



Data Contracts

JOCHEN CHRIST, ENTROPY DATA

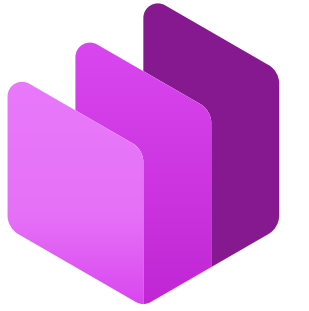
INNOQ



JOCHEN CHRIST

/in/jochenchrist

Hello, I am Jochen



Co-founder and CTO, Entropy Data
External INNOQ Consultant

Author datamesh-architecture.com

Author datacontract.com

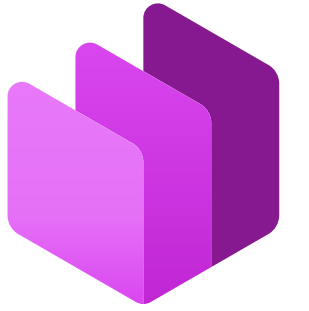
Maintainer Data Contract CLI

TSC Member, Bitol, Open Data Contract Standard



COMMERCIAL PRODUCT

Entropy Data



A solution for managing and sharing data products with data contracts, so everyone in an organization can use data to generate business value.

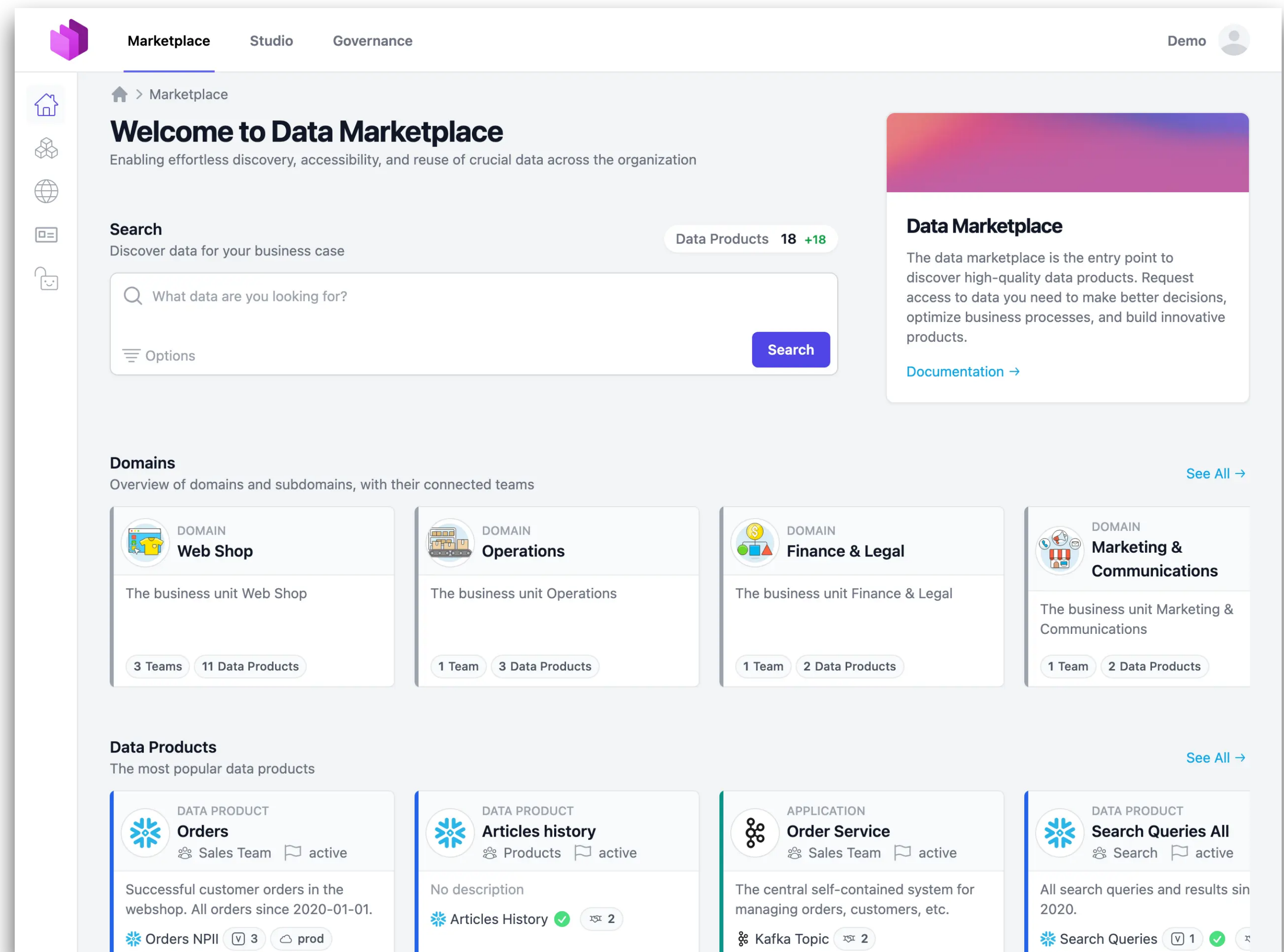
Data Product Marketplace. Self-service marketplace for humans and AI clients to discover, request, and consume data products, with context provided by data contracts.

Industry Standards. Native support for Open Data Product Standard (ODPS) and Open Data Contract Standard (ODCS) with user-friendly editors.

Support for Data Contract CLI. Automated testing of data contracts to enforce promised data quality and SLAs.

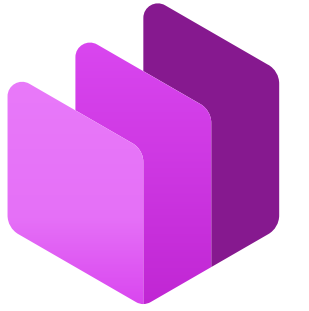
Integrations. Integration with all major data platforms (AWS, GCP, Microsoft, Databricks, Snowflake) and data governance tools.

UX/UI. It looks great.



OPEN SOURCE

Data Contract CLI



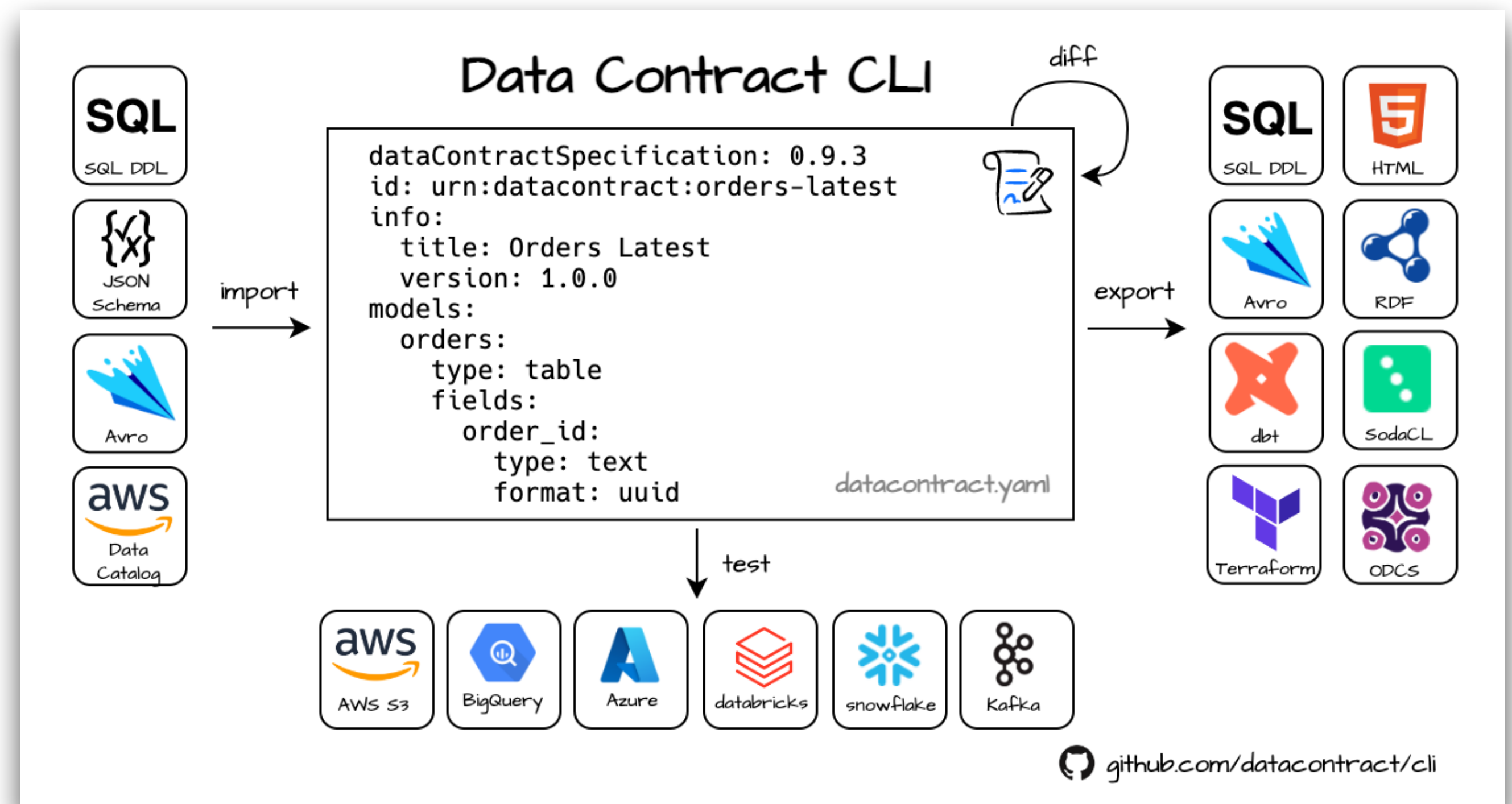
The popular Data Contract CLI is the most widely used developer tool to manage data contracts.

