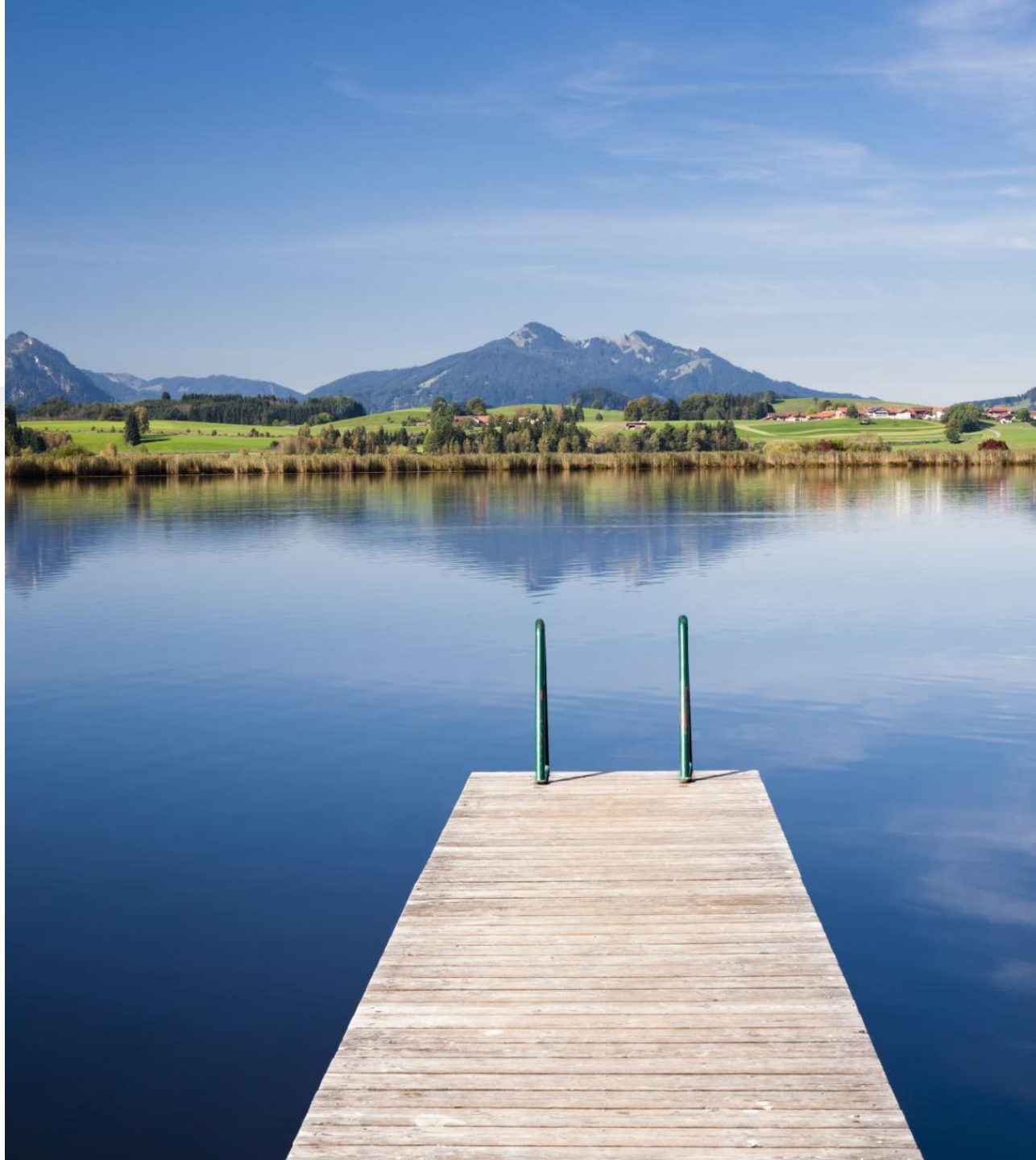
Data Lake, Delta Lake, Lakehouse and OneLake

- Jean Hayes
- Vijey Palaniappan
- Neeraj Jhaveri

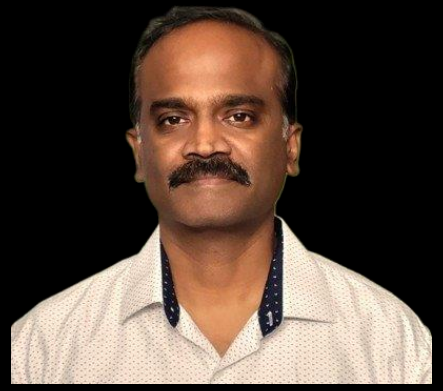


Jean Hayes

Jean Hayes, Senior Fast Track for Engineer, Data & AI Team

- Consulting in Microsoft BI/Analytics since 2007
- Passionate about helping customers build Data/AI workloads in the cloud
- Driven by learning new technologies and developing creative solutions
- Love my dog, family, friends, skiing, hiking, Orange Theory, Saint Thomas USVI

Vijey Palaniappan



Vijey Palaniappan is a Product Manager with a strong background in Computer Engineering and an MBA. With over 20 years of expertise in building data platforms, he is deeply passionate about Big Data and Data Warehousing.

Vijey is adept at defining product strategies, collaborating with stakeholders, and driving innovation through market research and competitive analysis. Vijey has a proven track record of guiding large cross-functional teams in the design, development, and successful launch of complex data platforms. His expertise enables organizations to make data-driven decisions, driving business growth and efficiency.

Neeraj Jhaveri

@jhaveri_neeraj

Neeraj Jhaveri is a Senior Engineer in Azure CXP with expertise in providing Data, Analytics, and AI solutions. He is one of the founding hosts of *Tales from the* Field on YouTube.

Neeraj as an MS in Computer Science from NYIT, and about 20 years of experience in the IT industry as an Architect/Team Leader. His background is in data warehousing, business intelligence and other analytics areas. As an IT leader he has managed and delivered Big Data and Business Intelligence Solutions while promoting an analytics driven culture.

Azure FastTrack CXP

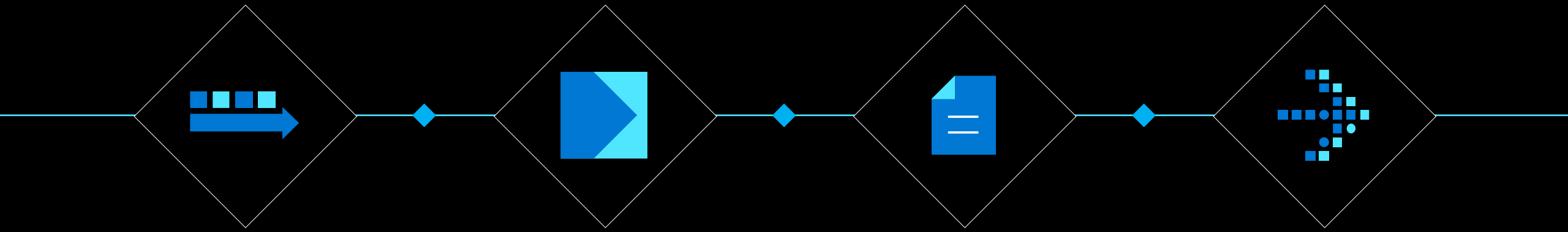
We support the following solutions:

- Datacenter migration
- Windows Server on Azure
- Linux on Azure
- SAP on Azure
- Business continuity and disaster recovery
- High-performance computing
- Internet of Things (IoT)
- Cloud-native apps
- DevOps
- App modernization
- Data modernization to Azure
- Security
- Globally distributed data
- Windows Virtual Desktop
- ASM Migration
- Azure Management and Governance
- Artificial Intelligence and Machine Learning
- Analytics

aka.ms/FTA

https://www.youtube.com/@Tales-from-the-Field

Agenda



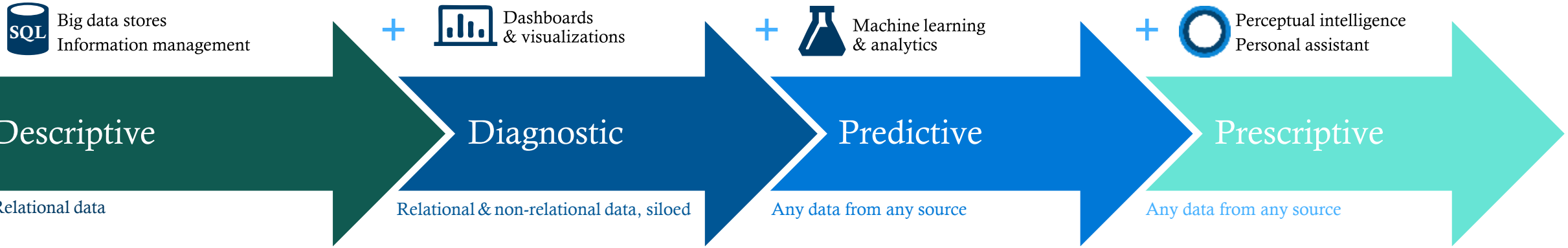
Data Lake

Delta Lake

Lakehouse

One Lake

ANALYTICS CAPABILITIES GROWTH PATH



Know your business better through manual reports built using structured department data.

Make smarter decisions using self-service tools to separately display relational and non-relational data sets available through a virtualized network.

Plan for the future by analyzing and modeling diverse data types integrated in real time through a hybrid environment.

Trigger automated rules to respond to analysis of data sets integrated in real time through a hybrid environment.

ACTIVITIES

- **Manual reports:** Request IT to pull data and build reports
- **Historical analysis:** Review business periodically and track against goals
- **Departmental data:** Combine data manually for business-level insights

- **Self-service & mobile BI:** Access anywhere, on any device
- **Visualizations and dashboards:** Answer questions by visualizing data
- **External data:** Capture big data for future use

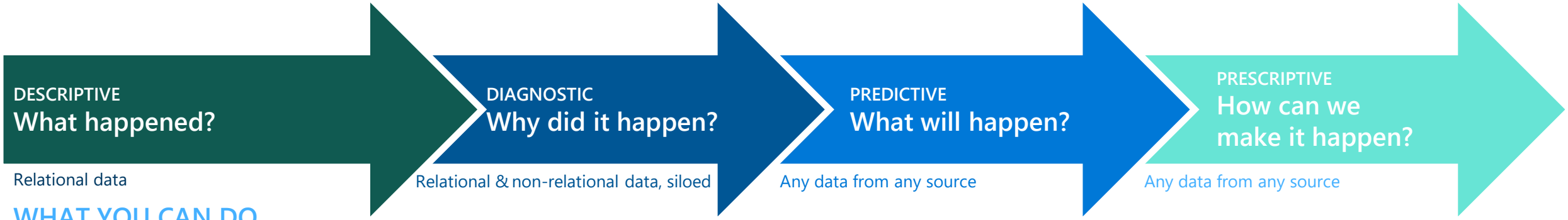
- **Real-time insights:** Track business changes up to the minute
- **Integrated data sources:** Combine and analyze all kinds of data
- **Predictive analytics:** Use complex modeling and analysis for predictions

