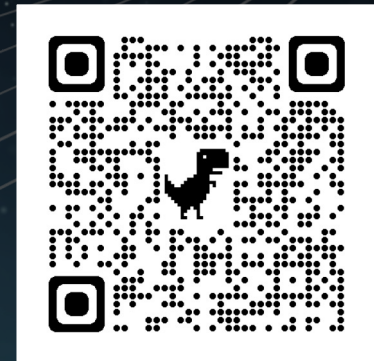


DATABRICKS 20
DATA JOURNEY 23 LATAM
EDICIÓN CASOS DE USO

Grupo de
Databricks en
Español
América Latina



***Bienvenidos, comenzamos
en unos minutos***

La sesión está siendo grabada

Databricks Lakehouse Journey

Diferentes **sesiones y demostraciones de casos** de uso por parte de expertos en Databricks y Solution Architects sobre los siguientes temas:

- 🚢 **Abril 13- Databricks Lakehouse: Data Mesh**
- 🚢 **Mayo 18 - Databricks Lakehouse: Desmitificando la migración a la nube con Databricks**



Dr Guillermo Schiava D'Albano
Databricks Sr Partner Technical Manager



Carlos Morillo
Databricks Solutions Architect



Data Mesh

Building a Data Mesh based on
Databricks Lakehouse

En Español

Dr Guillermo G Schiava D'Albano,
Sr Partner Technical Manager
(UK, IR, Benelux and Nordics)
©2022 Databricks Inc. — All rights reserved



Confidential

Do not share this information
without explicit permission from
Databricks.

TL,DR

Zhamak Dehghani's four key principles can be met within a **Lakehouse architecture**,

- Unity Catalog (greatly aids Federated Computational Governance).
- We can extend securely beyond organisation boundaries with Delta Sharing.

Data Mesh Blog Posts

Databricks Official Blog

- [Databricks Lakehouse and Data Mesh, Part 1](#)
 - Introductory concepts
- [Building a Data Mesh Based on the Databricks Lakehouse, Part 2](#)
 - Operating models
 - Implementation patterns
 - Data sharing ecosystem

Data & AI Summit 2022 talks

Data Mesh is an important topic for our customers

Customers

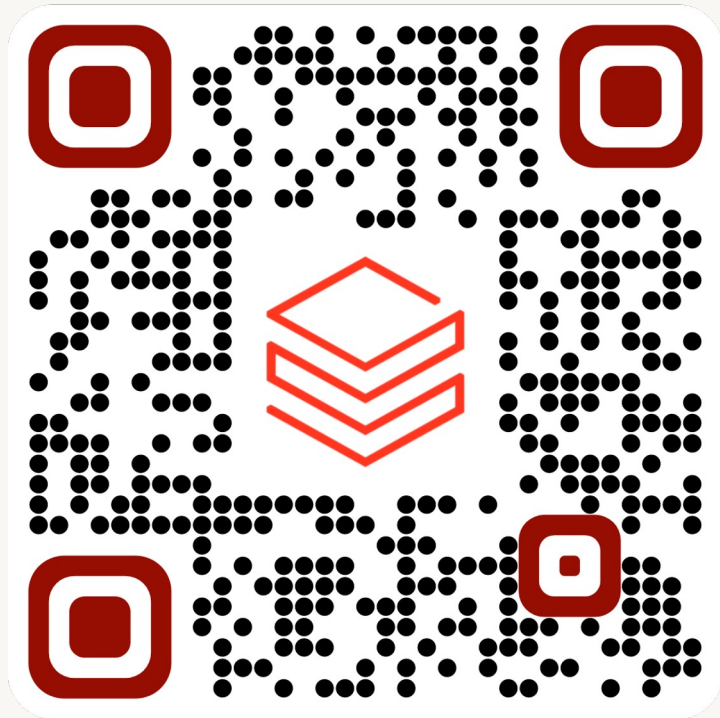
- McKesson, Accenture: [Data Mesh Implementation Patterns](#)
- HSBC: [Accidentally Building a Petabyte-Scale Cybersecurity Data Mesh in Azure With Delta Lake at HSBC](#)
- Zalando: [Data Lakehouse and Data Mesh—Two Sides of the Same Coin](#)
- GSK: [Data Warehousing on the Lakehouse](#)




























Databricks:

- [Automate Your Delta Lake or Practical Insights on Building Distributed Data Mesh](#)
- [Meshing about with Databricks](#)

All DAIS 2022 videos can be found at <https://www.youtube.com/c/Databricks/videos>

50+ Solution Accelerators



 Adverse Drug Event Detection	 Customer Lifetime Value	 Subscriber Churn Prediction	 Customer Retention
 Recommendation Engines	 Granular Demand Forecasting	 Genome-Wide Association Studies	 Safety Stock Analysis
 Customer Segmentation	 Alternative Data for Investing	 ESG Investing	 Predictive Maintenance (IoT)
 Risk / Value at Risk Calculation	 Quality of Service Video Streaming Analytics	 Ad Effectiveness With Forecasting and Attribution	 Threat Detection With DNS Analytics
 Disease Prediction	 Digital Pathology Image Analysis	 Anomaly Detection With Geo Clustering	 Reputation Risk
 Transaction Enrichment	 Rules-Based AI for Financial Fraud Prevention	 Product Match With Machine Learning	 Building Forward-Looking Intelligence With External Data
 Modernizing Investment Data Platforms	 Toxicity Detection in Gaming	 Cyber Analytics With Splunk Connector	 Multi-Touch Advertising Attribution



This presentation

This is a high level presentation:

- On Data Mesh, I will not be covering other flavours of Mesh-like implementations
- You are not alone lets us help you on implementing Data Mesh. All customers have a pair of Databricks AE-SA*
- About Databricks
- Introduction to Data Mesh Paradigm (high level with some low level tech)
- Semi Technical Introduction to the Lakehouse
- Implementing Data Mesh
- Q&A

* Amount of help would depend on the particular customer importance for you and the local team



Databricks

The Data + AI Company

Inventor and pioneer of the **data lakehouse**

Gartner recognized leader in both

- Database Management Systems
- Data Science and Machine Learning Platforms

Creator of highly successful OSS data projects: Delta Lake, Apache Spark, and MLflow

Raised over \$3B in investment

5900+ employees across the globe

Global adoption

Over 7000 customers, from F500 to unicorns



Advertisement

SHARE

INDUSTRY NEWS

Databricks Says It Has Surpassed \$1 Billion in Annualized Revenue

CEO says demand for AI-ready data is running high as companies brace for recession

By [Angus Loten](#)

Aug. 5, 2022 6:00 am ET | WSJ PRO



Total employee count

Based on LinkedIn data.

5,915
total employees

▲ 14%
6m growth

▲ 50%
1y growth

▲ 179%
2y growth



⌚ Median employee tenure · 1.3 years



bricks



Global adoption

Over 7000 customers



DELTA LAKE

REGENERON



S



If you place a bet in our career bet for Open Technologies

Recognized as a leader in two Gartner MQs

Data Science and ML Platforms



Cloud Database Management Systems



Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

©2021 Databricks Inc. — All rights reserved



Introduction to the Data Mesh paradigm

What is Data Mesh?

- *The shortest summary: **Treat data as a product**, not a by-product.* ←
- *By driving data product thinking and applying domain driven design to data, you can unlock significant value from your data.*
- ***Data needs to be owned by those who know it best.*** ←

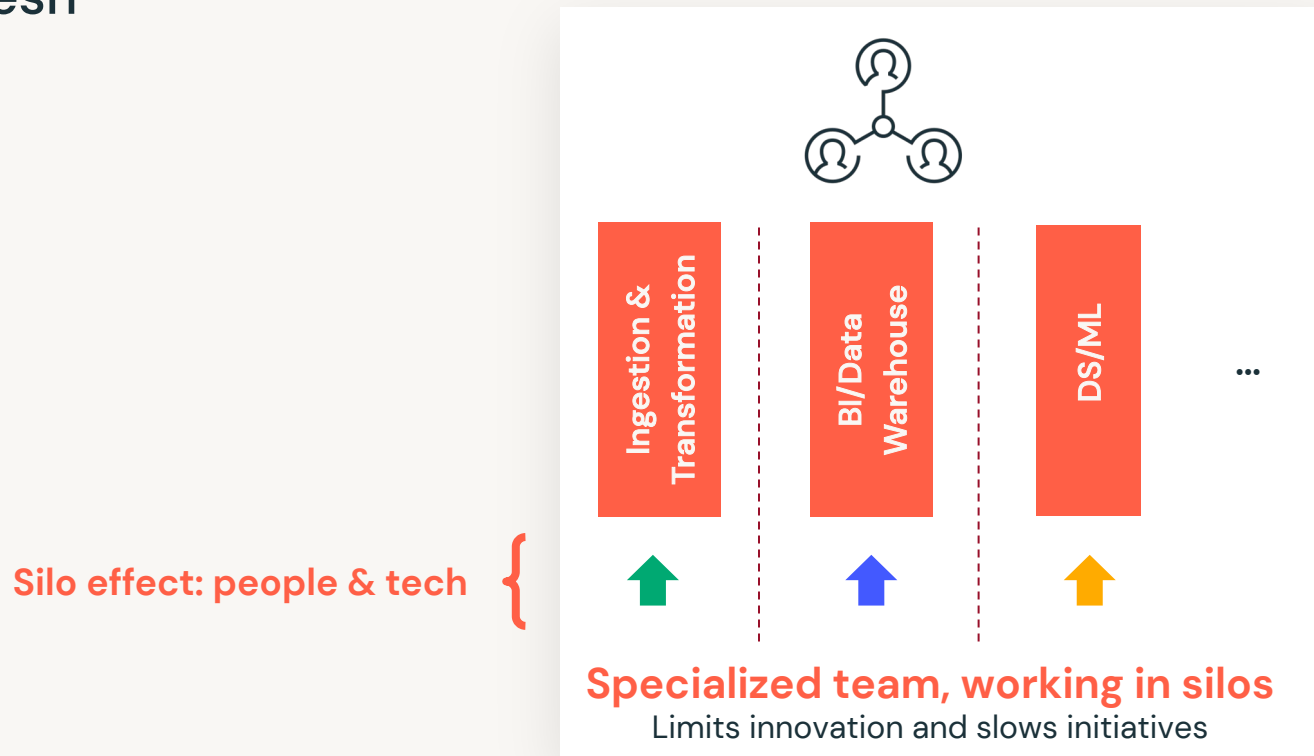
[Data Mesh Learning](#)
(aligned with Zhamak Dehghani)

[...for a more comprehensive introduction](#)

Data Mesh isn't a purchasable product

You can't buy an organizational process

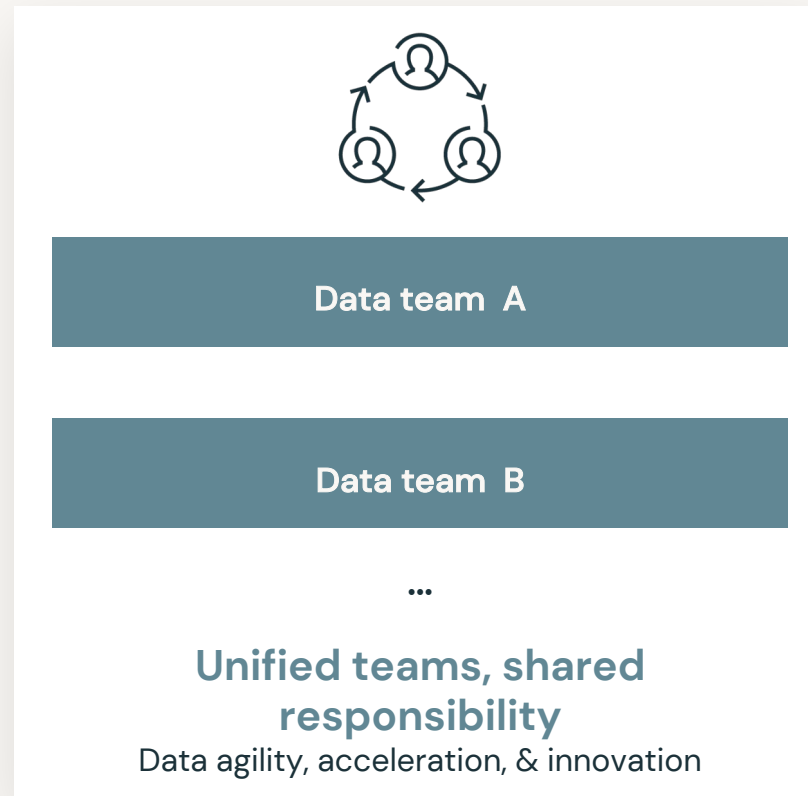
However, the right platform capabilities will ease (or not) the implementation of a data mesh



Data Mesh isn't a purchasable product

You can't buy an organizational process

However, the right platform capabilities will ease the implementation of a data mesh



Data Mesh isn't a purchasable product

You can't buy an organizational process

However, the right platform capabilities will ease the implementation of a data mesh

- Databricks Lakehouse has the **technical enablers** for teams to produce and consume data in a decentralized but governed way
 - Lakehouse is a polyglot technology that also works outside a Data Mesh concept ←
 - Lakehouse applies at all scales (startups to large orgs) ←
- Databricks **lowers the barrier to entry**, reducing reliance on central (monolithic) teams

* See also:

- [Common Misconceptions – \[Insert Vendor\] Offers a Data Mesh](#)
- [Common Data Mesh Misconceptions](#)

Databricks Lakehouse and Data Mesh

Complementary, not competing, paradigms

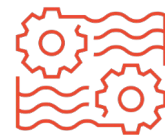
Databricks Lakehouse

Technological ecosystem to improve collaboration, quality, interoperability and productivity across Data & AI workloads

Data Mesh

Architectural and organizational paradigm to ensure value from data

Data Mesh
paradigm



Databricks
Lakehouse

Databricks capabilities for a Data Mesh

Addressing the 4 key pillars* with Databricks Lakehouse

#1 Domain ownership

Distributed architecture where domain teams, the **data producers**, can take responsibility for their data and its outcomes

- Open and flexible architecture enables [workspace/catalog](#) per domain
- Distributed ownership of data assets and [pipelines](#)

#2 Data as a product

Applying **product thinking** to analytical data, including providing quality data to **data consumers** beyond the source domain

- Open standards and formats for [FAIR](#) data
- ACID guarantees, versions, and audits with [Delta Lake](#)
- Fresh, high-quality data with [Delta Live Tables](#)

#3 Self-service infrastructure platform

Domain-agnostic approach to building, executing, and maintaining interoperable data products through common tools

- Unified platform serving all analytics workloads
- Managed orchestration with [Databricks Workflows](#)
- Auto-scaling/tuning, serverless
- Infrastructure as Code ([Terraform](#))

