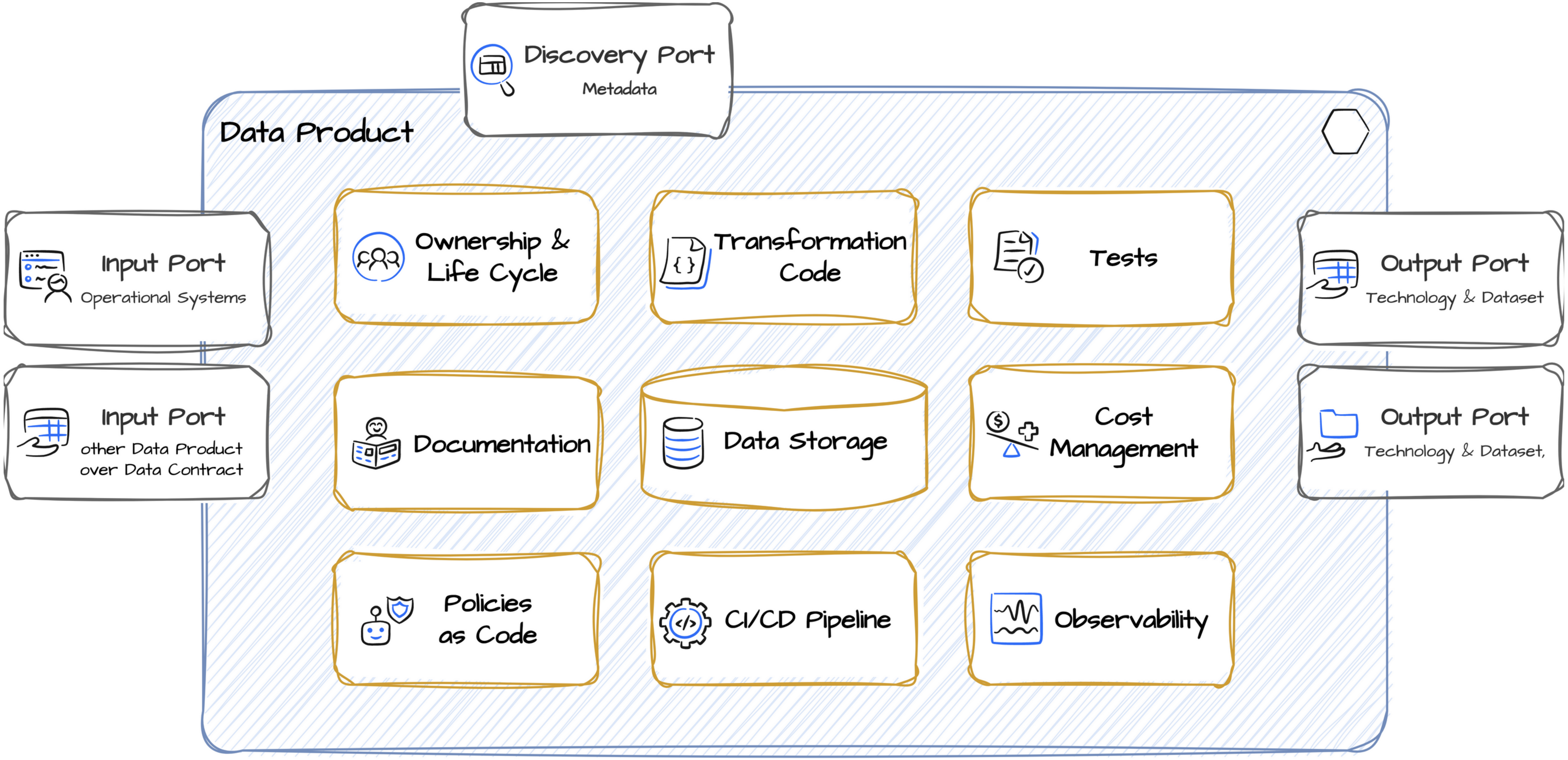
What is a data contract?

- Creates a link between a data producer and data consumers.
- Creates a link between business (logical representation of the data) and technology (its physical implementation).
- Describes meta meta data: rules, quality, and behavior.
- It's like an API Contract, but for data. And contains much more.

- **A data contract is the source of truth for your metadata.**

Slide used by courtesy of Jeon-George Perrin

Data Contracts Protect Output Ports



datamesh-architecture.com

What are the problems it solves?

- Normalizing and keeping documentation relevant.
- Bringing quality data in AI workflows. Describing service-level expectations.
- Easing data & tools integration.
- Ending painful data discovery.
- Enabling data product thinking.

Slide used by courtesy of Jeon-George Perrin

Data contracts [..] are a bit like tax returns. You have to do them. Many people don't feel like doing them and some people even try to avoid them altogether. For us, data contracts form the foundation of a living data ownership culture.

– David Brandstädter, Director Data Enablement
Lidl eCommerce

Slide used by courtesy of Jeon-George Perrin

What happened …

… with the rise of data products, the need for data contracts grew tremendously …

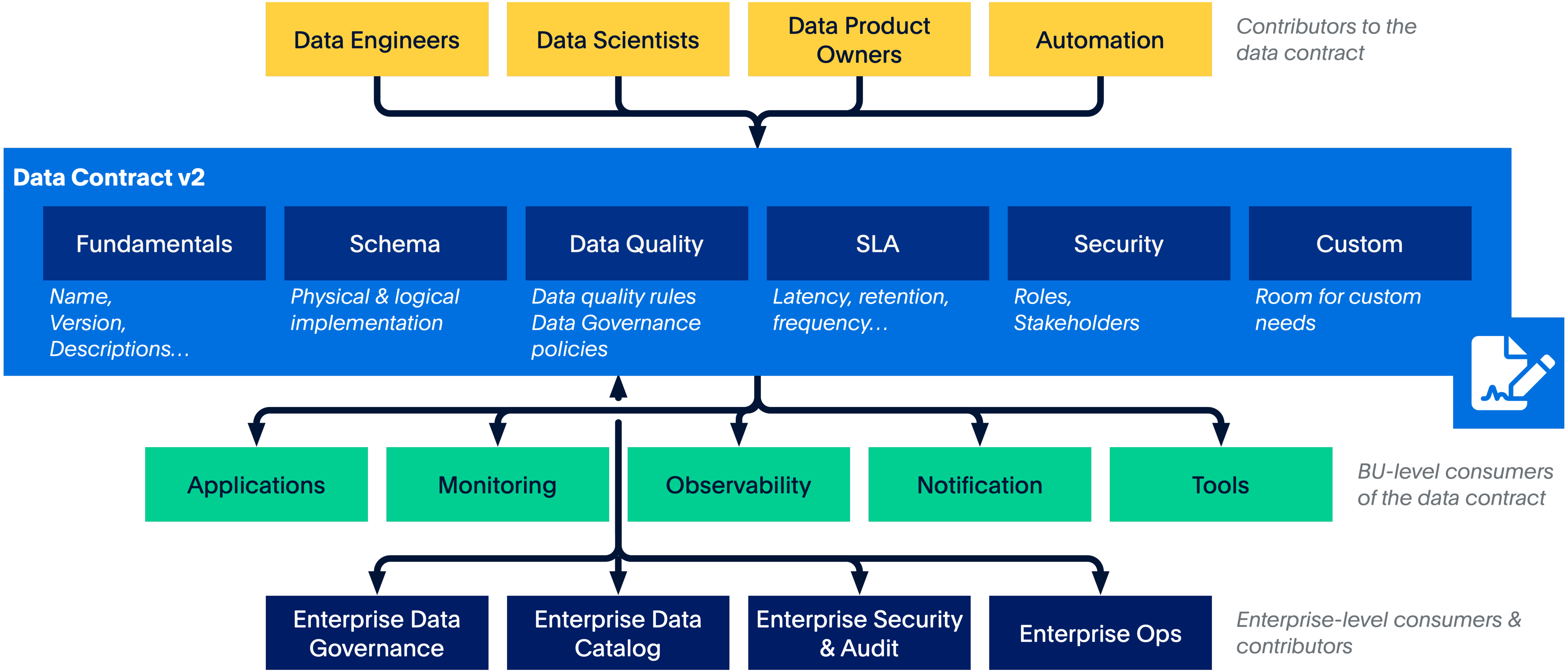
… and many many company created their own data contract format …

… and every vendor did the same …

Open Data Contract Standard

to the rescue (this time just-in-time)

