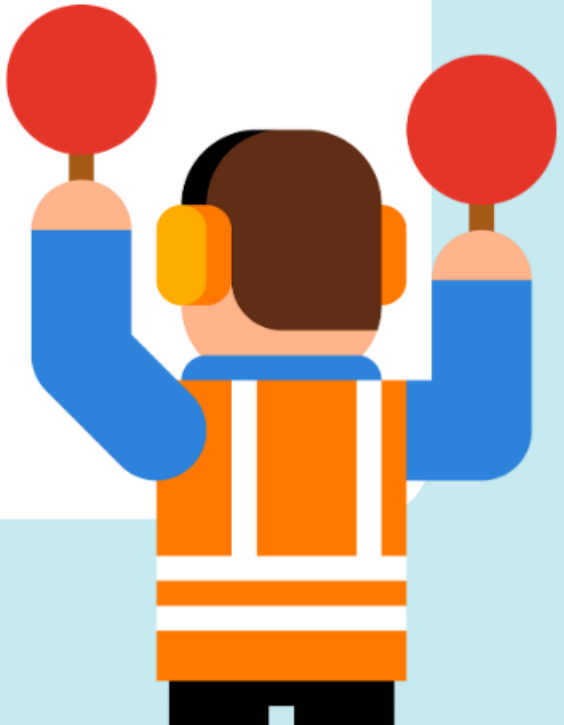
Analytics Architectures on Microsoft Fabric



sqlbits
2024

Freddie Santos

Working for Microsoft over the last 12 years

SQL Server PFE, Escalation Engineer for Azure SQLDW,
Technical Advisor for Azure Synapse now part of Fabric CAT

As a good Brazilian, I like Soccer.

Geek> Manga, Games, RPG and Anime… you name it

Brad Schacht

Principal Program Manager, Fabric CAT

Write books that put people to sleep

I like pizza (Absolutely ***NO*** pineapple!!)

YouTube: Tales from the Field

Blog: BradleySchacht.com

Agenda

How do we discuss analytics solutions?

What is Microsoft Fabric?

Data Warehouse

Data Lake

Lakehouse

Data Mesh

Types of analytics

Descriptive tell us what has happened.

X number of widgets were sold each day, at each location. Peaks and valleys are visible in the data.

Diagnostics tell us why something happened.

We can correlate certain widget sales different times of the year. We can also correlate the peaks and valleys of various widgets to major holidays and seasons.

Predictive tell us what could happen.

A surge in sales of aquatic widgets is forecasted to occur in a few weeks, just before kids begin them spring break from school.

Prescriptive tell us what we should do.

Increase the stock of aquatic widgets, move them to high traffic areas, and add more workers. This will result in more labor cost but that will be offset by increased sales.