- **Immediate innovation:** Make changes on the fly using real-time data
- **Asset feedback:** Monitor IoT data to identify new business models
- **Automated actions:** Trigger automatic actions to changing trends

TIME-TO-DATA:	WEEKS	OVERNIGHT	MINUTES	REAL-TIME

Bring the stages to life using sample scenarios

RETAIL



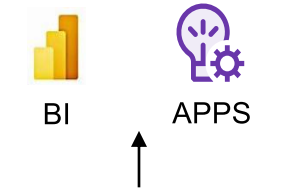
Relational data	Relational & non-relational data, siloed	Any data from any source	Any data from any source

WHAT YOU CAN DO

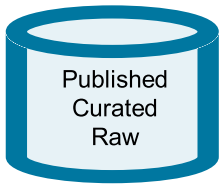
Calculate **store profitability** Track sales by **region or location**	**Visualize item-level profitability** by store	Calculate **customer profitability** Analyze likely impact of trends for **cross sell and upsell** opportunities	Use 360-degree customer view to boost **customer lifetime value** with **cross sell upsell** Automate **personalized offers** for deserted carts and make **product recommendations** for active carts	**Customer Lifetime Value & Cross sell, Upsell**

TIME-TO-DATA:	WEEKS	OVERNIGHT	MINUTES	REAL-TIME

Late 1980s – Mid 2000s



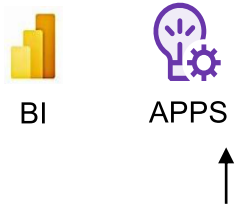
DATA WAREHOUSE



BATCH
ELT/ETL

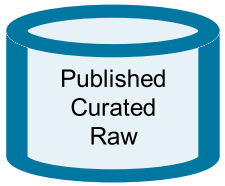
Internal / External
& 3rd Party Data

Mid 2000s – 2020



APIs ACCESS

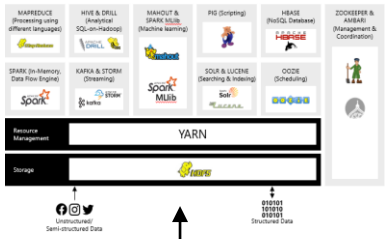
DATA WAREHOUSE



BATCH
ELT/ETL

Internal / External
& 3rd Party Data

DATA LAKE (Apache Hadoop,
Azure Storage/ADLS)



STREAMING

2021 - Present

**LAKEHOUSE – Powered by
Delta Lake, Iceberg, and Hudi**

BATCH ELT/ETL

Internal / External
& 3rd Party Data

STREAMING

DATA LAKE – WHAT?

- A storage repository that holds a **large amount of data** in its **native, raw** format.
- Data lake stores are **optimized for scaling** to terabytes and petabytes of data.
- The data typically comes from multiple **heterogeneous sources**, and may be structured, semi-structured, or unstructured.
- This approach differs from a traditional data warehouse, which transforms and processes the data at the time of ingestion.
- is usually a **single store of data** including raw copies of source system data, sensor data, social data etc.
- And **transformed data** used for tasks such as reporting, visualization, advanced analytics and machine learning.
- Can include **structured** data from relational databases (rows and columns), **semi-structured** data (CSV, logs, XML, JSON), **unstructured** data (emails, documents, PDFs) and **binary** data (images, audio, video).
- Poorly-managed data lakes have been facetiously called **data swamps**.

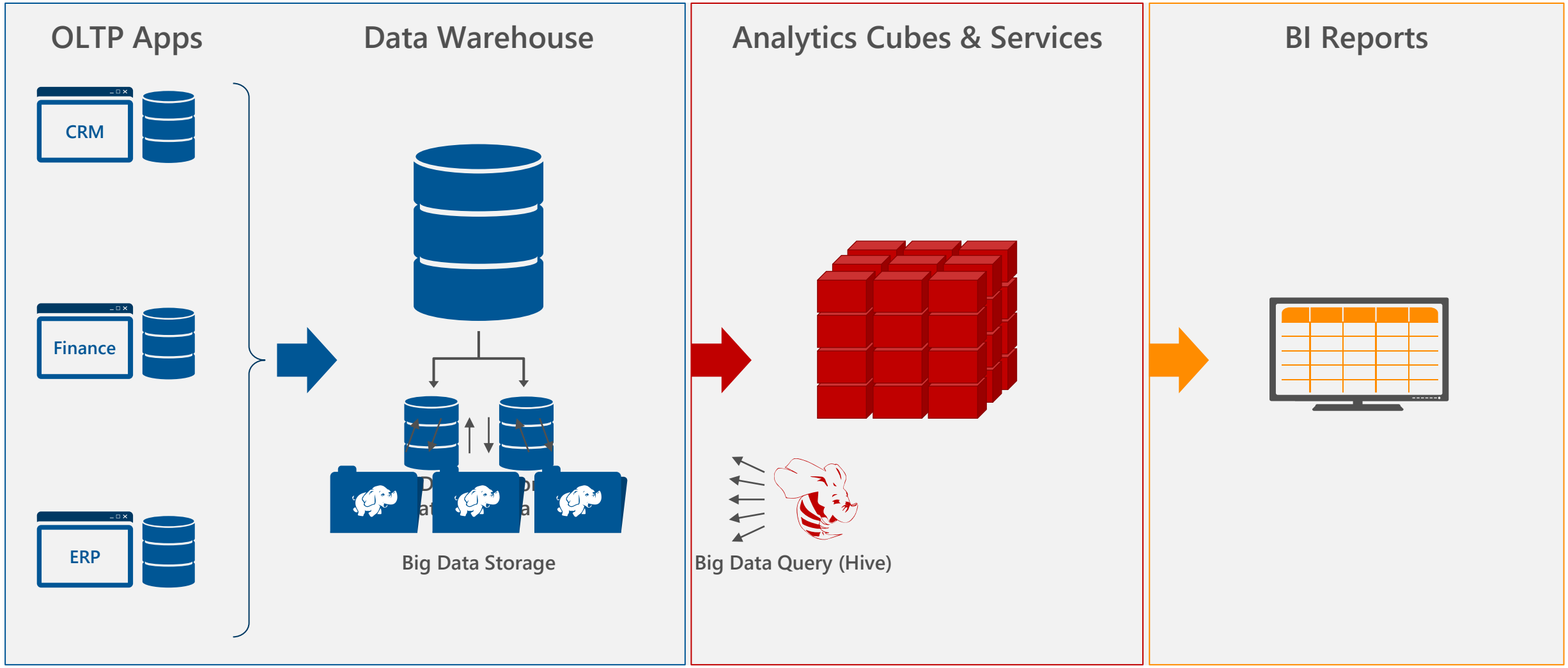
Data lakes - Azure Architecture Center | Microsoft Learn Data lake - Wikipedia

DATA LAKE – WHY?

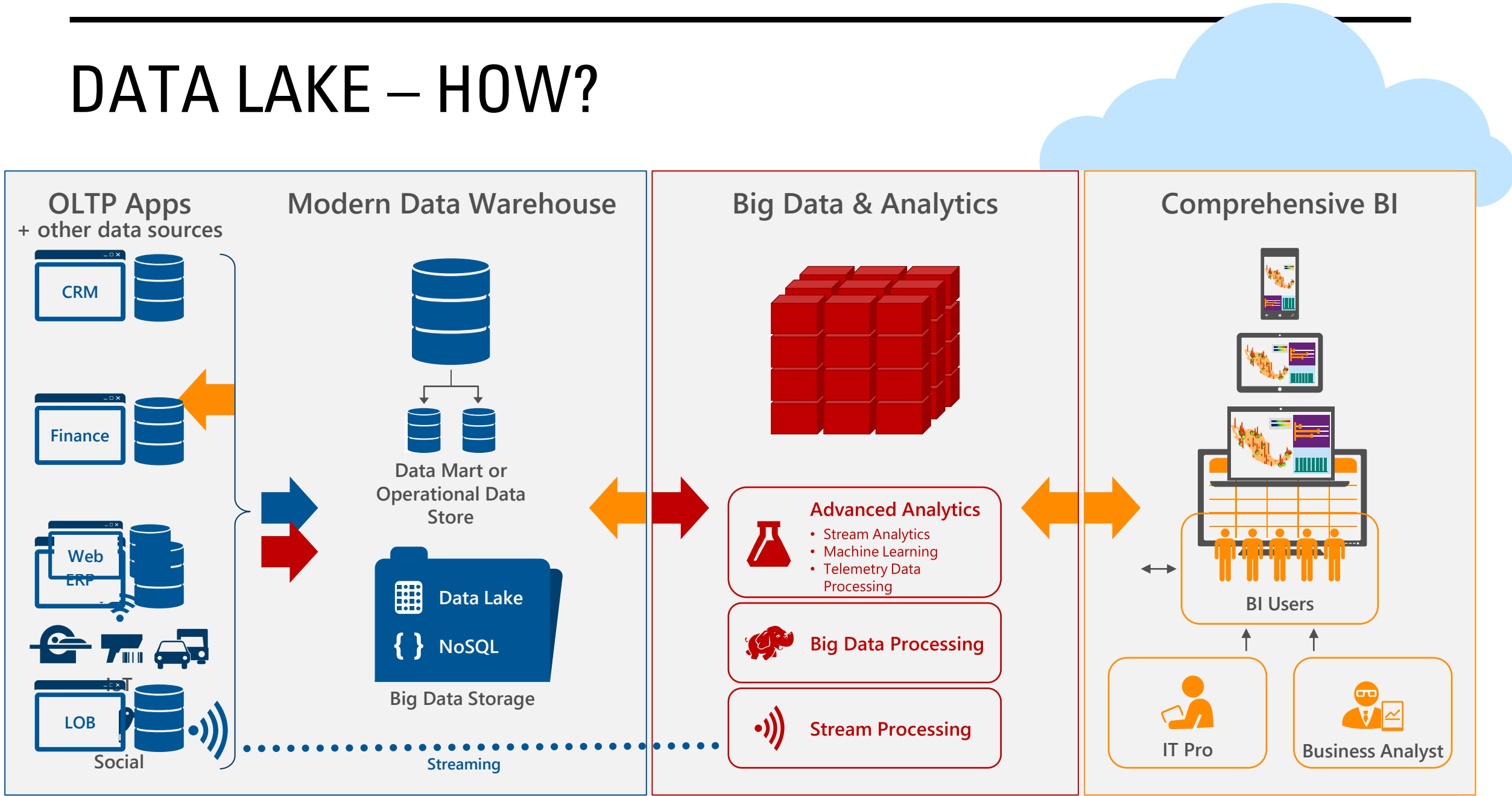
- Are also **highly durable** and **low cost**, because of their ability to scale and leverage object storage.
- Additionally**, advanced analytics** and **machine learning** on unstructured data are some of the most strategic priorities for enterprises today.
- The unique ability to ingest raw data in a variety of formats (structured, unstructured, semi-structured), along with the other benefits mentioned, makes a data lake the clear choice for data storage.
- When properly architected, data lakes enable the ability to:

Power data science and machine learning
Centralize, consolidate and catalogue your data
Quickly and seamlessly integrate diverse data sources and formats
Democratize your data by offering users self-service tools

DATA LAKE – HOW?



