Data Lakehouse, Data Mesh, and Data Fabric

(the alphabet soup of data architectures)

James Serra

Data & AI Solution Architect

Microsoft, Federal Civilian

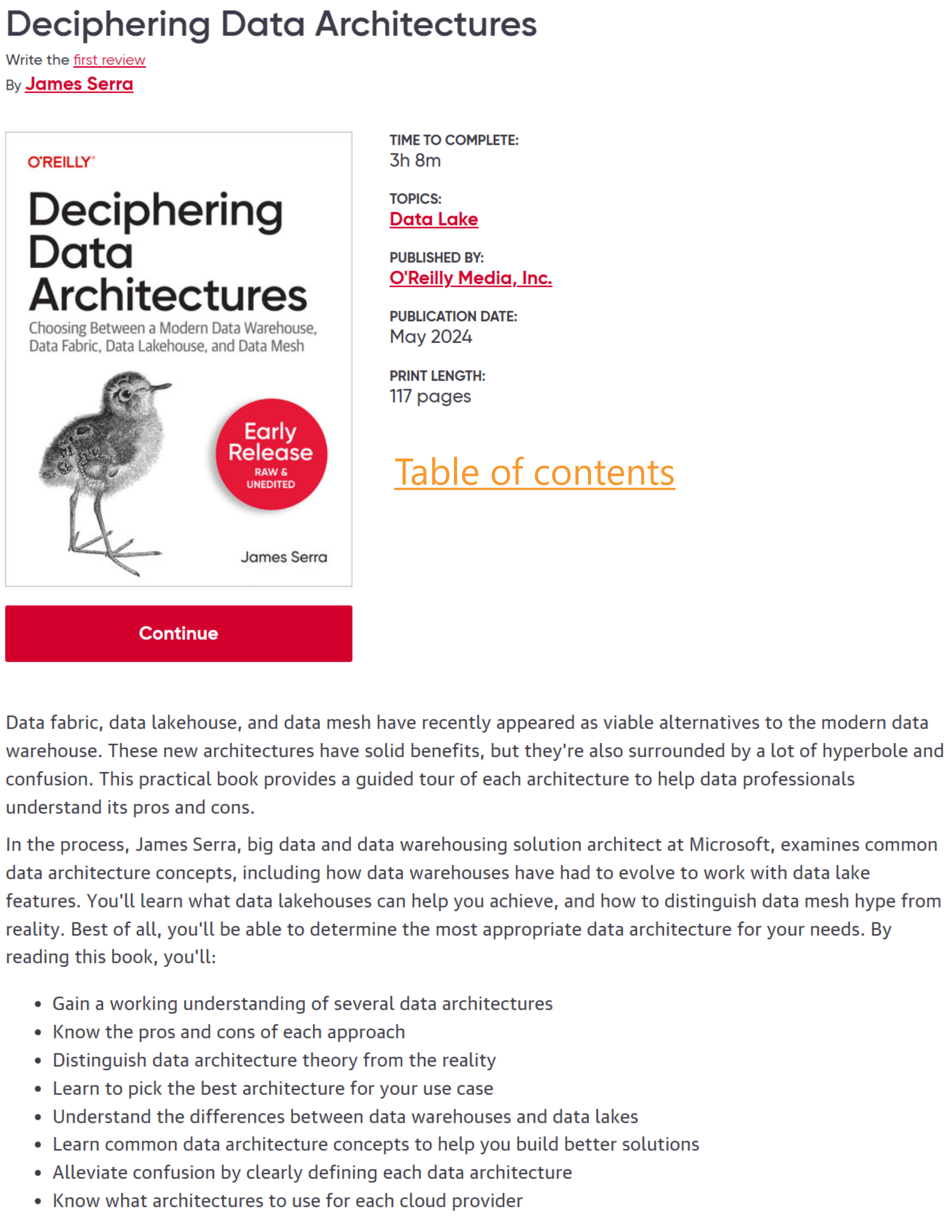
jamesserra3@gmail.com

Blog: JamesSerra.com

About Me

- Microsoft, Data & AI Solution Architect in Microsoft Federal Civilian
- At Microsoft for most of the last nine years as a Data & AI Architect , with a brief stop at EY
- In IT for 35 years, worked on many BI and DW projects
- Worked as desktop/web/database developer, DBA, BI and DW architect and developer, MDM architect, PDW/APS developer
- Been perm employee, contractor, consultant, business owner
- Presenter at PASS Summit, SQLBits, Enterprise Data World conference, Big Data Conference Europe, SQL Saturdays, Informatica World
- Blog at JamesSerra.com
- Former SQL Server MVP
- Author of book “Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh”

My upcoming book



Deciphering Data Architectures

Write the first review

By James Serra

O'REILLY

Deciphering Data Architectures

Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh

Early Release
RAW & UNEDITED

James Serra

Continue

TIME TO COMPLETE:
3h 8m

TOPICS:
Data Lake

PUBLISHED BY:
O'Reilly Media, Inc.

PUBLICATION DATE:
May 2024

PRINT LENGTH:
117 pages

Data fabric, data lakehouse, and data mesh have recently appeared as viable alternatives to the modern data warehouse. These new architectures have solid benefits, but they're also surrounded by a lot of hyperbole and confusion. This practical book provides a guided tour of each architecture to help data professionals understand its pros and cons.

In the process, James Serra, big data and data warehousing solution architect at Microsoft, examines common data architecture concepts, including how data warehouses have had to evolve to work with data lake features. You'll learn what data lakehouses can help you achieve, and how to distinguish data mesh hype from reality. Best of all, you'll be able to determine the most appropriate data architecture for your needs. By reading this book, you'll:

- Gain a working understanding of several data architectures
- Know the pros and cons of each approach
- Distinguish data architecture theory from the reality
- Learn to pick the best architecture for your use case
- Understand the differences between data warehouses and data lakes
- Learn common data architecture concepts to help you build better solutions
- Alleviate confusion by clearly defining each data architecture
- Know what architectures to use for each cloud provider

Table of contents

Seven chapters available now:
Deciphering Data Architectures (oreilly.com)

- Foundation
  - 1. Big Data (available)
  - 2. Types of Data Architectures (available)
  - 3. The Architecture Design Session (available)
- Common Data Architecture Concepts
  - 4. The Relational Data Warehouse (available)
  - 5. Data Lake (available)
  - 6. Data Storage Solutions and Processes (available)
  - 7. Approaches to Design (available)
  - 8. Approaches to Data Modeling
  - 9. Approaches to Data Ingestion
- Data Architectures
  - 10. Modern Data Warehouse (MDW)
  - 11. Data Fabric
  - 12. Data Lakehouse
  - 13. Data Mesh Foundation
  - 14. Data Mesh Adoption
- People, Process, and Technology
  - 15. People and process
  - 16. Technologies

Agenda

- Data Warehouse
- Data Lake
- Modern Data Warehouse
- Data Fabric
- Data Lakehouse
- Data Mesh

Note: These are James Serra's opinions and not that of Microsoft!

What is a Data Warehouse and why use one?

(or, why do we need a copy of the source data?)

A relational data warehouse is where you store data from multiple data sources to be used for historical and trend analysis reporting. It acts as a central repository for many subject areas and contains the "single version of truth". It is NOT to be used for OLTP applications.

Reasons for a data warehouse:

- Reduce stress on production system
- Optimized for read access, sequential disk scans
- Integrate many sources of data
- Keep historical records (no need to save hardcopy reports)
- Restructure/rename tables and fields, model data
- Protect against source system upgrades
- Use Master Data Management, including hierarchies
- No IT involvement needed for users to create reports
- Improve data quality and plugs holes in source systems
- One version of the truth
- Easy to create BI solutions on top of it (i.e. Power BI tabular model)
- Don’t need to provide security access for many users to the production systems
- Make better business decisions by getting greater insights into your company

Why You Need a Data Warehouse

What is a data lake and why use one?

A schema-on-read storage repository that holds a vast amount of raw data in its native format until it is needed.

Top reasons for a data lake:

- Store data with no modeling – “Schema on read”
- Frees up expensive enterprise data warehouse (EDW) resources for queries instead of using EDW resources for transformations (avoiding user contention)
- Quick user access to data for power users/data scientists (allowing for faster ROI)
- Data exploration to see if data valuable before writing ETL and schema for relational database, or use for one-time report
- Extreme performance for transformations by having multiple compute options each accessing different folders containing data

Data Lake with Data warehouse use cases

Data Lake

Staging & preparation

- Data scientists/Power users
- Batch processing
- Data refinement/cleaning
- ETL workloads
- Store older/backup data
- Sandbox for data exploration
- One-time reports
- Quick access to data
- Don’t know questions

Data Warehouse

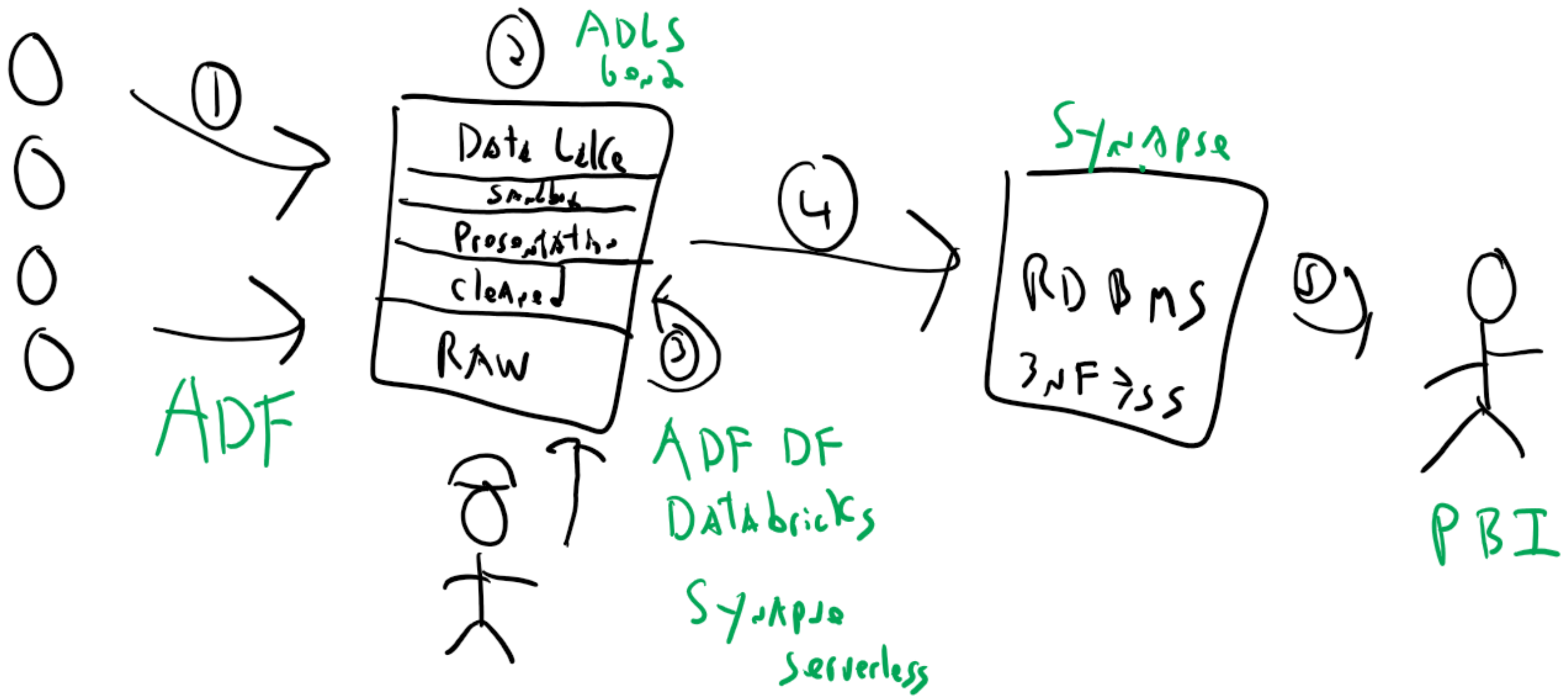
Serving, Security & Compliance

- Business people
- Low latency
- Complex joins
- Interactive ad-hoc query
- High number of users
- Additional security
- Large support for tools
- Dashboards
- Easily create reports (Self-service BI)
- Know questions

Modern Data Warehouse

Modern Data Warehouse (MDW)

1) Ingest
2) Store
3) Transform
4) Model
5) Visualize/ML



Data Fabric

A data fabric is a term used to describe the architecture of taking disparate systems and weaving them together, like fabric, to create a consistent layer on top of an organization’s data.