As it is maintained by us, we ensure tight integration with Entropy Data.

✔ **Enforce Data Contracts.** Validate data products against their contracts automatically during development and deployment. Run tests on your own servers to ensure data products conform to their specifications before they reach production, catching structural issues, quality problems, and compliance violations early.

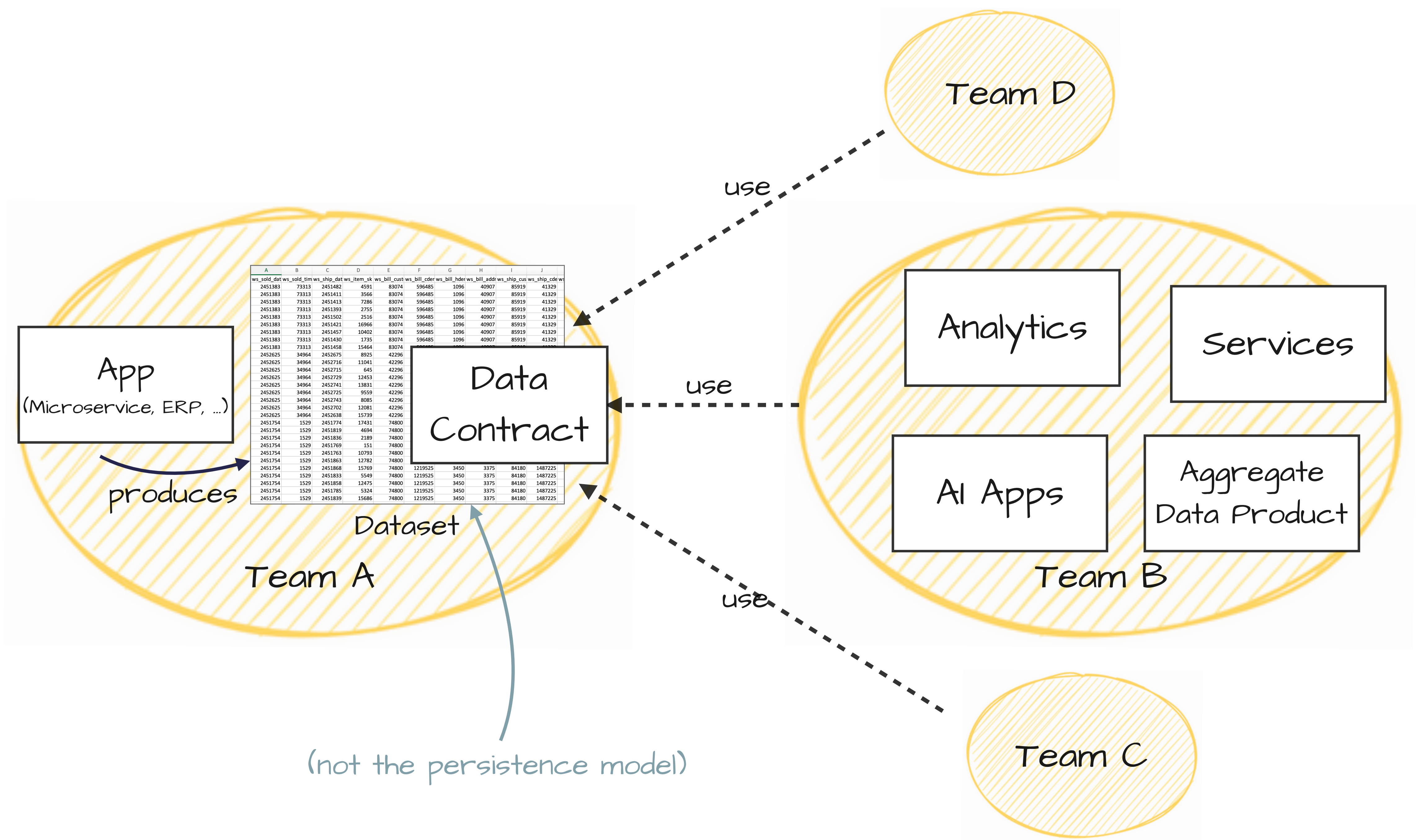
📄 **Generate Code.** Transform data contracts into code artifacts to bootstrap data product implementations. Generate SQL DDL, dbt files, JSON Schema, and more. Data Contract YAMLs are the metadata source of truth for provisioning other systems, such as data catalogs, semantic web technologies, and EAM tools.

🐙 **Open Source.** Python library that can be security assessed and it makes it is easy to extend.



Data Contract


```
apiVersion: v3.1.0
kind: DataContract
id: fulfillment_shipments_v2
name: Shipments v2
version: 2.0.0
status: draft
description:
  purpose: This data can be used for analytical purposes
  usage: Use this to analyze shipments
  limitations: Not suitable for real-time use cases
  authoritativeDefinitions:
    - type: Data Guidelines
      url: https://example.com/data-guidelines.html
schema:
- name: shipments2
  physicalType: table
  physicalName: shipments_v2
  description: "This table contains shipment data, including details about shipment IDs, associated orders, delive
  dataGranularityDescription: Not Aggregated
  properties:
    - name: shipment_id
      businessName: Shipment ID
      logicalType: string
      physicalType: uuid
      description: Unique identifier for each shipment.
      primaryKey: true
      classification: internal
      examples:
        - 123e4567-e89b-12d3-a456-426614174000
      authoritativeDefinitions:
        - type: definition
          url: 
```