DATA LAKE – HOW?



DATA LAKE ORGANISATION

Raw Data

Serves as landing area

In streaming patterns data is aggregated

Access is highly controlled

Is often archived after processing

Folders mirror ingestion patterns
Lock for Data Engineering Only

Enriched Data

Formats are harmonised

Common data model applied

Defined schema – data types are validated

Data is cleansed

Incremental change based

Processing errors captured

Folder Structure Mirrors Data Domain
Full control to data engineering with read access to BI Analysts/Data Scientists

Curated Data

Served to consumers as is

Data assets are governed and documented

Data has known structure

Folder Structure Mirrors Data Product
Full control to data engineering with read write access to BI Analysts/Data Scientists

Workspace Data

Consumers may ingest additional support sources

Space for groups to build and store data

Folder Structure Mirrors Team
Full control to data engineering
BI Analysts/Data Scientists

Ref: The Hitchhiker's Guide to the Data Lake

DATA LAKE STORAGE BEST PRACTICES

- Centralized (Single account) or federated data lake (Multiple accounts) implementation
- Plan zones(Raw, Enriched, Curated), folder structure and security groups upfront
- Raw data
  - stored in **original** format
  - Immutable & highly secure
  - life-cycle policies to reduce cost
  - In non-raw zone use **Parquet/Delta** format – performance & compression
- Each folder should contain the same file format
- Use consistent naming conventions
- Optimal file size
  - Large files are better than smaller files
    - more cost-effective & yields better analytic performance
    - Azure Data Lake Storage Gen2 is highly optimised to perform faster on larger files. This means that your analytics jobs will run faster, when operating on larger files, thus further reducing your TCO for running analytics jobs. files > 4 MB in size incurs a lower price for every 4 MB block of data read beyond the first 4 MB. eg to read a single file that is 16 MB is cheaper than reading 4 files that are 4 MB each.
    - Databricks / Spark recommendations – 64MB – 1GB

DATA LAKE SECURITY BEST PRACTICES

- Permissions:
  - assign relevant permissions as early as possible in your design/development/operationalization phase
  - Use RBAC to manage the resource – account admins
  - Use ACLS to view/edit/manage the data (POSIX style) – users & jobs
  - Restrict write permissions except for
    - Users' sandbox/work/private zone
    - Automated jobs using service principal
  - Use security groups, resist assigning ACLs to individuals
    - Changing permissions at the group level will not involve recursive operations (slow)
- Access to raw data should be restricted

CAUTIONARIES FOR ADLS

- Data Lake Storage limits
  - Max accounts per sub = 250 (Request for increase)
  - 5 PB default (Depending on region and workload up to 100+ PB on request)
  - Max request rate 20,000 per second per storage account (Upto 100K upon request)
  - See other limits here
- ADLS doesn't have a true inheritance model, use default ACLs
  - Changing the default ACL on a parent does not affect the access ACL or default ACL of child items that already exist.
- Scripting permission changes on existing folder requires recursion
- RBAC (filesystem level) evaluated at higher priority over ACLs
- Storage Blob Data Owner built-in role and SAS auth gain super-user access

DATA LAKE – PROS

- A single data platform for **real-time** and **batch analytics**
- **Cost** Effectiveness
- **Convenience**
- **Future** proofing
- Building a **staging area** for your data warehouse
- Audit log of all data ever ingested into your data ecosystem thanks to the **immutable staging** area
- Increase the **time-to-value** and **time-to-insights**

7 Reasons Why You Should Consider a Data Lake - Dashbird

DATA LAKE – CHALLENGES

- Lack of a **scheme** or descriptive metadata
- Lack of **semantic consistency** across the data
- It can be hard to guarantee the quality of the data
- **Governance**, **access controls** and **privacy** issues can be problems
- Integration of **relational** data
- Integrated or **holistic views** across the organisation
- **Dumping ground** for data that is never actually analysed or mined for insights

→ As a result, most of the data lakes in the enterprise have become **data swamps**

Introduction to Data Lakes - Databricks

DELTA LAKE

Delta Lake is an open-source storage format that brings ACID transactions to Apache Spark™ and big data workloads

- Atomicity
- Consistency
- Isolation
- Durability

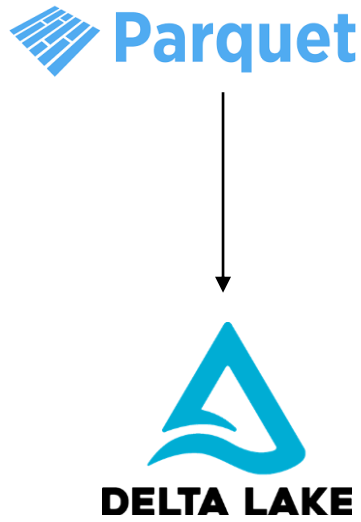
Works with your existing cloud data lake store

Data format is based upon Parquet

What are ACID Transactions? (databricks.com)

DELTA LAKE FILE FORMAT BASED UPON PARQUET

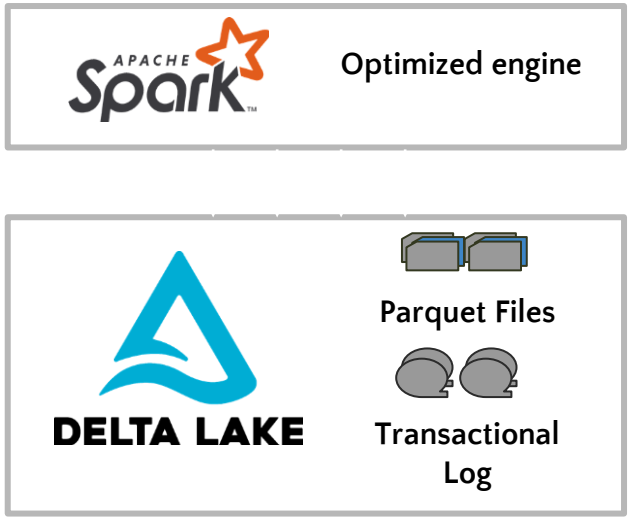
- Based on the Apache Parquet file format originally created for the Hadoop File System (HDFS)
- Apache Parquet is a **columnar data store** and takes advantage of **efficient compression** and **encoding** algorithms
- Apache Parquet is best suited for read operations
- Works best for batch loads vs singleton writes



DELTA LAKE

Delta Lake extends Parquet:

- FULL ACID TRANSACTIONS
- UNIFIED STREAMING AND BATCH
- SCHEMA ENFORCEMENT
- TIME TRAVEL/DATA SNAPSHOTS
- NATIVE SUPPORT FOR UPDATE/DELETE/MERGE



- Z-ORDER INDEXING AND STATS
- COMPACTION TO OPTIMIZE FILE SIZES
- DATA SKIPPING READS ONLY THE RELEVANT DATA
- CACHING


DATA LAKE…
DELTA LAKE …
WHERE DOES
LAKEHOUSE
FIT IN?

LAKEHOUSE – WHAT?

SOLUTION TO ADDRESS DATA LAKE SHORTCOMINGS

- **Data management architecture** that aims to simplify enterprise data infrastructure and accelerate innovation in an age when AI/ML is poised to disrupt every industry.
- Adds a transactional storage layer.
- Uses similar data structures and data management features as those in a data warehouse but instead runs them directly on cloud data lakes.
- Allows traditional analytics, data science and machine learning to coexist in the same system, all in an open format.

What Is a Lakehouse? - The Databricks Blog

LAKEHOUSE – WHY?

- Enables a wide range of new use cases for cross-functional enterprise-scale analytics, BI and machine learning projects that can unlock massive business value.
- Data analysts can harvest rich insights by querying the data lake using SQL
- Data Scientists can join and enrich data sets to generate ML models with ever greater accuracy
- Data engineers can build automated ETL pipelines
- And business intelligence analysts can create visual dashboards and reporting tools faster and easier than before.
- These use cases can all be performed on the data lake simultaneously, without lifting and shifting the data, even while new data is streaming in.

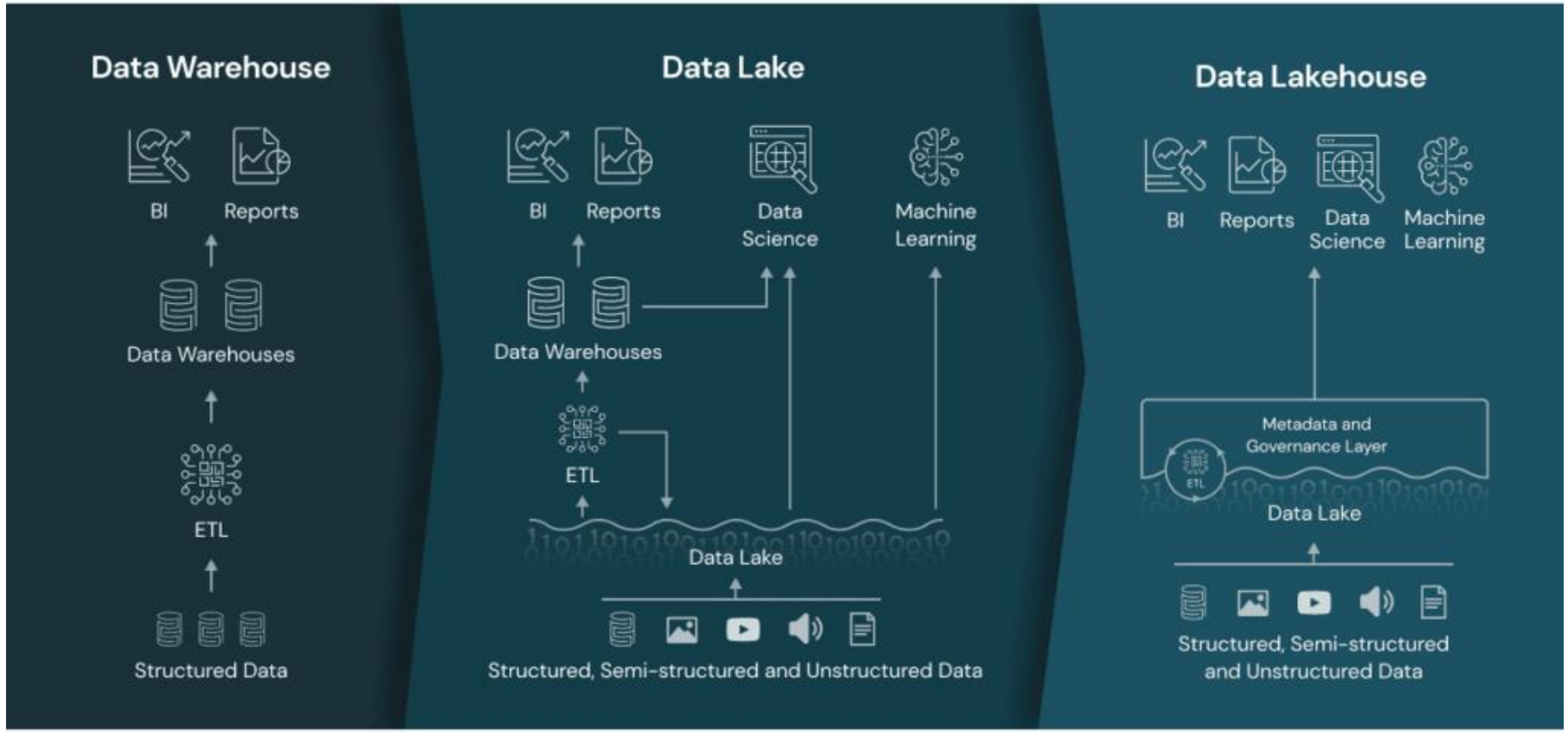
What Is a Lakehouse? - The Databricks Blog

LAKEHOUSE – WHY?