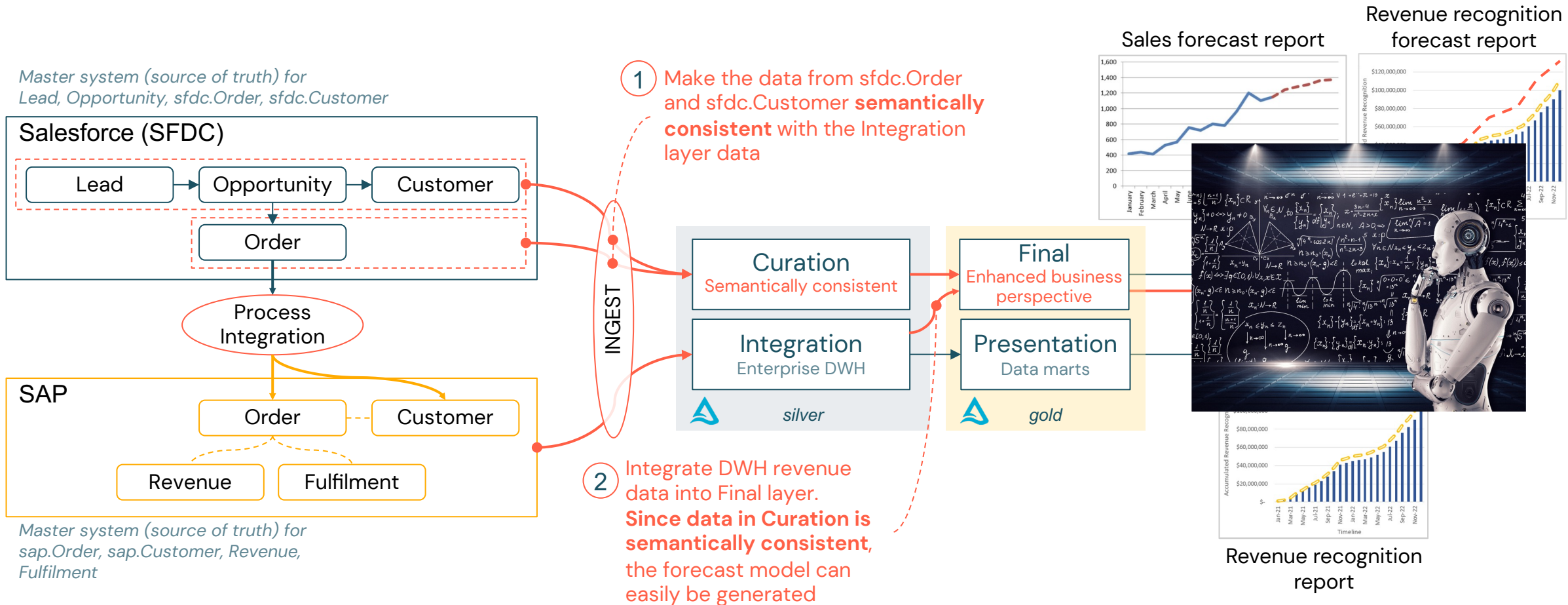
#4 Federated computational governance

Creating a data ecosystem that adheres to organisational rules and industry regulations through standardisation

- Discovery, access and lineage with [Unity Catalog](#)
- Global policy templates for access to data and [compute resources](#)

#2 Data as a product

Semantically consistent data allows integration of both worlds



Main Reasons for adopting the Data Mesh paradigm

Autonomy and accountability

- Avoid bottlenecks with central, monolithic platforms / processes
- Empower domains to be self-sufficient while respecting overarching governance rules

Improve data quality and usability

- Entrust teams that know the data and the domain best
- Domain-relevant quality and usability by design, not as an afterthought
- Data producers should delight their consumers and be rewarded

Accelerate (cross-domain) collaboration and productivity

- Simplify sharing and access to data across teams
- Encourage open standards, interoperability, and [FAIR data principles](#)

Main Reasons for adopting the Data Mesh paradigm

Autonomy and accountability

- Avoid bottlenecks with central, monolithic platforms / processes
- Empower domains to be self-sufficient while respecting overarching governance rules

Improve data quality and usability

Best practices stills applies

Accelerate (cross-domain) collaboration and productivity

- Simplify sharing and access to data across teams
- Encourage open standards, interoperability, and [FAIR data principles](#)

Data Mesh: Not a out of 'jail' card for best practices

Technical Debt (testing, clean code) affect Data Quality



Data Mesh (or any other paradigm) Scrum/Agile, etc applicable here

- Unit Testing (Databricks-IDE, Databricks-notebooks: here, here)
- Clean code, refactoring
- MLOps (MLFlow, Feature Store)
- CI/CD (Databricks Workflows or any other tool)

Raise your hand if you are Databricks Certified (at least associate DE/DS/SQL)

AI needs good Data -> Data needs best practices

Is High Quality Software Worth the Cost?

29 May 2019



Martin Fowler

Your job help here →

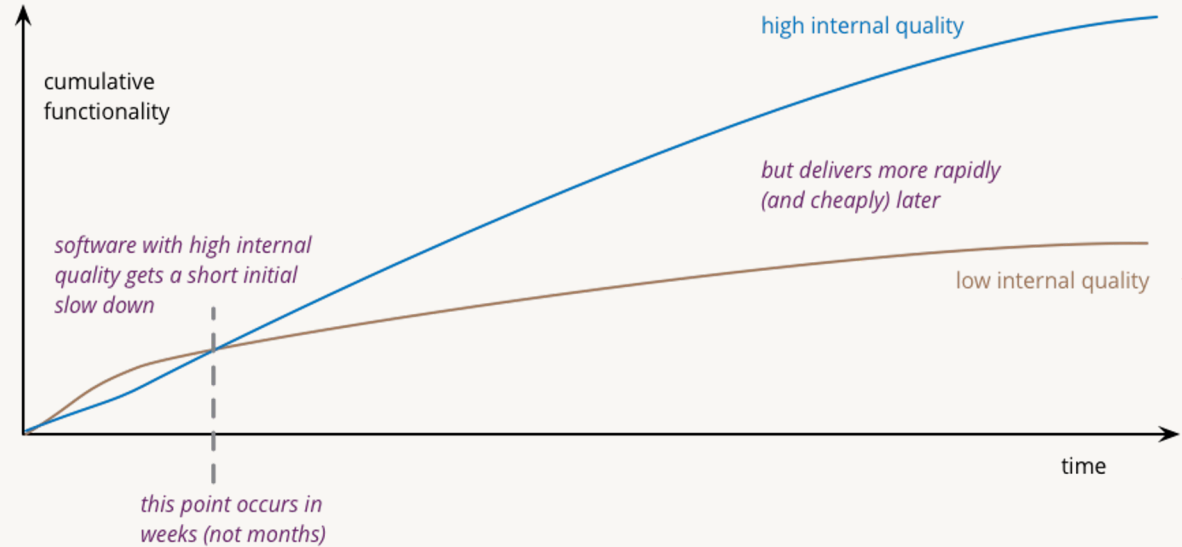
CONTENTS

We are used to a trade-off between quality and cost
Software quality means many things
At first glance, internal quality does not matter to customer
Internal quality makes it easier to enhance software
Customers do care that new features come quickly
Visualizing the impact of internal quality
Even the best teams create cruft
High quality software is cheaper to produce

SIDEBARS

Dora studies on elite teams

- PROGRAMMING STYLE
- PRODUCTIVITY
- PROJECT PLANNING
- TECHNICAL DEBT



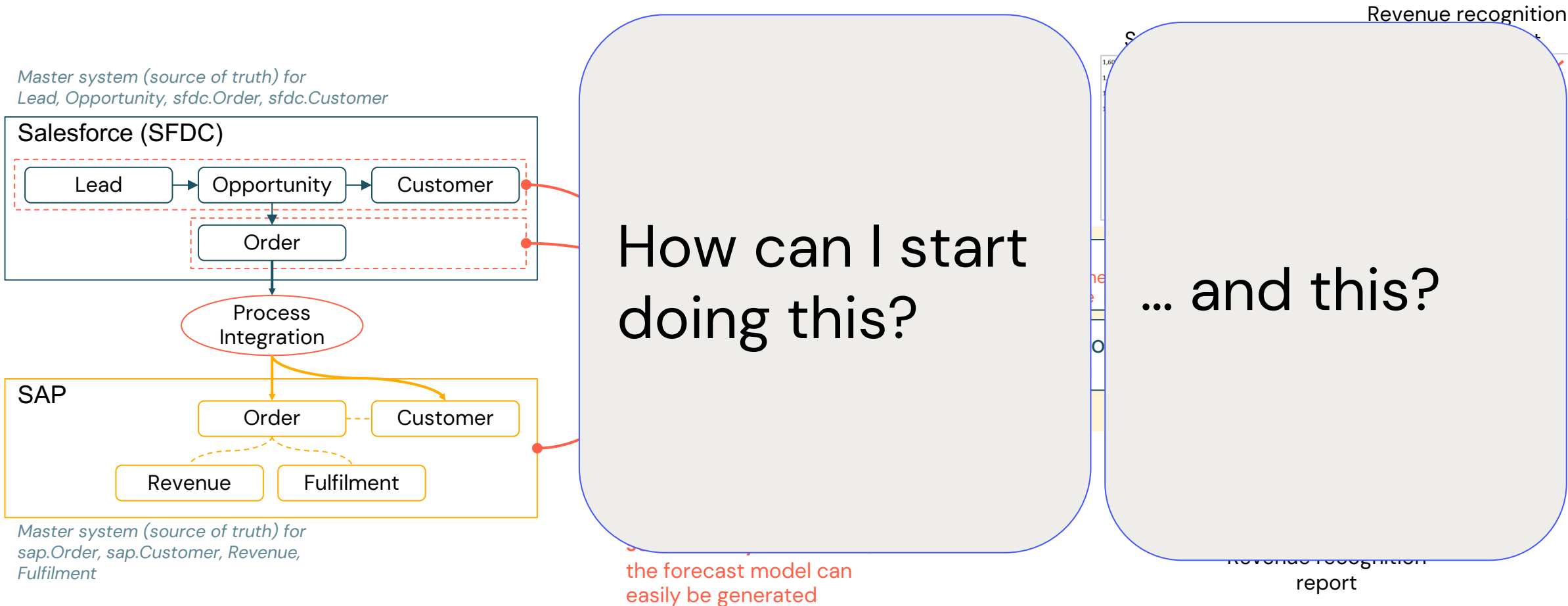
Data Mesh will not help you go fast if there is not High Internal Quality Software/Architecture

Your CoE for Databricks & Databricks Champions we can help you!

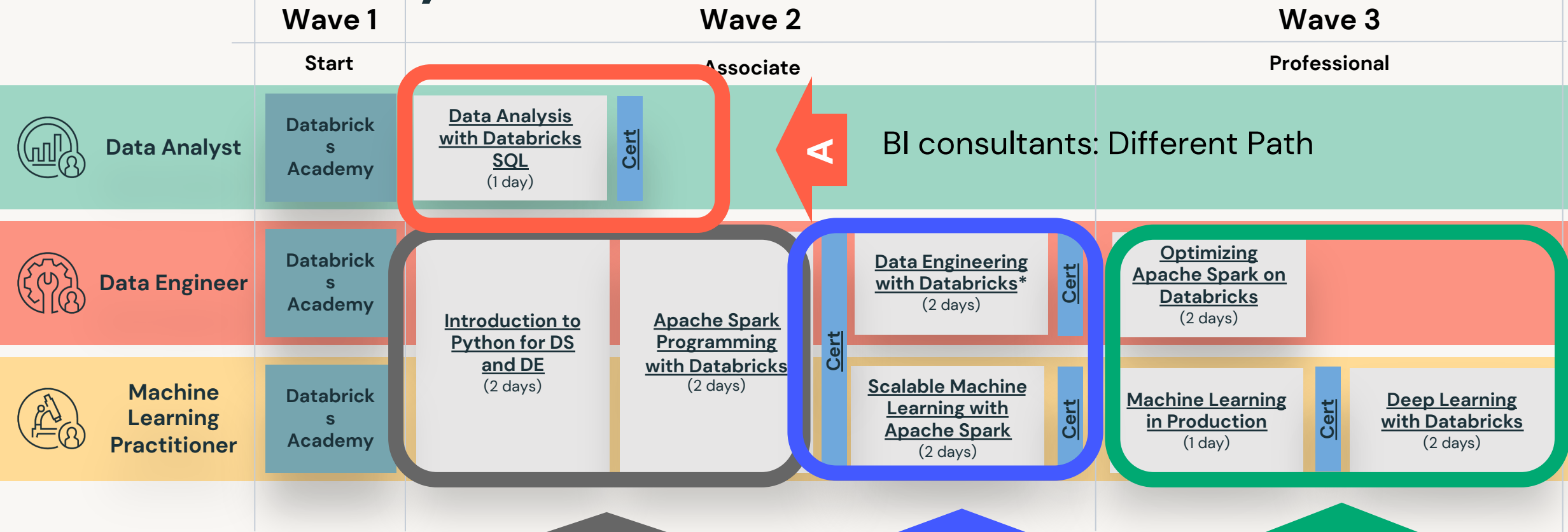


#2 Data as a product

Semantically consistent data allows integration of both worlds



TL,DR: Everyone DS/DE(one of DEDS-2), BI(A), every for DEDS-2 one DEDS-3



New to big data (do both)



All DE/DS consultants with at least one of these

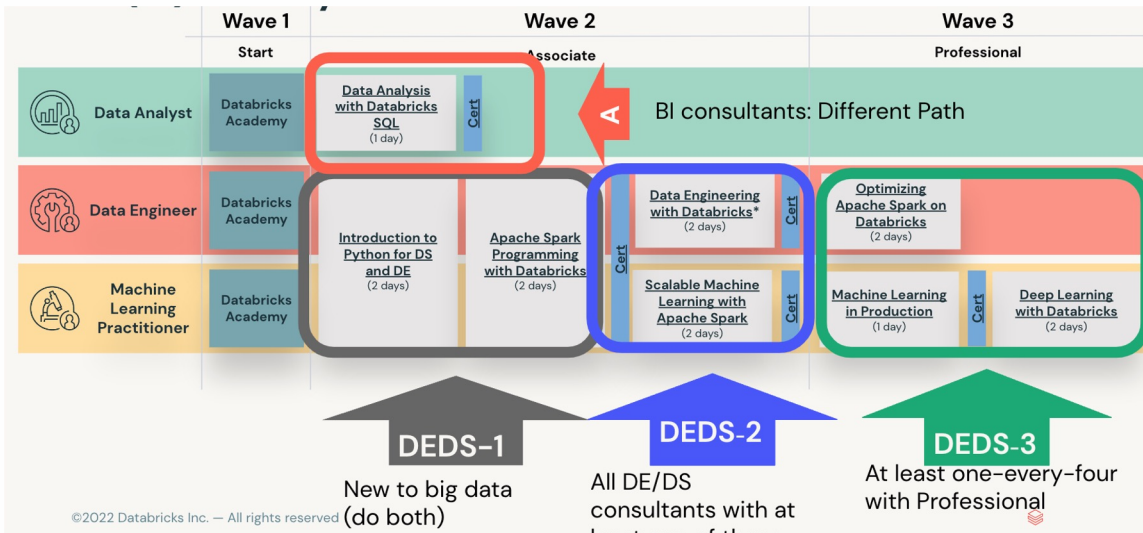


At least one-every-four with Professional

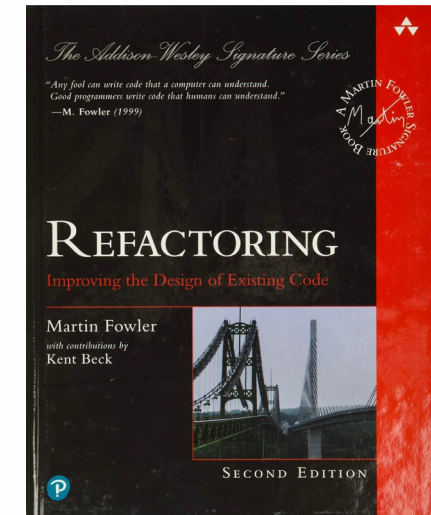
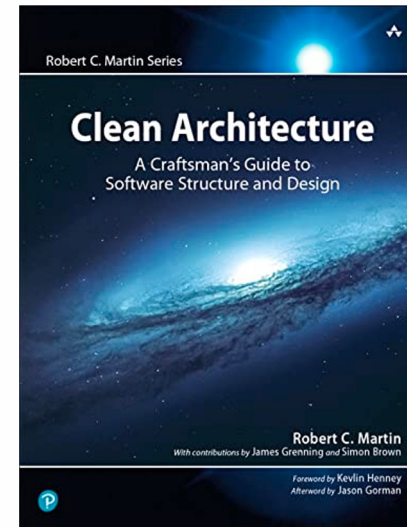
To work in Big Data you need two things (Total Quality, Jidoka)

