

Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh



James Serra

Data & AI Solution Architect

Microsoft, Federal Civilian

jamesserra3@gmail.com

Blog: JamesSerra.com





16 edycja konferencji SQLDay

13-15 maja 2024, WROCŁAW + ONLINE



partner platynowy



partner złoty



partner srebrny

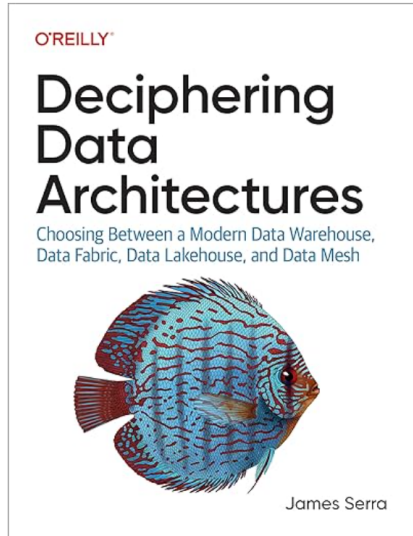


About Me



- 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 the book "Deciphering Data Architectures: Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse, and Data Mesh"

My book



Roll over image to zoom in



Deciphering Data Architectures 1st Edition, Kindle Edition

by **James Serra** (Author) | Format: Kindle Edition

5.0 ★★★★★ 3 ratings

#1 New Release in Data Modeling & Design

[See all formats and editions](#)

Book description

Editorial reviews

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

Google Play Ebook: 40% off coupon:
[P412MYVVMURSF](#)



Read entire book now with an O'Reilly subscription:
[Deciphering Data Architectures \(O'Reilly.com\)](#)

- Foundation
 - 1. Big Data
 - 2. Types of Data Architectures
 - 3. The Architecture Design Session
- Common Data Architecture Concepts
 - 4. The Relational Data Warehouse
 - 5. Data Lake
 - 6. Data Storage Solutions and Processes
 - 7. Approaches to Design
 - 8. Approaches to Data Modeling
 - 9. Approaches to Data Ingestion
- Data Architectures
 - 10. The Modern Data Warehouse
 - 11. Data Fabric
 - 12. Data Lakehouse
 - 13. Data Mesh Foundation
 - 14. Should You Adopt Data Mesh? Myths, Concerns, And The Future
- People, Process, and Technology
 - 15. People And Processes
 - 16. Technologies

[More details](#)

[Order](#) on Amazon!

Agenda

- Relational 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!

I tried understanding all this architecture stuff...

And ended up passed-out drunk in a Bar Mleczny parking lot



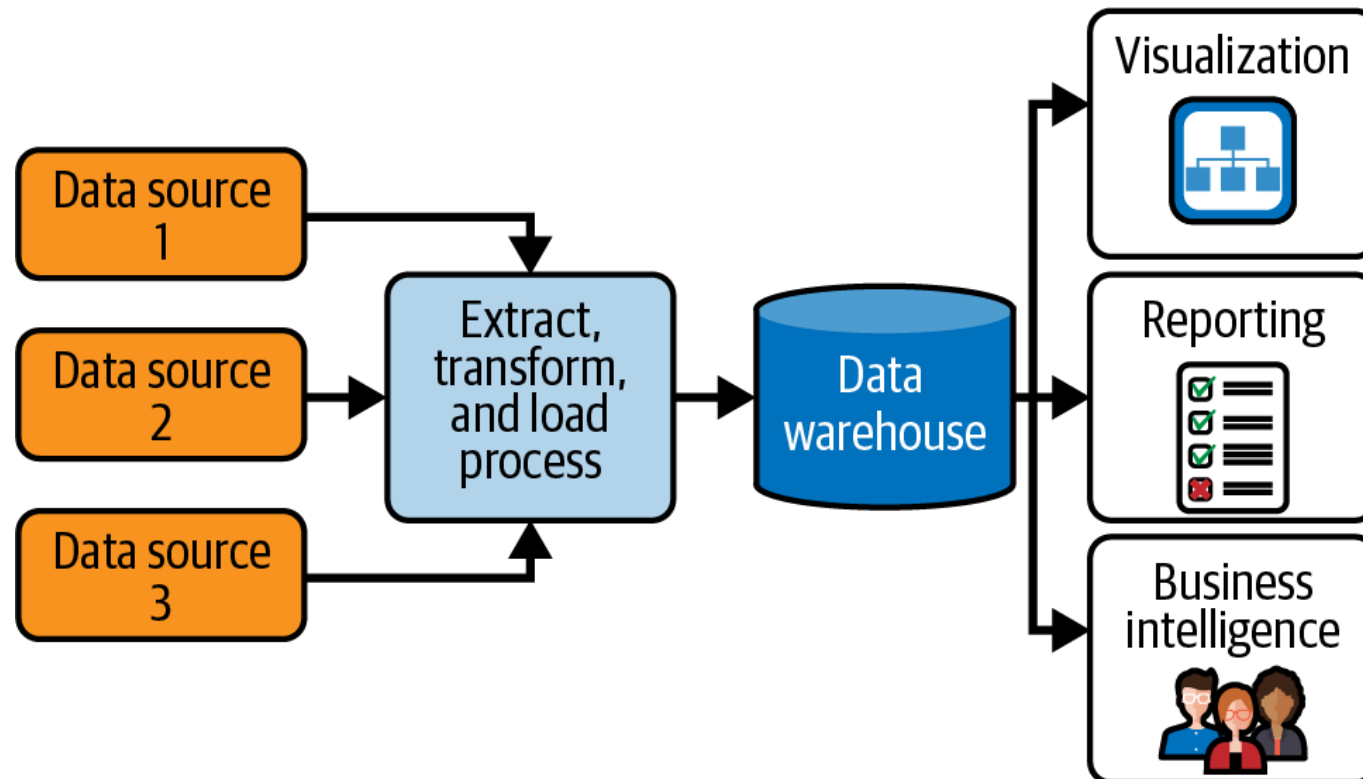
Let's prevent that from happening...

Relational Data Warehouse (RDW)

What is a Relational Data Warehouse?

(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 **to make better business decisions** by getting greater insights into your company. 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.



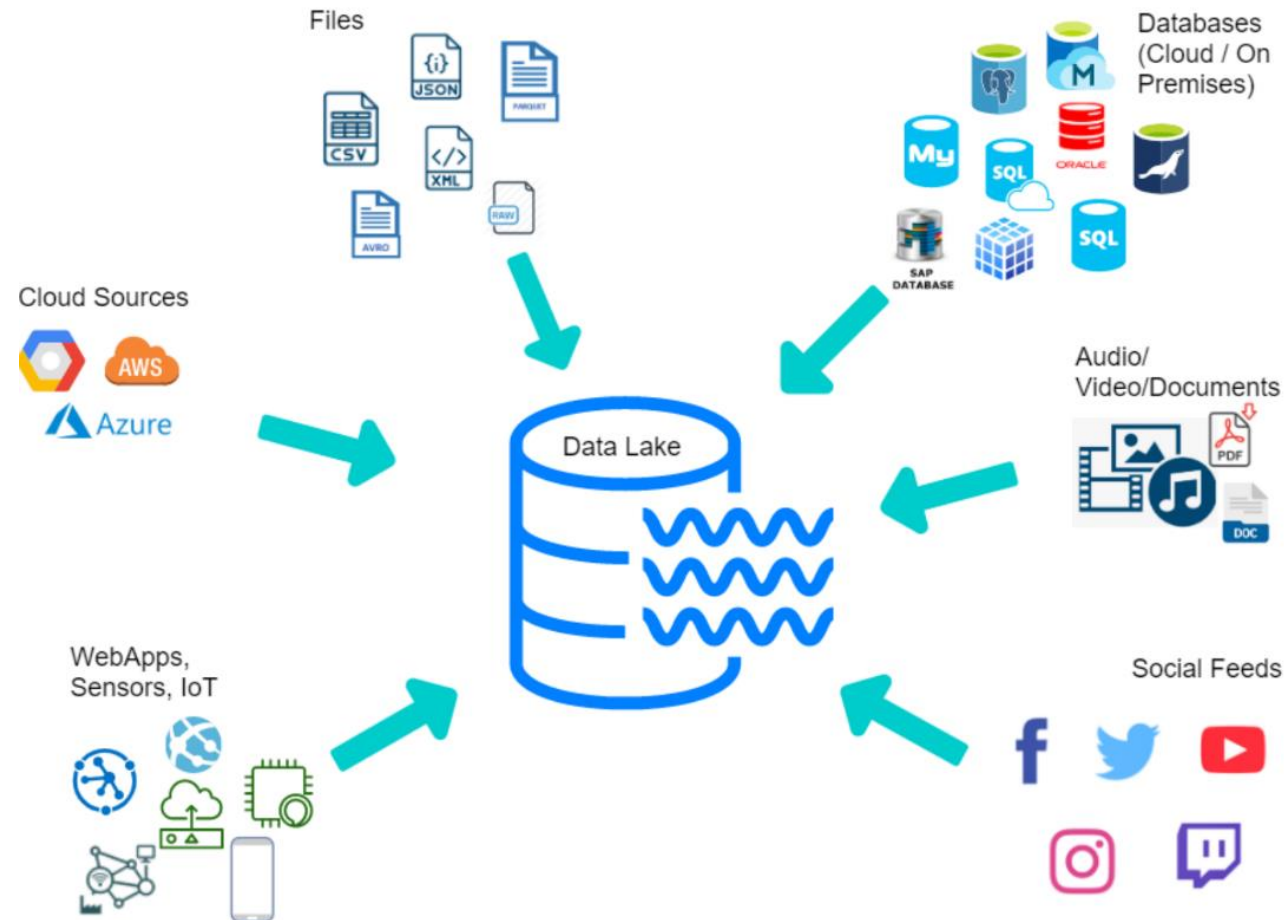
Why use a Relational Data Warehouse?