The V's of data analytics

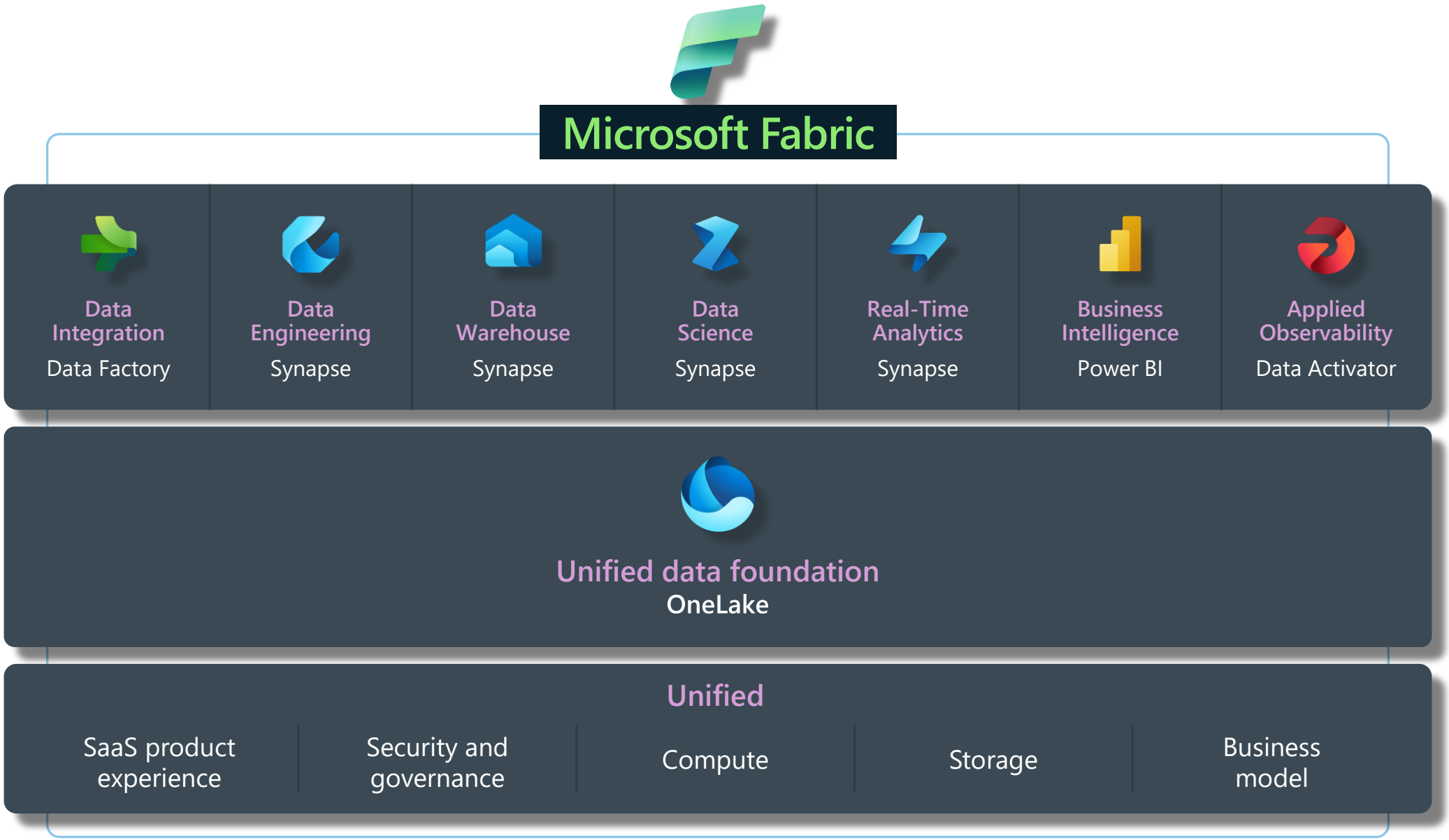
Volume – Amount of data

Velocity – Speed at which data is generated

Variety – Data types and sources

Veracity – Quality, accuracy, and authenticity

Value – What can be done with the data

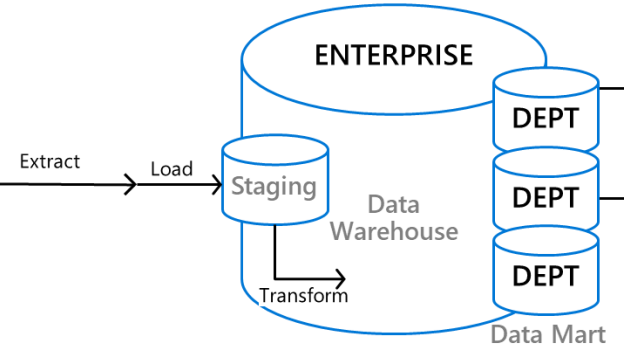

Microsoft Fabric
Data Integration
Data Factory
Data Engineering
Synapse
Data Warehouse
Synapse
Data Science
Synapse
Real-Time Analytics
Synapse
Business Intelligence
Power BI
Applied Observability
Data Activator
Unified data foundation
OneLake
Unified
SaaS product experience
Security and governance
Compute
Storage
Business model

What is your role?

Evolution of analytics

Late 1980's

Data Warehouse



Data warehouse

Benefits

Create a single “version of the truth”.

“What are the trends in the market?”

ETL into a consistent, organized format (relational database).

Fast and easy data exploration and visualization.

Kimball model introduced in 1996.

Challenges

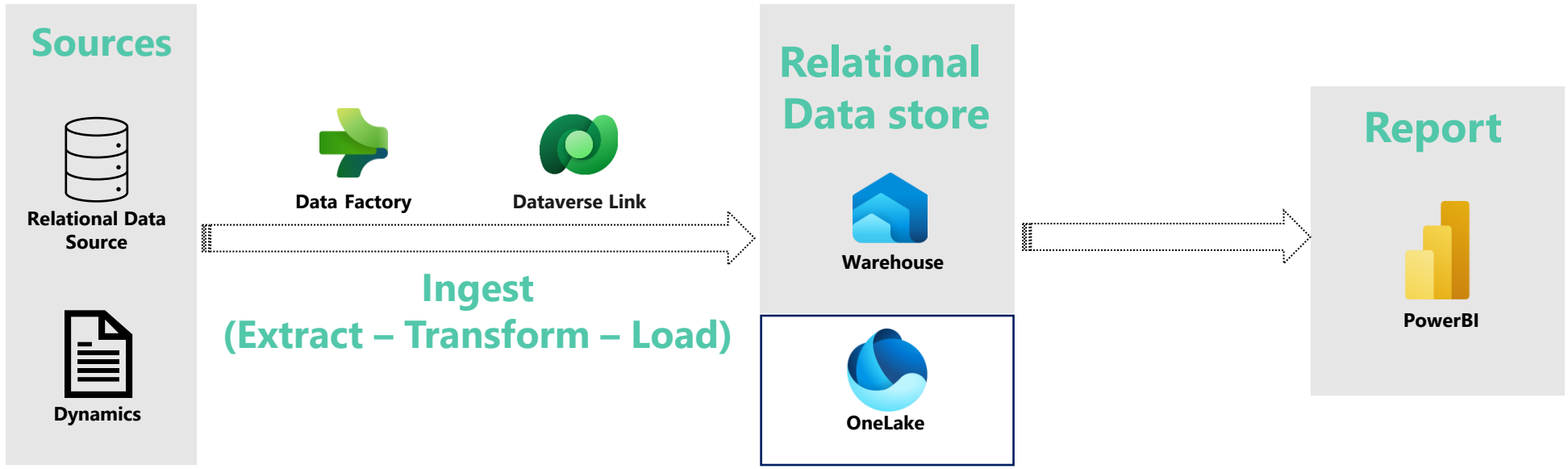
Long development cycles leading to project failure.

Data integration across systems.

Integrate new historical data points to the model.

Performance and scalability suffer when a single server gets overloaded.

Data warehouse on Fabric



How do we get that performance?

sqlbits
2024

Resource allocation

1. Stats updated if necessary
2. Cost calculated, resources requested
3. Resources provisioned
4. Query executed across nodes
5. Nodes torn down

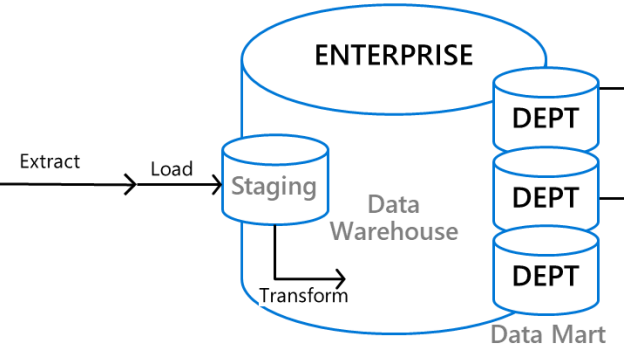
Bursting

Regardless of SKU, Fabric bursting will automatically allocate resources as needed to execute at maximum performance.