- Transaction support
- Schema enforcement and governance
- BI support
- Storage is decoupled from compute
- Openness
- Support for diverse data types ranging from unstructured to structured data
- Support for diverse workloads
- End-to-end streaming

What Is a Lakehouse? - The Databricks Blog

LAKEHOUSE – HOW?

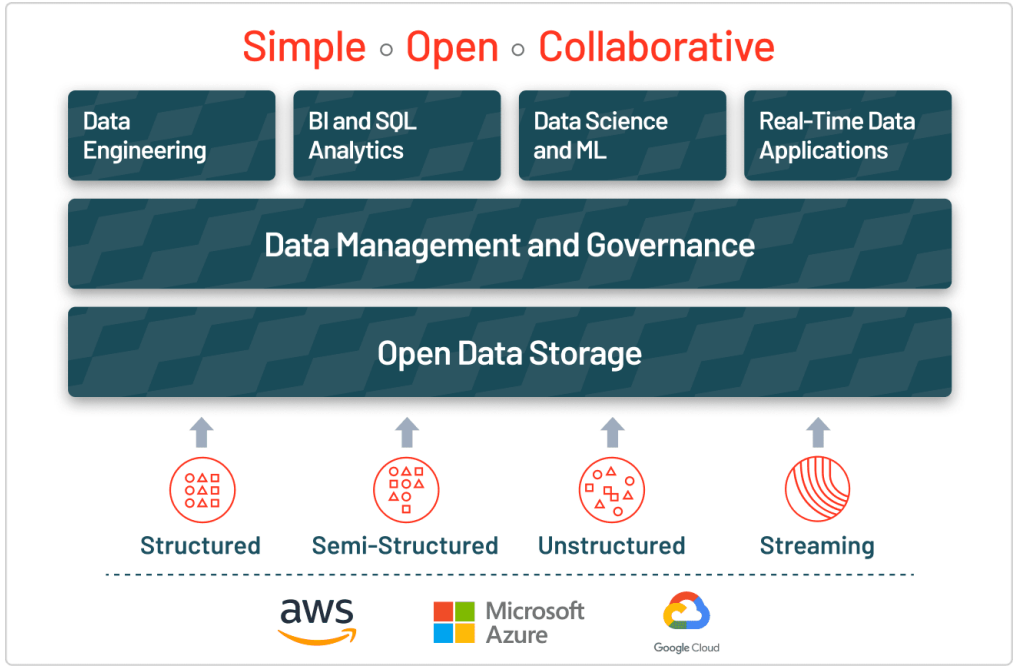


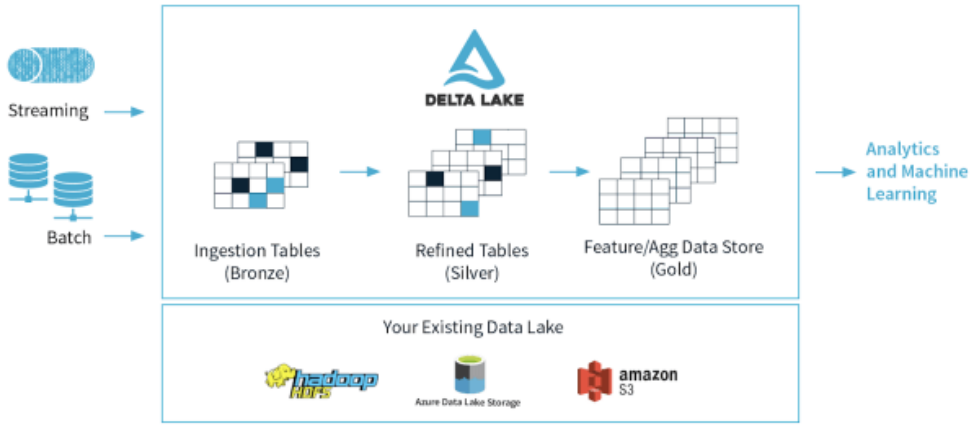
DATA LAKE VS. LAKEHOUSE VS. DATA WAREHOUSE

	Data lake	Data lakehouse	Data warehouse
Types of data	All types: Structured data, semi-structured data, unstructured (raw) data	All types: Structured data, semi-structured data, unstructured (raw) data	Structured data only
Cost	$	$	$$$
Format	Open format	Open format	Closed, proprietary format
Scalability	Scales to hold any amount of data at low cost, regardless of type	Scales to hold any amount of data at low cost, regardless of type	Scaling up becomes exponentially more expensive due to vendor costs
Intended users	Limited: Data scientists	Unified: Data analysts, data scientists, machine learning engineers	Limited: Data analysts
Reliability	Low quality, data swamp	High quality, reliable data	High quality, reliable data
Ease of use	Difficult: Exploring large amounts of raw data can be difficult without tools to organize and catalog the data	Simple: Provides simplicity and structure of a data warehouse with the broader use cases of a data lake	Simple: Structure of a data warehouse enables users to quickly and easily access data for reporting and analytics
Performance	Poor	High	High

BUILDING A LAKEHOUSE WITH DELTA LAKE

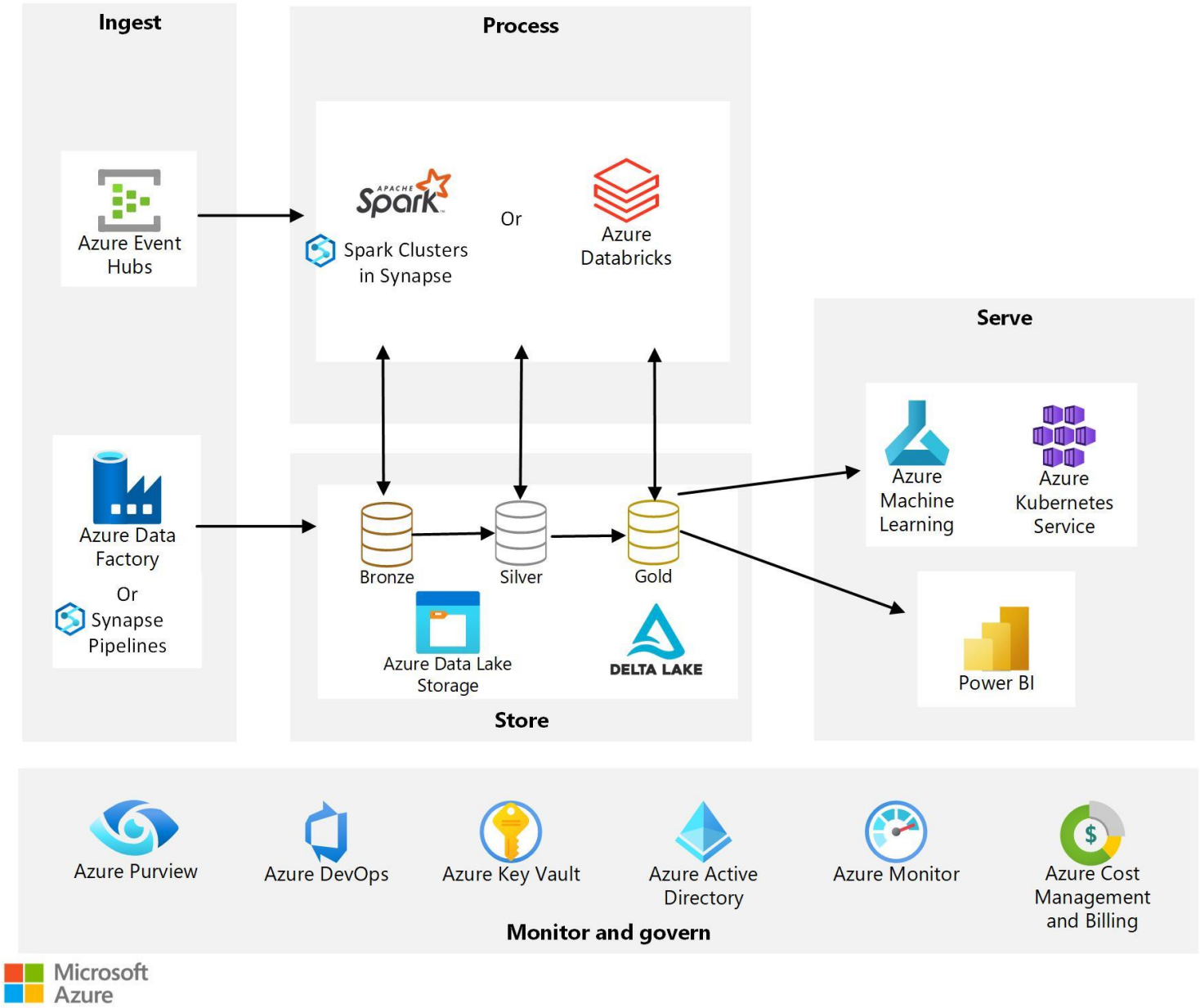
- Combines the best of both data lakes and data warehouses.
- Provides a reliable, single source of truth.
- Delivers quality, reliability, security and performance on your data lake for both streaming and batch operations
- Eliminates data silos and makes analytics accessible across the enterprise.
- Cost-efficient, highly scalable
- Gives self-serving analytics to end users.

