DATA+AI SUMMIT 2022

Meshing about with Databricks...

How to implement a Data Mesh on Databricks

ORGANIZED BY databricks

Som Natarajan
Solution Architect, Databricks

Jason Pohl
Dir of Data Management, Databricks

Agenda

Meshing About with Databricks

- Why Data Mesh?
- Why Data Lakehouse?
- Data Mesh Architecture with Databricks
- Data Mesh at a Top 10 pharmaceutical
- Demo

Why Data Mesh

Why Data Mesh?

Decentralization and distribution of responsibility

Domain Ownership

Decentralized & Autonomous Teams

Responsibility owned by those closest to data

Map to business org

Data as a Product

Data Product Owner

Serve Consumers as Customers

Measure Success of Products

Self-Serve Platform

Distributed and Scalable

Easily create & terminate resources on-demand

Compute & data locality

Federated Governance

Decentralized

Domain Self-Sovereignty

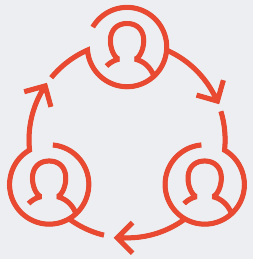
Interoperability

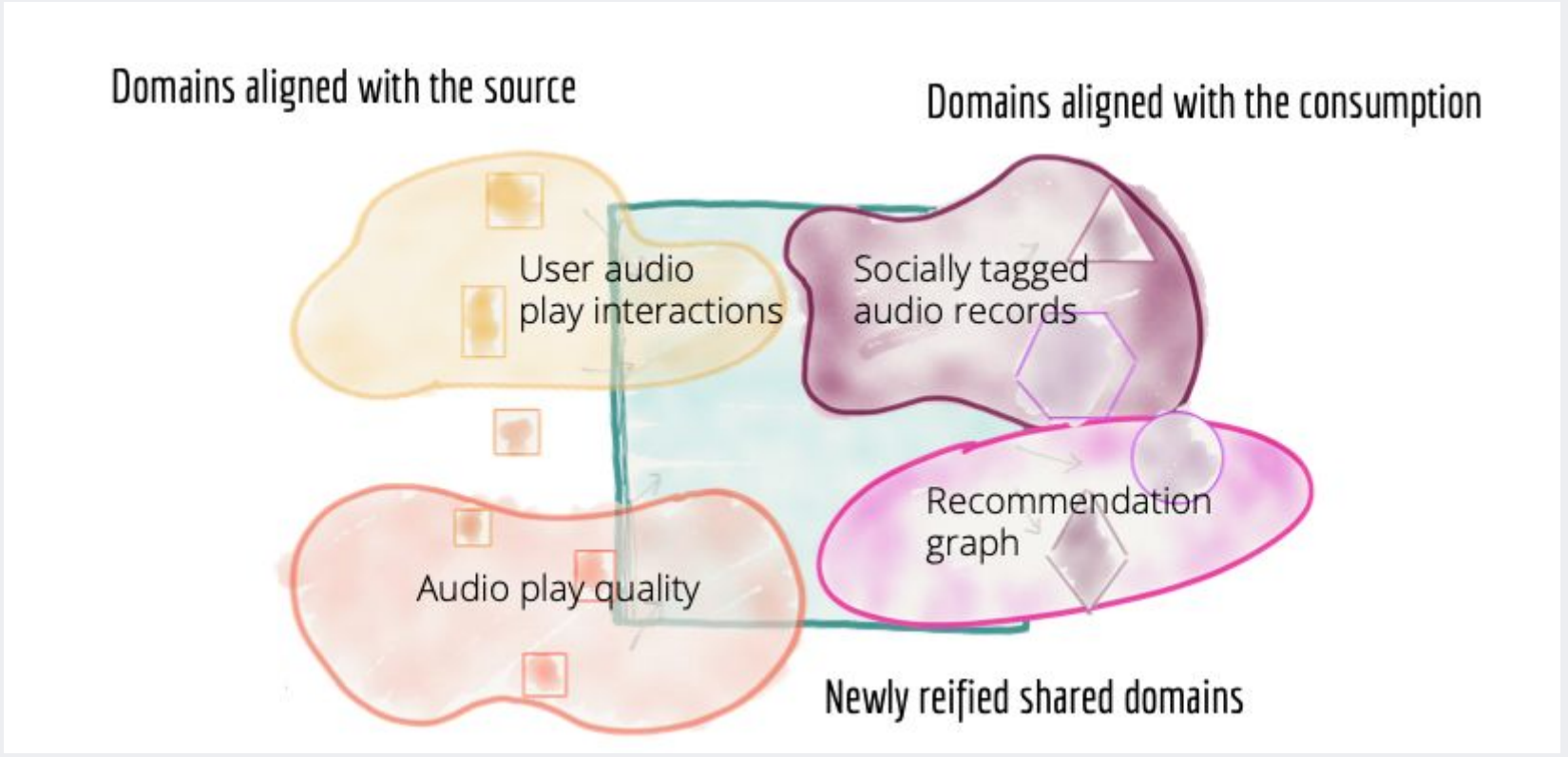
Global Standardization

Domain Data Teams

A team should autonomously own a domain capability

- Distributed Data Ownership
- Create & Own Data Products
- Implementation is decided by team



