



# MEDALLION FOR DATA MESH

EXPLORING WORKSPACE, CAPACITY,  
AND DOMAIN DESIGN

Sam Debruyne  
DataMinds Connect  
October 2024

Sam Debruyne - DataMinds Connect 2024



# Thank you, partners





# Who am I?

## Sam Debruyn

 Heist-op-den-Berg, BE

 Consultant / Data & Cloud Architect

 5 years in data

 10 years in software / architecture / cloud

 Fabric, Azure, modern data stack





# What we'll talk about



Data Mesh

Medallion

Workspaces & Capacities

Medallion & Data Mesh on Fabric

Capacity design for scalability

Access control & Domains

Sarah Debruyn - DataMinds Connect 2024



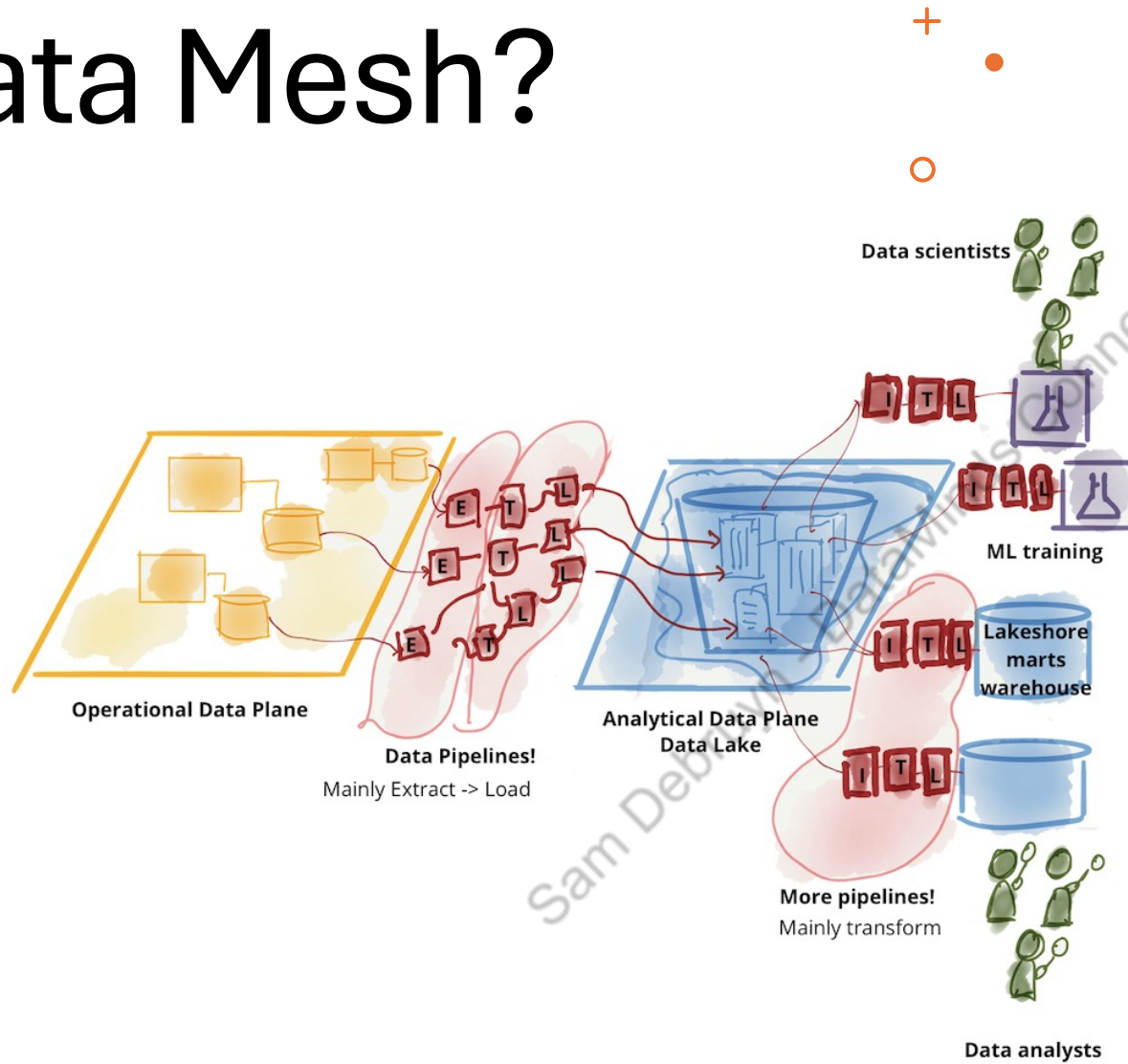


# Data mesh

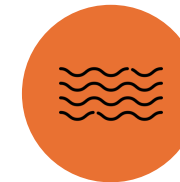
Sam Debruyn - DataMinds Connect 2024



# Data Mesh?

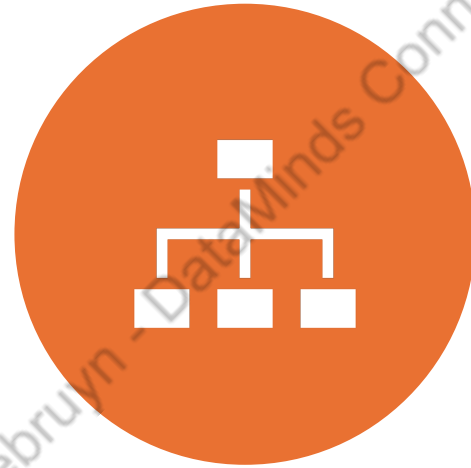


First introduced in 2019 by Zhamak Deghani at ThoughtWorks



Overcoming challenges of the monolithic data lake

