DEFI 4 – Comment favoriser l’exposition de la data (comment devenir une data driven company)

Idées clés

- Mesurer la valeur créé et l'usage
- Expliquer le why du data driven
- Un datalake est par nature transverse est utile si bcps de silos -> permet des use cases transverse jusque là impossible. Mais on ne met pas tout dans le lake seulement si on a un usage y compris le cas Michel de la compta
- Le lake n'est "jamais le source d'autres applis" sauf dans le cadre d'explo controllée, un aggregated data product. Les data sets exposes sont pour un usage connu

Les leviers - accélérateurs

- Un sponsor fort
- Catalog (owner, cycle de vie & lineage …)
- Langage commun (pas universel) au niveau domaine
- Data lineage
- Gouvernance fédérée et distribuée
- Team product IT & Biz
- Responsabilité (inclus les end points d'exposition) aux équipes produits
- Des plateformes techniques communes
- Documenter les patterns archi pour exposer la donnée (dans un lake ou pas).
- Systématiquement décrire la données (notamment pour le catalog)
- Avoir un data lab (explo/ inno) sourcé sur les lakes et controllés. Pas directement sur les sources
- Data market place est un canal (comme API/Topics)
- Design authority globale inter domaines

Les Freins - Pitfalls

- Rupture technologique app & data
- Diversité plateforme data = TCO ⇗
- Intéropérabilité tech
- Peu de skills (tech & data model) dans les équipes
- Pourquoi exposer? Quelle valeur?
- On explore pas dans l'absolu
- On evite de tout déverser dans un lake (RGPD)
- Se sourcer directement sur le modèle interne (couplage)
- Ownership & responsabilité

ROLES - MISSIONS

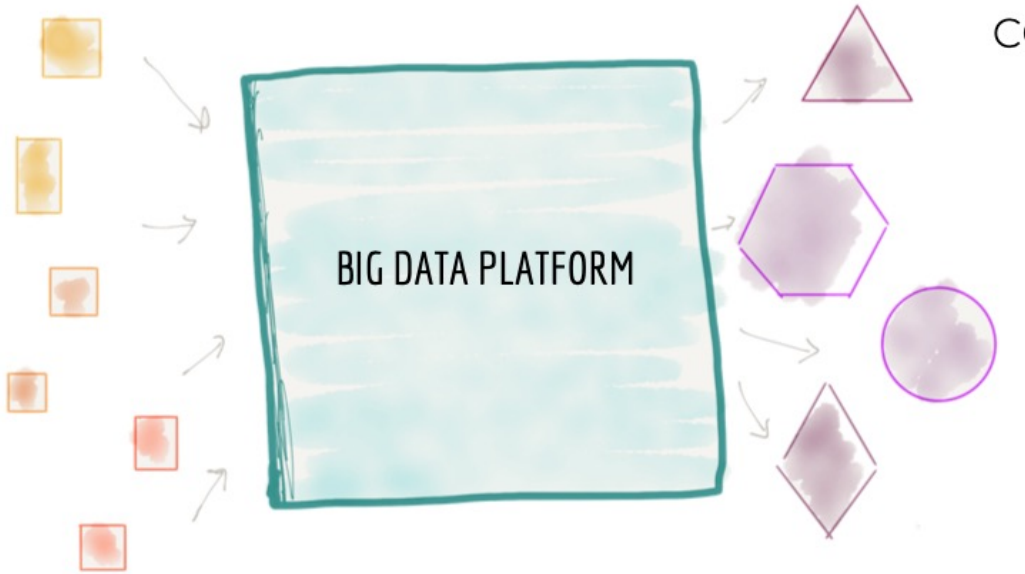
- Chief data office (attention gouvernance centralisée)
- Data / Domain owner
- Architecte mais à priori pas besoin de spécialiser pour la data
- Data engineer (maturité différente par rapport au soft eng)
- Data steward & custodian
- Urgent de remettre à plat les rôles data car il y en a trop
- Et si on appliquer une approche produit, les Product Owner / Product mgr remplaceraient une bonne partie de role (data owner déjà)

PRACTICES - TOOLS

- Catalogue de donnée
- Data plateform
- Continuous archi fonctionne quasiment tel quel si on applique l'approche produit sur la data