- Single version of the truth
- Reduce stress on the production system
- Optimized for read access
- Integrate multiple sources of data
- Run accurate historical reports (no need to save hardcopy reports)
- Restructure and rename tables and fields
- Protection against application upgrades
- Reduced security concerns
- Keep historical data
- Use Master Data Management, including hierarchies
- Improve data quality by plugging holes in source systems
- No IT involvement needed to create reports

Data Lake

What is a Data Lake?

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

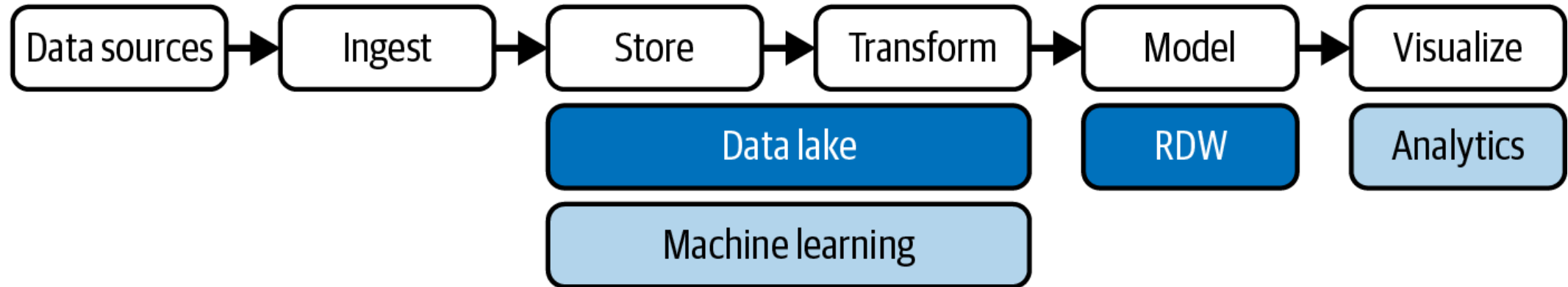


Why use a data lake?

- Store data with no modeling: Schema-on-read (RDW is schema-on-write)
 - Allows for quick user access to data for power users/data scientists (allowing for faster ROI)
 - Provides for data exploration to see if data valuable before writing ETL and schema for relational database, or use for one-time report/query
- Frees up expensive enterprise data warehouse (EDW) resources for queries instead of using EDW resources for transformations. Removes need for EDW maintenance window
- Extreme performance for transformations by having multiple compute options each accessing different folders containing data
- Stockpiling data cheaply
- Keep all history of data

Modern Data Warehouse (MDW)

MDW high-level architecture



Data Lake with Data warehouse use cases

Data Lake

Staging & preparation

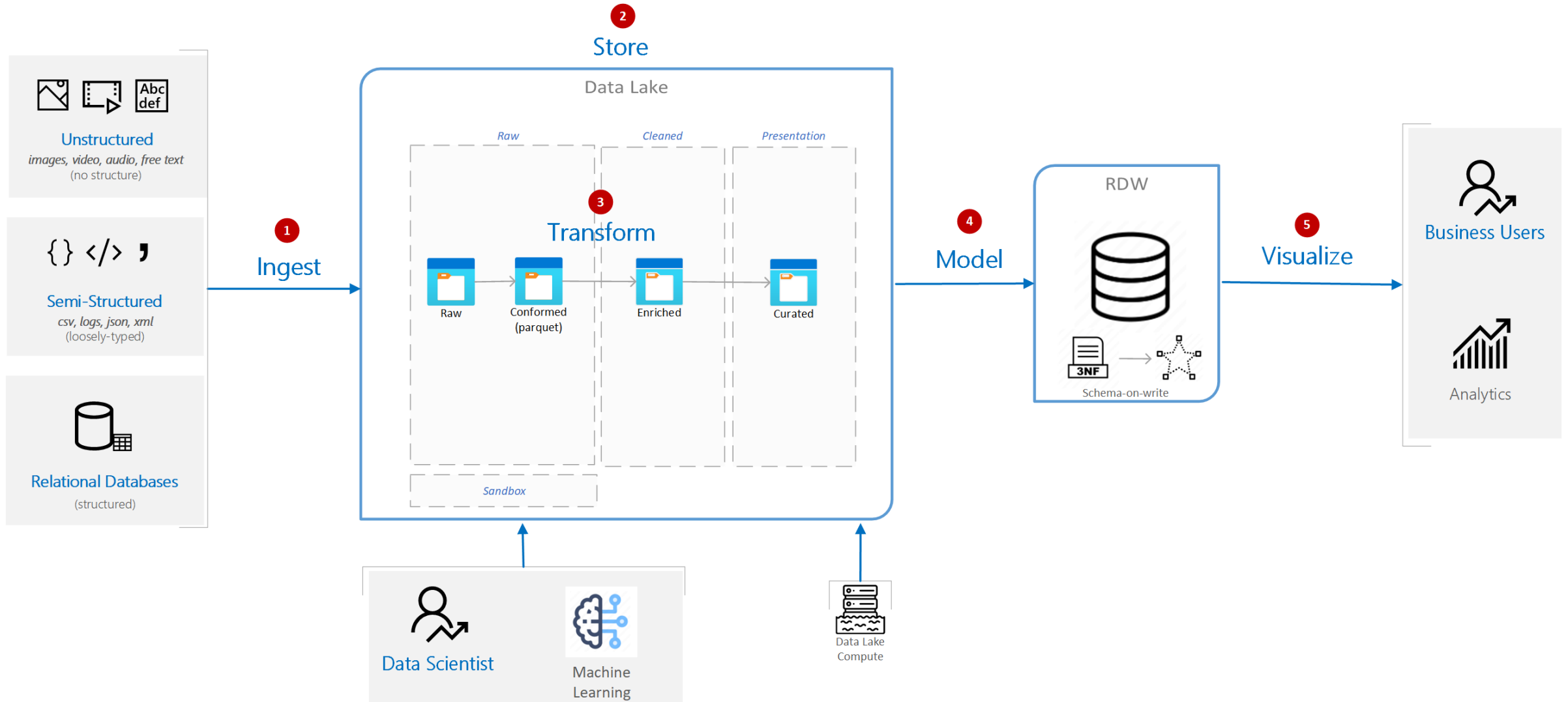
- Data scientists/Power users
- Batch processing
- Data refinement/cleaning
- ETL workloads
- Store older/backup data
- Sandbox for data exploration
- One-time reports
- Quick access to data
- Don't know questions

Relational 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 architecture



Data Fabric

What is Data Fabric?

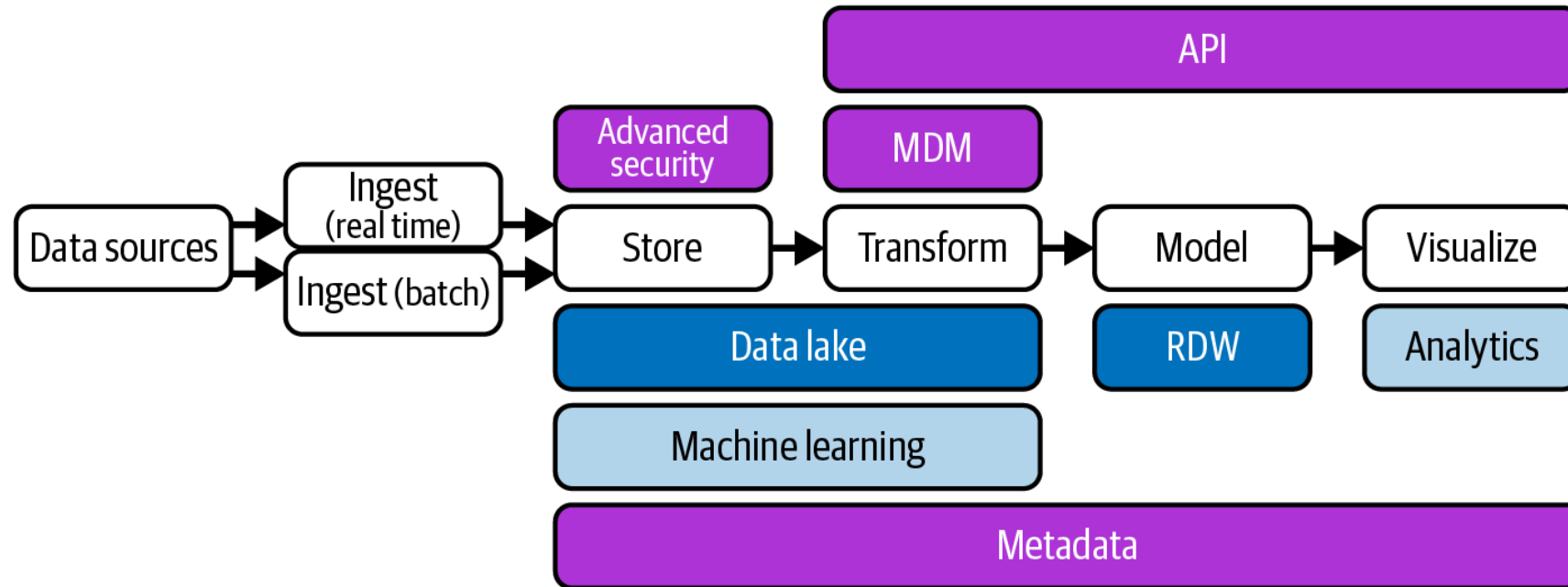
A data fabric is a term used to describe the architecture of taking disparate systems and weaving them together, like fabric, to create a consistent layer on top of an organization's data.