Data Contracts = API, but for Data

APIs



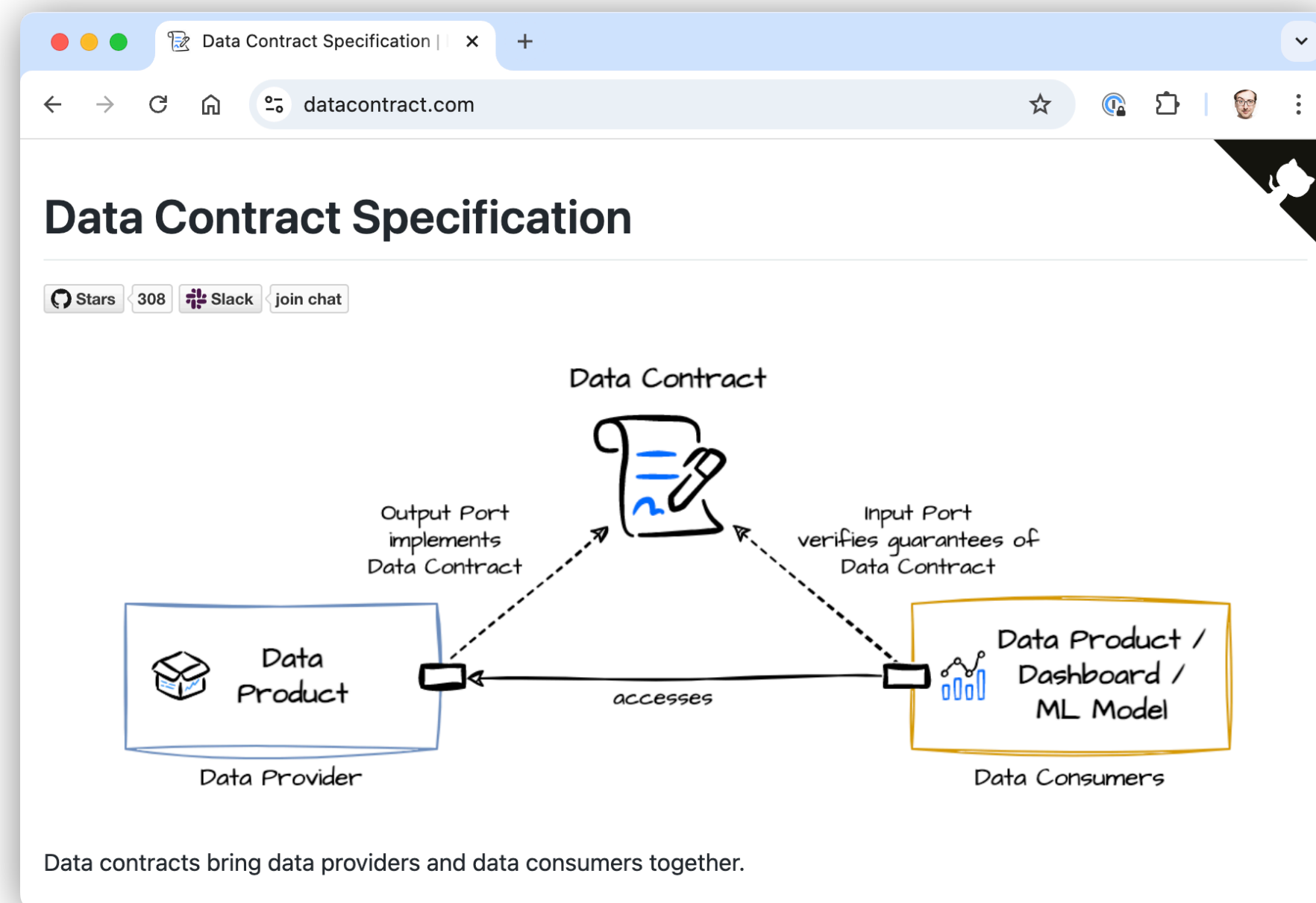
Messages



Data

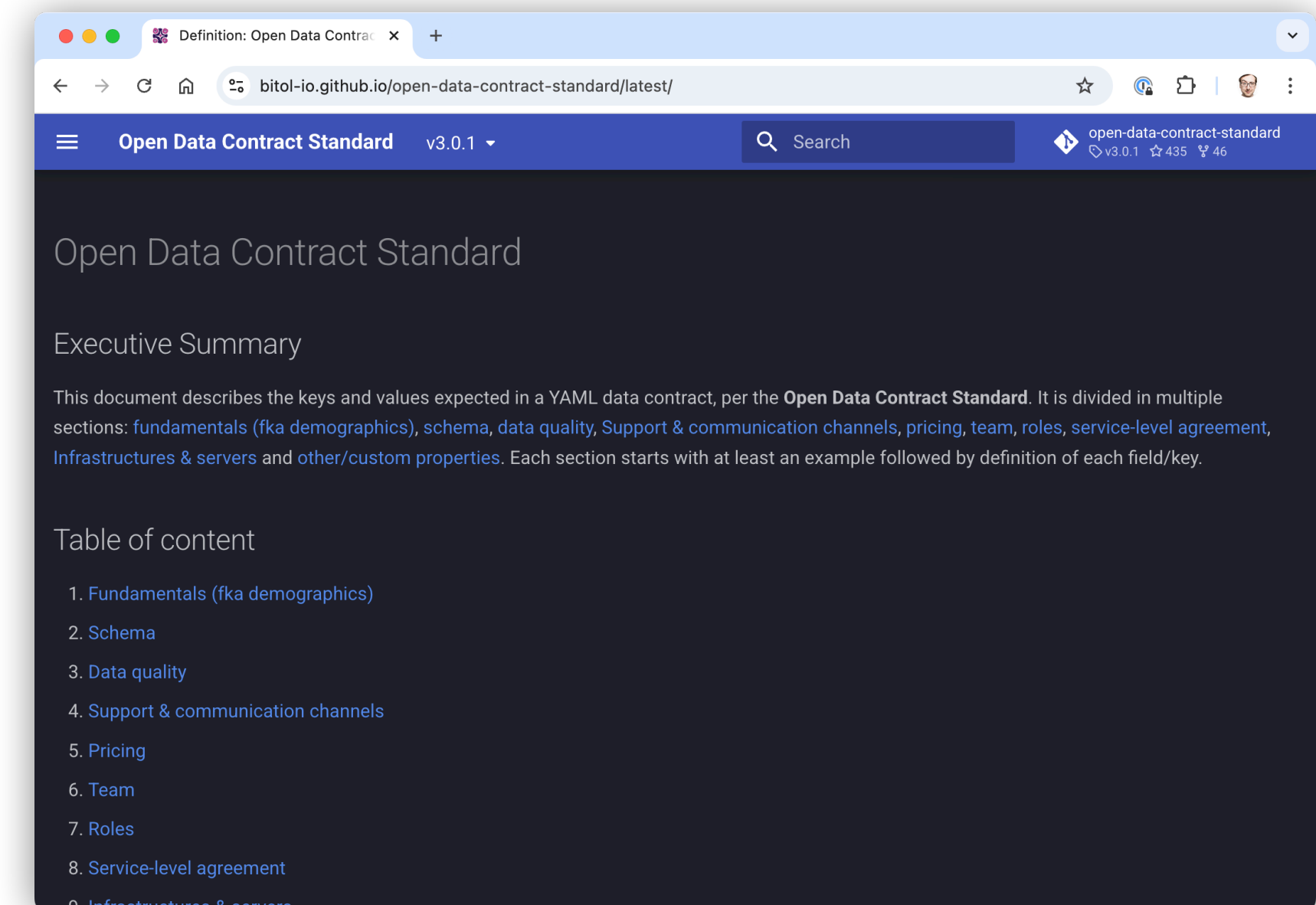


Data Contract Standards



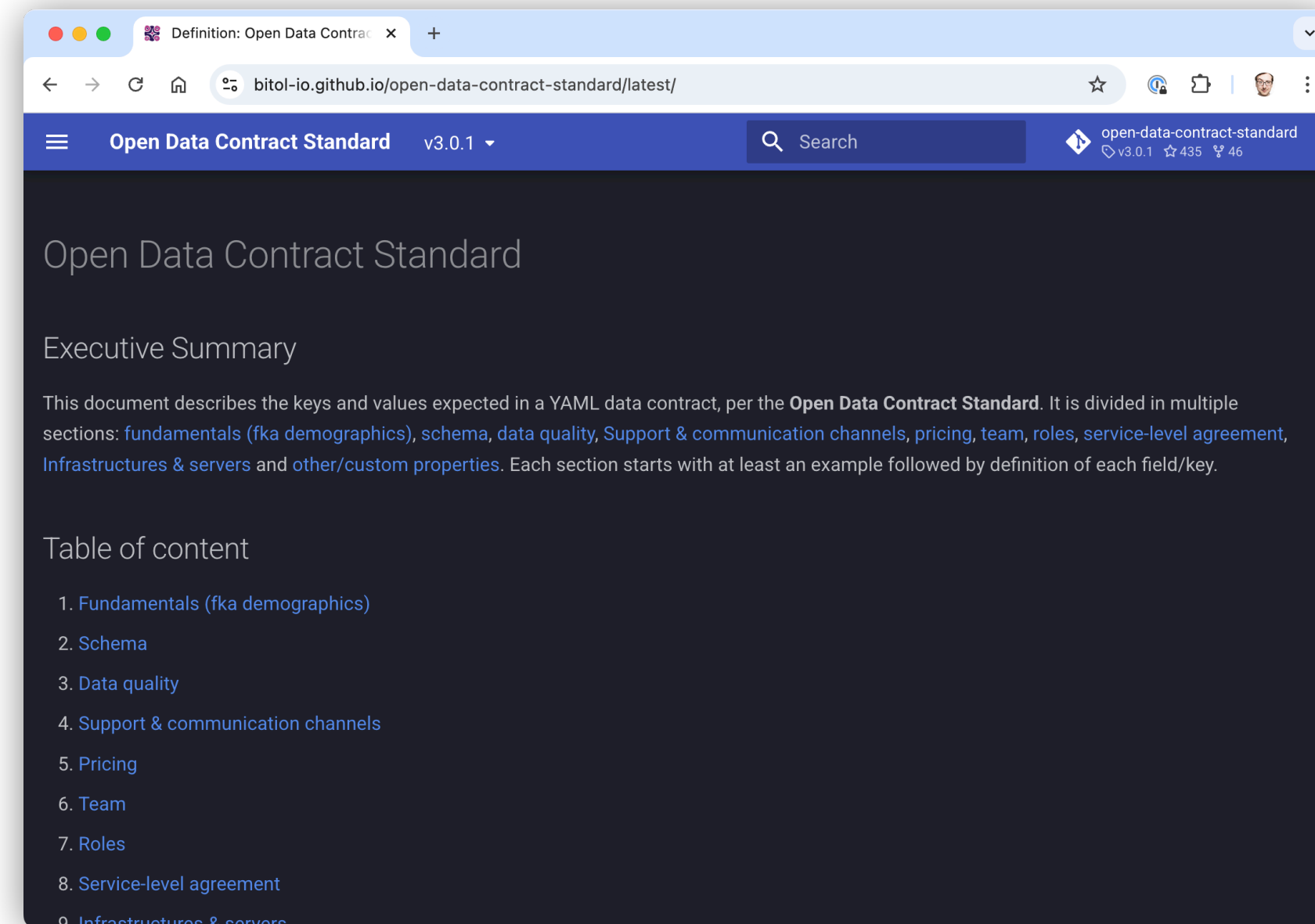
Data Contract Specification