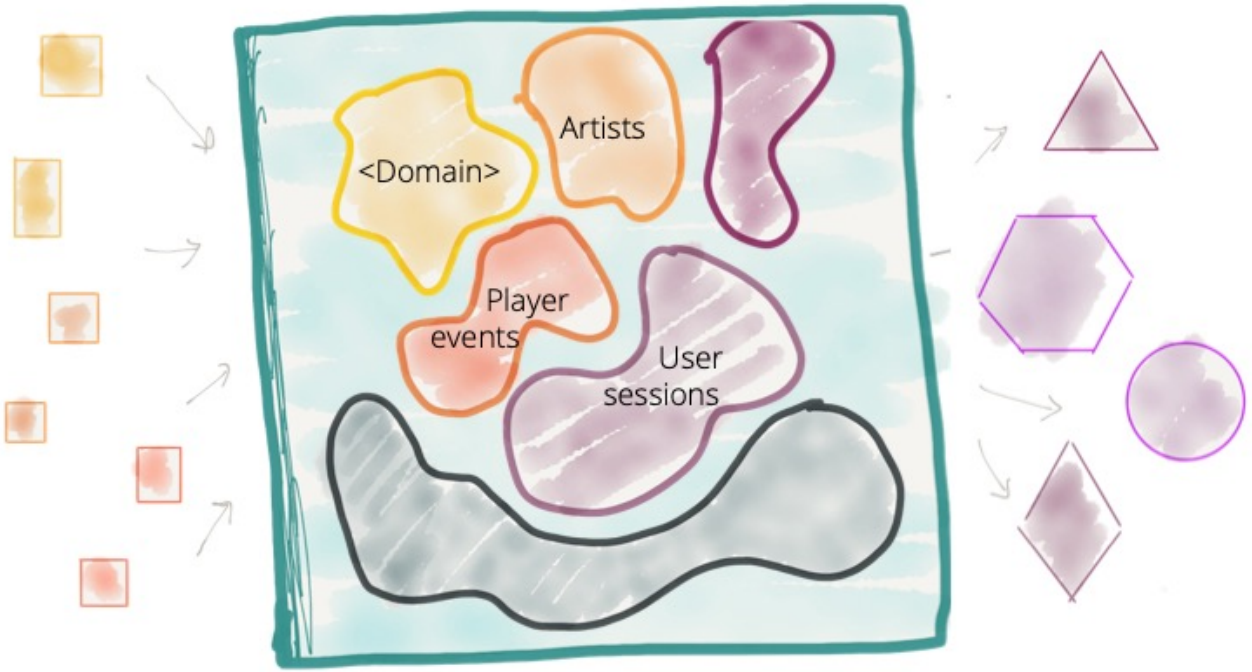
SOURCES

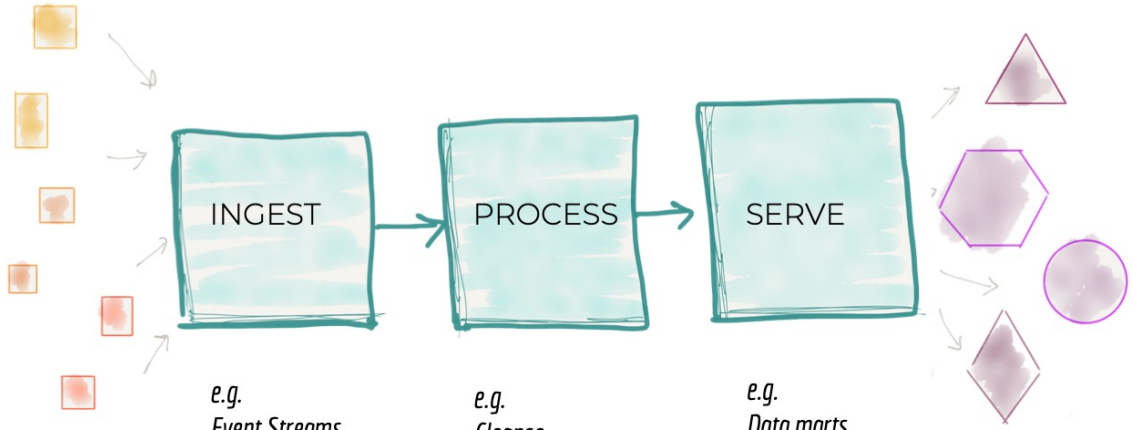
TO INGEST

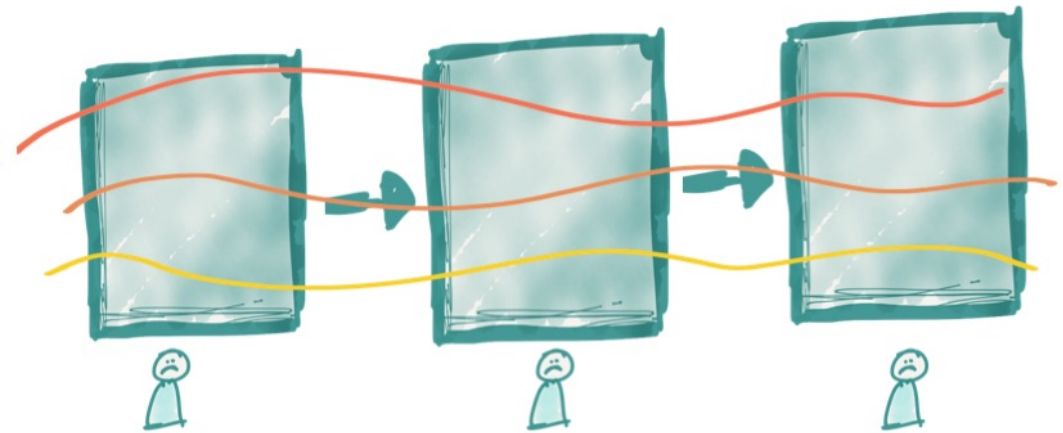


CONSUMERS

TO SERVE

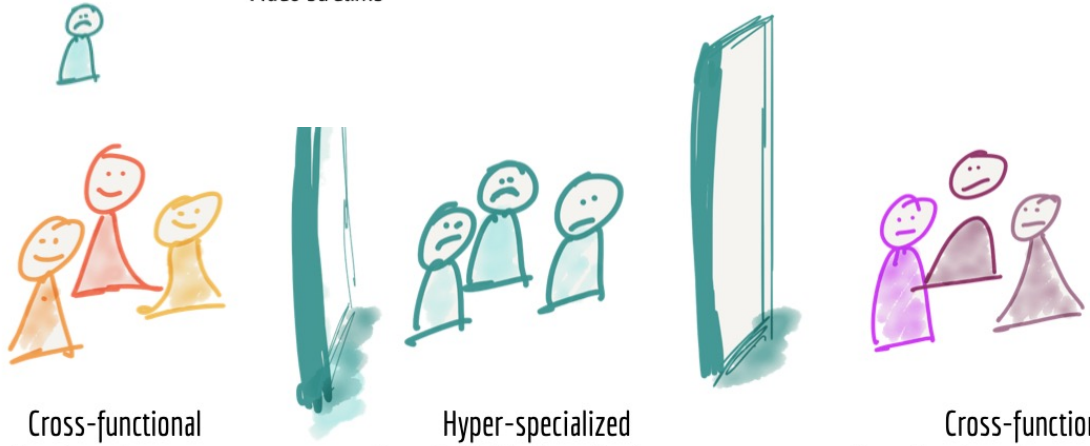






Features
&
Capabilities

e.g.
Podcasts,
Video streams



Cross-functional
Domain oriented source teams

Hyper-specialized
Data & ML Platform Engineers

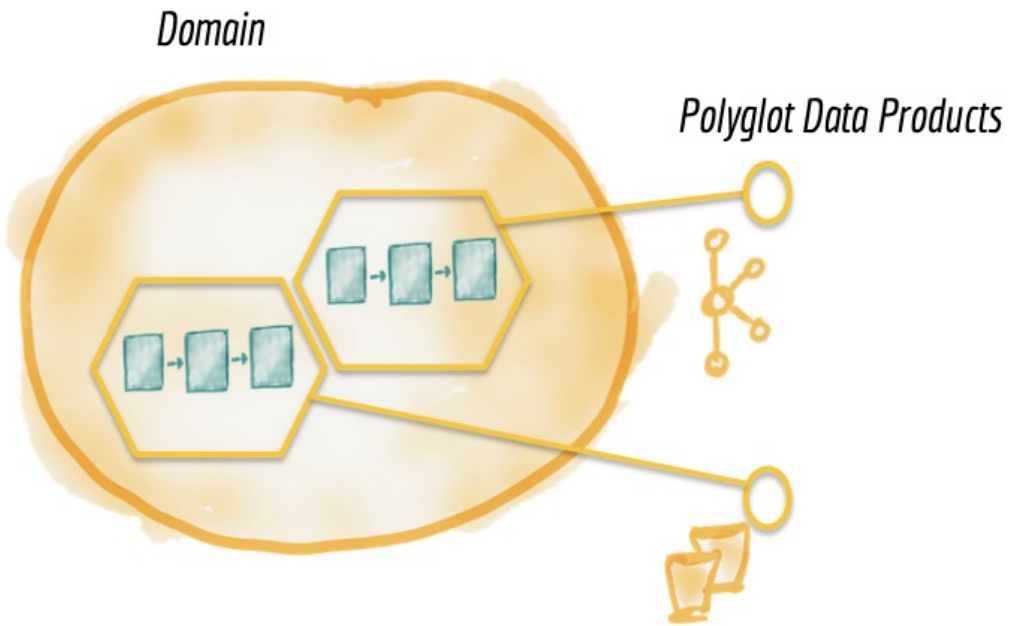
Cross-functional
Domain oriented consumer teams

Data

Distributed Domain Driven Architecture

Self-Serve Platform Design

Product Thinking



DISCOVERABLE

ADDRESSABLE

TRUSTWORTHY
(DEFINED & MONITORED SLOs)

