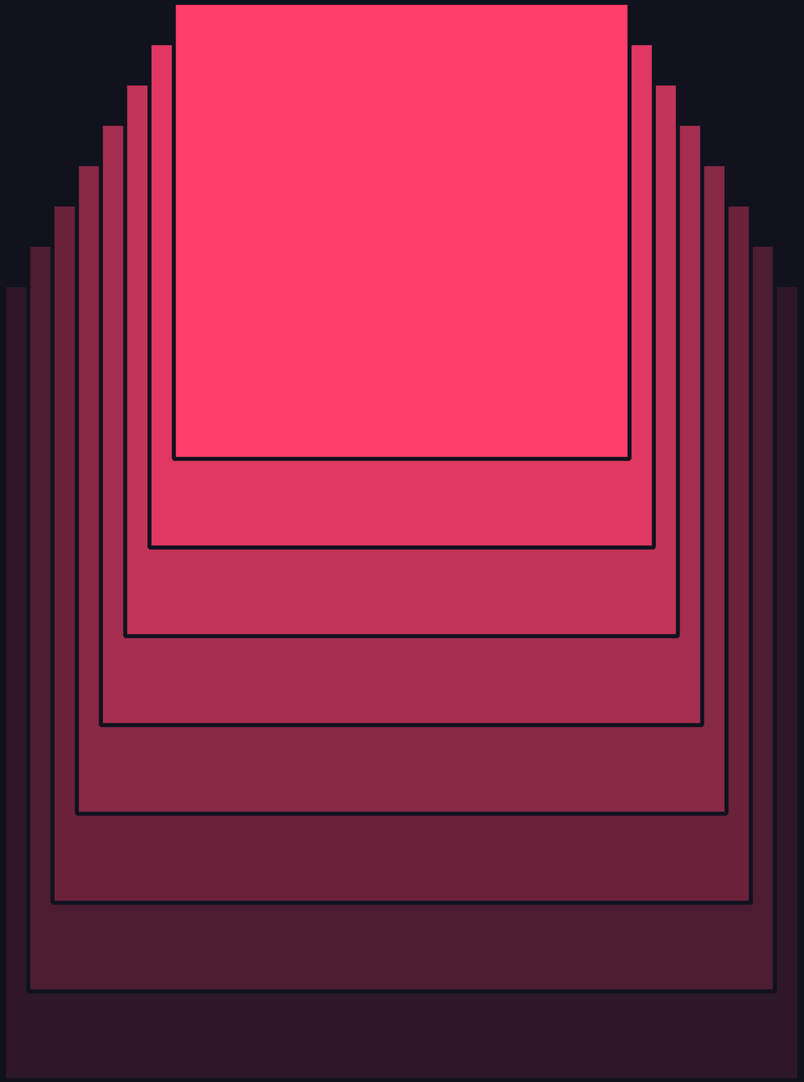
DATA+AI SUMMIT
BY databricks

BUILDING HIGH-QUALITY AND TRUSTED DATA PRODUCTS

WITH DATABRICKS



Karthik Subbarao | Pawarit Laosunthara

June 2024

Karthikeya Sampa Subbarao

Databricks EMEA

❑ Specialist Solutions Architect @ **Databricks**

❑ Experienced Software Engineer & Architect

❑ 10+ years in the Tech industry

❑ Data Architecture, Security & Governance SME

CONNECT

Pawarit Laosunthara

Databricks AMER

❑ Sr. Solutions Architect @ Databricks

❑ Working with Databricks’ largest customers in Logistics, Financial Services, Manufacturing

❑ Previous roles

  ❑ Tech Lead at Thoughtworks
  ❑ Data Scientist at Airbus

CONNECT

Agenda

- Data Products & Lifecycle
- Data Contracts & Governance
- Publishing & Discovery
- Demo
- Interoperability
- Takeaways

Product safe harbor statement

This information is provided to outline Databricks' general product direction and is for informational purposes only. Customers who purchase Databricks services should make their purchase decisions relying solely upon services, features, and functions that are currently available. Unreleased features or functionality described in forward-looking statements are subject to change at Databricks discretion and may not be delivered as planned or at all

Data Products

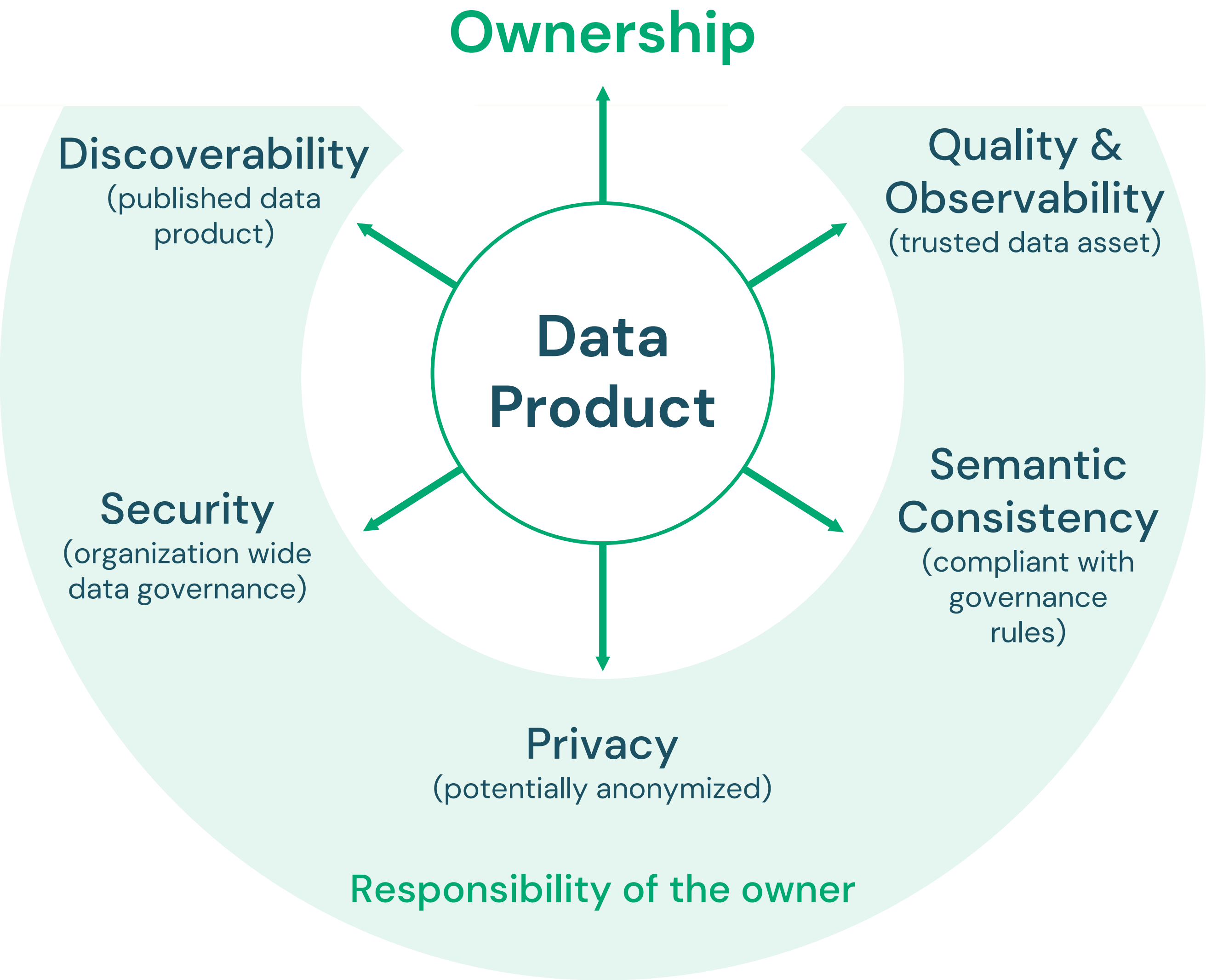
Data and “product thinking”

