SQL Saturday Toronto 2023

Toronto Data Professionals Community (TDPC)

SQL Saturday (#1064)

Our Sponsors

Microsoft

OMERS

Hitachi Solutions

ADASTRA

databricks

Community Support

Toronto Data Professionals Community (TDPC), one of the largest data professional's community in Toronto, host monthly event which offers interactive learning built by community and guided by trusted data experts.

TDPC Event Partners

Toronto Data Professionals Community (TDPC)

SQL Saturday (#1064)

Get Ready for Data Mesh and MS Fabric Domains

Toronto Data Professionals Community (TDPC)

SQL Saturday (#1064)

Intro

Mladen Celikovic

- Sr Data Engineer | Architect @ Cyderes | Herjavec Group
  https://www.cyderes.com
- About:
  - As a data champion, witnessed Microsoft data platform evolution since SQL 2000 and all subsequent versions through Business Intelligence to the latest and greatest modern world of Data Engineering, Architecture and Analytics. Passionate runner, sports lover, music fan.

- Email:
  - mladen.celikovic@cyderes.com
  - mladen.celikovic@bluedotanalytics.ca
- LinkedIn: https://www.linkedin.com/in/mladencelikovic
- Blog: https://sqlsrv4living.blogspot.com
- Twitter: https://twitter.com/MidWeekSQLBlues

Highlights

Data Mesh

- Core Concepts and Architecture
- Lakehouse
- Data Sharing and Security
- Data Governance
- Monitoring

MS Fabric

- Domains
- Demo with Business Use Cases

*Data Mesh

Core Data Mesh Concepts

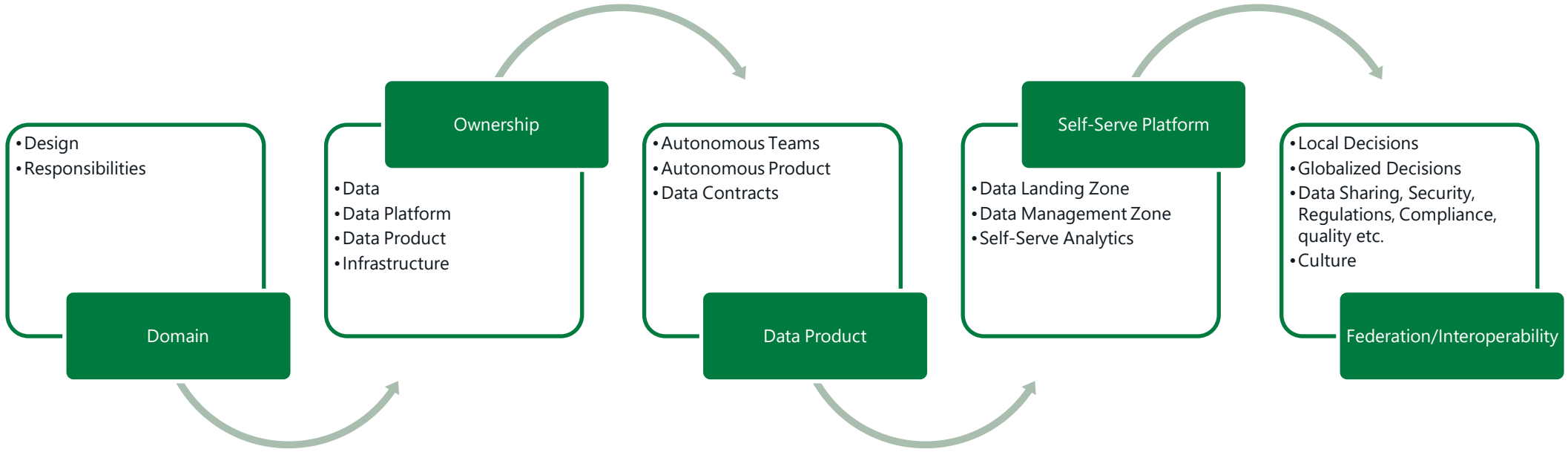
Data Domains and Ownership

Data Products

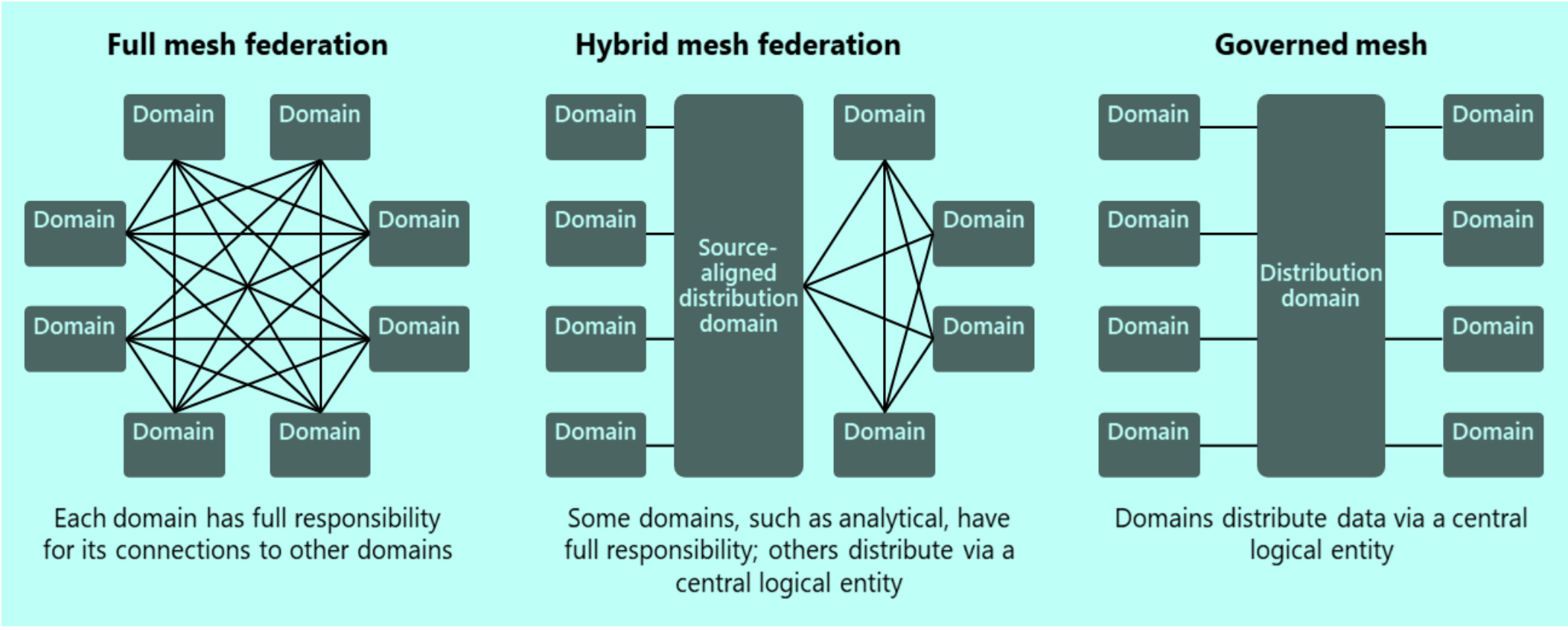
Data Platform

Federated Governance

Core Data Mesh Concepts



Design Considerations



Infrastructure

Design

Technology stack

Azure Architecture

Environments (Dev, QA, SIT, UAT)

Logical environments – Domains

Migration

Develop

Scripting:

Bicep,

Terraform,

PowerShell,

Python,

Etc.