SELF-DESCRIBING

INTER OPERABLE
(GOVERENED BY OPEN STANDARDS)

SECURE

What is a data product ?

Data product

Published data set that can be accessed by others

Enables consumers to perform cross-domain data analysis

Preferred format are files, tables or views but also all kind of API

Data quality expectations / SLA are described and monitored

Documented in a Data catalog

Metadata describing the product and the business terms

Access management process compliant with security rules

Addressable in a Data platform

Through a query engine with examples and best practices

With usage and access monitoring

Source or native data products

Aligned with the structure, lifecycle and semantic of the data source.

Should be developed by a team that is as close as possible to where the data is originally generated.

Aggregated data products

Aligned with a particular business concept aggregated from multiple domains

Providing the data as much as possible liked with the different dimentions

Fit for purpose data products

Data products that are transformed and modeled to fit a set of specific use cases

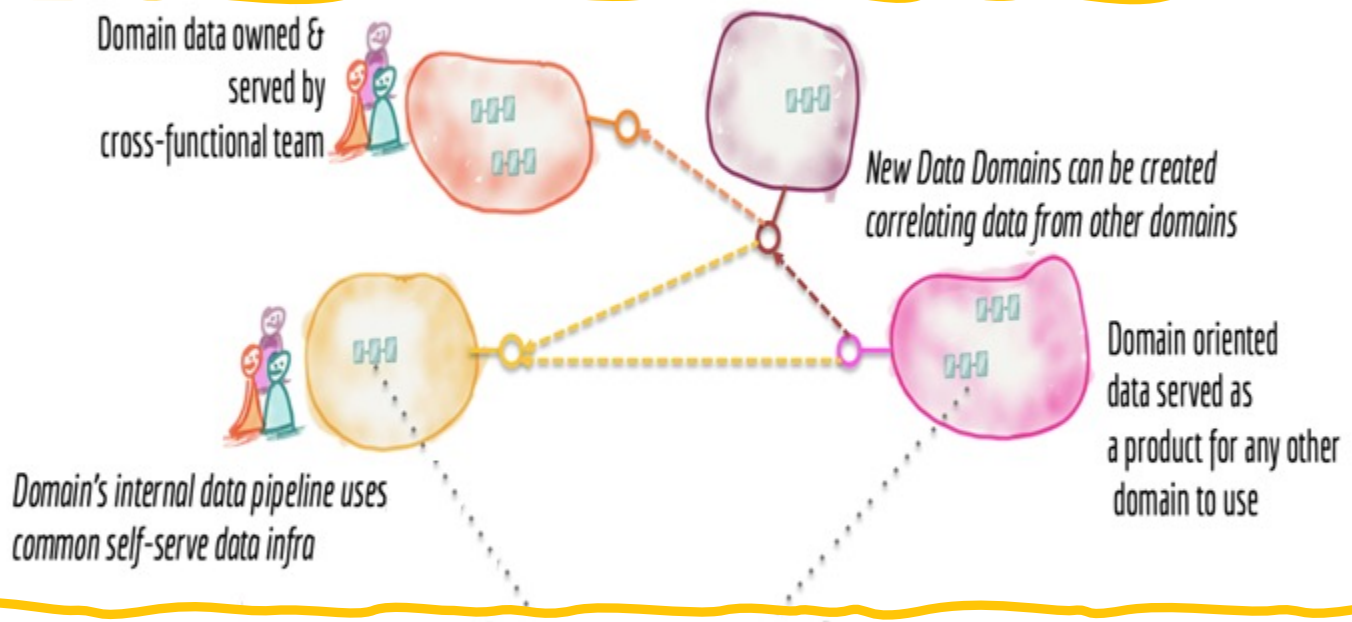
Often the result of a machine learning or analytics computation