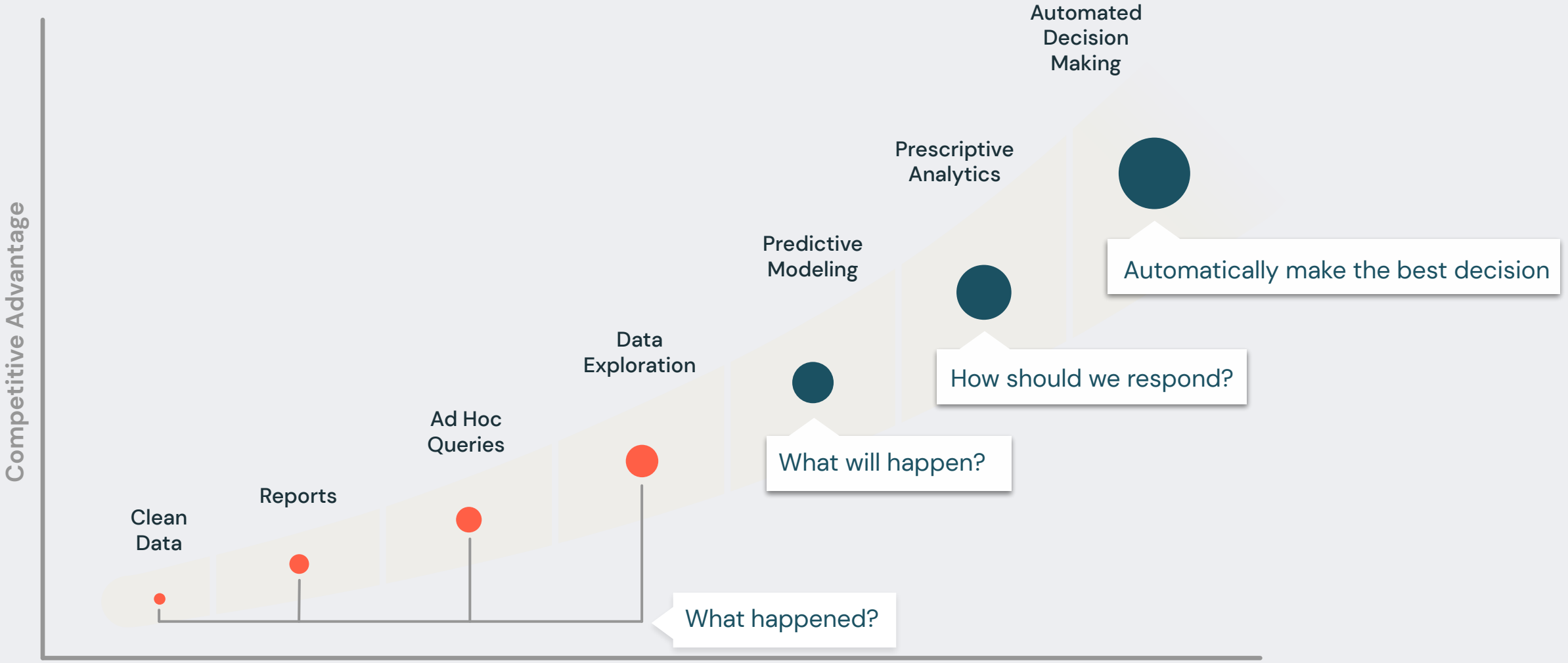


How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh
–Zhamak Dehghani

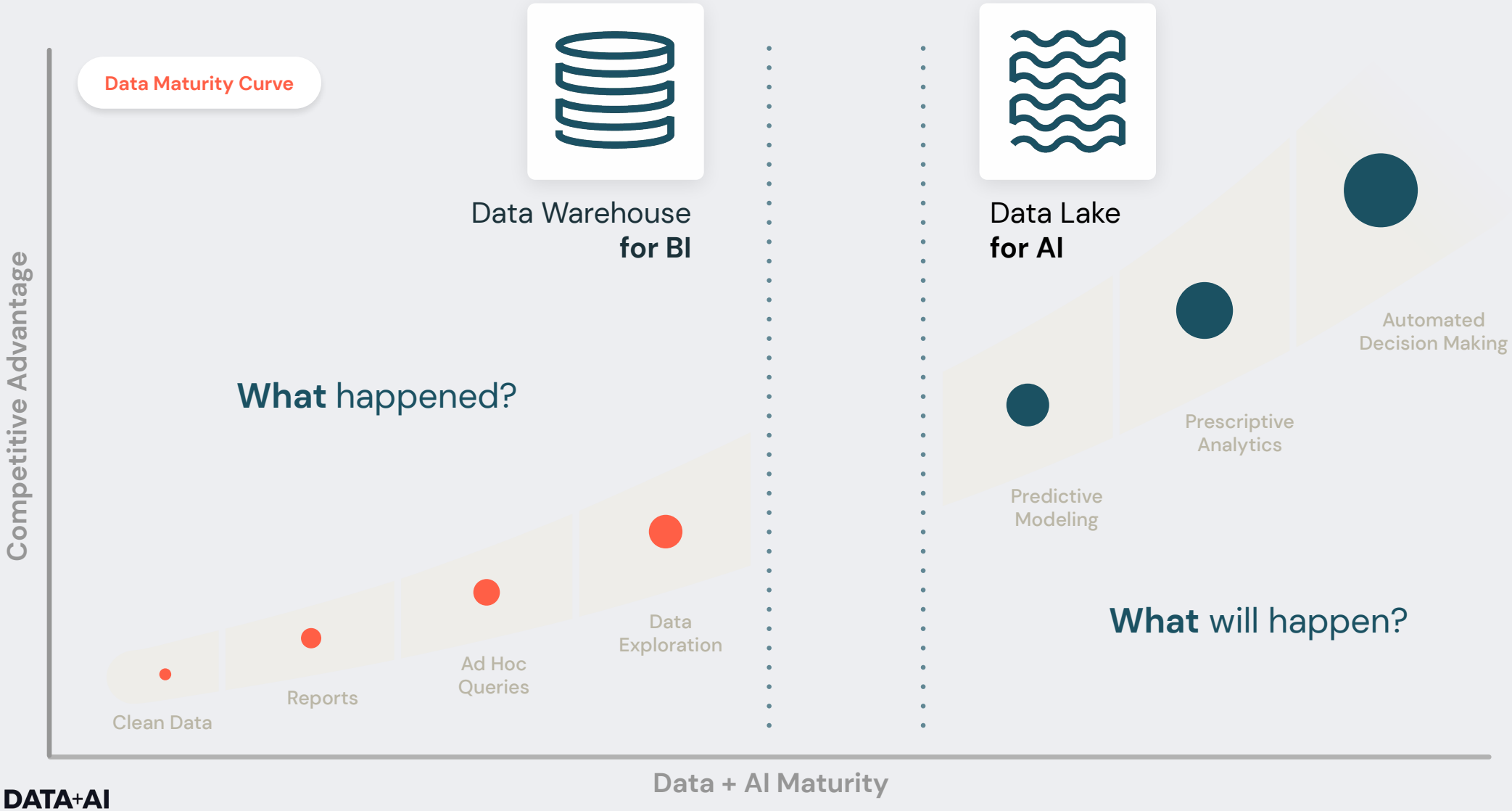
Why Data Lakehouse?

Data Maturity Curve

From hindsight to foresight



Incompatible data platforms emerged



Data Lakehouse

All of your use cases on all of your data with open standards

All Use Cases

Single platform for:

- Data Warehousing
- Data Engineering
- Data Science
- Machine Learning
- Streaming