Data Fabric adds to a modern data warehouse:

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

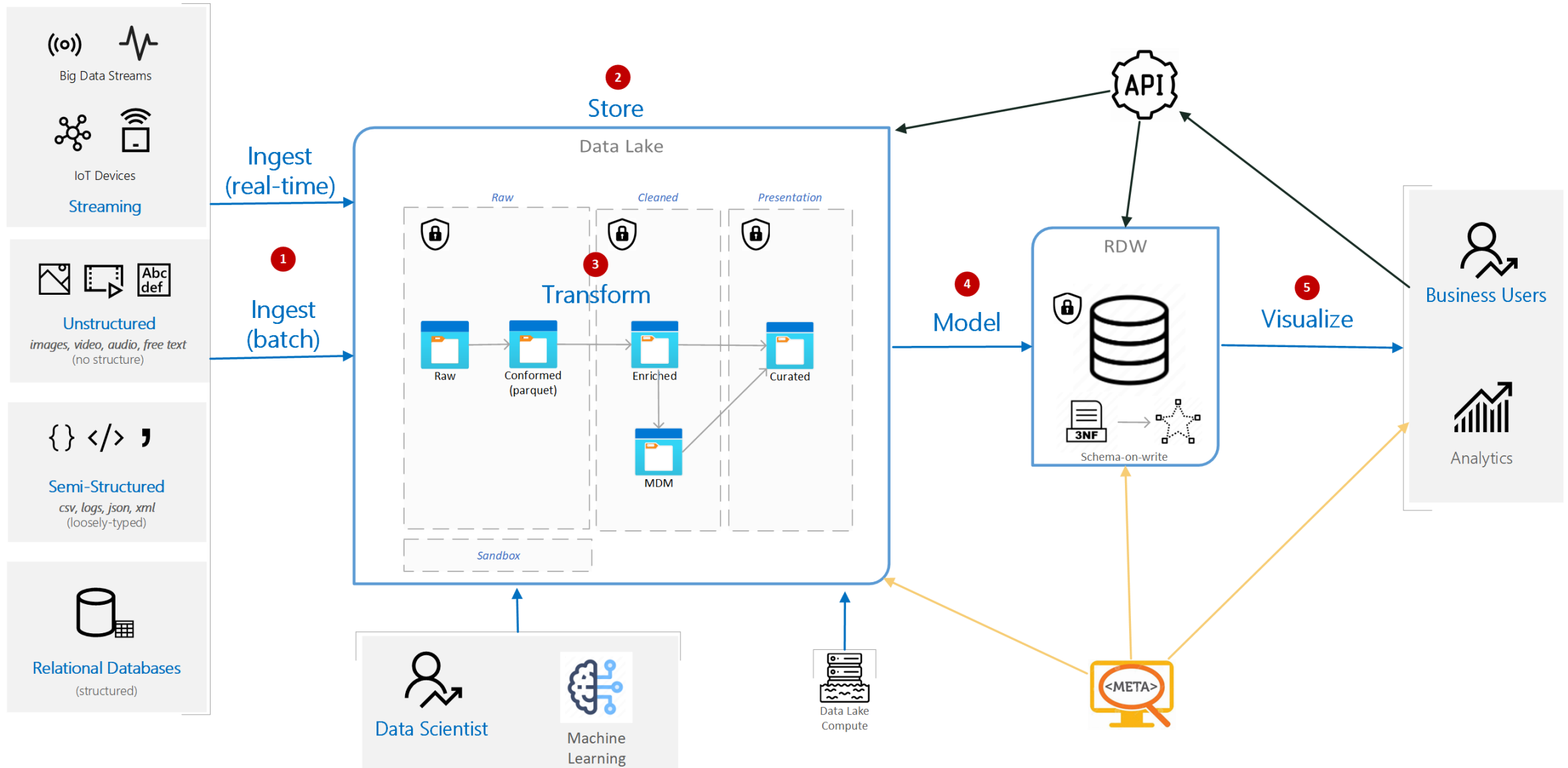
Bottom line: Data fabric provides additional technology to source more data, secure it, and make it available. *Think of it as an evolution of the MDW*

Data Fabric high-level architecture



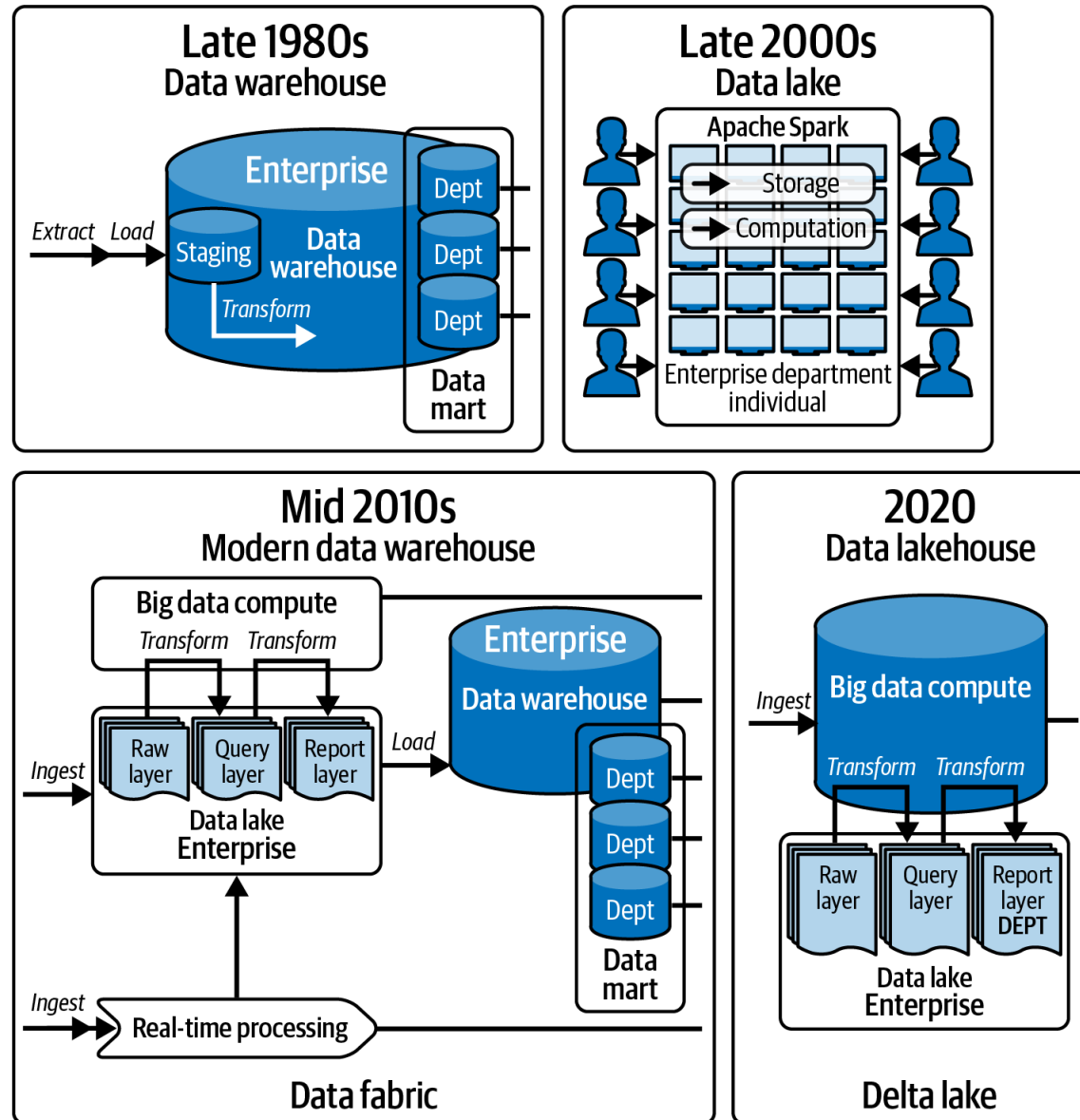
In purple are the data fabric features

Data Fabric architecture



Data Lakehouse

Data Lakehouse historical timeline



Delta Lake

A transactional storage software layer that runs on top of an existing data lake, adding RDW-like features.

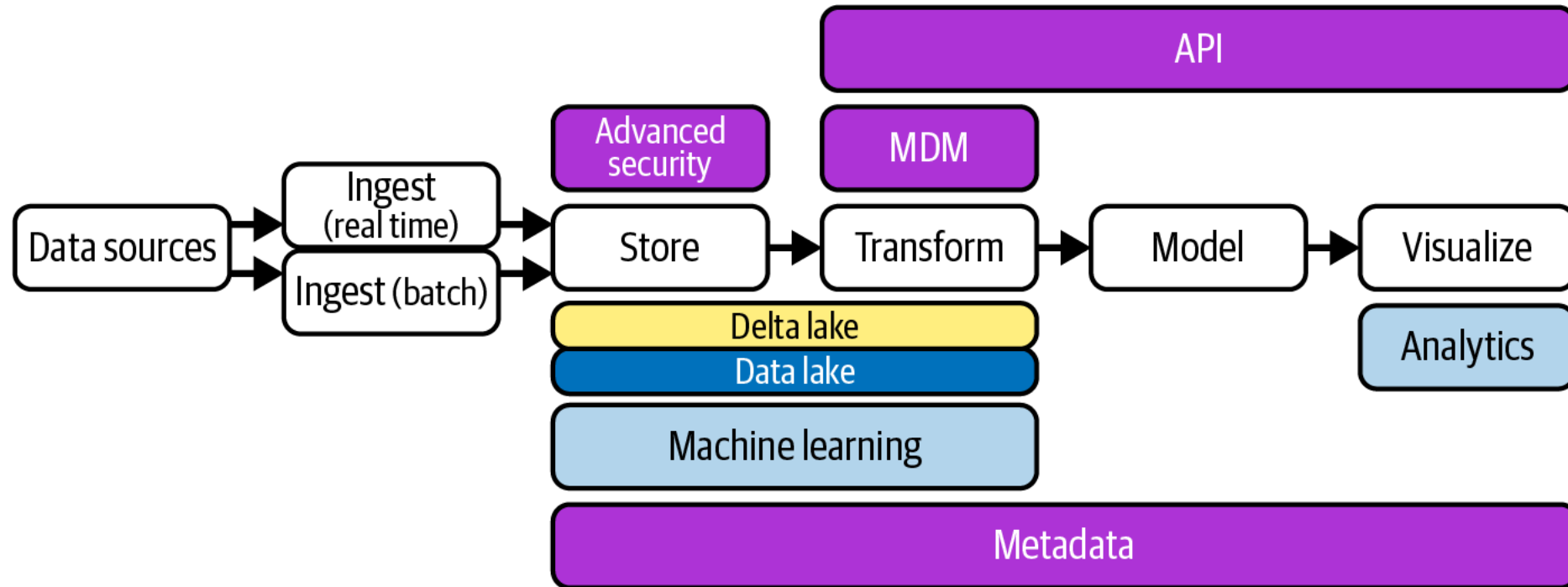
Top features:

- Supports SQL commands INSERT, DELETE, UPDATE, and MERGE
- ACID transactions for one table
- Time travel (data versioning enables rollbacks, audit trail)
- Streaming and batch unification
- Schema enforcement & schema evolution
- Performance improvements (Data skipping, caching, Z-Order, etc)
- Solve "small files" problem via OPTIMIZE command (compact/merge)

Spark:

```
df.write.format("delta").save(delta_table_path)
      instead of
df.write.format("parquet").save(delta_table_path)
```

Data Lakehouse high-level architecture



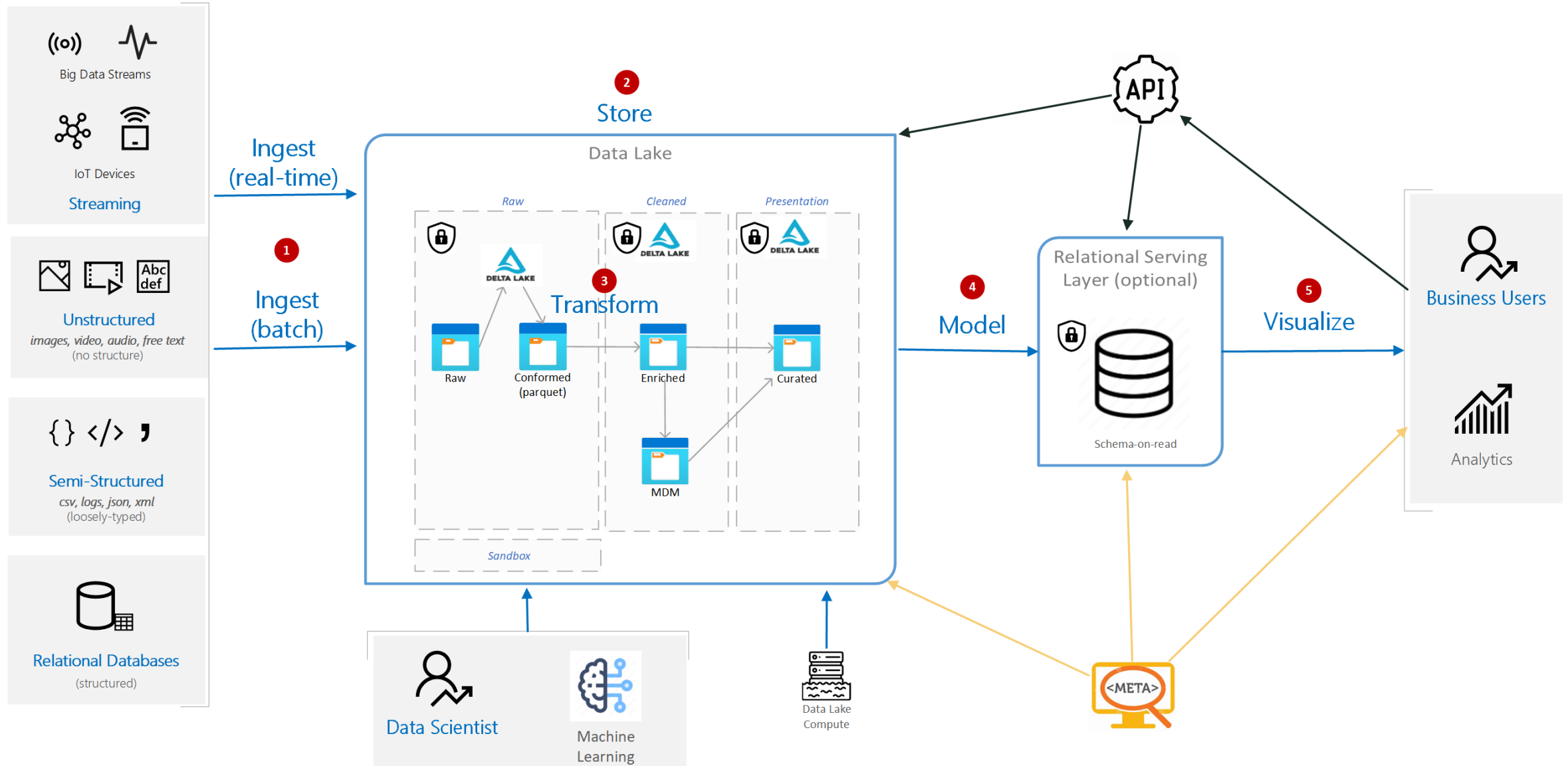
Added Delta lake, removed RDW

Use cases for Data Lakehouse

Today's data architectures commonly suffer from six problems:

- Reliability: Keeping the data lake and warehouse consistent
- Data staleness: Data in warehouse is older
- Limited support for advanced analytics: Data scientists prefer files
- Total cost of ownership: Extra cost for data copied to warehouse
- Data governance: More copies, more risk
- Complexity: More specialized skills needed for both a data lake and RDW

Data Lakehouse architecture



Opening a can of worms



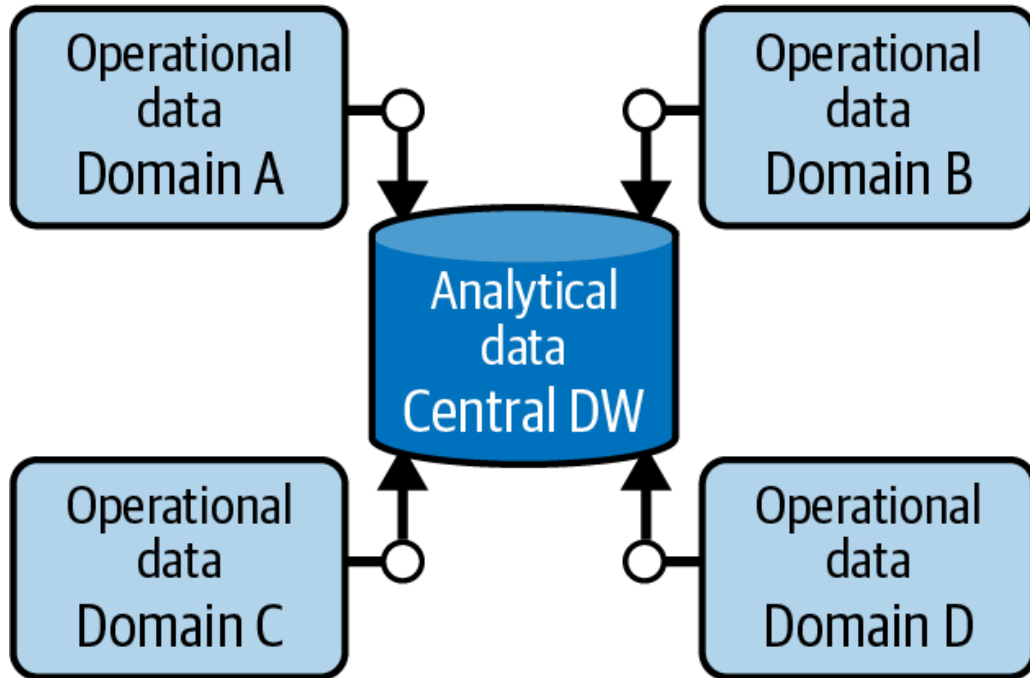
Concerns skipping relational data warehouse

- Speed: Relational database *queries* are faster, especially with Massively Parallel Processing (MPP) and with advanced indexing, advanced statistics, caching, advanced query plan optimization, materialized views, advanced join optimization
- Security: No row-level security (RLS), column-level security, data-at-rest encryption, column-level encryption, Transparent Data Encryption (TDE), dynamic data masking
- People are used to using a relational database (forced metadata layer)
- Complexity: Metadata separate from data, file-based world
- Concurrency: Multiple reads of a file at the same time can be slow
- Missing features: SQL Views, referential integrity, workload management, advanced auditing and compliance features (such as auditing trails, data retention policies, and compliance certifications), ACID against multiple tables
- Products must add delta lake support in order to use it