Understanding how actually big Data works

Follow best practices In industry



- You need to read :-p
- Do peer programming etc
- Not enough time to speak about this



More links [here](#)

Qualifying for a Data Mesh

Executive sponsorship*

- Data Mesh addresses

Business:
C&SIs are really good
at it

strategy and roadmap

Data-driven business units

-

Business:
C&SIs are really good
at it

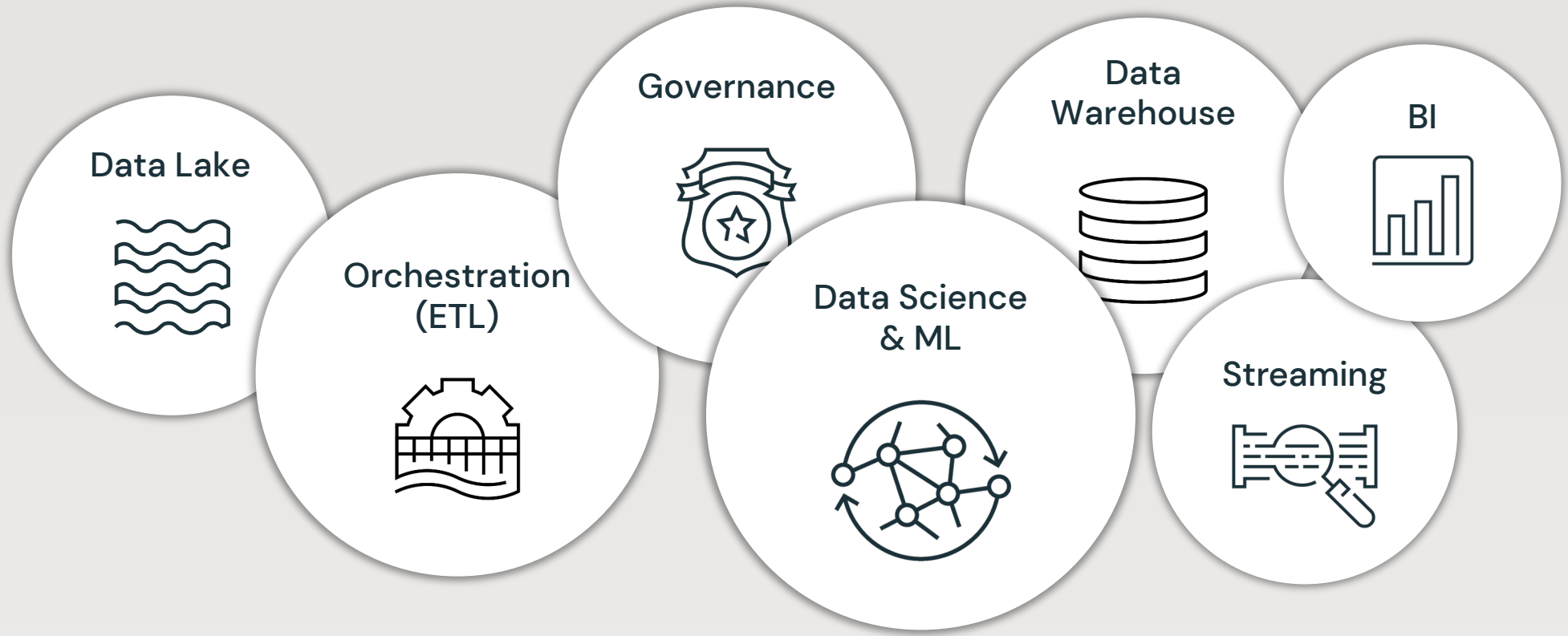
Organization size

- The organisation needs to be “large enough”.
- Enough DS/DE/BI in different business units so that the split into Data Domains actually makes sense



**We covered the basics
on Data Mesh... what
about Databricks?...**

... Semi Technical Introduction to the Lakehouse



Today, you stitch together too many platforms

It's all unnecessarily expensive and complex

Data
Lake

Data
Warehouse

Orchestration

Business
Intelligence

Data Science
& ML

Streaming

Governance

**Data silos drive high
operational costs**

**Inconsistent policies
reduce trust in the data**

**Disparate tools slow down
cross-team productivity**



A data lakehouse takes a different approach

One platform to support multiple personas



BI & Data
Warehousing



Data
Engineering



Data
Streaming



Data
Science & ML

One security and governance model for
all data access across the organization

One platform to store and manage all structured,
semi-structured, and unstructured data



Cloud Data Lake
All Raw Data
(Logs, Texts, Audio, Video, Images)





iPhone



Laptop



Watch



Music



Phone



Camera



GPS

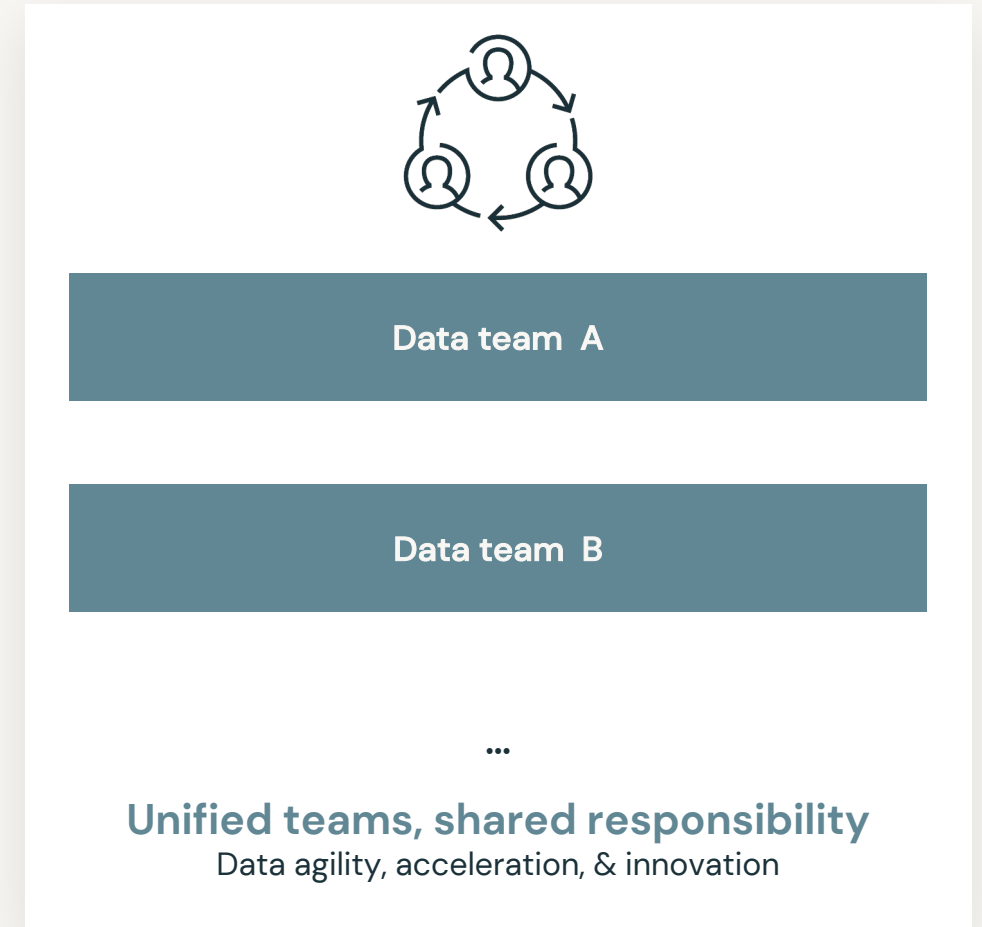
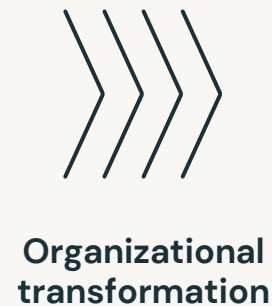
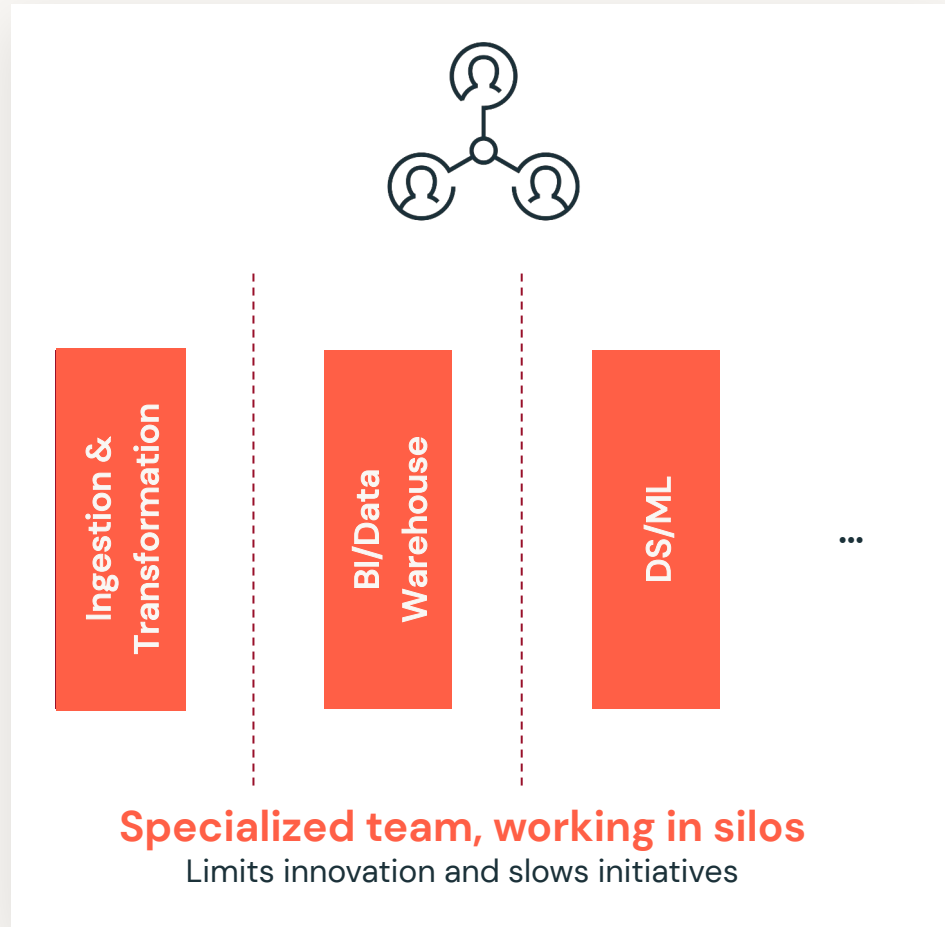
Our **strategy**
in one word



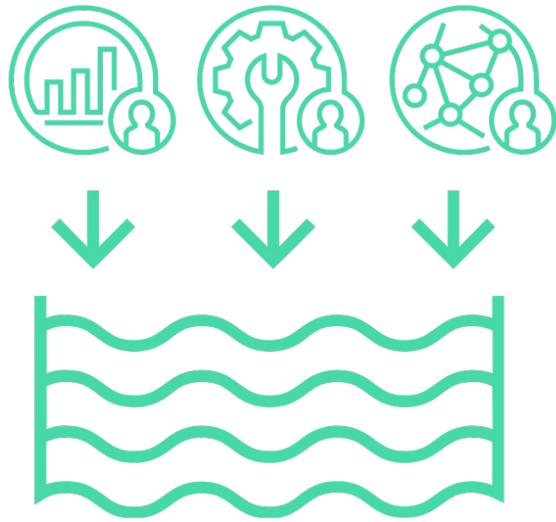
**Lower TCO +
Faster innovation**

Workforce Evolution Present State

The real meaning of openness



Exec Lakehouse



Guiding principles of the Lakehouse



Total cost optimization



Enhance the open network effect

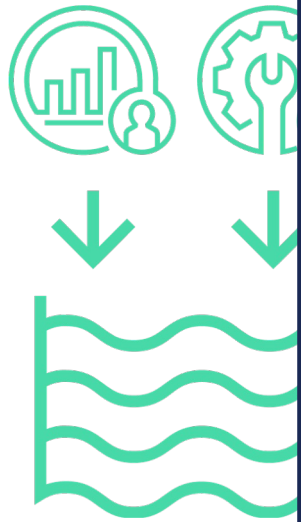


Cloud portable interoperability



Direct interactions between business data producers & business data consumers

Exec Lakehouse



Guid
princip
the Lake

Databricks releases Dolly 2.0, the first open, instruction-following LLM for commercial use



ight

Image by Canva Pro

optimization

open
ect

ole
lity

actions between
ta producers &
ta consumers

What is Dolly, and why you should care

5 min Video



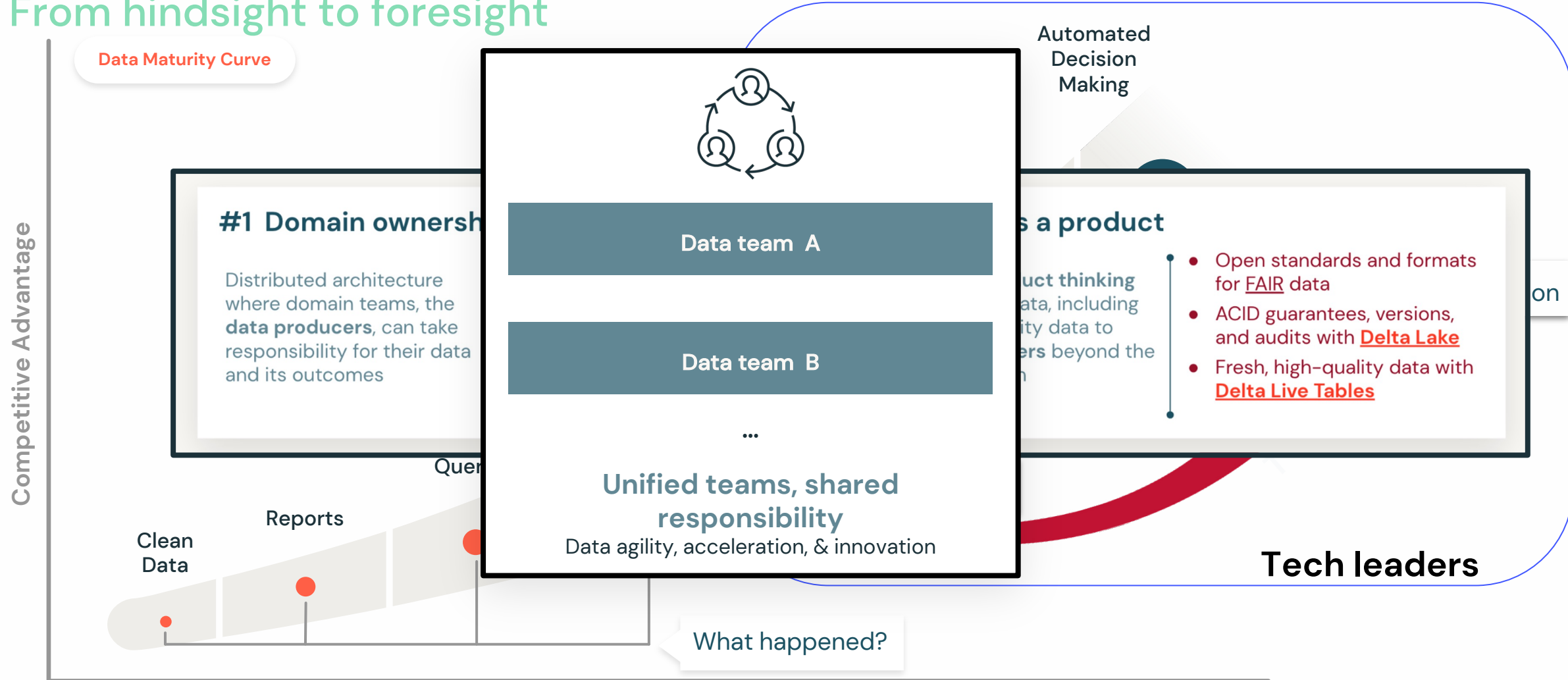
Why Databricks Dolly Is So Important to the Future of AI Adoption

