Microsoft

Hurtownia danych, data lake, a może lakehouse – czyli o pułapkach, wyzwaniach i jednorożcach świata analityki

Azure Club
https://aka.ms/azure-club

What we've heard about analytics

Barriers to achieve business outcomes

Ideally, organizations want to have.....

- Platform to actionable Insights to the business
- Robust data governance
- Ability to increase the value of hidden data
- To spend less time preparing data
- Improve Operational Efficiency

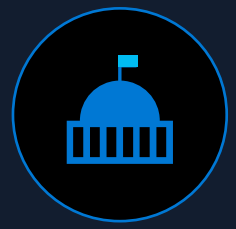
Roadblocks to deriving value from data

Complexity

- Elaborate architectures that don't scale.
- More teams engaging with data, not just engineering.

Uncertainty

- Constant evolution of tools and options.
- Lack of clear guidance on where to start and how to bring value.



Governance

- Prioritizing security prevents access to data and stifles innovation.
- Lack of consistent processes and policies create data silos.

Skilling

- On-prem knowledge and skill sets don't translate to cloud-based service.
- Need for skilling to adapt to new ways of working.

Time to value

- New solutions can take months to deploy and achieve ROI.
- Inability to provide real-time, trusted data to inform timely business decisions.

Develop and optimize your data strategy and platform

Leverage cloud scale analytics for best practices and executable resources

Cloud Adoption Framework
Align business and technical strategies for the cloud using Microsoft guidance, best practices, documentation, and tools.

Reference architectures
Leverage modular building blocks to grow your infrastructure over time and adapt the scenario to your use cases.

Assessments
Identify gaps in governance, evaluate workload performance, understand your DevOps capabilities, and more.

Well-Architected Framework
Follow key principles of architectural excellence to guide workload development and optimization.

Deployment templates
Deploy infrastructure and create new resources with repeatable templates and declarative syntax.

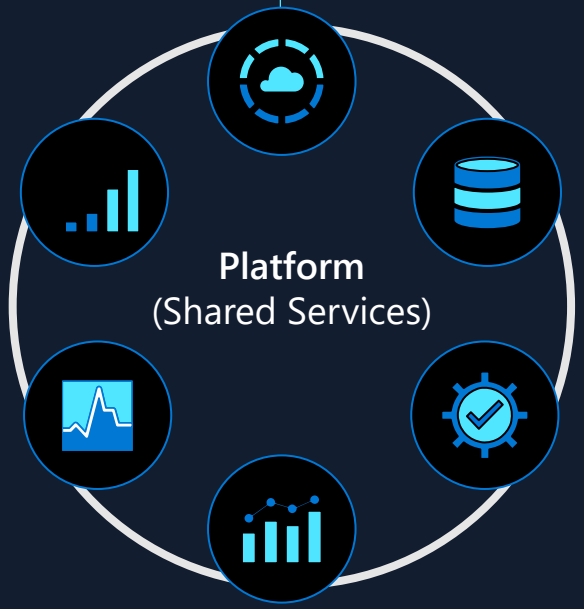
Skilling resources
Map the skills needed to drive your data strategy to your existing capabilities. Build new skills through learning resources and certifications.

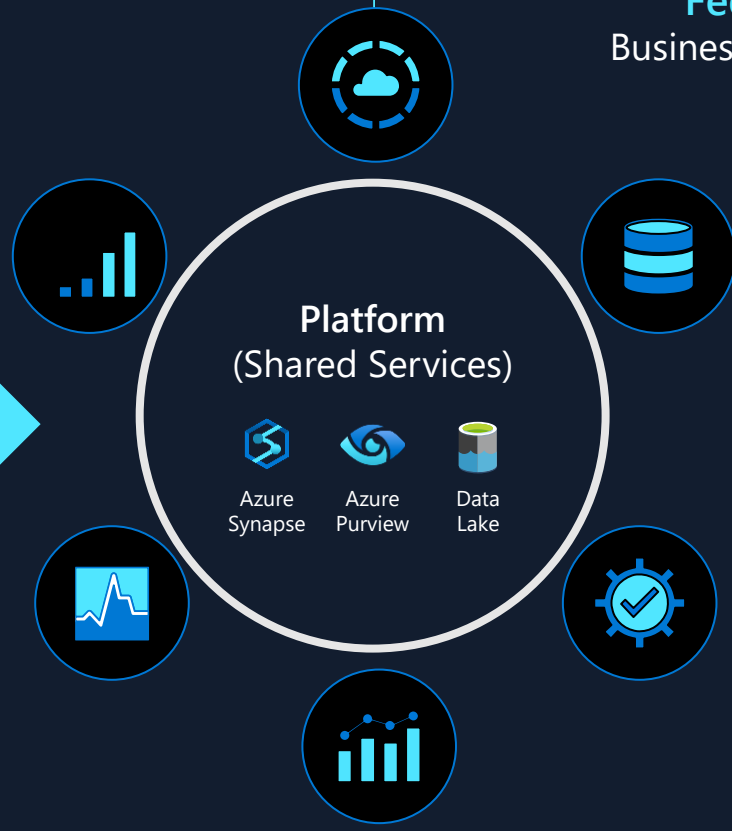
Empower teams by facilitating data self-service

Increase data agility and accelerate time to value.

Centralized data ownership
IT centrally owns data and all services

- Centralized architectures
- Competing business needs
- Workarounds to enable agility



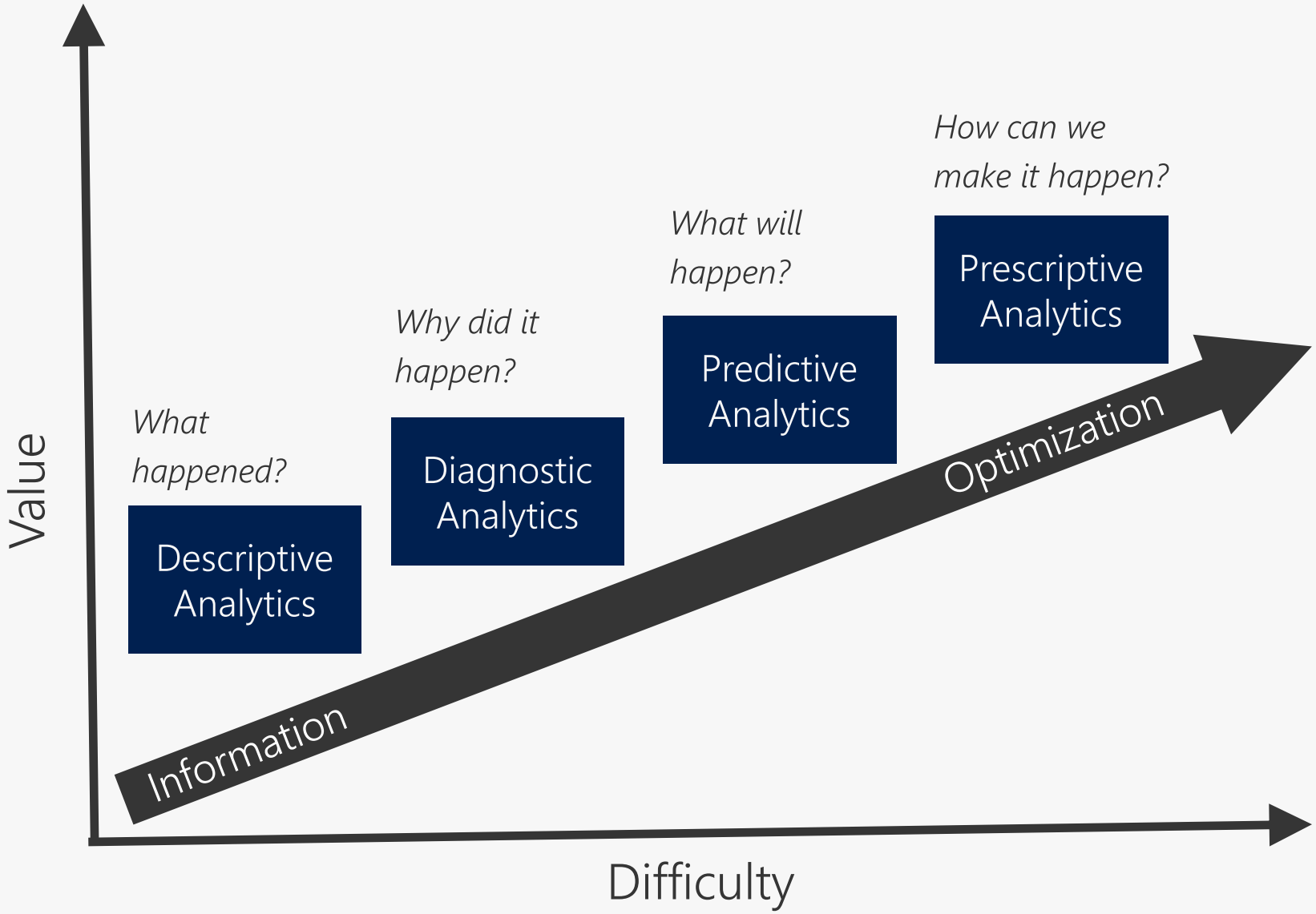


Federated data ownership
Business owns data-driven projects

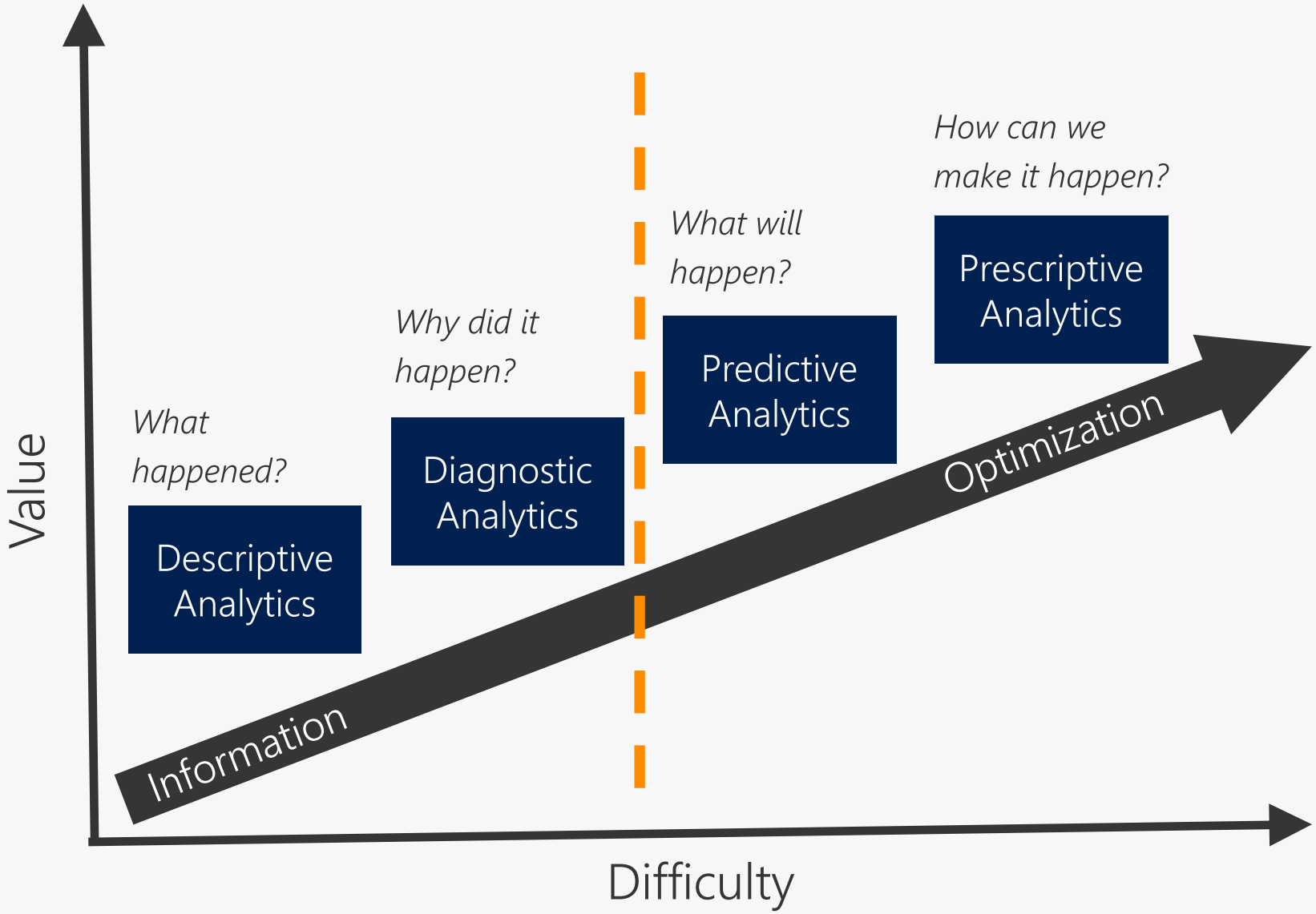
- Common governance and management
- Improves quality and security
- Increases efficiency and access