# The 4 Principles of the Data Mesh

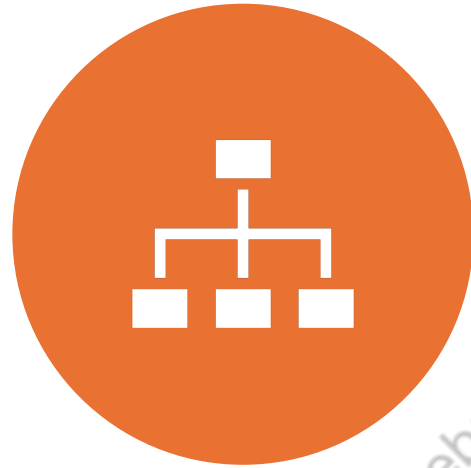


DOMAIN-ORIENTED  
DECENTRALIZED DATA OWNERSHIP

Sam Debruyn - DataMinds Connect 2024



# The 4 Principles of the Data Mesh



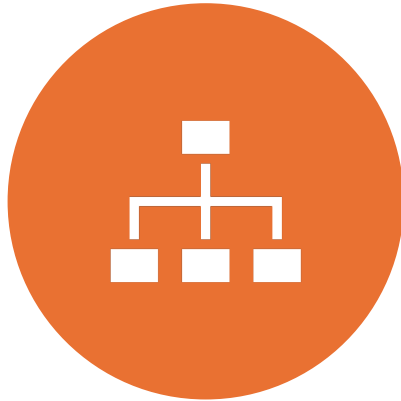
DOMAIN-ORIENTED  
DECENTRALIZED DATA OWNERSHIP



DATA PRODUCT THINKING

Sam Debruyn - DataMinds Connect 2024

# The 4 Principles of the Data Mesh



DOMAIN-ORIENTED  
DECENTRALIZED DATA OWNERSHIP



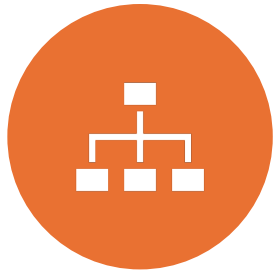
DATA PRODUCT THINKING



SELF-SERVICE ANALYTICS

Sam Debruyne - DataMinds Connect 2024

# The 4 Principles of the Data Mesh



DOMAIN-ORIENTED  
DECENTRALIZED DATA  
OWNERSHIP



DATA PRODUCT  
THINKING



SELF-SERVICE  
ANALYTICS



FEDERATED  
GOVERNANCE

Sam Debruyn - DataMinds Connect 2024



O'REILLY®

# Data Mesh

Delivering Data-Driven Value at Scale



Zhamak Dehghani

## More content on Data Mesh

[Microsoft Cloud Adoption Framework](#)