Data Fabric adds to a modern data warehouse:

- Data access
- Data policies
- Metadata catalog/Lineage
- Master Data Management (MDM)
- Data virtualization
- Real-time processing
- Data scientist tools
- APIs
- Building blocks/Services
- Products

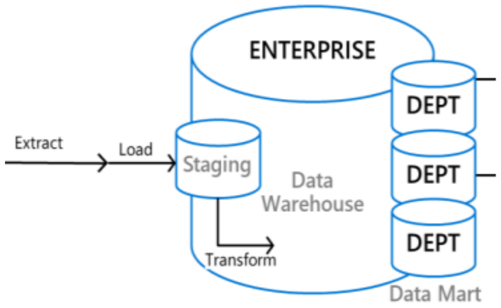
Bottom line: Additional technology to source more data, secure it, and make it available

Data Fabric defined

Data Lakehouse

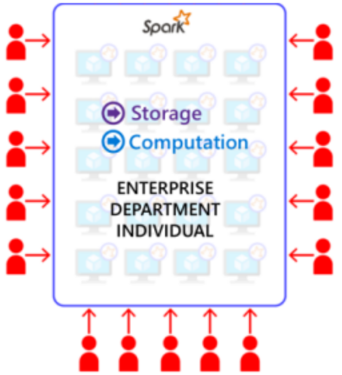
Late 1980s

Data Warehouse



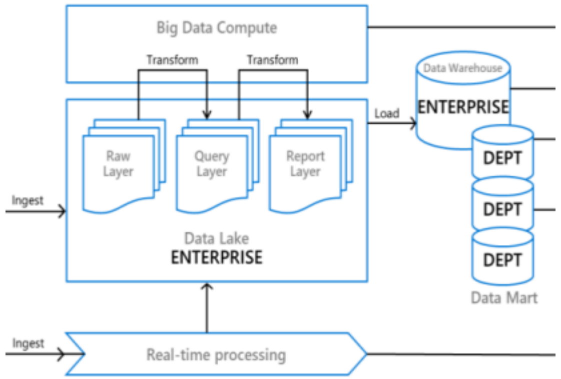
Late 2000s

Data Lake



Mid 2010s

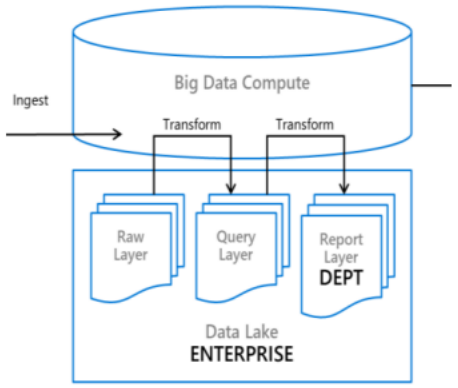
Cloud Data Platform



Modern Data Warehouse

2020

Data Lakehouse



Delta Lake

Delta Lake

A transactional storage software layer that runs on top of cloud storage (ADLS Gen2)

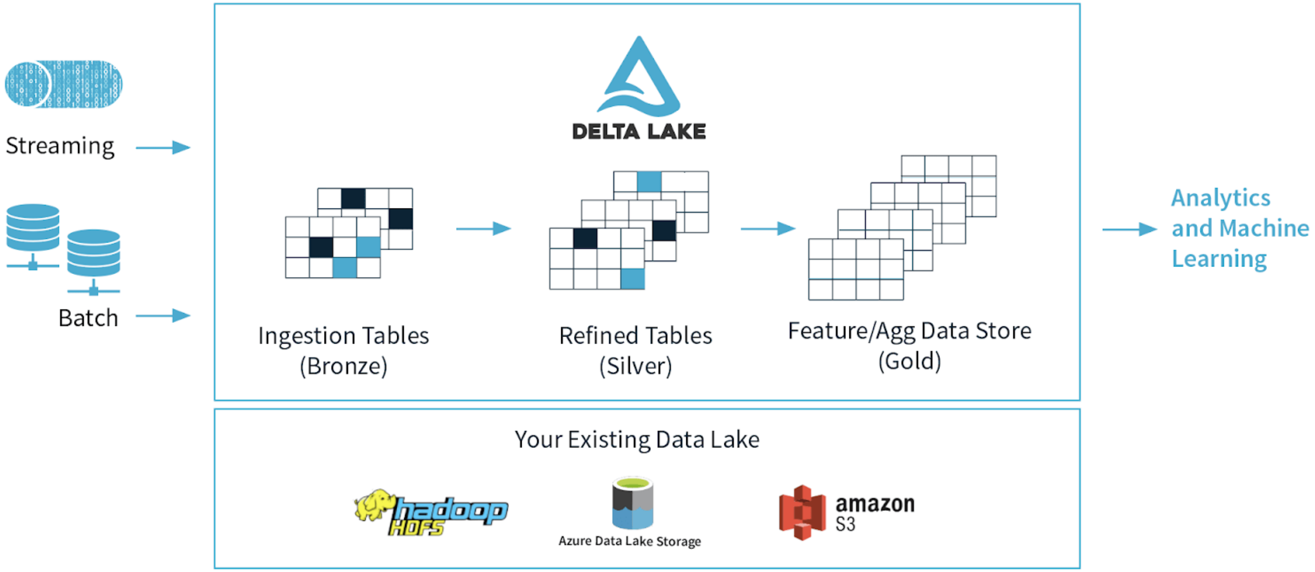
Top features:

- ACID transactions
- Time travel (data versioning enables rollbacks, audit trail)
- Streaming and batch unification
- Schema enforcement
- Supports commands DELETE, UPDATE, and MERGE
- Performance improvements (Z-Order)
- Solve “small files” problem
  via OPTIMIZE command (compact/merge)

```
df.write.format("delta").save(delta_table_path)
```

instead of

```
df.write.format(“parquet").save(delta_table_path)
```



Databricks Delta Lake

Use cases for Data Lakehouse

Today’s data architectures commonly suffer from four problems:

- Reliability: Keeping the data lake and warehouse consistent
- Data staleness: Data in warehouse is older
- Limited support for advanced analytics: Top ML systems don’t work well on warehouses
- Total cost of ownership: Extra cost for data copied to warehouse

Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics

Opening a can of worms

Concerns skipping relational database

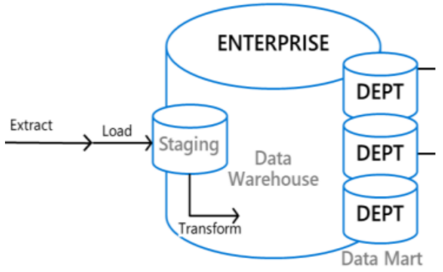
- Speed: Relational database *queries* faster, especially Massively Parallel Processing (MPP) with clustered columnstore indexes, statistics, caching, query plans, materialized views; joining tables
- Security: No row-level security (RLS), column-level security, dynamic data masking
- Complexity: Metadata separate from data, file-based world
- Concurrency: Multiple reads of a file at the same time
- Missing features: Referential integrity, TDE, workload management, MDM, ACID against multiple tables; using certain features lock you into Spark
- Having to use Spark SQL instead of T-SQL (if using Databricks)
- Products must add delta lake support in order to use it
- People are used to using a relational database (forced metadata layer)

Azure Synapse and Microsoft Fabric: starting to see data lake only solutions because can use T-SQL, Power BI (speed, RLS), cost savings with Serverless

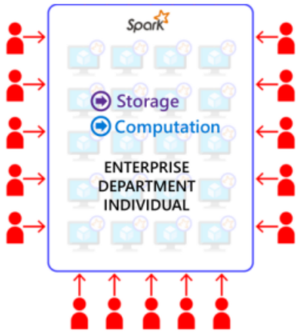
[Data Lakehouse & Synapse]

Data Mesh

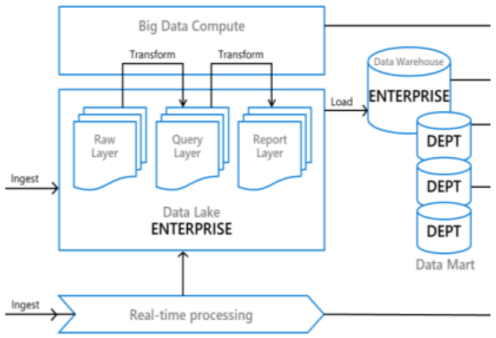
Late 1980s
Data Warehouse



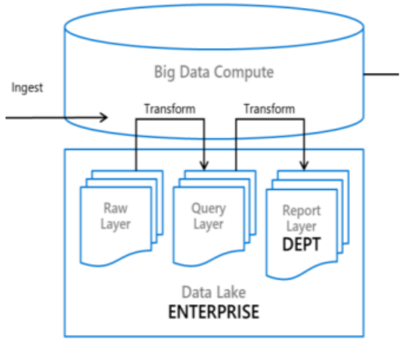
Late 2000s
Data Lake



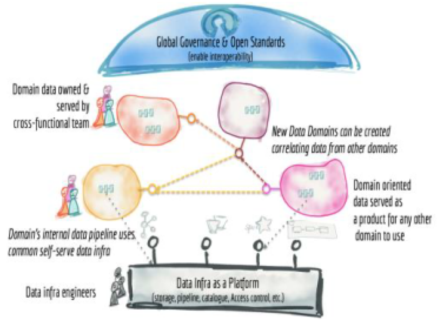
Mid 2010s
Cloud Data Platform



2020
Data Lakehouse



2021
Data Mesh??


Global Governance & Open Standards
(enable interoperability)
Domain data owned & served by cross-functional team
New Data Domains can be created correlating data from other domains
Domain oriented data served as a product for any other domain to use
Domain's internal data pipeline uses common self-serve data infra
Data infra engineers
Data Infra as a Platform
(storage, pipeline, catalogue, Access control, etc.)