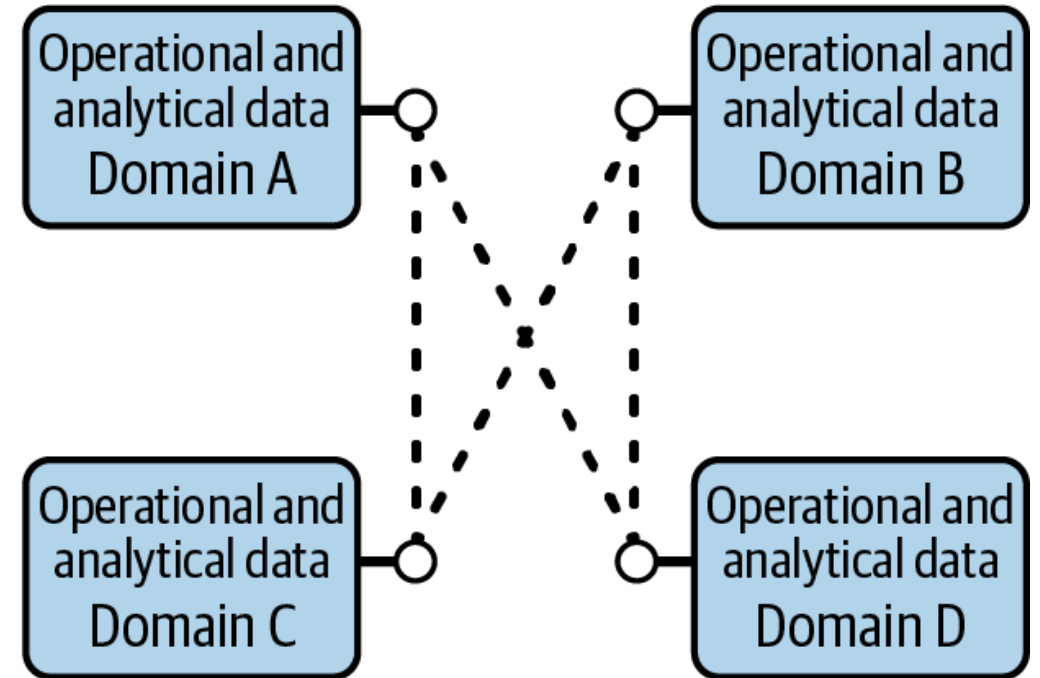
New technology addressing some of these concerns. Just put limited data in RDW that needs it

Data Mesh

Traditional centralized vs Data Mesh decentralized



MDW, Data Fabric, Data Lakehouse



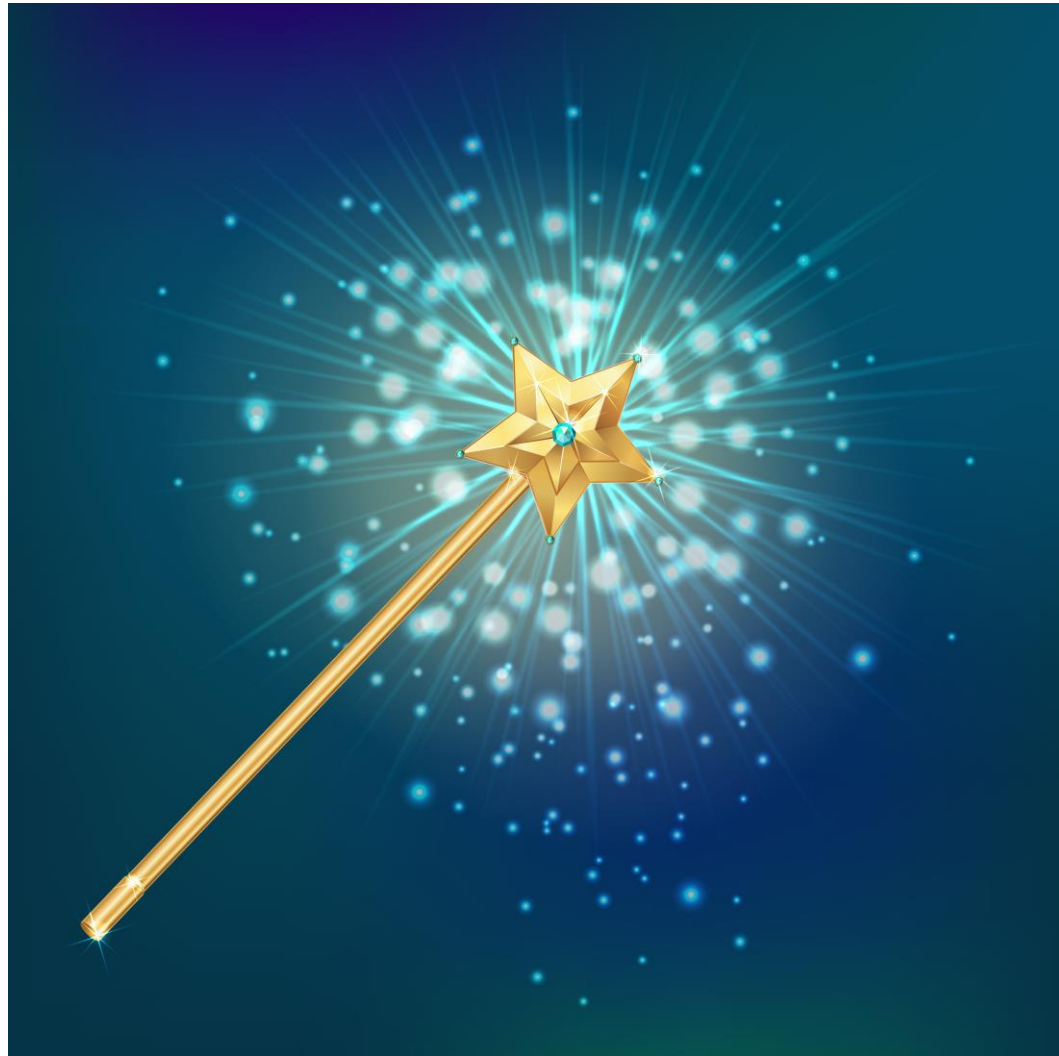
Data Mesh

Data Mesh

Data Mesh is a concept, not technology

*It is a **huge** organizational and cultural shift*

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:

Limited quantity available
146 sold / [See feedback](#)

Price: **US \$3,950.00**

[\\$164.58 for 24 months with PayPal Credit*](#)

[Buy It Now](#)

[Add to cart](#)

[♥ Add to Watchlist](#)

☐ [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 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 - 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)

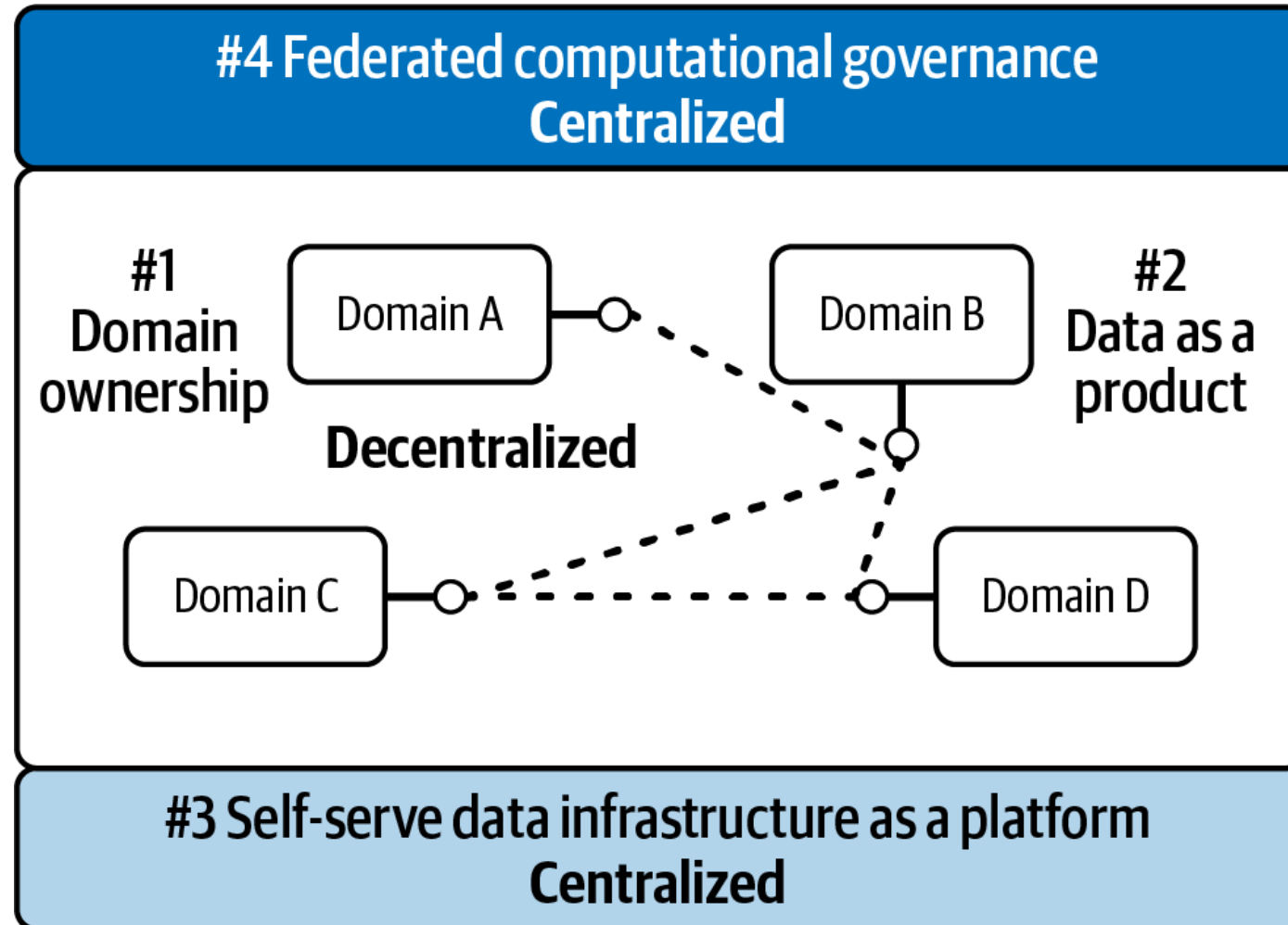
#3) Self-serve data infrastructure as a platform