Let’s take a step back…

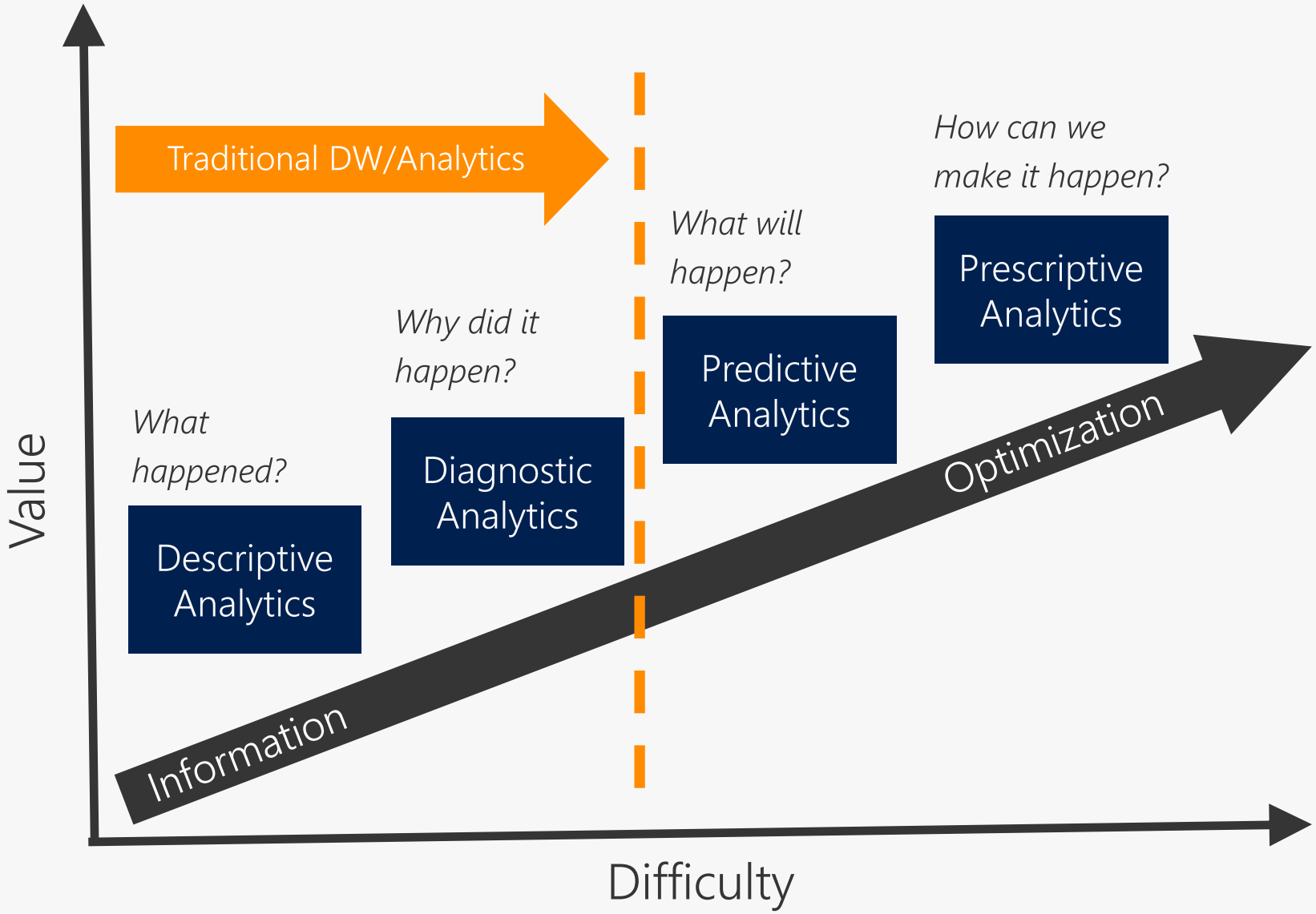
Types of Analytics



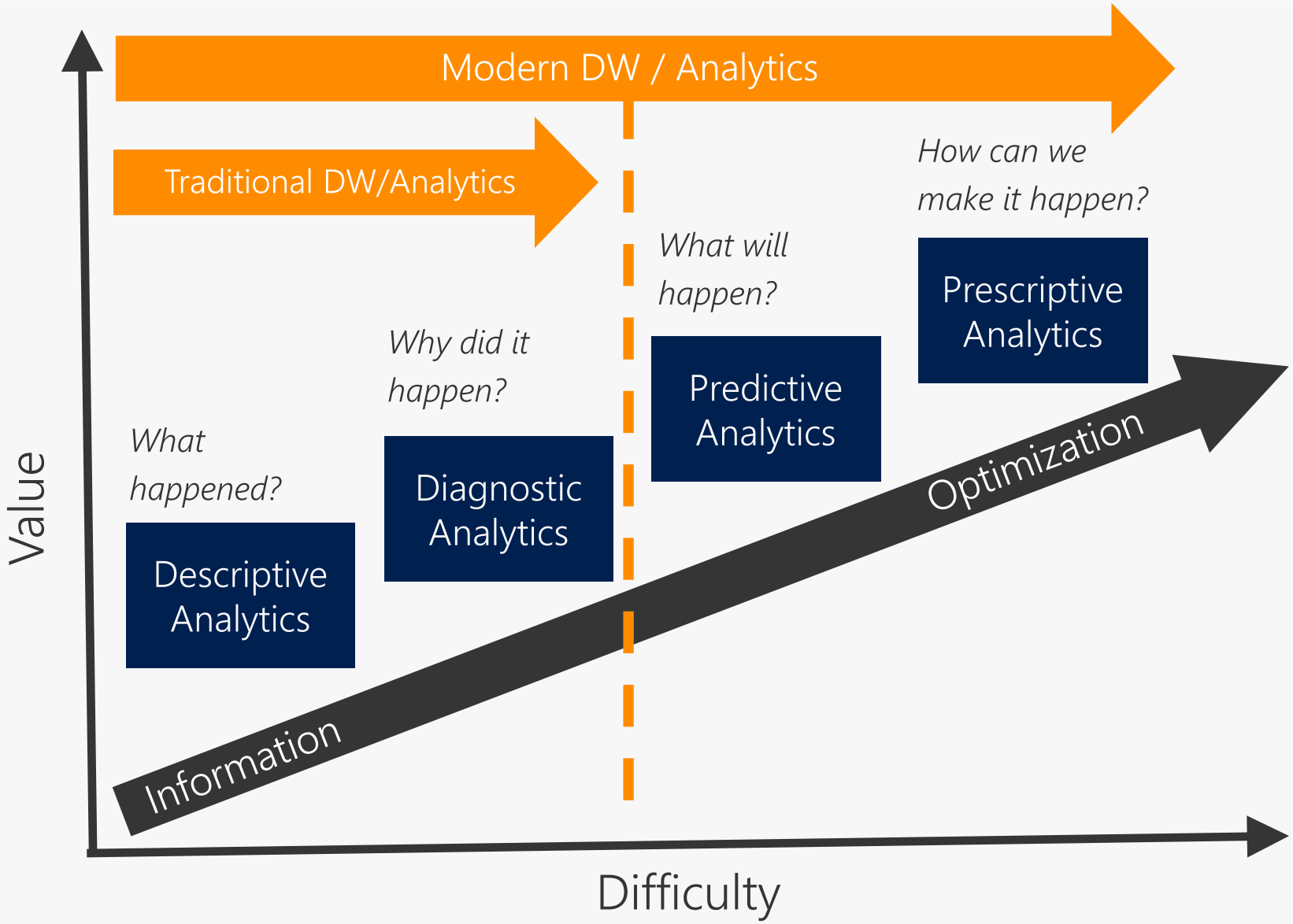
Types of Analytics



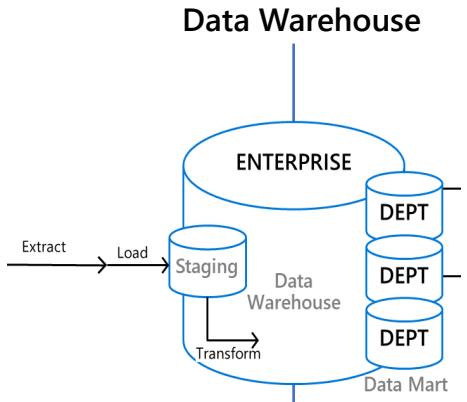
Types of Analytics

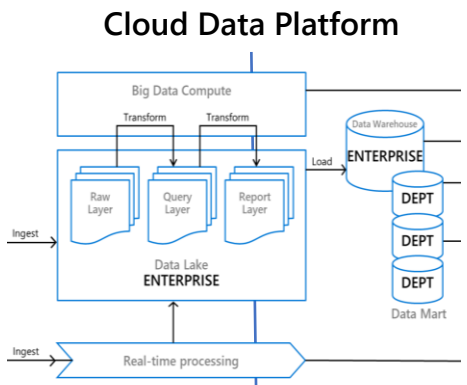


Types of Analytics

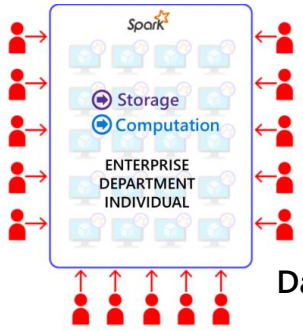


Evolution of Data & Analytics Architecture

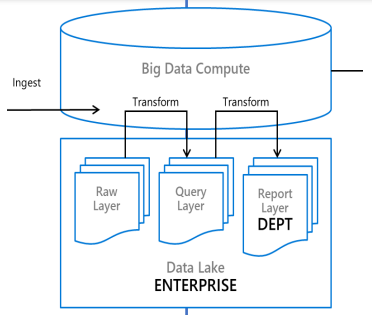




1980 1990 2000 2010 2020



Data Lake



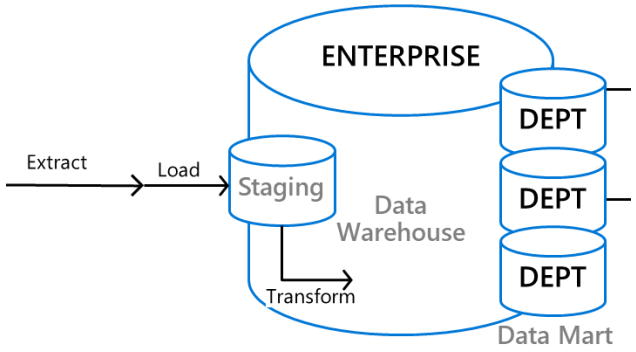
Data Lakehouse

Centralized

Evolution of Data & Analytics Architecture

Late 1980s

Data Warehouse



- **Central repository**
- Typically used for business reporting and analysis
- **Highly structured**
- Data is **cleaned and deduplicated**

PROs:

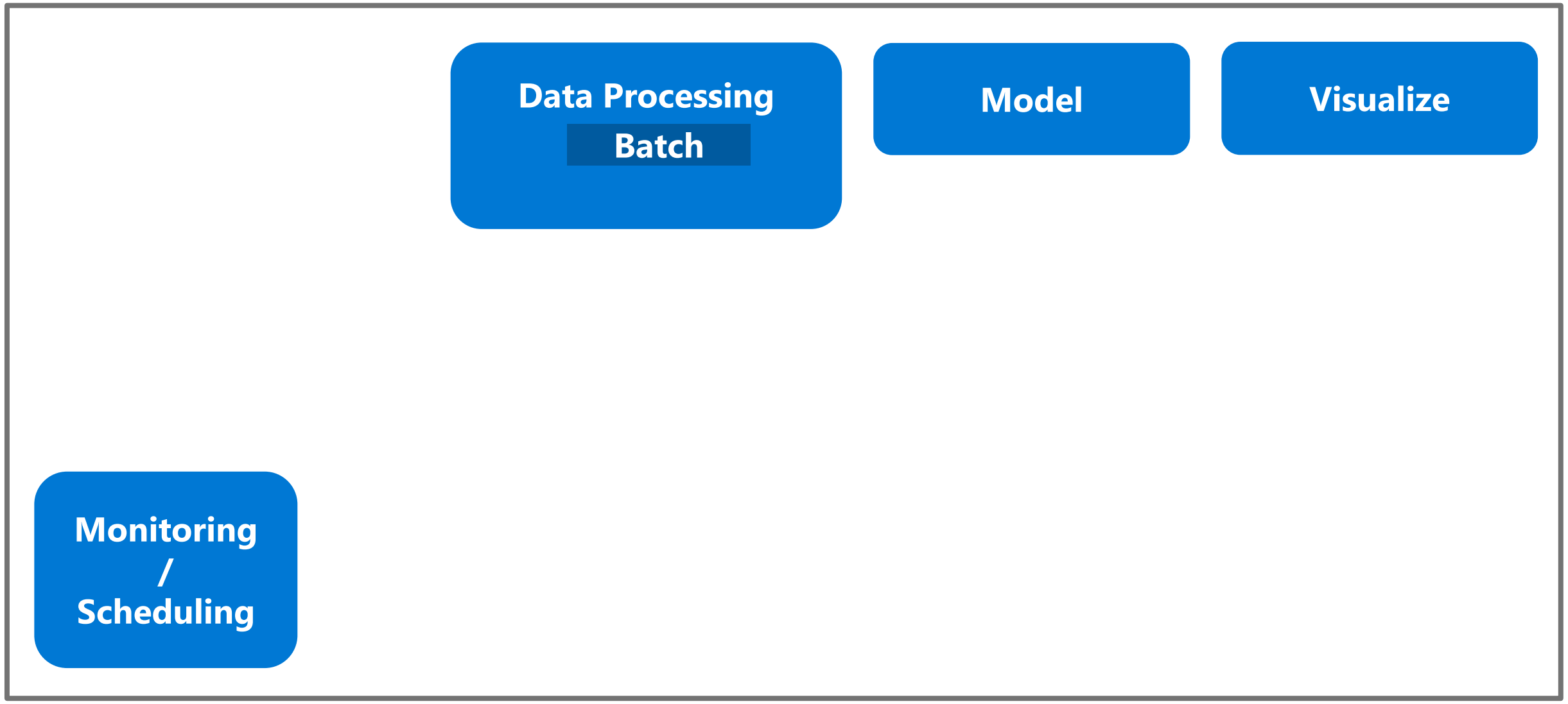
- Solves the problem of silos of operational data
- Decision support for authorized business users + executives
- Improved data quality allows for consistent business steering