Build and Deploy

Code Repository

CI/CD

Execute

Networking

Migration

Integration

Extensibility

Scalability

Environment Configuration

Support

Access

Security

Data Sharing Policy

RBAC

Data Privacy

Monitor

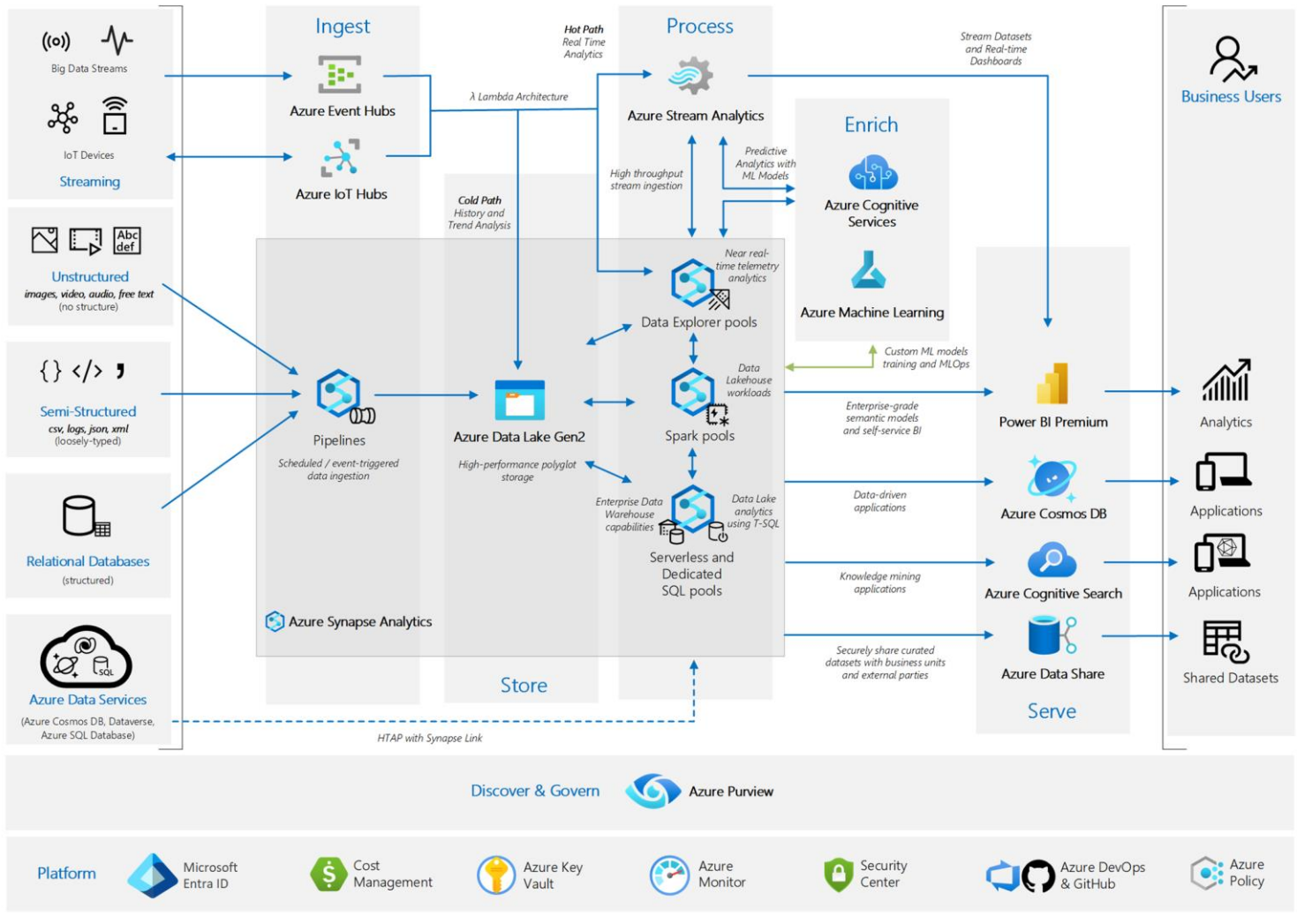
Logs

Audit

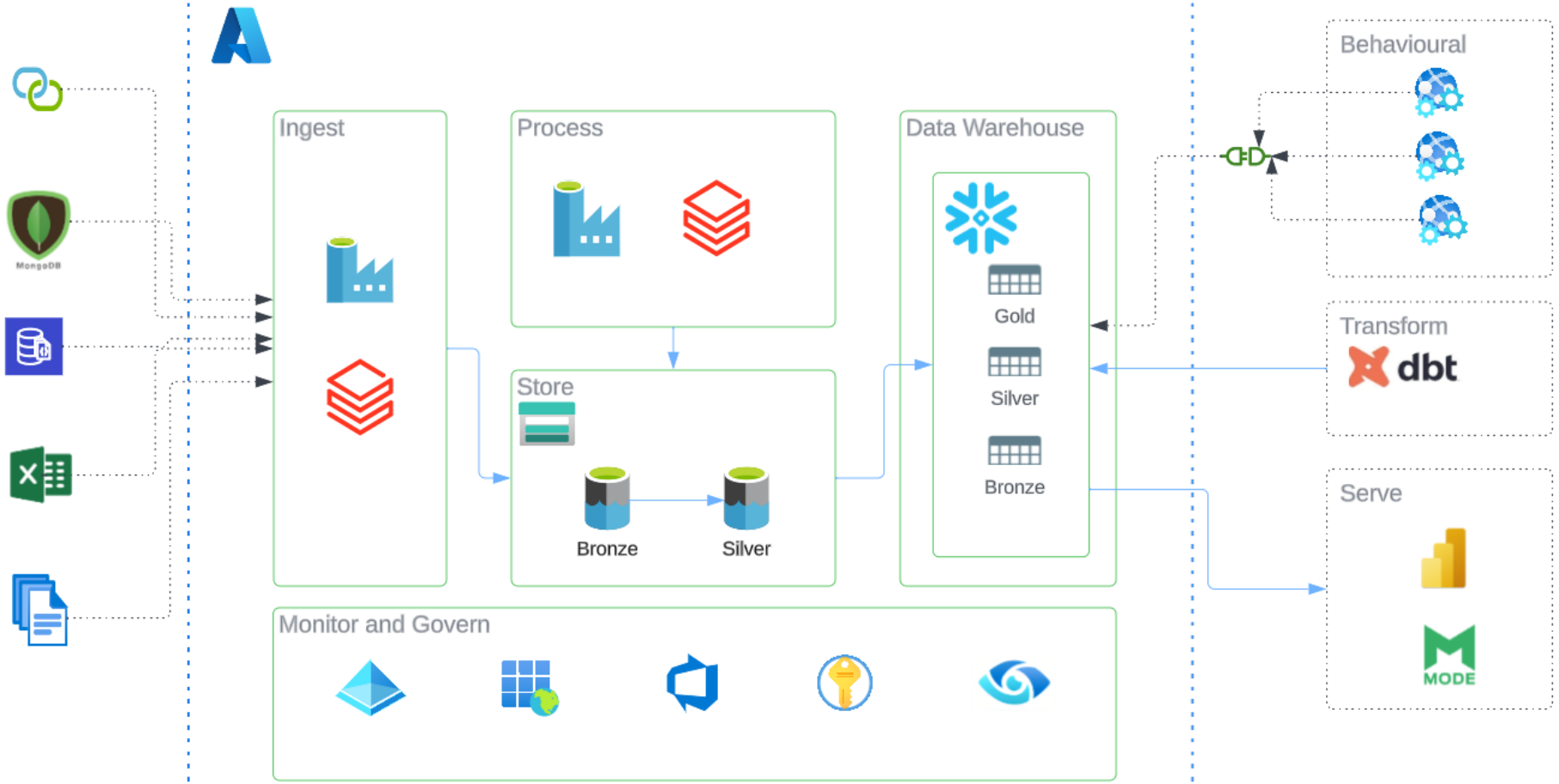
Performance

Cost Management

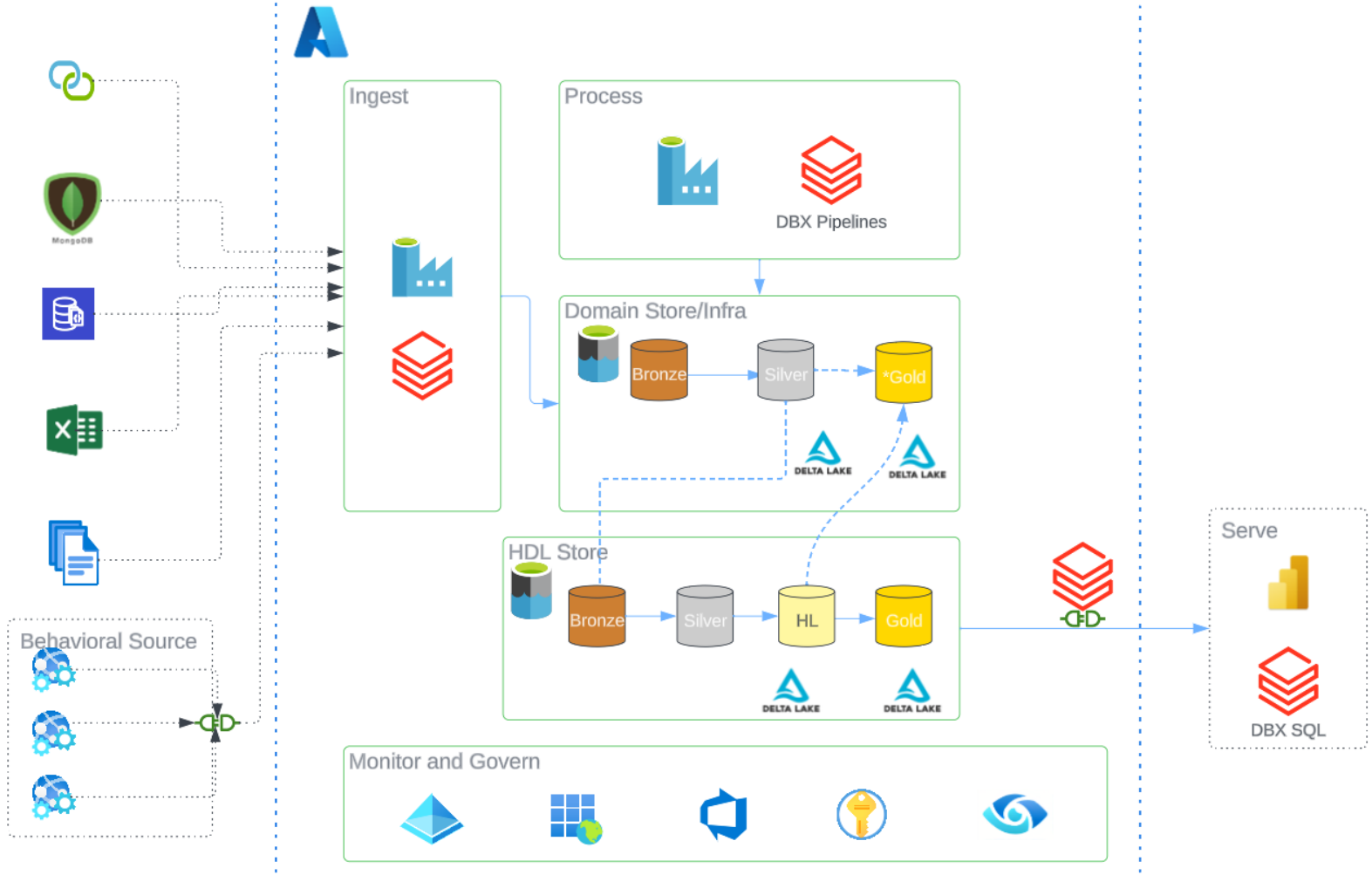
Architecture – Analytics



Architecture – Real World Example



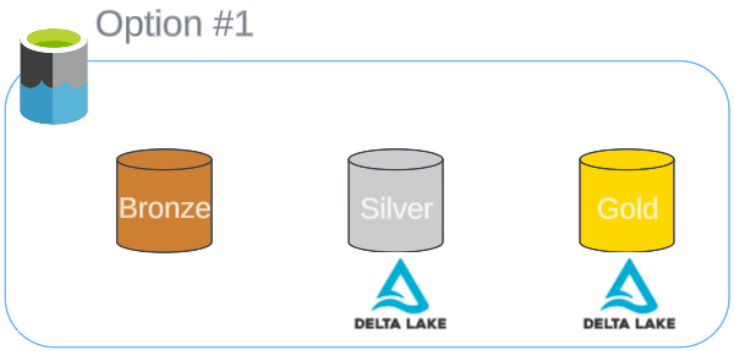
Architecture – Real World Example

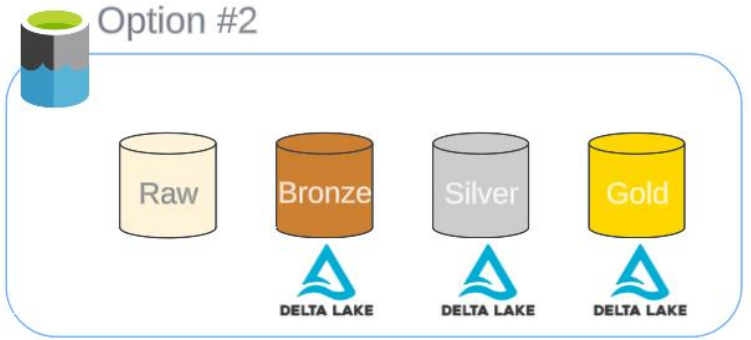


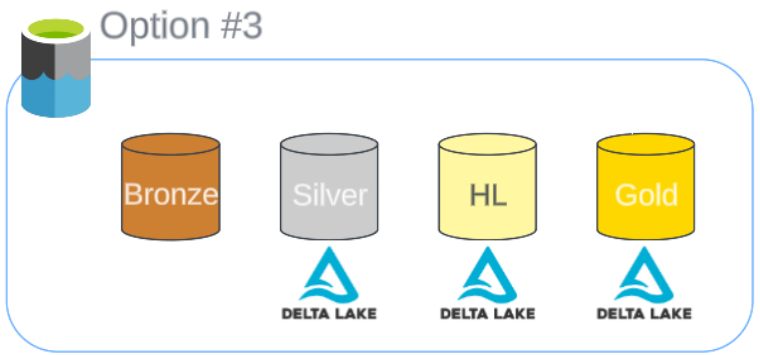
Architecture – contd.

- Previously optimized architecture:
  - Created simplified and linear “left to right” data flow
  - Removed technology redundancy (Snowflake, Databricks, dbt)
  - Avoided medallion layer/data redundancy (bronze, silver, gold)