Domain team Is responsible of the **operations**

We measure the data product success through **data usage**, **number of data consumer** and their **satisfaction** !

Global Governance & Open Standards
(enable interoperability)

Distributed data domains

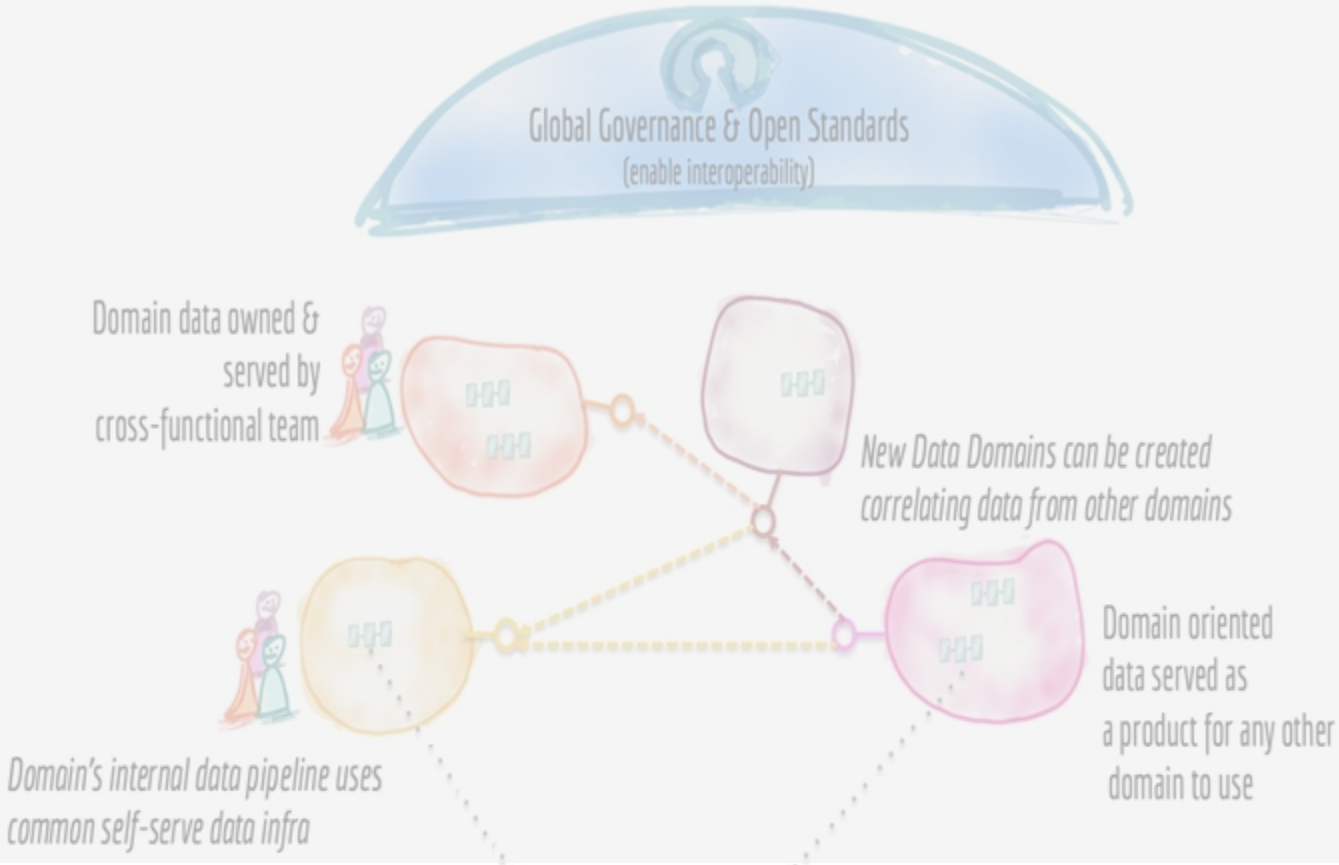


Data infra engineers

Data Infra as a Platform
(storage, pipeline, catalogue, Access control, etc.)

In 2021, we have identified the data domains and measured their exposition

In 2022, we need to reinforce the link between the domains and their data products and measure their usage