CONs:

- Design + deployment is large undertaking
- Flood of data integration processes is hard to maintain
- Vendor lock-in (hard to migrate)

=> high cost, complex, not scalable

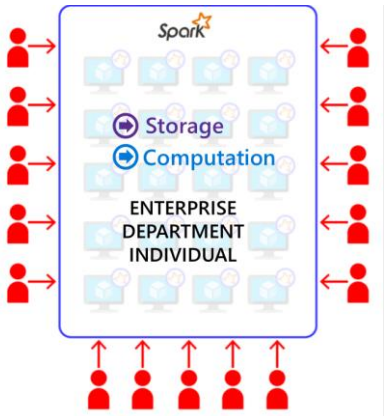
Data Architecture – Building Blocks



Evolution of Data & Analytics Architecture

Late 2000s

Data Lake



- **Central repository**
- Data is stored in it's **natural format**
- Can store all kinds of data, **structured + unstructured**
- **Schema-on-read**

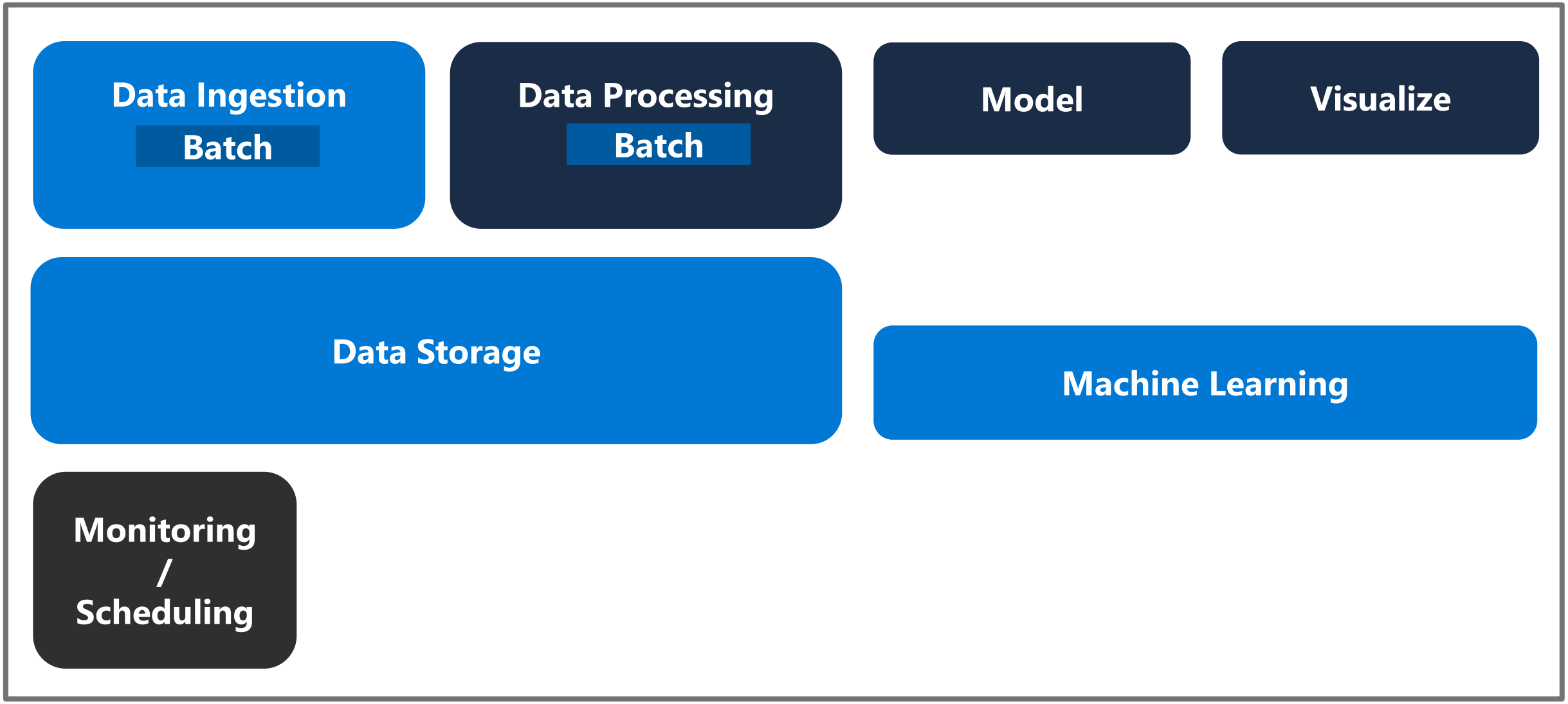
PROs:

- Holds all enterprise data, not just the one with specific need
- Supports Data Science + ML
- Flexible
- No vendor lock-in - **open data formats** and protocols

CONs:

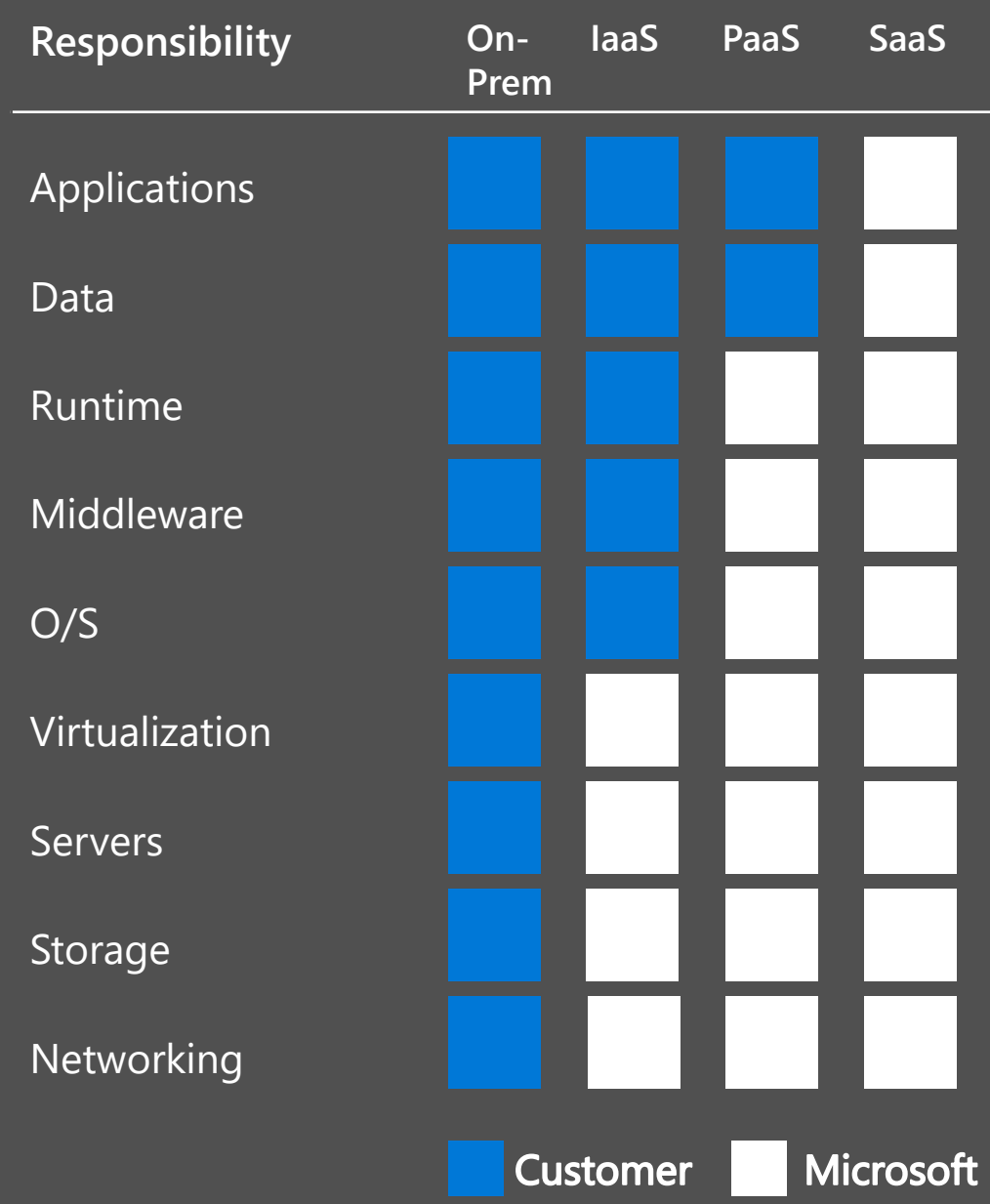
- Hard to govern and maintain
- Data might be inconsistent and unclean

Data Architecture – Building Blocks



Moving to the cloud

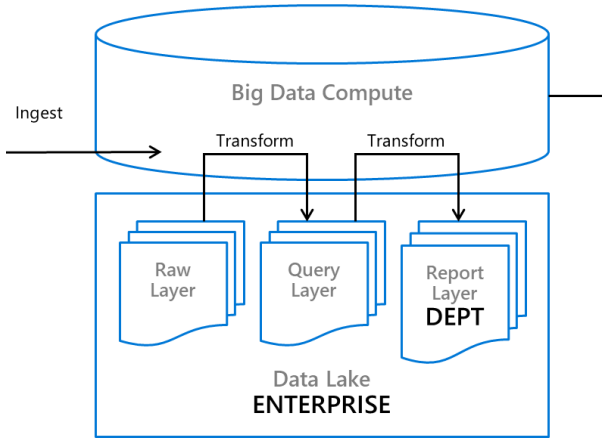
- Lift and Shift existing data solutions to the cloud
  - ✓ Scaling
  - ✓ No capital expenditure / hardware
- Platform as a Service Offerings
  - ✓ take advantage of new features instantly
  - ✓ Less operational effort
  - ✓ Backups out of the box
  - ✓ Default Replication
  - ✓ Built-in usage of cloud scaling capabilities
  - ✓ Elasticity
  - ✓ Serverless
- Easier access in distributed data landscapes



Evolution of Data & Analytics Architecture

Early 2020s

Data Lakehouse



- **Central repository**
- Combines the **best of DWH and Data Lake**
- Uses **cheap storage** of the Data Lake
- Has similar Data Management features like DWH
- Separate **scaling of storage and compute**