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 architecture



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

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

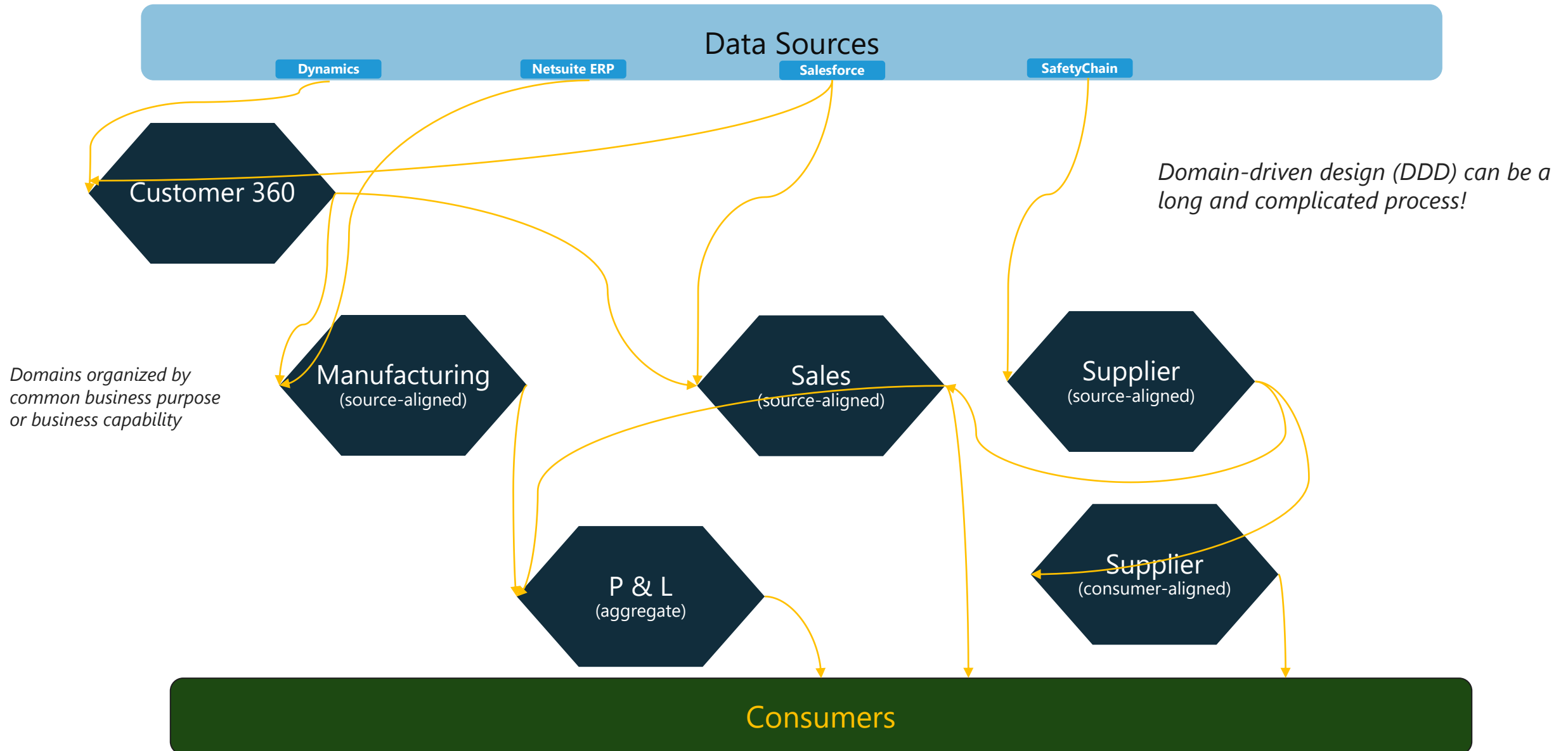
Public health research

Infectious disease modeling

Health behavior change

Chronic disease management

Data Mesh – Logical Architecture



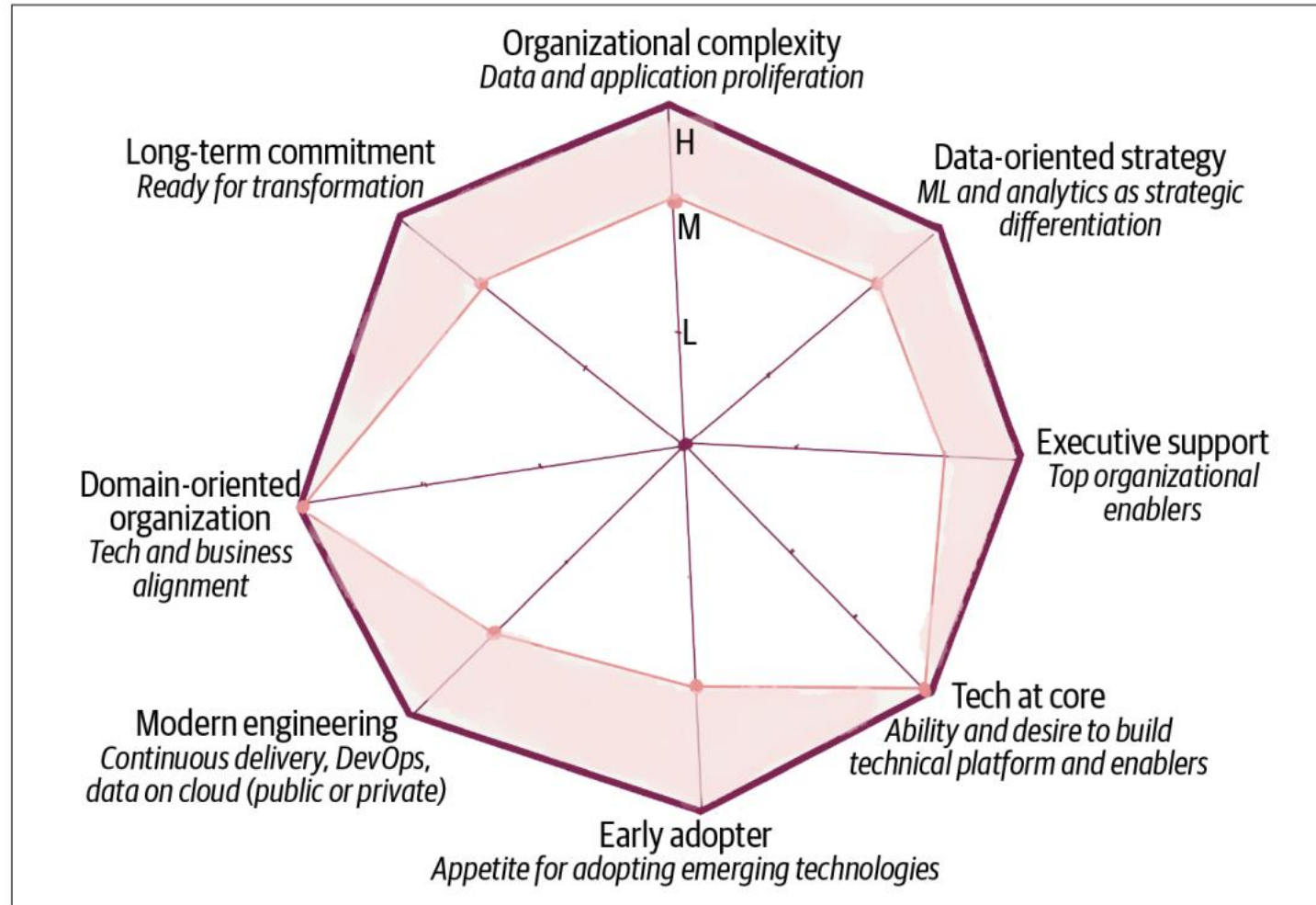
Concerns with Data Mesh

- No standard definition of a data mesh
- Huge investment in organizational change and technical implementation
- Performance problem 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 combining data

[Data Mesh: Centralized vs decentralized data architecture](#)

[Data Mesh: Centralized ownership vs decentralized ownership](#)

Should you adopt data mesh today?



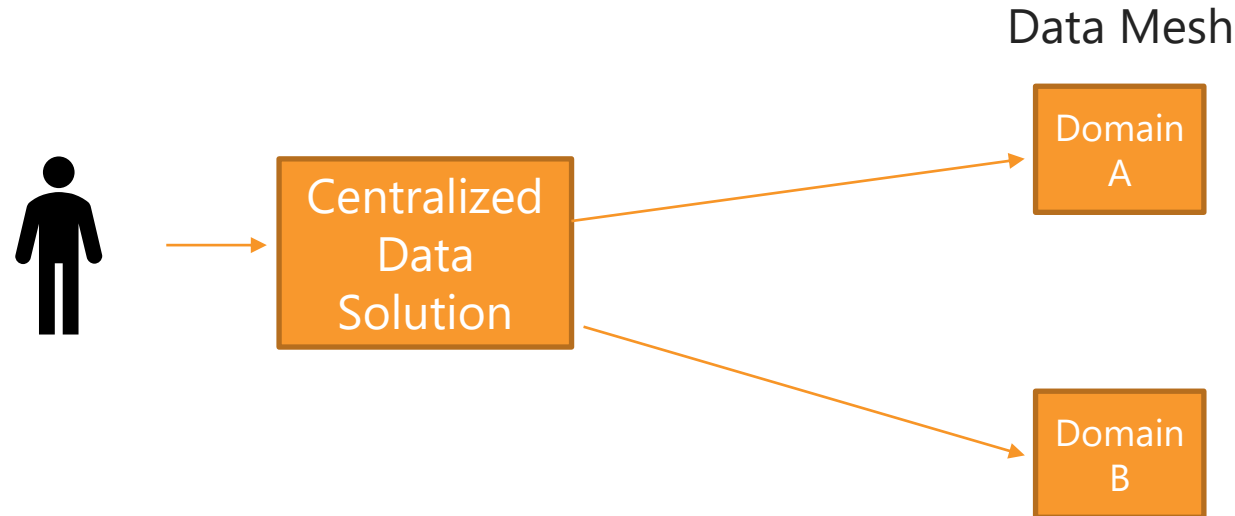
Need to score medium or high in ALL categories

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 via a hub-and-spoke model:

- Start by using new data to create new data mesh domains
- Supplement those domains with your current centralized data solution
- Slowly migrate your centralized data into data mesh domains over time
- Paves the way for a cultural shift over time