  - Ingesting Behavioural source straight to data lake
  - Added harmonized data layer for intermediate business transformations

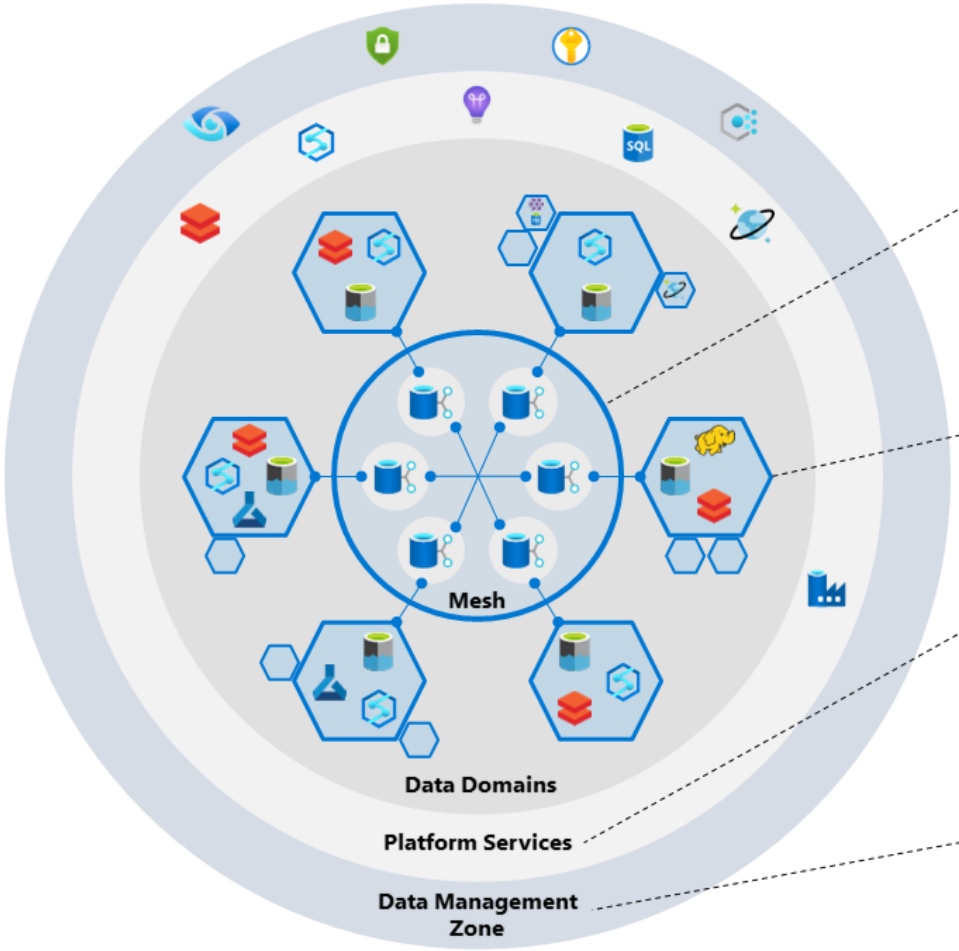
Lakehouse


Option #1
Bronze
Silver
DELTA LAKE
Gold
DELTA LAKE


Option #2
Raw
Bronze
DELTA LAKE
Silver
DELTA LAKE
Gold
DELTA LAKE


Option #3
Bronze
Silver
DELTA LAKE
HL
DELTA LAKE
Gold
DELTA LAKE

Architecture – Data Mesh Example



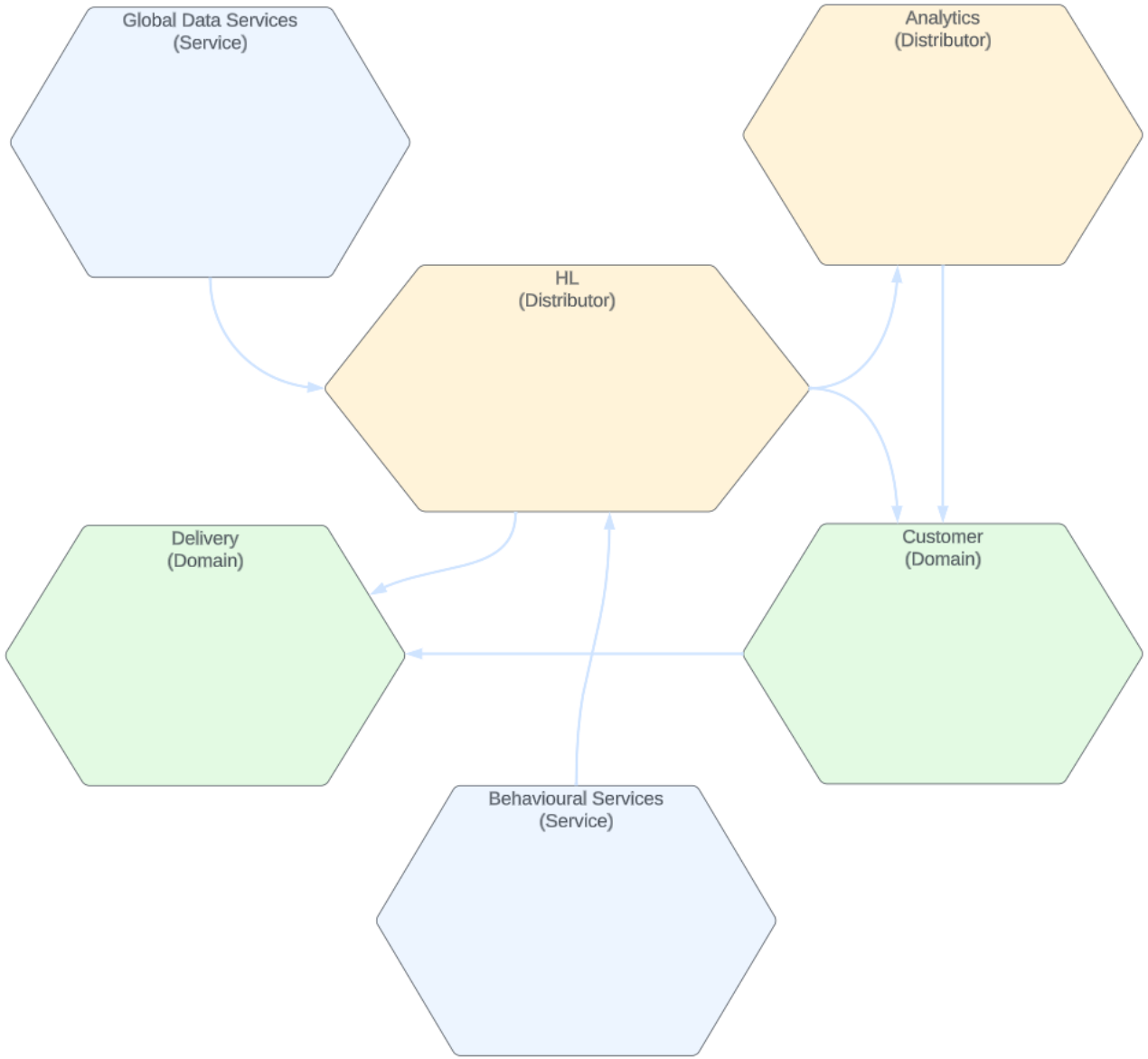
The **data mesh** intelligently distributes data products between data domains. Read data stores share compute resources. This reduces costs, solves interoperability concerns, and better addresses time-variant and non-volatile concerns of large data consumers.

Data domains operate their own applications or analytics platforms, whilst adhering to common policies and standards.