2023



Data Engineers

Data Scientists

Data Product Owners

Automation

Contributors to the data contract

Data Contract v2

Fundamentals

Name, Version, Descriptions…

Schema

Physical & logical implementation

Data Quality

Data quality rules Data Governance policies

SLA

Latency, retention, frequency…

Security

Roles, Stakeholders

Custom

Room for custom needs

Applications

Monitoring

Observability

Notification

Tools

BU-level consumers of the data contract

Enterprise Data Governance

Enterprise Data Catalog

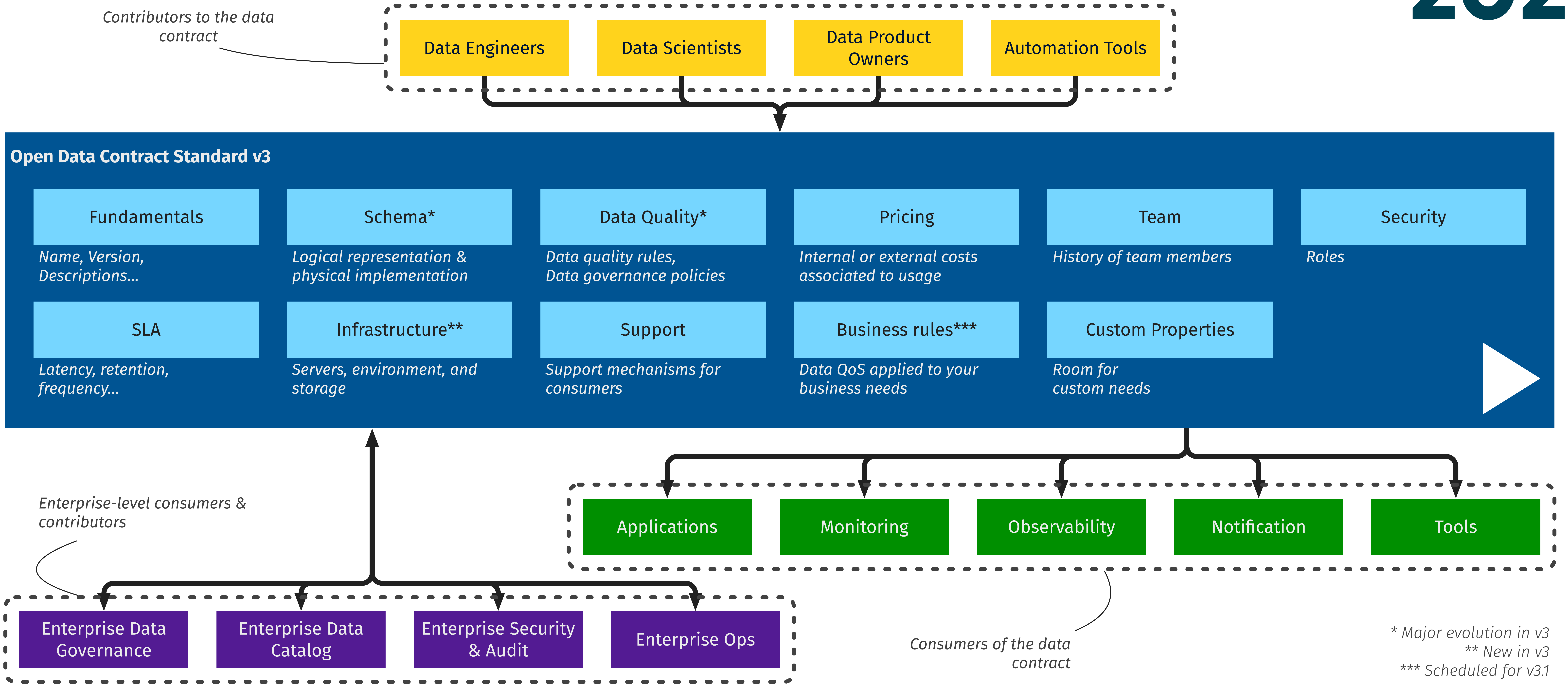
Enterprise Security & Audit

Enterprise Ops

Enterprise-level consumers & contributors

https://github.com/paypal/data-contract-template

2025



Contributors to the data contract

Data Engineers

Data Scientists

Data Product Owners

Automation Tools

Open Data Contract Standard v3

Fundamentals

Name, Version, Descriptions...

Schema*

Logical representation & physical implementation

Data Quality*

Data quality rules, Data governance policies

Pricing

Internal or external costs associated to usage

Team

History of team members

Security

Roles

SLA

Latency, retention, frequency...

Infrastructure**

Servers, environment, and storage

Support

Support mechanisms for consumers

Business rules***

Data QoS applied to your business needs

Custom Properties

Room for custom needs

Enterprise-level consumers & contributors

Enterprise Data Governance

Enterprise Data Catalog

Enterprise Security & Audit

Enterprise Ops

Applications

Monitoring

Observability

Notification

Tools

Consumers of the data contract

** Major evolution in v3*
*** New in v3*
**** Scheduled for v3.1*

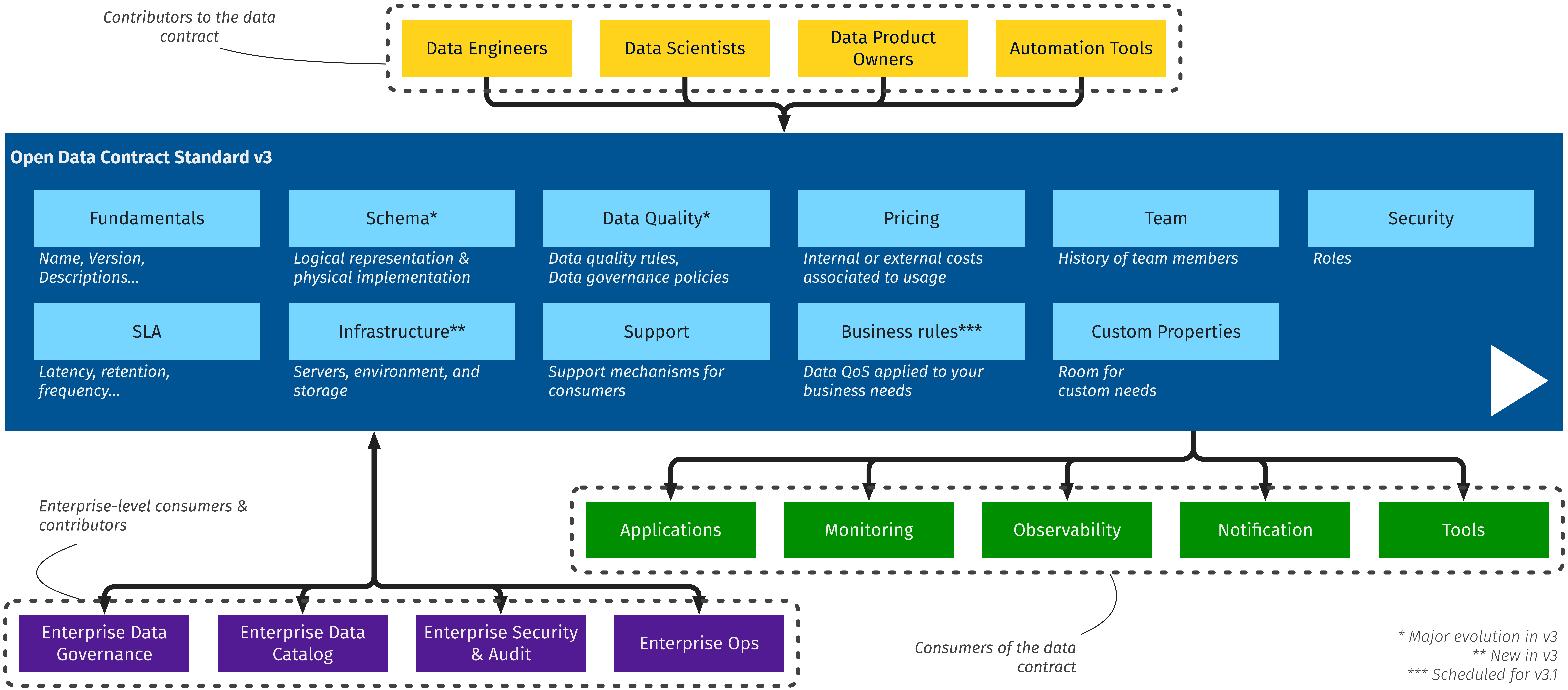
https://github.com/bitol-io/open-data-contract-standard

ODCS (Open Data Contract Standard) can be found & used at 30+ companies:

- **End users:** peer-to-peer payment leader, major cable company, major retailers, SMB to Fortune 500.
- Software **Vendors**: several data-oriented startups and vendors in Europe, NA, and APAC.
- **Service providers** in Europe and NA.
- A **going community** behind the standard.
- **Academia** (JADS, HTW).

Industry-led governance.