Centralization

Decentralization

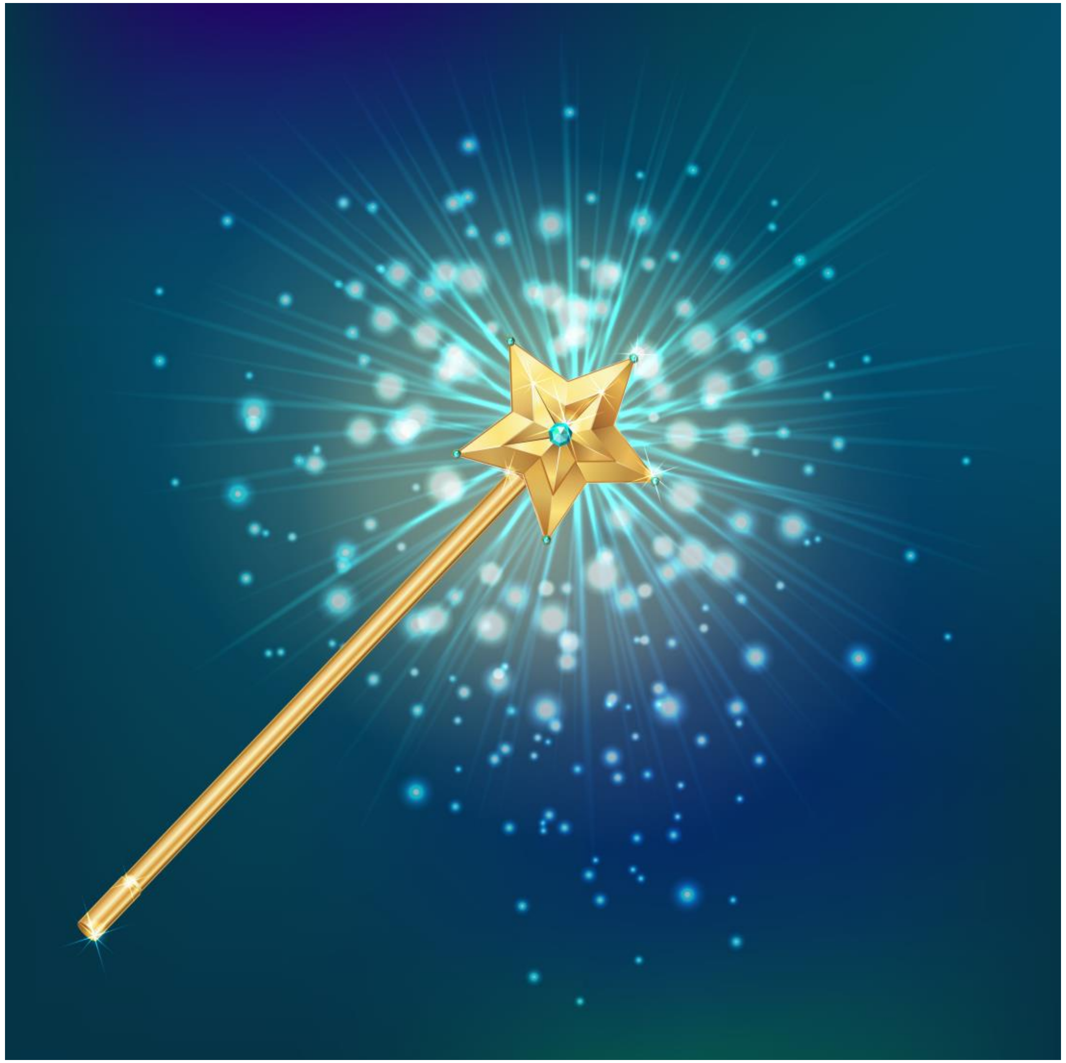
Data Mesh

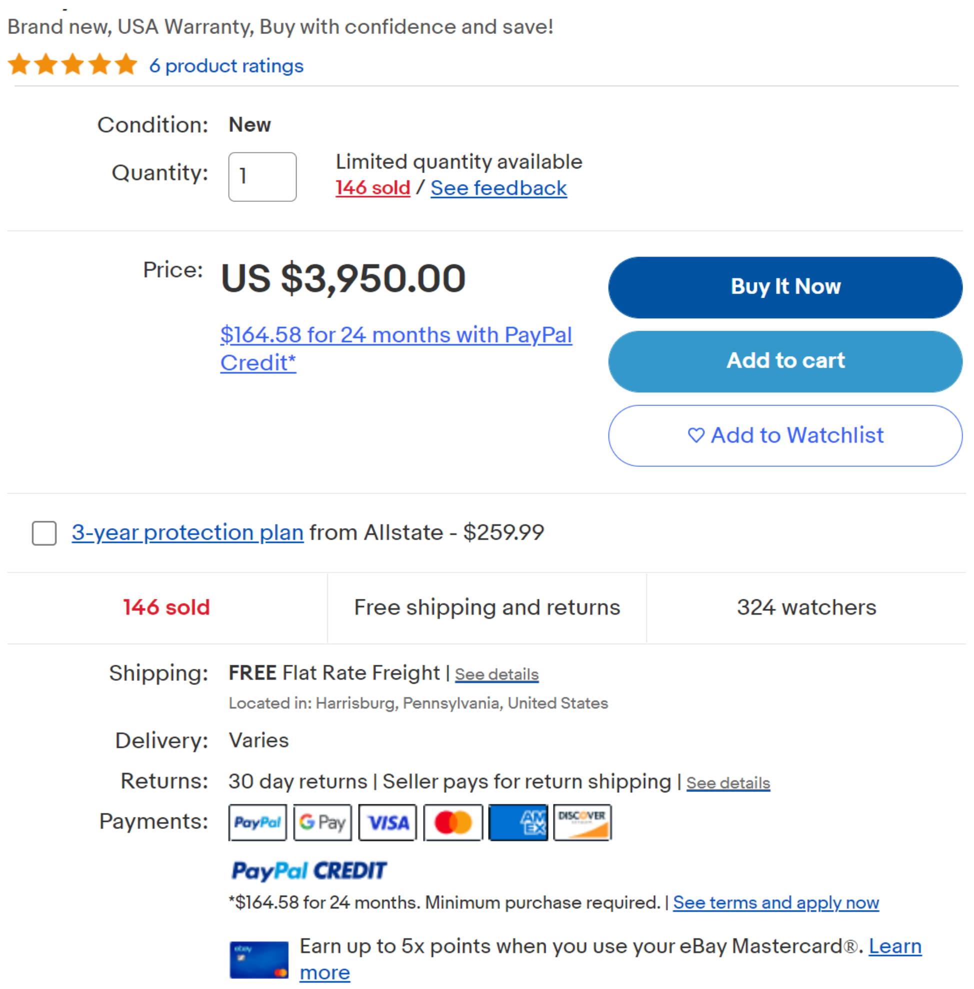
Data Mesh is a concept, not technology

It is an organizational and cultural shift

Credit to Zhamak Dehghani

Data Mesh is not a magic wand that you can buy on ebay





Brand new, USA Warranty, Buy with confidence and save!

★★★★★ 6 product ratings

Condition:	New	
Quantity:	1	Limited quantity available 146 sold / See feedback

Price: US $3,950.00

$164.58 for 24 months with PayPal Credit*

□ 3-year protection plan from Allstate - $259.99

146 sold	Free shipping and returns	324 watchers

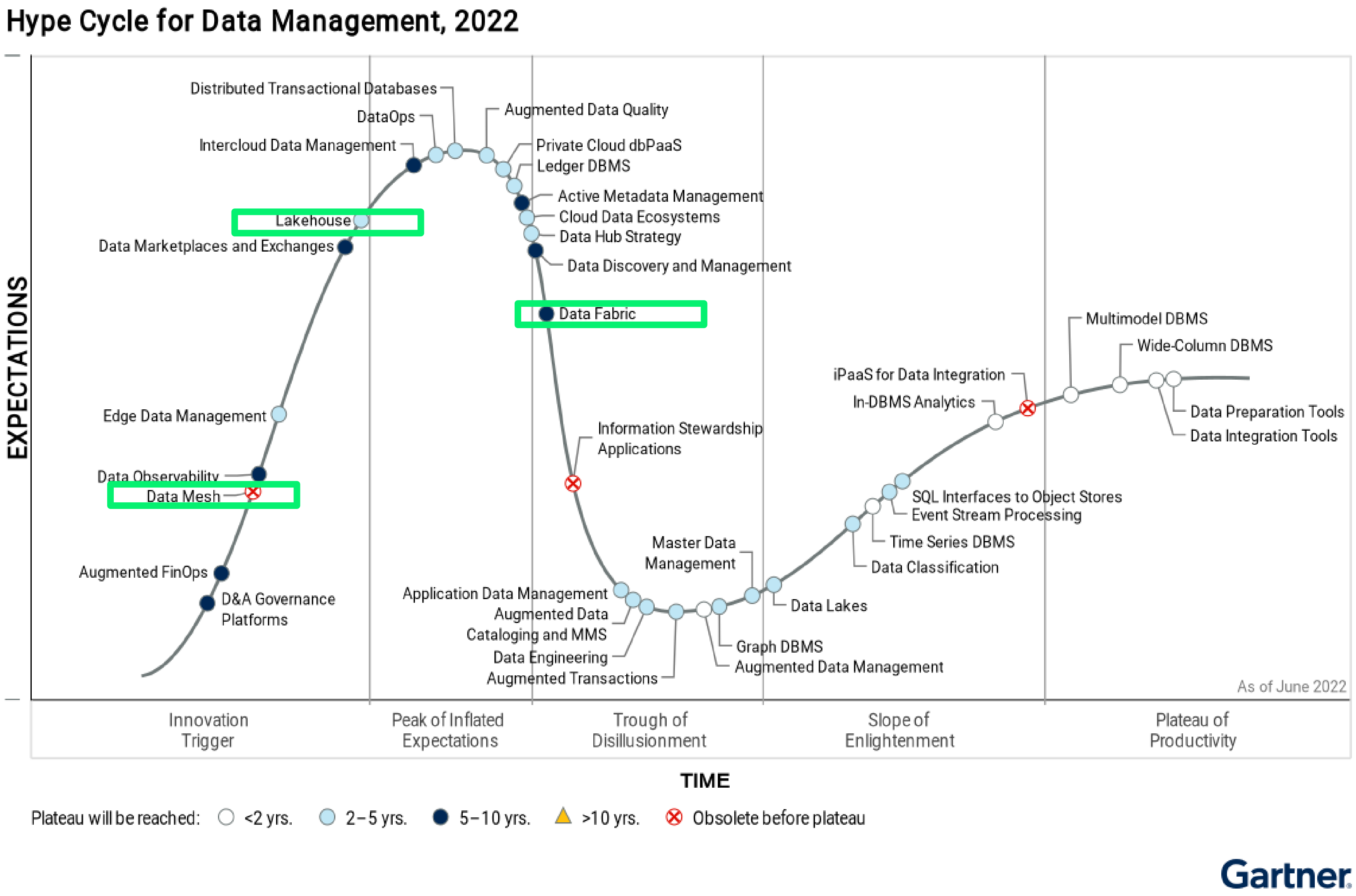
Shipping:	**FREE** Flat Rate Freight | See details Located in: Harrisburg, Pennsylvania, United States
Delivery:	Varies
Returns:	30 day returns | Seller pays for return shipping | See details
Payments:	PayPal G Pay VISA AMEX DISCOVER