[Initial blog post on data mesh](#)

[Second blog post on data mesh](#)

[Free PDF copy of the Data Mesh book](#)  
(thanks to Starburst)

- +
  - • 

# Medallion layers

Sam Debruyn - DataMinds Connect 2024





# The 3 Layers of the Medallion Architecture

## **Raw/bronze**

**Purpose:** all data in its original form without transformations or quality checks. Source of truth for historical data and reprocessing if needed.





# The 3 Layers of the Medallion Architecture

## **Cleansed/silver**

**Purpose:** ensure consistency and quality. Data is cleansed, transformed, and enriched.



# The 3 Layers of the Medallion Architecture

## **Curated/gold**

**Purpose:** high-quality data supporting business reporting, advanced analytics. Pre-aggregated and tailored to analytical needs.

central data team

connecting new data sources

data cleansing

data governance

domain-oriented team

data modelling

data quality

# Data platform architectural design questions

Which data mesh design principles should be applied at which level?

The answer = different for every organization

What are the key elements guiding your decision?

- Historical reasons
- Data maturity
- Expectations from every department
- Plans for upskilling and/or upstaffing
- Tools which act as enablers

You often start from the data products in the gold layer and work your way back.

So... how do we  
bring these  
concepts together?

Sam Debruyn - DataMinds Connect 2024

Data mesh and medallion with Fabric

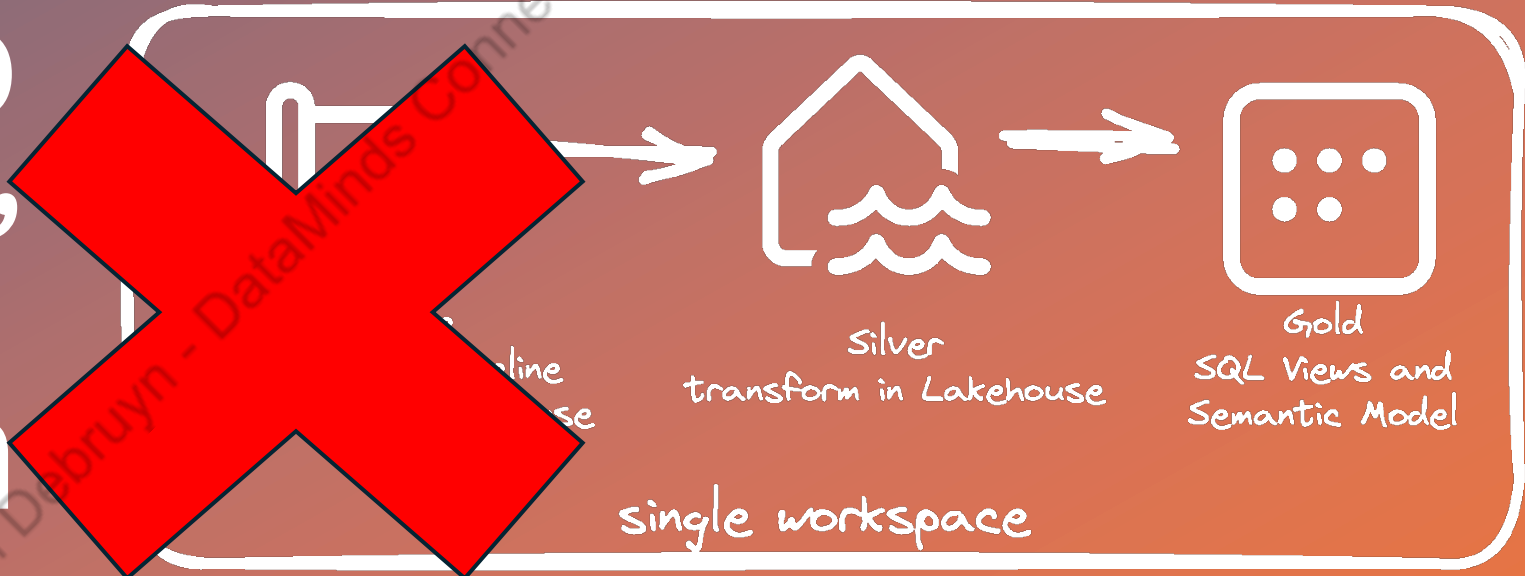


# Let's look at Workspace design for medallion and data mesh in Microsoft Fabric

Sam Debruyn - DataMinds Connect 2024



# The “Hello World” version



Sam Debruyn - DataMinds Connect 2024



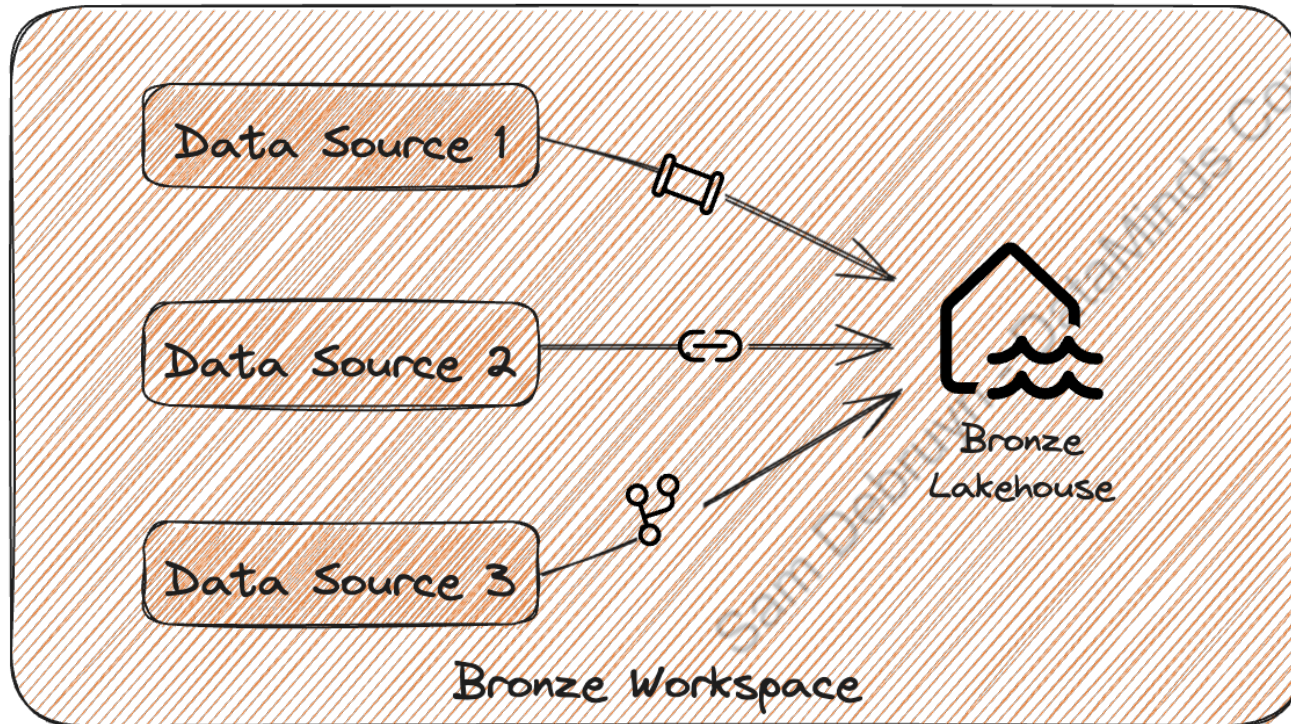


# Bronze

Sam Debruyn - DataMinds Connect 2024



# Bronze



Ingestion is a complex task

Data sources are ingested into a Bronze Lakehouse in their raw/source format

No business knowledge required

Managed by central data team with specialized data engineers



# Shortcuts: Fabric cornerstones

Sam Debruyn - DataMinds.Com 2024

Virtual / logical link to a dataset in Delta Lake or Iceberg format on

- Azure Data Lake Storage Gen2
- AWS S3
- Google Cloud Storage
- Fabric OneLake
- ...