datacontract.com



bitol-io.github.io/open-data-contract-standard/

Data Contract Standard



bitol-io.github.io/open-data-contract-standard/

Approaches

Data-first

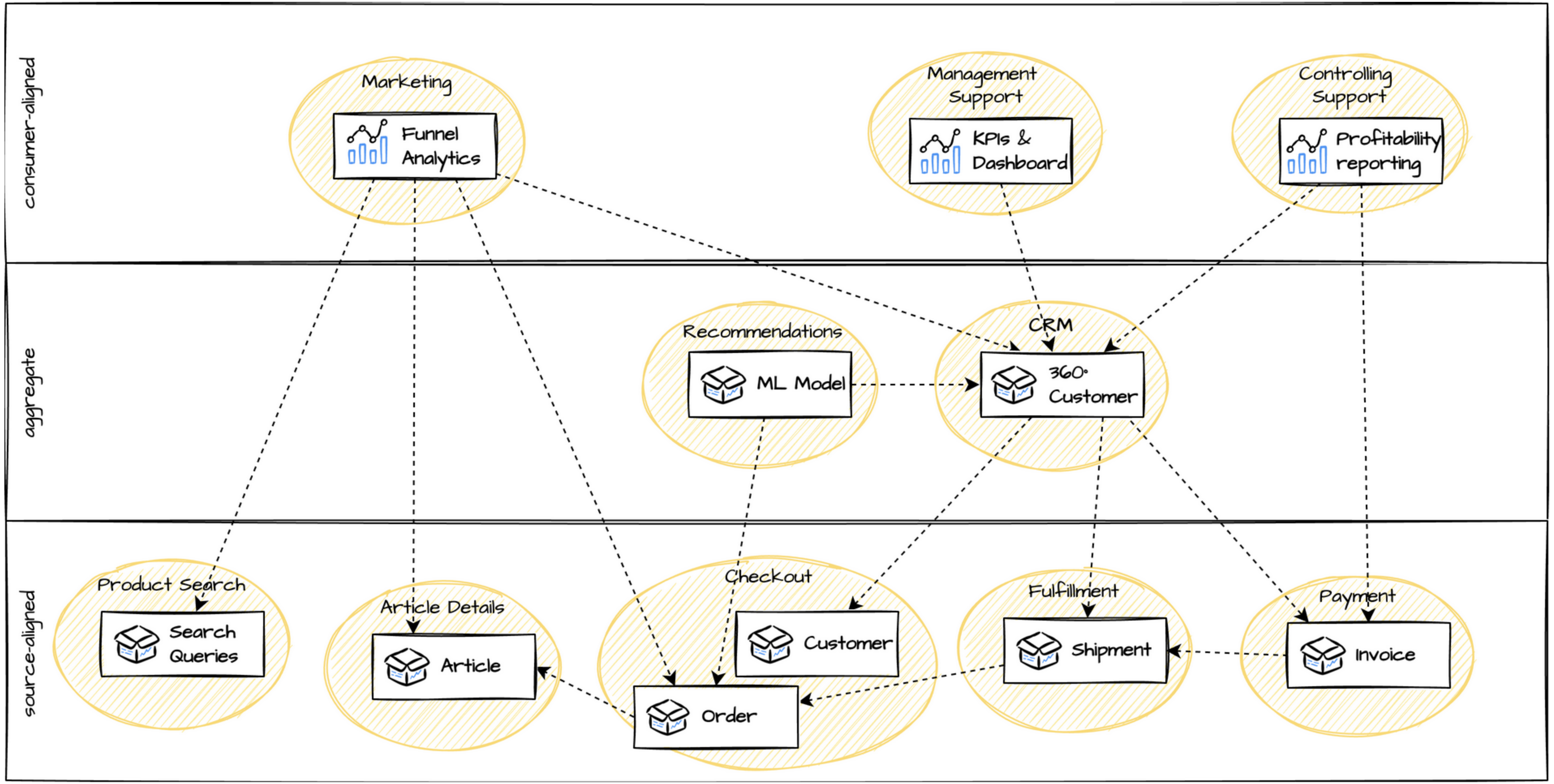
You have a data set.
Now you add the metadata.