All Data

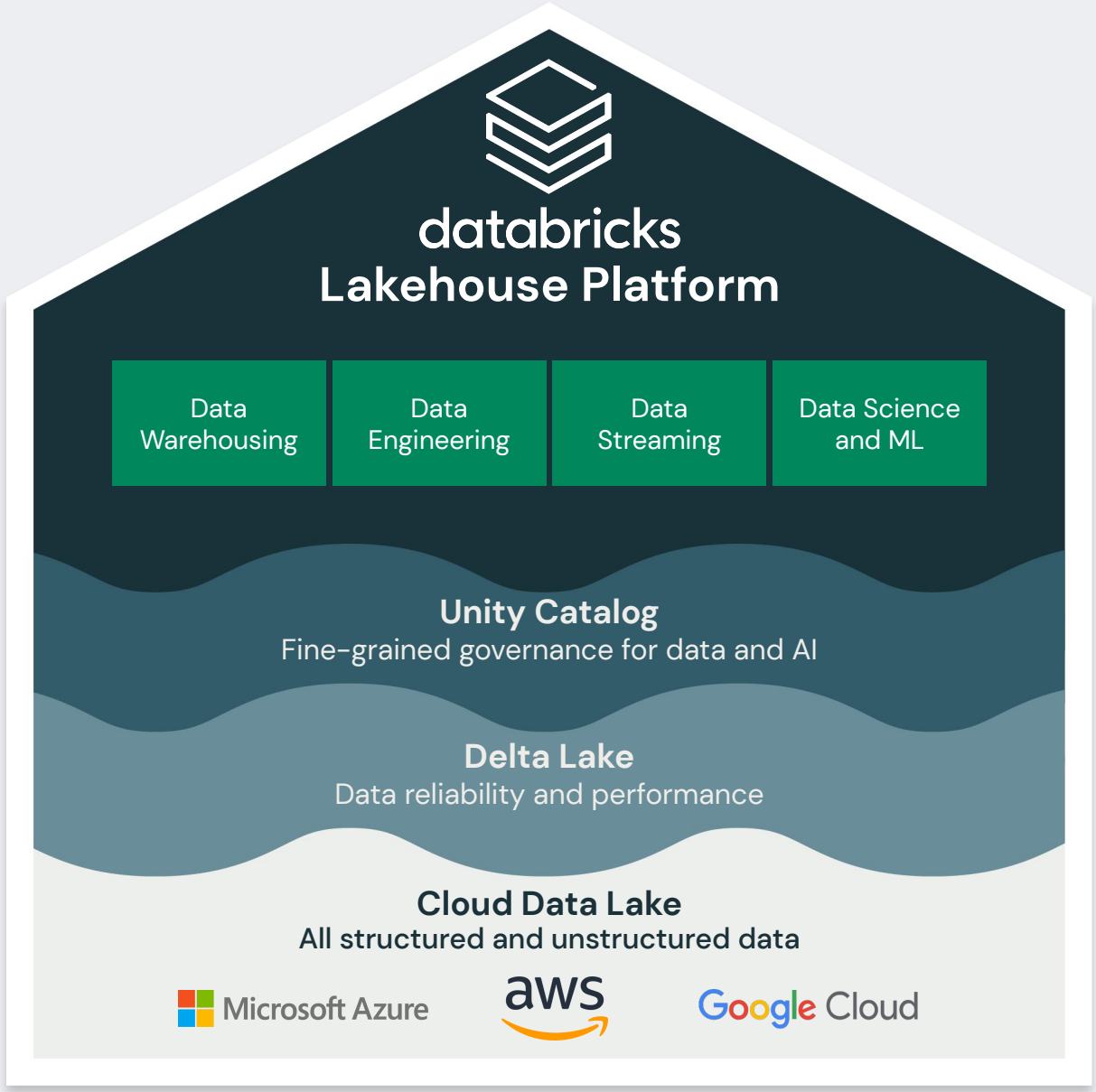
Scale with the cloud

- Structured Data
- Semi-Structured
  - JSON
  - XML
- Unstructured Data
  - Images
  - Videos
  - Audio
  - Text

Open Standards

Portability & Access

- Data Applications
- ML Libraries
- Compute Engine
- Data Sharing
- Governance Layer
- Storage Layer