PayPal CREDIT

*$164.58 for 24 months. Minimum purchase required. | See terms and apply now

Earn up to 5x points when you use your eBay Mastercard®. Learn more

Data Mesh in hype cycle



Data Mesh

Analysis By: Mark Beyer, Ehtisham Zaidi, Robert Thanaraj

Benefit Rating: Low

Market Penetration: 1% to 5% of target audience

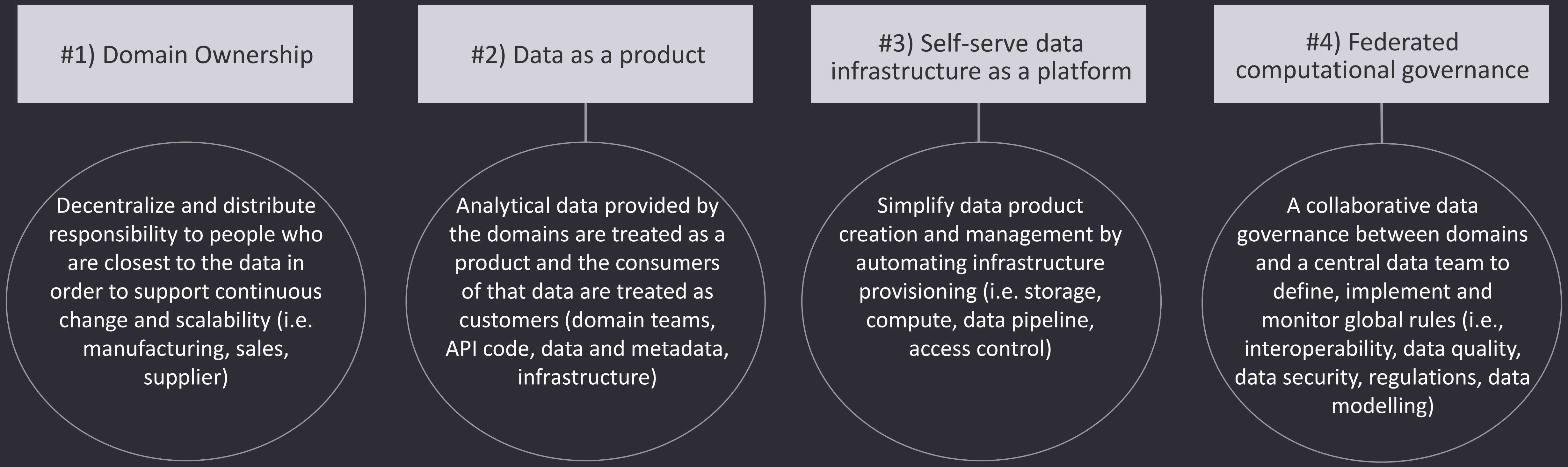
Maturity: Embryonic

Gartner Reprint

Data Mesh - Overview

A data mesh is a decentralized approach to managing data, where multiple teams within a company are responsible for their own data, promoting collaboration and flexibility. By implementing data mesh principles, the quality and accuracy of data can be enhanced, resulting in increased trust among businesses to utilize data more extensively for informed decision-making.

Data Mesh Principles



#1) Domain Ownership

Decentralize and distribute responsibility to people who are closest to the data in order to support continuous change and scalability (i.e. manufacturing, sales, supplier)

#2) Data as a product

Analytical data provided by the domains are treated as a product and the consumers of that data are treated as customers (domain teams, API code, data and metadata, infrastructure)

#3) Self-serve data infrastructure as a platform

Simplify data product creation and management by automating infrastructure provisioning (i.e. storage, compute, data pipeline, access control)

#4) Federated computational governance

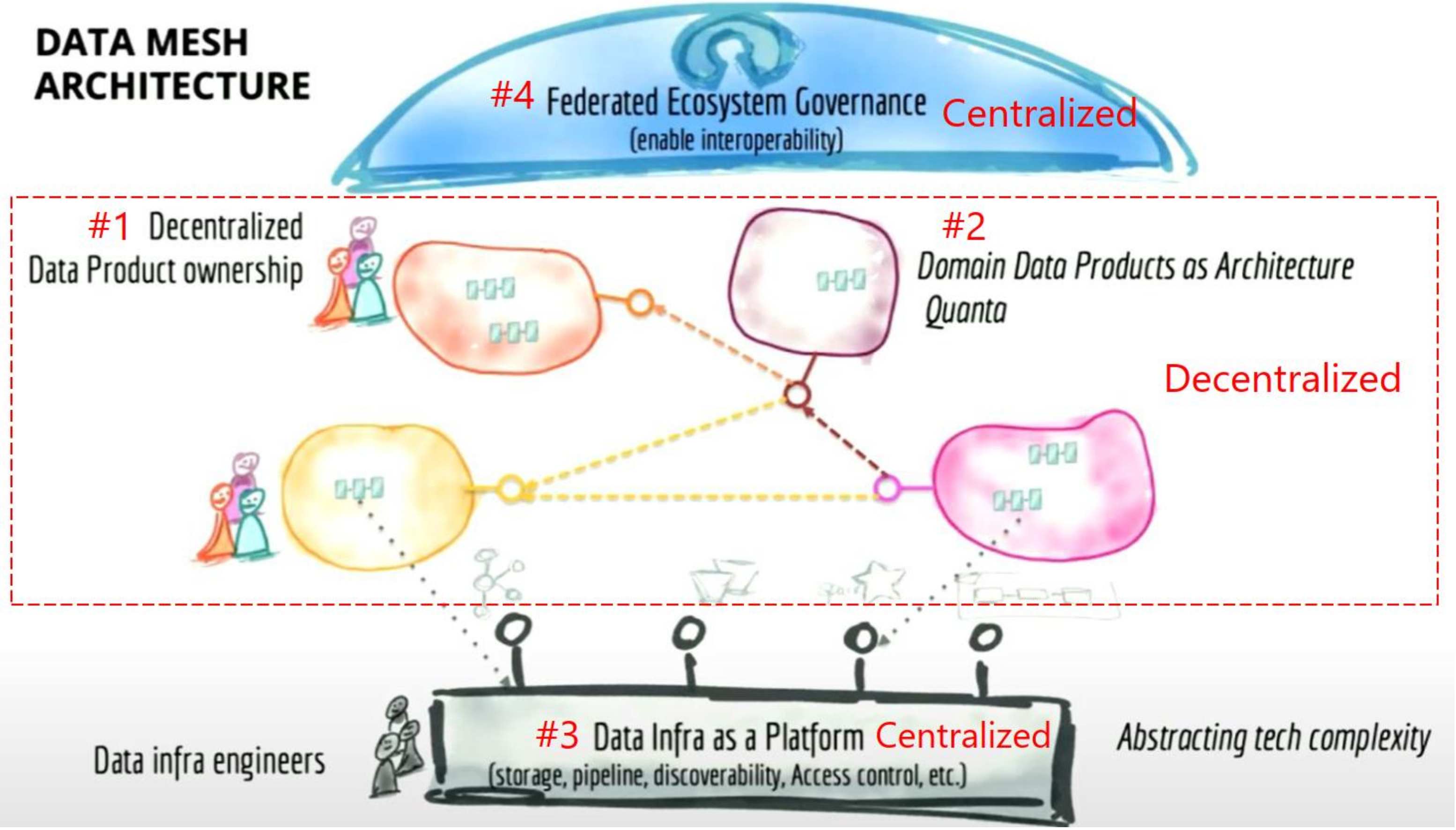
A collaborative data governance between domains and a central data team to define, implement and monitor global rules (i.e., interoperability, data quality, data security, regulations, data modelling)

Data Mesh

Data Mesh is a concept, not a product

It's a mindset shift where you go from:

- Centralized ownership to decentralized ownership
- Pipelines as first-class concern to domain data as first-class concern
- Data as a by-product to data as a product
- A siloed data engineering team to cross-functional domain-data teams
- A centralized data lake/warehouse to an ecosystem of data products


DATA MESH ARCHITECTURE
#4 Federated Ecosystem Governance
(enable interoperability)
Centralized
#1 Decentralized Data Product ownership
#2 Domain Data Products as Architecture Quanta
Decentralized
Data infra engineers
#3 Data Infra as a Platform Centralized
(storage, pipeline, discoverability, Access control, etc.)
Abstracting tech complexity

Credit to Zhamak Dehghani, Slack: data-mesh-learning.slack.com

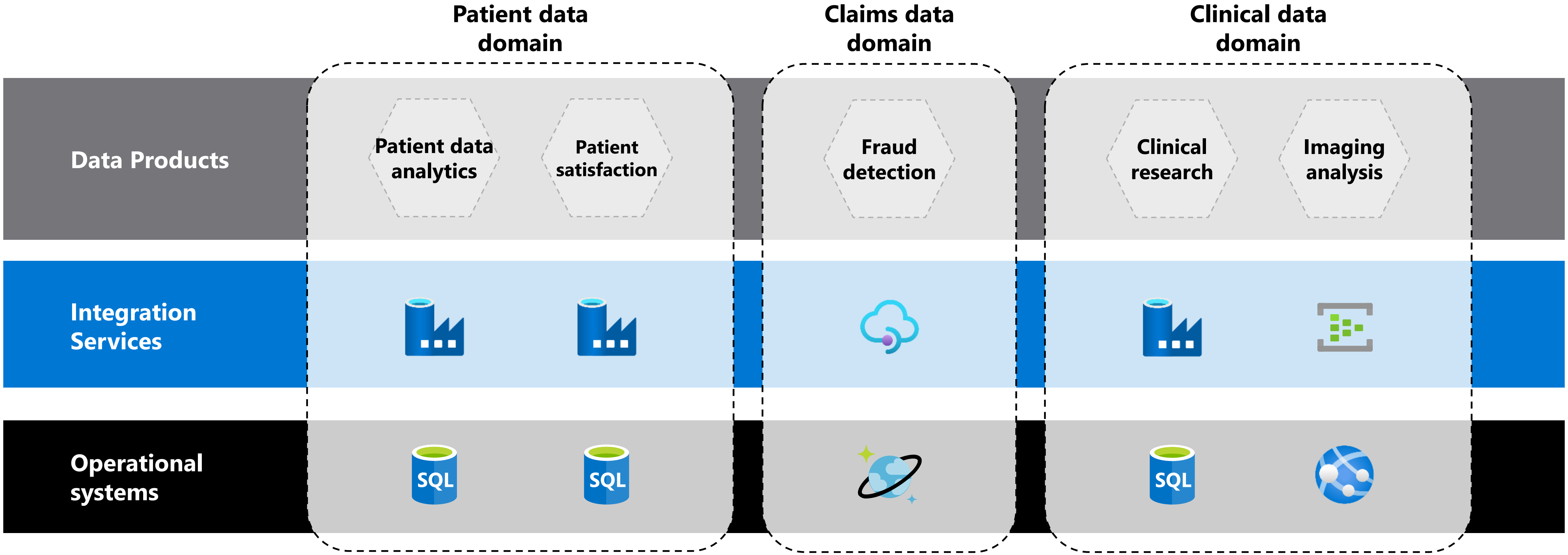
Use cases for Data Mesh

Data mesh tries to solve four challenges with a centralized data lake/warehouse:

- Lack of ownership: who owns the data – the data source team or the infrastructure team?
- Lack of quality: the infrastructure team is responsible for quality but does not know the data well
- Organizational scaling: the central team becomes the bottleneck, such as with an enterprise data lake/warehouse
- Technical scaling: current big data solutions can’t keep up with additional data requirements

What are Data Domains?