[youtube.com](https://www.youtube.com)

Why Companies move to Lakehouse

Tech leaders are to the right of the Data Maturity Curve

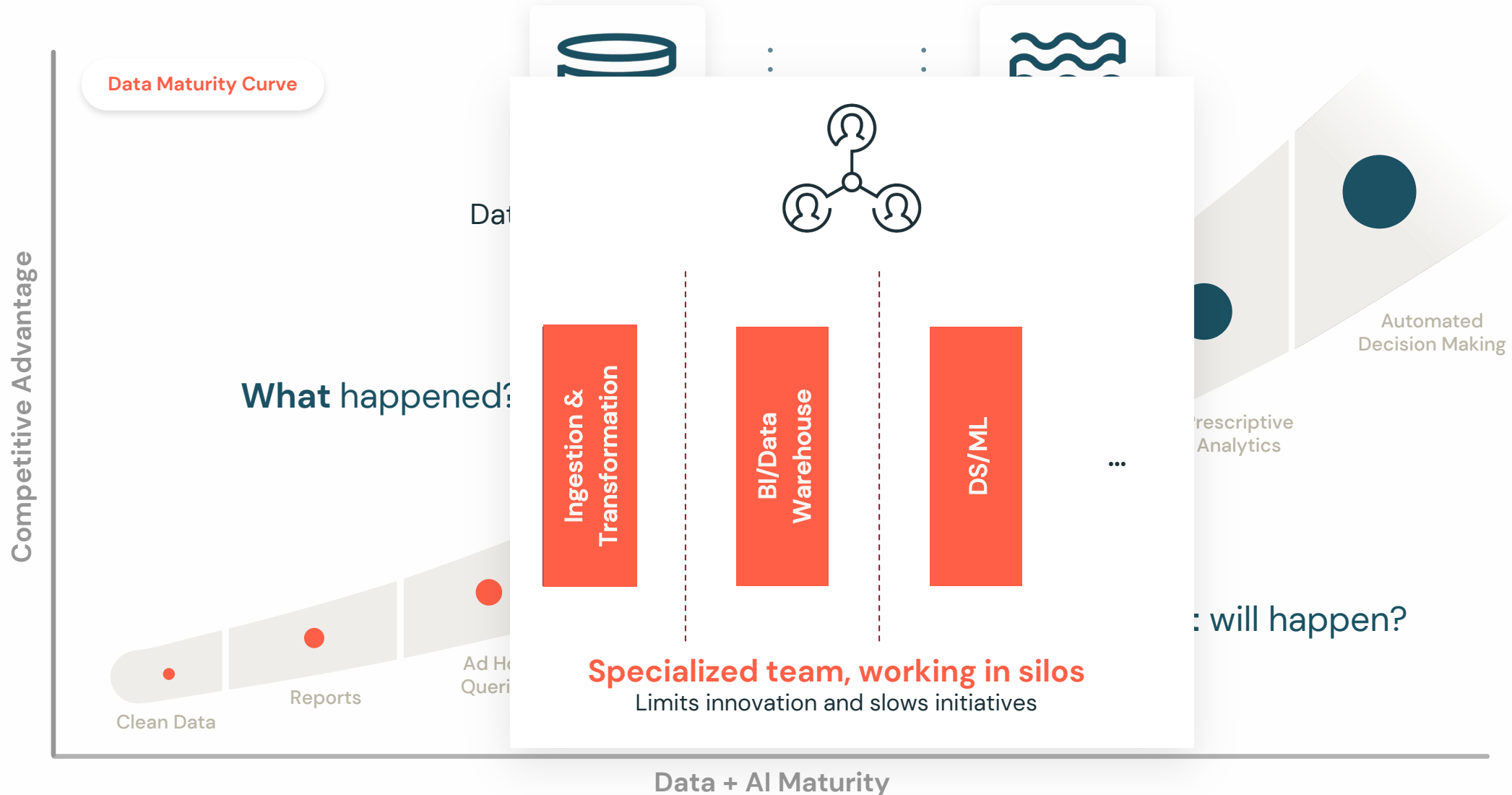
From hindsight to foresight



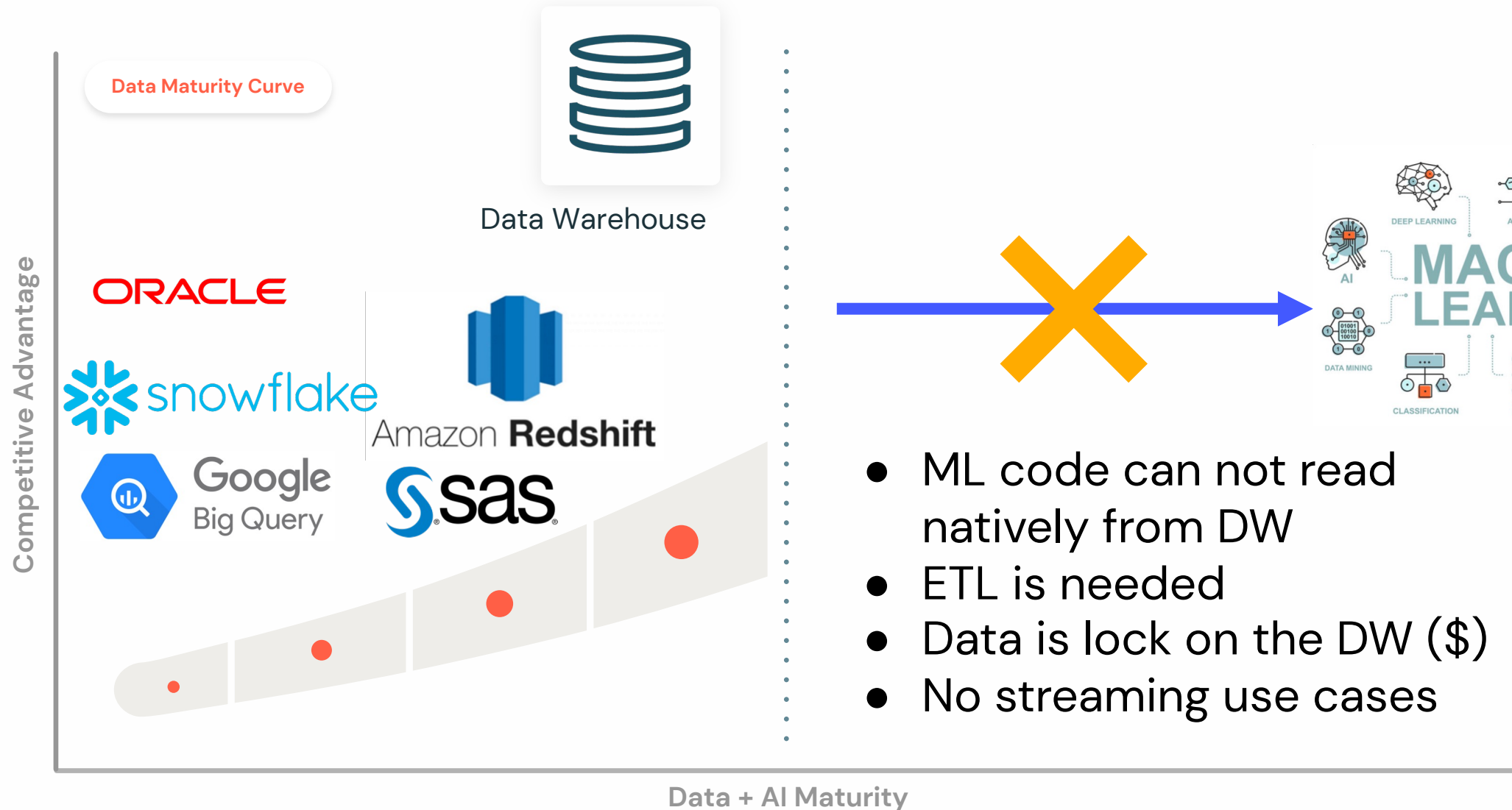
But...
most companies
still struggles
to find success
at scale



Two incompatible architectures get in the way

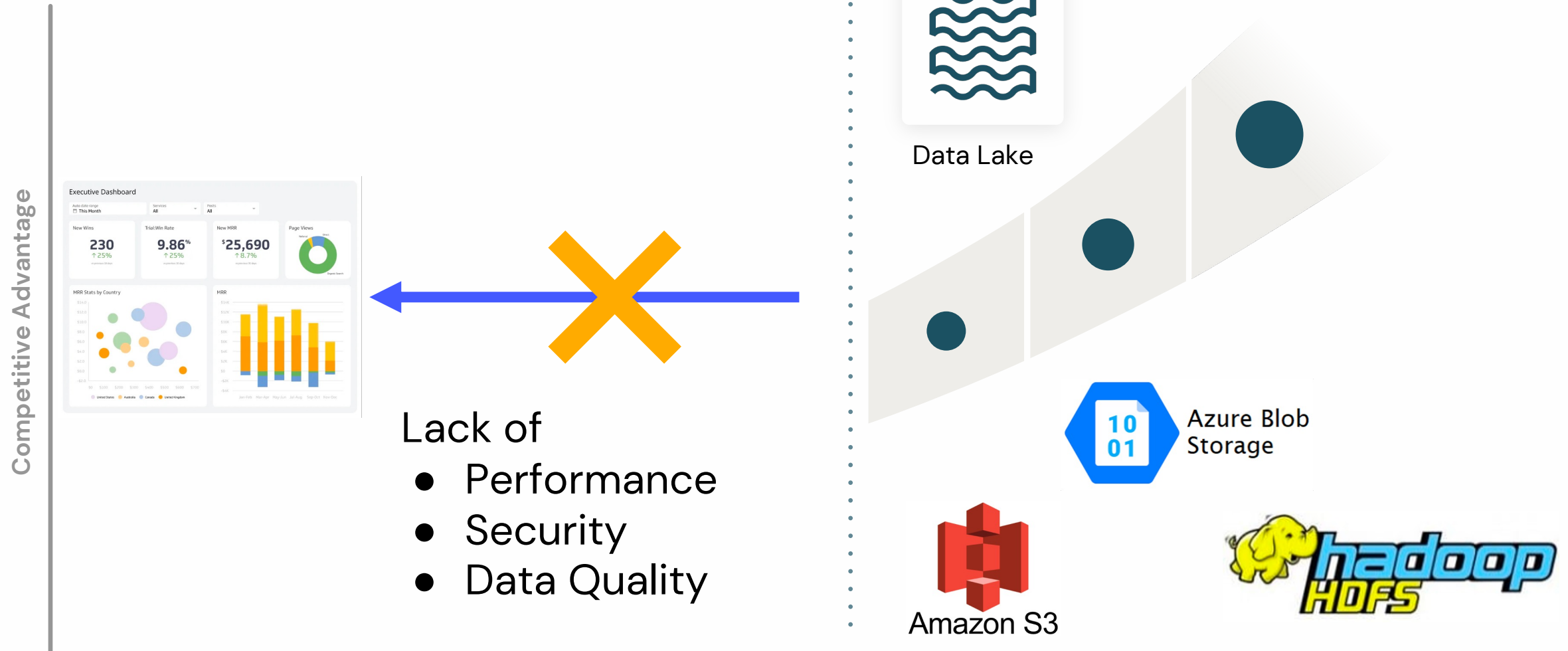


Machine Learning on DW



- ML code can not read natively from DW
- ETL is needed
- Data is lock on the DW (\$)
- No streaming use cases