To publish data as data products, “product thinking” needs to be applied

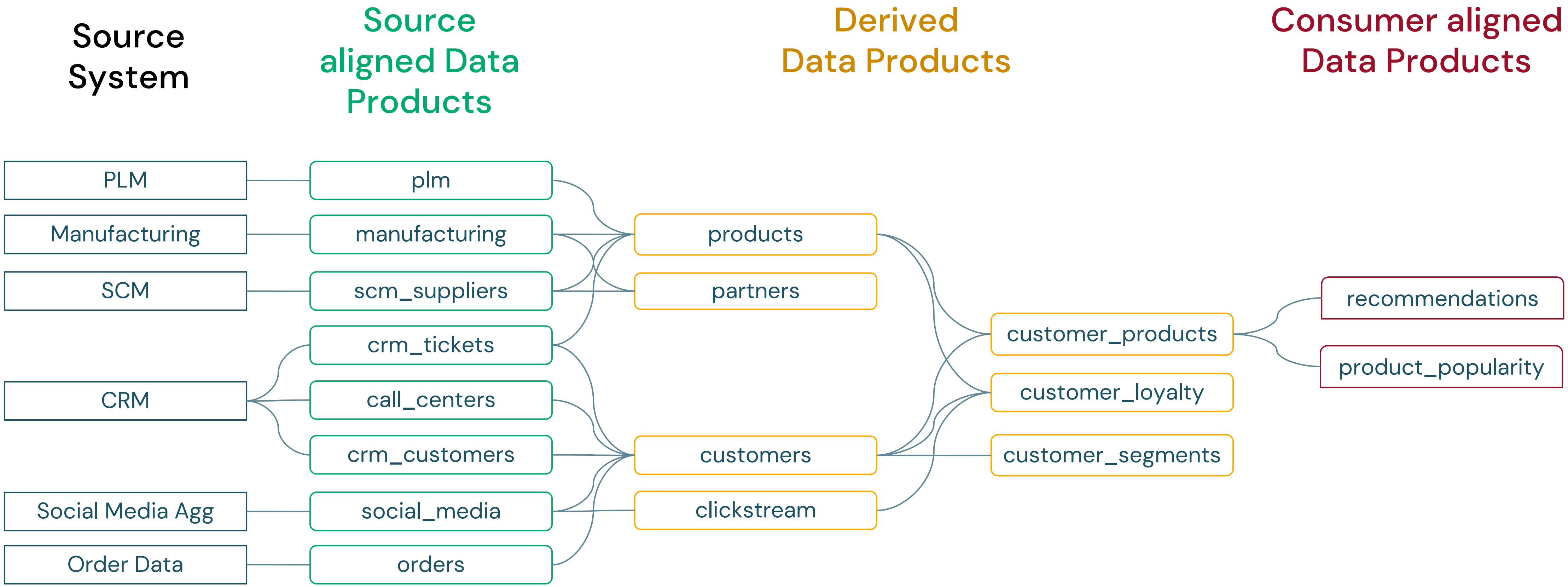
Data products should:

- have an owner and is built for specific audiences
- follow a defined product life cycle
- be defined and described by data contracts
- be published following an agreed governance process

Adding Data Product attributes to the concept



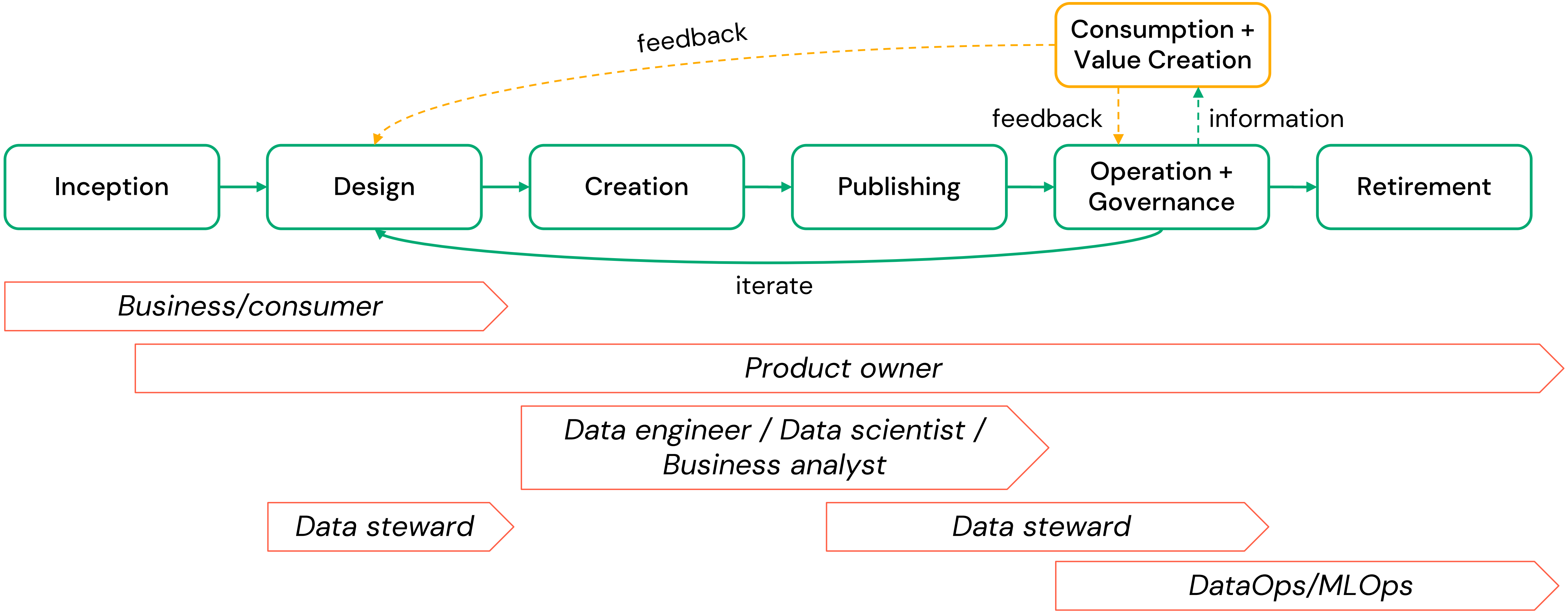
Data product hierarchy



Data Product Lifecycle

Data Products

Typical Lifecycle



Consumption + Value Creation

feedback

feedback

information

Inception

Design

Creation

Publishing

Operation + Governance

Retirement

iterate

Business/consumer

Product owner

Data engineer / Data scientist / Business analyst

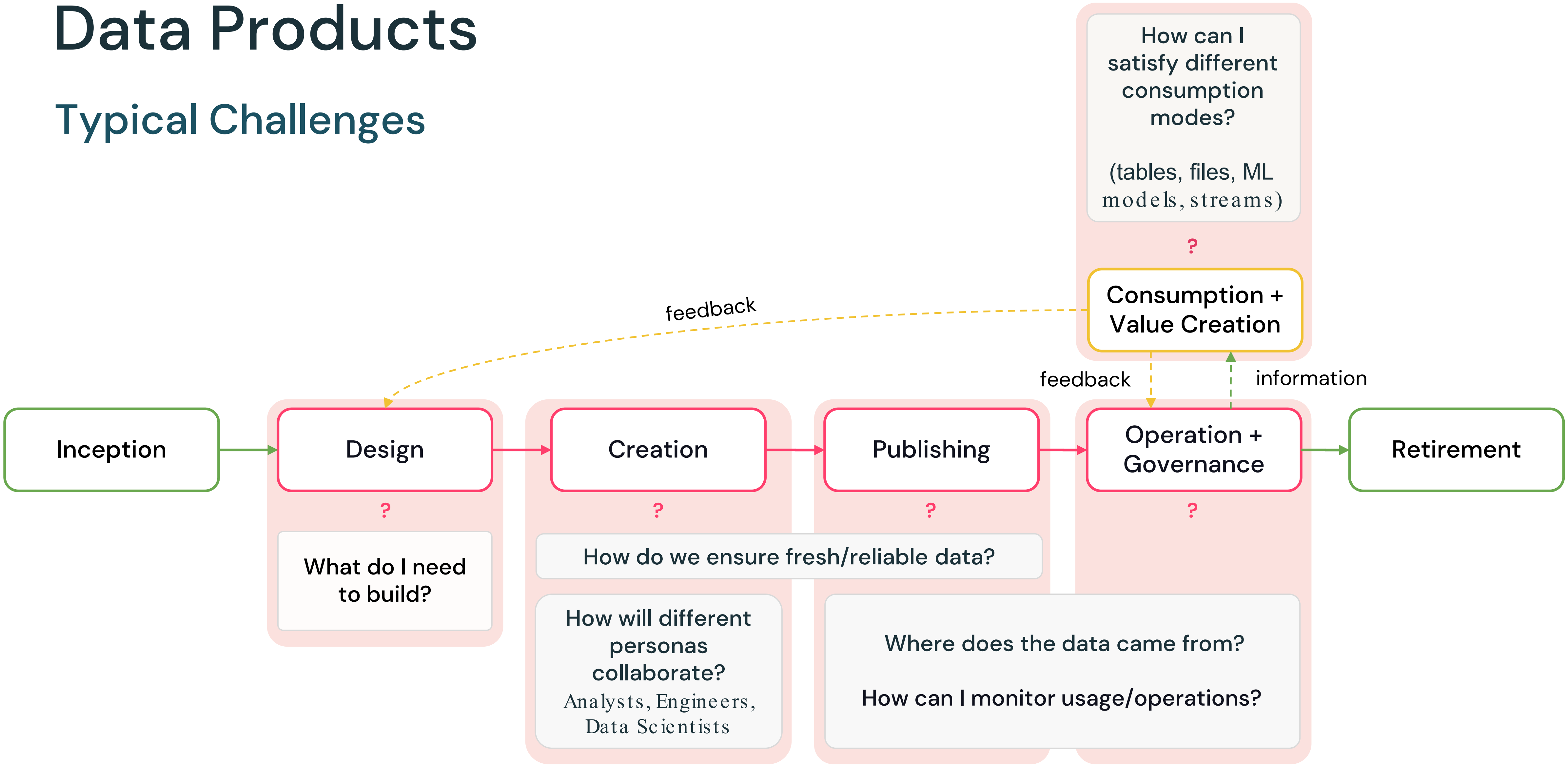
Data steward

Data steward

DataOps/MLOps

Data Products

Typical Challenges



How can I satisfy different consumption modes?