- A domain is simply a collection of people typically organized around a ***common business purpose or business capability***. Use ***domain-driven design (DDD)***
- Create and serve data products to other domains and end users, independently from other domains
- Ensure data is accessible, usable, available, and meets the quality criteria defined (abide by a contract)
- Evolve data products based on user feedback. Retire data products when they become irrelevant



Example healthcare domains and products within them

Patient data domain

- Patient data analytics
- Patient engagement
- Population Health Management
- Clinical decision support systems
- Patient matching and ID resolution
- Clinical research
- Patient outcome prediction
- Patient satisfaction

Clinical data domain

- Clinical research
- Clinical decision support
- Clinical analytics
- Clinical trial management
- Imaging analysis
- Disease registries
- Real-world evidence
- Clinical trial matching

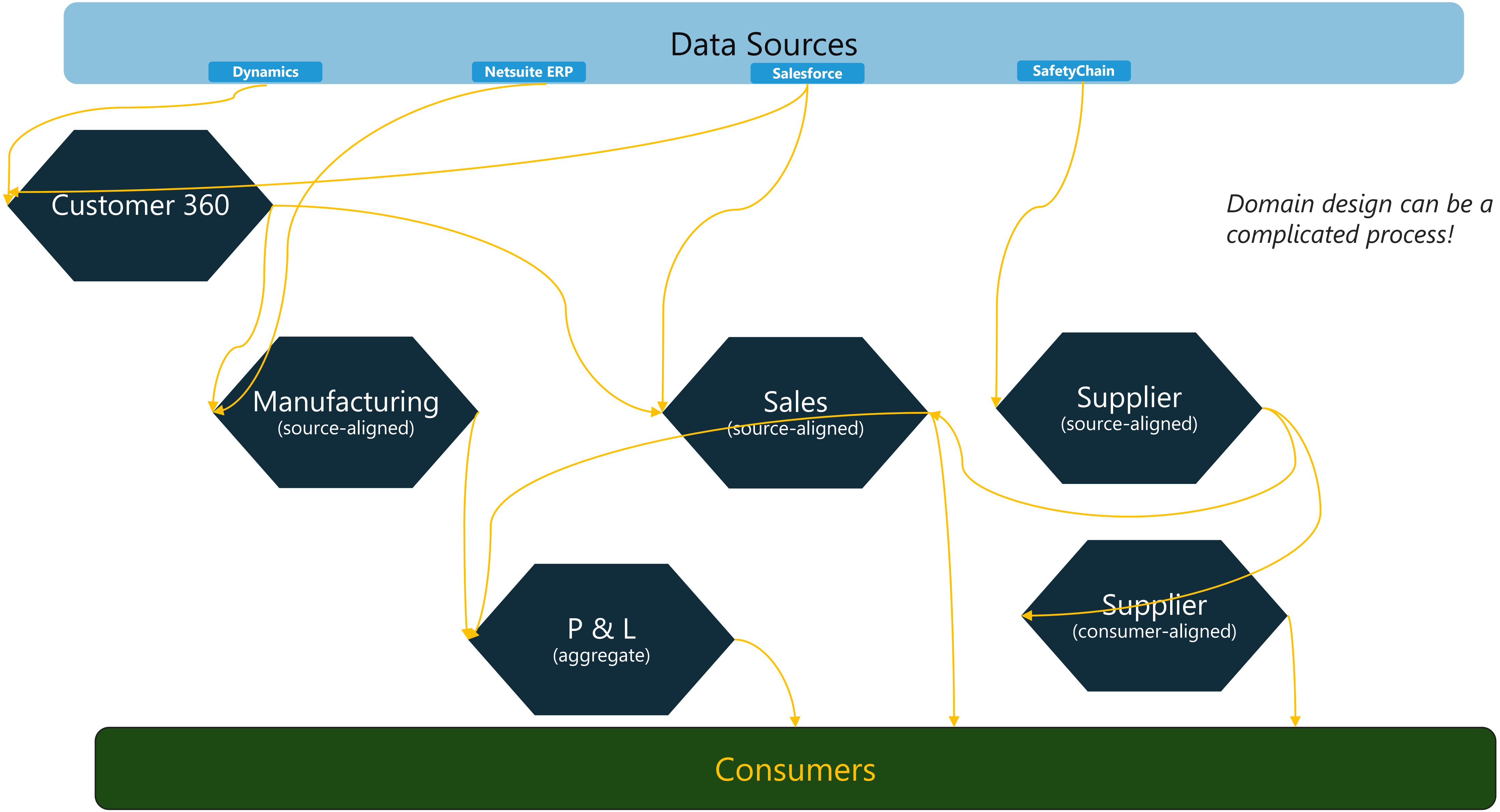
Claims data domain

- Claims analytics
- Claims management
- Fraud detection
- Payment processing
- Patient financial management
- Provider network analysis
- Claims denial management
- Health plan selection

Public health data domain

- Disease surveillance
- Health equity
- Population health management
- Environmental health tracking
- Public health research
- Infectious disease modeling
- Health behavior change
- Chronic disease management

Data Mesh – Logical Architecture



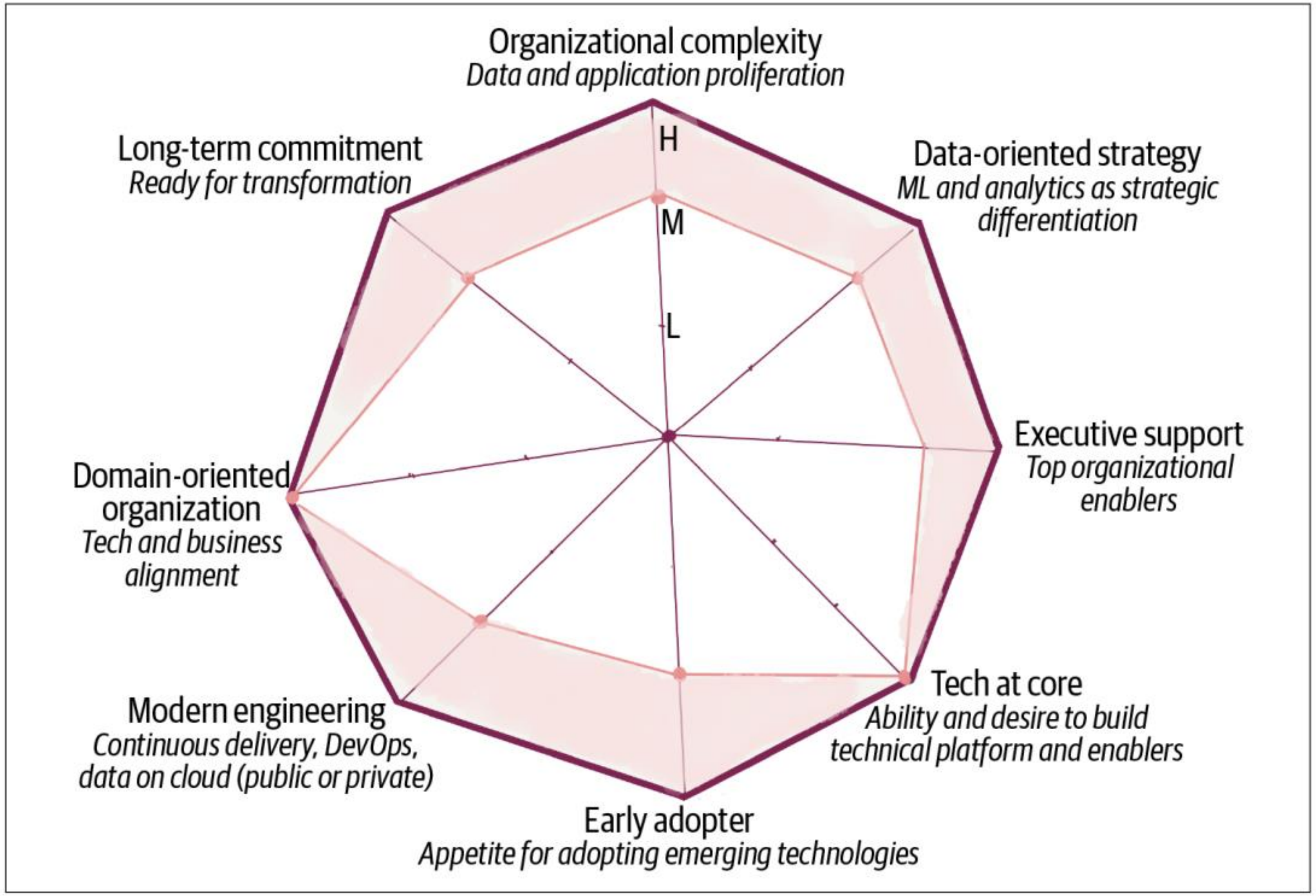
Domain design can be a long and complicated process!

Concerns with Data Mesh

- No standard definition of a data mesh
- Huge investment in organizational change and technical implementation
- Performance of combining data from multiple domains
- Duplication of data for performance reasons
- Getting quality engineering people for each domain
- Inconsistent technical implementations for the domains
- Domains don't want to wait for a data mesh
- Need incentives for each domain to counter extra work
- Self-serve approach of data requests could be challenging
- Duplication of data and ingestion platform
- Creation of data silos for domains not able to join data mesh
- Not seeing the big picture for combing data

Data Mesh: Centralized vs decentralized data architecture
Data Mesh: Centralized ownership vs decentralized ownership

Should you adopt data mesh today?