Slide used by courtesy of Jeon-George Perrin



Contributors to the data contract

Data Engineers

Data Scientists

Data Product Owners

Automation Tools

Open Data Contract Standard v3

Fundamentals

Name, Version, Descriptions…

Schema*

Logical representation & physical implementation

Data Quality*

Data quality rules, Data governance policies

Pricing

Internal or external costs associated to usage

Team

History of team members

Security

Roles

SLA

Latency, retention, frequency…

Infrastructure**

Servers, environment, and storage

Support

Support mechanisms for consumers

Business rules***

Data QoS applied to your business needs

Custom Properties

Room for custom needs

Applications

Monitoring

Observability

Notification

Tools

Consumers of the data contract

Enterprise-level consumers & contributors

Enterprise Data Governance

Enterprise Data Catalog

Enterprise Security & Audit

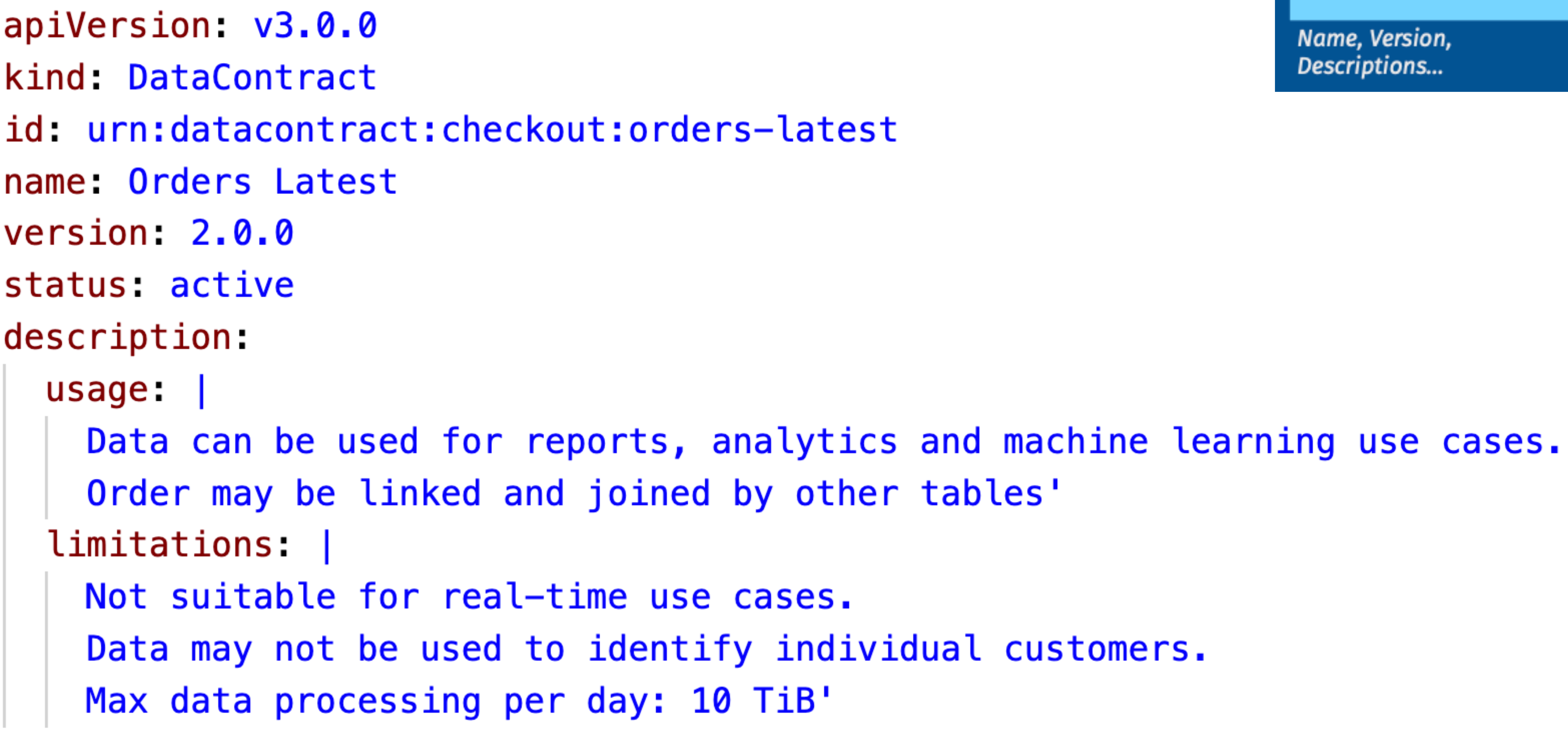
Enterprise Ops

** Major evolution in v3*
*** New in v3*
**** Scheduled for v3.1*

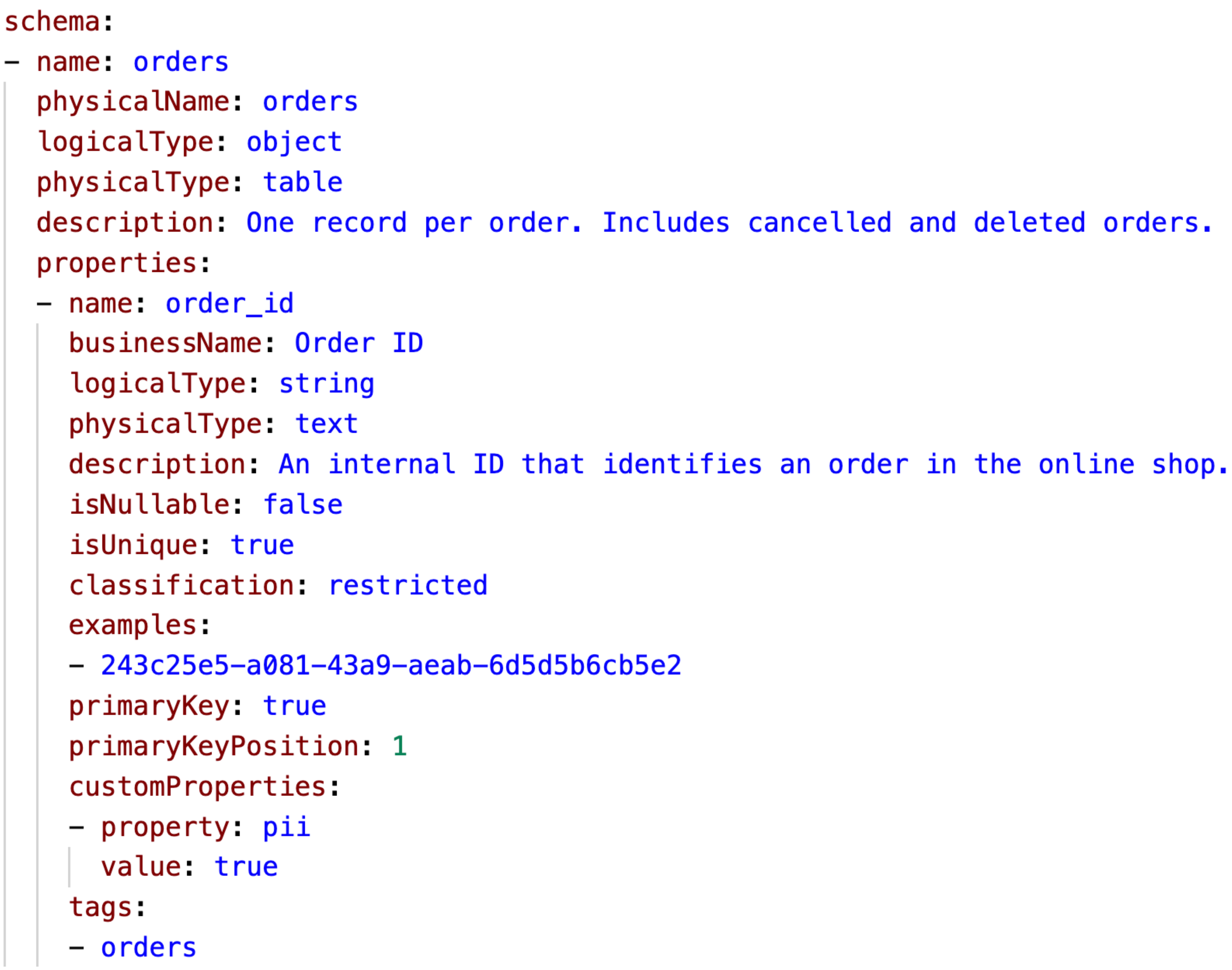
https://github.com/bitol-io/open-data-contract-standard

Fundamentals

Name, Version, Descriptions...