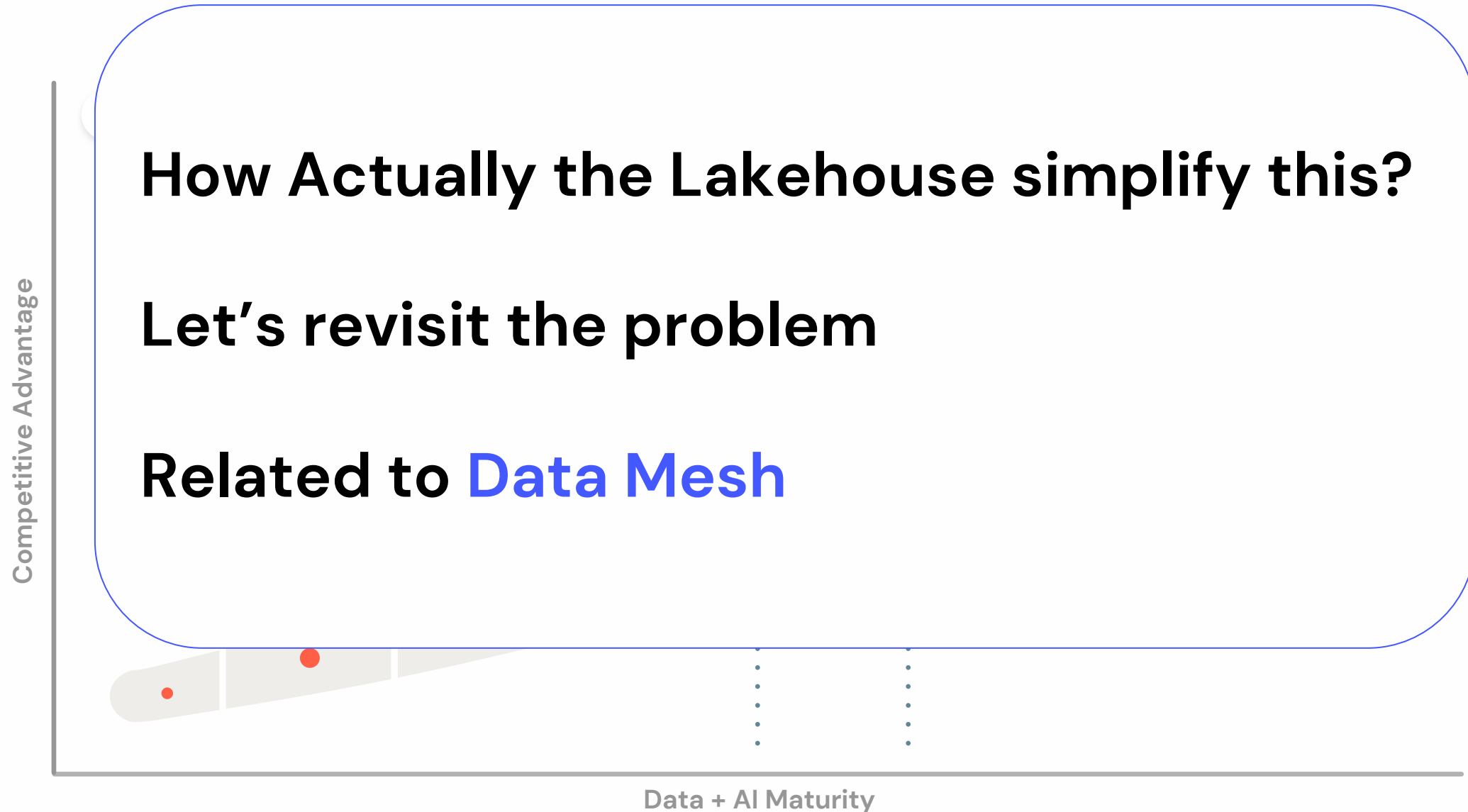
BI on top of the Data Lake



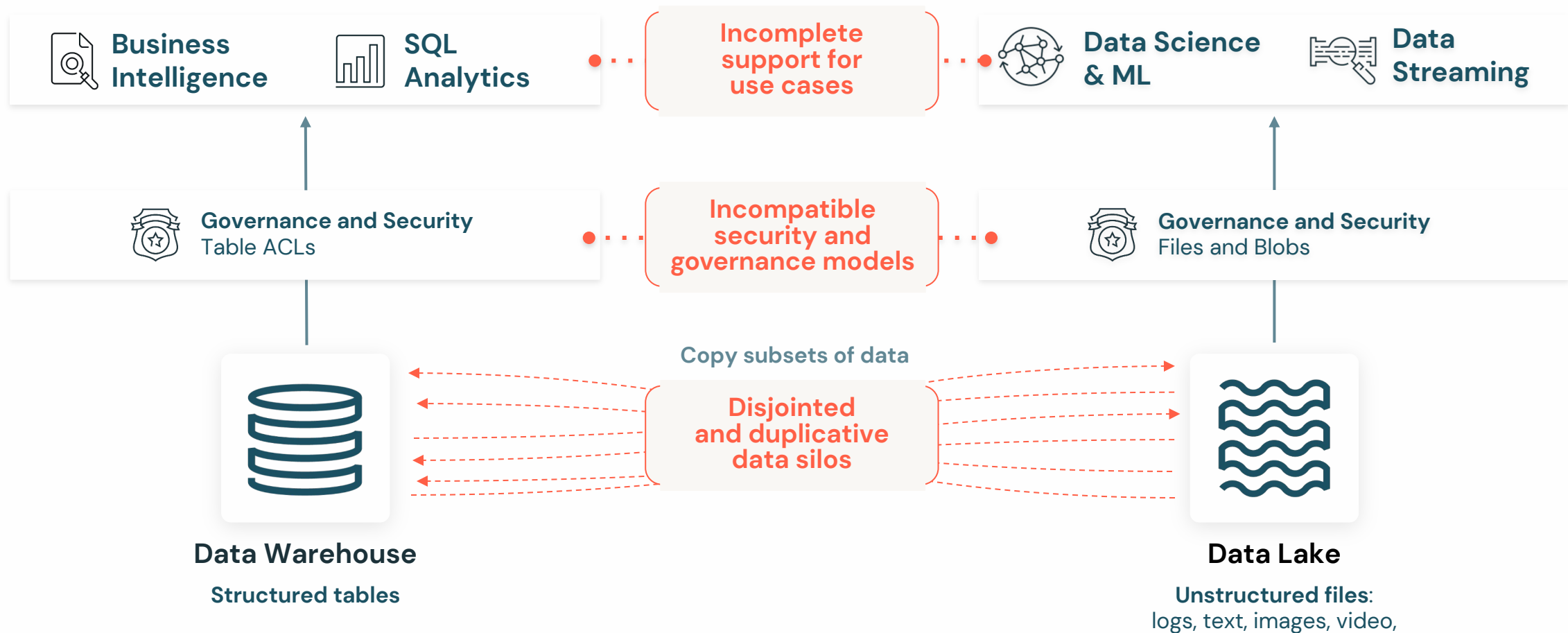
**OK Guillermo but what
about security, etc?**



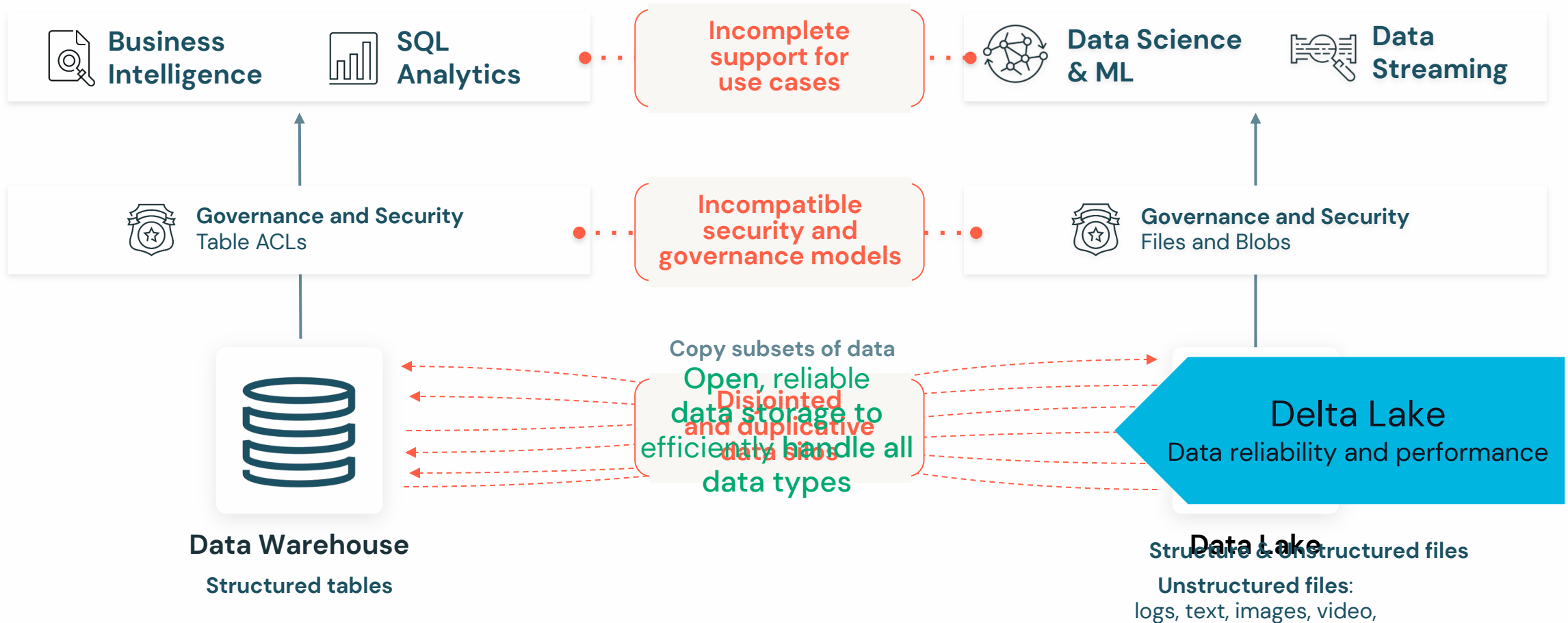
Data flows in both directions



Realizing this requires two disparate, incompatible data platforms

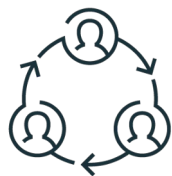


There is no need to have two disparate platforms



There is no need to have two disparate platforms

This is the Lakehouse



Data team A

Data team B

...

Unified teams, shared responsibility

Data agility, acceleration, & innovation

All ML, SQL, BI,
and Streaming
use cases

Incomplete
use cases
use cases

One security
and governance
approach for all data
assets on all clouds

Incompatible
security and
governance models

Open, reliable
data storage to
efficiently handle all
data types



Structure & Unstructured files

Databricks releases Dolly 2.0, the

#1 Domain ownership

Distributed architecture where domain teams, the **data producers**, can take responsibility for their data and its outcomes

- Open and flexible architecture enables [workspace/catalog](#) per domain
- Distributed ownership of data assets and [pipelines](#)

#2 Data as a product

Applying **product thinking** to analytical data, including providing quality data to **data consumers** beyond the source domain

- Open standards and formats for [FAIR](#) data
- ACID guarantees, versions, and audits with [Delta Lake](#)
- Fresh, high-quality data with [Delta Live Tables](#)

#3 Self-service infrastructure platform

Domain-agnostic approach to building, executing, and maintaining interoperable data products through common tools

- Unified platform serving all analytics workloads
- Managed orchestration with [Databricks Workflows](#)
- Auto-scaling/tuning, serverless
- Infrastructure as Code ([Terraform](#))

#4 Federated computational governance

Creating a data ecosystem that adheres to organisational rules and industry regulations through standardisation

- Discovery, access and lineage with [Unity Catalog](#)
- Global policy templates for access to data and [compute resources](#)

Databricks thrives within your modern data stack

BI and Dashboards

Power BI, Tableau, Looker, MicroStrategy, ThoughtSpot, Qlik

Machine Learning

MathWorks, Labelbox, John Snow LABS, Azure Machine Learning, H2O.ai, Amazon SageMaker

Data Science

PyCharm, Jupyter, R Studio, HEX

Data Governance

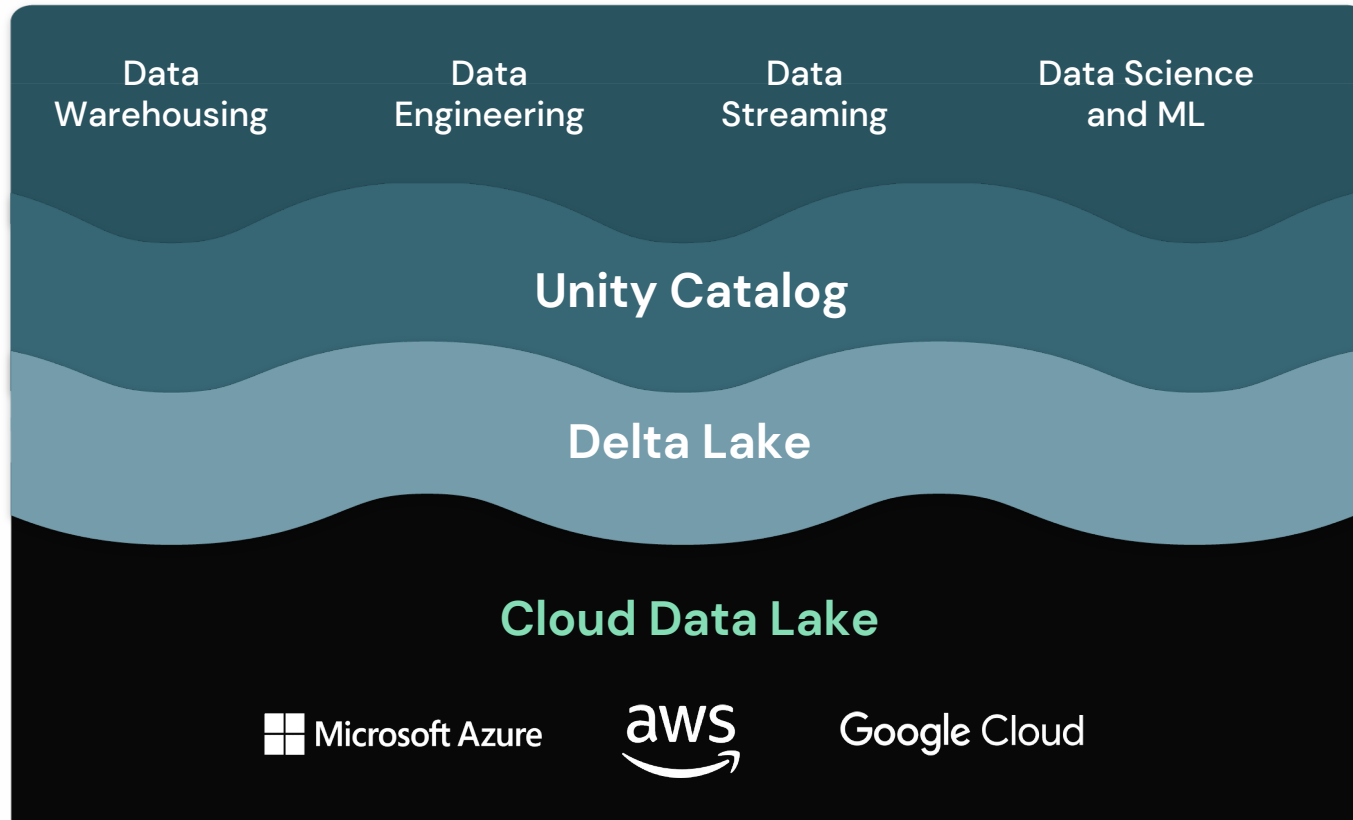
Collibra, MMUTA, PRIVAERA, Quest, Alation, AZURE PURVIEW

Data Pipelines

dbt Labs, MATILLION, Azure Data Factory, Informatica, Prophecy

Data Ingestion

Fivetran, arcion, CONFLUENT, Rivery, Airbyte, Qlik



databricks

BRICKBUILDER SOLUTION

accenture
avanade
Capgemini
Cognizant
Deloitte.
slalom
Infosys
Navigate your next
TREDENCE
Lovelytics



iPhone



Laptop



Watch



Music





Phone



Camera



GPS



OK... let's go back...