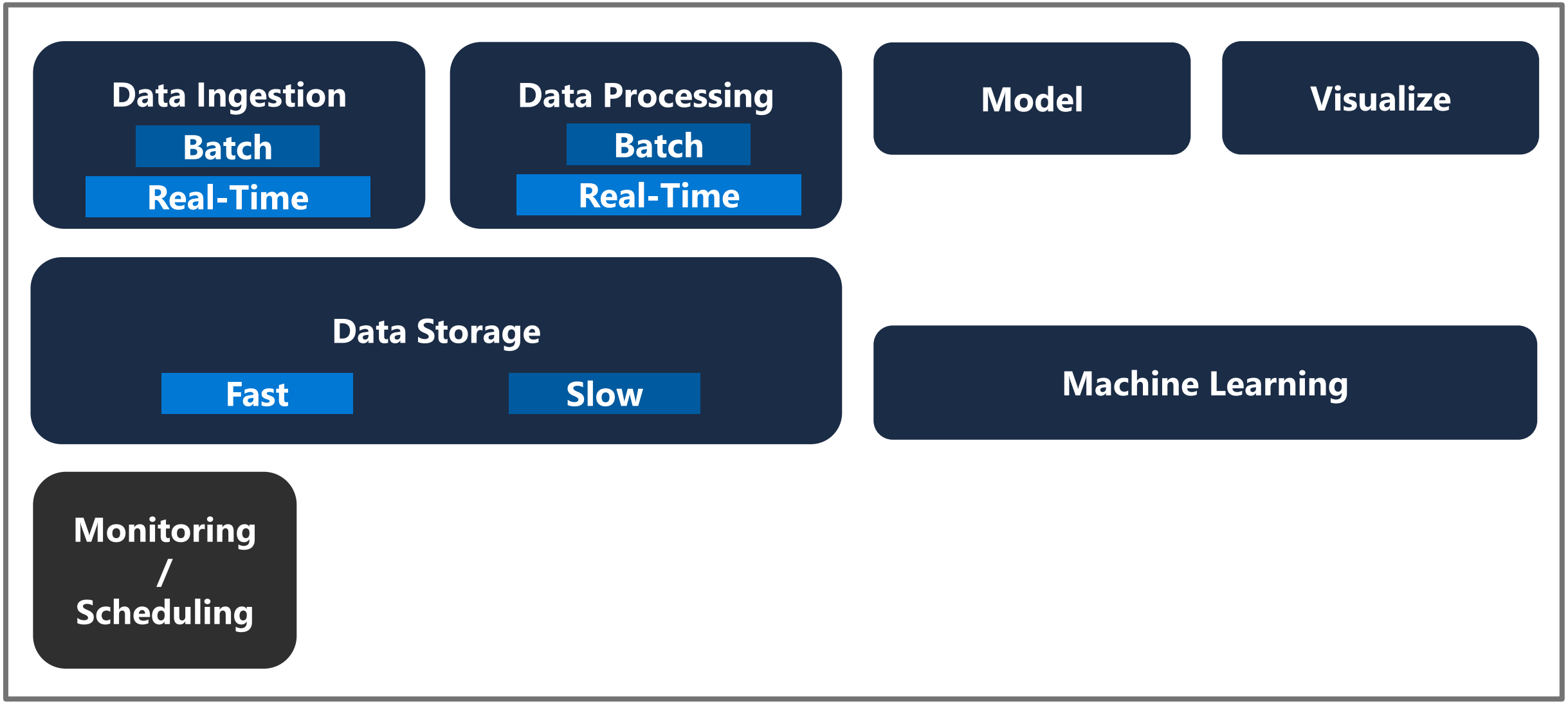
PROs:

- Cloud-native
- Separate scaling of storage and compute
- Supports diverse workloads

CONs:

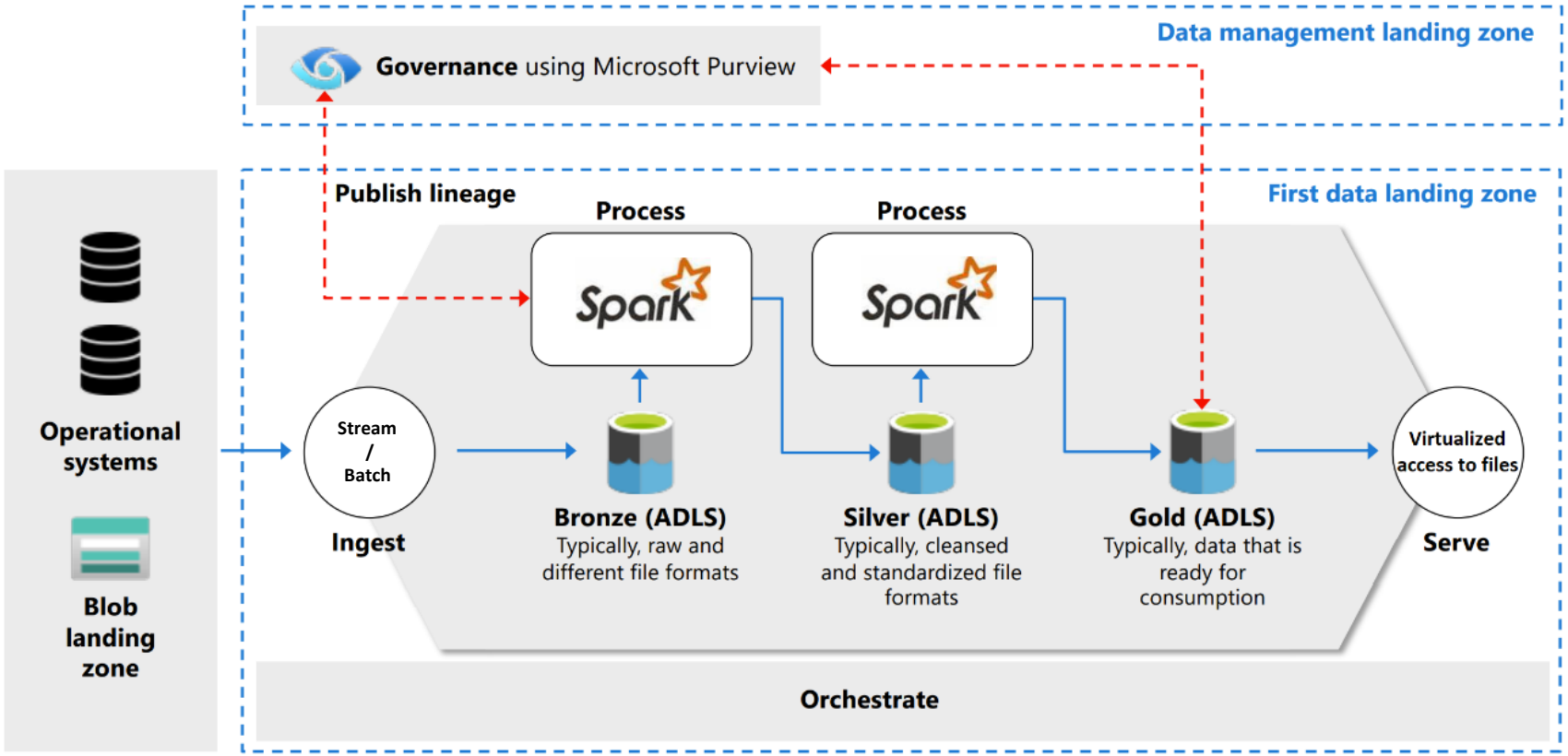
- Data teams might be bottleneck
- Data consumers, data teams and data producers are separated => worse data quality

Data Architecture – Building Blocks



Cost Optimization | Operational Excellence | Performance Efficiency | Reliability | Security

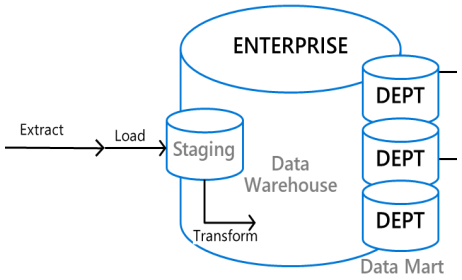
Where We Land:
Lake-centric (Medallion)



Evolution of Data & Analytics Architecture

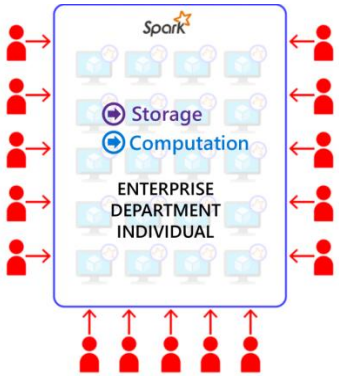
Late 1980s

Data Warehouse



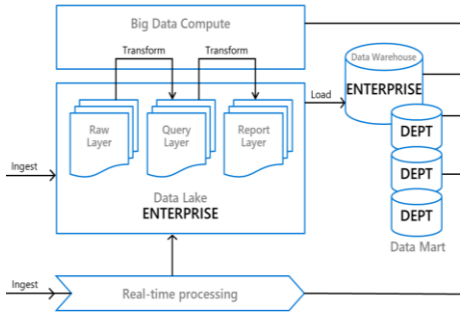
Late 2000s

Data Lake



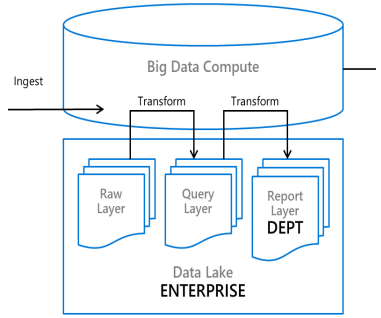
Mid 2010s

Cloud Data Platform



2020

Data Lakehouse

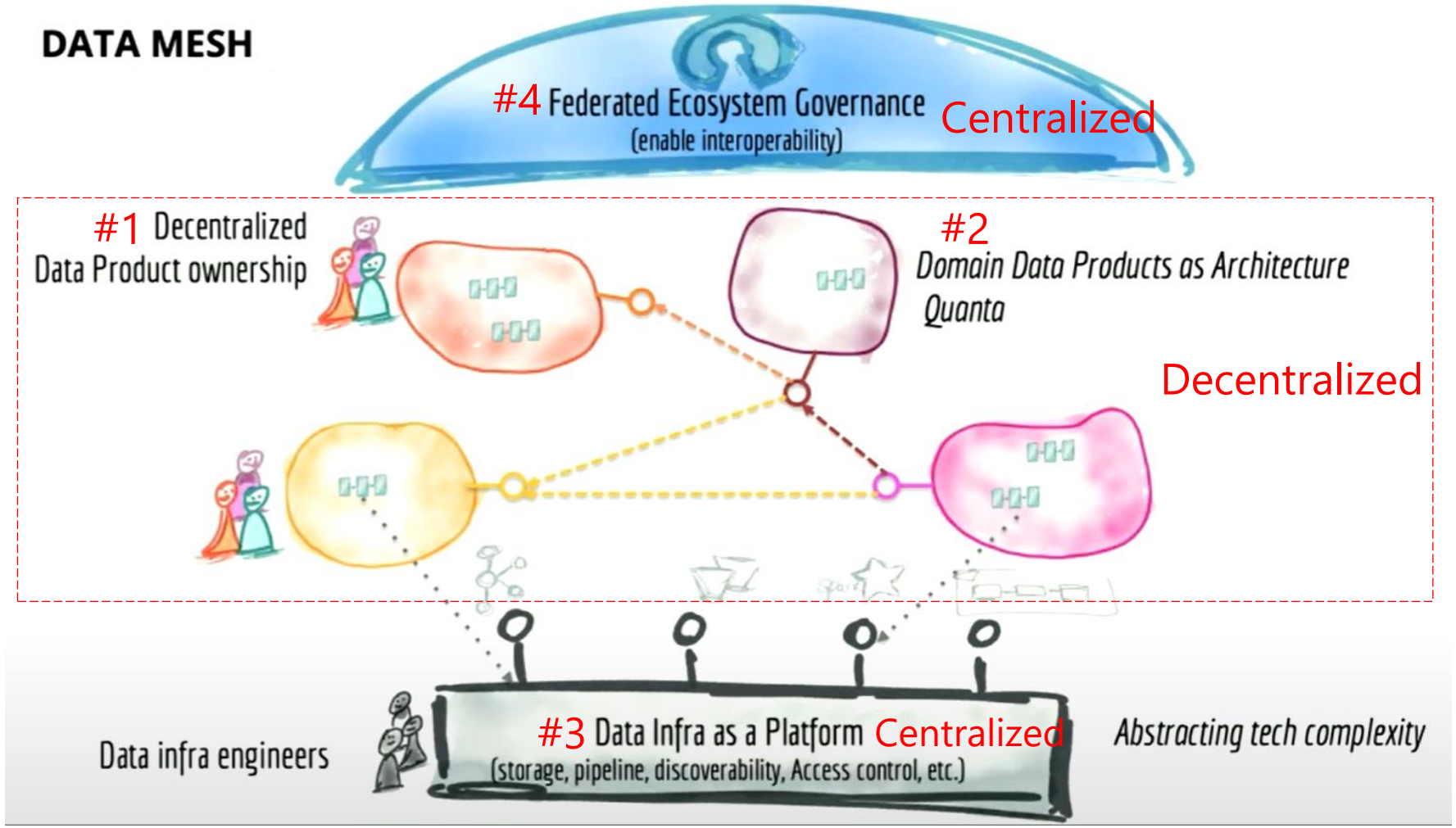


Emerging

Data Mesh?