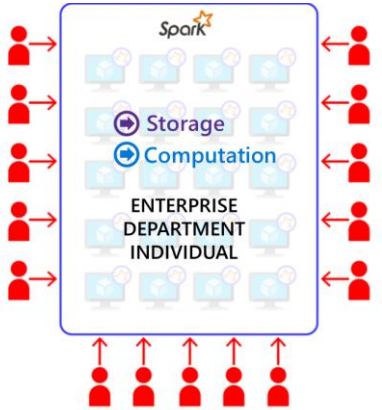
Resource isolation

Evolution of analytics

Late 1980's
Data Warehouse



2011
Data Lake



Data lake

Scalability: Data lakes can handle large data volumes that grow and fluctuate based on data inputs. Making them a good option for businesses with rapidly increasing data storage needs.

Low cost: Data lakes use technologies that are cost-effective for organizations.

Agility: Data scientists can prepare and analyze data models rapidly using data lakes.

Flexibility: Data lakes can store structured, semi-structured, and unstructured data in its native format, which makes it easier to integrate with different types of applications.

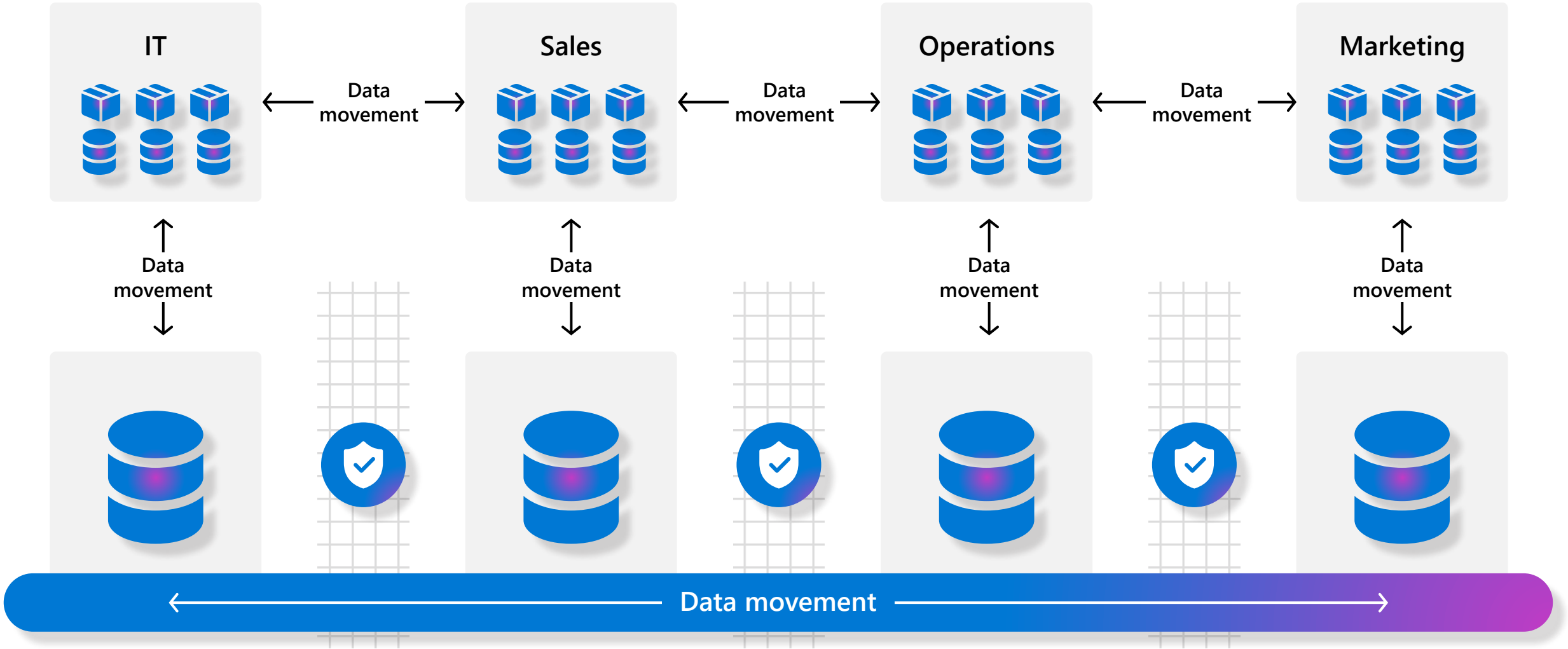
Data exploration: Data lakes allow users to explore raw data without the need for pre-defined schemas or structures of a Relational Store.