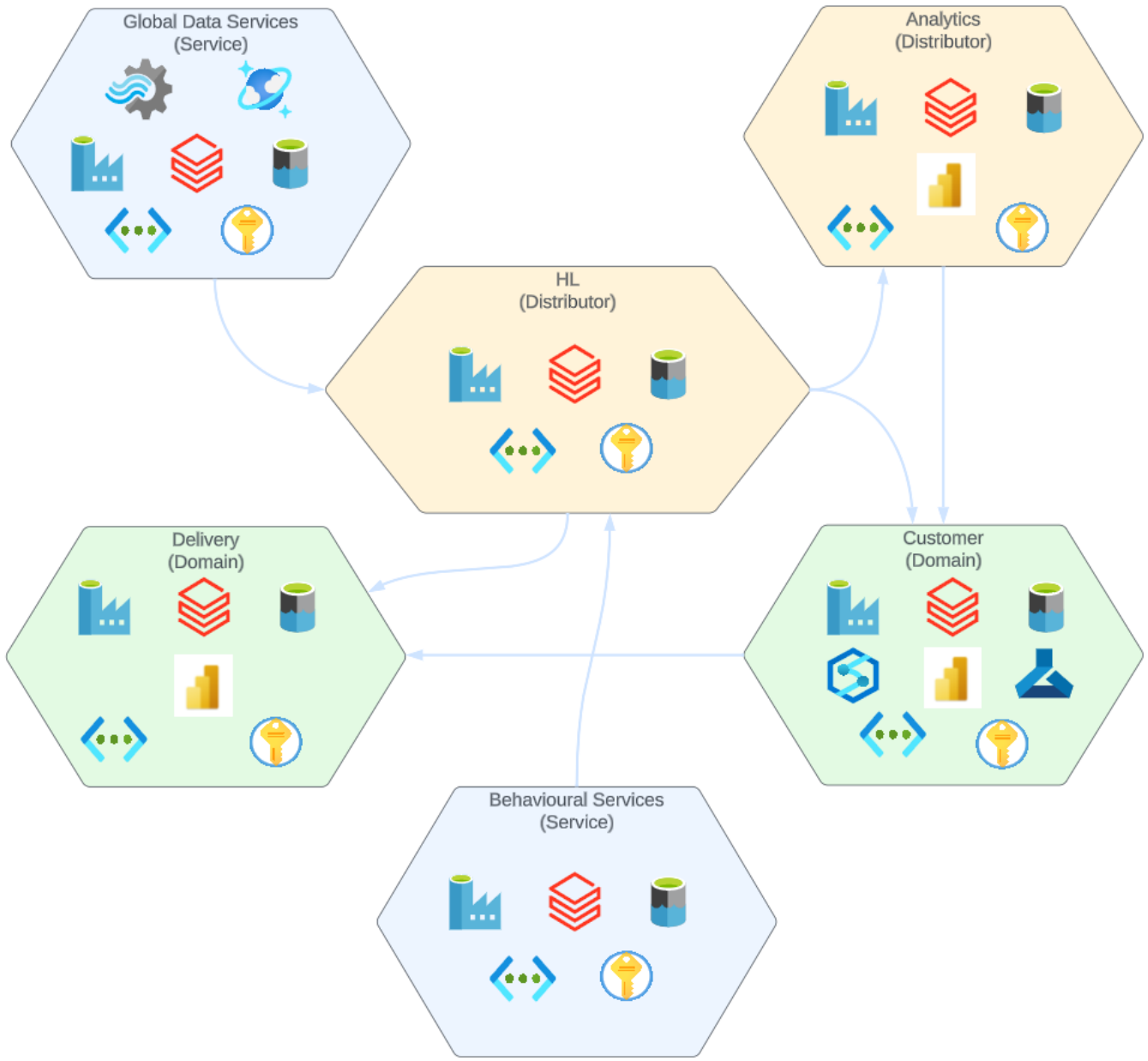
The central **platform services** defines blueprints that encompass baseline security, policies, capabilities, and standards.

A key concept for every enterprise-scale analytics and AI implementation is having one **data management zone**. This subscription, which is required for data management, contains resources that'll be shared across all landing zones.

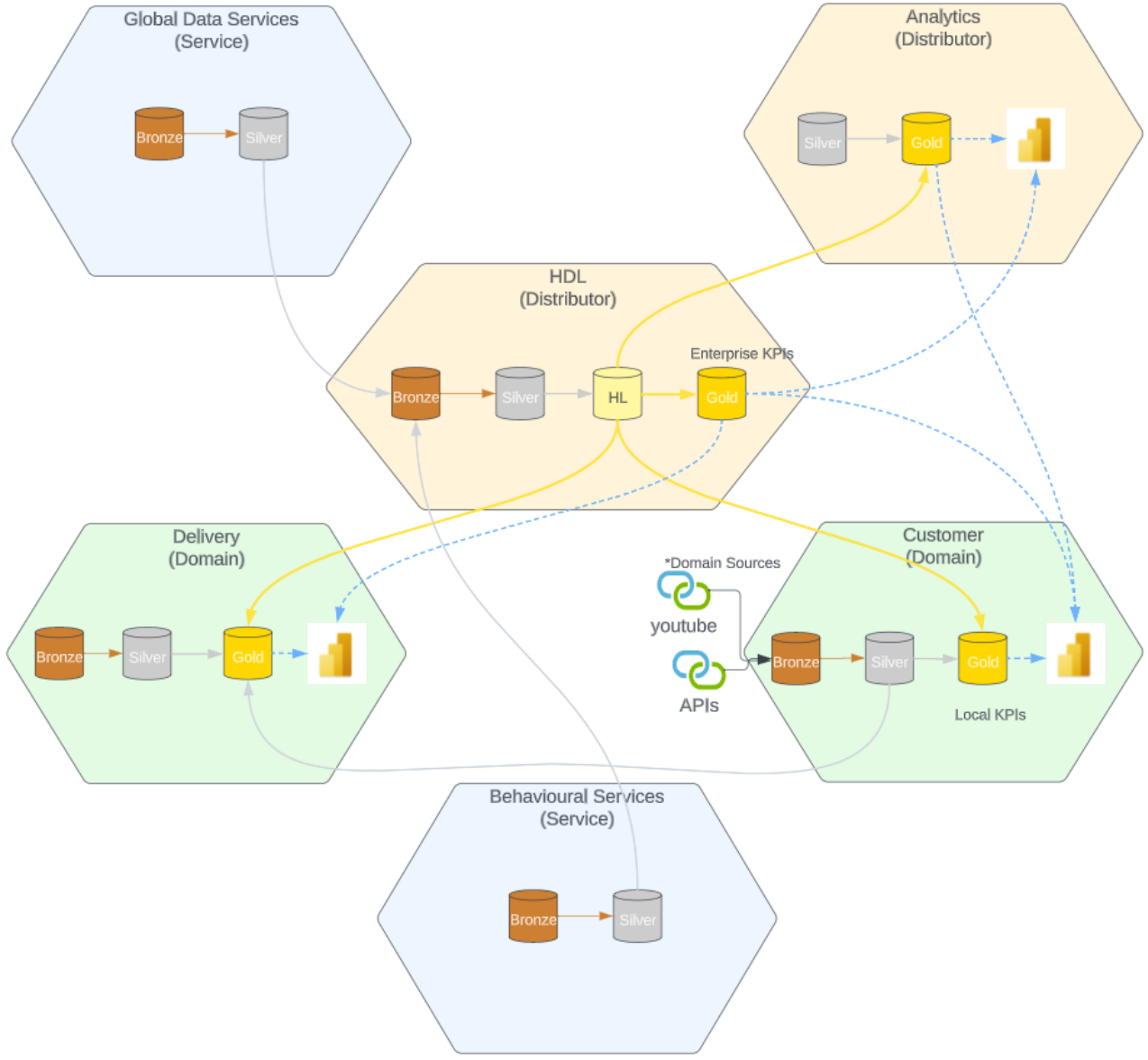
Domains


Global Data Services
(Service)
Analytics
(Distributor)
HL
(Distributor)
Delivery
(Domain)
Customer
(Domain)
Behavioural Services
(Service)

Infra

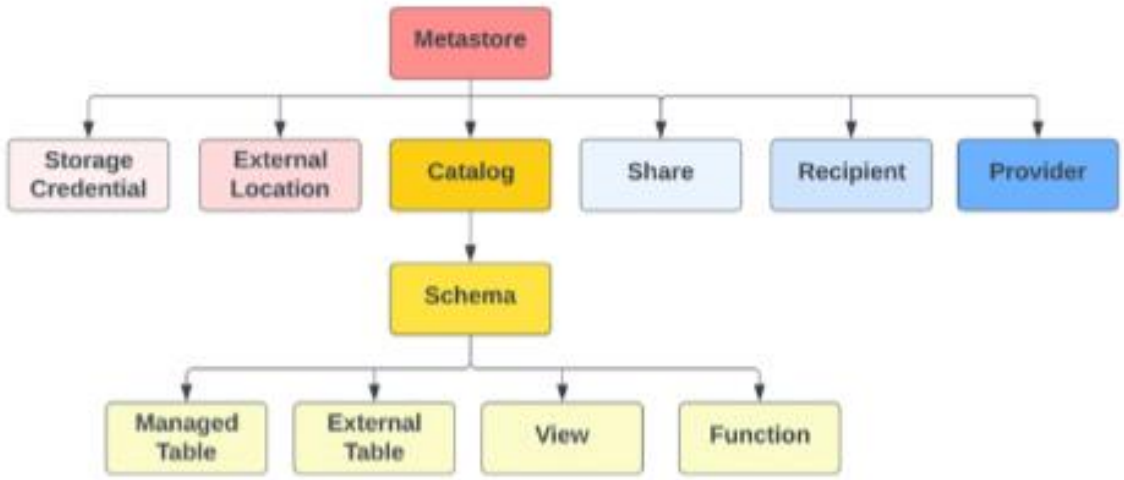

Global Data Services
(Service)
Analytics
(Distributor)
HL
(Distributor)
Delivery
(Domain)
Customer
(Domain)
Behavioural Services
(Service)