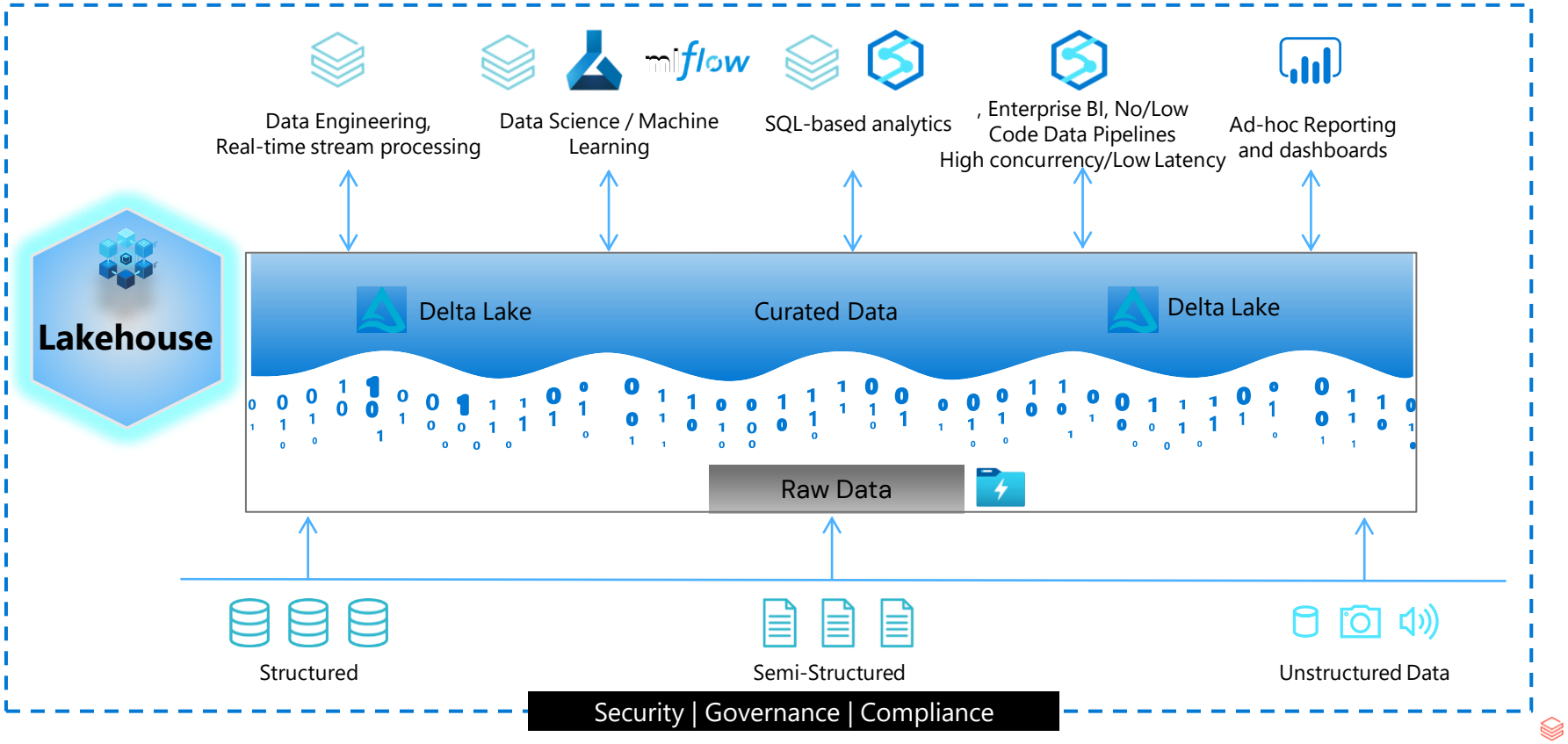




[Data Lakehouse defined | James Serra's Blog]

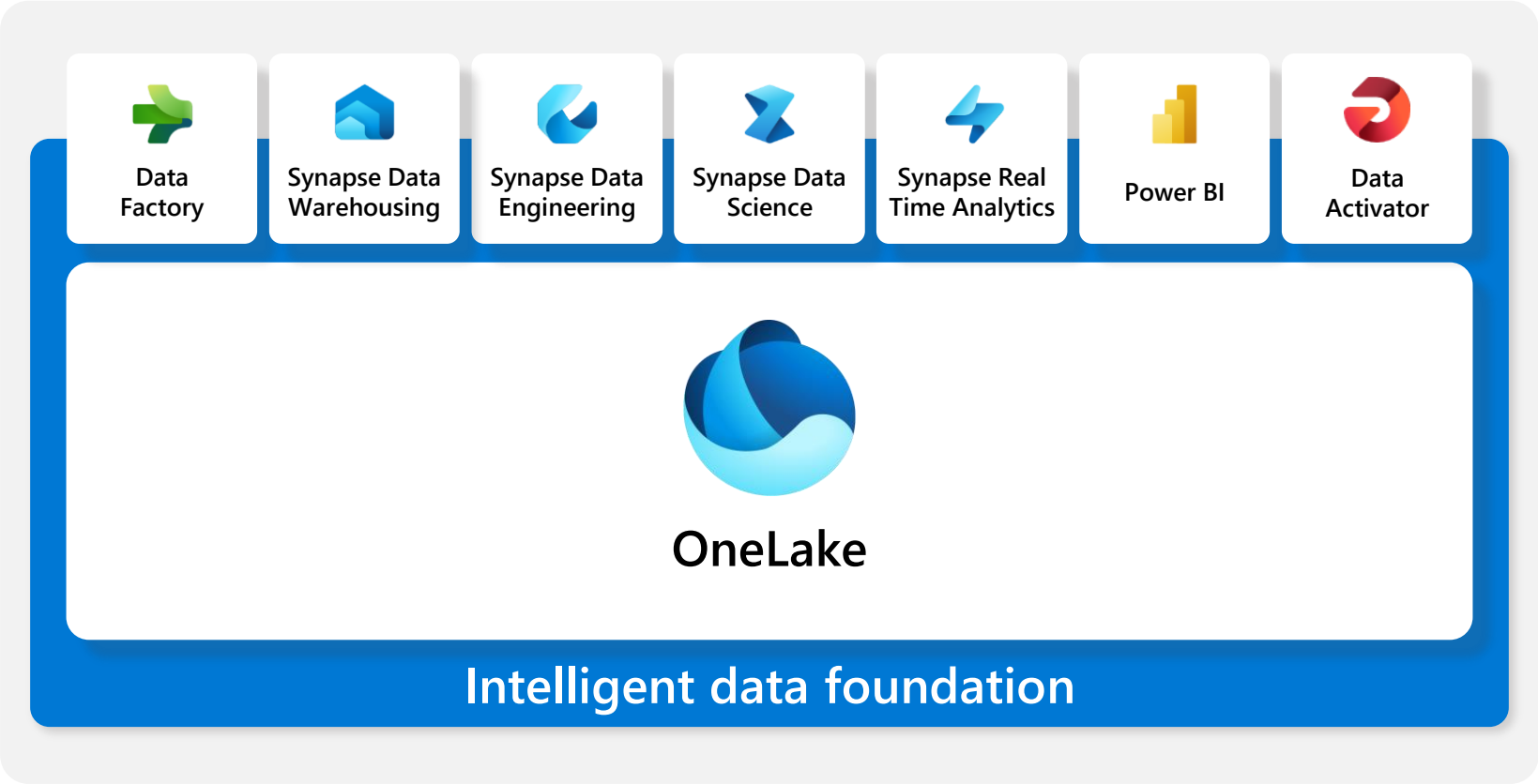
SAMPLE LAKEHOUSE ARCHITECTURE




Data Engineering,
Real-time stream processing
Data Science / Machine
Learning
SQL-based analytics
, Enterprise BI, No/Low
Code Data Pipelines
High concurrency/Low Latency
Ad-hoc Reporting
and dashboards
Lakehouse
Delta Lake
Curated Data
Delta Lake
Raw Data
Structured
Semi-Structured
Unstructured Data
Security | Governance | Compliance

Microsoft Fabric

The data platform for the era of AI



OneLake for all data

"The OneDrive for data"

OneLake provides a data lake as a service without you needing to build it

OneLake for all data

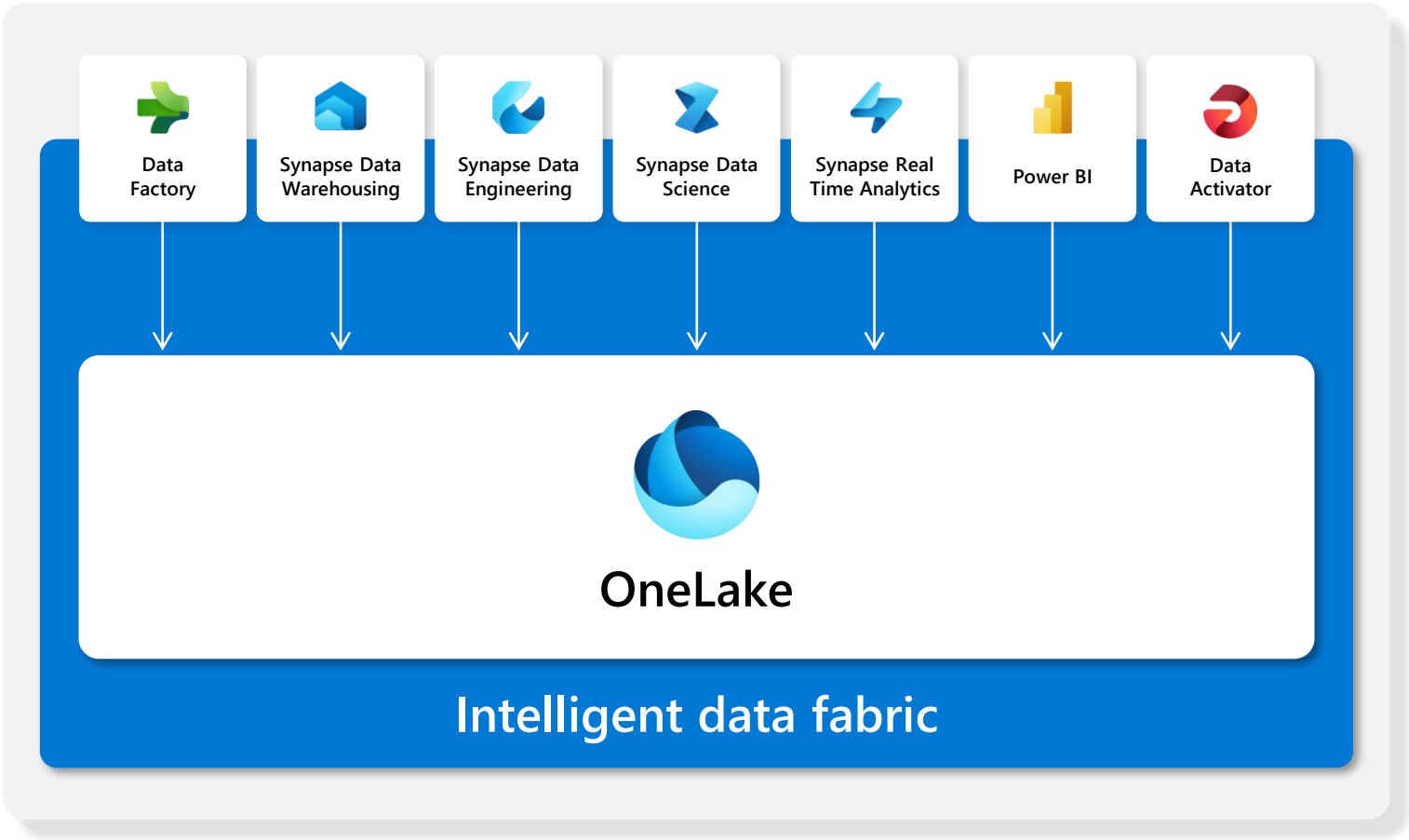
"The OneDrive for data"

OneLake

One Copy

One Security

OneLake Data Hub



OneLake for all data

"The OneDrive for data"