Data lake challenges

- Ungoverned
- Unstandardized Data
- Hard to find
- Hard to use
- Hard to Secure

Data Swap

Common Problems



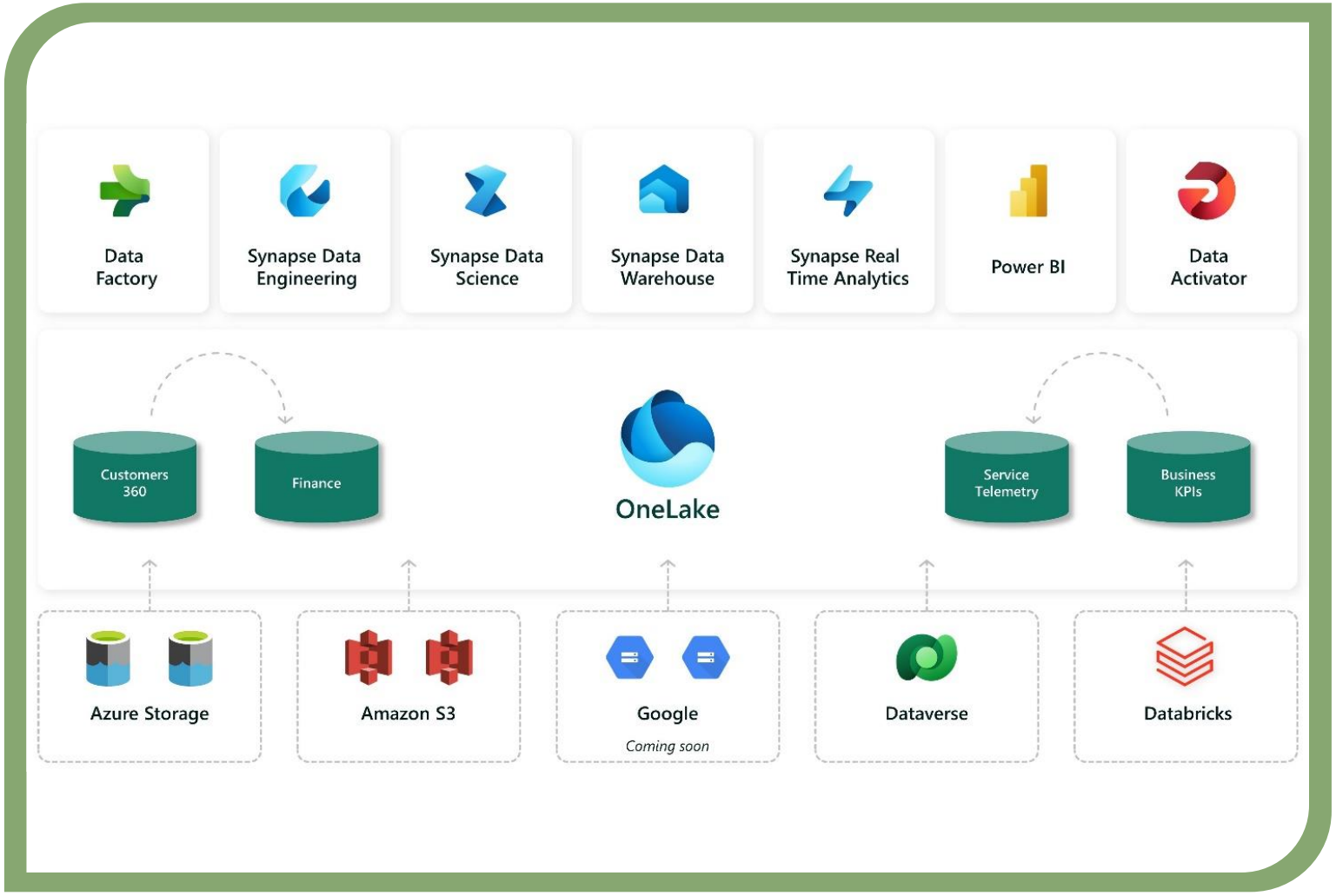
Data lake on Fabric

Shortcuts unify data without copying or moving existing data.

This means that data can be used multiple times without data duplication.

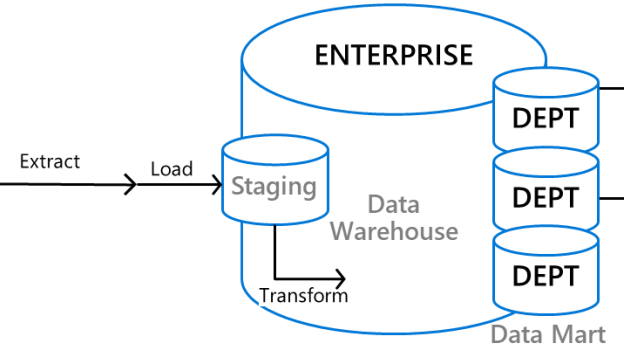
Key Capabilities:

- Create shortcuts within Microsoft Fabric to consolidate data across artifacts or workspaces, without changing ownership of the data
- With shortcuts, data throughout OneLake can be composed together without any data movement
- Shortcuts can be external that links external data sources (ADLS, S3, Google Cloud) or internal, from logical entities such as tables, or physical as folders.
- With support for industry standard APIs, OneLake data can be directly accessed by any application or service