(tables, files, ML models, streams)

?

Consumption + Value Creation

feedback

feedback

information

Inception

Design

Creation

Publishing

Operation + Governance

Retirement

?

?

?

?

What do I need to build?

How do we ensure fresh/reliable data?

How will different personas collaborate?

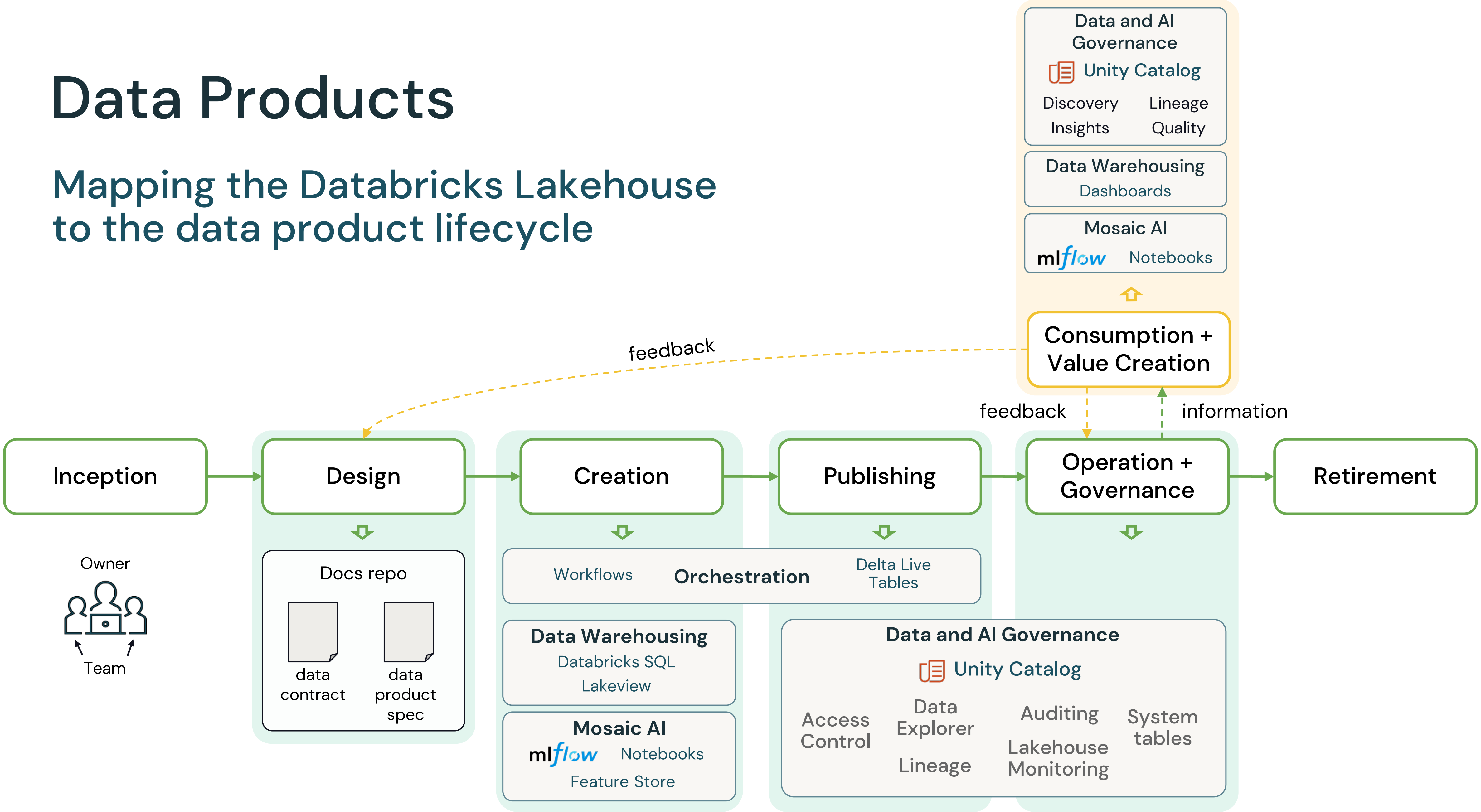
Analysts, Engineers, Data Scientists

Where does the data came from?

How can I monitor usage/operations?

Data Products

Mapping the Databricks Lakehouse to the data product lifecycle



Data Contract and Governance

Data Contract

Data description

name, owner, description, source systems, …

Data SLAs

last update, expiration dates, retention time, usage restrictions, code of conduct, re-sharing conditions, …

Data schema

tables, columns, anonymization and encryption info, …

Security

who is allowed to use the data product

Data quality

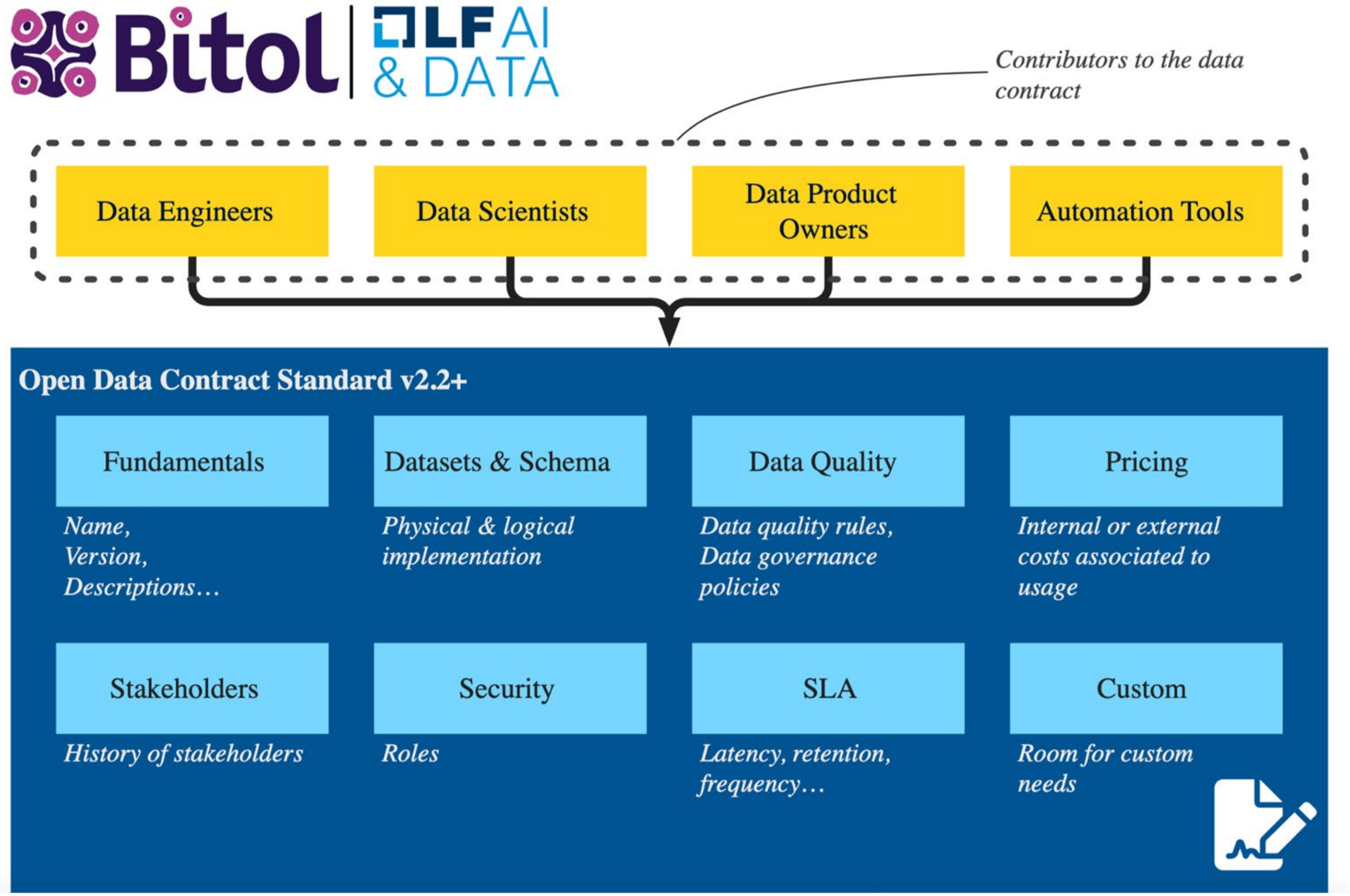
applied quality checks, quality metrics, …