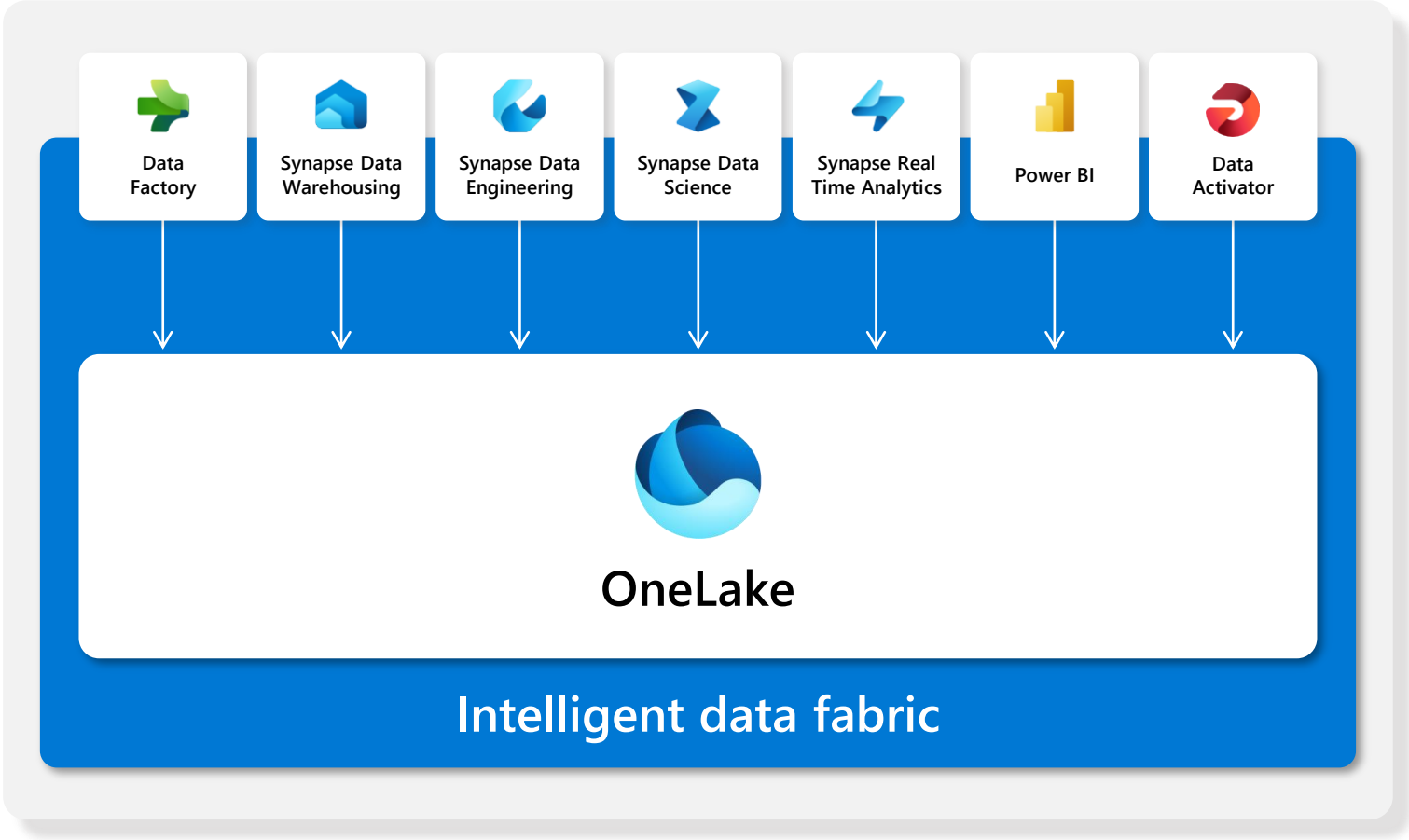
OneLake

- A single unified logical SaaS data lake for the whole organization (no silos)
- Organize data into domains
- Foundation for all Fabric data items
- Provides full and open access through industry standard APIs and formats to any application (no lock-in)

One Copy

One Security

OneLake Data Hub



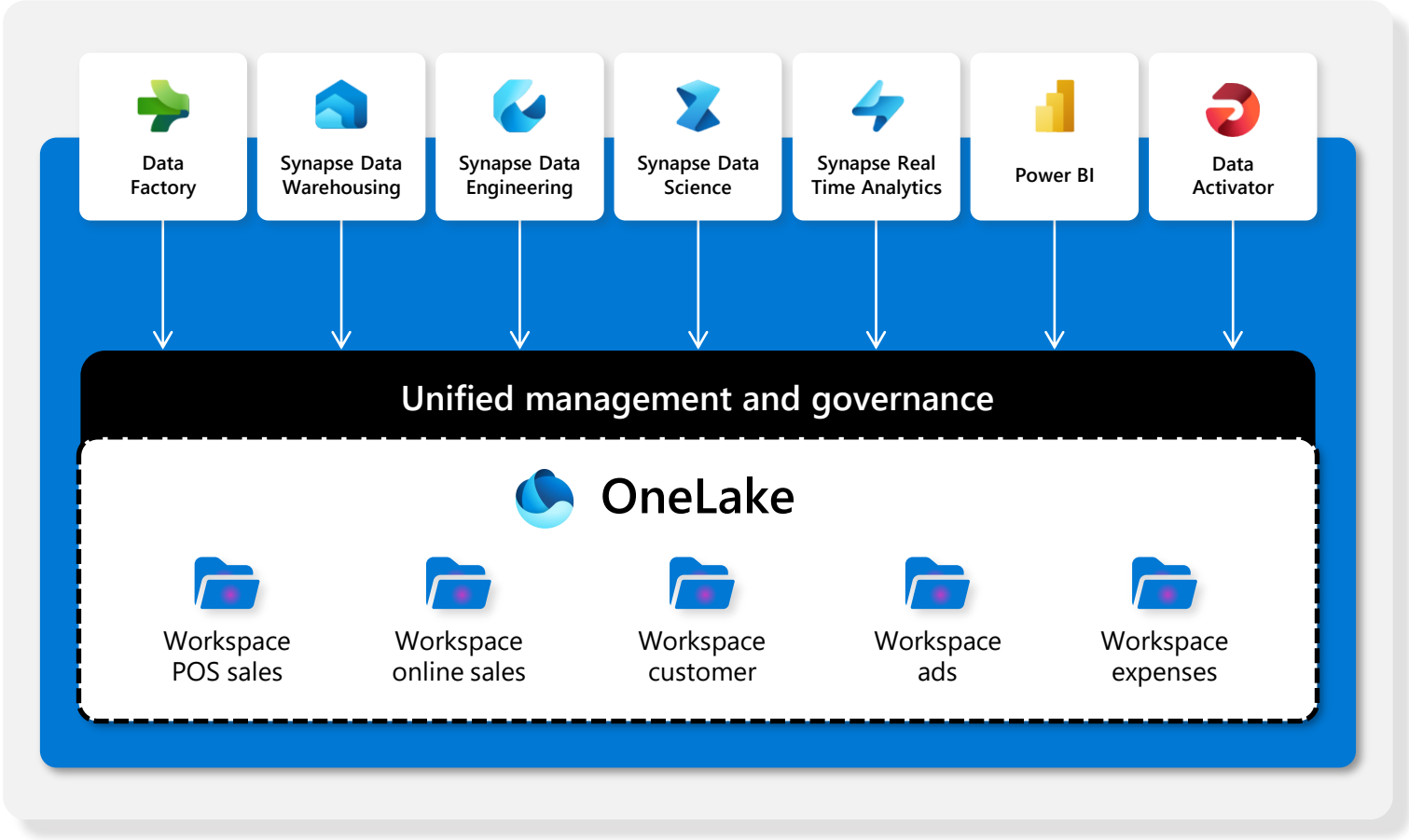
A single unified SaaS data lake

"No silos"

OneLake comes automatically provisioned with every Fabric tenant with no infrastructure to manage.

Any data in OneLake works with out-of-the-box governance such as data lineage, data protection, certification, catalog integration, etc. All data is ultimately under the control of a tenant admin.

OneLake enables distributed ownership. Different workspaces allow different parts of the organization to work independently while still contributing to the same data lake. Each workspace can have its own administrator, access control, region and capacity for billing.



OneLake for all data

"The OneDrive for data"

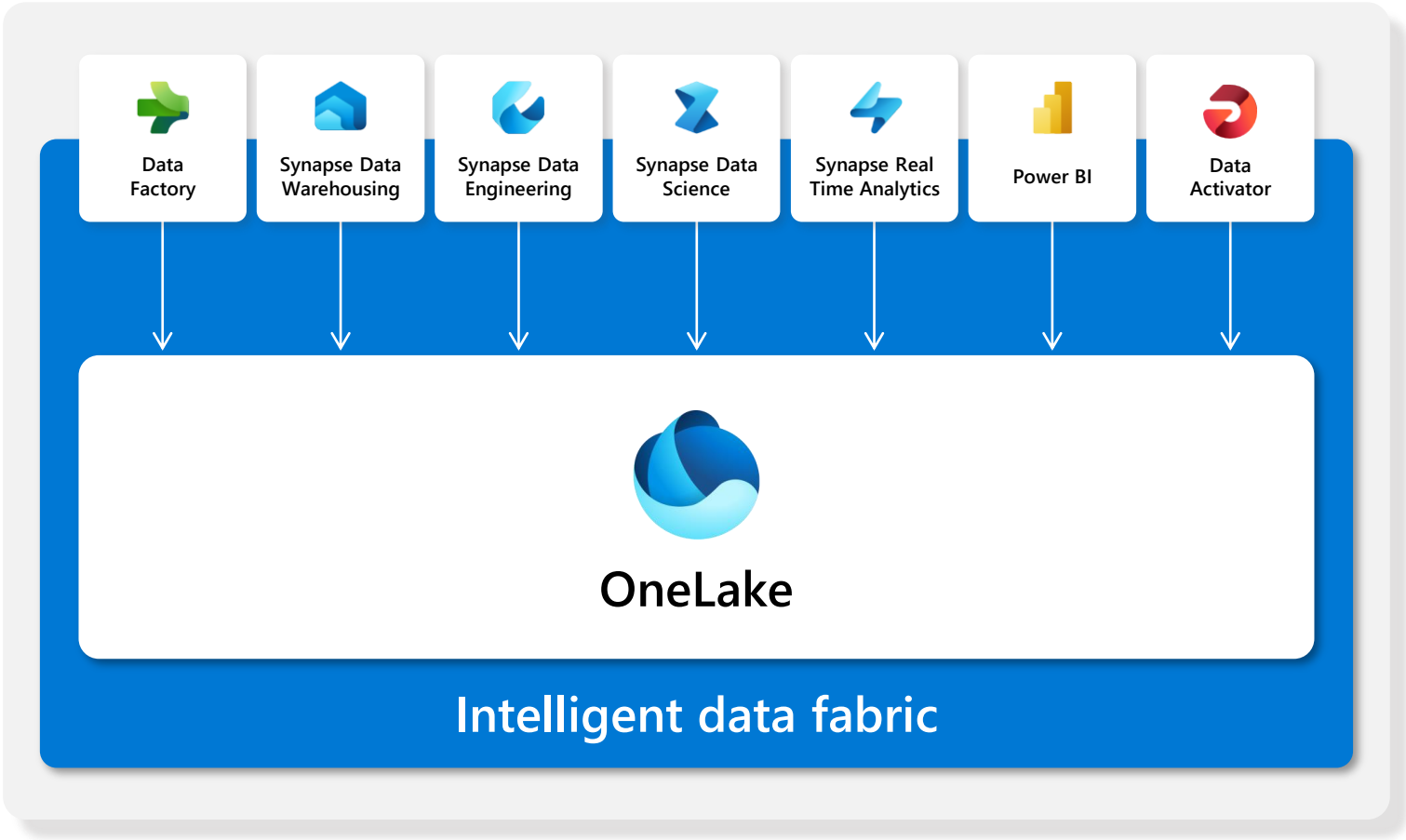
OneLake

One Copy

- Virtualize data across domains and clouds into a single logical lake with shortcuts
- The One Copy of data for all the analytical engines of Fabric without moving or duplicating data