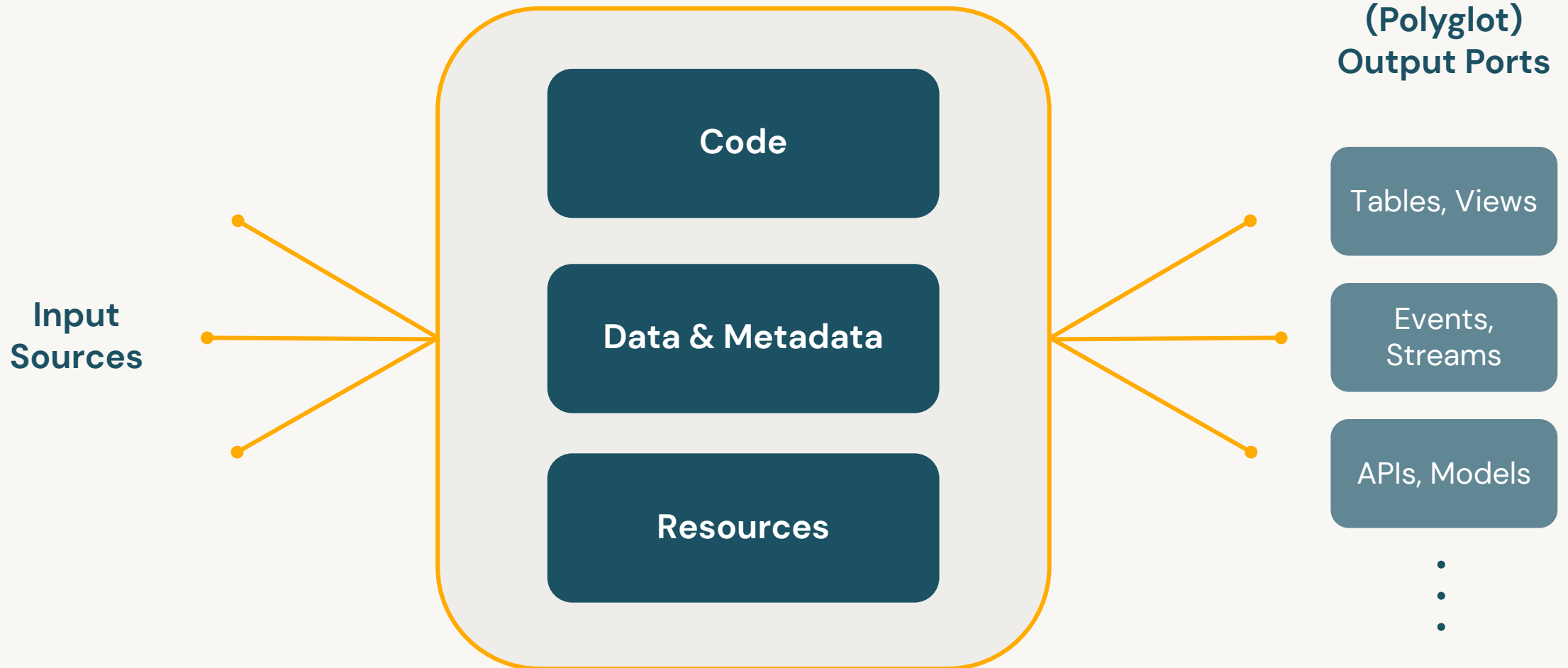
Data Domains,

Data Products and

Governance

Anatomy of a Data Product (1/2)

Not just datasets. Not just tables or files



Anatomy of a Data Product (2/2)

Not just datasets. Not just tables or files

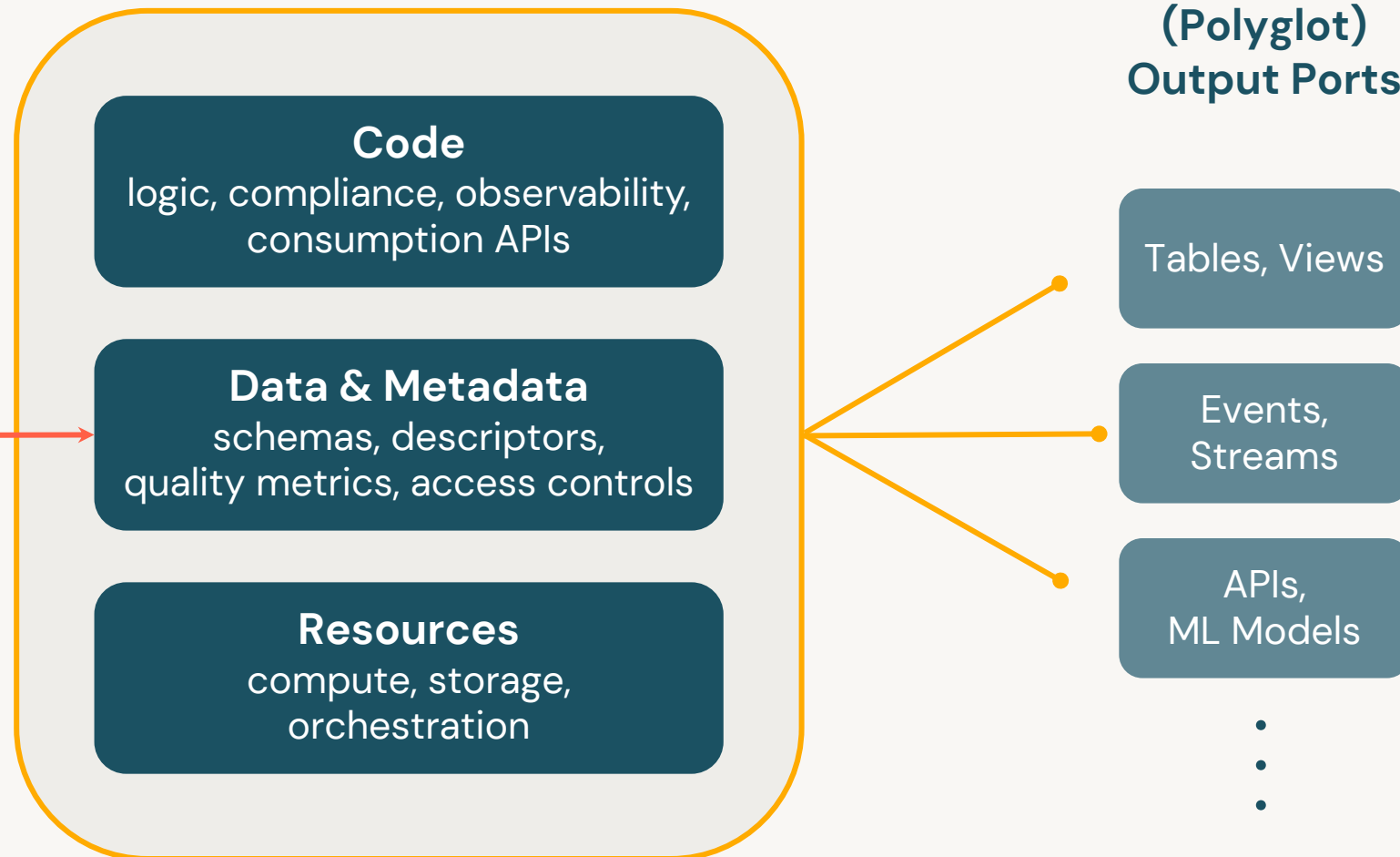
Data & Metadata

“Depending on the nature of the domain ... data can be served as events, batch files, relational tables, graphs, etc., while maintaining the same semantic.”

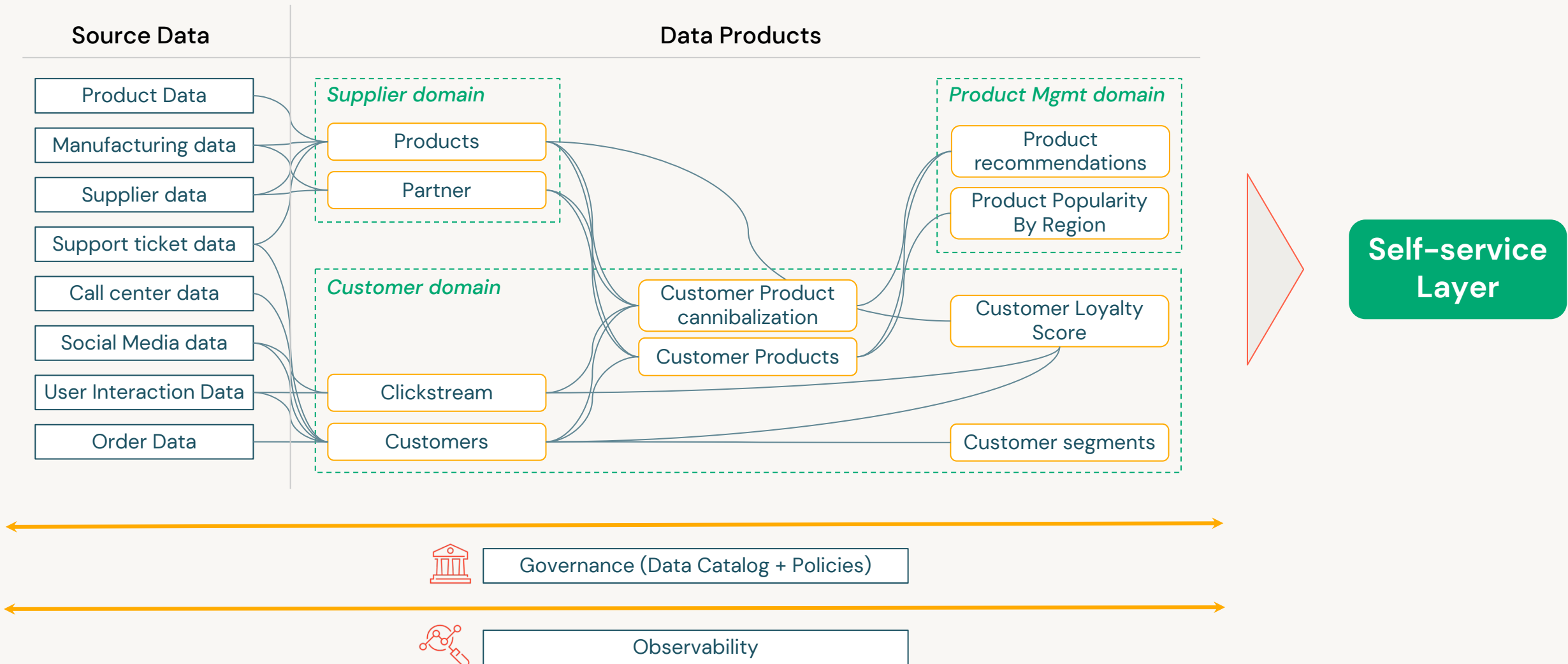
[Source: Zhamak Dehghani](#)

Lakehouse unifies:

- batch and streaming semantics
- consumption and governance of tables and files

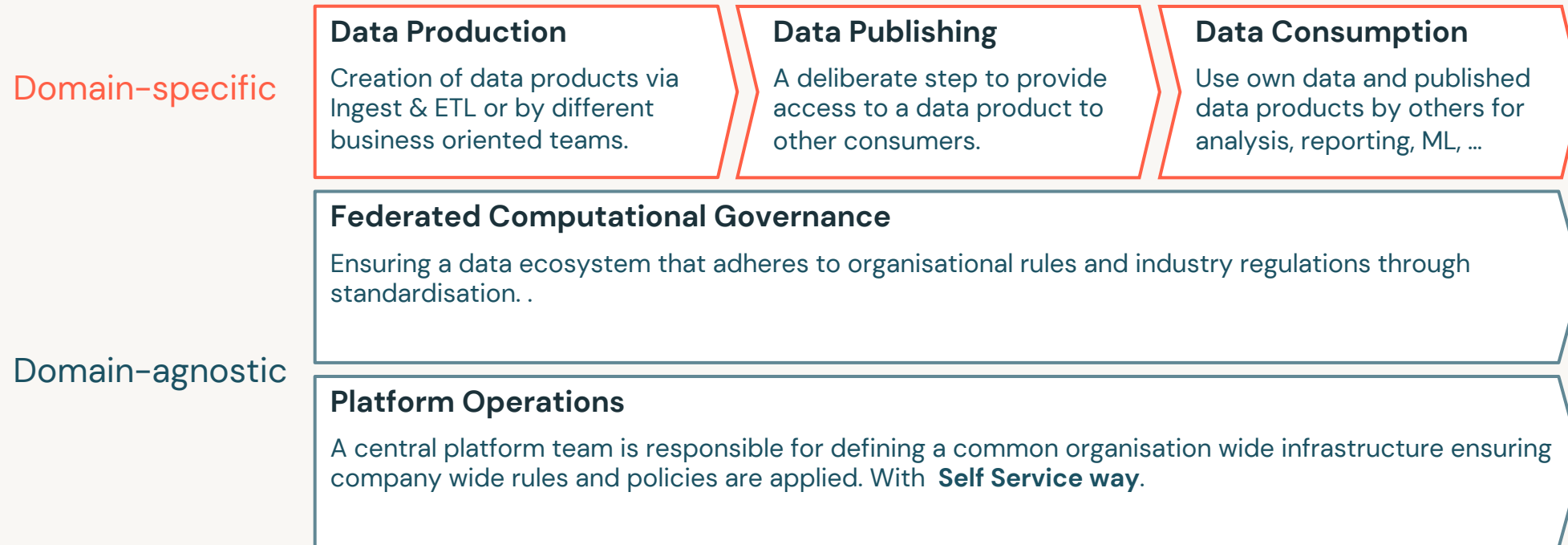


Data Domains and Data Products



Five core processes

To be taken into account when defining a data mesh architecture

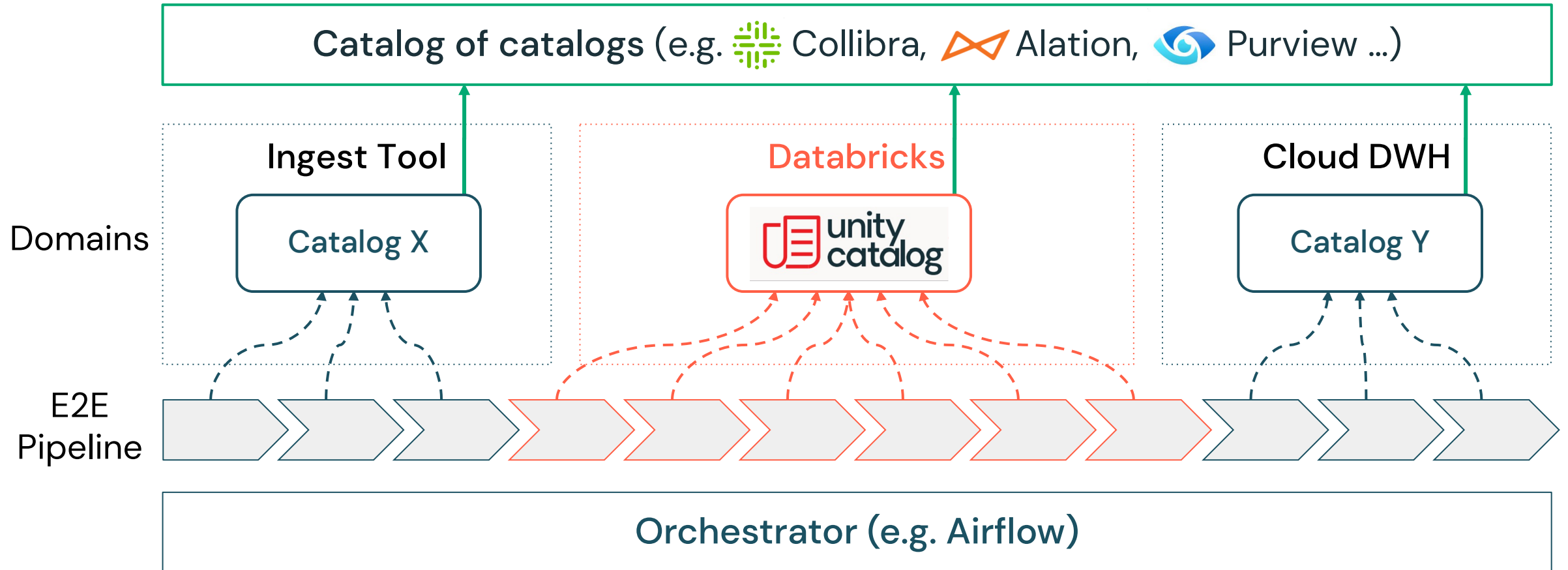




Adopting Data Mesh: Implementation based on Databricks Lakehouse



Enterprise Level Catalog integration



Lineage information flow:

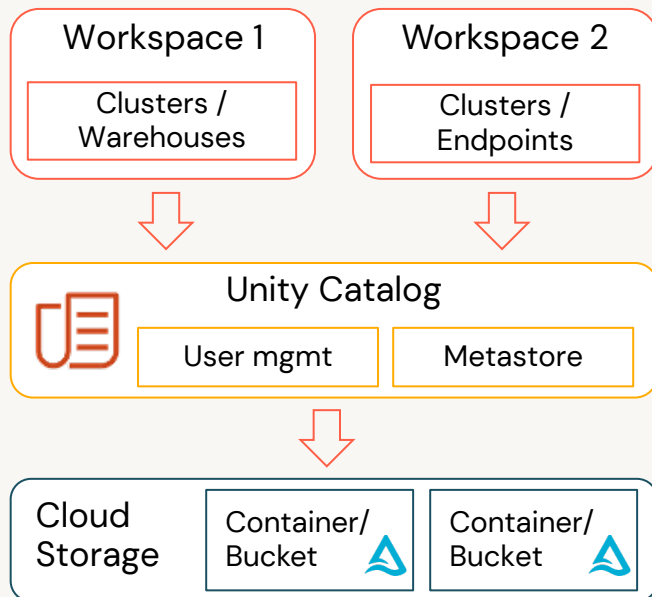
 Pipeline step sending lineage to domain's catalog (e.g. UC)

 Domain's catalog to global catalog of catalogs

Unity Catalog and data products

Unity Catalog

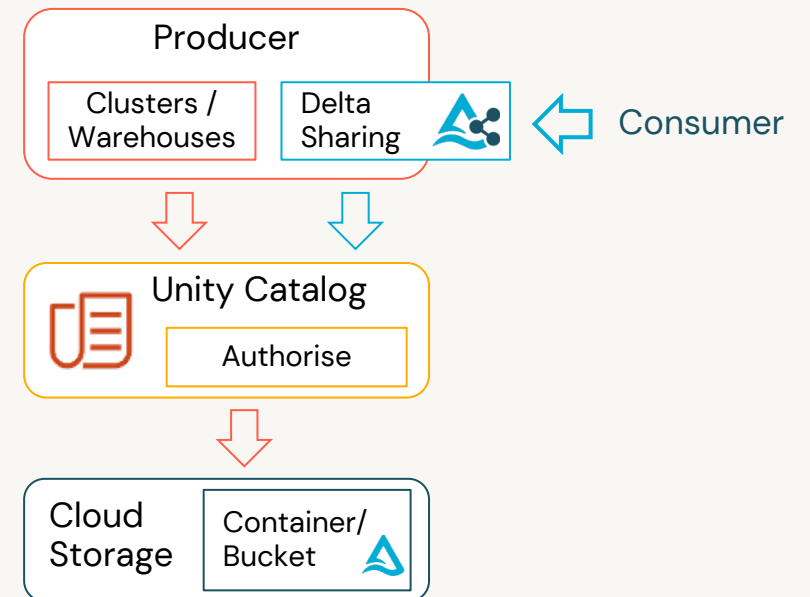
Govern once. Secure everywhere.
Standardisation and discovery.