Data Sharing

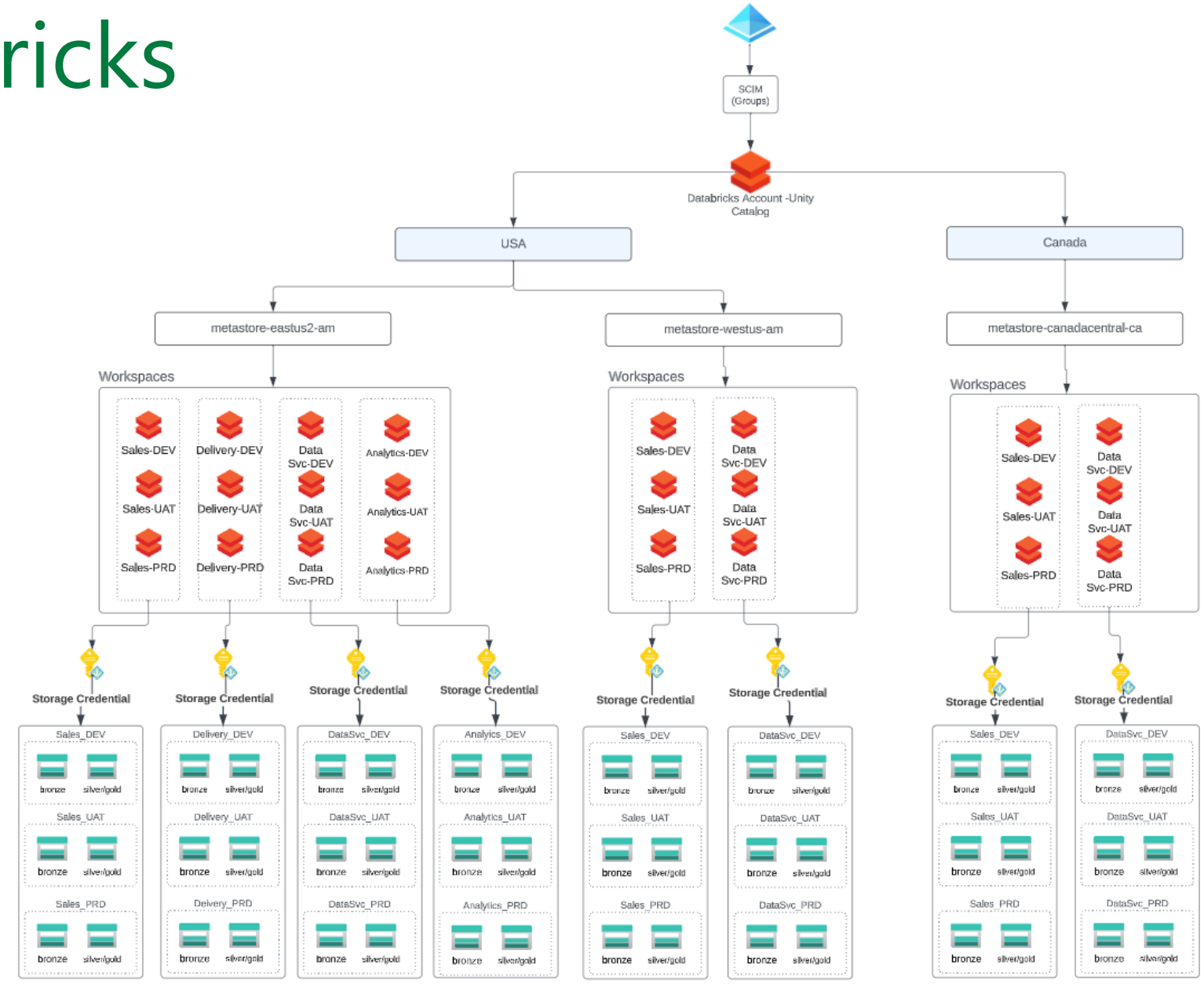


Databricks

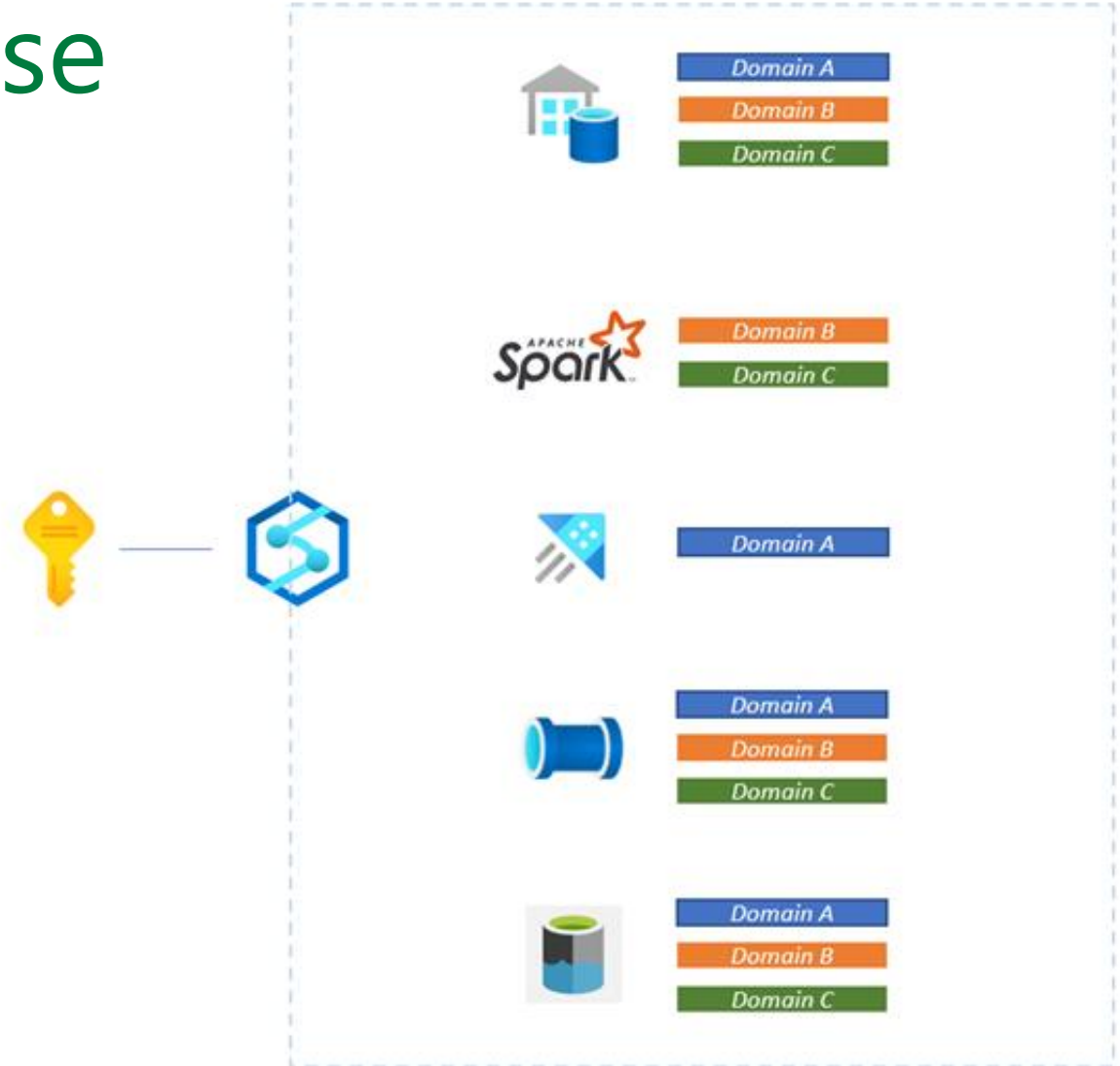

WITHOUT UNITY CATALOG
Databricks Workspace 1
Access Controls
User Management
Metastore
Clusters
SQL Endpoints
Databricks Workspace 2
Access Controls
User Management
Metastore
Clusters
SQL Endpoints
WITH UNITY CATALOG
Unity Catalog
Access Controls
User Management
Metastore
Databricks Workspace 1
Clusters
SQL Endpoints
Databricks Workspace 2
Clusters
SQL Endpoints


Metastore
Storage Credential
External Location
Catalog
Share
Recipient
Provider
Schema
Managed Table
External Table
View
Function