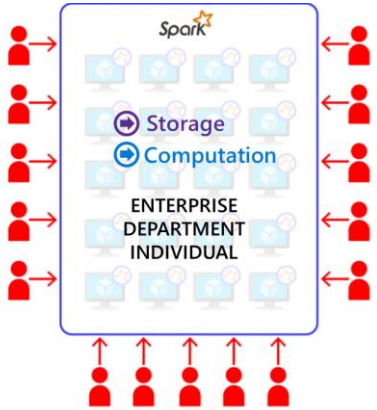


Evolution of analytics

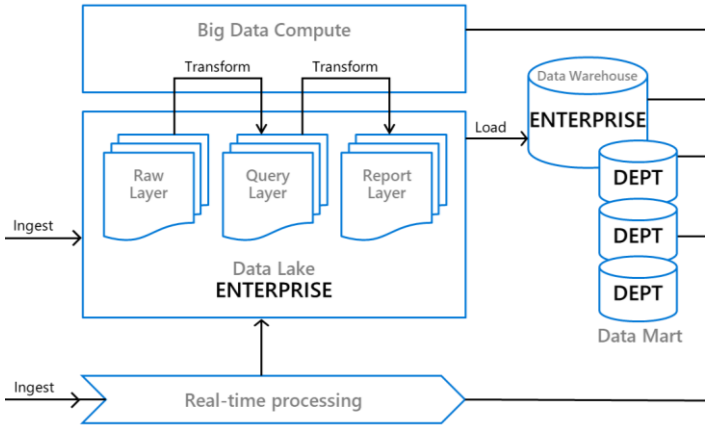
Late 1980's
Data Warehouse



2011
Data Lake



Mid 2010's
Cloud Data Platform



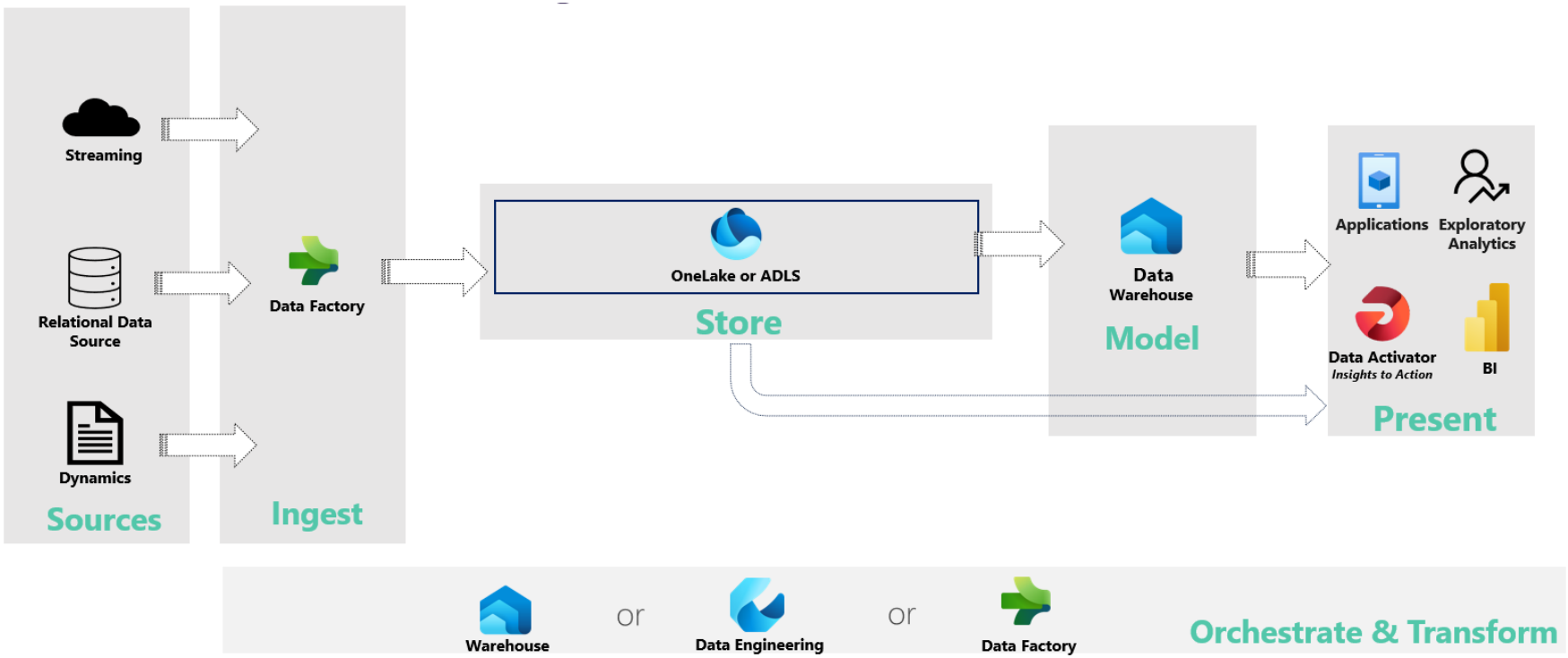
Cloud data platform

Flexibility	Scalability	Accessibility
Similar concepts	Any size	Lower barrier of entry (cost and difficulty)
New tools	On-demand scale	New technology available to everyone
Multi-region	Multi-region	

Cloud data platform challenges

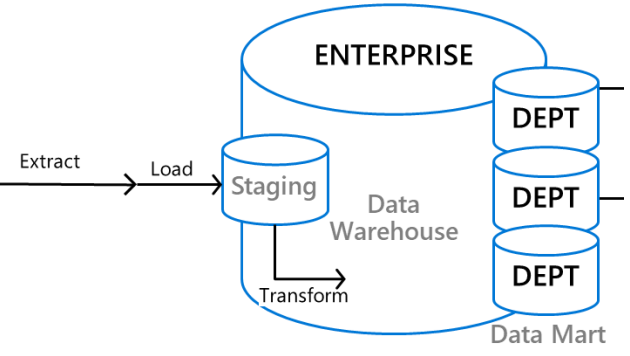
Complexity	Data Integration	Security	Vendor Lock-in
Multiple components Management and orchestrating	Sources span on-prem, cloud, streaming, APIs, etc.	Networking Encryption Mixed identities Auditing	Limited interoperability Limited portability of data and tools

Cloud data platform on Fabric

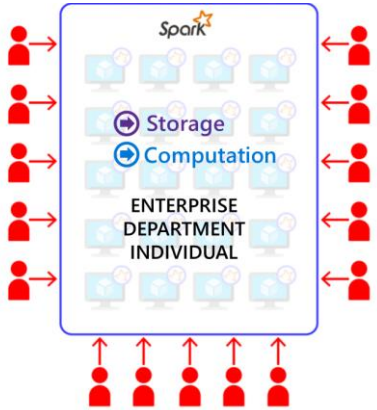


Evolution of analytics

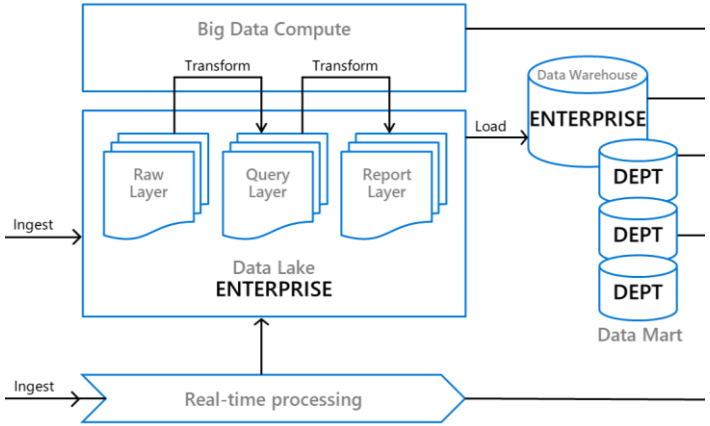
Late 1980's
Data Warehouse



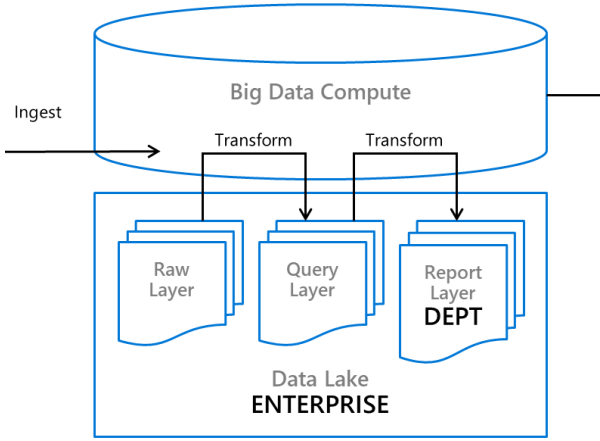
2011
Data Lake



Mid 2010's
Cloud Data Platform



2020
Data Lakehouse



Data lakehouse

Data Lake + SQL + Performance + Python + Data Science

Low TCO

Open-source format

Flexibility

Unlimited rainbows, butterflies, unicorns, and free pizza for all
(no pineapples, seriously, why people, just why?).