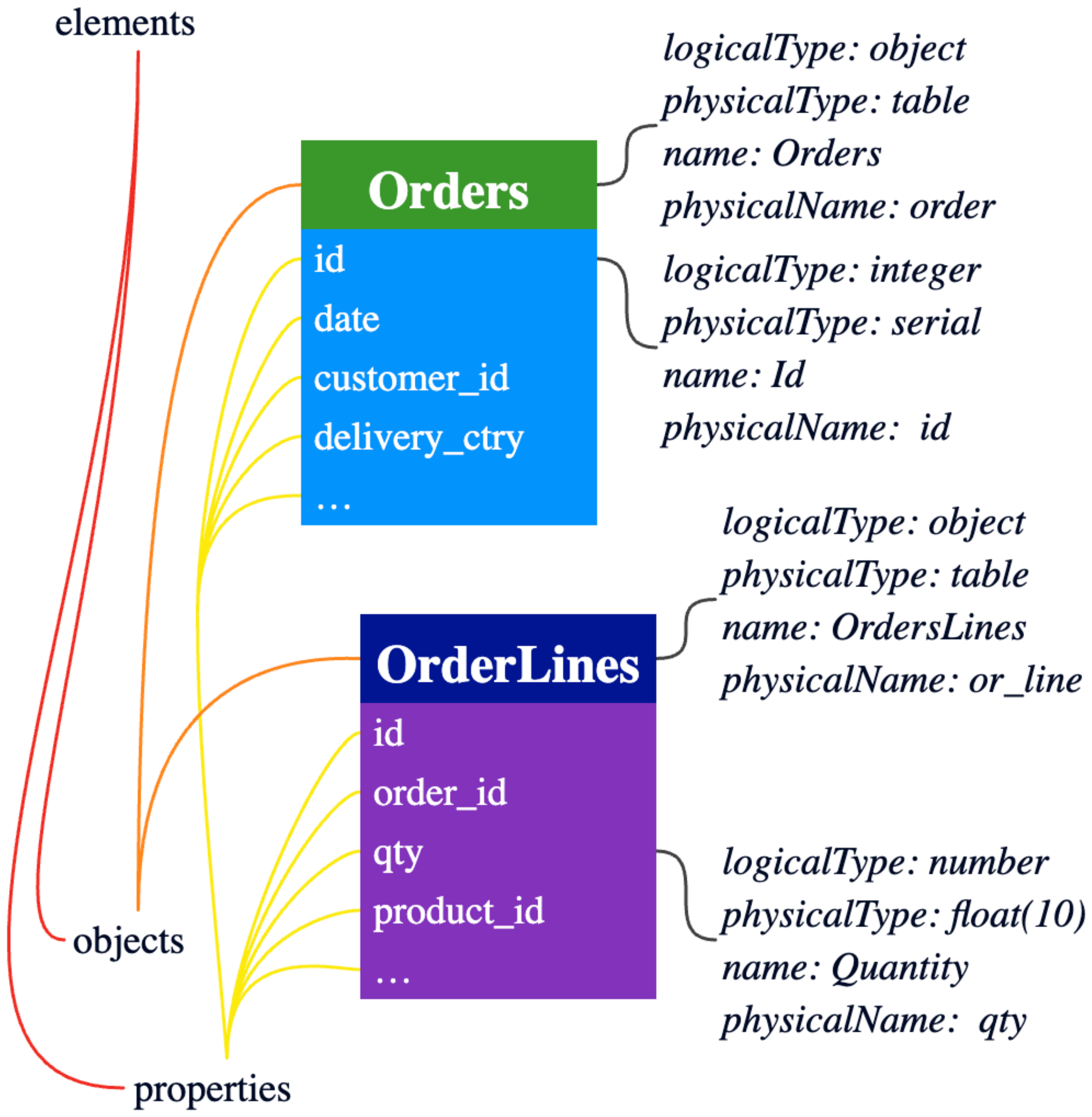
```yaml
apiVersion: v3.0.0
kind: DataContract
id: urn:datacontract:checkout:orders-latest
name: Orders Latest
version: 2.0.0
status: active
description:
  usage: |
    Data can be used for reports, analytics and machine learning use cases.
    Order may be linked and joined by other tables'
  limitations: |
    Not suitable for real-time use cases.
    Data may not be used to identify individual customers.
    Max data processing per day: 10 TiB'
```



```yaml
schema:
- name: orders
  physicalName: orders
  logicalType: object
  physicalType: table
  description: One record per order. Includes cancelled and deleted orders.
  properties:
  - name: order_id
    businessName: Order ID
    logicalType: string
    physicalType: text
    description: An internal ID that identifies an order in the online shop.
    isNullable: false
    isUnique: true
    classification: restricted
    examples:
    - 243c25e5-a081-43a9-aeab-6d5d5b6cb5e2
    primaryKey: true
    primaryKeyPosition: 1
    customProperties:
    - property: pii
      value: true
    tags:
    - orders
```

Schema*

Logical representation & physical implementation



elements

Orders

id

date

customer_id

delivery_ctry

…

logicalType: object
physicalType: table
name: Orders
physicalName: order

logicalType: integer
physicalType: serial
name: Id
physicalName: id

OrderLines

id

order_id

qty

product_id

…

logicalType: object
physicalType: table
name: OrdersLines
physicalName: or_line

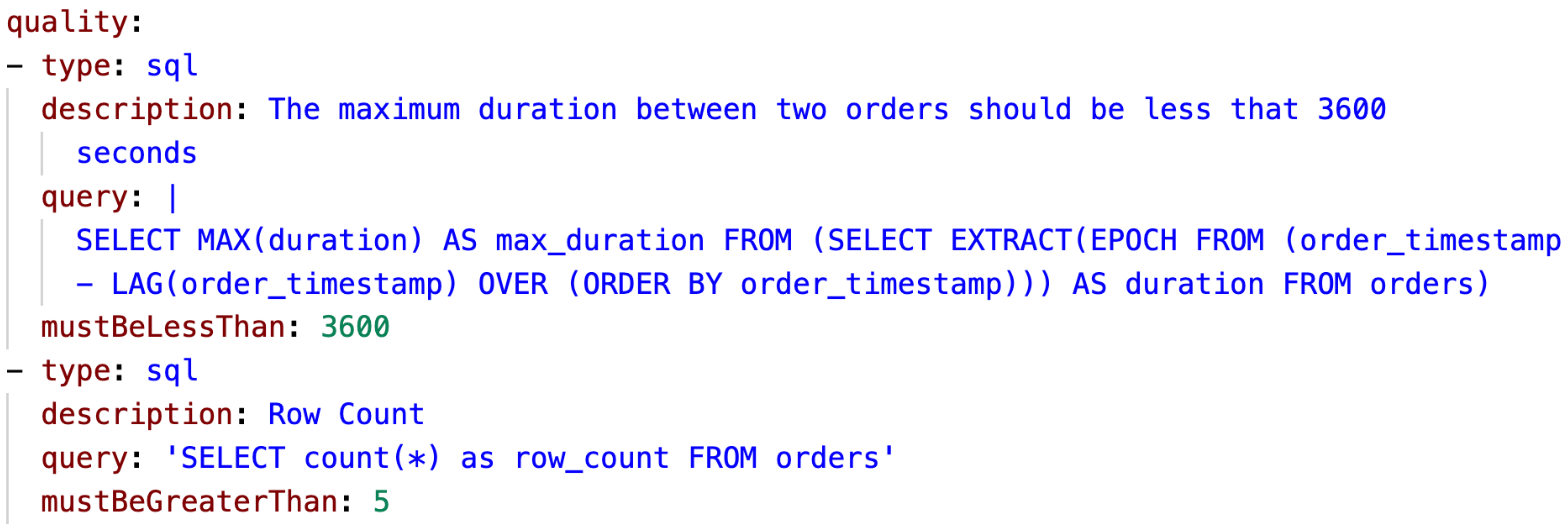
logicalType: number
physicalType: float(10)
name: Quantity
physicalName: qty

objects

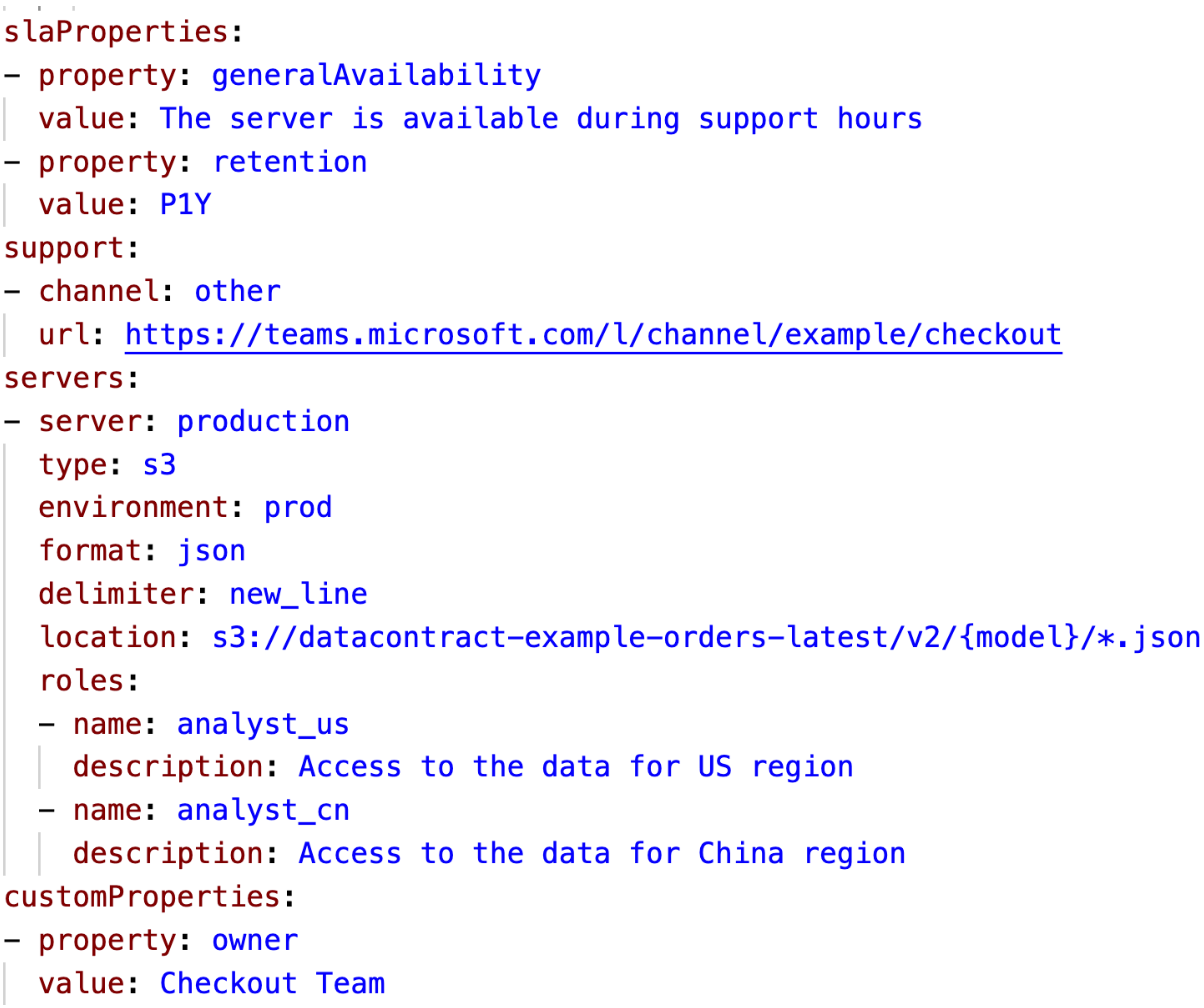
properties

Data Quality*

Data quality rules,
Data governance policies



```yaml
quality:
- type: sql
  description: The maximum duration between two orders should be less that 3600
    seconds
  query: |
    SELECT MAX(duration) AS max_duration FROM (SELECT EXTRACT(EPOCH FROM (order_timestamp
    - LAG(order_timestamp) OVER (ORDER BY order_timestamp))) AS duration FROM orders)
  mustBeLessThan: 3600
- type: sql
  description: Row Count
  query: 'SELECT count(*) as row_count FROM orders'
  mustBeGreaterThan: 5
```



```yaml
slaProperties:
- property: generalAvailability
  value: The server is available during support hours
- property: retention
  value: P1Y
support:
- channel: other
  url: https://teams.microsoft.com/l/channel/example/checkout
servers:
- server: production
  type: s3
  environment: prod
  format: json
  delimiter: new_line
  location: s3://datacontract-example-orders-latest/v2/{model}/*.json
  roles:
  - name: analyst_us
    description: Access to the data for US region
  - name: analyst_cn
    description: Access to the data for China region
customProperties:
- property: owner
  value: Checkout Team
```