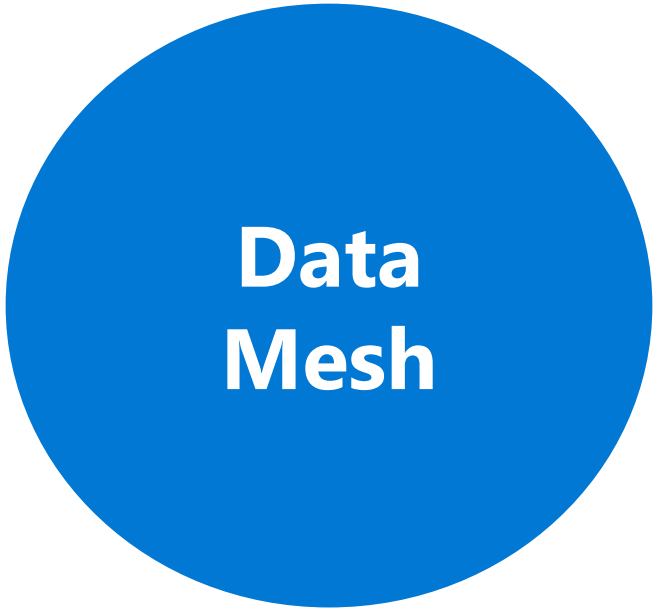
Centralization

Operating Model



Credit: Zhamak Dehghani, Thoughtworks

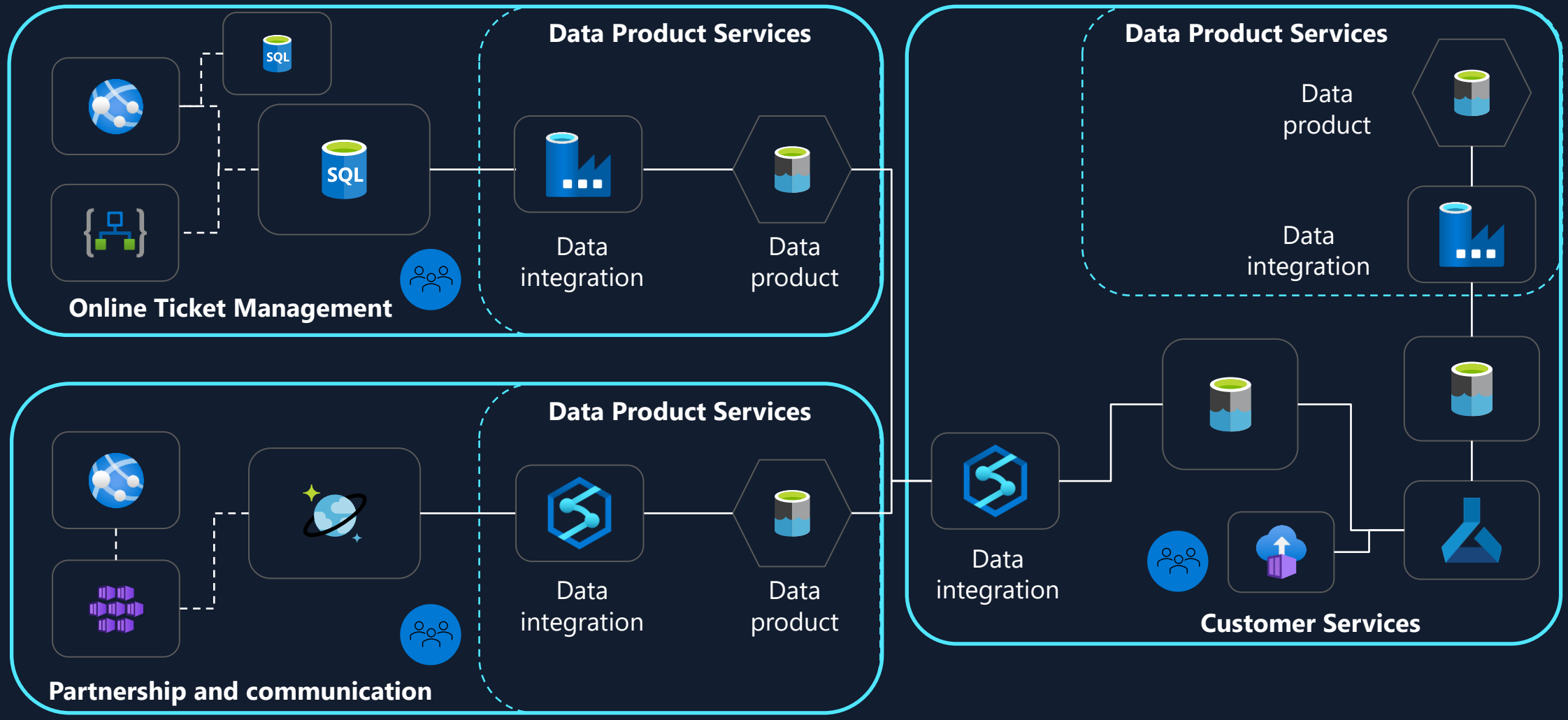
A new paradigm



- Decentralized data architecture
- introduces a new organizational perspective
- apply domain-driven design and product thinking to data and analytics
- Producers of the data are responsible for the data
- Data Team manages common data infrastructure

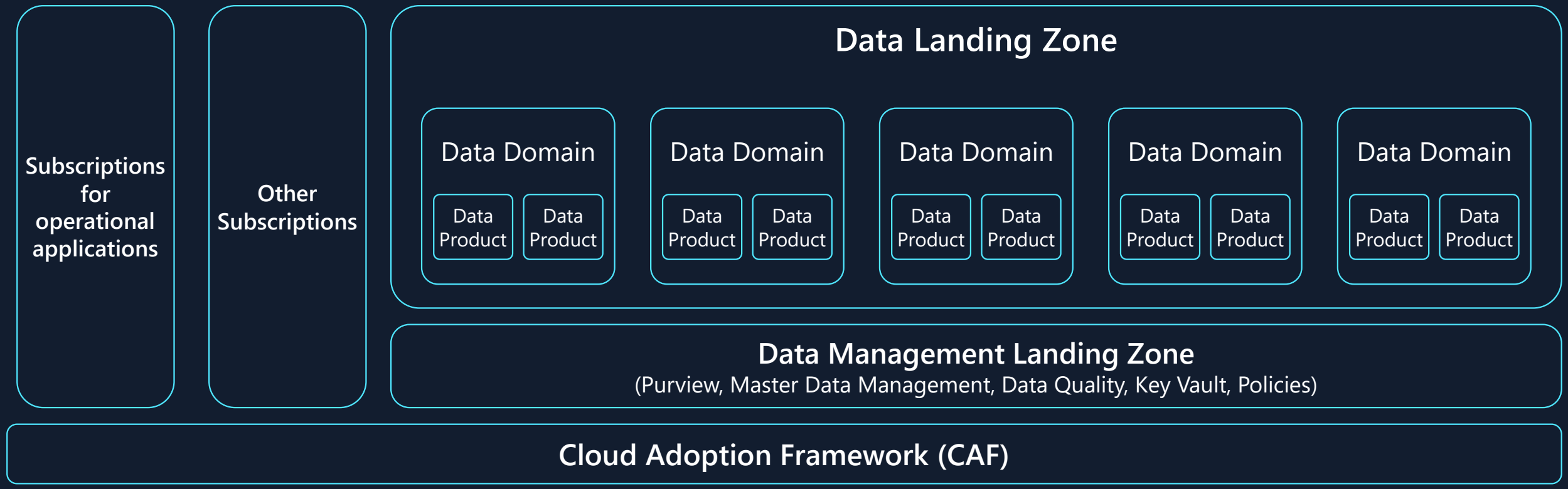
Data Domain & Data Product Decomposition