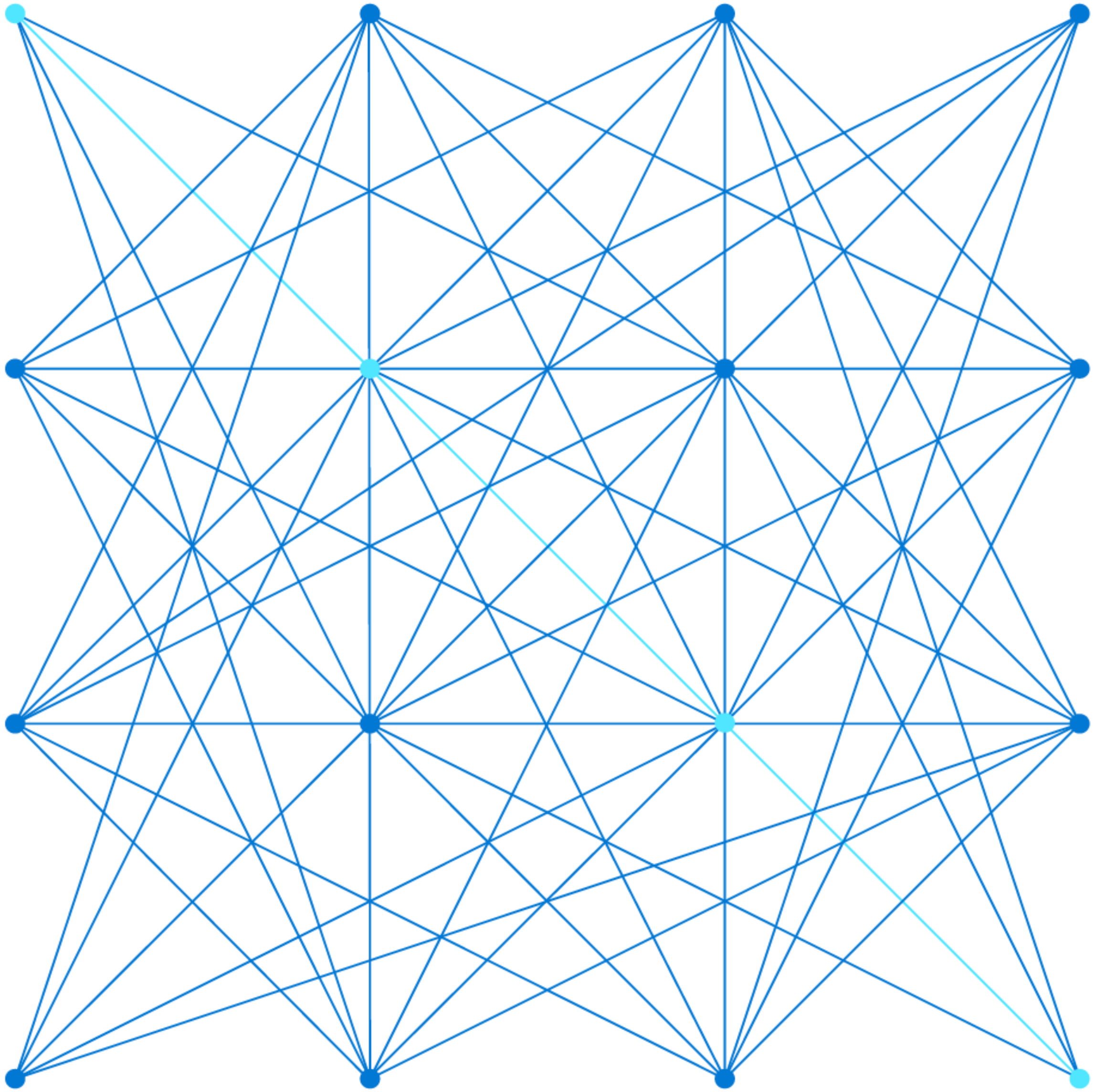


Need to score medium or high in ALL categories

From book "Data Mesh – Delivering Data-Driven Value at Scale" by Zhamak Dehghani, Page 272

Microsoft Azure

Data Mesh on Azure



Cloud Scale Analytics

Cloud Scale Analytics is an architecture approach and reference implementation that enables effective construction and operationalization of landing zones on Azure, at scale and aligned with Azure Roadmap and Cloud Adoption Framework.

What is Cloud Scale Analytics?

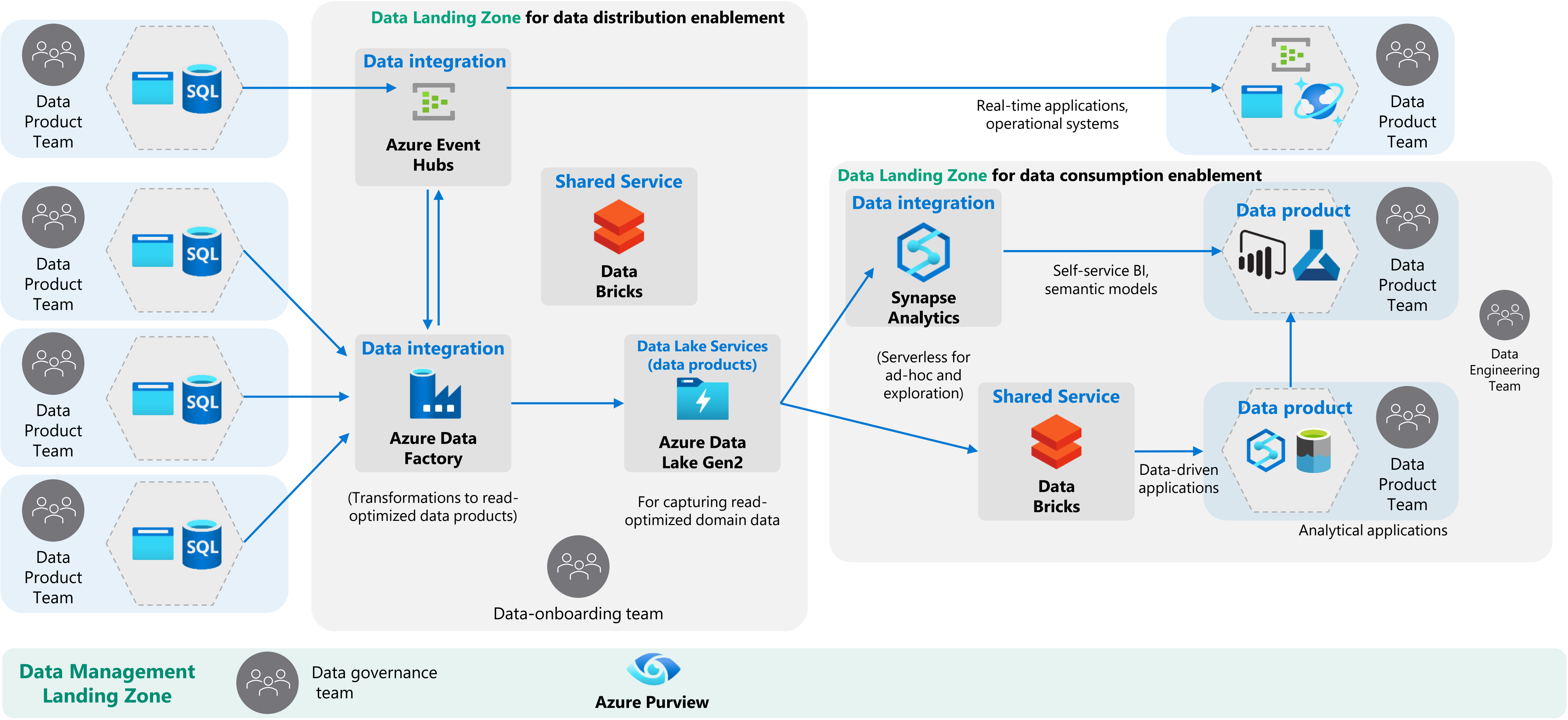
A scalable analytics framework designed to enable customers building a data platform.

- Follows a hub-and-spoke model for centralized governance and controls while allowing logical or functional business units to operate individual Landing Zones to facilitate their analytics workloads. The centralized Data Management Subscription allows for collaboration and information sharing without compromising security
- *Supports multiple topologies ranging across Data Centric, Lakehouse, Data Fabric and Data Mesh*
- Based on inputs from the Microsoft Program Group and a diverse international group of specialists working with a range of customers
- Separate guidance tailored to Small-Medium and Large enterprises
- ~80% prescribed viewpoint with 20% client customization
- More info: Cloud-scale analytics - Microsoft Cloud Adoption Framework for Azure - Cloud Adoption Framework | Microsoft Learn

Helps you to create:

Data Landing Zone
Data Management Landing Zone

Example reference architecture for governed mesh; using landing zones to optimize distribution and consumption of data



Data Product Team

Data Product Team

Data Product Team

Data Product Team

Data Landing Zone for data distribution enablement

Data integration

Azure Event Hubs

Shared Service

Data Bricks

Data integration

Azure Data Factory

(Transformations to read-optimized data products)

Data Lake Services (data products)

Azure Data Lake Gen2

For capturing read-optimized domain data

Data-onboarding team

Real-time applications, operational systems

Data Product Team

Data Landing Zone for data consumption enablement

Data integration

Synapse Analytics

(Serverless for ad-hoc and exploration)

Self-service BI, semantic models

Data product

Data Product Team

Data Engineering Team

Shared Service

Data Bricks

Data-driven applications

Data product

Data Product Team

Analytical applications

Data Management Landing Zone

Data governance team

Azure Purview

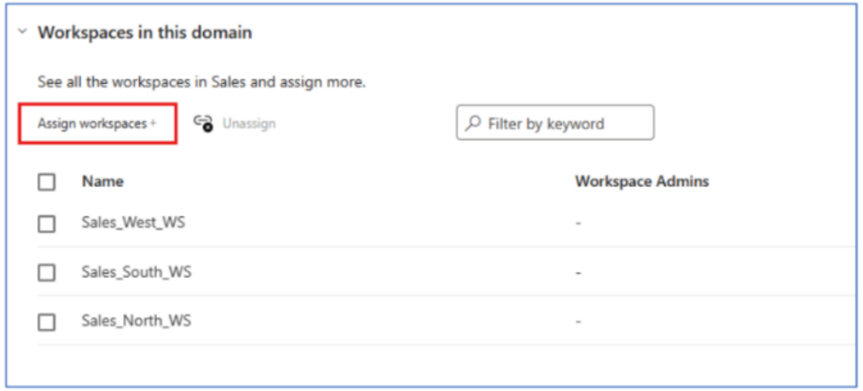
Microsoft Fabric and Data Mesh

- Can logically organize data into domains
- PBI workspaces are associated with domains – all the items in the workspace become part of the domain (they receive a domain attribute as part of their metadata
- Data consumers can filter and find content by domain
- Future releases will enable federated governance, which means some of the governance currently controlled at the tenant level will move to domain-level control
- Use low-code to make it easier for domain teams to build solution
- OneLake technology to help