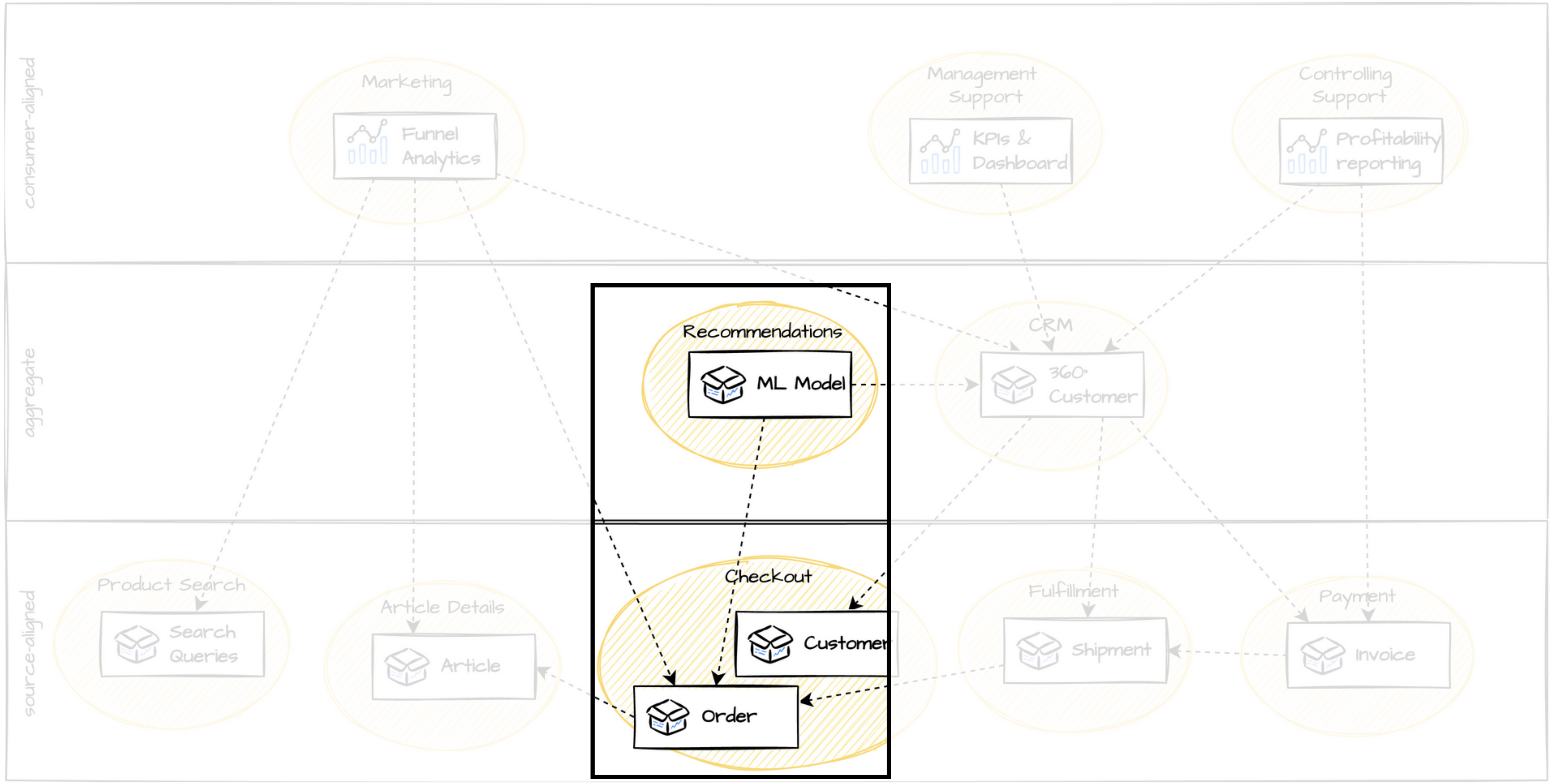
Contract-first

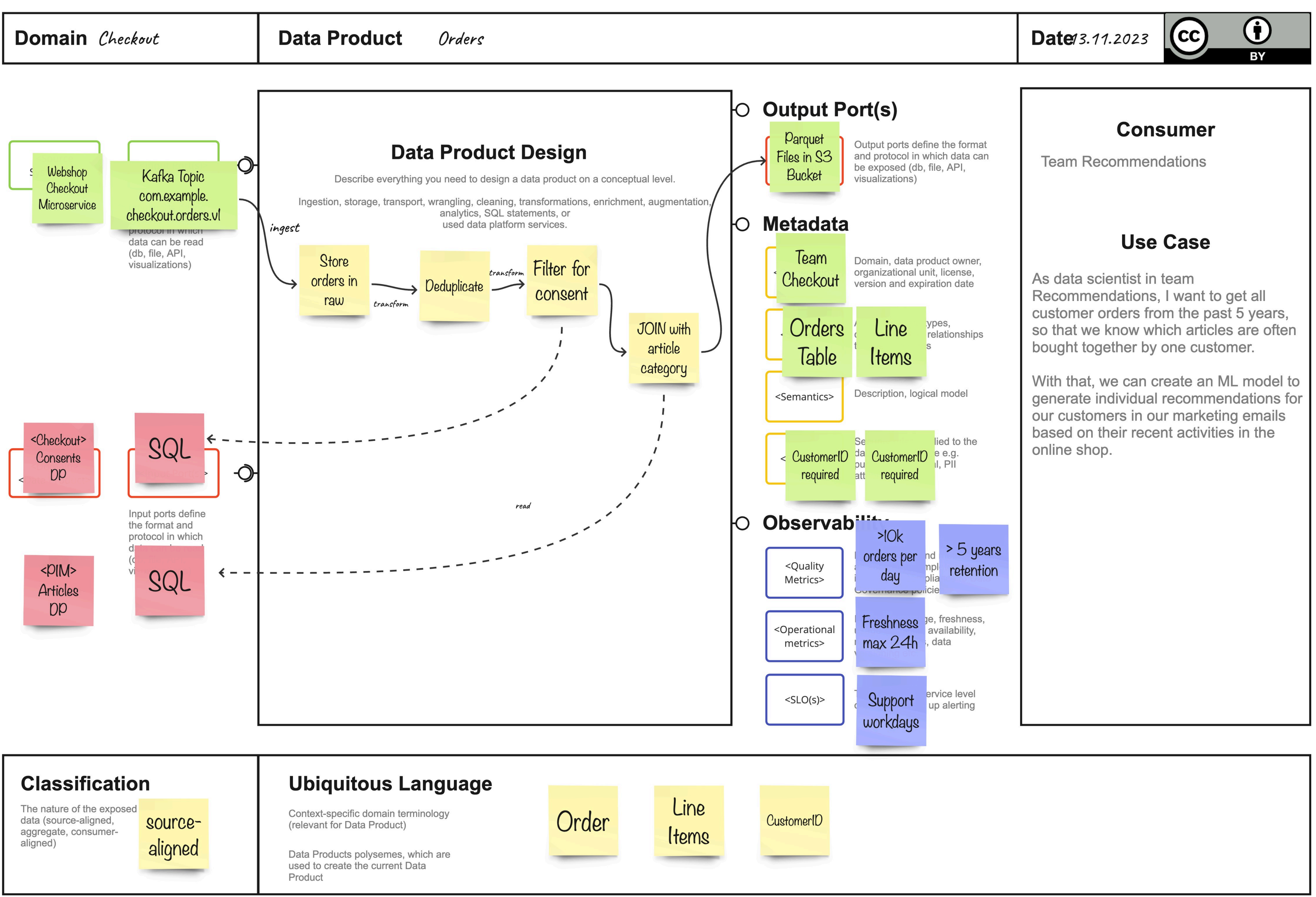
You define requirements as data contract.
Then you build a data product.

Example

Data Product Orders for Team Recommendations







Drive Data Product Development by Business Case

"As data scientist in team Recommendations, I want to get all webshop orders from the past 5 years, so that we know which articles (categories) are often bought together by one customer.

With that, we can create an ML model to generate individual recommendations for our customers in our marketing emails based on their recent activities in the online shop."



Sophia, Data Scientist

Ownership

```
apiVersion: v3.1.0
kind: DataContract
id: orders
name: Orders
version: 1.0.0
status: active
tenant: company-A
team:
|  name: sales-team
tags:
- dataLocation: EU
```


Data Model

```
schema:
- name: shipments_v2
  physicalType: table
  description: "This table contains shipment data, including details about shipment IDs."
  properties:
  - name: shipment_id
    businessName: Shipment ID
    description: Unique identifier for each shipment.
    logicalType: string
    physicalType: uuid
    primaryKey: true
    partitioned: false
    classification: internal
    examples:
    - 123e4567-e89b-12d3-a456-426614174000
    tags:
    - pii:false
    authoritativeDefinitions:
    - type: definition
      url: https://demo.datamesh-manager.com/demo948744545264/definitions/fulfillment/shipment\_id
  - name: order_id
    physicalName: oid
    physicalType: text
    primaryKey: false
```

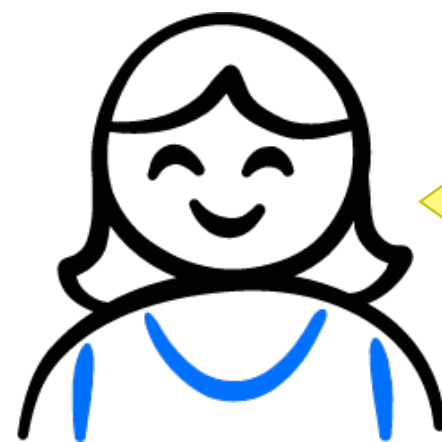
Quality

```
- name: delivery_date
  businessName: Delivery Date
  description: The actual or expected delivery date of the shipment.
  logicalType: date
  physicalType: timestamp_tz
  required: false
  classification: internal
  examples:
    - 2023-10-01T10:00:00Z
  quality:
    - type: sql
      description: The delivery_date is expected to be not before and not more than 30 days after the order_date in 95% of all shipments.
      query: |
        SELECT
        (
          COUNT(*) FILTER (WHERE delivery_date >= order_date AND delivery_date <= DATEADD(DAY, 30, order_date)) * 100.0
          / COUNT(*)
        ) AS percentage_within_range
      FROM shipments_v2
      mustBeGreaterThan: 95
```


Drive Data Product Development by Business Case

"As data scientist in team Recommendations, I want to get all webshop orders from the past 5 years, so that we know which articles (categories) are often bought together by one customer.

With that, we can create an ML model to generate individual recommendations for our customers in our marketing emails based on their recent activities in the online shop."



Ah, good to know.
That is OK for us.

Sophia, Data Scientist



"Sorry, we cannot give you **all** orders, but only orders when the customer expressed their consent for analytical use. That are around 80% of all orders.