Databricks Lakehouse Platform

Simple

Unify your data warehousing and AI use cases on a single platform

Open

Built on open source and open standards

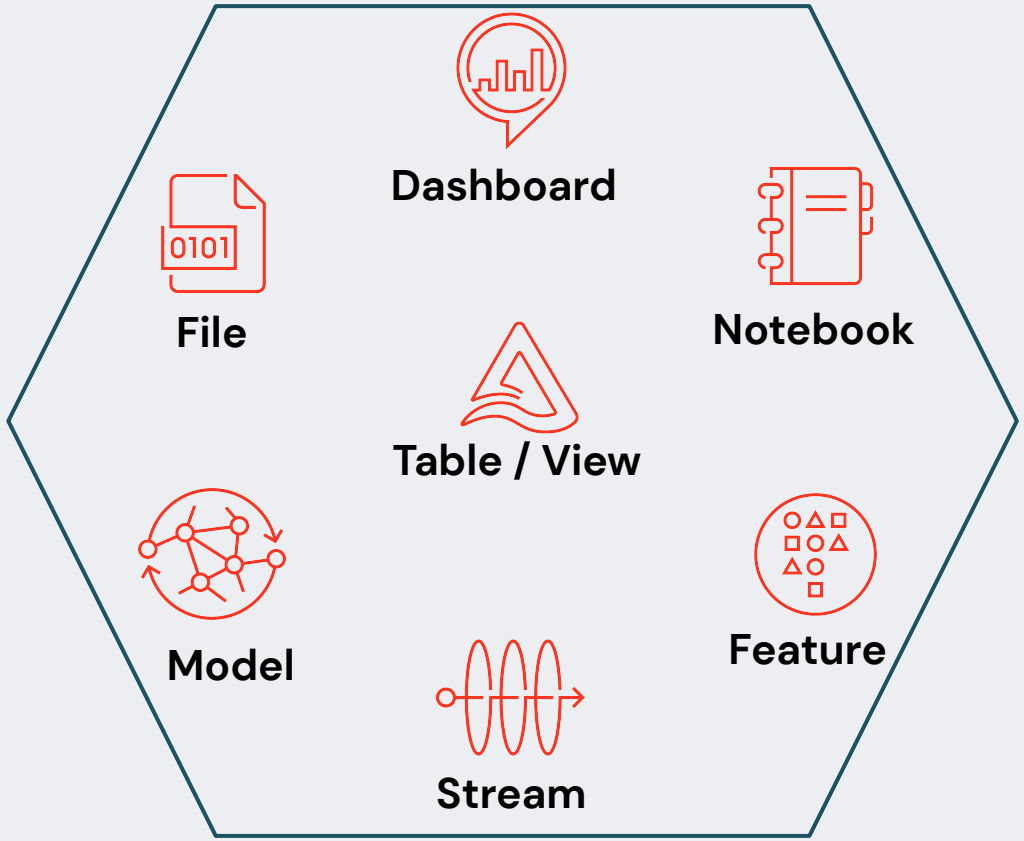
Multicloud

One consistent data platform across clouds

Data Mesh Architecture with Databricks

Data as a Product

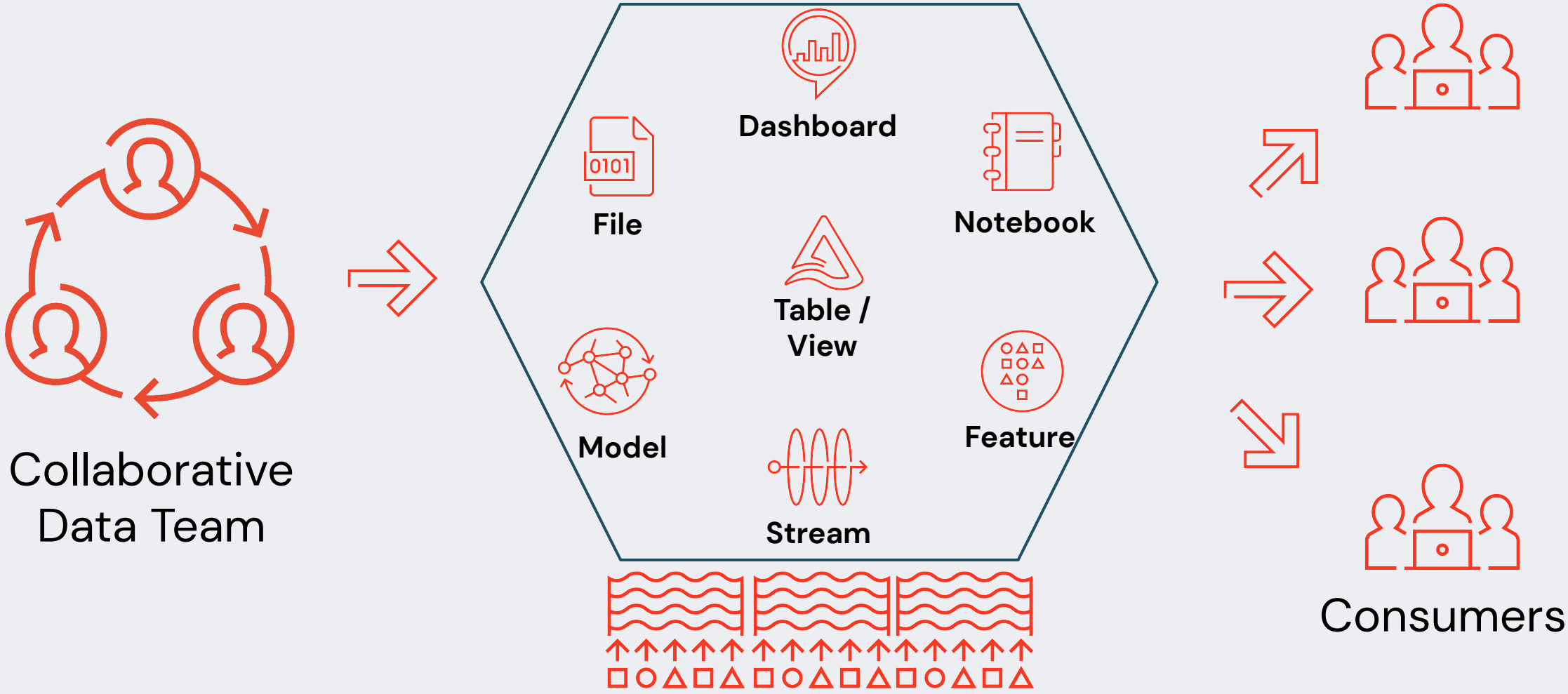
Domain data teams must apply product thinking