Explanatory add-ons (optional)

notebook, dashboard, sample code, …

Bitol – Linux Foundation AI & Data sandbox project

Example: Open Data Contract Standard (ODCS)



Contributors to the data contract

Data Engineers

Data Scientists

Data Product Owners

Automation Tools

Open Data Contract Standard v2.2+

Fundamentals

Name, Version, Descriptions…

Datasets & Schema

Physical & logical implementation

Data Quality

Data quality rules, Data governance policies

Pricing

Internal or external costs associated to usage

Stakeholders

History of stakeholders

Security

Roles

SLA

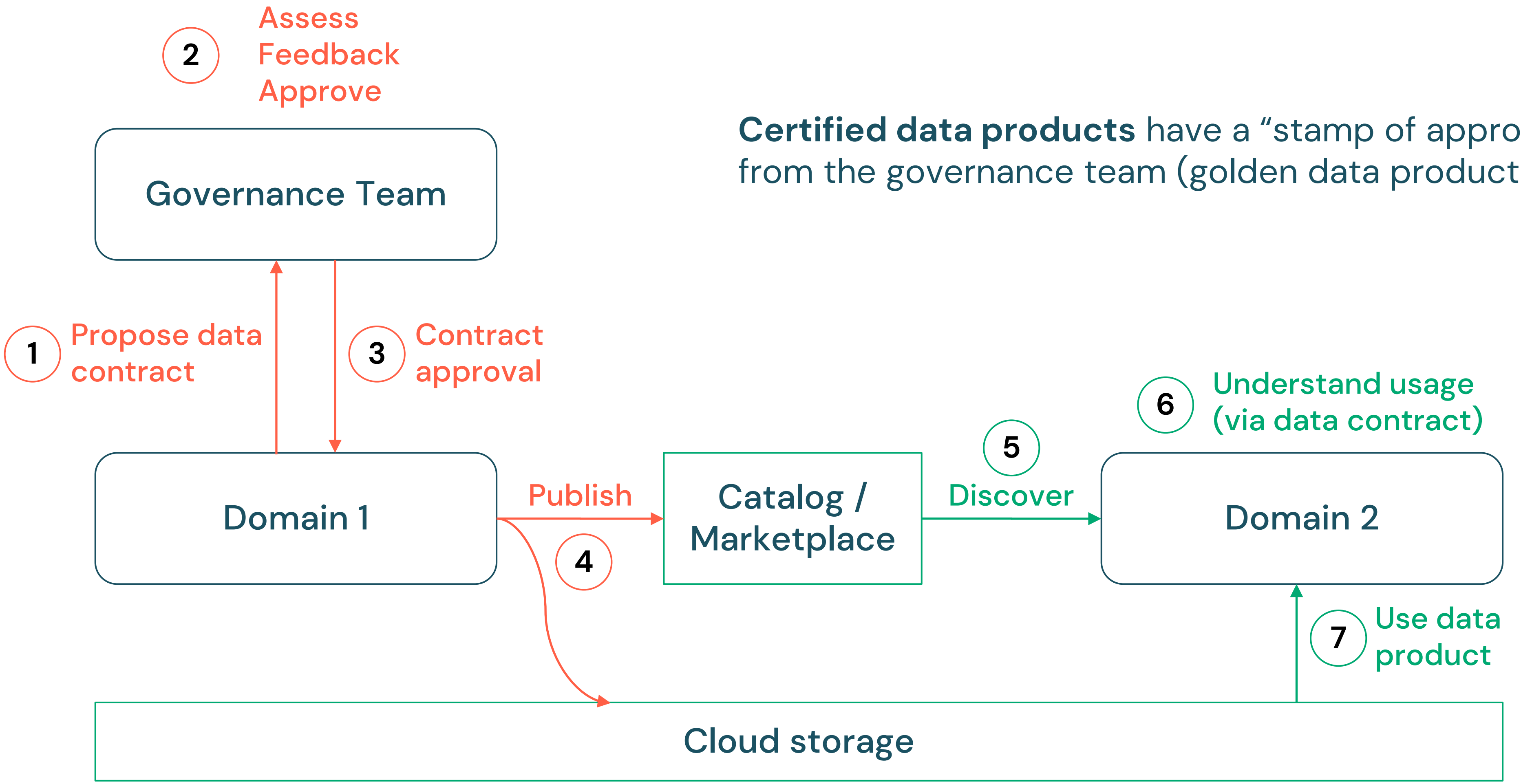
Latency, retention, frequency…

Custom

Room for custom needs

Data Product Certification

Example process to achieve standards and consistency



Certified data products have a “stamp of approval” from the governance team (golden data products)

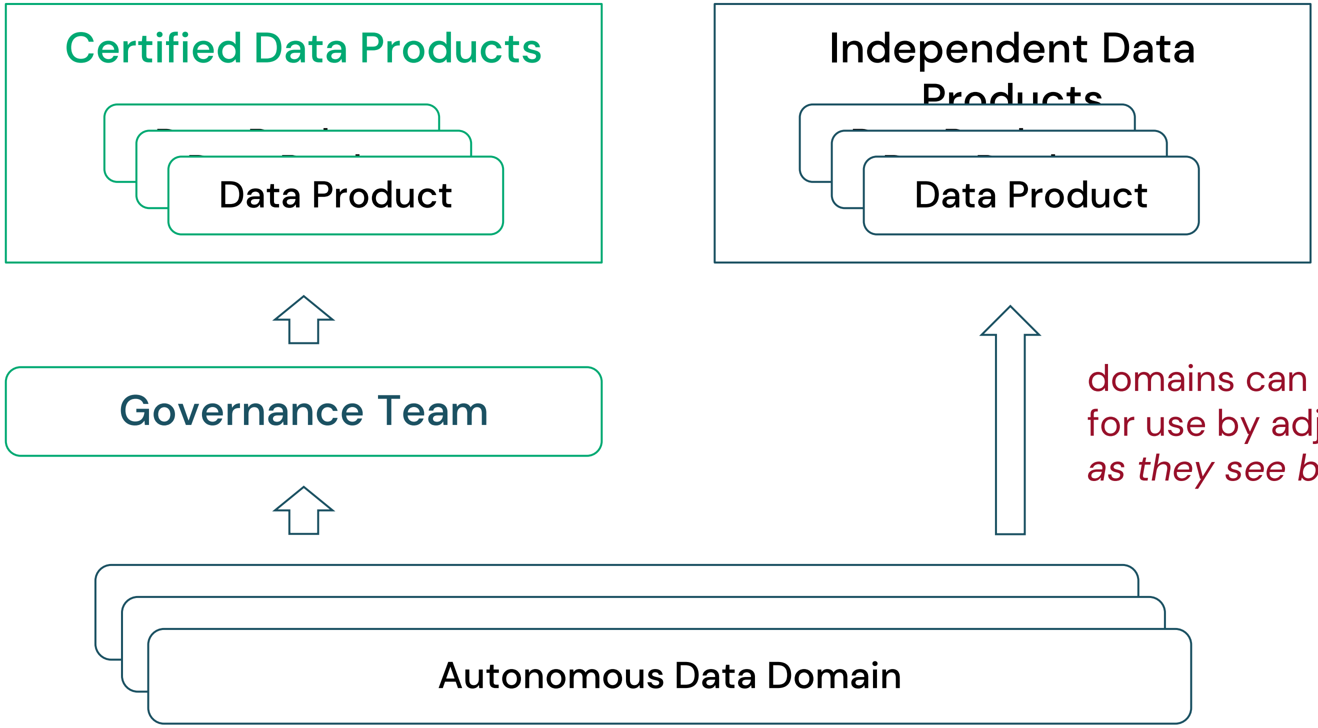
Independent and certified data products

Balance between centralization and autonomy

high quality data products
Semantically consistent
easy to integrate with
other data products