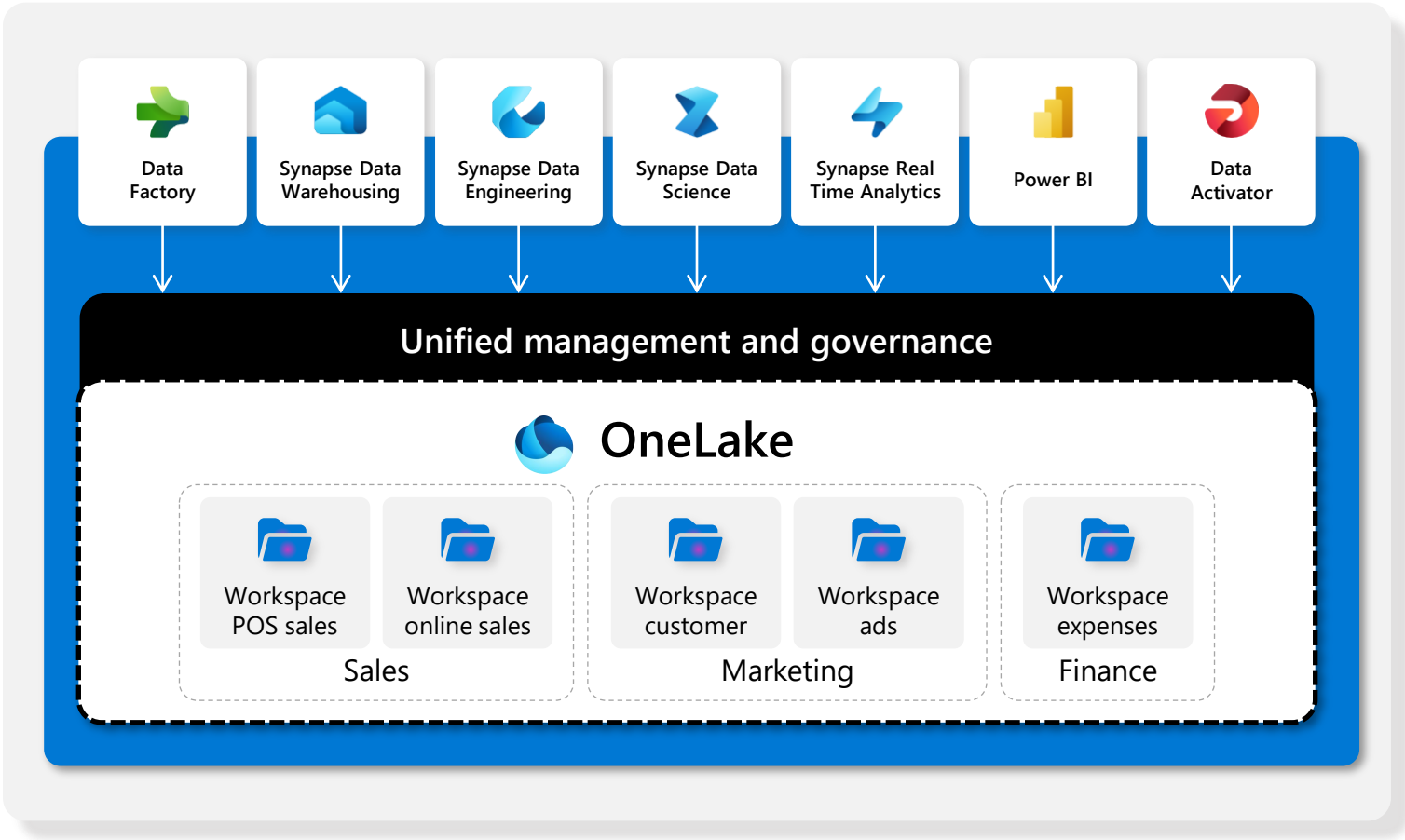
One Security

OneLake Data Hub



OneLake gives a true data mesh as a service

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

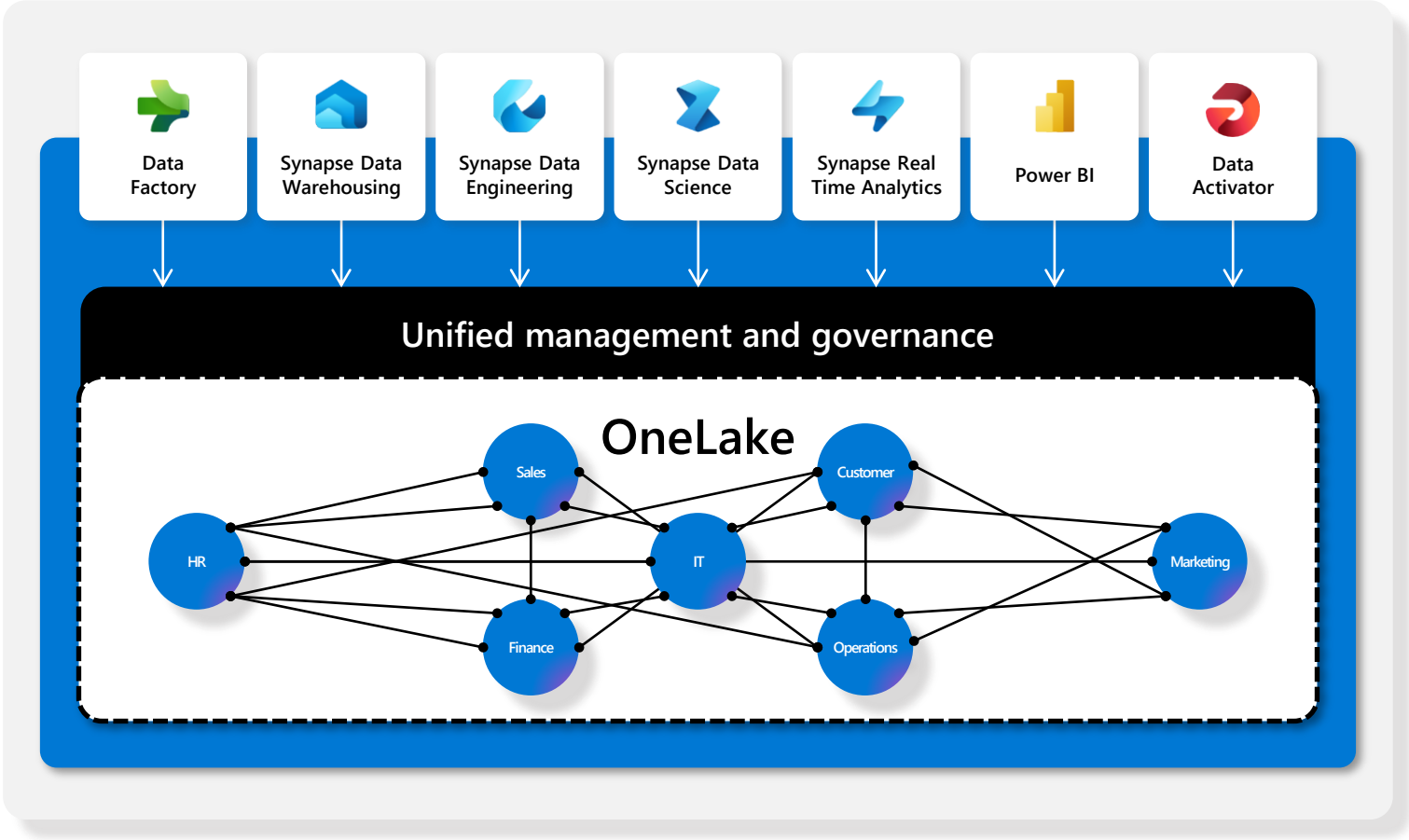


OneLake gives a true data mesh as a service

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.



Shortcuts virtualize data across domains and clouds

No data movements or duplication

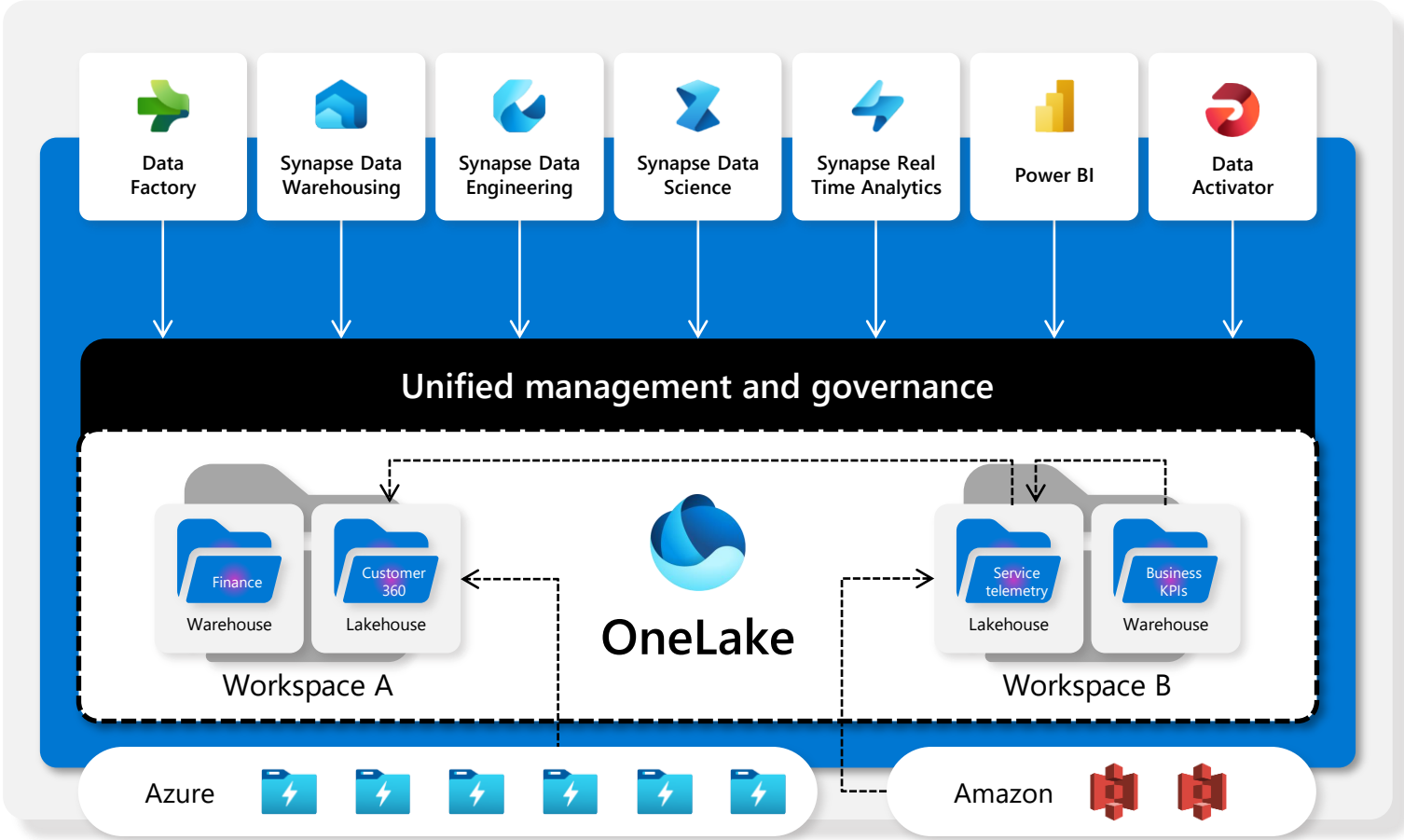
A shortcut is a symbolic link which points from one data location to another

Create a shortcut to make data from a warehouse part of your lakehouse

Create a shortcut within Fabric to consolidate data across items or workspaces without changing the ownership of the data. Data can be reused multiple times without data duplication.