Data Mesh principles adoption estimate:

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

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

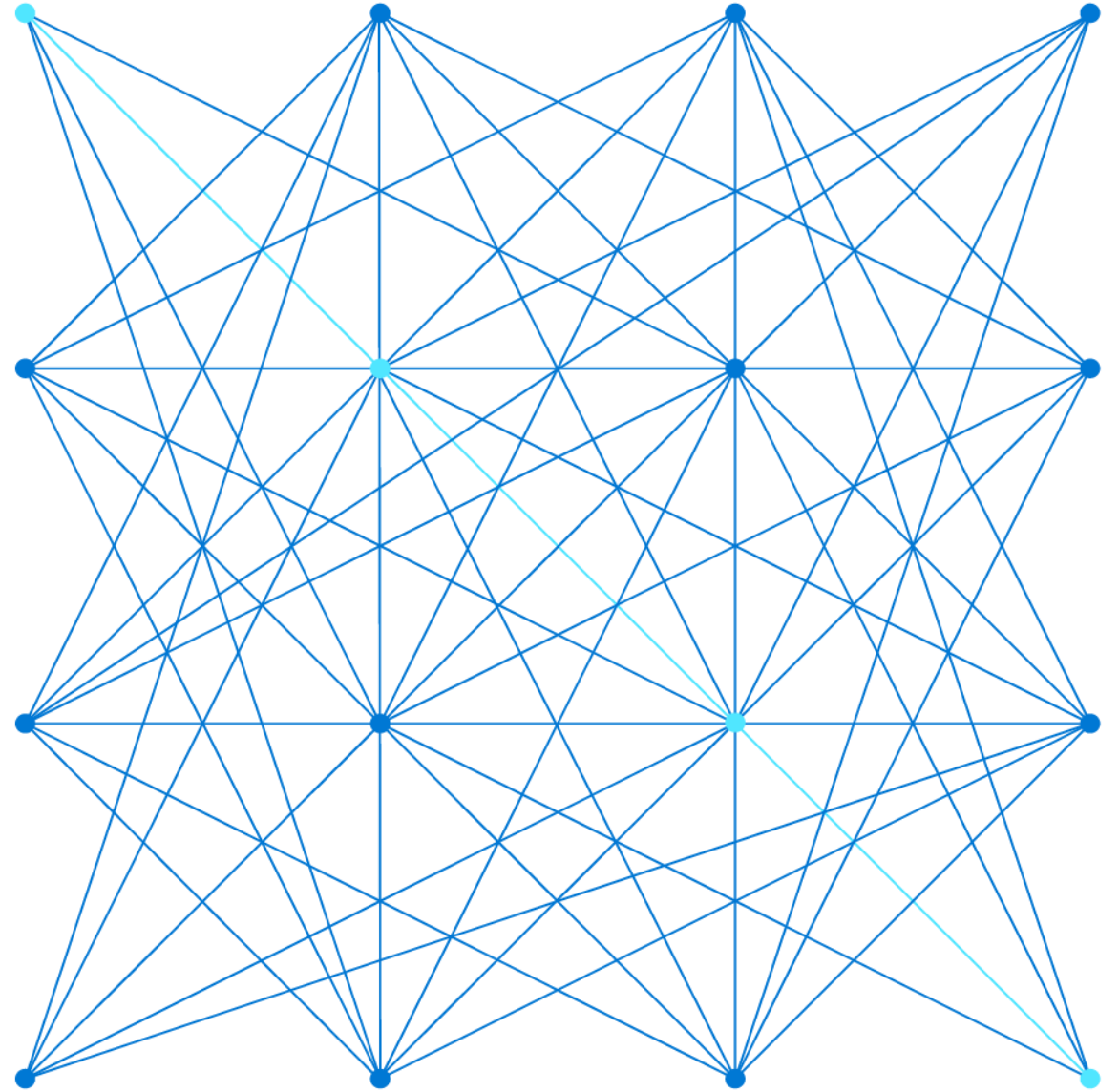
When to use each architecture?

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); can skip data lake if very small amount of data
- Data fabric: Need to ingest many different data sources (size, speed, type). Large migration effort (i.e., many SP's would need to be rewritten if no RDW)
- 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 pain points with scalability and can afford a long timeline. Each domain will use one of the three architectures

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



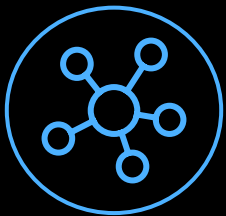
Microsoft Position:



Microsoft focuses on organizations challenges and strategic vision vs landing a framework



Microsoft provides the NorthStar for implementing Modern Analytics Ecosystem (technology) in alignment to people and processes



Supports aspects of Data Mesh and Data Fabric but with the caution that we're not purists of either framework\solution

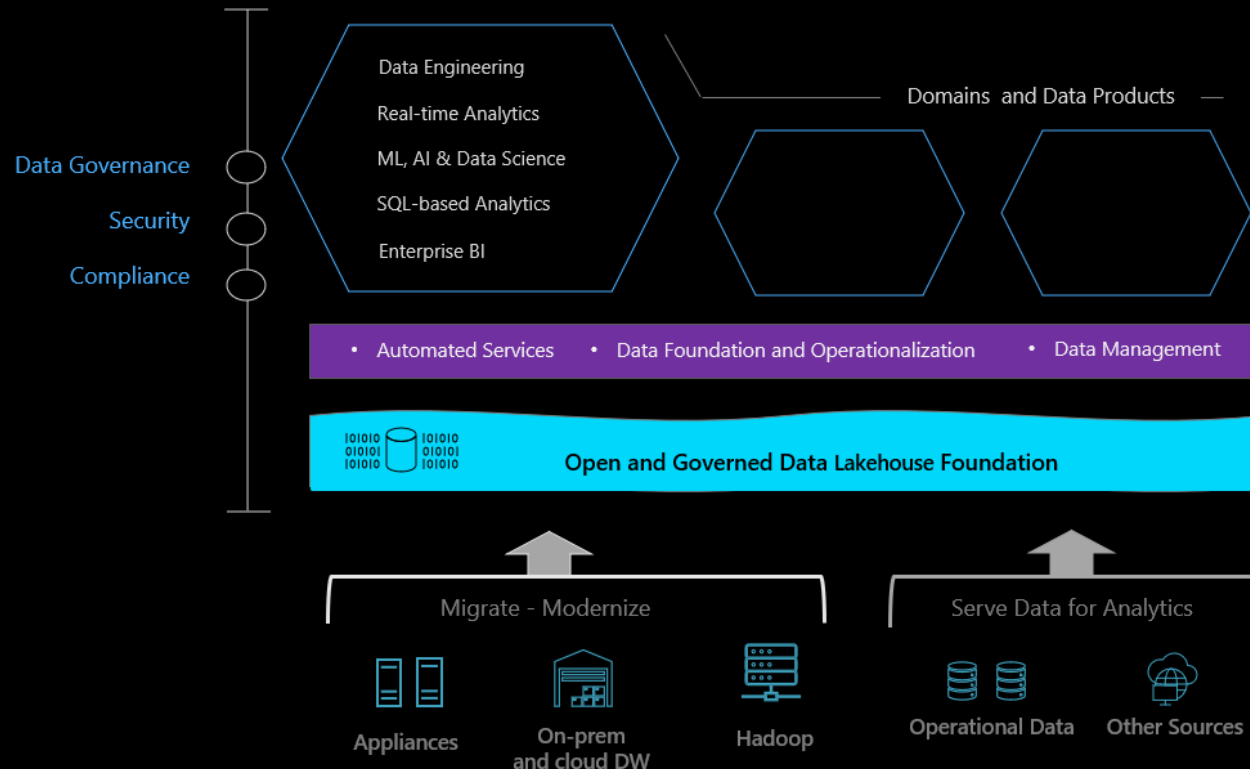
Modern Analytics and Governance (MAG) at Scale -

A transformative approach to building a modern, end-to-end analytics ecosystem aligned to data mesh, data fabric, and data hub

Call it MAG at Scale instead of data mesh!

Modern Analytics and Governance at Scale

Microsoft's Hybrid Approach to **data mesh**, **data fabric** and **data hub**



MS Packaged Solution

- **Enterprise Data Strategy**
 - Culture Transformation
 - Align Process + People + Technology
 - Enable Autonomous Lines of Businesses
 - Organization Change Management
 - Platform and Data Ownership
- **Composable Solution**
 - Enterprise Data Governance
 - Data Management Foundation
 - Domains and Data Products
- **Technical Architecture**
 - (Azure Cloud Adoption Framework)
 - Azure Landing Zones
 - Azure Data Services
 - IPs and Accelerators (Microsoft + Partners)
- **Implementation Stages**
 - Discovery, MVPs, Phases

