TEADAL

Towards Serverless Data Exchange Within Federations

TU Wien: Boris Sedlak, Victor Casamayor Pujol, Praveen Kumar Donta, Schahram Dustdar
TU Berlin: Sebastian Werner, Karl Wolf, Frank Pallas, Stefan Tai
POLIMI: Matteo Falconi, Pierluigi Plebani

WWW.TEADAL.EU

26/06/2023

Problem Description - General

- Data regarded as the new “oil”
  - Data silo vs active data exchange

Free image under CC license from https://upload.wikimedia.org/wikipedia/commons/3/38/Oil_Barrel_graphic.png

Problem Description - General

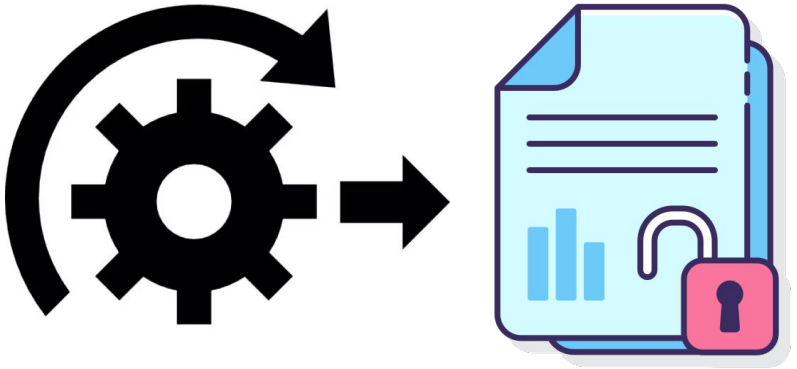
- Data regarded as the new “oil”
  - Data silo vs active data exchange
- Contains **confidential** or **private** information

Free image under CC license from https://freeiconshop.com/wp-content/uploads/edd/lock-outline-filled.png

Problem Description - General

- Data regarded as the new “oil”
  - Data silo vs active data exchange
- Contains **confidential** or **private** information

- Transformations introduce **data friction**

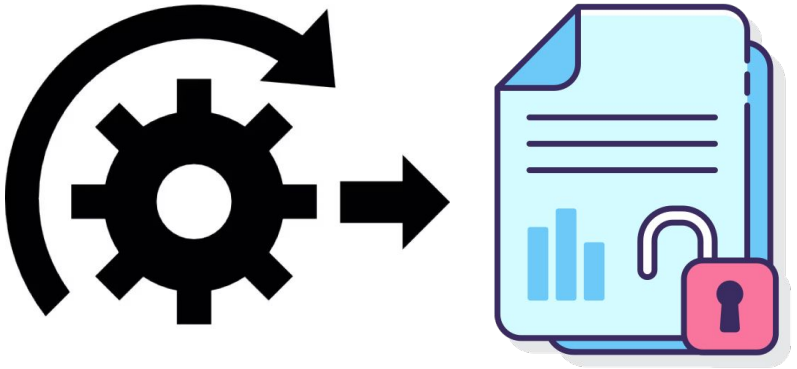


Free image under CC license from https://img.freepik.com/free-icon/development-progress_318-31445.jpg

Problem Description - General

- Data regarded as the new “oil”
  - Data silo vs active data exchange
- Contains **confidential** or **private** information