Example: Collaboration between different domains



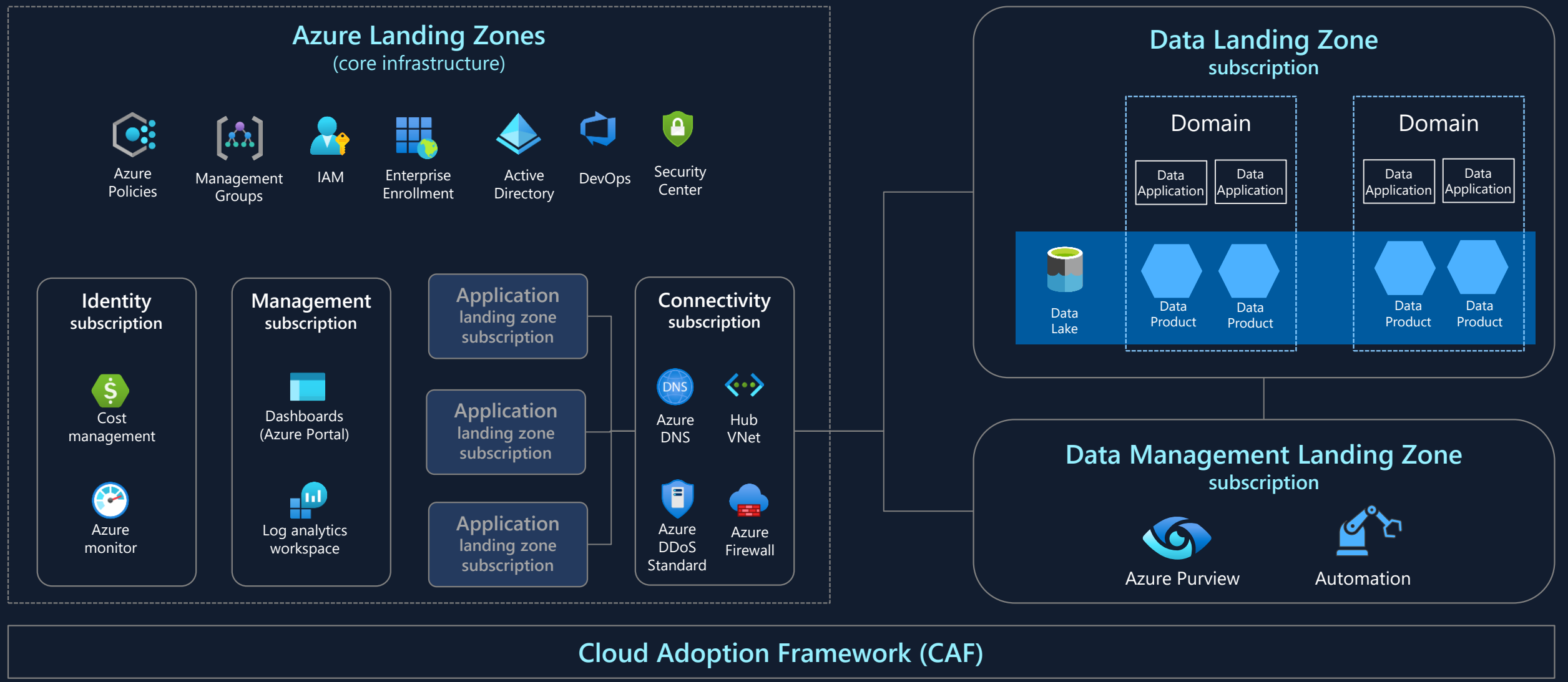
Data Mesh Pattern

Single landing zone

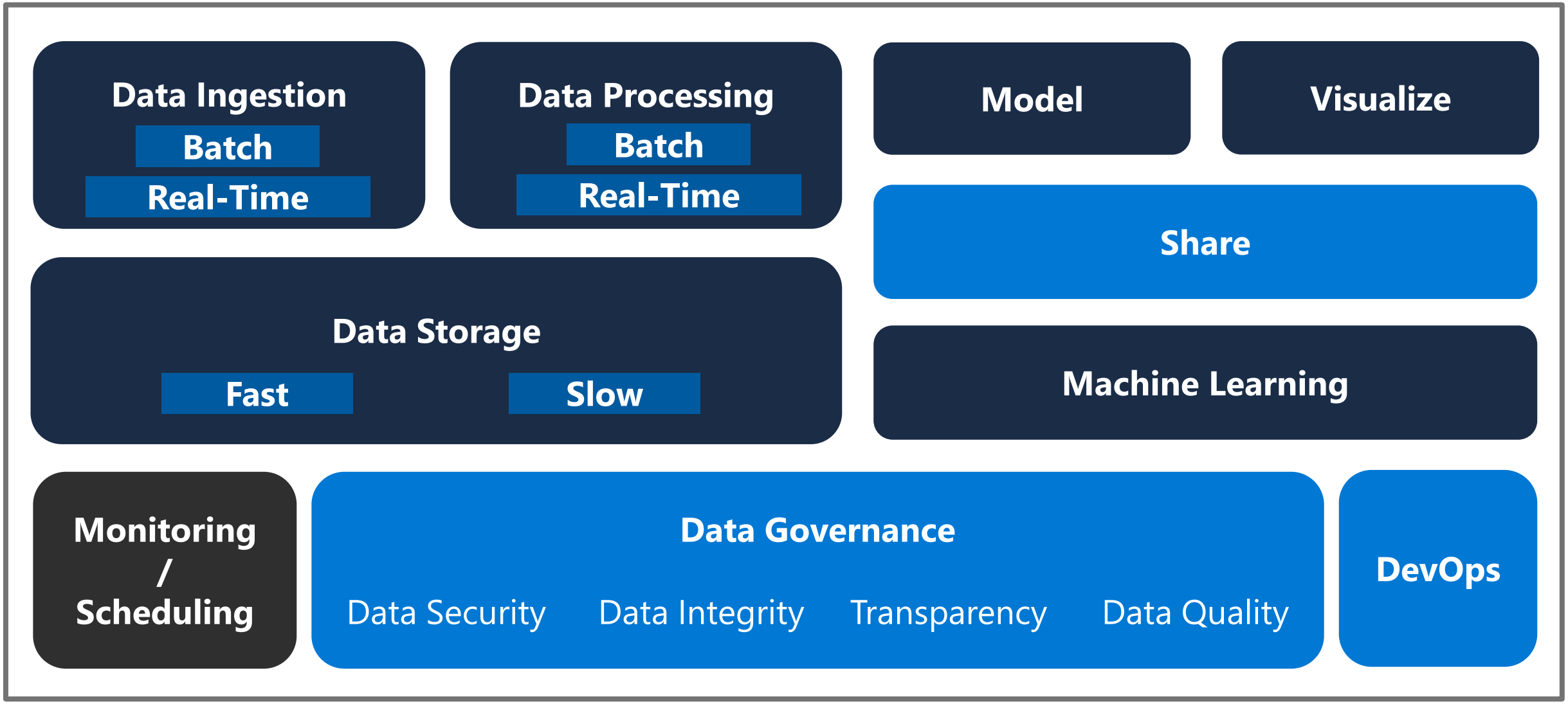


Microsoft

Implementation Scenario: Single Data Landing Zone w/ Multiple Domains



Data Architecture – Building Blocks

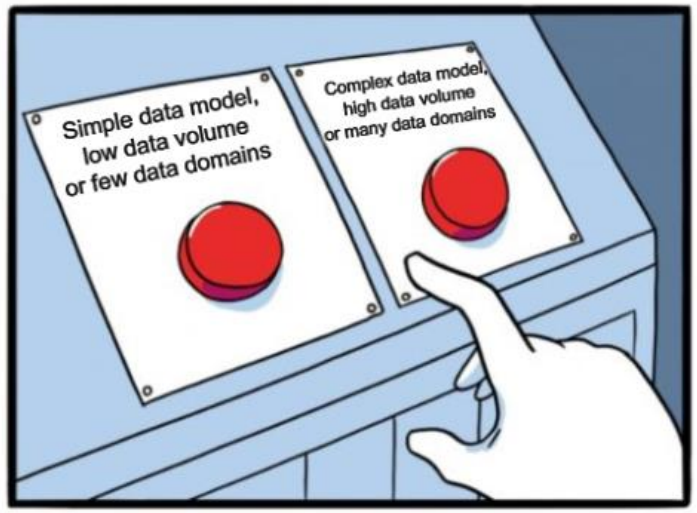


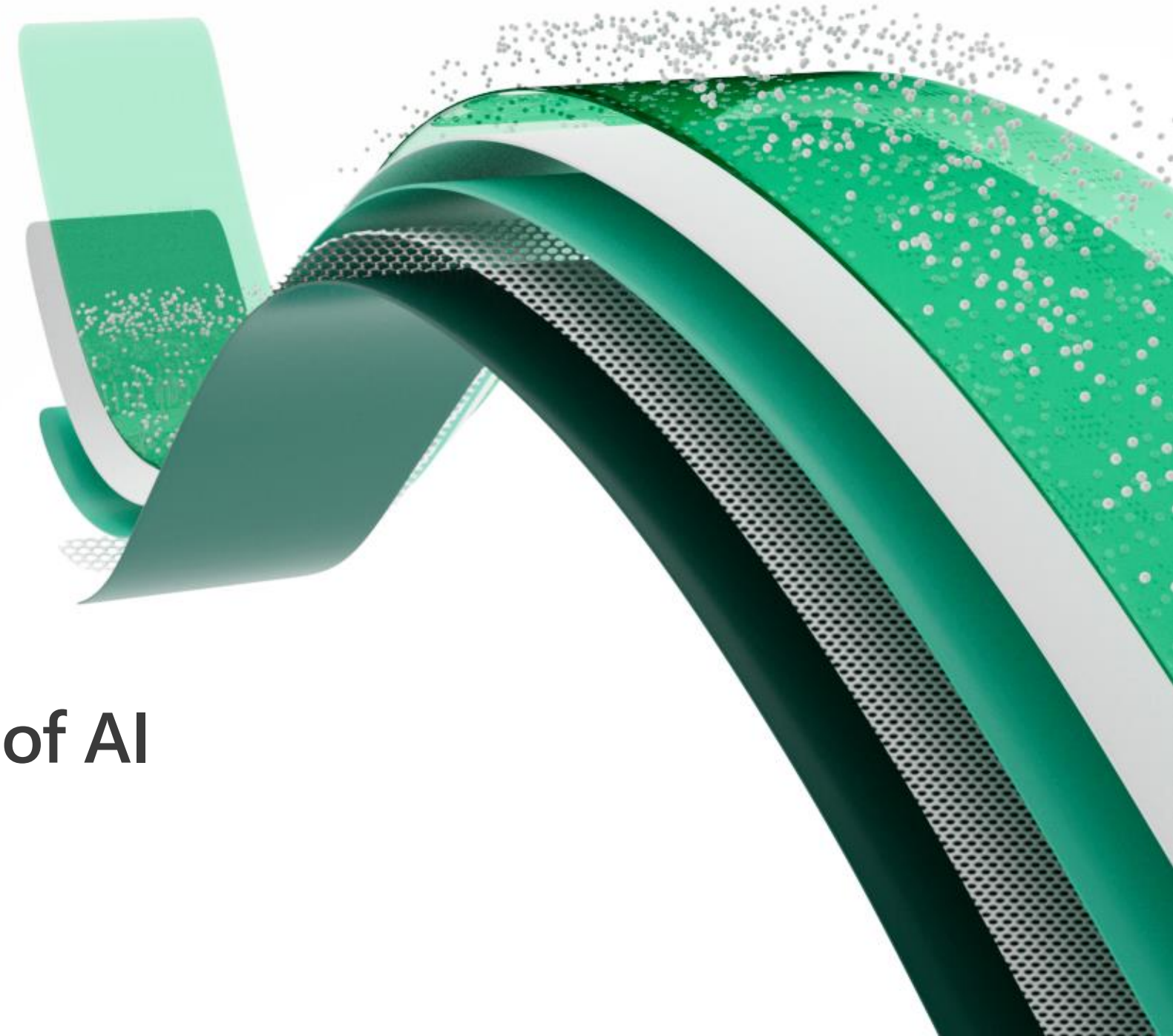
Cost Optimization | Operational Excellence | Performance Efficiency | Reliability | Security

Is Data Mesh a Replacement for All?

- Data Mesh = business operating model for analytics and data

- NOT a replacement of the Data Warehouse





Analytics in the Era of AI

Microsoft Fabric

The data platform for the era of AI

Microsoft Fabric

The data platform for the era of AI

Complete Analytics Platform

Lake Centric and Open

Empower Every Business User

AI Powered

Microsoft Fabric

The data platform for the era of AI

Complete Analytics Platform

- Everything, unified
- SaaS-ified
- Secured and governed

Lake centric and open

- OneLake
- One Copy
- Open at every tier

Empower Every Business User

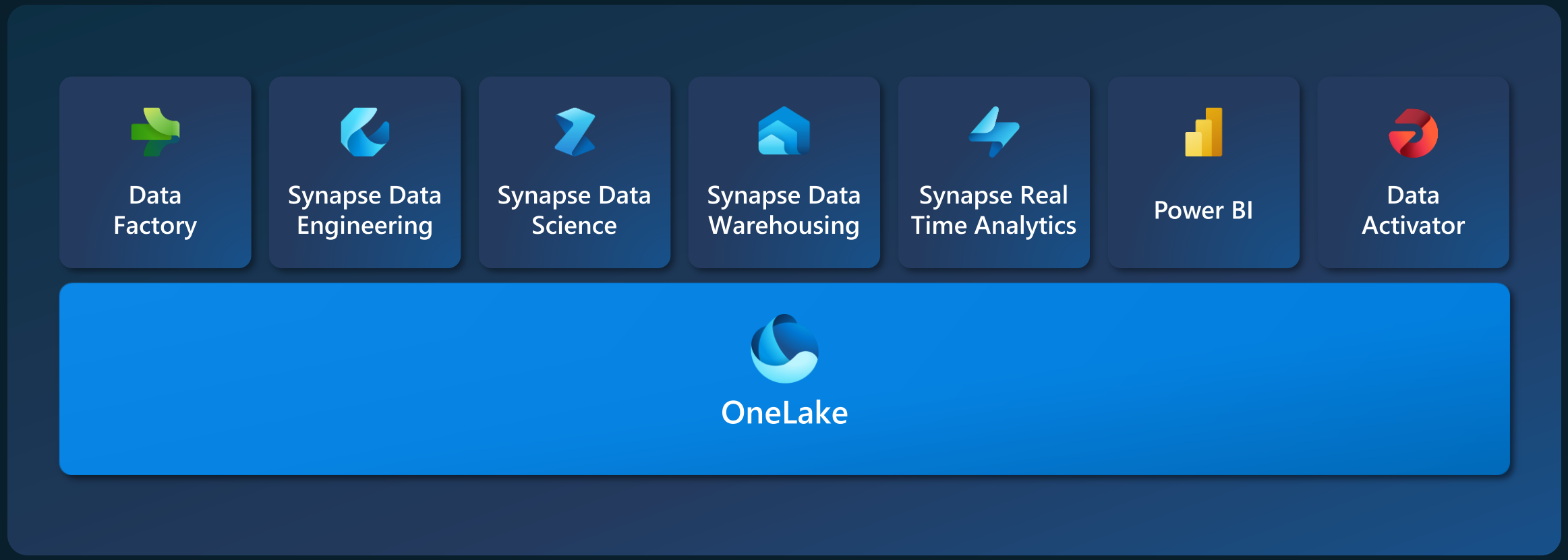
- Familiar and intuitive
- Built into Microsoft 365
- Insight to action

AI Powered

- Copilot accelerated
- ChatGPT on your data
- AI driven insights

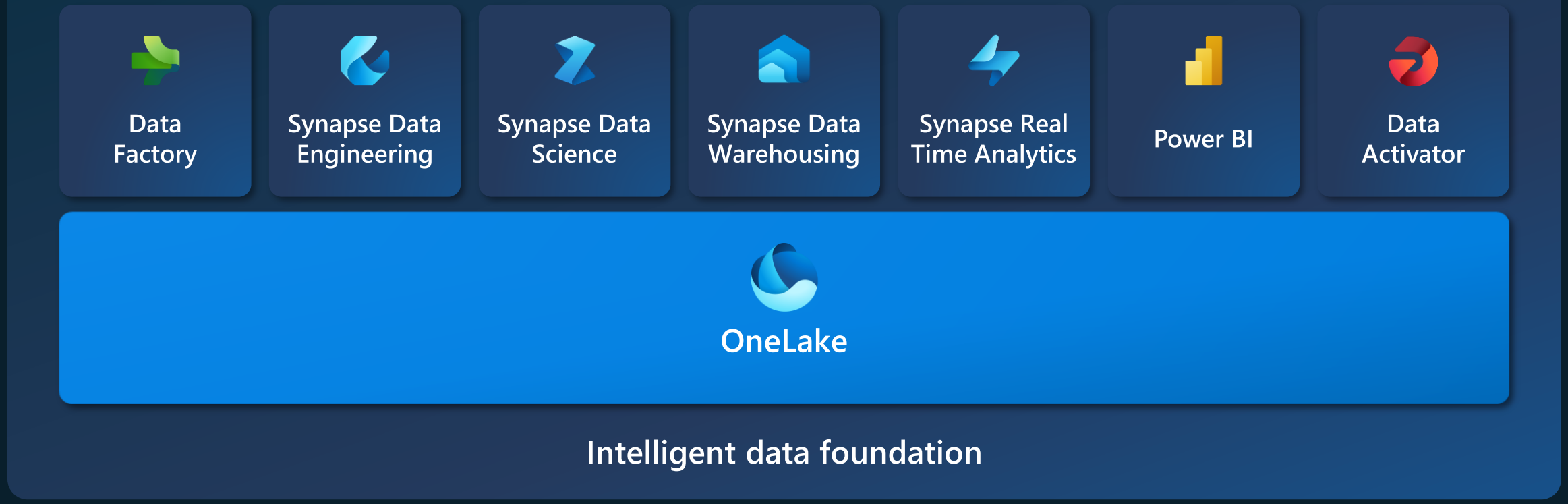
Microsoft Fabric

The unified data platform for the era of AI



Microsoft Fabric

The data platform for the era of AI



Next steps

1. Join the keynote and live simulcast
   aka.ms/BuildWithAnalytics
2. Try Microsoft Fabric
   aka.ms/try-fabric
3. Engage in the community
   aka.ms/fabric-community

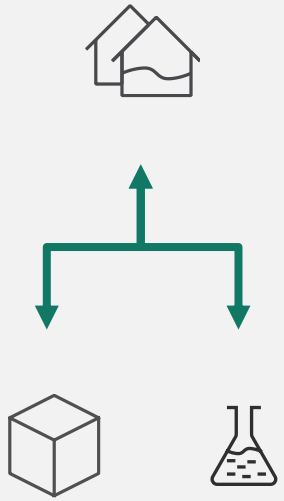
Common analytics scenarios addressed by Microsoft Fabric