Data Mesh four principles:

1. Domain ownership: partial
2. Data as a product: very little
3. Self-serve data infrastructure as a platform: some
4. Federated computational governance: future

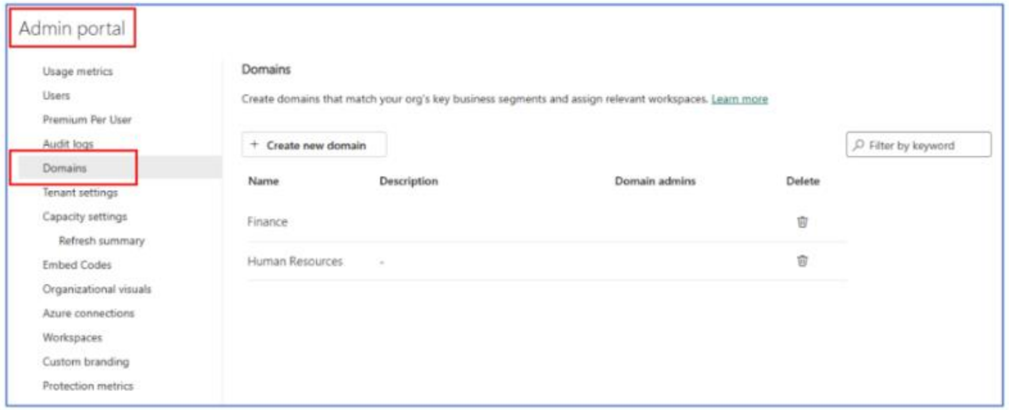
[Data Mesh with Fabric.docx (sharepoint.com)]



Workspaces in this domain

See all the workspaces in Sales and assign more.

□	Name	Workspace Admins
□	Sales_West_WS	-
□	Sales_South_WS	-
□	Sales_North_WS	-

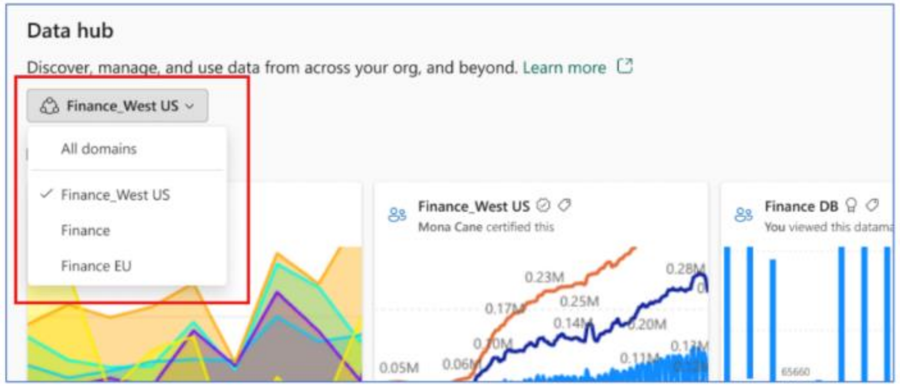


Admin portal

Domains

Create domains that match your org's key business segments and assign relevant workspaces. Learn more

Name	Description	Domain admins	Delete
Finance			🗑
Human Resources	-		🗑



Data hub

Discover, manage, and use data from across your org, and beyond. Learn more

OneLake for all domains

OneLake gives a true data mesh as a service

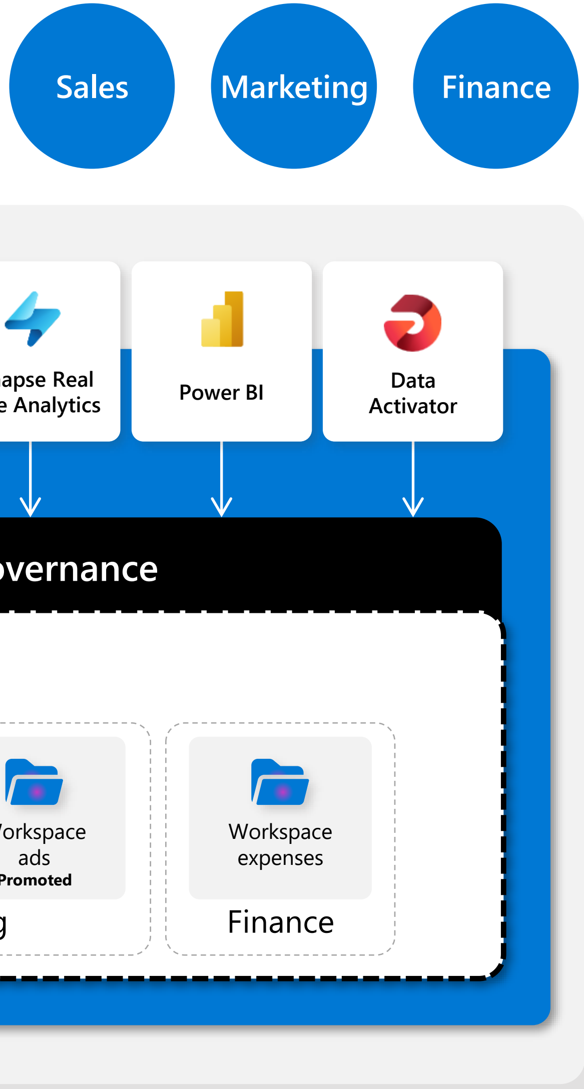
Introducing domains as an integral part of Fabric: A domain is a way to logically group together all the data in an organization relevant to an area or field, according to business needs

Domains are defined with domain admins and contributors who can associate workspaces and group them together under a relevant domain

Federated governance can be achieved by delegating settings to domain admins, thus allowing them to achieve more granular control over their business area

Domains simplify discovery and consumption of data across the organization, thus allowing business optimized consumption

Avoid data swamps by endorsing certain data as certified or promoted, thus encouraging reuse.

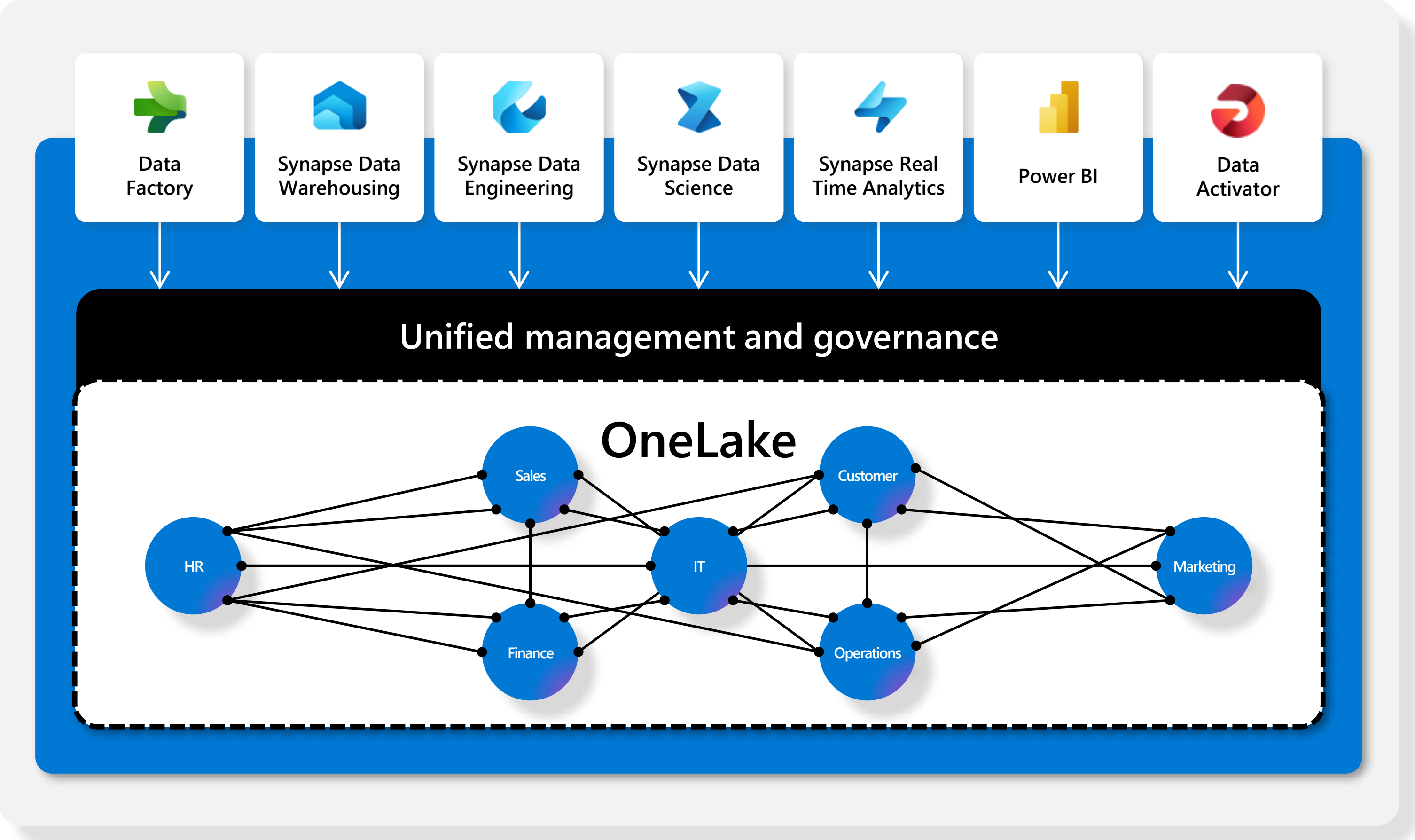


OneLake gives a true data mesh as a service

One Copy enables data to be used across domains, clouds and engines

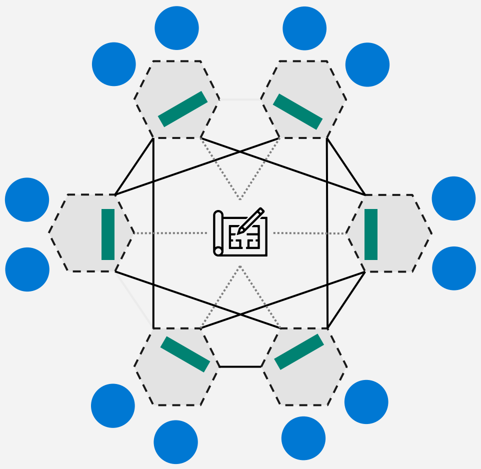
An organization will have many data domains with many workspaces with different data owners. However, a single data product can span multiple domains.

Shortcuts provide the connections between domains so that data can be virtualized into a single data product without data movement, data duplication or changing the ownership of the data.