[How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh](#)
–Zhamak Dehghani

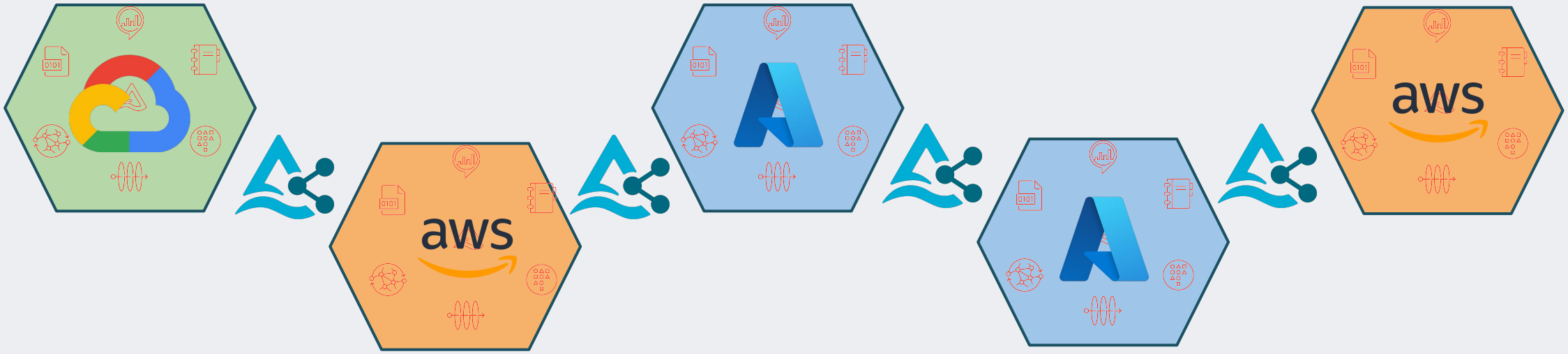
Databricks Workspace

A Home for Distributed Domain Data Teams



Self-Serve Data Platform

Identity, Governance, Discoverability, and Cost Control



Identity

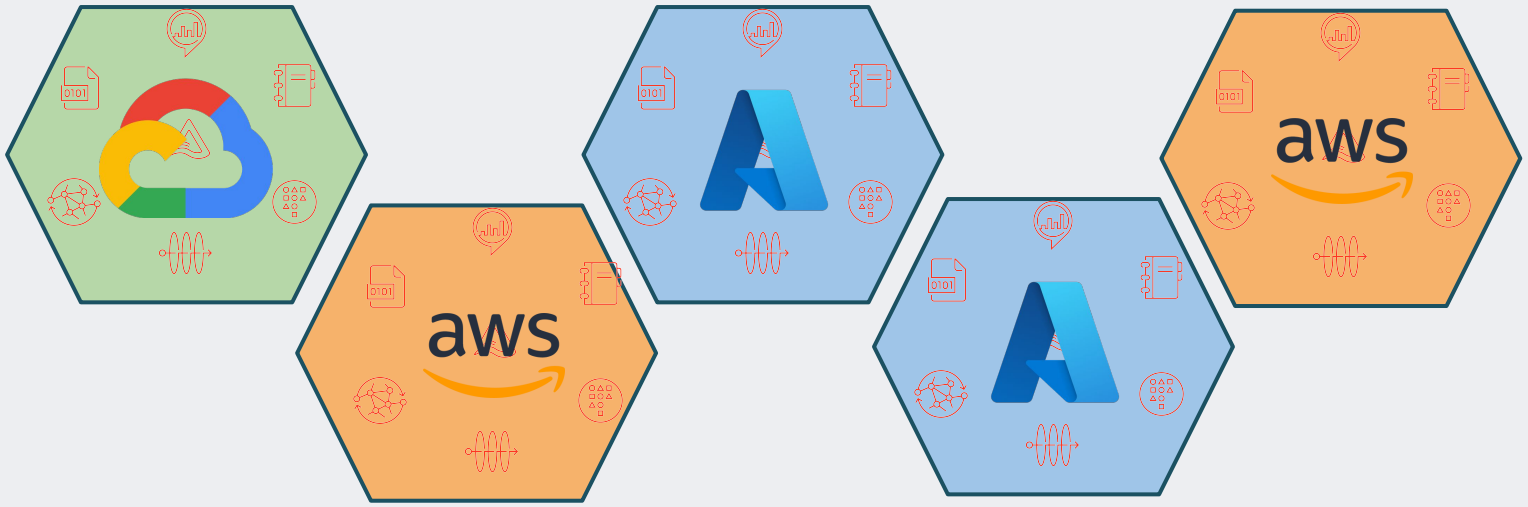
Audit

Catalog

Cost Control

External Data Sharing

Share data with any recipient; internal or external



Identity

Audit

Catalog

Cost Control

Power BI

Spark

pandas

tableau

...