SLA

Latency, retention, frequency...

Infrastructure**

Servers, environment, and storage

Custom Properties

Room for custom needs

Code Generation

- Java
- Python in Pydantic
- dbt Models and Sources
- SQL DDL and Queries

Test

- Compare contract with real data
- Breaking data detection in PR
- Breaking metadata detection
- Continuous Monitoring
- Consumer-driven Contract Testing

Metadata Distribution

- Metastores: Hive, ...
- Data Catalogs: Colibra, ...
- Data Contract Catalog: Data Mesh Manager, ...
- Software Catalogs: LeanIX, ...

Automate all the things!

Infrastructure Provisioning

- Output Port (S3 Bucket, ...)
- Input Port (dbt sources.yml)
- Transformations (anonymisation)
- Access Control (IAM permissions)

Collaboration

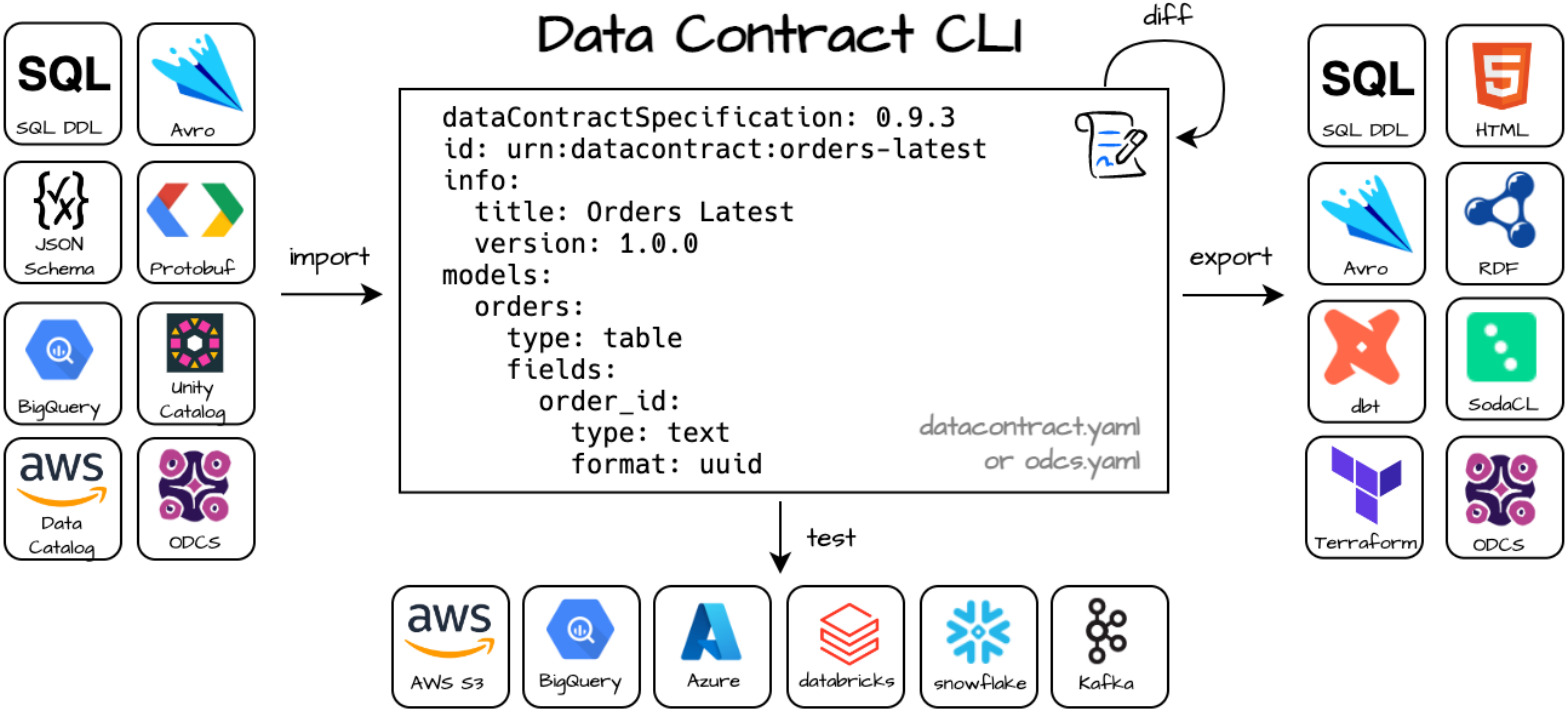
- Contract-First (in workshop)
- Data-First (import from ...)
- Semantics

Governance

- Policies (naming conventions, ...)
- Schema Evolution (Notice period)
- Usage Agreements
- Approval Workflows

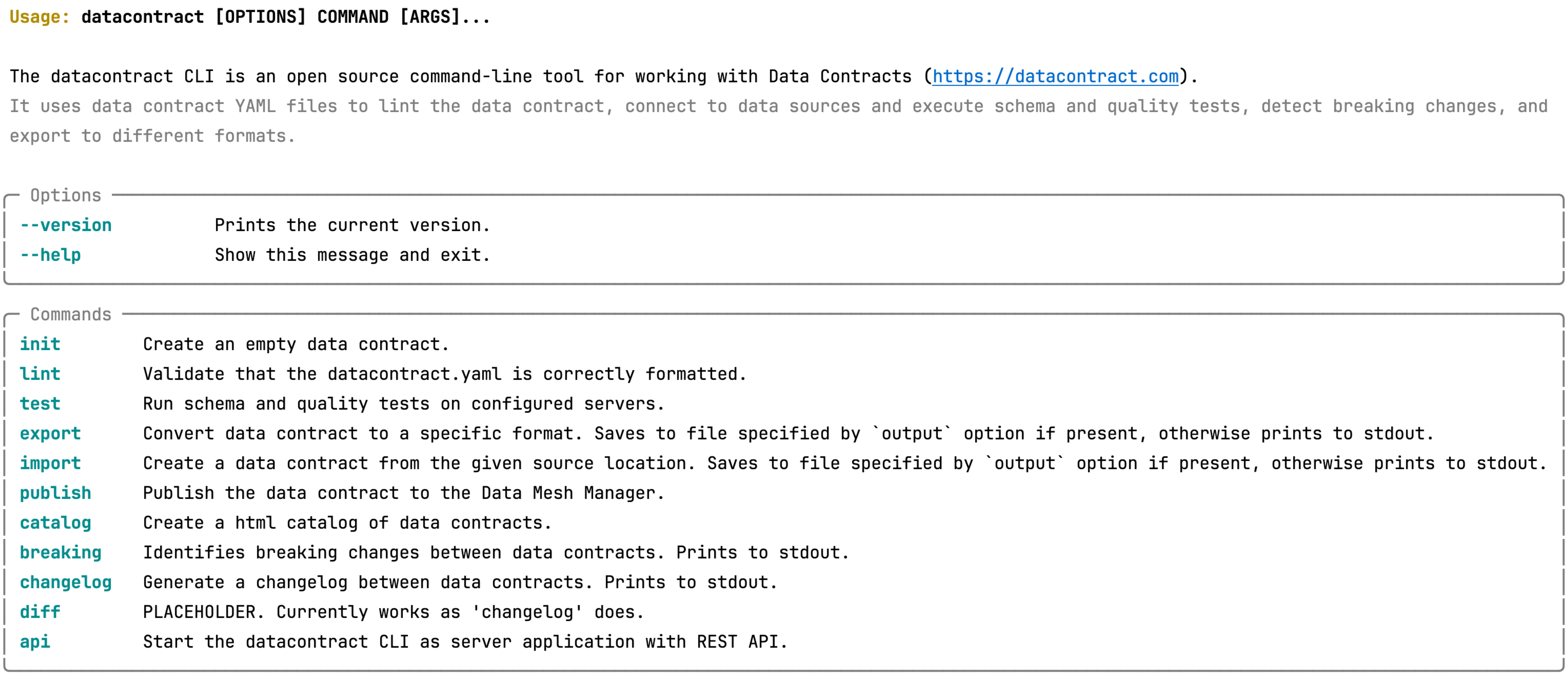
Data Contract CLI

Data Contract CLI



```yaml
dataContractSpecification: 0.9.3
id: urn:datacontract:orders-latest
info:
  title: Orders Latest
  version: 1.0.0
models:
  orders:
    type: table
    fields:
      order_id:
        type: text
        format: uuid
```