Becomes a “native” table in a Fabric Lakehouse

💡 **NEW:** Schema Shortcuts – link a folder with multiple datasets as a schema in a Lakehouse

• 💡 **HINT:** create Shortcuts using the Fabric APIs

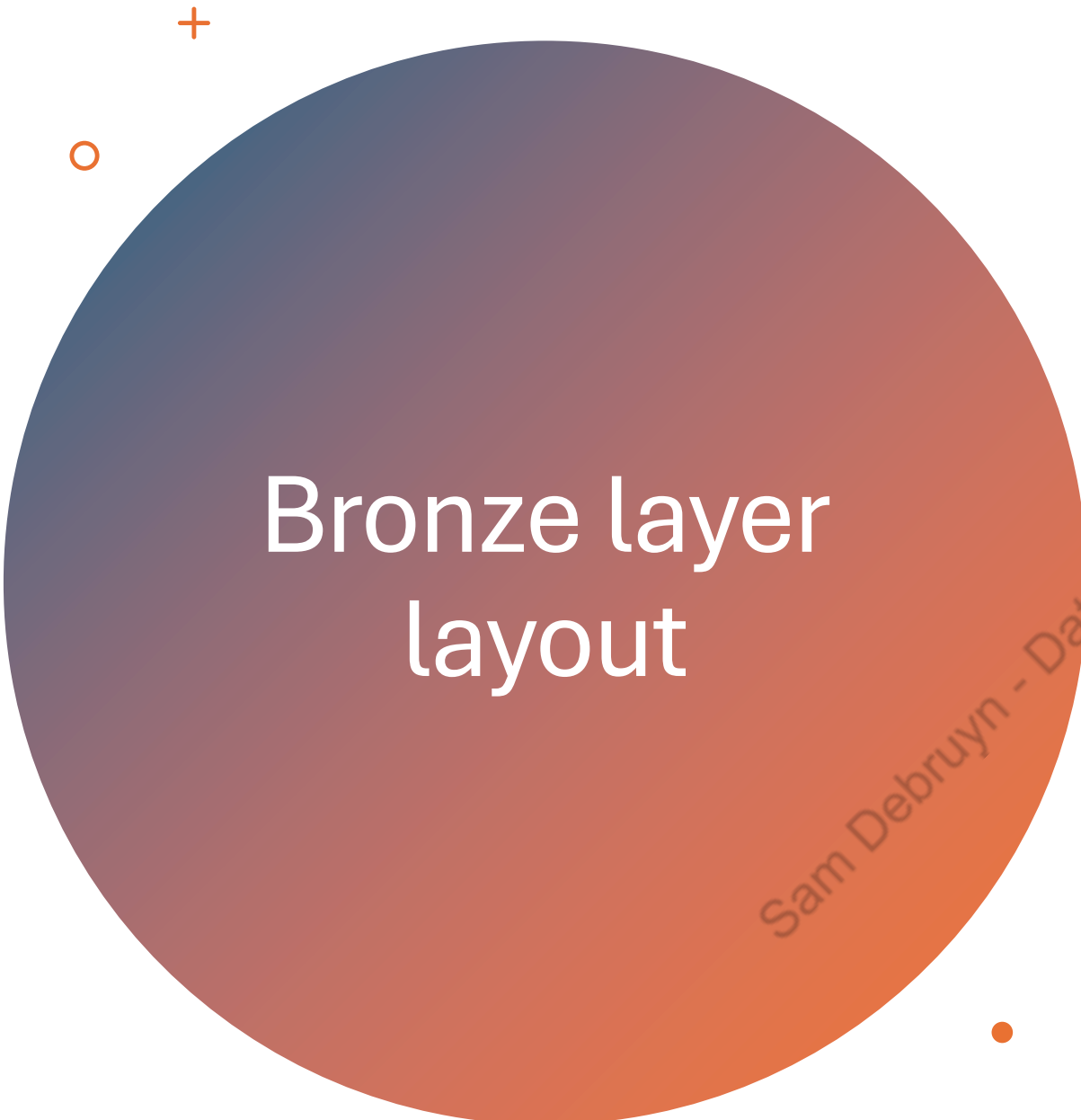
+

•

○

# Other ways to ingest data into Bronze

- Data Factory / Copy Activity
- **Copy Job**
- **Database Mirroring**
- Dataflow Gen2
- Notebooks / Spark Jobs
- ADLS APIs
- OneLake File Explorer
- DWH SQL APIs: COPY INTO / OPENROWSET