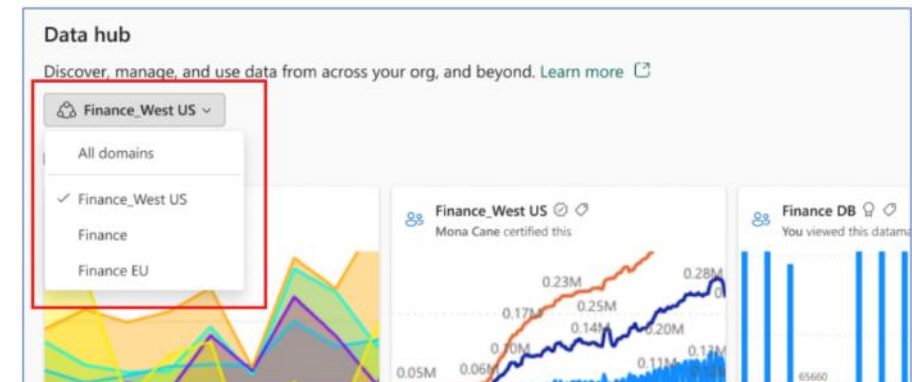
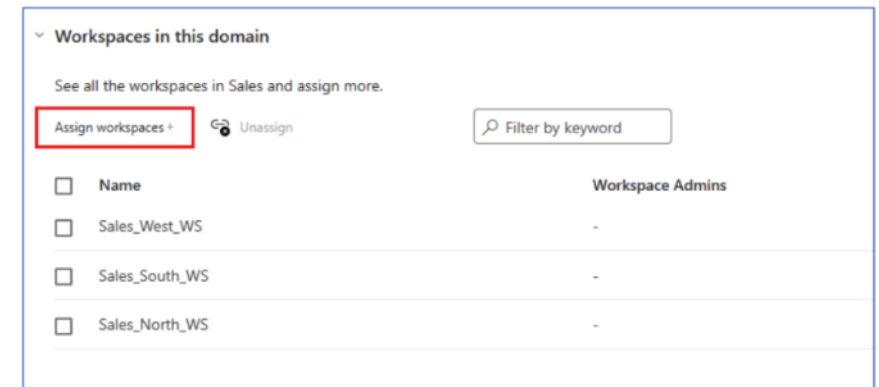
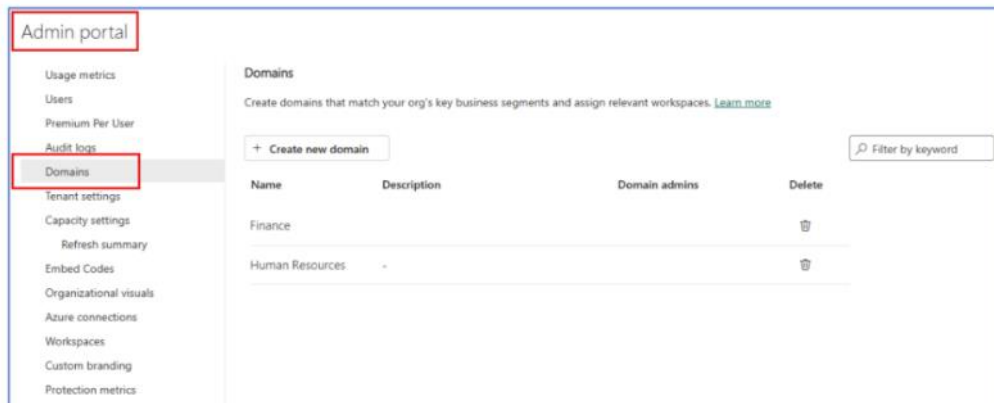
Microsoft Fabric and Data Mesh

- Can logically organize data into domains and subdomains: [Domains - Microsoft Fabric | Microsoft Learn](#)
- PBI workspaces are associated with domains or subdomains – 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:

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

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



Microsoft Purview and Data Mesh

- Public preview of the new and fully reimagined Microsoft Purview data governance solution
- Business domains -> Data products -> Assets (i.e. tables, files, Power BI reports)

Business domains

Organize data products into meaningful groups and link them to business concepts. [Learn more about business domains](#)

8 items

Filter by keyword

+ New business domain

Business domain name	Owner	Status
Care Providers	NDMP 4 owners	Published
Claims	NDMP 4 owners	Published
Corporate Functions	NDMP 5 owners	Published
Finance	NDMP 6 owners	Published
Fraud Services	NDMP 4 owners	Published
Human Resources	NDMP 4 owners	Published
Personal Health	NDMA 3 owners	Published
Sales	NDMSM 14 owners	Published

Details

Roles

Data estate mappings

S Sales

Business domains

Set to draft

Mark as expired

Manage policies

Type

Parent

Owner

Status

Data quality score

Health actions

Functional unit

C Corporate Functions

NDMSM 14 owners

Published

Healthy 88.2

93

The sales team plays a crucial role within an organization. Their functions encompass various aspects of the sales process. They actively prospect and generate leads, evaluating their fit for the product or service. Salespeople must have a solid grasp of the product, enabling them to engage in consultative selling tailored to customer needs. Their ultimate goal is to negotiate and close deals, converting leads into paying clients. Post-sale, they focus on account management, maintaining client relationships and identifying upsell opportunities.

What matters most to the sales team? Revenue generation tops the list, as they strive to meet or exceed quotas. Additionally, they prioritize customer satisfaction, aiming to deliver value and foster repeat business. Monitoring the health of their sales pipeline, staying informed about market trends and competitors, and collaborating effectively with other teams are also key concerns. Lastly, salespeople care about their personal growth, continuously improving their skills and contributing to overall company success.

Business concepts

Data products

9

Go to data products

Glossary terms

41

View all

OKRs

3

View all

Critical data elements

4

View all

Data products

S Sales

Manage groups of data assets packaged together for specific use cases. [Learn more about data products](#)

+ New data product

Refresh

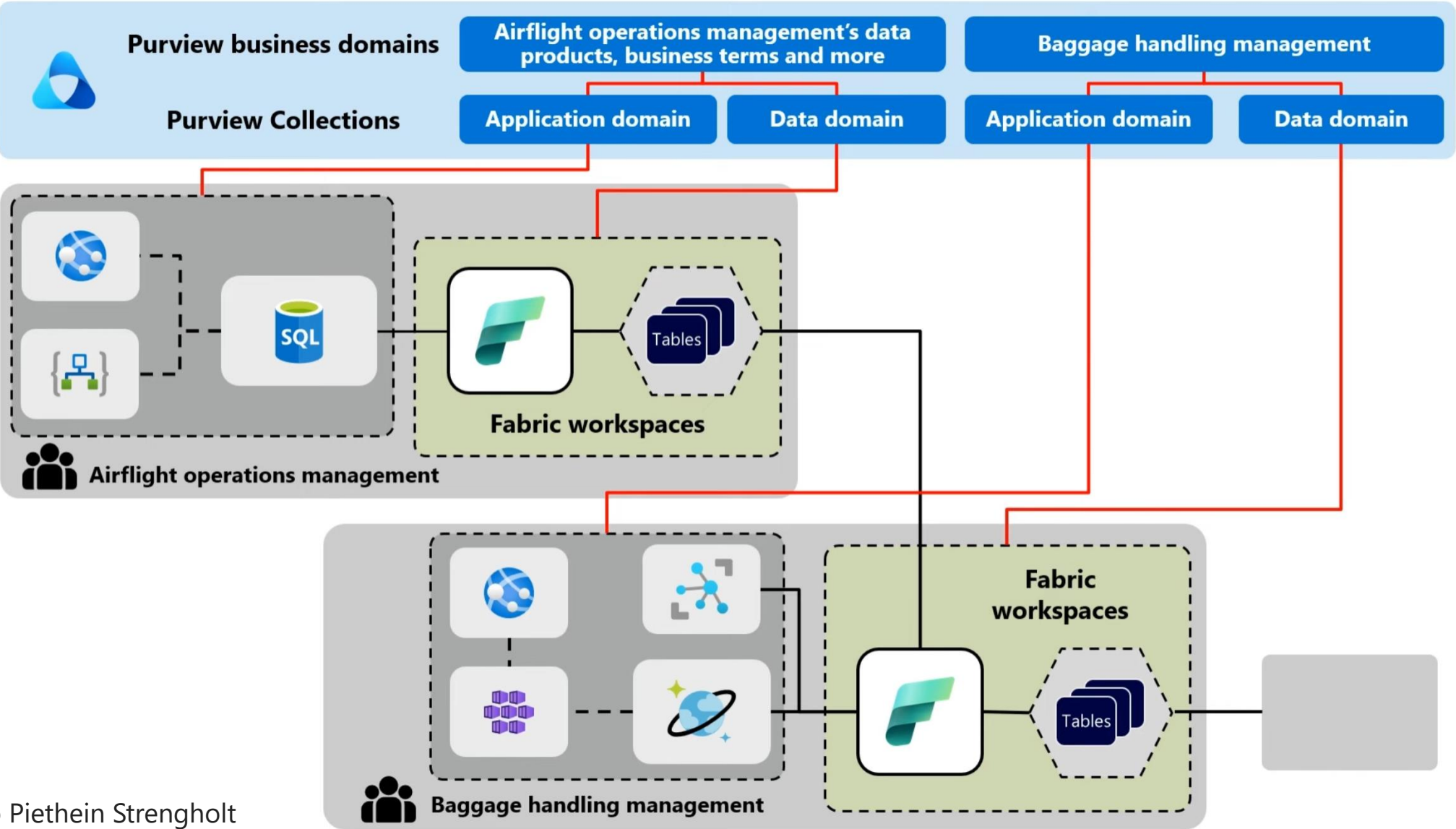
Status: All

Owner: All

Showing 1-9 of 9

Data product name	Business domain	Data assets	Owner	Status	Data quality score
Canada Sales Revenue Insights...	S Sales	4	MU Mahesh Uppala	Draft	--
Commercial Customer Accounts	S Sales	8	NDMSMU 3 owners	Draft	Healthy 93.1
Customer Master List	S Sales	9	NDMSM 3 owners	Published	Healthy 97.2
DE Sales Revenue Insights 2023	S Sales	1	MU A 2 owners	Published	--
Global Sales Revenue for 2023...	S Sales	13	NDMA 3 owners	Published	Healthy 100
North East Region Sales Reven...	S Sales	1	MU A 2 owners	Published	--
Opportunities conversion to s...	S Sales	4	NDMSM 4 owners	Draft	Fair 63.1
South West Region Sales Reve...	S Sales	3	MU A 2 owners	Draft	--
US Sales Revenue Insights 2023	S Sales	3	MU A 2 owners	Published	--

Purview and Fabric and Data Mesh



Q & A



James Serra, Microsoft, Data & AI Solution Architect

Email me at: jameserra3@gmail.com

Follow me at: @JamesSerra

Link to me at: www.linkedin.com/in/JamesSerra

Visit my blog at: JamesSerra.com

(email me for the deck)