github.com/datacontract/cli



```
Usage: datacontract [OPTIONS] COMMAND [ARGS]...

The datacontract CLI is an open source command-line tool for working with Data Contracts (https://datacontract.com).
It uses data contract YAML files to lint the data contract, connect to data sources and execute schema and quality tests, detect breaking changes, and
export to different formats.

╭─ Options ────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────╮
│ --version            Prints the current version.                                                                                      │
│ --help               Show this message and exit.                                                                                      │
╰───────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────╯
╭─ Commands ────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────╮
│ init         Create an empty data contract.                                                                                           │
│ lint         Validate that the datacontract.yaml is correctly formatted.                                                              │
│ test         Run schema and quality tests on configured servers.                                                                      │
│ export       Convert data contract to a specific format. Saves to file specified by `output` option if present, otherwise prints to stdout. │
│ import       Create a data contract from the given source location. Saves to file specified by `output` option if present, otherwise prints to stdout. │
│ publish      Publish the data contract to the Data Mesh Manager.                                                                      │
│ catalog      Create a html catalog of data contracts.                                                                                 │
│ breaking     Identifies breaking changes between data contracts. Prints to stdout.                                                    │
│ changelog    Generate a changelog between data contracts. Prints to stdout.                                                           │
│ diff         PLACEHOLDER. Currently works as 'changelog' does.                                                                        │
│ api          Start the datacontract CLI as server application with REST API.                                                          │
╰───────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────╯
```

DEMO

Modern Data Governance

Responsibility

Data Product Owner

Data is owned decentralized by business & IT experts where data is generated
Product owners are responsible for what happens with their data

Concepts & Tools

Data Contracts

Define the syntax, semantics, quality, and terms of use as YAML

Data Marketplace

Data discovery with a self-service access request workflow

Global Policies

The conventions and rules of play for data on the data platform

Automation

Contract Enforcement

Test that data products correctly implement the data contract

Automated Permission Granting

Give table access based on access request approvals

AI-based Policy Checking

Check that policies are correctly adopted by data product owners

In summary: data contracts bring API thinking and specification into the data world.

Swagger was first released in 20211 and is everywhere today.
ODCS was first released in 2025 and will be everywhere in ???

My prediction: This will be huge!

Bonus-Part: Data Governance with AI (with Arif)

Modern Data Governance

Responsibility

Data Product Owner

Data is owned decentralized by business & IT experts where data is generated
Product owners are responsible for what happens with their data

Concepts & Tools

Data Contracts

Define the syntax, semantics, quality, and terms of use as YAML

Data Marketplace

Data discovery with a self-service access request workflow

Global Policies

The conventions and rules of play for data on the data platform

Automation

Contract Enforcement

Test that data products correctly implement the data contract

Automated Permission Granting

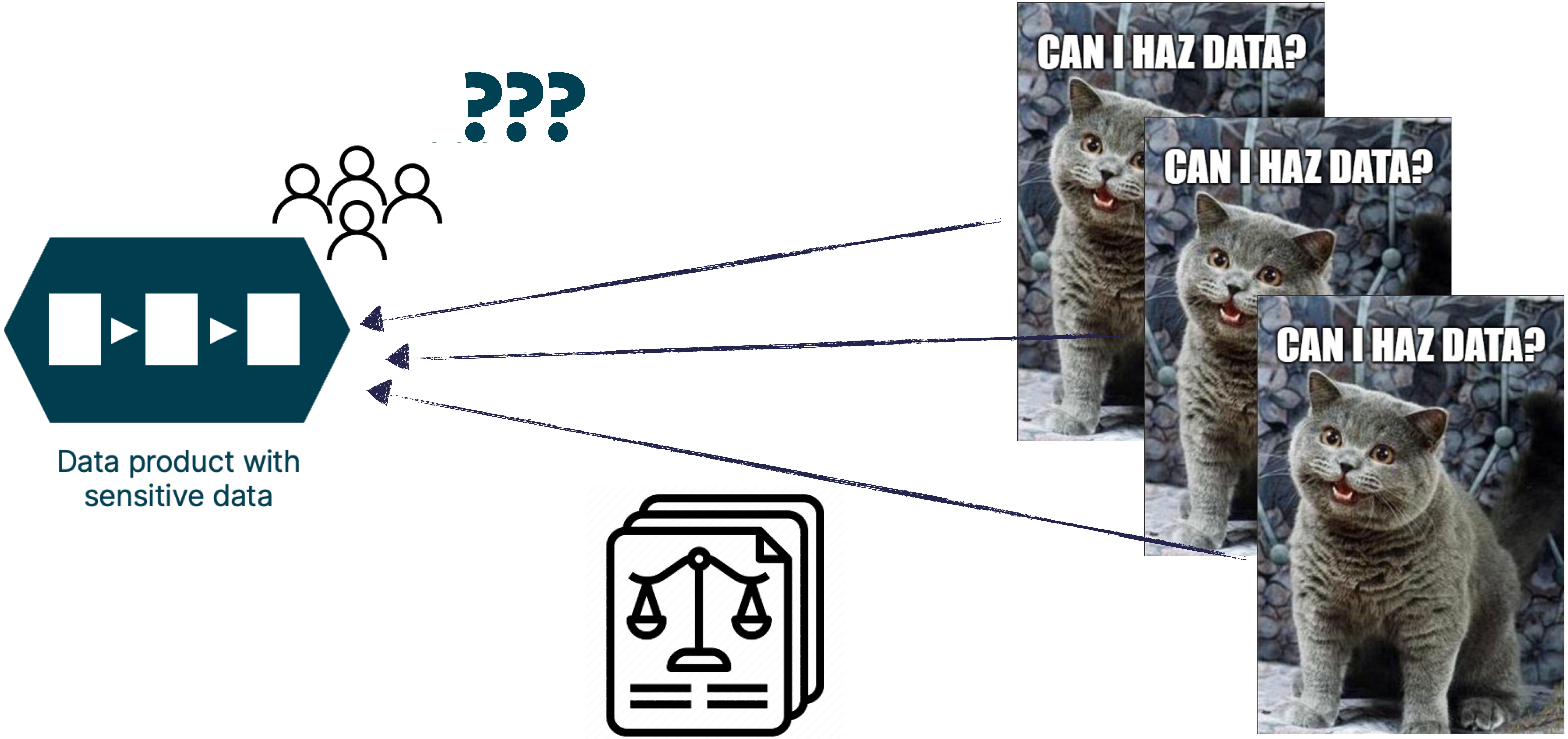
Give table access based on access request approvals

AI-based Policy Checking

Check that policies are correctly adopted by data product owners

datamesh-manager.com

Use Case: Data Access Requests

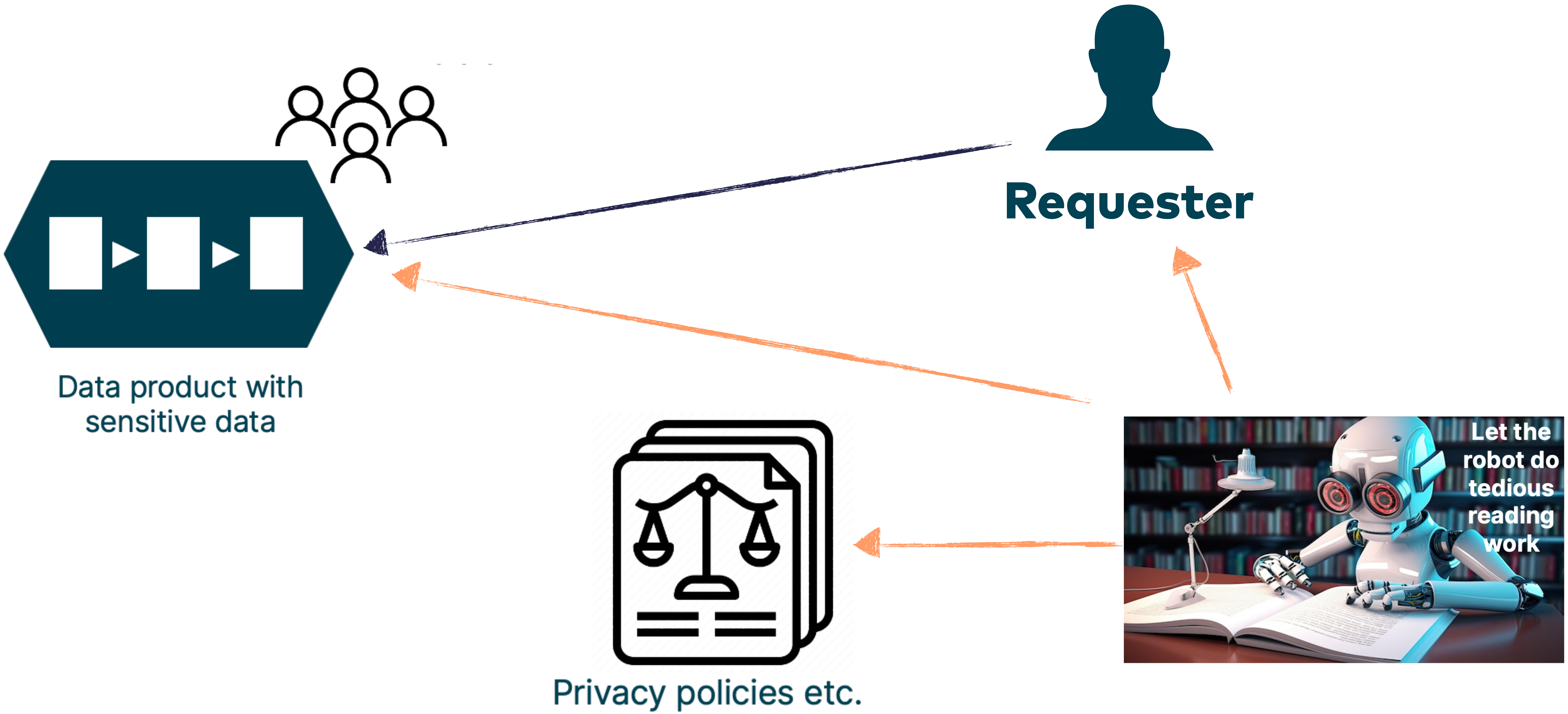


???