Delta Lake

ACID transactions

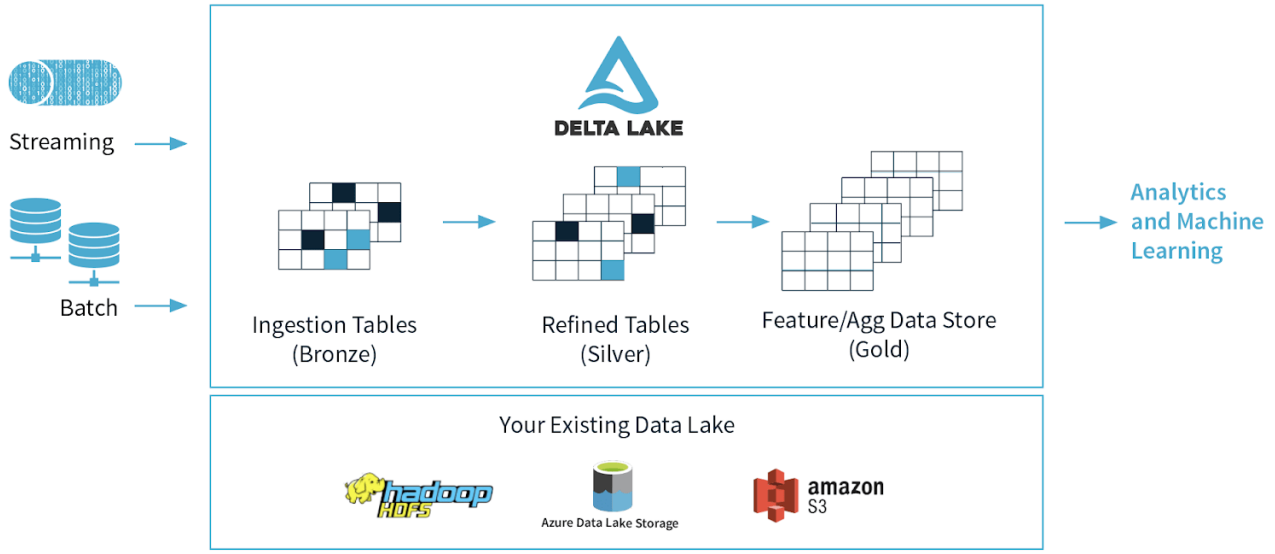
Time travel

Schema enforcement

Append, update and delete

Performance improvement

Streaming and batch unification



Data lakehouse challenges

Reporting speed
(especially vs. MPP)

Security

Missing functionality
(foreign keys, TDE, workload management, Spark/SQL separation)