Any Sharing Client

Delta Sharing Recipient

Data Mesh implementation

(Top 10 Pharmaceutical customer for Databricks)

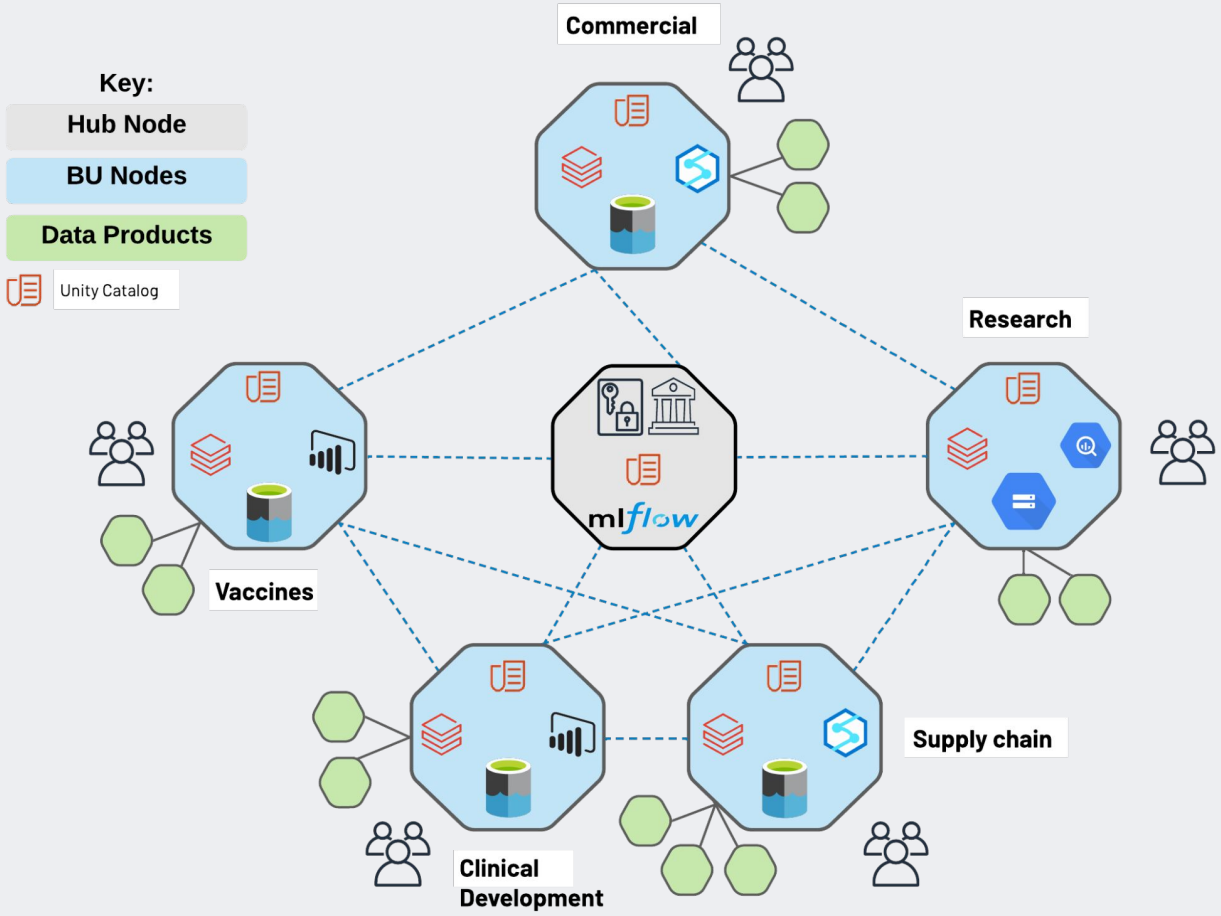
Key tenets governing the mesh architecture

Areas of focus for organizations looking to embrace data mesh

1. Plan for ***chargebacks*** upfront amongst different business units (BUs)
2. Plan for ***multi-cloud*** which is increasingly becoming pertinent for growing enterprises
3. Invest in ***automation*** which helps with faster adoption and reduced costs
4. Identify ***technology gaps*** early as different BUs will have varying maturity models
5. Invest in ***enterprise governance*** capabilities to enable easier collaboration amongst BUs
6. Realise data mesh is not just a ***technology arc*** but requires organisational and skills upgrade

A typical data mesh reference architecture

Collaborating with the cloud providers to deploy the mesh



Key highlights:

- **Centralized services** operating out of the core node enforcing standard security policies, access control, auditing, monitoring and cost control
- Mesh core node will enforce c**entral data cataloguing services** to discover data assets and manages **common data ingestion** pipelines and houses enterprise level datasets
- **BU nodes** set up in their **own subscriptions** with each BU node houses its own l**ake storage and Databricks workspace(s)**
- **Operating model** includes the platform group (platform and data ops) and the data node groups (domain ops and data product teams)
- Each ops team has a **well-defined set of responsibilities** both centrally as well as individually within the BUs

What has been achieved thus far?