Existing ADLS gen2 storage accounts and Amazon S3 buckets can be managed externally to Fabric and Microsoft while still being virtualized into OneLake with shortcuts

All data is mapped to a unified namespace and can be accessed using the same APIs including the ADLS Gen2 DFS APIs

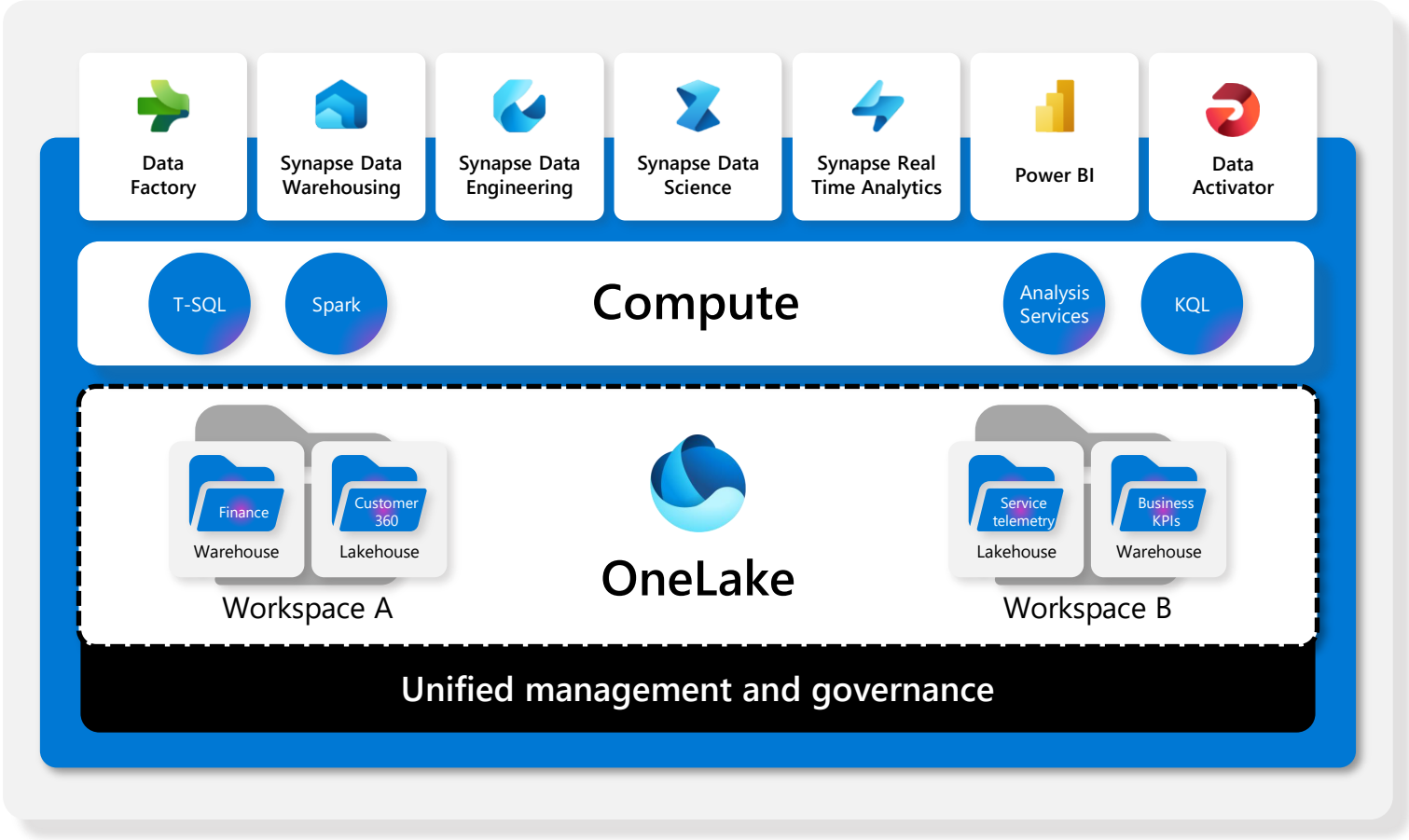


One Copy for all computes

Real separation of compute and storage

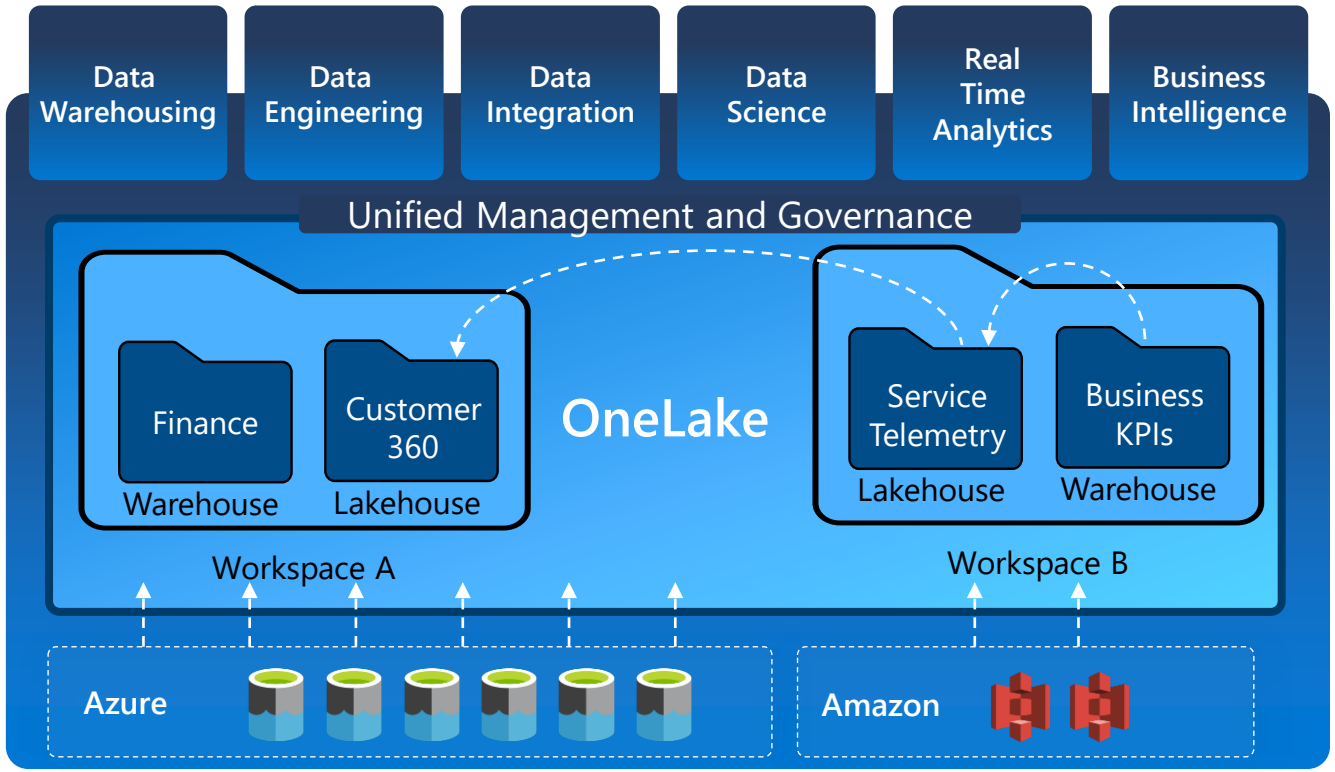
Compute powers the applications and experiences in Fabric. The compute is separate from the storage.

Multiple compute engines are available, and all engines can access the same data without needing to import or export it. You are able to choose the right engine for the right job.



Shortcuts virtualize data across domains and clouds

No data movements or duplication



A shortcut is a symbolic link which points from one data location to another.

Create a shortcut to make data from a warehouse part of your lakehouse.

Create a shortcut within Fabric to consolidate data across items or workspaces without changing the ownership of the data. Data can be reused multiple times without data duplication.

Existing storage accounts can be managed externally to Fabric and Microsoft while still being virtualized into OneLake with shortcuts.

All data is mapped to a unified namespace and can be accessed using the same APIs including the ADLS gen2 DFS APIs.

CONFUSION?

LAKEHOUSE

ONE LAKE

POWER BI

BIG DATA

PAAS

DATA WAREHOUSE

DELTA LAKE

STORAGE

ANALYTICS

ADLS GEN 2

COMPUTE

SAAS

DATA LAKE

Demo

Datalake, Deltalake, Lakehouse within Azure and Microsoft Fabric

Azure (PaaS)

- Azure Data Lake Storage Gen2
- Synapse for Data Orchestration
- Load and transform with tool of choice
  - Synapse Notebooks with Spark Clusters
  - Azure Databricks
  - Dataflow
- Data landed in ADLS Gen 2, Synapse Lakehouse, or Azure Databricks Live Delta

Microsoft Fabric (SaaS) - In Preview

- Shortcuts to ADLS Datalake
- Fabric Data Pipelines for Orchestration
- Same options for transforming data as Azure
- Data landed in OneLake

Resources

Azure Data Lake

- Create a storage account for Azure Data Lake Storage Gen2
- Use Azure Storage Explorer with Azure Data Lake Storage Gen2

Azure Delta Lake

- What is Delta Lake? - Azure Databricks

Azure Synapse for Orchestration, Spark Pools and Spark Notebook Development

- Orchestrating data movement and transformation in Azure Data Factory and Synapse Pipelines
- Apache Spark in Azure Synapse Analytics overview
- How to use Synapse notebooks

Microsoft Fabric

- What is a lakehouse?
- What is OneLake?
- How to use notebooks
- Access Fabric data locally with OneLake file explorer


Microsoft