Learning curve

Lakehouse on Fabric

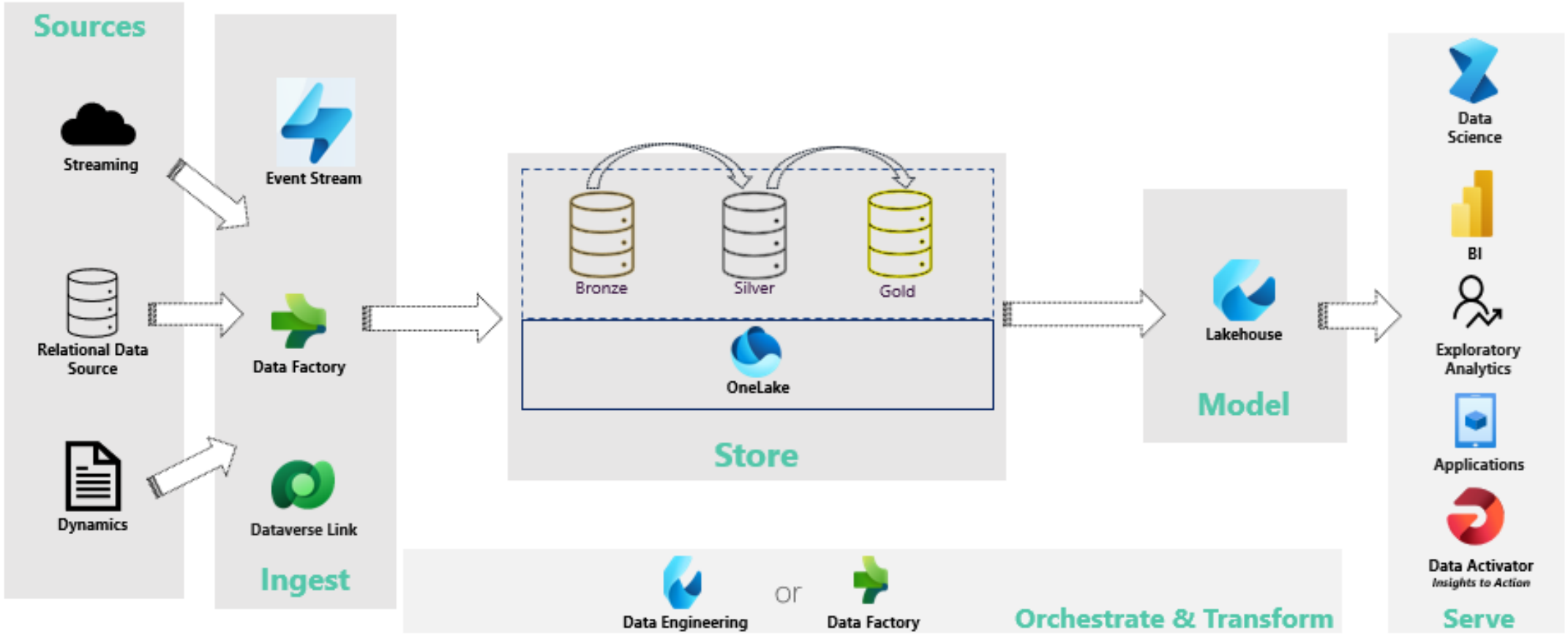
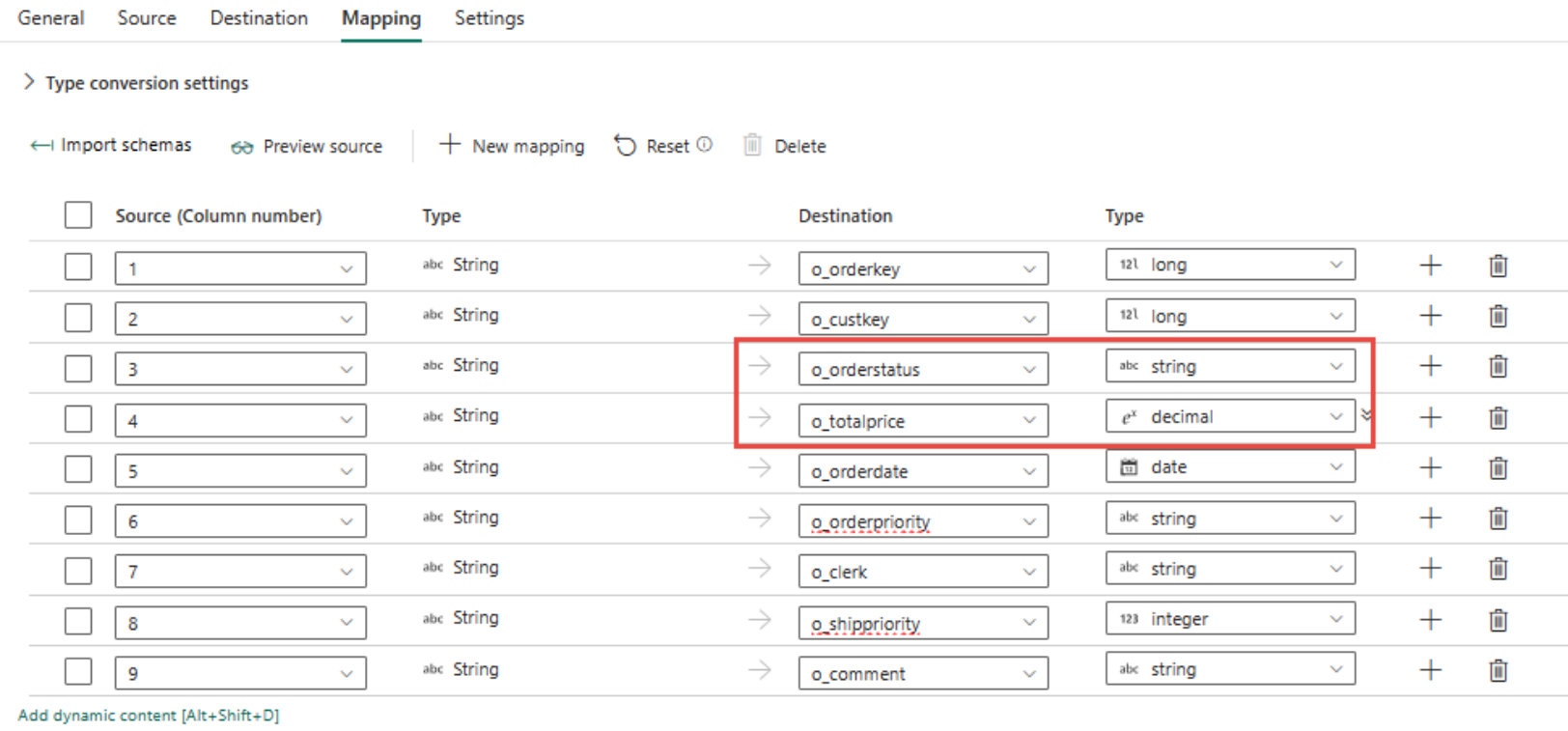
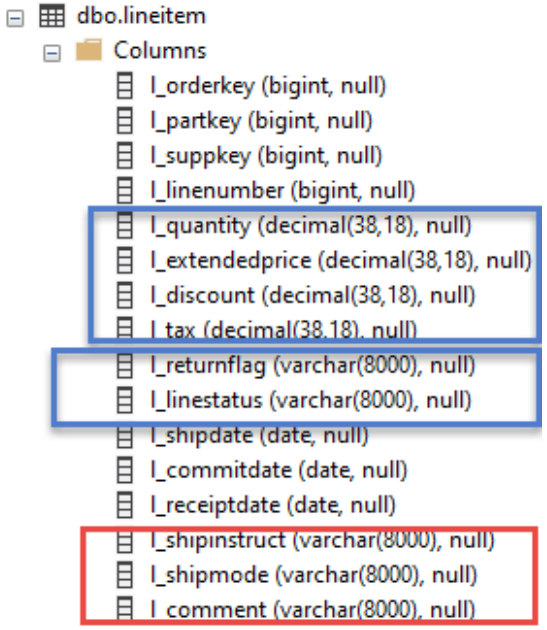