John, Product Owner Team Checkout

Terms of Use

description:

purpose: analytical

usage: The data can be used for analytical purposes and AI use cases

limitations: |

| As the dataset is filtered, these dataset cannot be used for financial KPIs

authoritativeDefinitions:

- type: GDPR Policies

| url: <https://example.com/gdpr-policies.html>

Service Levels

```
servicelevels:  
  availability:  
    description: The server is available during support hours  
    percentage: 99.9%  
  retention:  
    description: Data is retained for one year  
    period: P1Y  
    unlimited: false  
  frequency:  
    description: Data is delivered once a day  
    type: batch  
    cron: 0 0 * * *  
  support:  
    description: The data is available during typical business hours at headquarters  
    time: 9am to 5pm in EST on business days  
    responseTime: 1h  
  backup:  
    description: Data is backed up once a week, every Sunday at 0:00 UTC.  
    interval: weekly  
    cron: 0 0 * * 0  
    recoveryTime: 24 hours  
    recoveryPoint: 1 week
```

Servers (Physical Endpoints)

```
servers:  
  production:  
    type: BigQuery  
    project: acme_sales_prod  
    dataset: orders_latest_pii_v1
```



```
apiVersion: v3.1.0
kind: DataContract
id: fulfillment_shipments_v2
name: Shipments v2
version: 2.0.0
status: draft
description:
  purpose: This data can be used for analytical purposes
  usage: Use this to analyze shipments
  limitations: Not suitable for real-time use cases
  authoritativeDefinitions:
    - type: Data Guidelines
      url: https://example.com/data-guidelines.html
schema:
- name: shipments2
  physicalType: table
  physicalName: shipments_v2
  description: "This table contains shipment data, including details about shipment IDs, associated orders, delive
  dataGranularityDescription: Not Aggregated
  properties:
    - name: shipment_id
      businessName: Shipment ID
      logicalType: string
      physicalType: uuid
      description: Unique identifier for each shipment.
      primaryKey: true
      classification: internal
      examples:
        - 123e4567-e89b-12d3-a456-426614174000
      authoritativeDefinitions:
        - type: definition
          url: 
```

Excel Template

AutoSave

fulfillment_shipments_v1 (20)

Search (Cmd + Ctrl + U)

HomeInsertDrawPage LayoutFormulasDataReviewViewAutomate

CommentsShare

A26

x

✓

fx

A

B

C

D

E

F

1**Schema**

2This section describes the data model for one table, message, or object with their properties.

3Copy this sheet for every model in your data contract.

4

5Name

6Type

7Description

8Business Name

9Physical Name

10Data Granularity

11Tags

12

shipments
table
This table contains shipment data, including details about shipment IDs, associated orders, de
Shipments
shipments_v1
Not Aggregated
pii

Property	Business Name	Logical Type	Physical Type	Example(s)	Description
shipment_id	Shipment ID	string	uuid	123e4567-e89b-12d3-a456-426614174000	Unique identifier for each shipment.
order_id	Order ID	string	text	ORD12345	Identifier for the order associated with the shipment.
delivery_date	Delivery Date	date	timestamp_tz	2023-10-01T10:00:00Z	The actual or expected delivery date of the shipment.
carrier	Carrier	string	text	FedEx,UPS	The shipping carrier used for the delivery.
tracking_number	Tracking Number	string	text	1Z999AA10123456784	Tracking number provided by the carrier.
status	Status	string	text	Delivered,In Transit	Current status of the shipment.
inline_object_definition	Inline Object Definition	object	json	{"destination": "New York"}	A JSON representation of additional shipment info
address	Shipment Address	object	JSON		Shipping address details.
address.street	Street	string	text	Marienplatz 1	Street address.

InstructionsFundamentalsSchema shipmentsSupportTeamRolesSLAServersPricingCustom Properties+

ReadyAccessibility: Investigate

<https://github.com/datacontract/open-data-contract-standard-excel-template>

**A data contract
is not a contract**

**A data contract
is not a contract in a legal sense**

**A data contract
is an offer**

**A data contract
is an offer
for consumers to use data with
guarantees under specified
terms and conditions**

**A data contract
is an interface specification**

Benefits

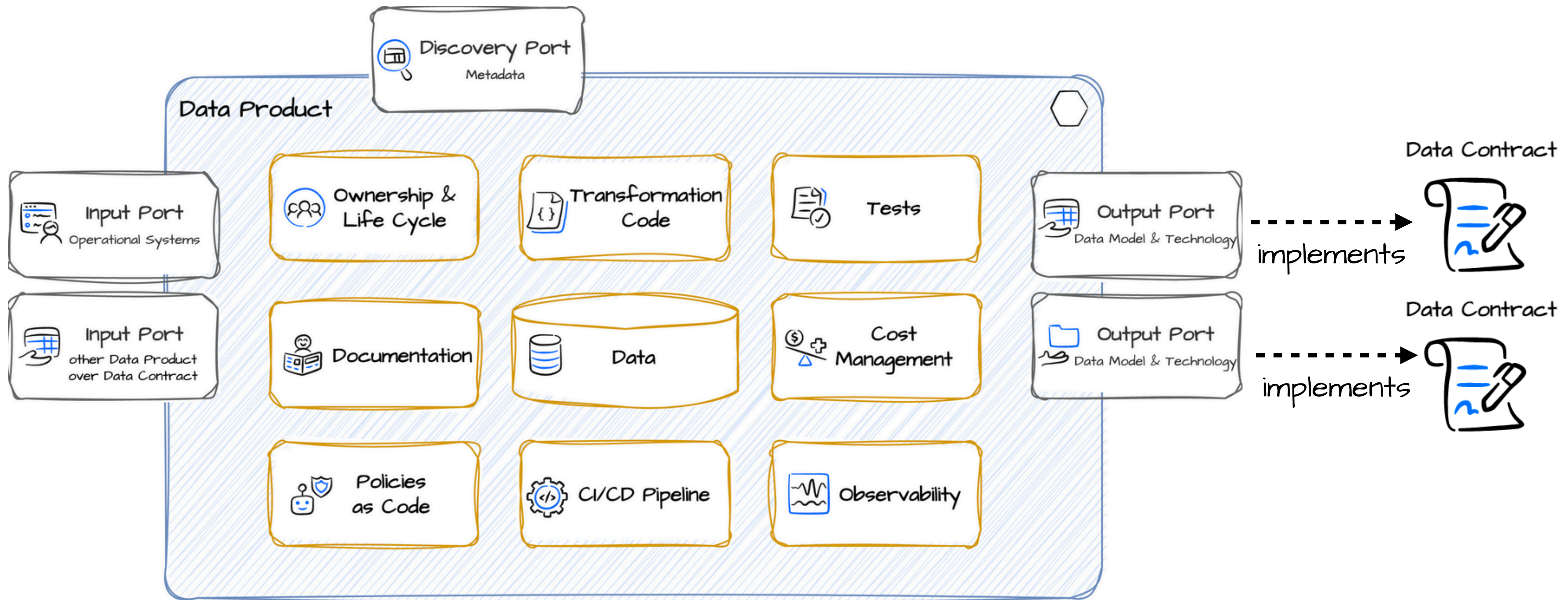
Why Data Contracts became important?

Data Contracts are the Metadata Foundation for Sharing Data with other Teams (and AI)

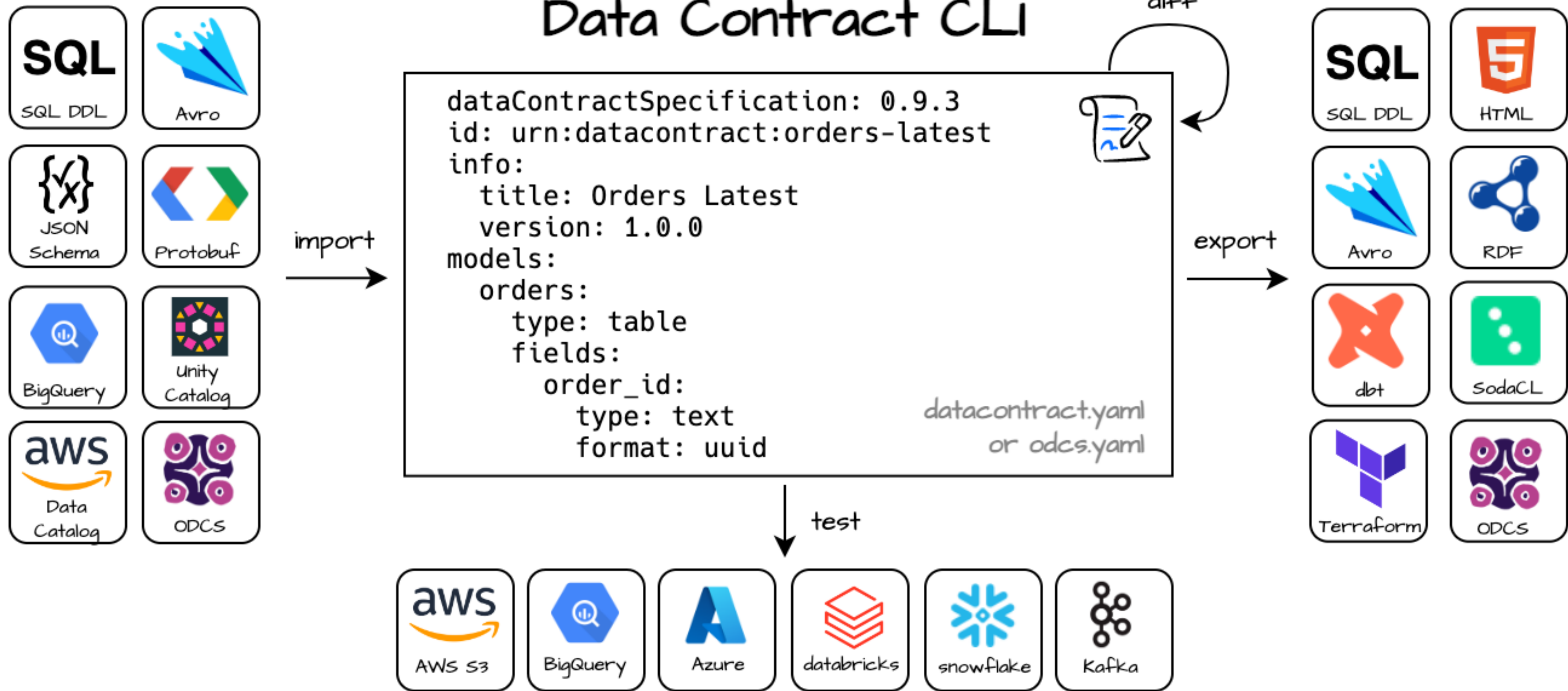
- **Communication:** Collaboration & Requirements Engineering (Excel!)
- **Trust:** Data Contract Testing (CLI!)
- **Discover:** Enterprise Data Marketplace (Entropy Data!)
- **Context:** Semantics, Structure, and SLAs (MCP!)
- **Governance:** Terms of Use (Entropy Data MCP!)