agree on rules and policies
for certified data products

approve and govern
data contracts



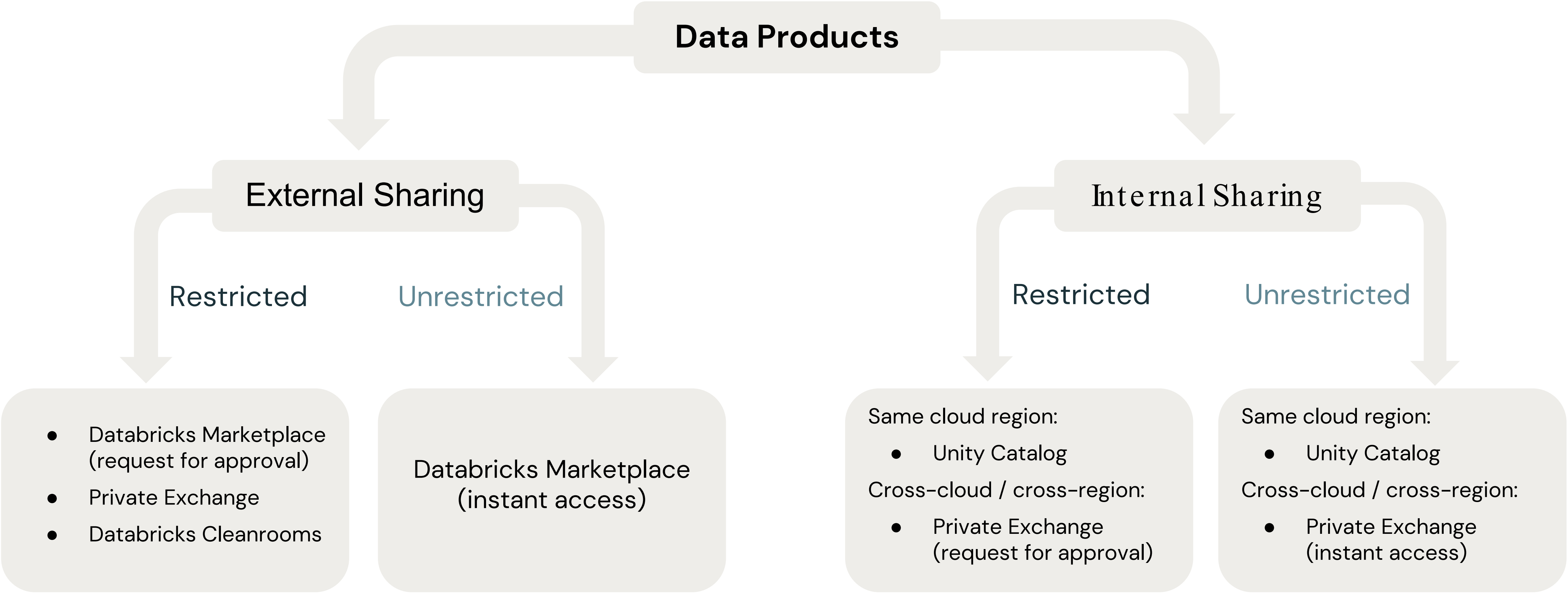
good quality
...but no guarantee that they
can be combined easily

domains can publish data products
for use by adjacent teams
as they see best

Data Product Publishing and Discovery

Publishing use cases for data products

Share them internally / externally and with or without restrictions



Data Products

External Sharing

Restricted

- Databricks Marketplace (request for approval)
- Private Exchange
- Databricks Cleanrooms

Unrestricted

Databricks Marketplace (instant access)

Internal Sharing

Restricted

Same cloud region:

- Unity Catalog

Cross-cloud / cross-region:

- Private Exchange (request for approval)

Unrestricted

Same cloud region:

- Unity Catalog

Cross-cloud / cross-region:

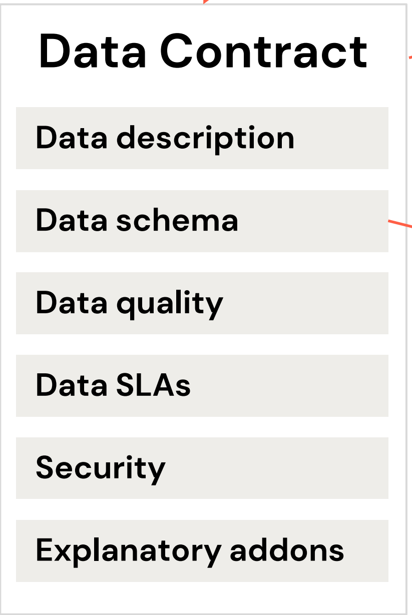
- Private Exchange (instant access)

Data Contract with Unity Catalog metadata

Ownership

Tags derived from data contract, e.g. PII

Markdown allows hyperlinks to link back to external data contract*

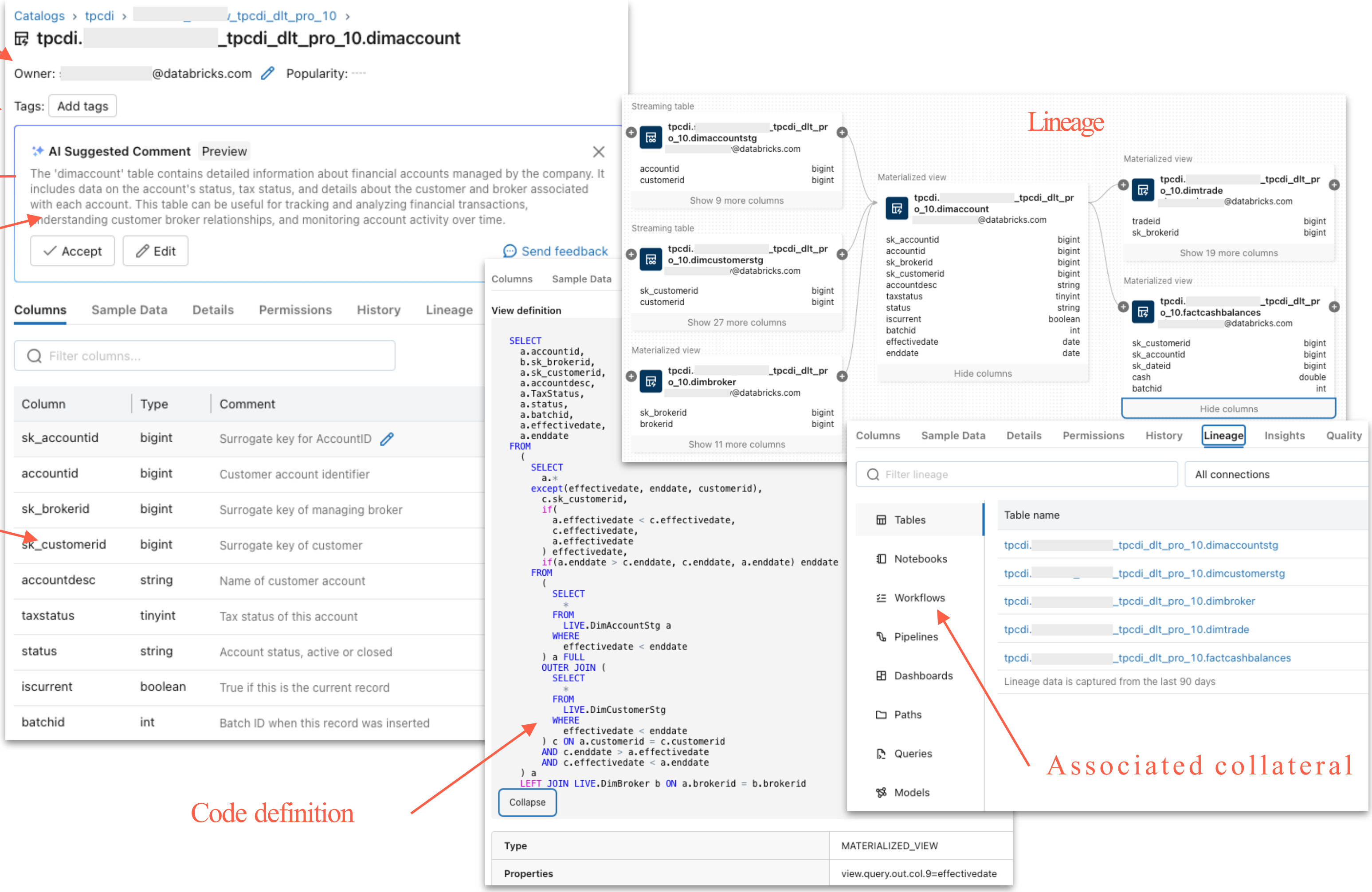


Data Contract

- Data description
- Data schema
- Data quality
- Data SLAs
- Security
- Explanatory addons

Stored externally

Summaries and important details



Catalogs › tpcdi › [redacted]_tpcdi_dlt_pro_10 ›

tpcdi.[redacted]_tpcdi_dlt_pro_10.dimaccount

Owner: [redacted]@databricks.com Popularity:

Tags: Add tags

AI Suggested Comment Preview

The 'dimaccount' table contains detailed information about financial accounts managed by the company. It includes data on the account's status, tax status, and details about the customer and broker associated with each account. This table can be useful for tracking and analyzing financial transactions, understanding customer broker relationships, and monitoring account activity over time.