Table definition matters

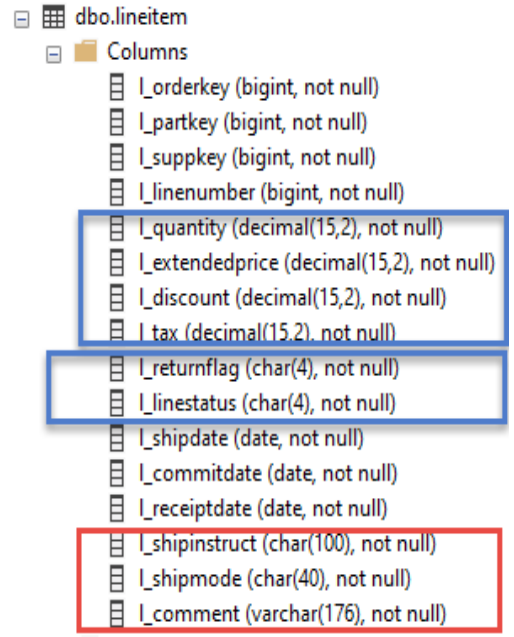


sqlbits
2024

Table definition matters



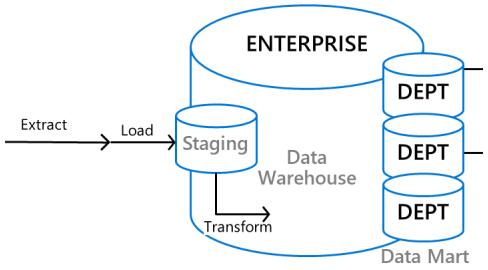
8 minutes 10 seconds



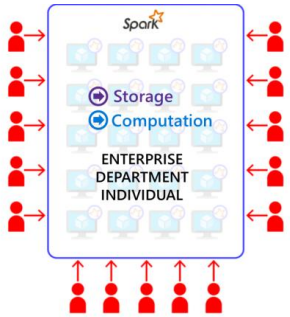
5 minutes 15 seconds

Evolution of analytics

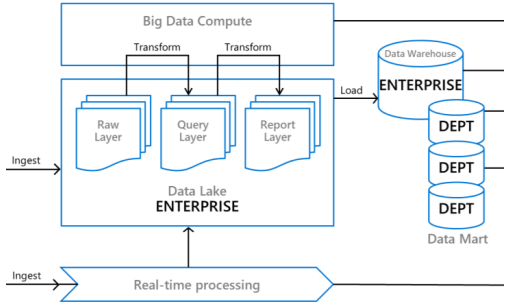
Late 1980's
Data Warehouse



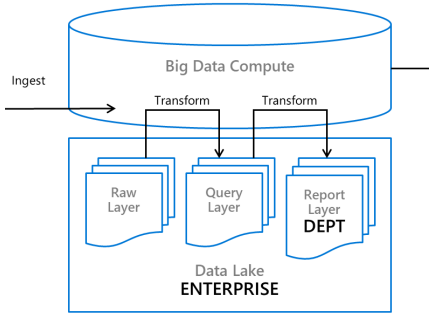
2011
Data Lake



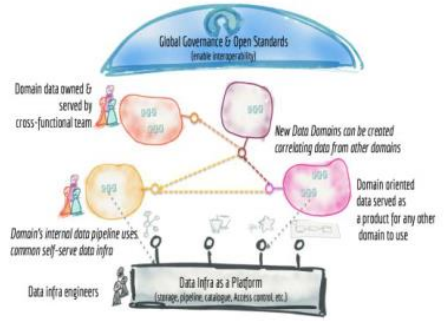
Mid 2010's
Cloud Data Platform



2020
Data Lakehouse



2021
Data Mesh



Data mesh

Domain-oriented: Decentralized data ownership and architecture.

Data as a product: This principle treats data as a product that is designed, built, and managed by the domain team.

Self-service: Provides a self-service platform for domain teams to build, test, and deploy their data products.

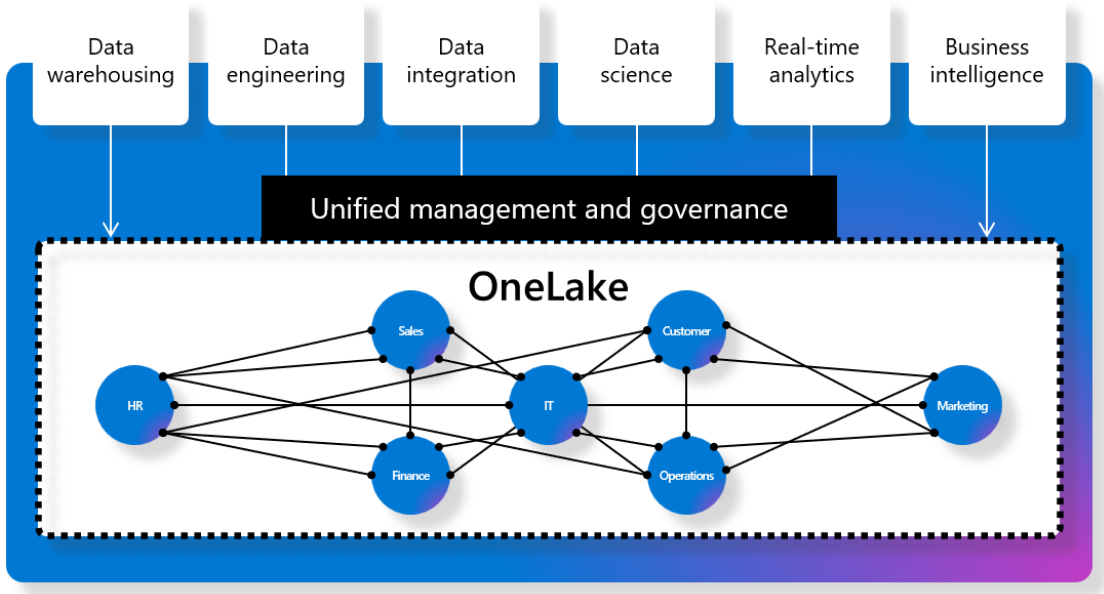
Federated computational governance: Domain teams have the autonomy to govern their data products.

Data mesh on Fabric

One Copy enables data to be used across domains

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.



Data mesh challenges

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.

Self-serve approach of data requests could be challenging.

Creation of data silos for domains not able to join data mesh.

What is the answer?

sqlbits
2024

Comparing the options

Architecture	Volume	Velocity	Variety	Veracity	Value
Data Warehouse	100s TB	🔴	🔴	🟡	🟢
Data Lake	PB+	🟢	🟢	🟡	🔴
Cloud Data Platform	PB+	🟢	🟢	🟡	🟡
Data Lakehouse	TB to PB	🟢	🟡	🟡	🟡
Data Mesh	PB+	🟢	🟢	🟡	🟡

Questions

Thank you

sqlbits
2024