Contract Testing

A data product implements a data contract



Data Contract CLI



datacontract test

```
$ datacontract test datacontract.yaml
```

order_total:

description: Total amount the smallest monetary unit (e.g., cents)

type: long

required: true

customer_id:

description: Unique identifier for the customer.

type: text

minLength: 10

maxLength: 20

customer_email_address:

description: The email address, as entered by the customer. The email

type: text

format: email

required: true

order_total:

description: Total amount the smallest monetary unit (e.g., cents).

type: long

required: true

examples:

- 9999

quality:

- type: sql

description: 95% of all order total values are expected to be between 10 and 499 EUR.

query: |

SELECT quantile_cont(order_total, 0.95) AS percentile_95

FROM orders

mustBeBetween: [1000, 49900]

Data Contract Testing

jochen — -zsh

[jochen@Jochens-MacBook-Pro-2 ~ % datacontract test https://datacontract.com/examples/orders-latest/datacontract.yaml
Testing <https://datacontract.com/examples/orders-latest/datacontract.yaml>

Result	Check	Field	Details
passed	Check that JSON has valid schema	orders	All JSON entries are valid.
passed	Check that JSON has valid schema	line_items	All JSON entries are valid.
passed	Check that field order_id is present	orders	
passed	Check that field order_timestamp is present	orders	
passed	Check that field order_total is present	orders	
passed	Check that field customer_id is present	orders	
passed	Check that field customer_email_address is present	orders	
passed	Check that field processed_timestamp is present	orders	
passed	row_count >= 5	orders	
passed	Check that required field order_id has no null values	orders.order_id	
passed	Check that unique field order_id has no duplicate values	orders.order_id	
passed	duplicate_count(order_id) = 0	orders.order_id	
passed	Check that required field order_timestamp has no null values	orders.order_timestamp	
passed	Check that required field order_total has no null values	orders.order_total	
passed	Check that required field customer_email_address has no null values	orders.customer_email_address	
passed	Check that required field processed_timestamp has no null values	orders.processed_timestamp	
passed	Check that field lines_item_id is present	line_items	
passed	Check that field order_id is present	line_items	
passed	Check that field sku is present	line_items	
passed	values in (order_id) must exist in orders (order_id)	line_items.order_id	
passed	row_count >= 5	line_items	
passed	Check that required field lines_item_id has no null values	line_items.lines_item_id	
passed	Check that unique field lines_item_id has no duplicate values	line_items.lines_item_id	

data contract is valid. Run 23 checks. Took 6.776398 seconds.
jochen@Jochens-MacBook-Pro-2 ~ %

Jobs

datacontract-test

Data Contract Test Results

Run details

Usage

Workflow file

GitHub Actions / Data Contract Test Results

succeeded 1 minute ago in 0s

27 passed, 1 failed and 0 skipped

tests 27 passed, 1 failed

./TEST-datacontract.xml

28 tests were completed in 4s with 27 passed, 1 failed and 0 skipped.

Test suite	Passed	Failed	Skipped	Time
snowflake_orders_pii_v2	27	1		4s

snowflake_orders_pii_v2

```
orders.ORDER_ID
  ✓ orders__ORDER_ID__field_is_present
  ✓ orders__ORDER_ID__field_type
  ✓ orders__ORDER_ID__field_required
orders.PREVIOUS_ORDER_ID
  ✓ orders__PREVIOUS_ORDER_ID__field_is_present
  ✗ orders__PREVIOUS_ORDER_ID__field_type
    Type Mismatch, Expected Type: NUMBER; Actual Type: TEXT
orders.CUSTOMER_ID
  ✓ orders__CUSTOMER_ID__field_is_present
  ✓ orders__CUSTOMER_ID__field_type
  ✓ orders__CUSTOMER_ID__field_required
```



Data Contract Manager

x

+

localhost:8080/demo129500424135/dataproducts/orders/outputports/60107bb0-6711-4d30-8357-d2acac96fc74/checks

☆

Search

Data Contracts

Data Products

Data Governance

AI

More

▼

ACME

🏠

>

Data Products

>

Orders

>

Output Ports

>

snowflake_orders_pii_v2

>

Checks

Checks

d8445796-d5f0-4b31-9e78-a932dc2a5977

Info

Information about the checks

Overall Result

✓

Passed

Last Run

9 minutes ago

Server

[orders/snowflake_orders_pii_v2](#)

Data Contract

Defines the syntax, semantics, and quality

Orders

All order-created events, with PII.

Checks

Test summary

Result	Check	Details
<div><div>✓</div>Passed</div>	Check that the datacontract.yaml is valid	
<div><div>✓</div>Passed</div>	Check that the field order_id exists <div><div></div>orders<div><div></div>orders_id</div></div>	
<div><div>✓</div>Passed</div>	Check that the values in field order_id are unique <div><div></div>orders<div><div></div>orders_id</div></div>	
<div><div>✓</div>Passed</div>	Custom Query: select count(*) from orders > 1000000 <div><div></div>orders</div>	
<div><div>✓</div>Passed</div>	freshness	

Enterprise Data Marketplace

(Commercial Tooling*)

*** Advertisement**



Home > Marketplace > Data Products

Data Products

Discover data products for your business case and request access.

Add Data Product

Search

Owner

Type

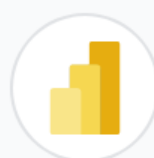
Status

Output Port

Environment

Tag

Download



DATA CONSUMER

Article Profitability Analysis

Controlling Team active

A PowerBI report that combines revenue, cost, and margin to deliver actionable insights into the profitability of each article. It is designed to help...

demo



APPLICATION

Article Service

Products active

The central self-contained system for articles.

Articles API 1

demo



DATA PRODUCT

Articles history

Products active

No description

Articles History 2

demo



DATA PRODUCT

Articles latest

Products active

No description

Articles Latest 1

demo



DATA PRODUCT

Customer Cohorts

Marketing active

No description

Customer Cohorts 1

demo



DATA PRODUCT

Customers

Sales Team active

Customers are registered users and unregistered guest buyers in the online shop. Dataset all customers since launch of the online shop.