Column	Type	Comment
sk_accountid	bigint	Surrogate key for AccountID
accountid	bigint	Customer account identifier
sk_brokerid	bigint	Surrogate key of managing broker
sk_customerid	bigint	Surrogate key of customer
accountdesc	string	Name of customer account
taxstatus	tinyint	Tax status of this account
status	string	Account status, active or closed
iscurrent	boolean	True if this is the current record
batchid	int	Batch ID when this record was inserted

View definition

```
SELECT
  a.accountid,
  b.sk_brokerid,
  a.sk_customerid,
  a.accountdesc,
  a.TaxStatus,
  a.status,
  a.batchid,
  a.effectivedate,
  a.enddate
FROM
  (
    SELECT
      a.*
    except(effectivedate, enddate, customerid),
      c.sk_customerid,
      if(
        a.effectivedate < c.effectivedate,
        c.effectivedate,
        a.effectivedate
      ) effectivedate,
      if(a.enddate > c.enddate, c.enddate, a.enddate) enddate
    FROM
      (
        SELECT
          *
        FROM
          LIVE.DimAccountStg a
        WHERE
          effectivedate < enddate
      ) a FULL
      OUTER JOIN (
        SELECT
          *
        FROM
          LIVE.DimCustomerStg
        WHERE
          effectivedate < enddate
      ) c ON a.customerid = c.customerid
      AND c.enddate > a.effectivedate
      AND c.effectivedate < a.enddate
  ) a
  LEFT JOIN LIVE.DimBroker b ON a.brokerid = b.brokerid
```

Type	MATERIALIZED_VIEW
Properties	view.query.out.col.9=effectivedate

Code definition

Lineage

Streaming table
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimaccountstg
[redacted]@databricks.com

accountid	bigint
customerid	bigint

Show 9 more columns

Streaming table
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimcustomerstg
[redacted]@databricks.com

sk_customerid	bigint
customerid	bigint

Show 27 more columns

Materialized view
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimbroker
[redacted]@databricks.com

sk_brokerid	bigint
brokerid	bigint

Show 11 more columns

Materialized view
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimaccount
[redacted]@databricks.com

sk_accountid	bigint
accountid	bigint
sk_brokerid	bigint
sk_customerid	bigint
accountdesc	string
taxstatus	tinyint
status	string
iscurrent	boolean
batchid	int
effectivedate	date
enddate	date

Hide columns

Materialized view
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimtrade
[redacted]@databricks.com

tradeid	bigint
sk_brokerid	bigint

Show 19 more columns

Materialized view
tpcdi.[redacted]_tpcdi_dlt_pro_10.factcashbalances
[redacted]@databricks.com

sk_customerid	bigint
sk_accountid	bigint
sk_dateid	bigint
cash	double
batchid	int

Hide columns

Table name
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimaccountstg
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimcustomerstg
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimbroker
tpcdi.[redacted]_tpcdi_dlt_pro_10.dimtrade
tpcdi.[redacted]_tpcdi_dlt_pro_10.factcashbalances

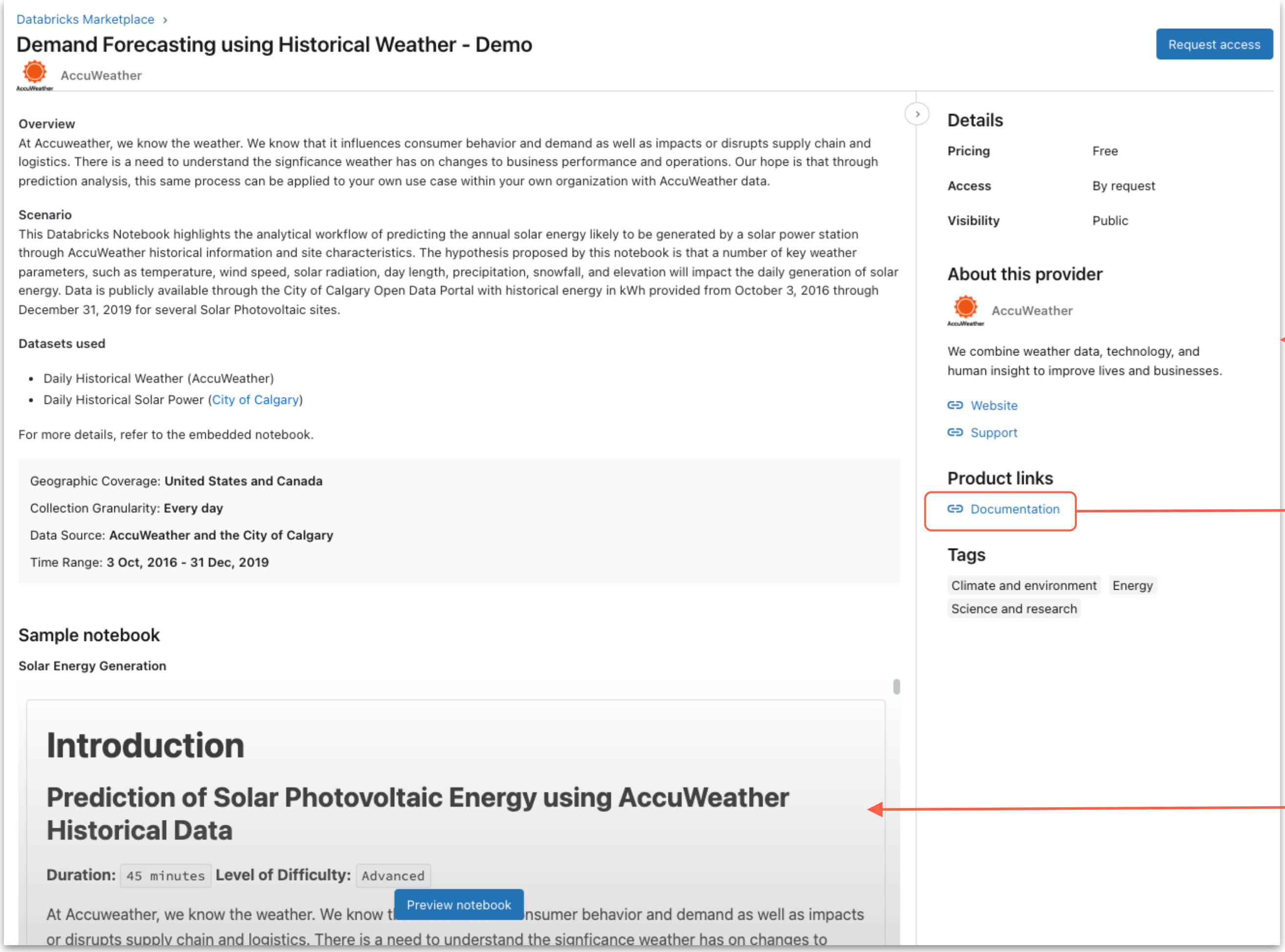
Lineage data is captured from the last 90 days

Associated collateral

* consider using notebooks in the platform

Data Contract with Databricks Marketplace

Share summaries and important details and link full data contract



Demand Forecasting using Historical Weather - Demo

Overview
At Accuweather, we know the weather. We know that it influences consumer behavior and demand as well as impacts or disrupts supply chain and logistics. There is a need to understand the signficance weather has on changes to business performance and operations. Our hope is that through prediction analysis, this same process can be applied to your own use case within your own organization with AccuWeather data.

Scenario
This Databricks Notebook highlights the analytical workflow of predicting the annual solar energy likely to be generated by a solar power station through AccuWeather historical information and site characteristics. The hypothesis proposed by this notebook is that a number of key weather parameters, such as temperature, wind speed, solar radiation, day length, precipitation, snowfall, and elevation will impact the daily generation of solar energy. Data is publicly available through the City of Calgary Open Data Portal with historical energy in kWh provided from October 3, 2016 through December 31, 2019 for several Solar Photovoltaic sites.

Datasets used

- Daily Historical Weather (AccuWeather)
- Daily Historical Solar Power (City of Calgary)

For more details, refer to the embedded notebook.