⇒ Data as a **global** product

+ Delta Sharing

Authorise access to data beyond
organisational boundaries.



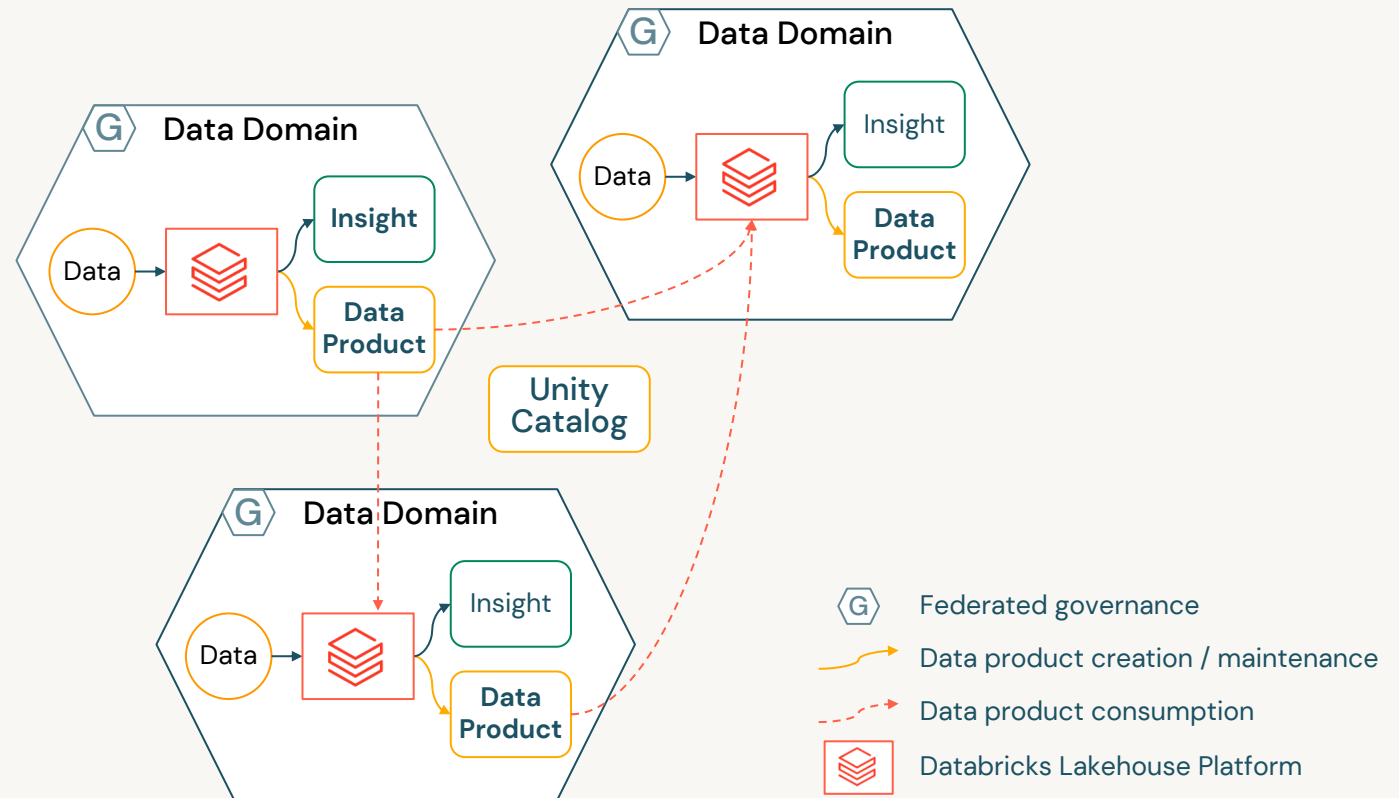
⇒ Data as an **external** product

Databricks Lakehouse and Data Mesh

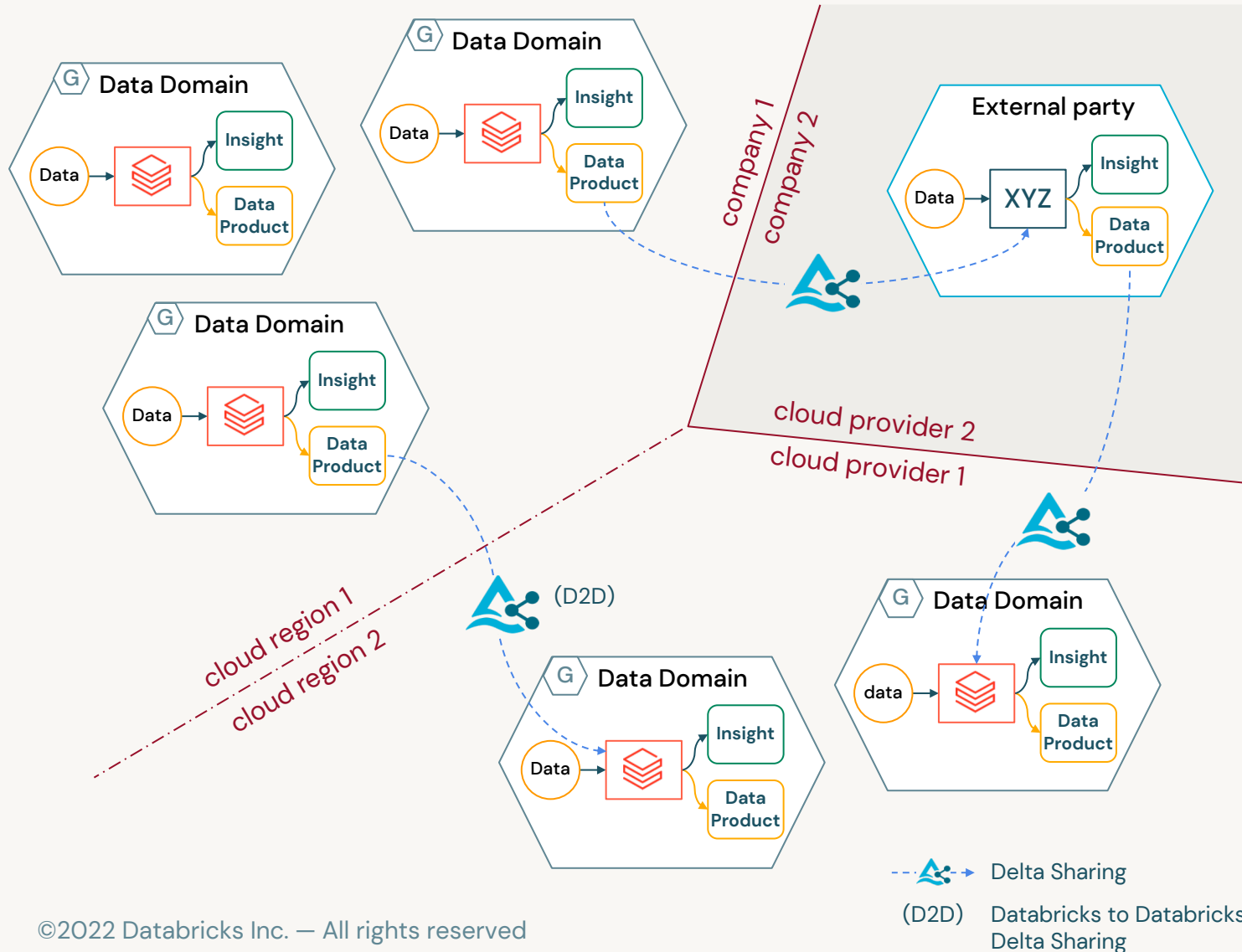
Business units are organized in **Data Domains** and own their respective sources, data and metadata. Using Databricks Lakehouse, they create domain specific insights from data and offer **Data Products** to other domains using **Unity Catalog** a common data catalog. Compliance to organisational rules and industry regulations is ensured via **Federated Governance**.

With the Databricks Lakehouse Platform, data domains can be created on different levels:

- One Workspace using clusters to isolate domains
- Using a separate Workspace per data domain
- Data domains being full Lakehouses



Scaling the Data Mesh

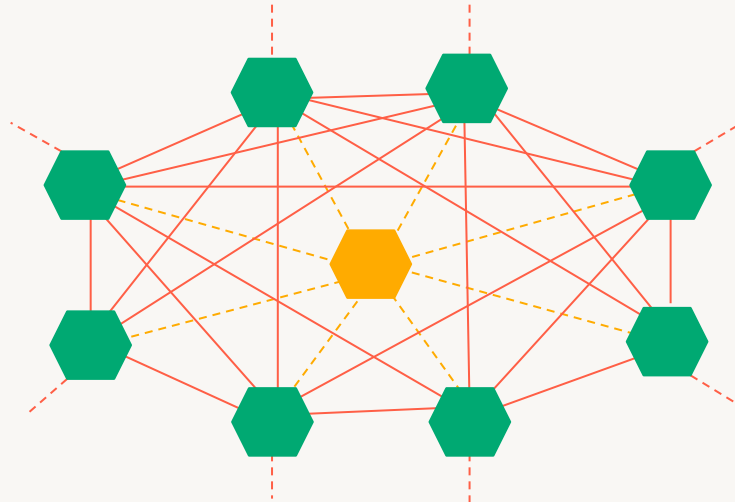
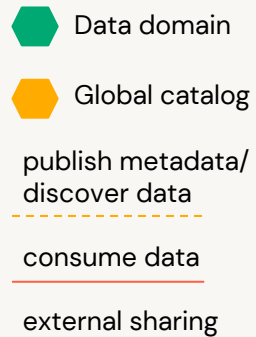


- A data mesh based on Databricks can be extended across cloud regions and cloud providers
- **Delta Sharing** is an open protocol to share data products between the domains across business and technical boundaries
- The Delta Sharing protocol is vendor agnostic, hence the different domains of the data mesh
 - do not need to use the same technology stack
 - can even be different companies like business partners

Organising like a Data Mesh (1/3)

Balancing autonomy with complexity

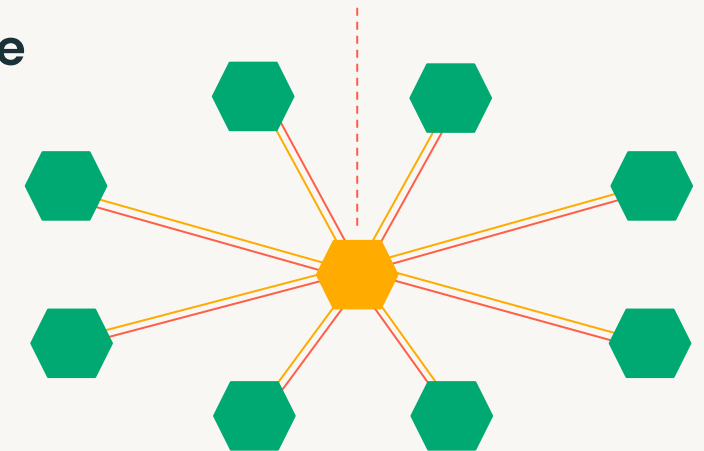
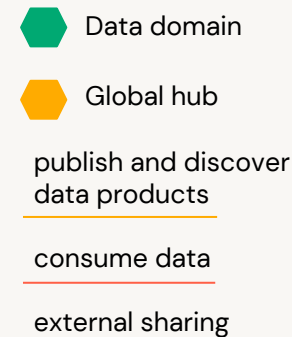
Harmonised



Global catalog for discovery, each domain hosts and serves its own data products

- Truest interpretation of data mesh principles
- Requires each domain to have skills to manage end-to-end data lifecycle
- May create inefficiencies if there is a high-level of data re-use across many domains and sharing

Hub-and-Spoke



Global data hub for publishing, discovery and serving of shareable data products

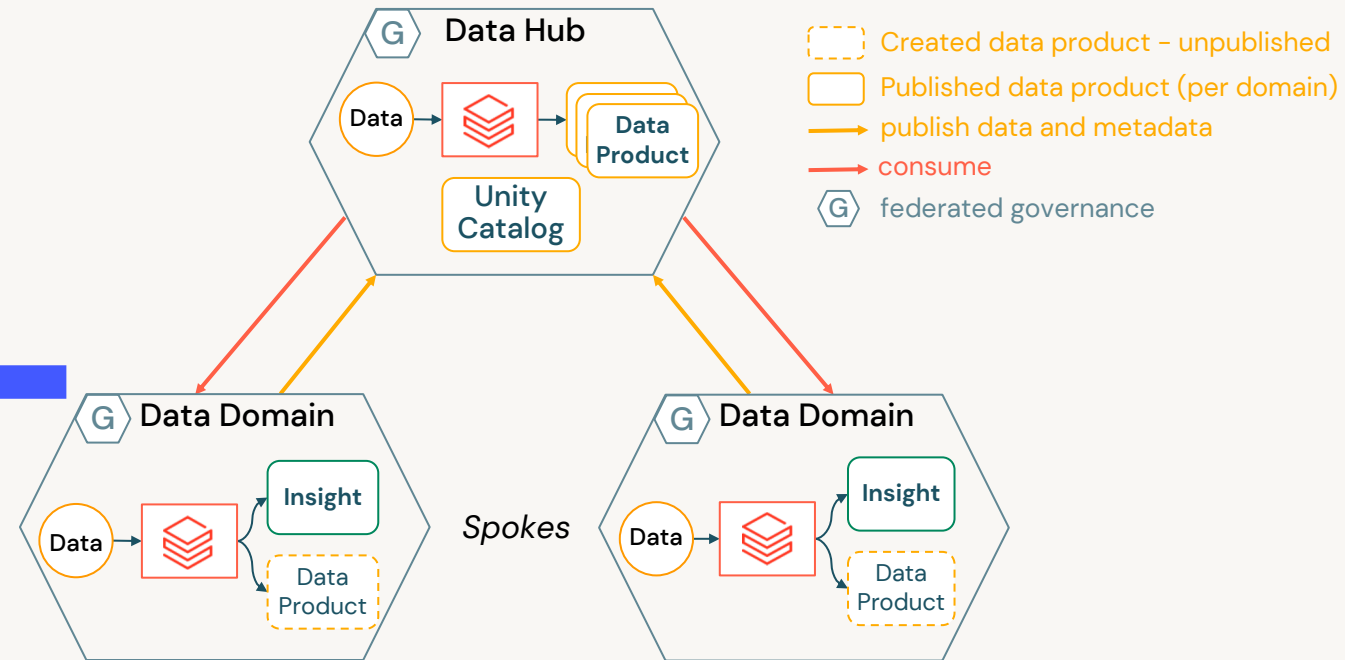
- Hybrid data mesh with some centralisation
- Requires each data spoke (domain) to publish shareable data products to a global data hub
- Can reduce data sharing and management overheads when there are a large number of domains

Organising the Data Mesh

Hub and Spoke Data Mesh

Approach

- Data domains (spokes) will create domain specific high quality data products
- Data products will be published to the data hub
- The data hub provides **generic services**
 - platform operations for data domains.
 - centralized data publishing via **self service** (data QA is kept in the domains)
 - Data catalog via Unity Catalog features
 - **generic data services** like time travel, historization, ...
- The data hub can also act as a data domain additionally offer own data products



Implications

- Full automation and self-service will help to avoid the data hub getting a bottleneck for data product publishing.
- To keep independence, data products in the data hub are partitioned per domain (no integration).
- The advantage is that data domains can benefit from centrally developed and deployed data services without own effort.