Data product with sensitive data

Privacy policies etc.

LLMs to the rescue!

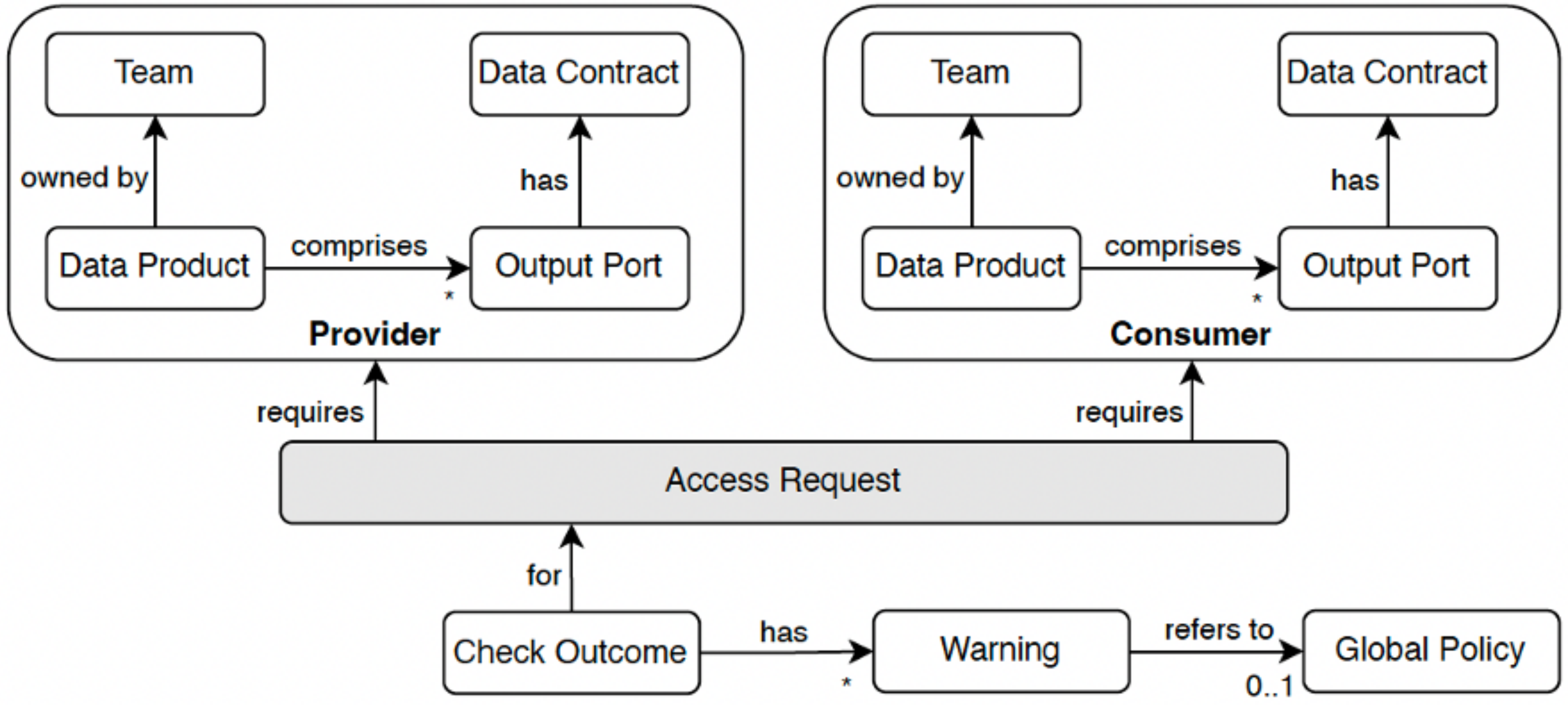


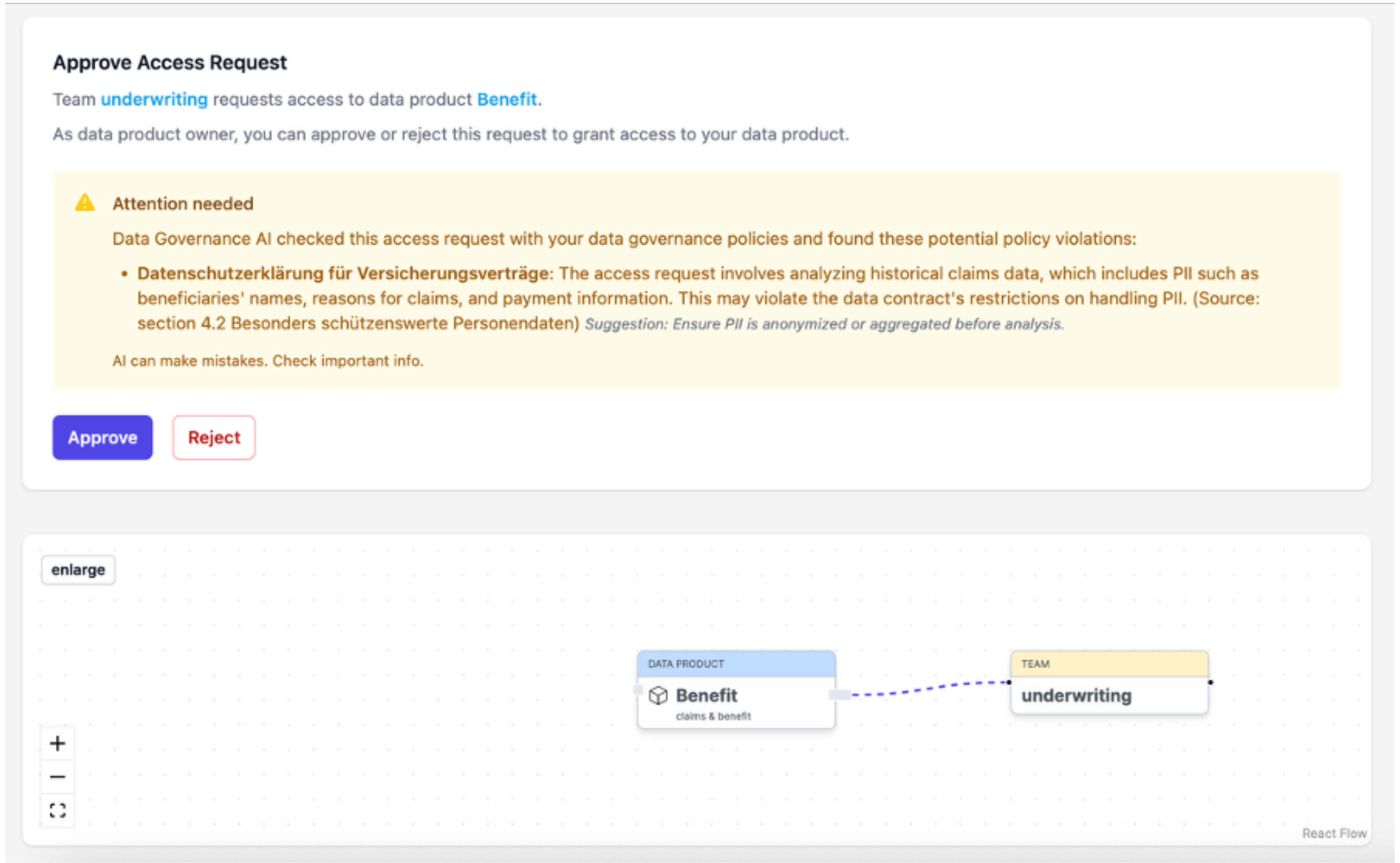
Requester

Data product with sensitive data

Privacy policies etc.

Let the robot do tedious reading work





Approve Access Request

Team underwriting requests access to data product Benefit.

As data product owner, you can approve or reject this request to grant access to your data product.