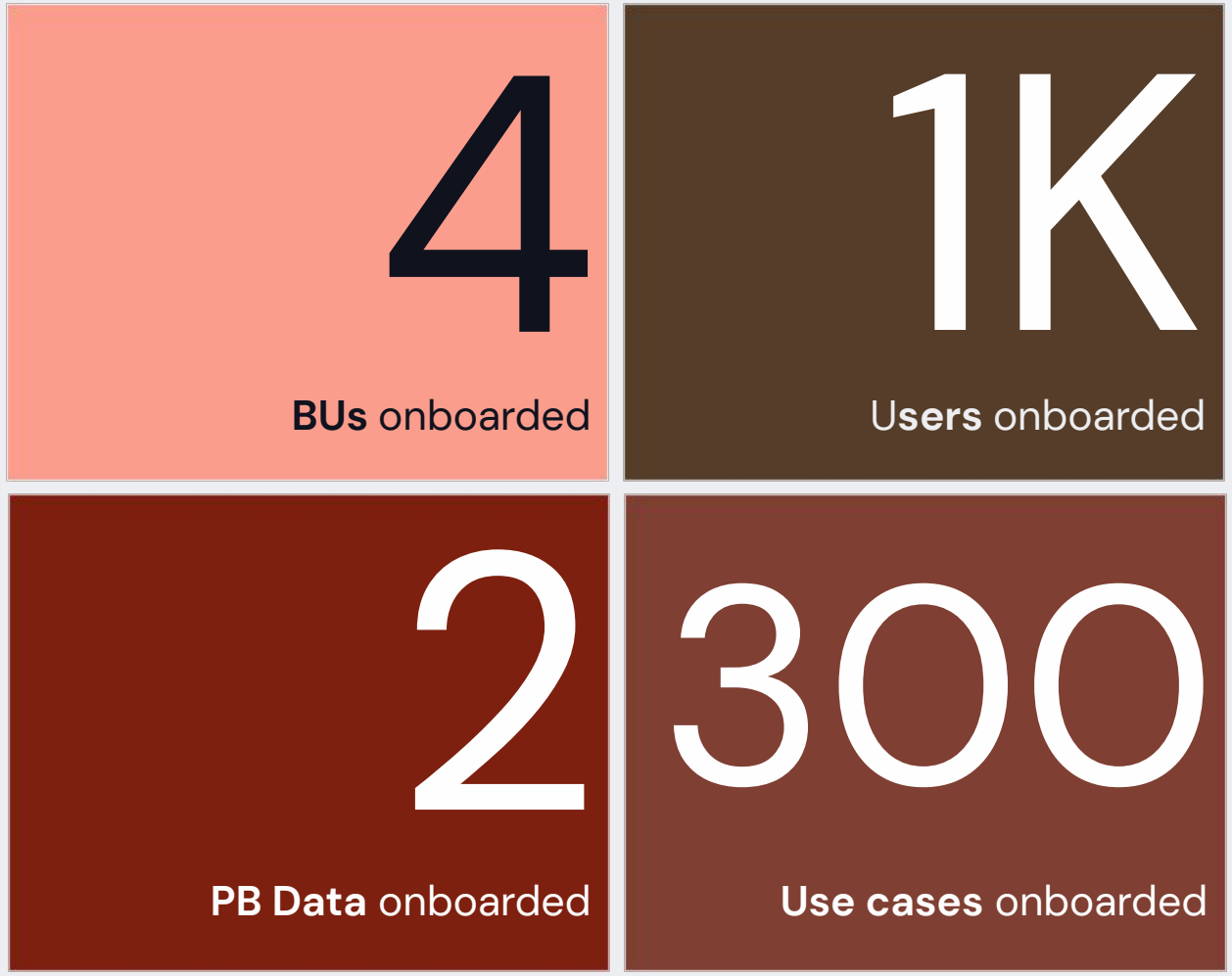
The road to data mesh is a long and continuous one…

What went right?

- Identifying t**echnology and platform champion(s) early** to navigate the organizational challenges and guide business stakeholders
- **Productization of data evangelised early** within the participating BUs with key stakeholders identified to champion the change
- Setting up **centralised services** which helped reduce the technical burden on individual BUs

What challenges lie ahead?

- Onboarding of BUs and use cases has slowed down owing to **organizational changes required** to work within the mesh architecture
- Multi-cloud mesh requires **consistent end user experience**
- Collaboration and sharing of data assets **requires new capabilities** to be onboarded to data mesh



Demo

Other Data Mesh Talks

Be Sure to Check out…

Automate Your Delta Lake or Practical Insights on Building Distributed Data Mesh

by Serge Smertin (Databricks)

Data Mesh Implementation Patterns

by Sankalan Bhattacharjee (Accenture), Ken Gravenor (McKesson)

Accelerating Hybrid Data Mesh Implementation

by Timur Mehmedbasic (Avanade)

Accidentally Building a Petabyte-Scale Cybersecurity Data Mesh in Azure With Delta Lake at HSBC

by Ryan Harris (HSBC)

Data Lakehouse and Data Mesh—Two Sides of the Same Coin

by Max Schultze (Zalando), Arif Wider (Thoughtworks)

DATA+AI
SUMMIT 2022

Thank you

Jason Pohl & Som Natarajan
Databricks