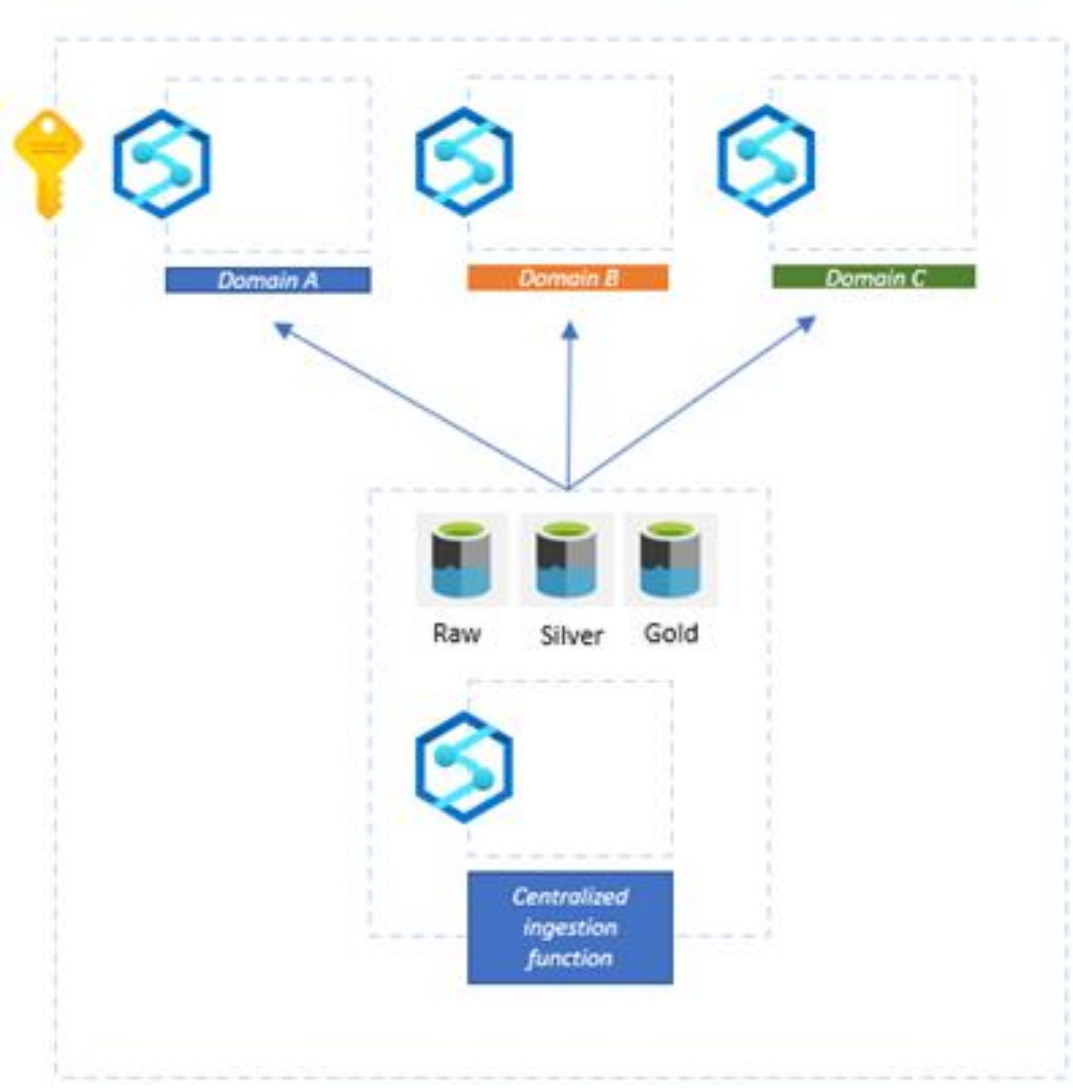
Databricks


SCIM (Groups)
Databricks Account -Unity Catalog
USA
Canada
metastore-eastus2-am
metastore-westus-am
metastore-canadacentral-ca
Workspaces
Sales-DEV
Sales-UAT
Sales-PRD
Delivery-DEV
Delivery-UAT
Delivery-PRD
Data Svc-DEV
Data Svc-UAT
Data Svc-PRD
Analytics-DEV
Analytics-UAT
Analytics-PRD
Storage Credential
Sales_DEV
Sales_UAT
Sales_PRD
Delivery_DEV
Delivery_UAT
Delivery_PRD
DataSvc_DEV
DataSvc_UAT
DataSvc_PRD
Analytics_DEV
Analytics_UAT
Analytics_PRD
bronze
silver/gold

Azure Synapse



Azure Synapse

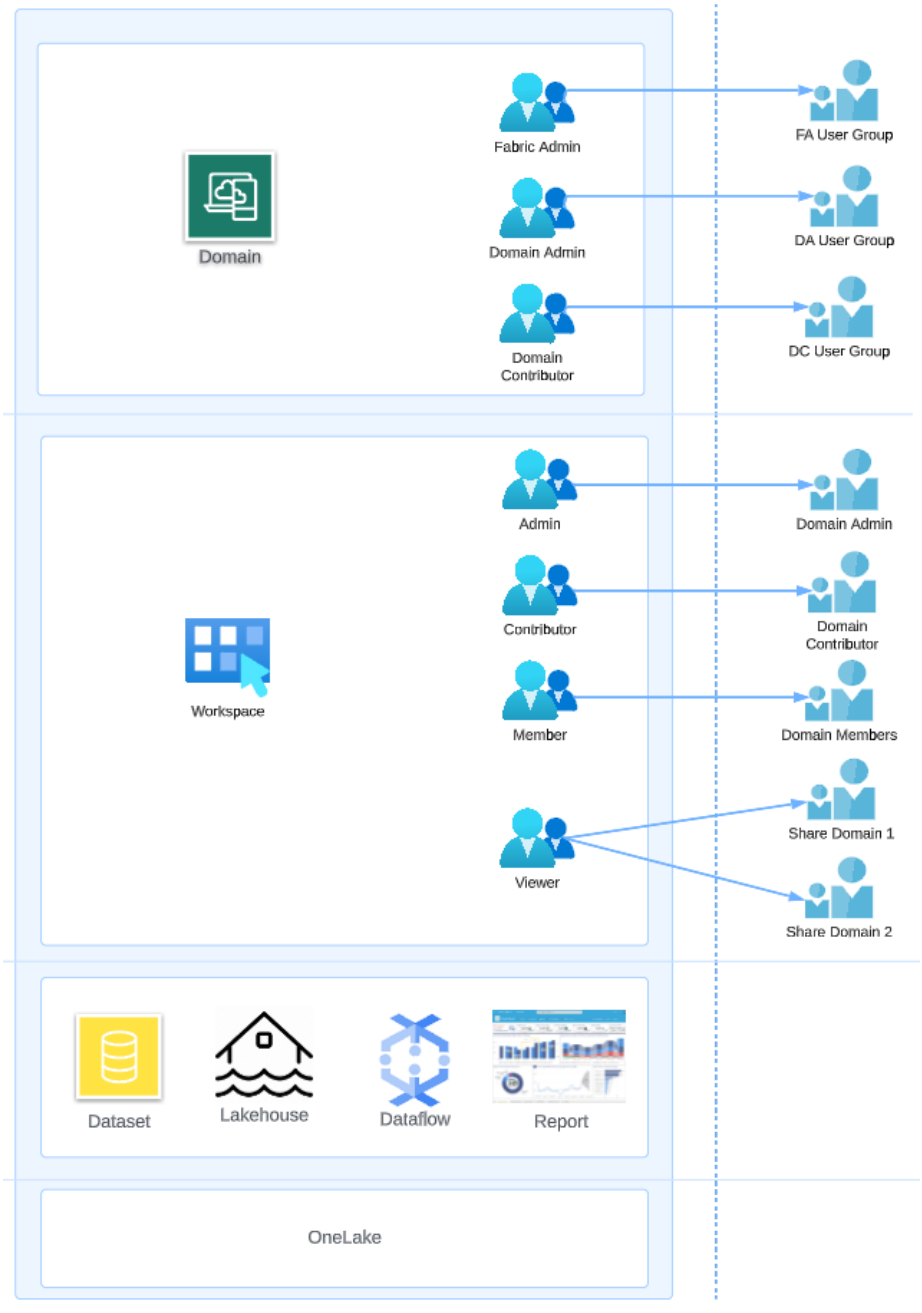


Microsoft Fabric

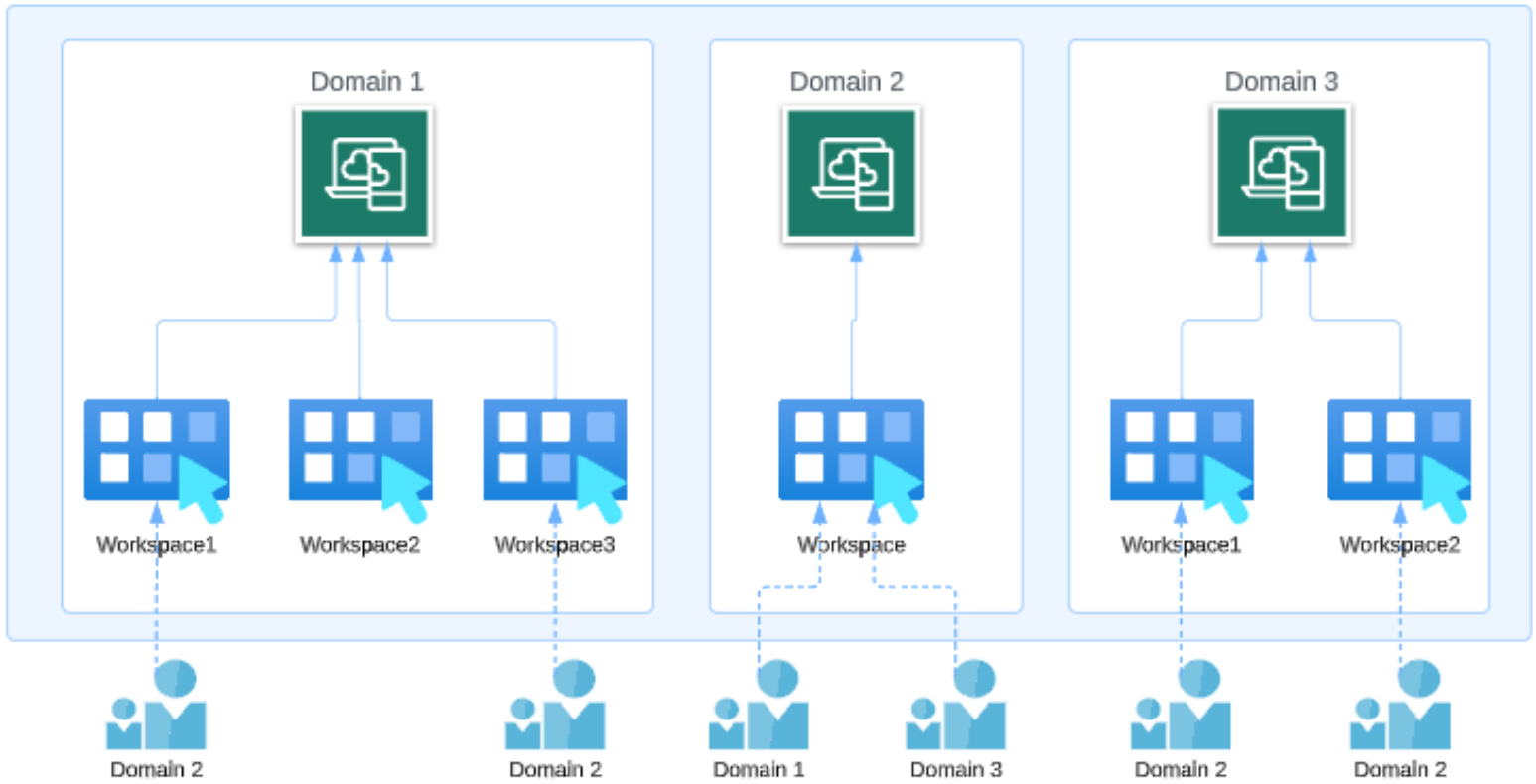
MS Fabric Domains (Preview)

- Logical grouping of domain's entities
- Easy to organize, filter, find and determine the proper data sharing and access
- OneLake Shortcuts – easy way of accessing data without duplication
- Domain access roles:
  - Fabric Admin
  - Domain Admin
  - Domain Contributor

Microsoft Fabric



Microsoft Fabric



Roles and Users

Roles:

Admin

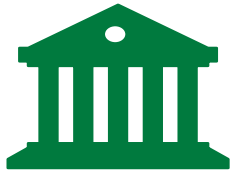
Analyst

Engineer

ETL

Scientist

Reporting End Users



Users:

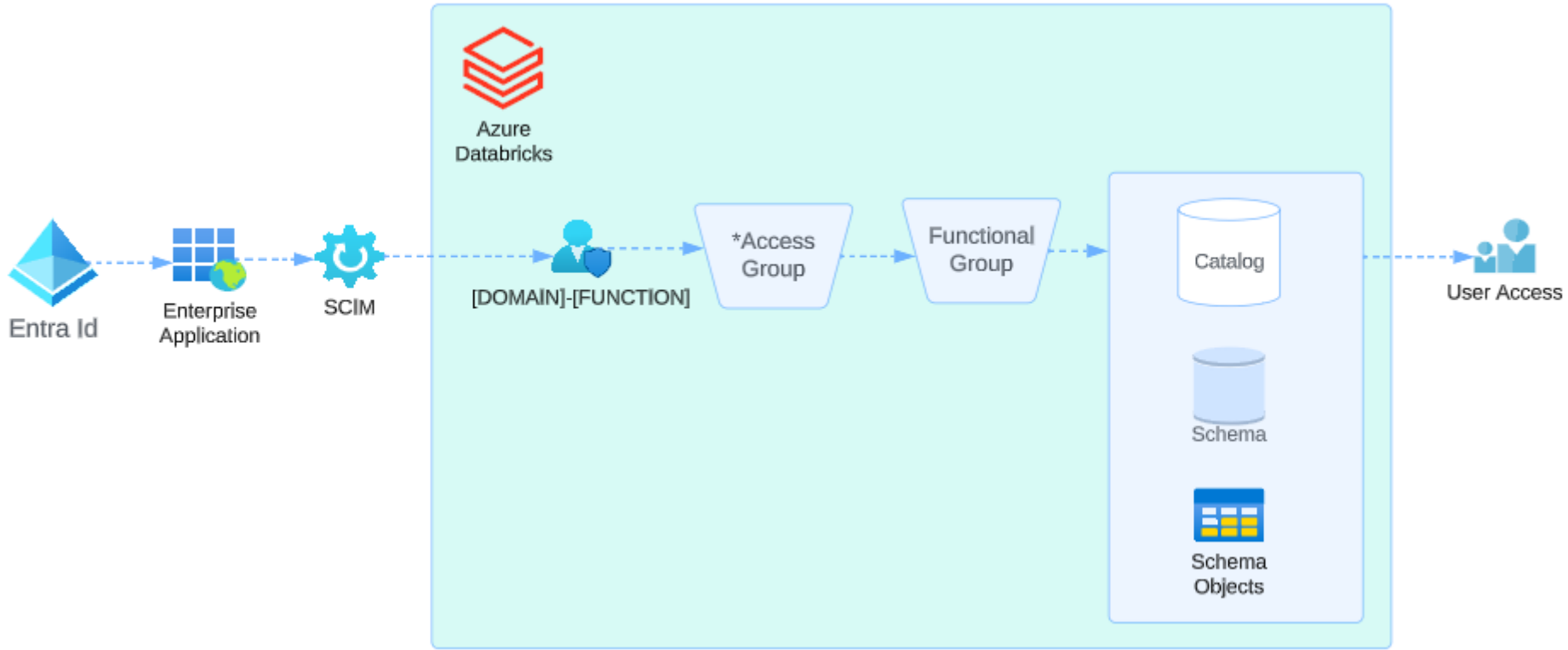
Domains

Internal

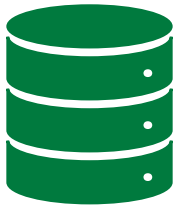
Distributed (GDPR and other regulations, Different Regions, etc.)

External

RBAC Security



Monitoring



Data

Pipelines

Catalog

Lineage

Quality

Platform

Logs

Audit

Performance

Costs

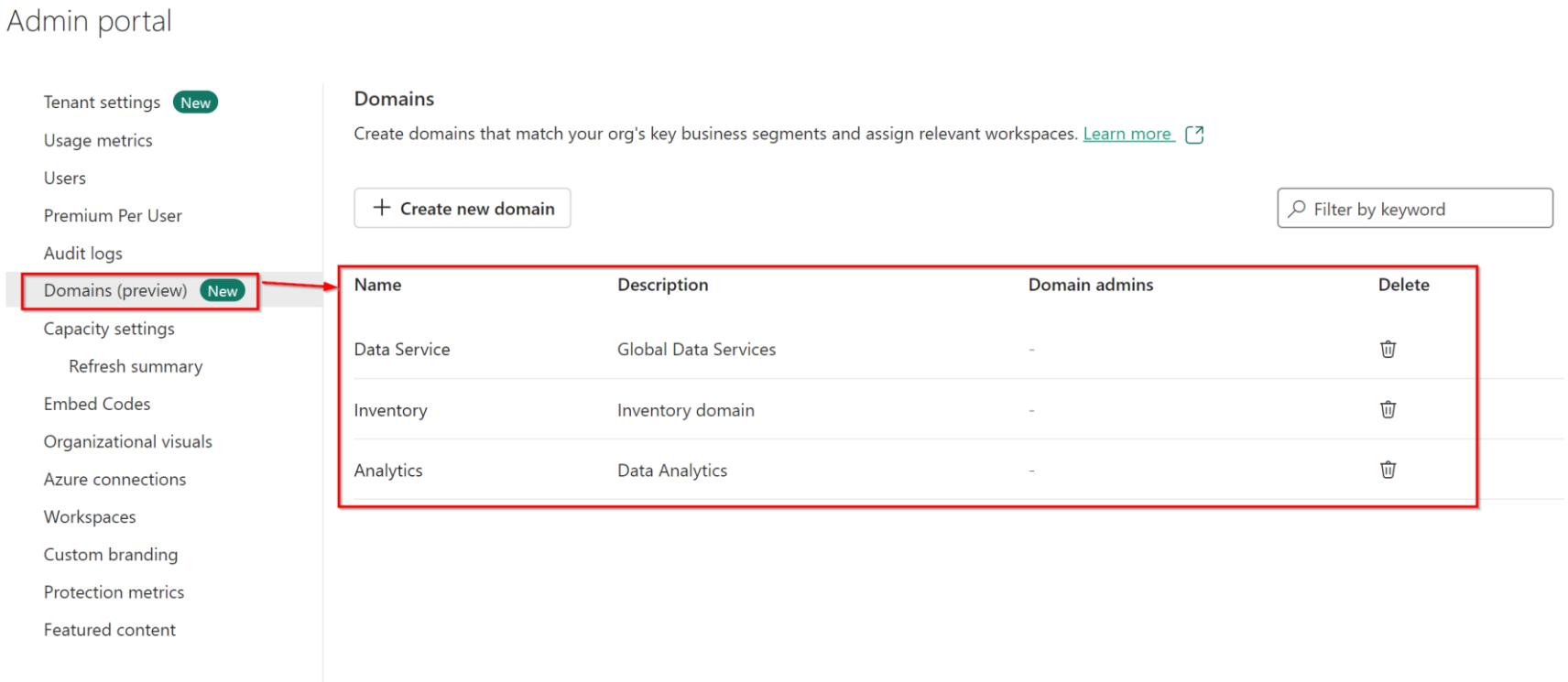
Data Governance

- Data Catalog (Purview)
- Data Sharing Policies – Data Products vs Usage Matrix
- Security and Privacy Policies
- Data Quality Guidelines
- *Separate Governance team or distributed responsibility

MS Fabric Domains - Demo

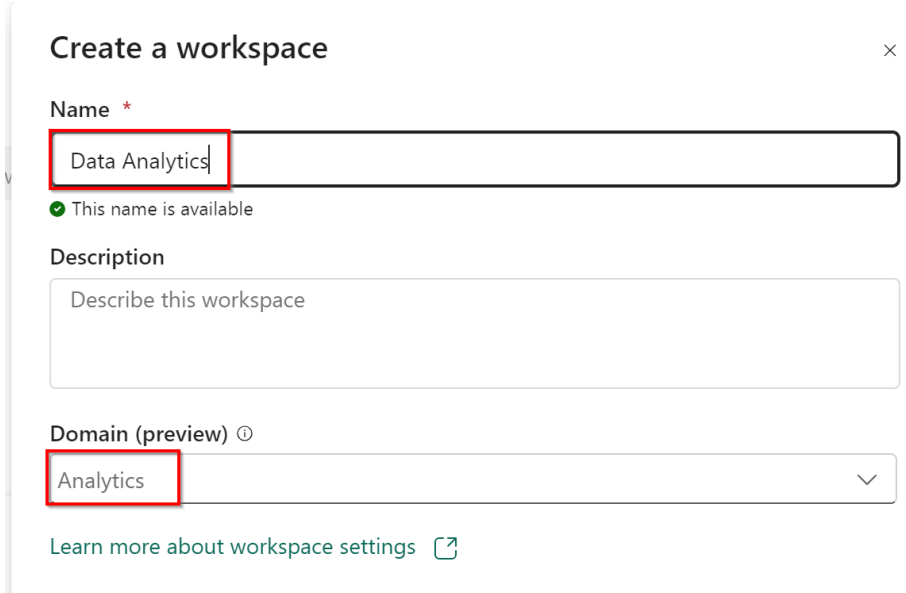
MS Fabric - Demo

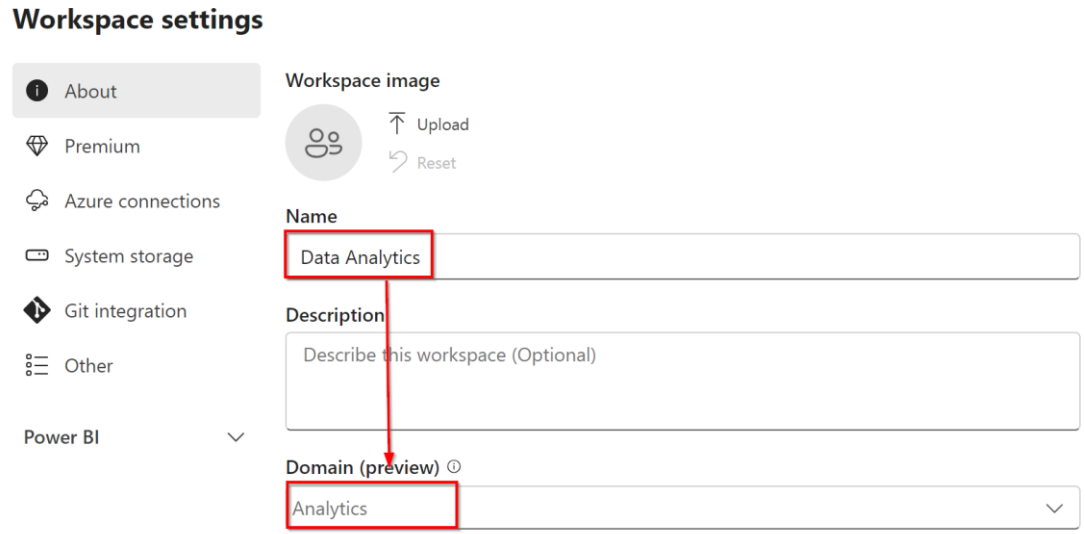
- Create domain: Setting -> Admin Portal -> Domains:



MS Fabric - Demo

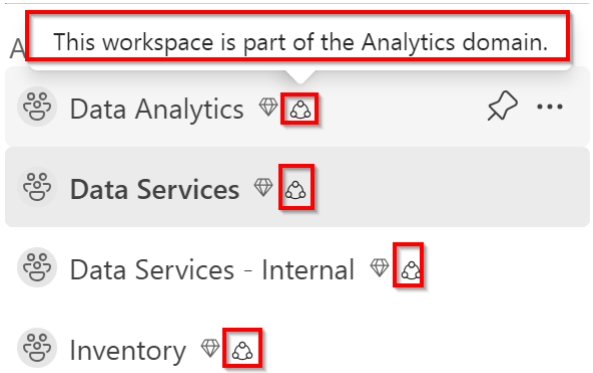
- Assign Domain to Workspace in one of the two ways:





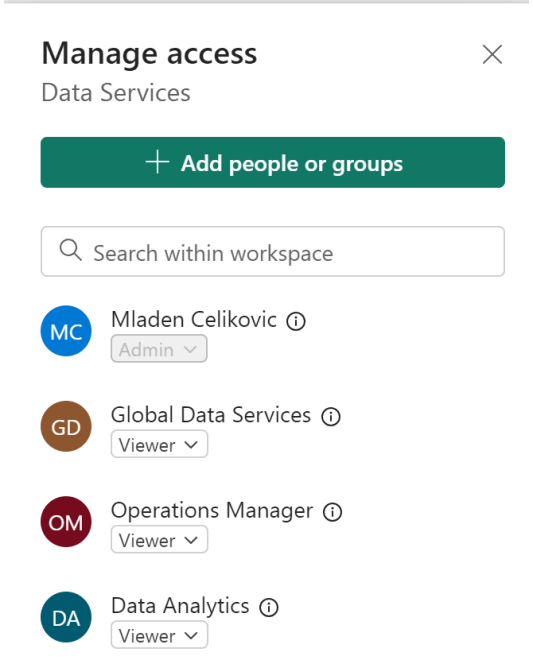
MS Fabric - Demo

- Workspaces with assigned domains:

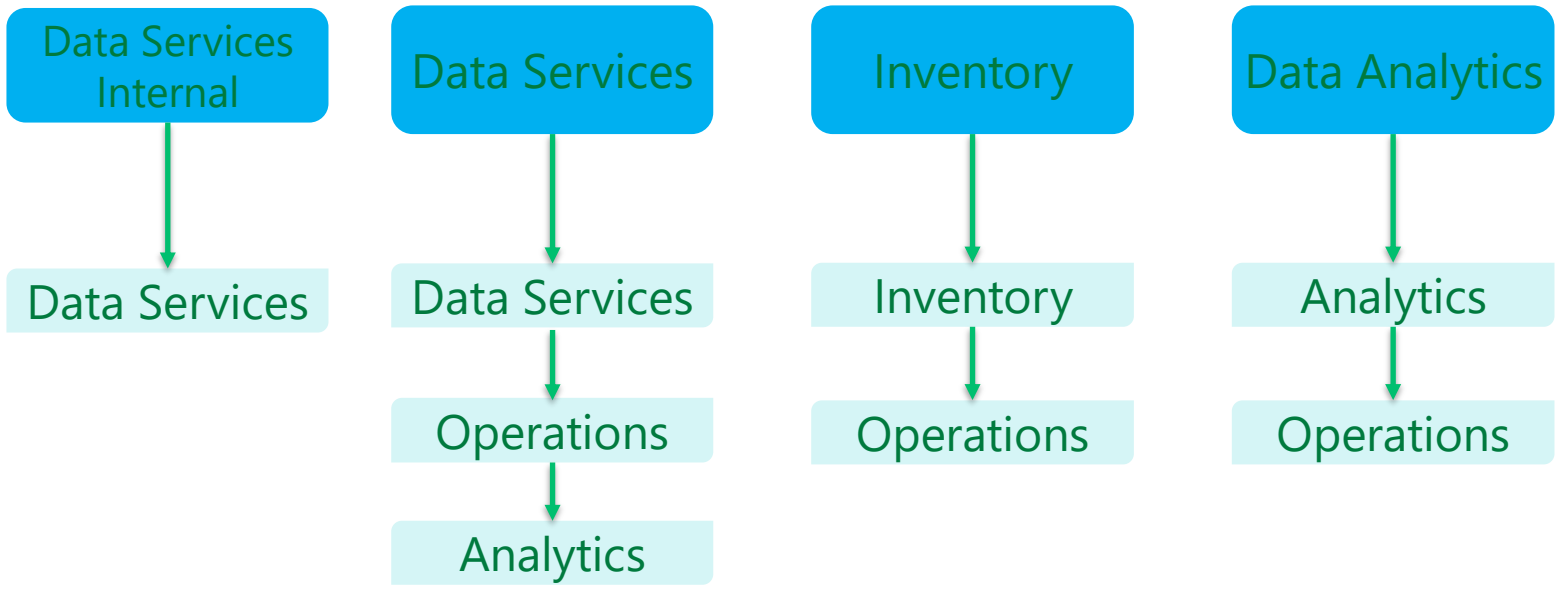


MS Fabric - Demo

Assign access to users:



Example of user access:



MS Fabric - Demo

- Spark access to Power BI – data analyst and scientist friendly
- DAX example:

```
%pip install semantic-link
```

✓ -Command executed in < 1 ms by Mladen Celikovic on 4:12:37 PM, 10/21/23

Output is hidden

```
import sempy.fabric as fab
```

✓ -Command executed in 8 sec 339 ms by Mladen Celikovic on 10:56:36 PM, 10/19/23

```
fab.evaluate_dax("Sales & Returns Sample", """
    EVALUATE Age
""")
```

✓ -Command executed in 8 sec 474 ms by Mladen Celikovic on 10:57:05 PM, 10/19/23

	Age[Age]	Age[Age Bucket]
0	15	15 to 20 years
1	16	15 to 20 years
2	17	15 to 20 years

```
fab.evaluate_dax("Sales & Returns Sample", """
    EVALUATE
        SUMMARIZECOLUMNS(
            'Product'[Category],
            'Calendar'[Year],
            'Calendar'[Month],
            "Net Sales", [Net Sales],
            "Net Sales PM", [Net Sales PM],
            "Net Sales Variance %", [Net Sales Variance %]*100
        )
""")
```

✓ -Command executed in 1 sec 723 ms by Mladen Celikovic on 11:22:40 PM, 10/19/23

	Product[Category]	Calendar[Year]	Calendar[Month]	[Net Sales]	[Net Sales PM]	[Net Sales Variance %]
0	Office 365	2019	1	124652	<NA>	-100.0
1	Office 365	2019	2	94232	124652	-24.403941
2	Office 365	2019	3	40415	94232	-57.111172
3	Office 365	2019	4	67036	40415	65.869108
4	Office 365	2019	5	145062	67036	116.394176
5	Office 365	2019	6	267280	145062	84.252251

MS Fabric - Demo

- There is also a SQL way:

```
1   spark.conf.set("spark.sql.catalog.pbi", "com.microsoft.azure.synapse.ml.powerbi.PowerBICatalog")
```

✓ -Command executed in 348 ms by Mladen Celikovic on 11:24:48 PM, 10/19/23

```
1   display(spark.sql("SHOW TABLES FROM pbi"))
```

✓ -Command executed in 2 sec 542 ms by Mladen Celikovic on 11:26:01 PM, 10/19/23

Table | Chart | Export results

Index	namespace	tableName	isTemporary
1	`Sales & Returns Sample`	Design DAX	false
2	`Sales & Returns Sample`	Sales	false
3	`Sales & Returns Sample`	DateTableTemplate_81b13e4d...	false
4	`Sales & Returns Sample`	_Metrics	false
5	`Sales & Returns Sample`	Tooltip Info	false

```
1   df = spark.sql("""
2       SELECT
3           `Product[Category]`,
4           `Calendar[Year]`,
5           `Calendar[Month]`,
6           `Net Sales`,
7           `Net Sales PM`,
8           `Net Sales Variance %`*100 AS `Net Sales Var %`
9       FROM pbi.`Sales & Returns Sample`.`_Metrics`
10  """)
11
12  display(df)
```

✓ -Command executed in 978 ms by Mladen Celikovic on 11:36:16 PM, 10/19/23 PySpark (Python)

Table | Chart | Export results

Index	Product[Category]	Calendar[Year]	Calendar[Month]	Net Sales	Net Sales PM
5	Power Platform	201902	2	46537	70324
6	XBOX	201902	2	47040	51760
7	Office 365	201903	3	40415	94232
8	Power Platform	201903	3	19574	46537
9	XBOX	201903	3	19400	47040

Resources

- Data mesh: https://martinfowler.com/articles/data-mesh-principles.html
- MS Adoption Framework: https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/architectures/data-domains
- Synapse Data Mesh: https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/data-mesh-a-perspective-on-using-azure-synapse-analytics-to/ba-p/3644657
- MS Fabric domains: https://learn.microsoft.com/en-us/fabric/governance/domains
- One lake shortcuts: https://learn.microsoft.com/en-us/fabric/onelake/onelake-shortcuts
- Power BI read by Spark: https://learn.microsoft.com/en-us/fabric/data-science/read-write-power-bi

Thank You

Toronto Data Professionals Community (TDPC)

SQL Saturday (#1064)