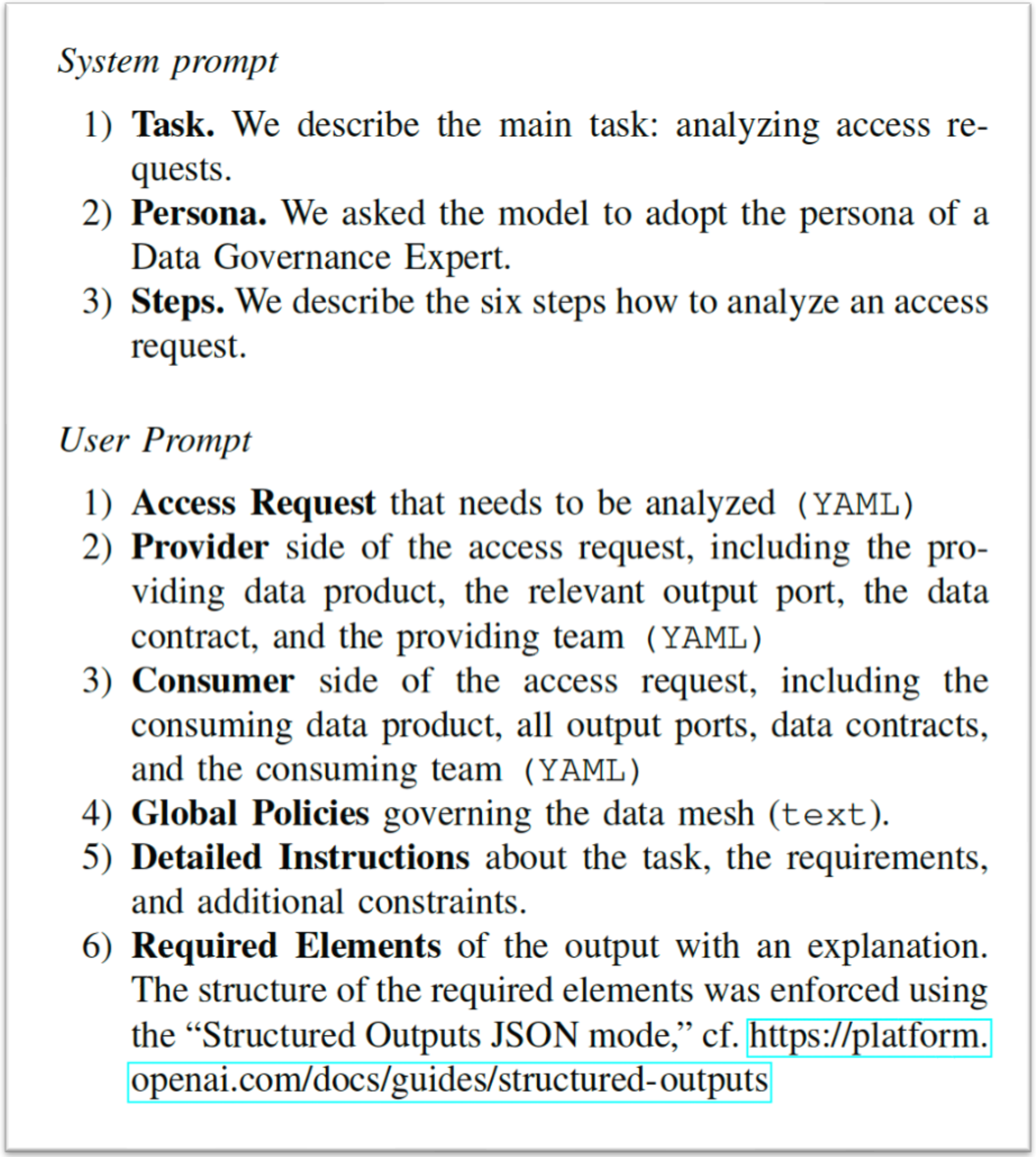
⚠ **Attention needed**

Data Governance AI checked this access request with your data governance policies and found these potential policy violations:

- **Datenschutzerklärung für Versicherungsverträge:** The access request involves analyzing historical claims data, which includes PII such as beneficiaries' names, reasons for claims, and payment information. This may violate the data contract's restrictions on handling PII. (Source: section 4.2 Besonders schützenswerte Personendaten) *Suggestion: Ensure PII is anonymized or aggregated before analysis.*

AI can make mistakes. Check important info.

Approve
Reject
enlarge
DATA PRODUCT
Benefit
claims & benefit
TEAM
underwriting
+
−
React Flow



System prompt

1) **Task.** We describe the main task: analyzing access requests.
2) **Persona.** We asked the model to adopt the persona of a Data Governance Expert.
3) **Steps.** We describe the six steps how to analyze an access request.

User Prompt

1) **Access Request** that needs to be analyzed `(YAML)`
2) **Provider** side of the access request, including the providing data product, the relevant output port, the data contract, and the providing team `(YAML)`
3) **Consumer** side of the access request, including the consuming data product, all output ports, data contracts, and the consuming team `(YAML)`
4) **Global Policies** governing the data mesh (`text`).
5) **Detailed Instructions** about the task, the requirements, and additional constraints.
6) **Required Elements** of the output with an explanation. The structure of the required elements was enforced using the "Structured Outputs JSON mode," cf. https://platform.openai.com/docs/guides/structured-outputs

In summary: Data Goverance perfectly fits to the strengths of LLMs and can help in federation data management.

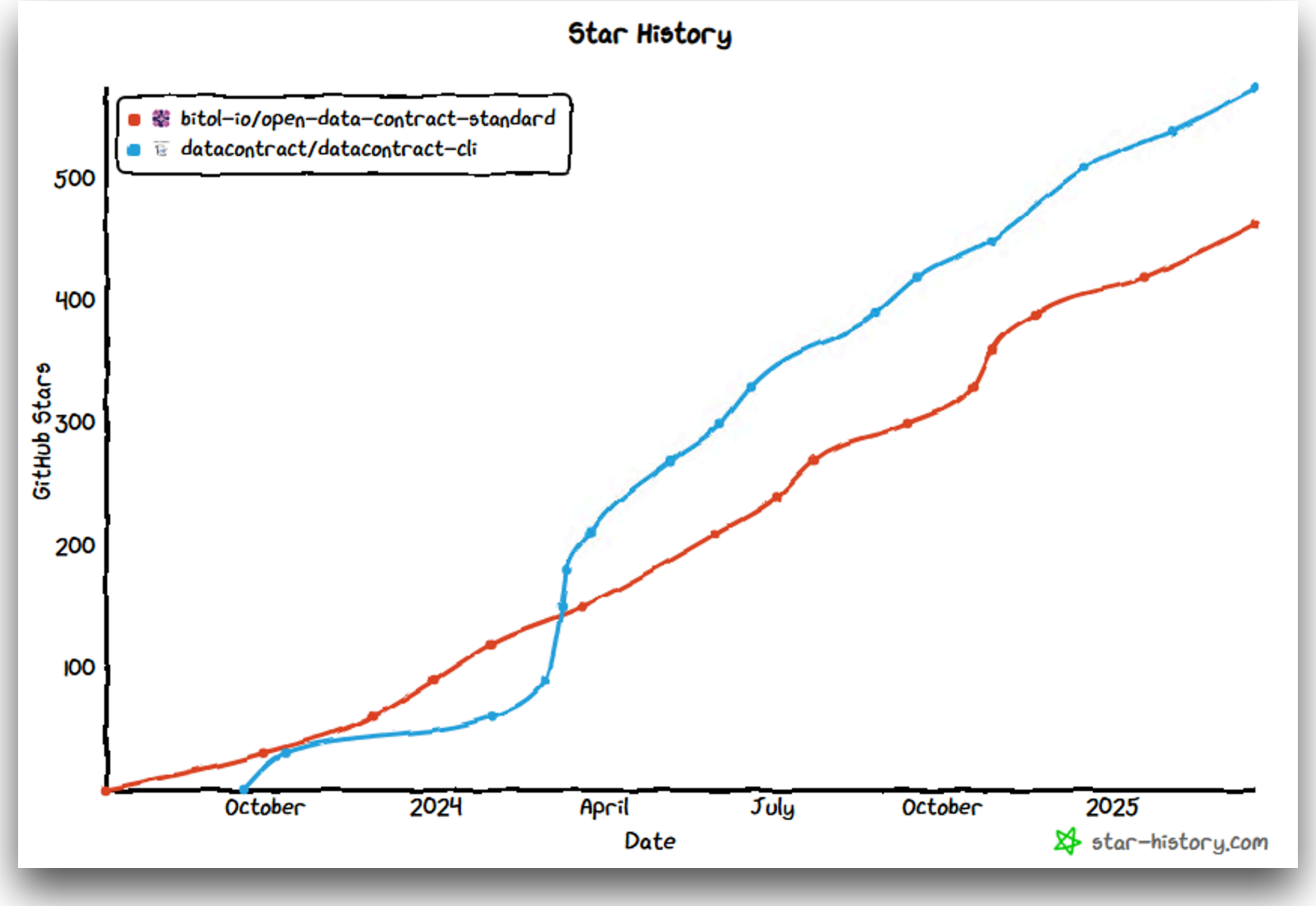
Thank You! Want to learn more?

- Talk to me at this event, or later via LinkedIn   /in/simonharrer

- Try out cli.datacontract.com with an ODCS data contract

- And, of course, give us a star on Github at

  github.com/datacontract/datacontract-cli
  github.com/bitol-io/open-data-contract-standard


Star History
🟥 bitol-io/open-data-contract-standard
🟦 datacontract/datacontract-cli
GitHub Stars
October
April
July
October
Date
star-history.com