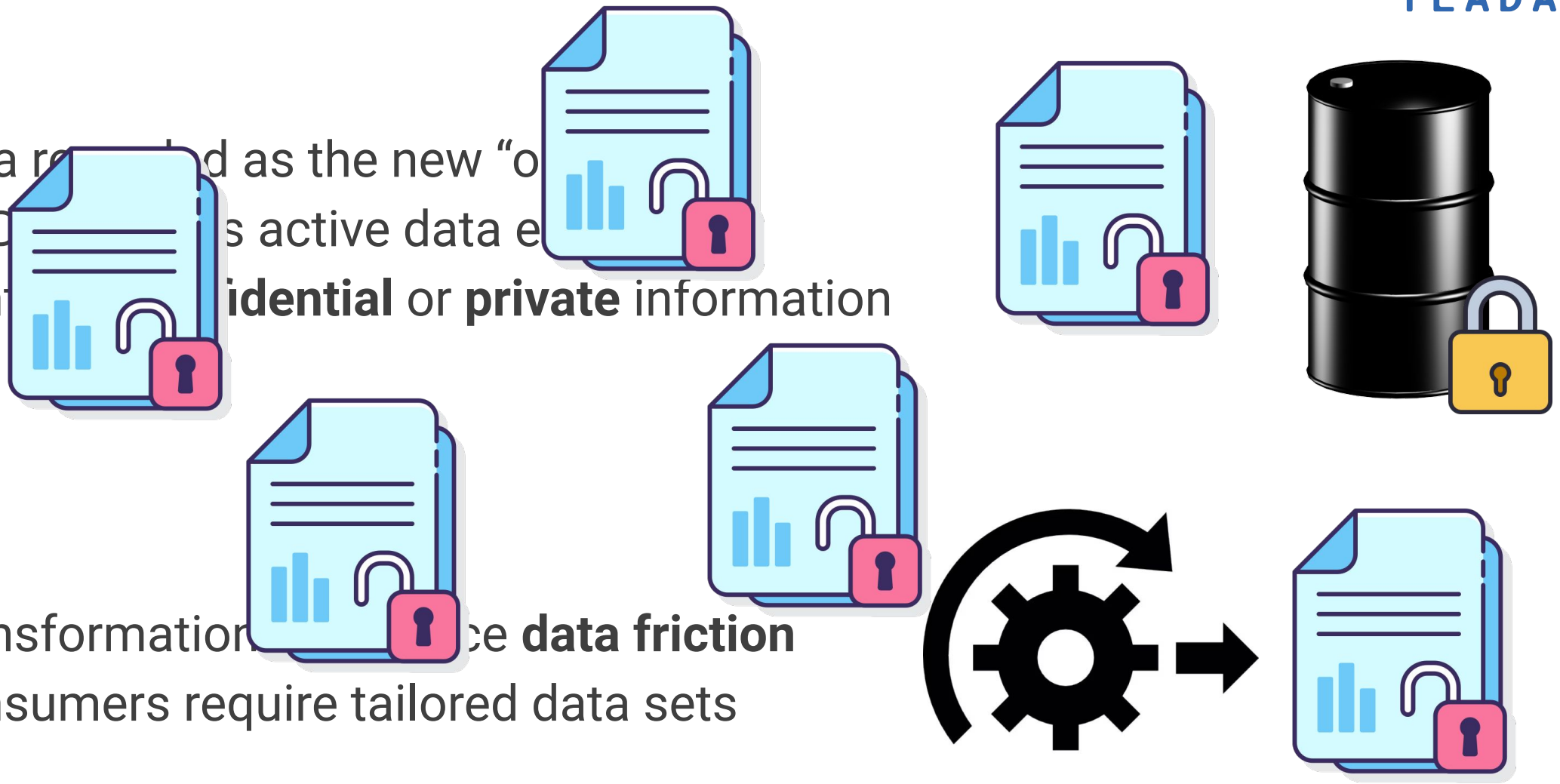
- Transformations introduce **data friction**
- Consumers require tailored data sets



Free image under CC license from https://img.freepik.com/free-icon/development-progress_318-31445.jpg

Problem Description - General

- Data re[illegible]d as the new "o[illegible]
  - D[illegible]s active data e[illegible]
- Cont[illegible]idential** or **private** information



- Transformation[illegible]ce **data friction**
- Consumers require tailored data sets

Free image under CC license from https://cdn-icons-png.flaticon.com/512/2037/2037155.png