Geographic Coverage: **United States and Canada**

Collection Granularity: **Every day**

Data Source: **AccuWeather and the City of Calgary**

Time Range: **3 Oct, 2016 - 31 Dec, 2019**

Sample notebook

Solar Energy Generation

Introduction

Prediction of Solar Photovoltaic Energy using AccuWeather Historical Data

Duration: `45 minutes` **Level of Difficulty:** `Advanced`

At Accuweather, we know the weather. We know t[Preview notebook]nsumer behavior and demand as well as impacts or disrupts supply chain and logistics. There is a need to understand the signficance weather has on changes to

Details

Pricing	Free
Access	By request
Visibility	Public

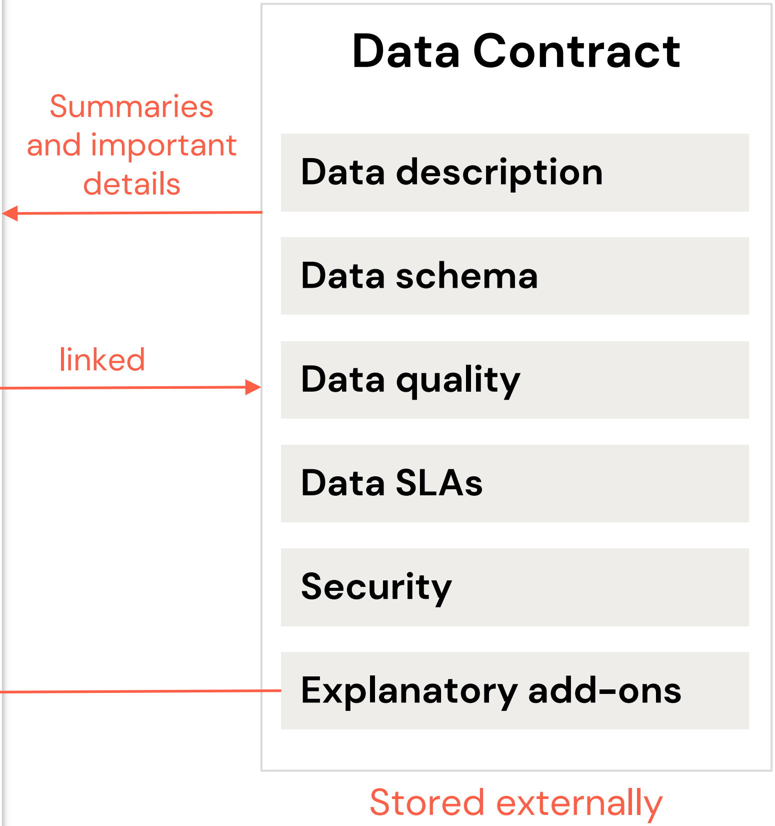
About this provider

AccuWeather

We combine weather data, technology, and human insight to improve lives and businesses.

Product links

Tags



Summaries and important details

linked

Data Contract

- Data description
- Data schema
- Data quality
- Data SLAs
- Security
- Explanatory add-ons

Stored externally

Demo

Demo

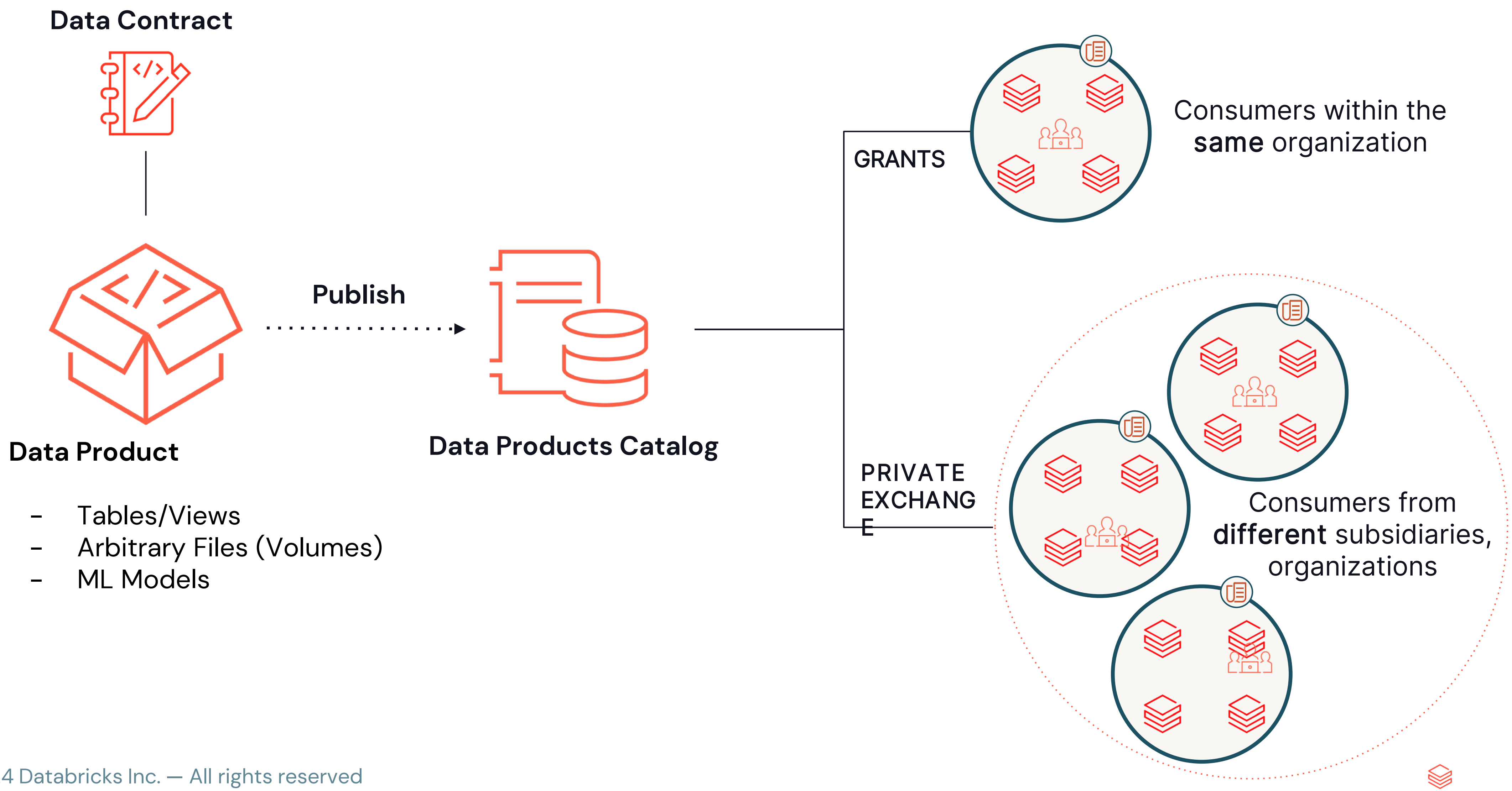
What you’ll see

➔ How consumers can discover data products

➔ How a data contract can look like in Databricks

➔ How to monitor the quality of the data products

Demo



Data Contract

Data Product

- Tables/Views
- Arbitrary Files (Volumes)
- ML Models

Publish

Data Products Catalog

GRANTS

Consumers within the **same** organization

PRIVATE EXCHANG
E

Consumers from **different** subsidiaries, organizations

Interoperability

Reality of Enterprise Data Platforms

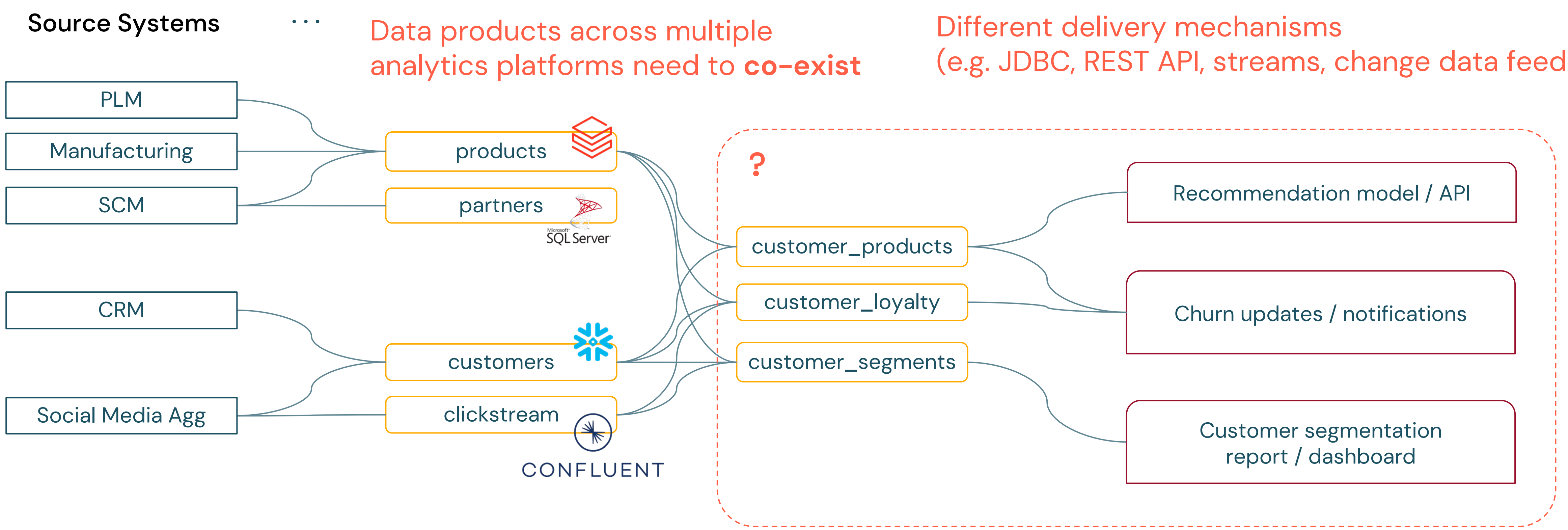
Data products come in many shapes and forms