Common analytics scenarios

Lakehouse

Data Warehouse

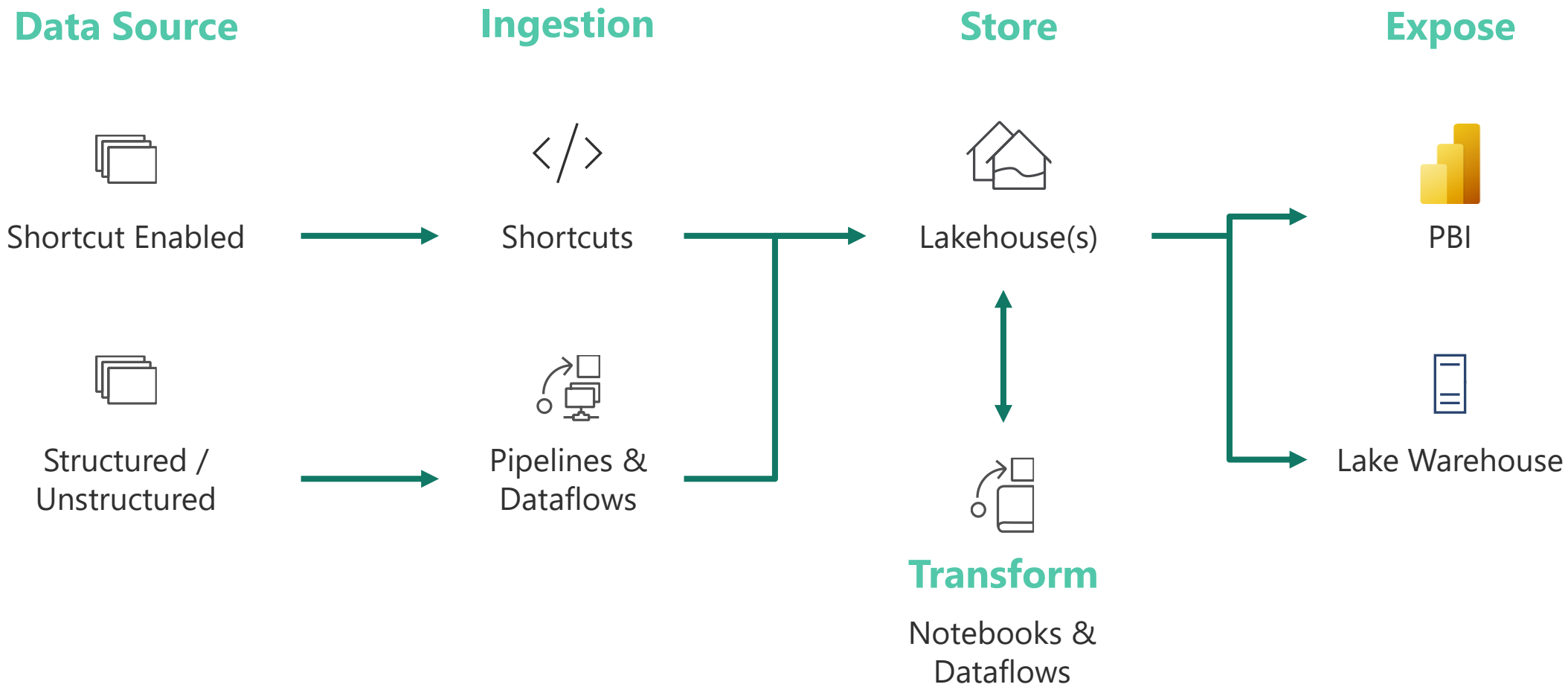
Data Science



Real Time Analytics

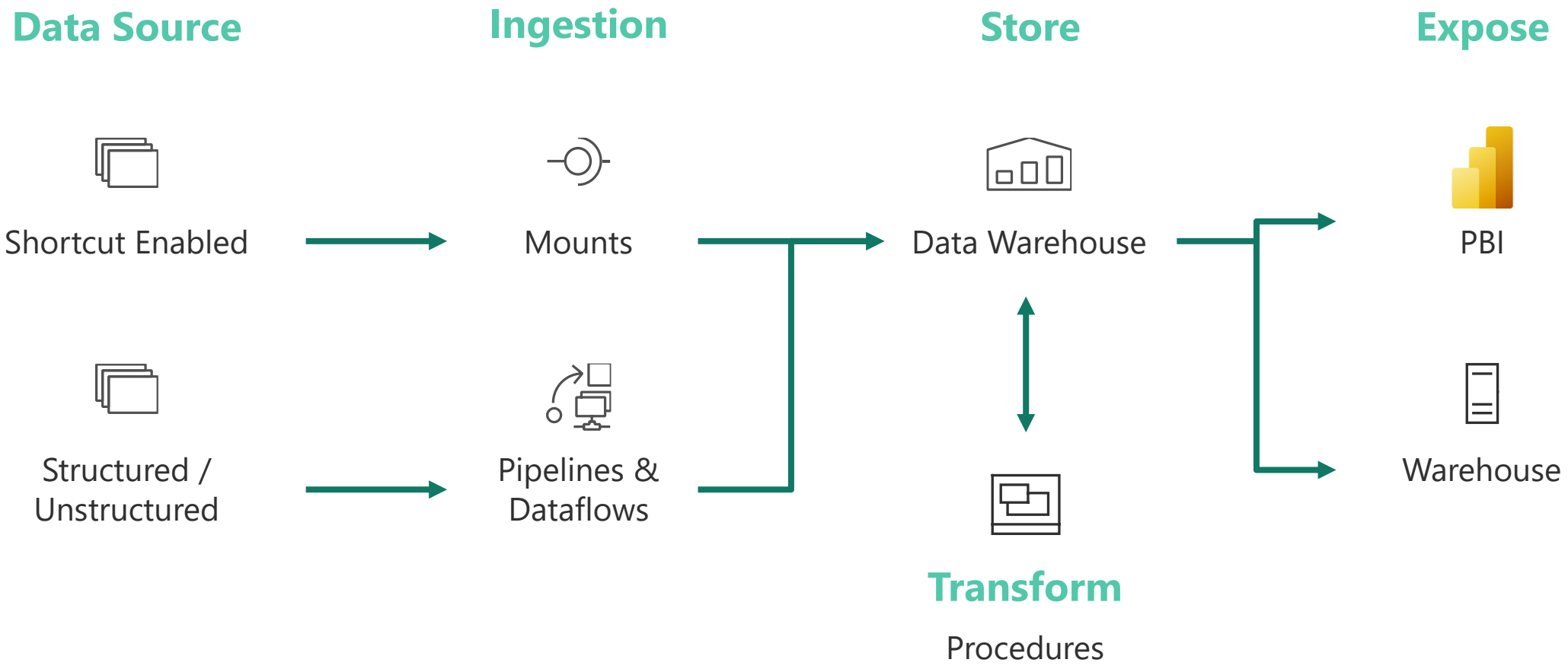
Microsoft Fabric Integration with Azure Synapse, Databricks, etc.

Lakehouse



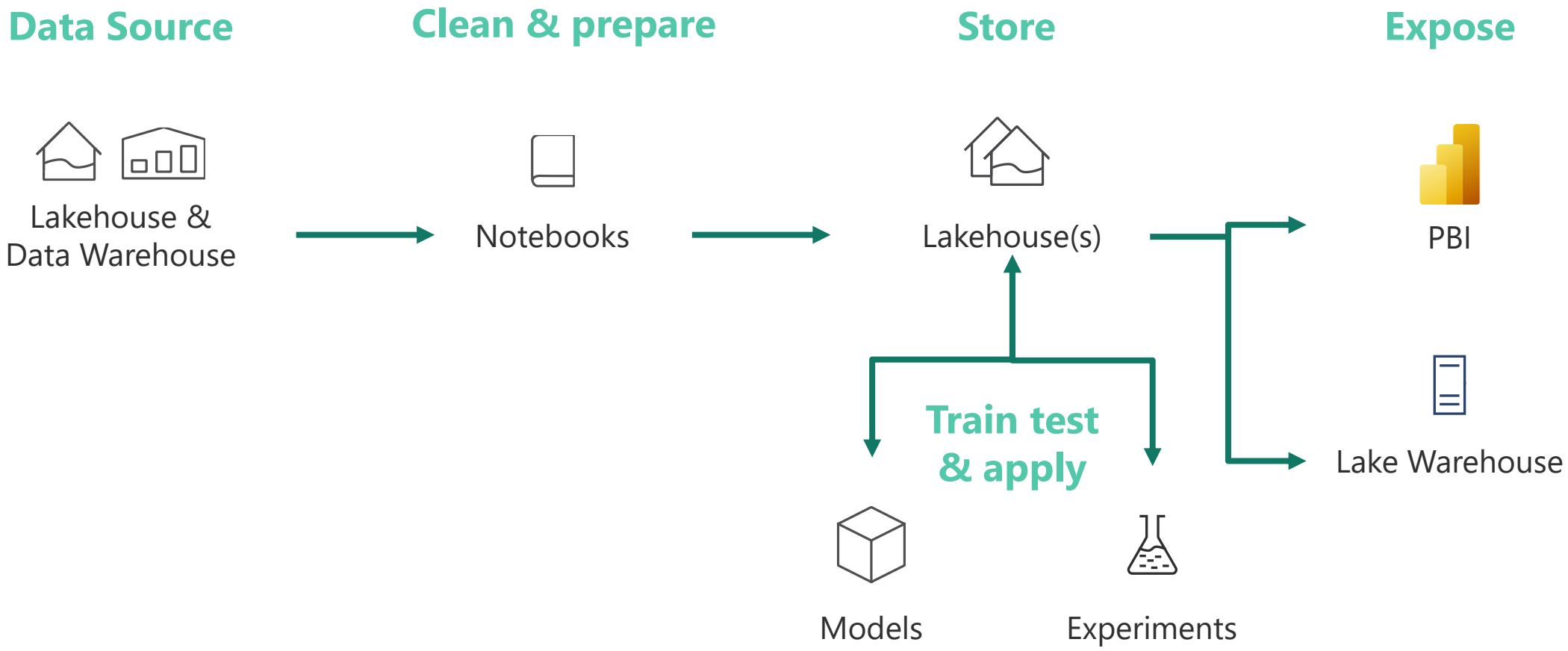
[Lakehouse end-to-end scenario: overview and architecture - Microsoft Fabric]

Data warehouse



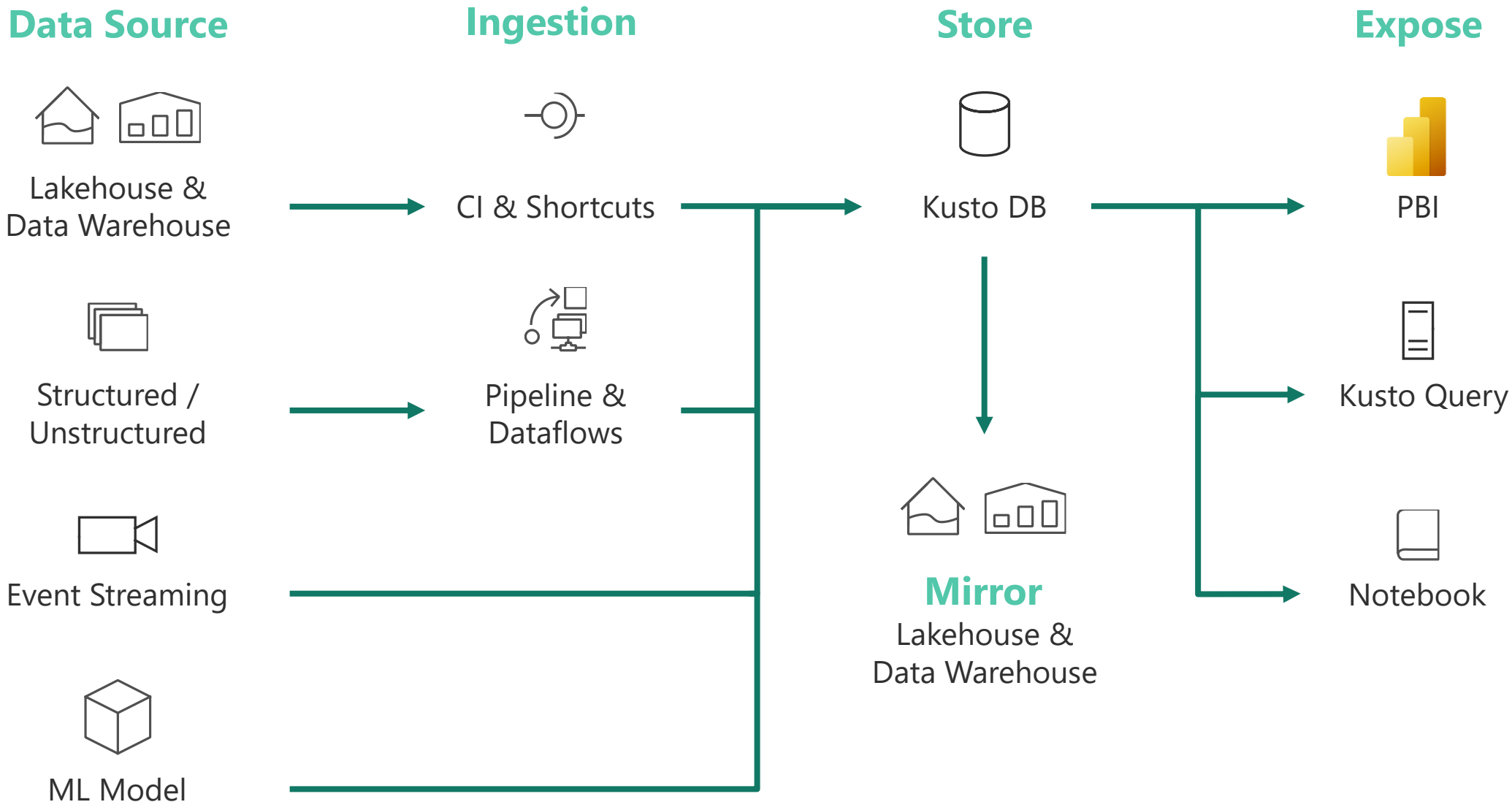
[Data warehouse tutorial - create a workspace - Microsoft Fabric]