Problem Description - Use Case

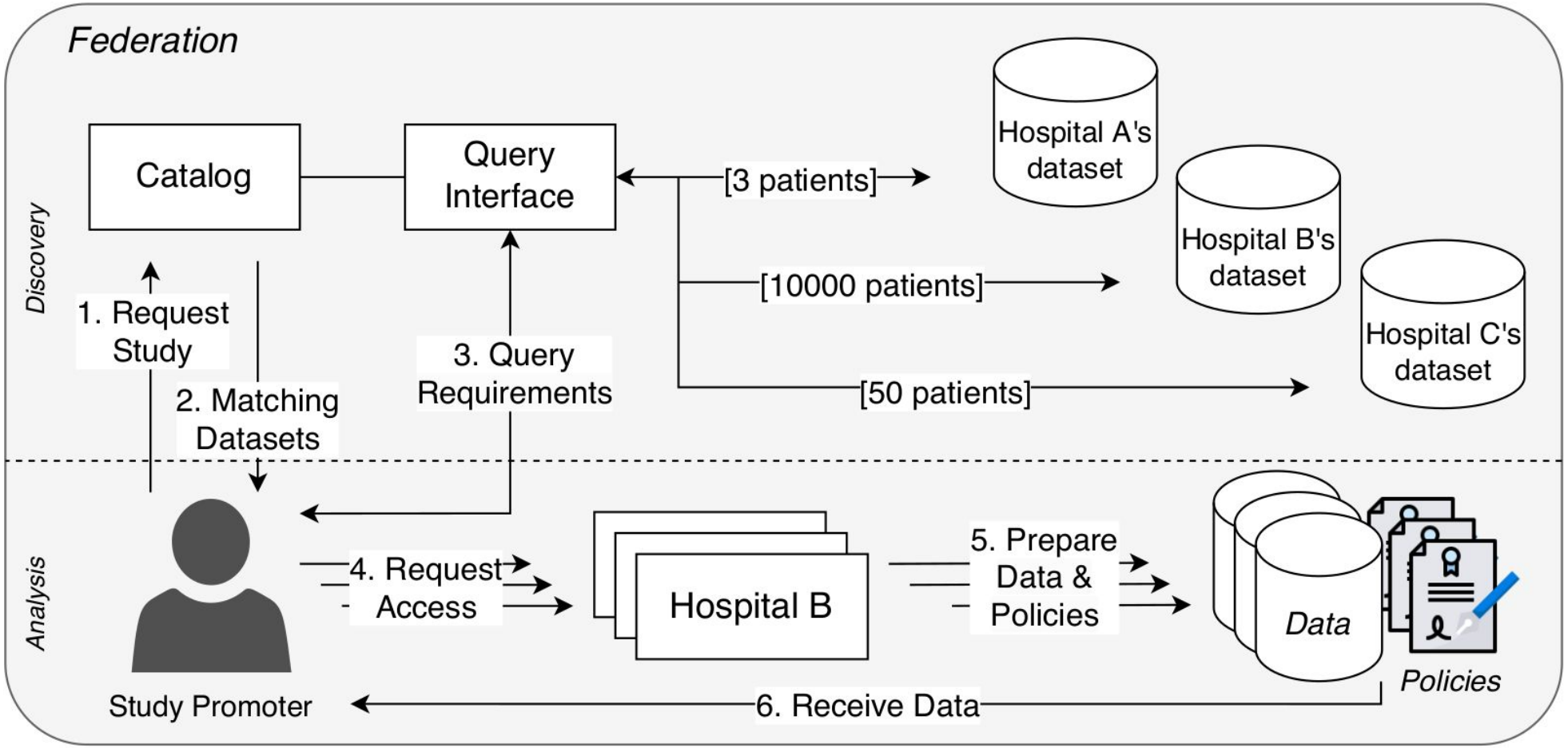
- Study promoter / hospital wants to conduct a **joint study**
- Cumbersome to retrieve data from multiple providers
  - General norms (e.g., **GDPR**), incompatible data formats, etc
  - **Manual negotiation** of usage agreements

Problem Description - Use Case

- Study promoter / hospital wants to conduct a **joint study**
- Cumbersome to retrieve data from multiple providers
  - General norms (e.g., **GDPR**), incompatible data formats, etc
  - **Manual negotiation** of usage agreements