Challenge #1: Ecosystem

Data products across multiple analytics platforms need to **co-exist**

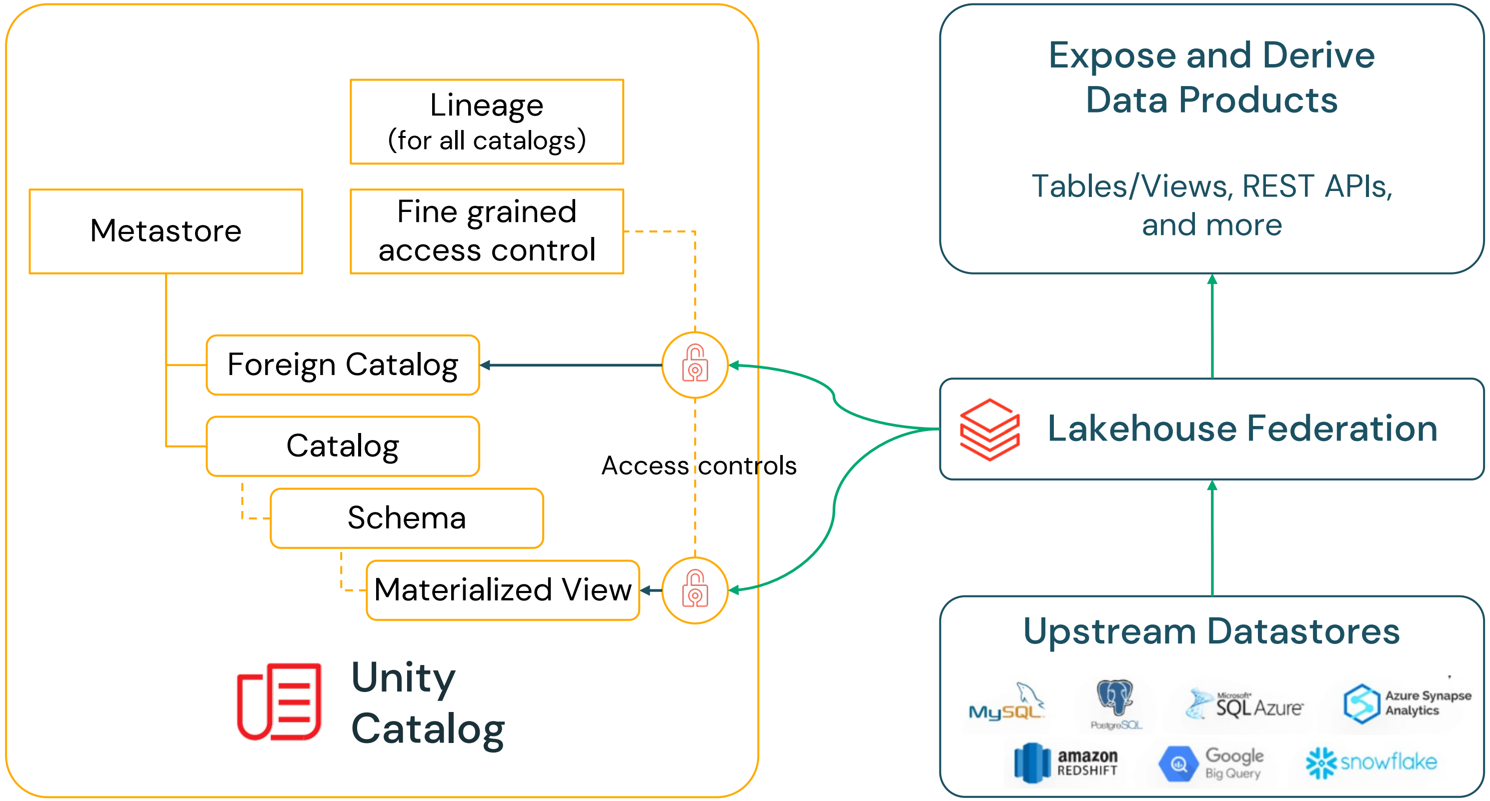
Challenge #2: Consumption Modes

Different delivery mechanisms (e.g. JDBC, REST API, streams, change data feed)



Build on top of data products from other datastores

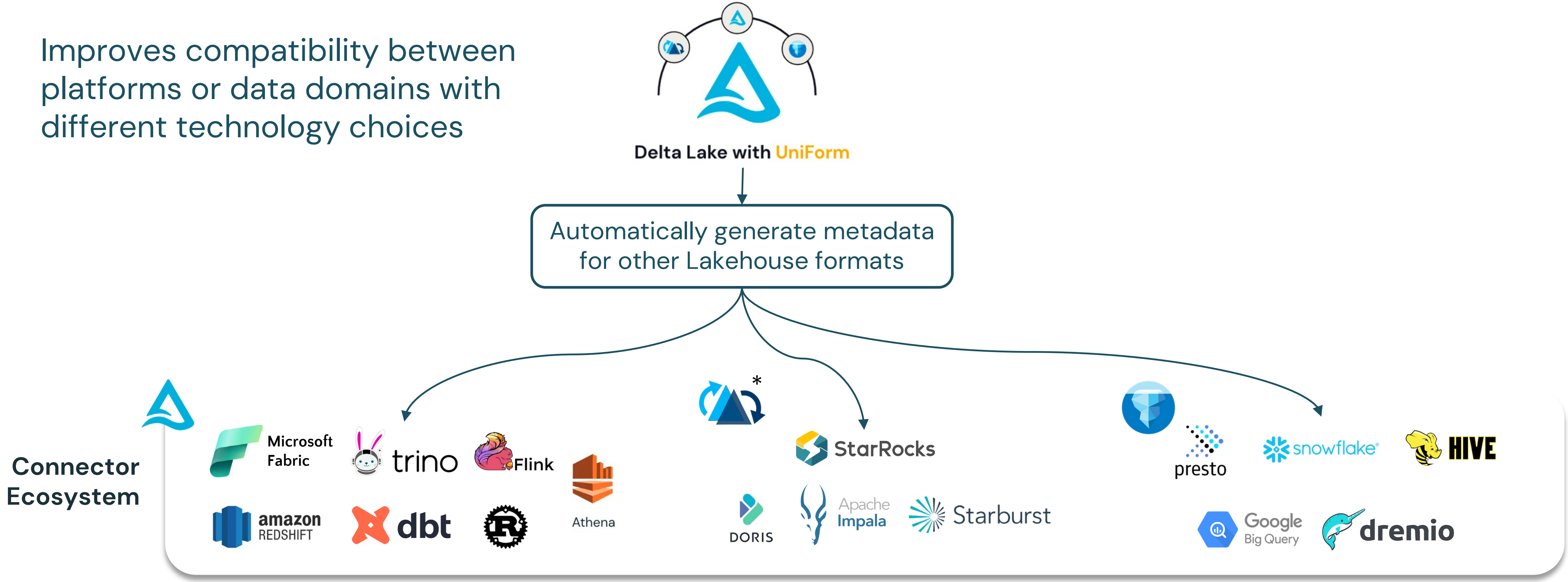
Bridge siloes with Lakehouse Federation



Interoperability with other Lakehouse formats

UniForm - Universal Format

Improves compatibility between platforms or data domains with different technology choices



* Hudi support in roadmap

Takeaways

Building high-quality data products with Databricks

Data Products should be:

1. Discoverable
    - Take advantage of Unity Catalog's AI-assisted documentation, search
2. Reliable and Transparent
    - Prevent quality issues with Delta Live Tables
    - Monitor with Lakehouse Monitoring
3. Well governed (access controls, PII, auditable)
    - Leverage Unity Catalog's row/column-level security + tagging
    - Activate **system tables** for auditability, lineage

Learn more at the summit!

Databricks
Events App

Tells us what you think

- We kindly request your valuable feedback on this session.
- Please take a moment to rate and share your thoughts about it.
- You can conveniently provide your feedback and rating through the **Mobile App**.

What to do next?

- Discover more related sessions in the mobile app!
- Visit the Demo Booth: Experience innovation firsthand!
- More Activities: Engage and connect further at the Databricks Zone!

Get trained and certified

- Visit the Learning Hub Experience at Moscone West, 2nd Floor!
- Take complimentary certification at the event; come by the Certified Lounge
- Visit our Databricks Learning website for more training, courses and workshops! databricks.com/learn

Thank You

databricks