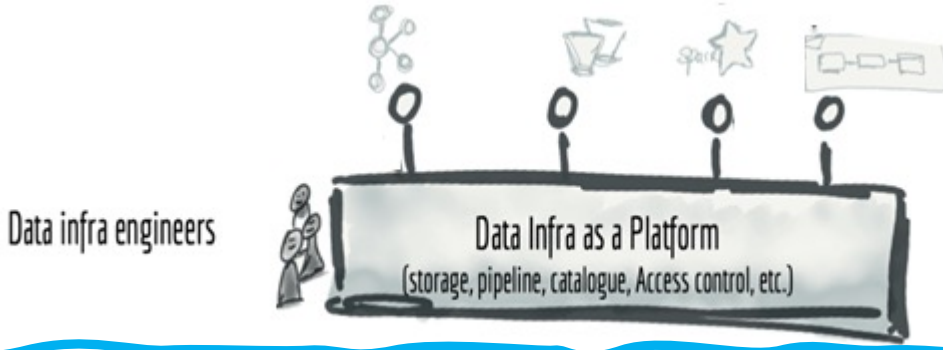
Data Platforms

Several platforms sharing their knowledge and practices

Focus on a common « experience » and « capabilities » provided to domains

One Data World :
to ensure data platform capabilities governance

- One common backlog and roadmap
- Regular demo and sharing between teams



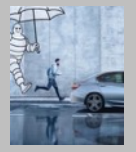
Data catalog

D1
data platform

Protecting core secrets

Factory
Dataware

Deployed in the **plant**



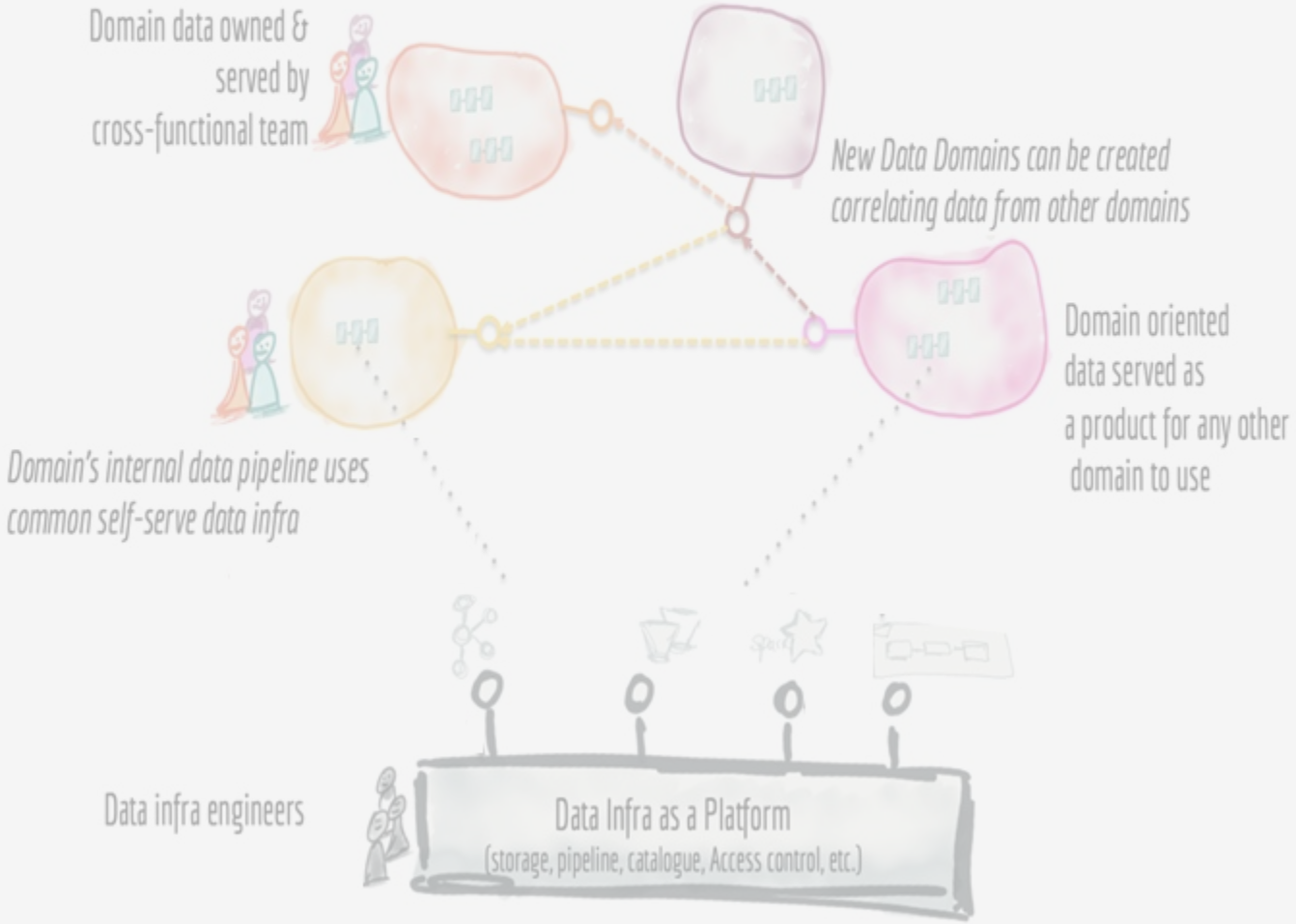
Corporate
data lake

Transverse data platform

Ultim
data lake

Vehicle and Tire usage

Federated governance



Data catalog deployment crystalized data governance Maturation

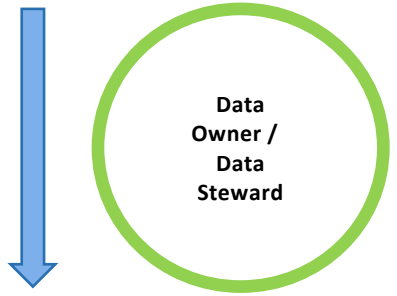
Federated Governance
Support a distributed orgnization

Collibra

Data catalog as a one stop shop for data

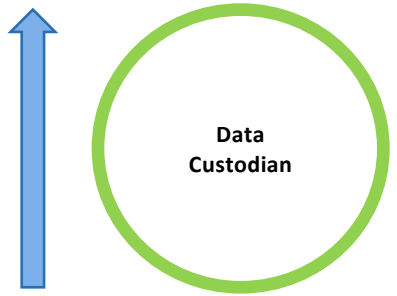
Business Data Model

1. Communities and data domains (like Stock, Product etc)
2. Business terms, data concepts, data attributes and policies



Operational Data Model

3. Logical : Logical view to abstract physical model, and highlight interconnection between products
4. Physical : List of physical files and tables



Beyond the technical deployment of a tool, Data Catalog crystallize Data Governance maturation and structure the governance

- Data Owners define and align themselves on their perimeter. They take ownership of the Enterprise Data Model and a common language by defining Domains and Business Terms
- Data custodians describes all the physical tables and datasets that are automatically scanned by Data Catalog.

The 2 approaches come together and make it possible to bridge the gap between the physical model and the logical model.