Databricks

AWS

ITV

Digital transformation for rapid value creation: ITV's Data Mesh journey

< Who we work with

Data mesh

Client story

Digital transformation for rapid value creation: ITV's Data Mesh journey

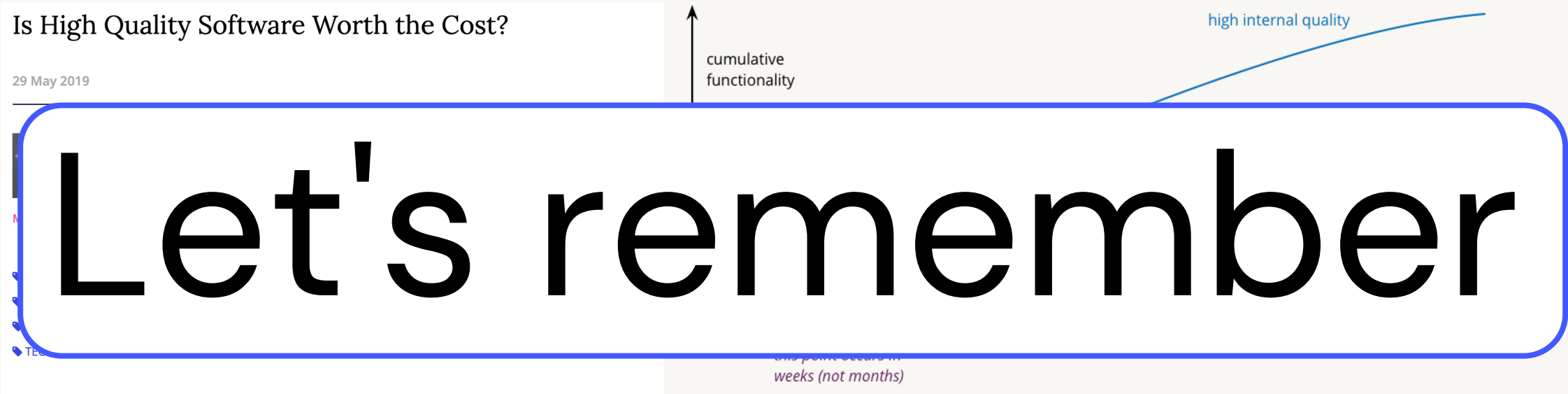
5 min on Global
household name
(flat pack) 

From my own experience

- Only possible if IT work together with Business (not a tool)
- Central mandate to work on a Mesh Architecture, giving time and allow for failure.
- Holistic approach: the key team (center of excellence)
 - This team define the tech and best practices
 - This team and Databricks interact on weekly bases
 - This team interacts with cloud vendors and C&SIs
- Analyze and integrate tools
 - Colibra, Pureview, etc
- Central repository of best practices (Confluence, etc)
- Working on the edges is not simple (skills)
... peer 'architecture', training, support...



AI needs good Data → Data needs best practices



Your CoE for Databricks & Databricks Champions we can help you!

AI needs good Data → Data needs best practices

Is High Quality Software Worth the Cost?

29 May 2019



Martin Fowler

Your job help here →

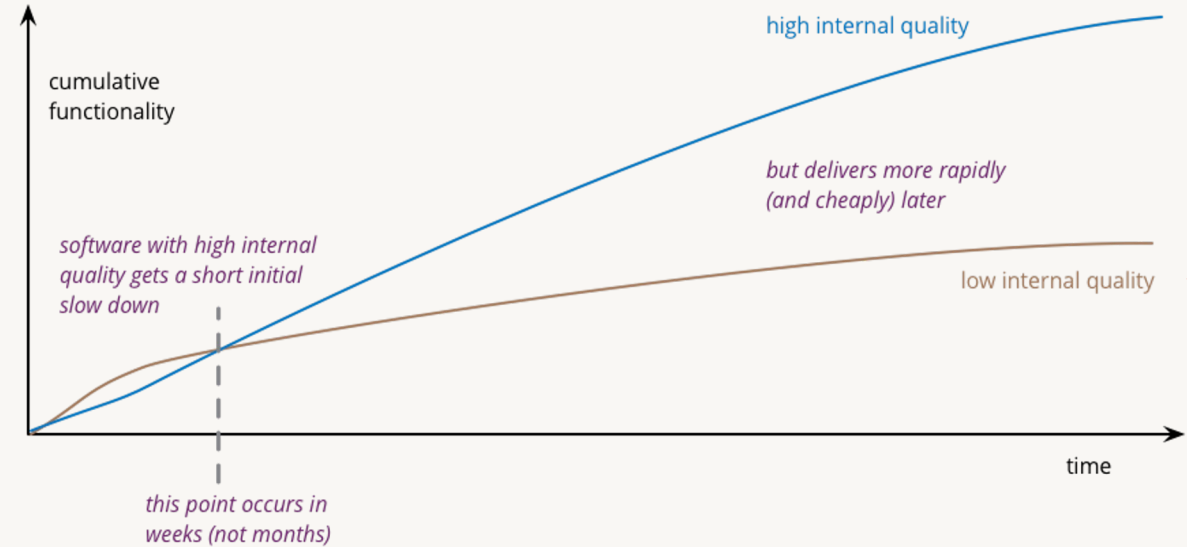
CONTENTS

We are used to a trade-off between quality and cost
Software quality means many things
At first glance, internal quality does not matter to customer
Internal quality makes it easier to enhance software
Customers do care that new features come quickly
Visualizing the impact of internal quality
Even the best teams create cruft
High quality software is cheaper to produce

SIDEBARS

Dora studies on elite teams

- PROGRAMMING STYLE
- PRODUCTIVITY
- PROJECT PLANNING
- TECHNICAL DEBT



Data Mesh will not help you go fast if there is not High Internal Quality Software/Architecture

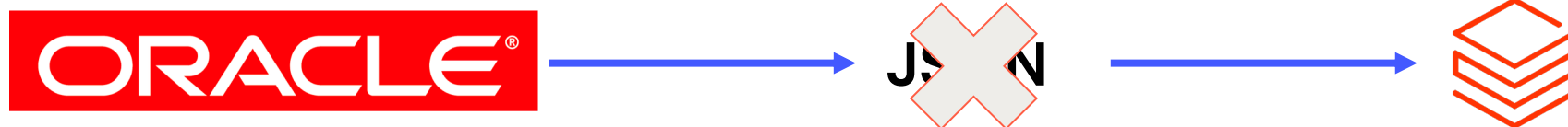
Your CoE for Databricks & Databricks Champions we can help you!

How *not* to Migrate: Migration Oracle into Databricks

- Client DE/DS/BI did not have Databricks certifications
- Partner (C&SI) DE/DS/BI did not have Databricks certs

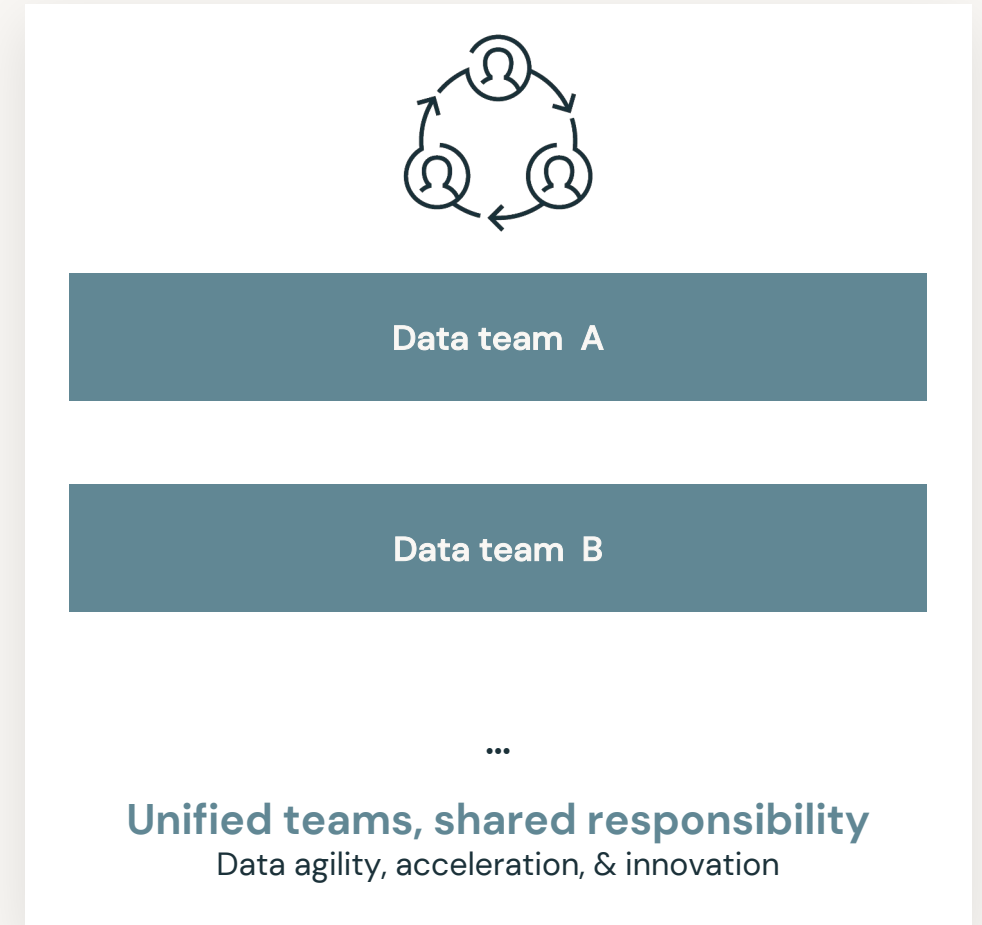
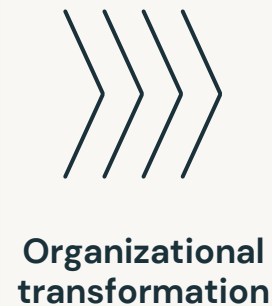
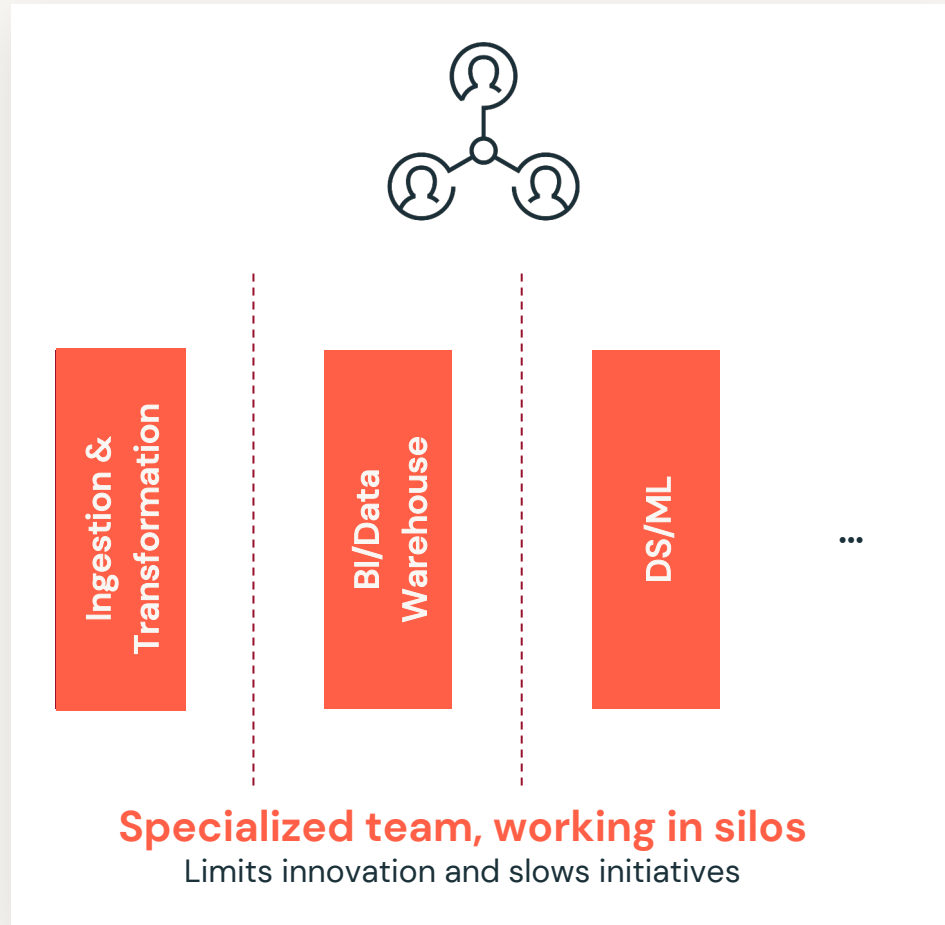
How *not* to Migrate: Migration Oracle into Databricks

- First try failure
- **How you store the data** influence the performance of the Lakehouse
 - 500 G stored on Azure Blob storage as **json files**, compressed into 9G files !!!
- Performance is **CPU + Storage**
- TL,DR use **Delta Lake**



Workforce Evolution Present State

The real meaning of openness



From my own experience

- Lighthouse project/team
 - Small iterations



Questions

¿Qué temas les interesaría que cubriéramos en nuestras siguientes sesiones?

GRACIAS!

Para aquellos que aún no son clientes de Databricks,
comience su prueba gratuita de 14 días
<https://www.databricks.com/try-databricks>



Thank you