- Lack a mechanism to *discover* data sets and *agree* on their provision

Envisioned Solution - Use Case



Envisioned Solution - General

- Automatic matching of requirements for owner and consumer
- Data transformed according to agreements
- Provisioning of storage / computing resources (e.g., ad hoc or premises)

- For **data providers**, alleviate the burden of data sharing
- For **data consumers**, ensure that data is served as desired

Contributions

- Federated Data Products

  Identify five **lifecycle phases** that data products pass through when sharing them within a federation

- Serverless Data Exchange

  Apply serverless principles for **processing** and **storage** of data

Federated Data Product

- Data product as self-served data set (**data mesh**)
- Domain experts <--> Platform providers
- Described with **usage policies** and shared

- Cross-enterprise sharing opens issues (e.g., **identity**, resources)
- Must be supported by a **underlying platform**
- But first, let’s focus on the federated data product!

Free image under CC license from https://media.istockphoto.com/id/1469124165/vector/cloud-database-black-glyph-icon.jpg

Federated Data Product - Lifecycle (1/2)

1. Data onboarding
   - Persist the data product according to storage policies
   - Domain experts supply general policies (e.g. privacy transformations)

2. Publishing
   - Register the data product in the federation-wide catalog
   - Support consumer-aware policies (e.g. # records)
   - Sync catalog information (incl. policies) with data product

3. Sharing
   - Agree how data is served to consumers
   - Include constraints (e.g. policies, transformations, time, # access)
   - Sign contract and provide to all parties

Federated Data Product - Lifecycle (2/2)

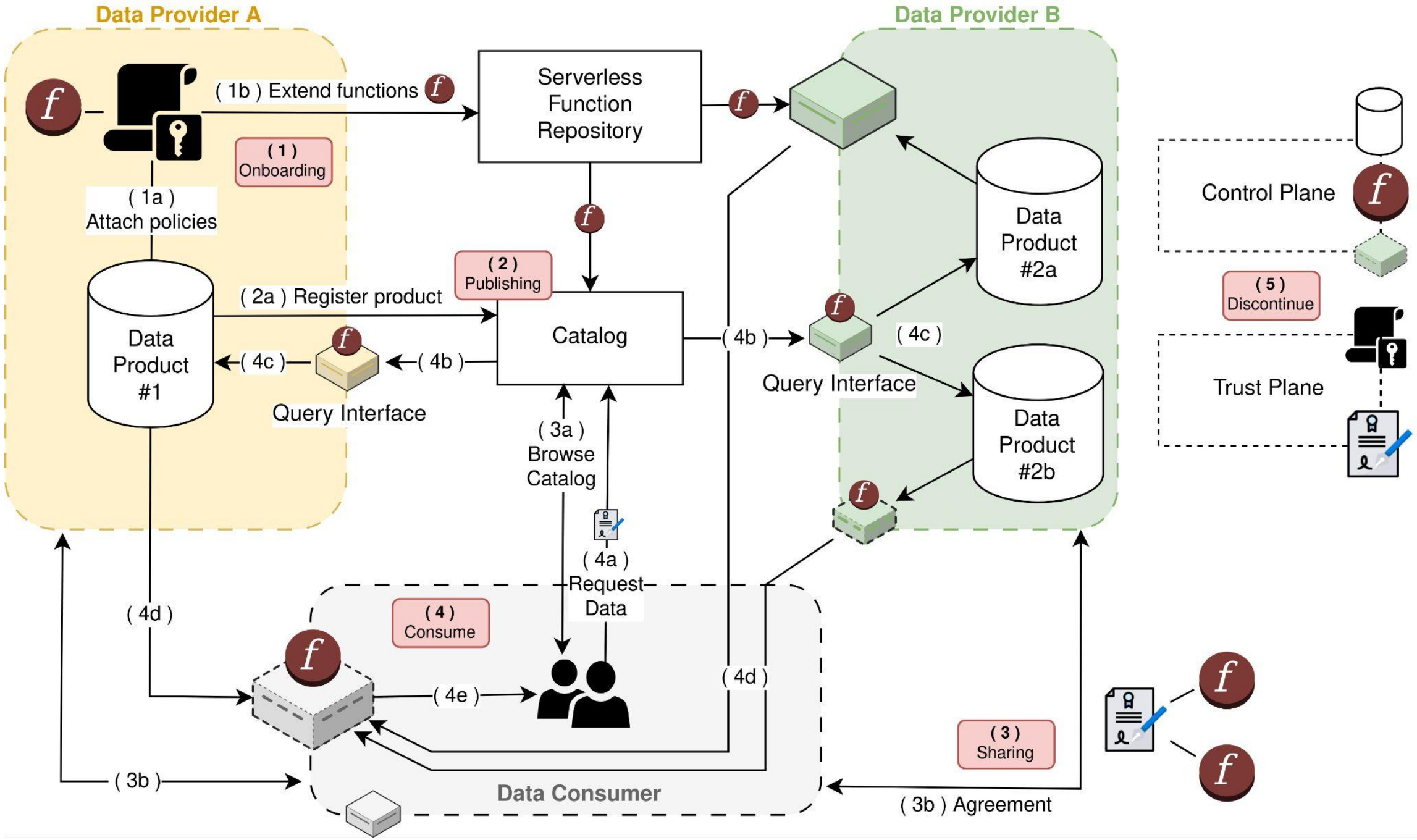
4. Consumption
   - Run compulsory operations (e.g. transformations)
   - Document all interactions with the dataset (audit)
   - Optimize processing by moving data and/or processing