Data science



[Data science tutorial - get started - Microsoft Fabric]

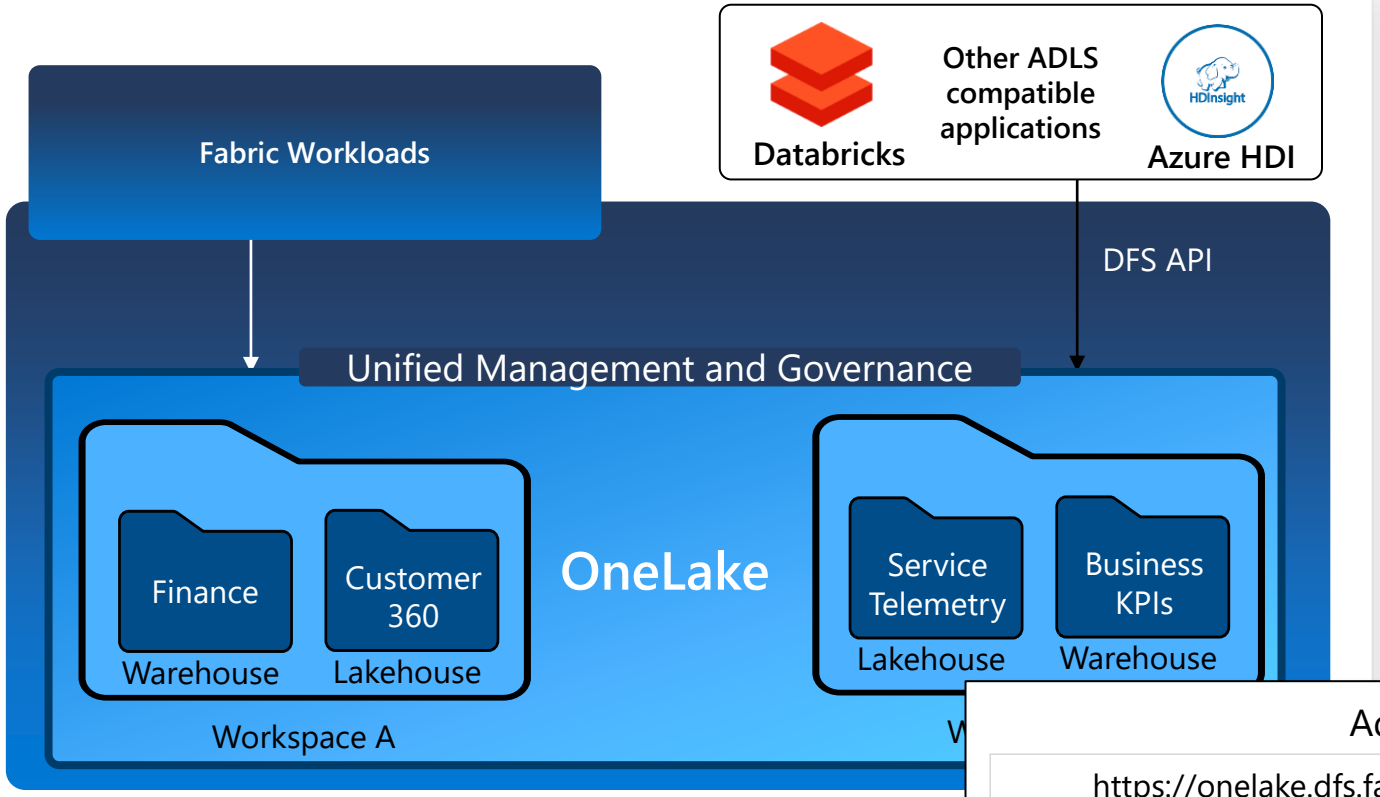
Real time analytics



[Real-Time Analytics Tutorial- Introduction - Microsoft Fabric]

Open access to data in OneLake

No lock-in with industry standard APIs and open file formats



By supporting the ADLS gen2 DFS APIs and SDKs, OneLake is compatible with existing ADLS applications.

Tenants will appear as one big storage account with different workspaces appearing as different containers with data organized into folders.

Underlying physical storage is virtualized away. OneLake ensures proper scale and performance.

Addressing Fabric workspaces and items as ADLS:

https://onelake.dfs.fabric.microsoft.com/{workspace-Name}/{itemName.itemType}/

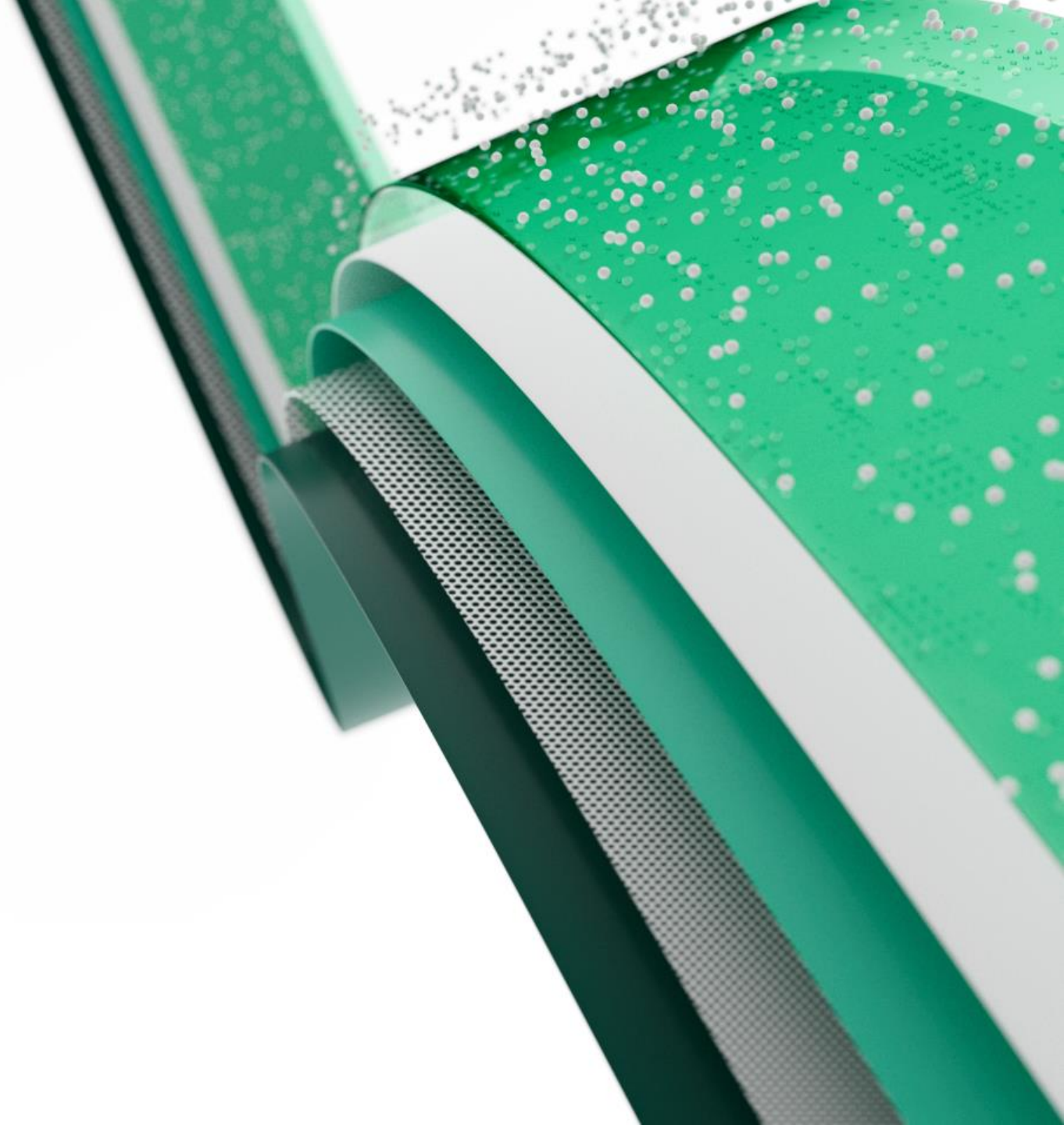
Ex. abfss://myWorkspace@onelake.dfs.fabric.microsoft.com/
abfss://WorkspaceName@onelake.dfs.fabric.microsoft.com/LakehouseName.lakehouse/Tables

[Integrate OneLake with Azure Databricks - Microsoft Fabric]

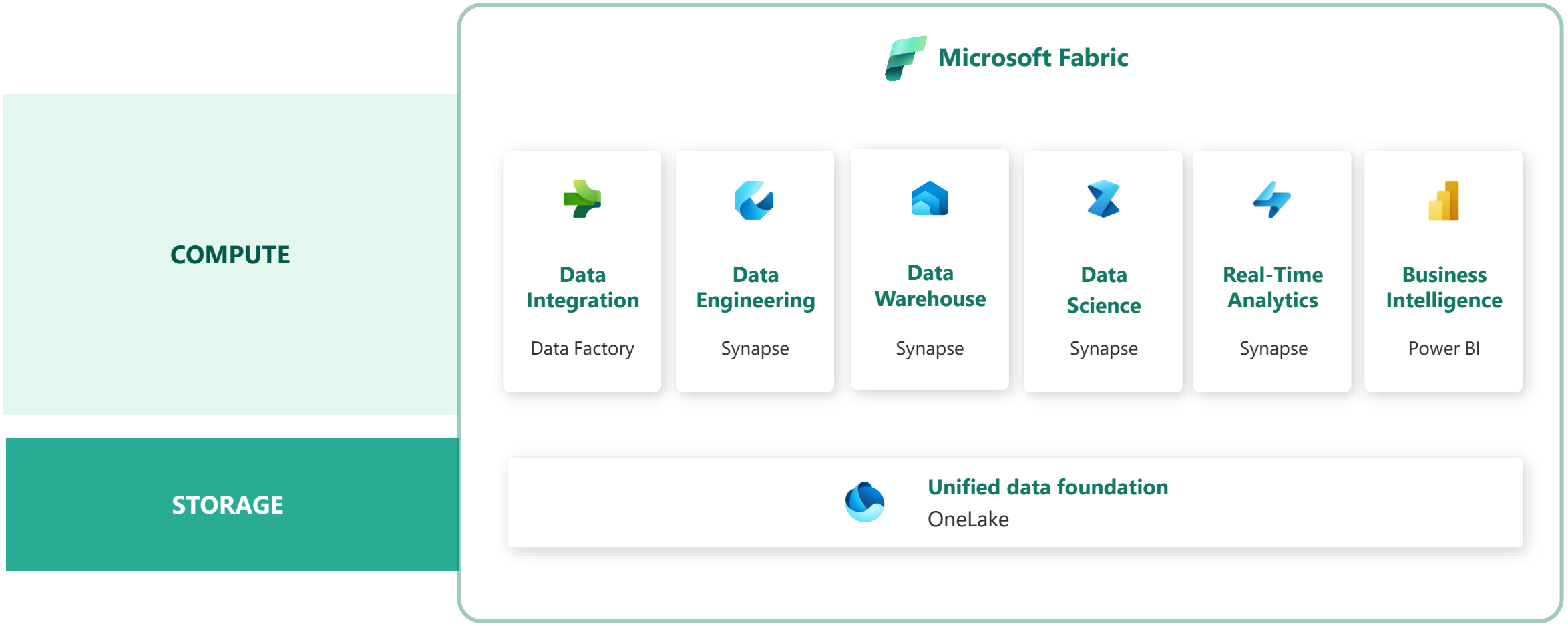
Microsoft

Microsoft Fabric

Business model overview



Microsoft Fabric business model overview



Microsoft Fabric simplicity

Microsoft Fabric is a unified product for all your data and analytics workloads. Rather than provisioning and managing separate compute for each workload, with Microsoft Fabric, your bill is determined by two variables: the amount of compute you provision and the amount of storage you use.

COMPUTE

A shared pool of capacity that powers all capabilities in Microsoft Fabric, from data modeling and data warehousing to business intelligence.

Pay-as-you-go

STORAGE

A single place to store all data.
Pay-as-you-go ($ per GB / month).

It’s time to see Microsoft Fabric in reality