Governance Topologies : Different Approaches

Mesh Type 2

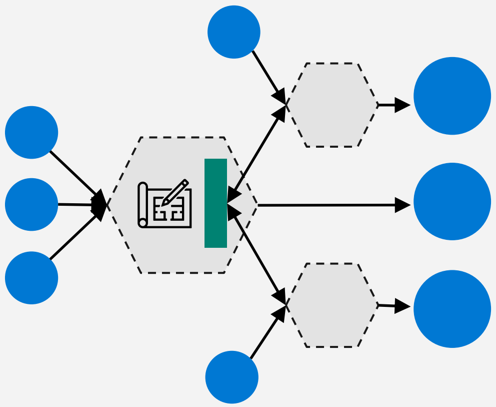


- Domains use the same technology
- Each domain has its own storage that is the same technology

Centralised (Control)

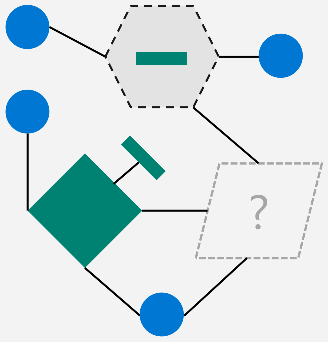
Distributed (Agility)

Mesh Type 1



- Domains use the same technology
- Data is kept in one enterprise data lake with each domain getting its own container/folder

Mesh Type 3



- Domains can use any technology they want
- Each domain has its own storage that can be any technology

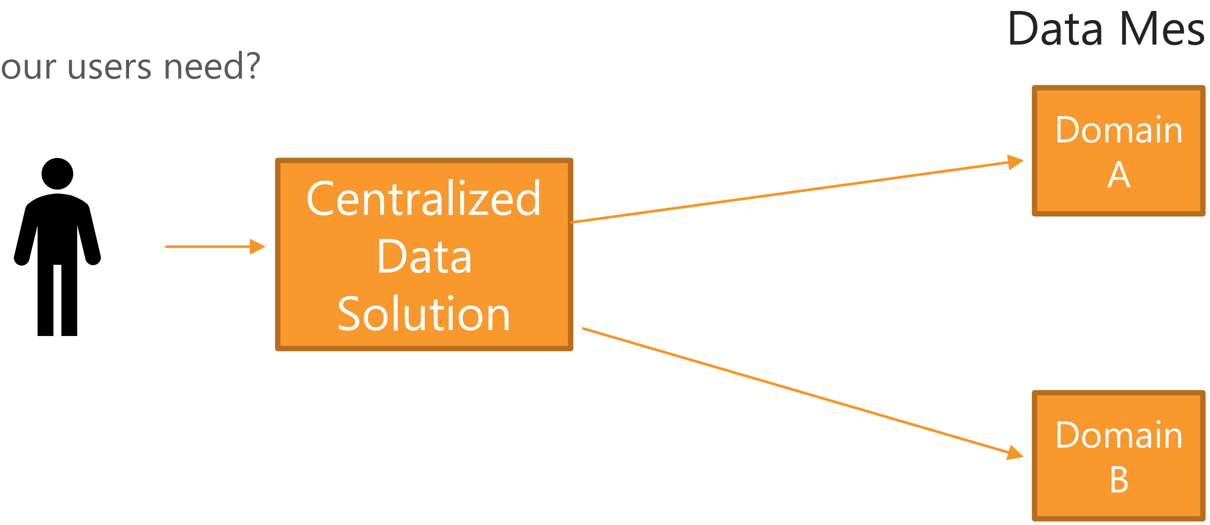
Data Mesh Future

This view is my own and not that of Microsoft!

In the end, I predict data mesh will become an extension to a centralized data solution for a small percentage of solutions.

There will be a very small percentage of solutions that are 100% true to the pure data mesh concept (assuming mesh type 1 and 2 are true to the data mesh concept). *Ask ten people what a data mesh is and you will get eleven answers!* Some of the concepts of a data mesh will be used in a larger percentage of solutions.

Always ask: What do your users need?



1) **Domain ownership (90%)**
2) Data as a product (70%)
3) Self-serve data infrastructure as a platform (30%)
4) Federated computational governance (50%)

Rethinking the Data Mesh Architecture: Apply it Piecemeal (eckerson.com)

Data Mesh concepts help with a better way of thinking how to get value out of data

When to use what architecture?

A very high-level use case for each architecture (in ascending cost and complexity):

- Modern data warehouse: Small amount of data; if used to relational data warehouses (RDW)
- Data fabric: Need to ingest many different data sources (size, speed, type)
- Data lakehouse: Use it until you can’t – then copy some data to RDW
- Data mesh: Very large, domain-oriented company, that is having major paint points with scalability and can afford a long timeline

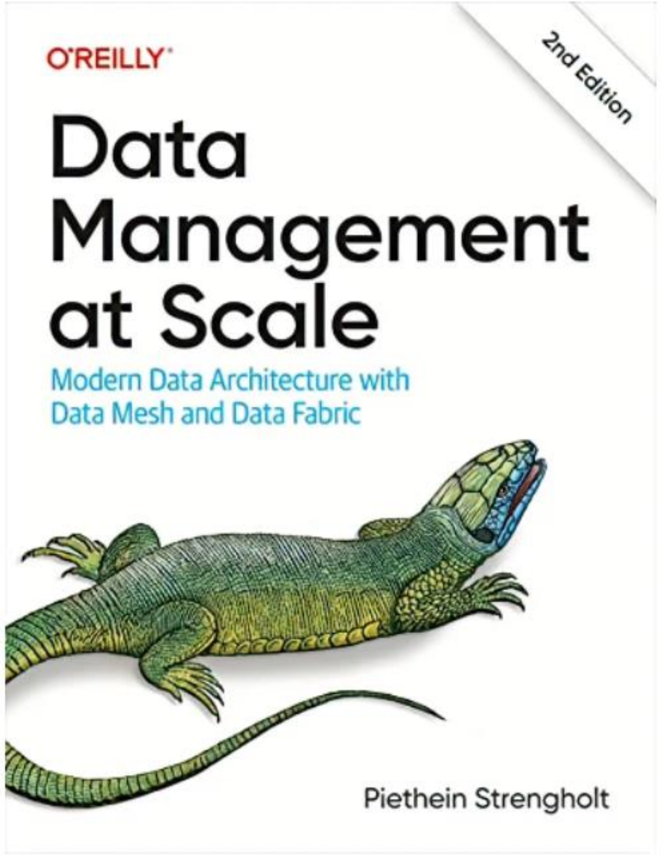
Most companies will use pieces of each architecture to build a solution adapted to their specific needs for data (use cases) and their business capabilities.

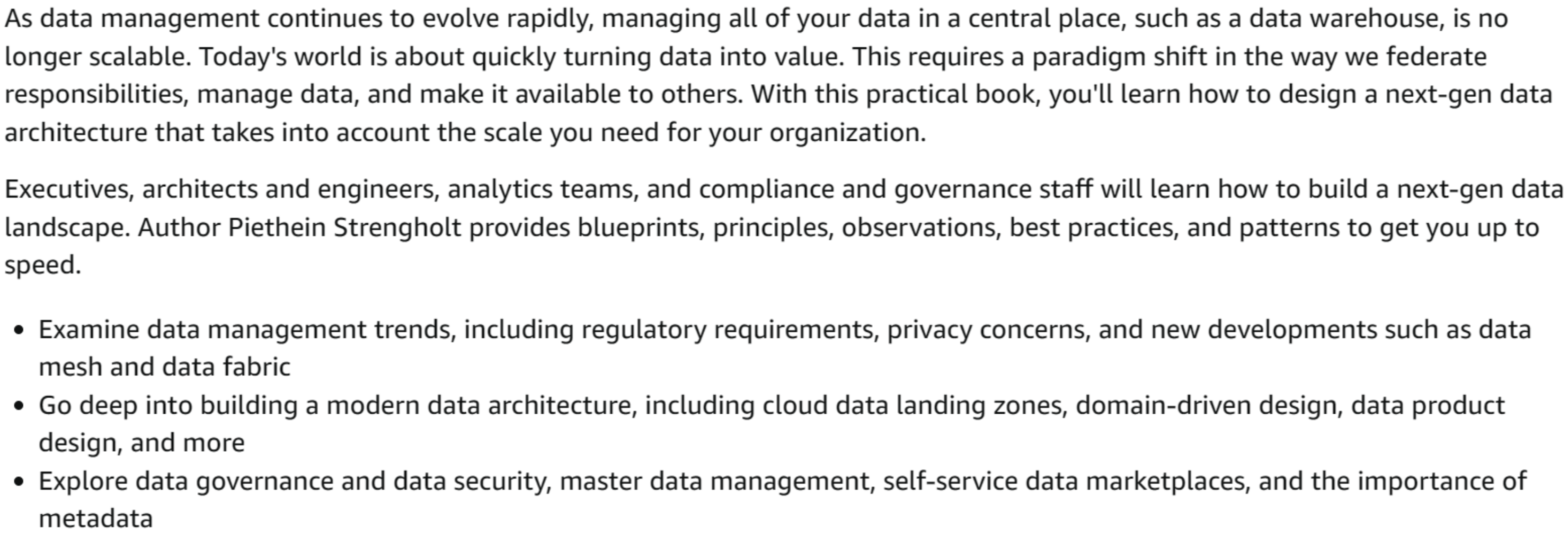
Data Mesh on Azure Resources

- Piethein Strengholt: Blog - Implementing Data Mesh on Azure, Blog – Data Mesh topologies, Book - Data Management at Scale: Best Practices for Enterprise Architecture
- Cloud Adoption Framework: Azure data management and analytics scenario
- Data Management & Analytics Scenario - Data Management Zone: Github
- Data Management & Analytics Scenario - Data Landing Zone: Github
- Enterprise-Scale - Reference Implementation: Github
- Microsoft doc: A financial institution scenario for data mesh
- Provision three Azure Data Lake Storage Gen2 accounts for each data landing zone
- Overview of Azure Data Lake Storage for the data management and analytics scenario
- The best practices for organizing Synapse workspaces and lakehouses

- Data Mesh, Data Fabric, Data Lakehouse – video from Toronto Data Professional Community on 2/15/23

Data Management at Scale





As data management continues to evolve rapidly, managing all of your data in a central place, such as a data warehouse, is no longer scalable. Today's world is about quickly turning data into value. This requires a paradigm shift in the way we federate responsibilities, manage data, and make it available to others. With this practical book, you'll learn how to design a next-gen data architecture that takes into account the scale you need for your organization.

Executives, architects and engineers, analytics teams, and compliance and governance staff will learn how to build a next-gen data landscape. Author Piethein Strengholt provides blueprints, principles, observations, best practices, and patterns to get you up to speed.

- Examine data management trends, including regulatory requirements, privacy concerns, and new developments such as data mesh and data fabric
- Go deep into building a modern data architecture, including cloud data landing zones, domain-driven design, data product design, and more
- Explore data governance and data security, master data management, self-service data marketplaces, and the importance of metadata

http://aka.ms/datamanagementatscale (free copy)

Amazon link

Q & A

James Serra, Microsoft, Data & AI Solution Architect
Email me at: jamesserra3@gmail.com
Follow me at: @JamesSerra
Link to me at: www.linkedin.com/in/JamesSerra
Visit my blog at: JamesSerra.com