Customers History NPII 1

Customers History PII 1

Customers Latest NPII 1

Customers Latest PII 1



DATA CONSUMER

Funnel Analytics

Marketing active

No description

demo



DATA CONSUMER

Monthly Target Performance Report

Controlling Team active

No description



APPLICATION

OpenSearch

Search active

The index for the ecommerce search

Elasticdump 1

demo



Home > Studio > Data Contracts > Orders v2



Orders v2

snowflake_orders_pii_v2 2.0.0

Sales Team Connected to Git active

PII

Sensitive

Data Contract Specification 1.1.0

Watching 6

Edit

Request Access

enlarge

apply layout

+

-

⌂

orders		
ORDER_ID	string	
PREVIOUS_ORDER_ID	string	
CUSTOMER_ID	string	
EMAIL	string	
PHONE_NUMBER	text	
INVOICE_ADDRESS	object	
ADDRESS_LINE	text	
CITY_LINE	text	
ORDER_DATE	timestamp	
ORDER_TOTAL	decimal	

line_items		
LINE_ITEM_ID	string	
ORDER_ID	string	
ARTICLE_SKU	string	

React Flow

Info

Information about the data contract

Description

All order-created events, with PII.

Tags

popular

Terms

Terms and conditions of the data contract

Usage

Max. 10x queries per day

Billing

\$1000 / month

Limitations

Not suitable for real-time use cases

Notice Period

3 months

Data Product

The data product with this data contract



Orders PII (Orders)

ORDER_DB.ORDERS_PII_V2

Passed 32 checks

4 months ago



Passed

Run Data Contract Test

Use [Data Contract CLI API](#) to run data contract tests.

Versions

Shows details about other major versions of this data contract

Orders v2



Create v3

Show Changes

Request Access

demo.datamesh-manager.com/demo882420088379/datausageagreements/request-acc...

Request Access

Orders

on output port

snowflake_orders_npii_v2

The system will create a data usage agreement for the team Checkout to approve.

Consumer

Required

Data Product

Request access for one of your data products.

Team

Request access for everybody in your domain team.

User

Request access for yourself.

Consumer Data Product

Required

Recommendations ML Model (Team Recommendations)

Select your data product that needs to access and use the provided data.

Purpose

Required

We need to get all webshop orders from the past 5 years so that we know which articles (categories) are often bought together by one customer. With that, we can create an ML model to generate individual recommendations for our customers in our marketing emails based on their recent activities in the online shop.

Why do you want access and what do you want to do with the data?

Cancel

Customize

Request access

Approve Access Request

Team **Recommendations** requests access to data product **Orders**.

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


Approve

Reject

Data Platform

Status of the data platform integration

Status

 Permissions granted

Updated

1 minutes ago

Role

agreement_55gCnBIHYdUWnUJJKYbYIs_role

Agent

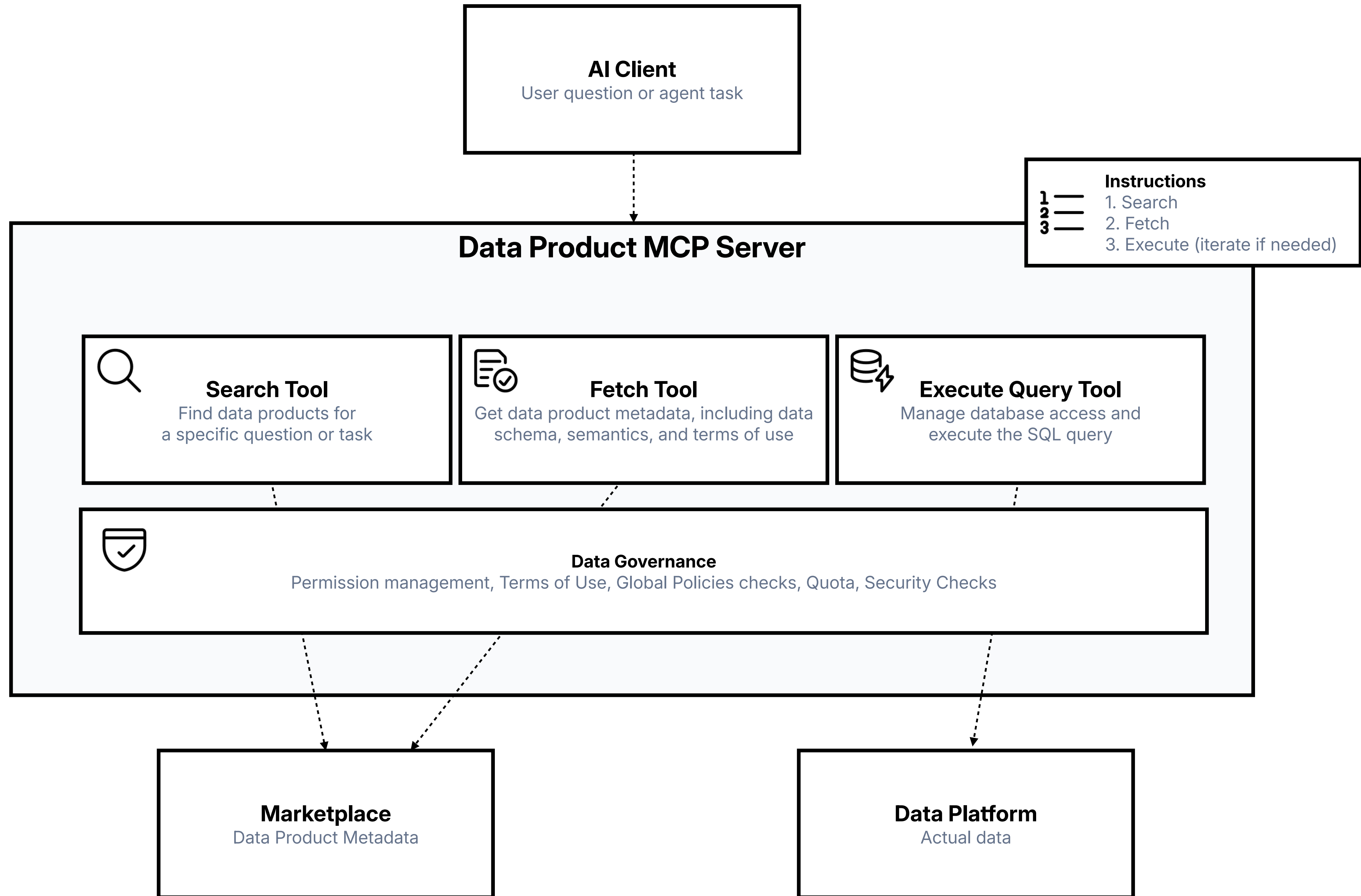
Data Mesh Manager Platform Agent v0.1

Details

```
CREATE ROLE agreement_55gCnBIHYdUWnUJJKYbYIs_role;  
GRANT ROLE op_orders_snowflake_orders_npii_v2_role TO ROLE  
agreement_55gCnBIHYdUWnUJJKYbYIs_role;  
GRANT ROLE agreement_55gCnBIHYdUWnUJJKYbYIs_role TO ROLE ;
```

Connect AI
with Business Data
(securely)





 Good afternoon, Jochen



Sonnet 4.5



 Good afternoon, Jochen

Create a CSV for mailchimp with our top 5000 customers.



Sonnet 4.5 ▾



Data Governance ✨

(Commercial Tooling)

Terms & Conditions

terms:

usage: "The data can be used for analytical and data science use cases, as the customer has expressed their consent."

limitations: |

As the dataset is filtered, these data set cannot be used to aggregate financial KPIs.

Not suited for real-time use cases.

billing: \$1000 per month

noticePeriod: P3M

Data Contracts Cover Governance Aspects

- **Ownership** Who is responsible for providing data? **Ownership (Info)**
- **Quality** What quality do we provide? **Quality Attributes & Contract Testing**
- **Compliance** What are we allowed to do with data? **Terms and Conditions**
- **Legal** In which region may data be stored? **Terms and Conditions**
- **Privacy** What is personal (PII) data? **Data Model Attribute**
- **Classification** What sensitivity level has my data? **Data Model Attributes**
- **Security** Who has access to which data and why? --> **Access Requests**

Data Contracts are the Metadata Foundation for Sharing Data with other Teams (and AI)

- **Communication:** Collaboration & Requirements Engineering (Excel!)
- **Trust:** Data Contract Testing (CLI!)
- **Discover:** Enterprise Data Marketplace (Entropy Data!)
- **Context:** Semantics, Structure, and SLAs (MCP!)
- **Governance:** Terms of Use (MCP!)

Summary

Data Contract

~ API for data

Data Contract

promotes collaboration

Data Contract

enforces data quality

Data Contract

is not a contract

Links

Open Source

- <https://bitol-io.github.io/open-data-contract-standard/latest/>
- <https://bitol-io.github.io/open-data-product-standard/latest/>
- cli.datacontract.com

Commercial

- entropy-data.com
- <https://entropy-data.com/learn/mcp>



Data Contracts

JOCHEN CHRIST, ENTROPY DATA

INNOQ



JOCHEN CHRIST

/in/jochenchrist