5. Discontinue
   - Remove data product from catalog, (inform consumers)
   - Delete data product (and all copies) from all locations

Serverless Data Exchange

- Supports the exchange of federated data products

- **Serverless functions** – can be included as part of agreements, supplied by domain experts which can define multiple implementations
- **Serverless processing** – happens somewhere in the federations, based on serverless functions, can be optimized by moving processing
- **Serverless data** – is stored somewhere in the federation, it can be fragmented from a storage perspective, but is offered as one logical product to the consumer. Can be optimized by moving (copies of) data


Data Provider A
Data Provider B
(1b) Extend functions
Serverless Function Repository
(1) Onboarding
(1a) Attach policies
(2) Publishing
(2a) Register product
Catalog
Data Product #1
(4c)
(4b)
Query Interface
(3a) Browse Catalog
(4a) Request Data
Data Product #2a
Data Product #2b
(4b)
(4c)
Query Interface
Control Plane
(5) Discontinue
Trust Plane
(4d)
(4) Consume
(4e)
(4d)
(3) Sharing
(3b) Agreement
(3b)
Data Consumer

Possible Issues

- Very dependent on *Control plane* and *Trust plane*
  - CP – How to allocate **resources** (i.e., storage, processing, functions)?
  - TP – How to technically ensure **trust** between parties?
- So far very agnostic in terms of tools and technologies

Summary

- Data sharing is **impeded** by constraints and semantics
- Transformations introduce excessive **data friction**

- Data is exchanged as **federated data products**
  - Domain experts define policies (i.e. **serverless functions**) in upfront, or supply them in agreements

- Serverless data exchange supports this exchange
  - **Control plane** provides resources (i.e. processing, storage, functions)
  - **Trust plane** assures identities and that policies are respected
  - Data is **transformed** ad hoc according to policies attached to it (Though this can be optimized)

Summary

- Data sharing is **impeded** by semantics and constraints
- Transformations introduce excessive **data friction**

- Data is exchanged as **federated data products**
  - Domain experts define policies (i.e. **serverless functions**) in upfront, or supply them in agreements

- Serverless data exchange supports this exchange
  - **Control plane** provides resources (i.e. processing, storage, functions)
  - **Trust plane** assures identities and that policies are respected
  - Data is **transformed** ad hoc according to policies attached to it (Though this can be optimized)

TEADAL

TEADAL.EU

@TEADAL_eu

@TEADAL

TEADAL project is funded by the EU's Horizon Europe programme under Grant Agreement number 101070186