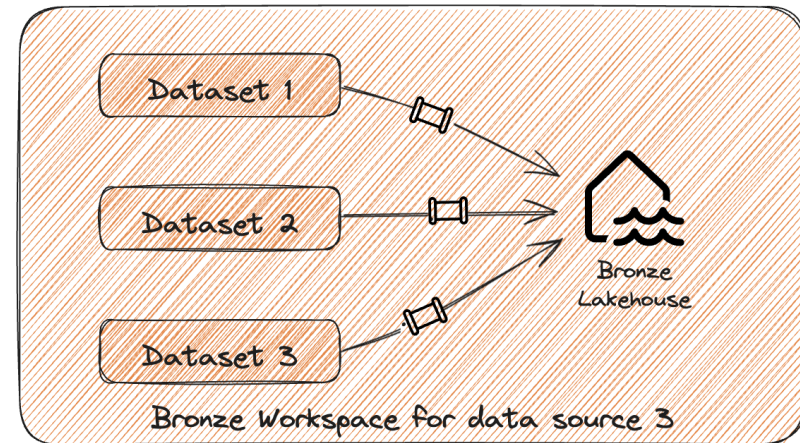
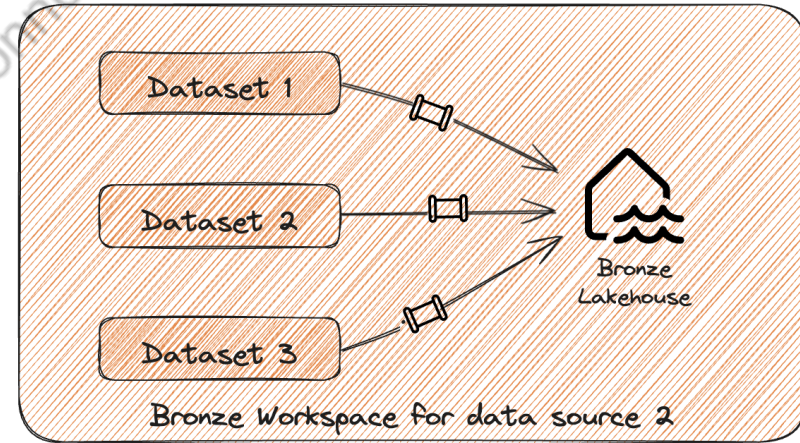
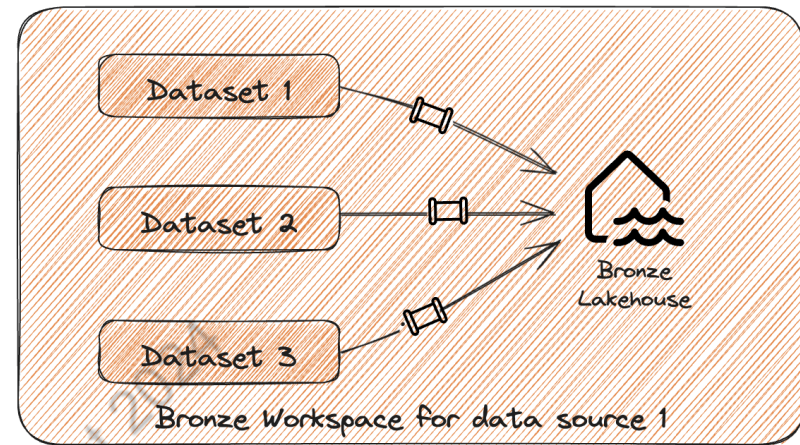
# Bronze layer layout

Multiple options, depending on data platform size & complexity:

- Single Workspace, Single Lakehouse, 1 schema per source system
- Single Workspace, 1 Lakehouse per source system
- 1 Workspace per source system



# Bronze layer layout





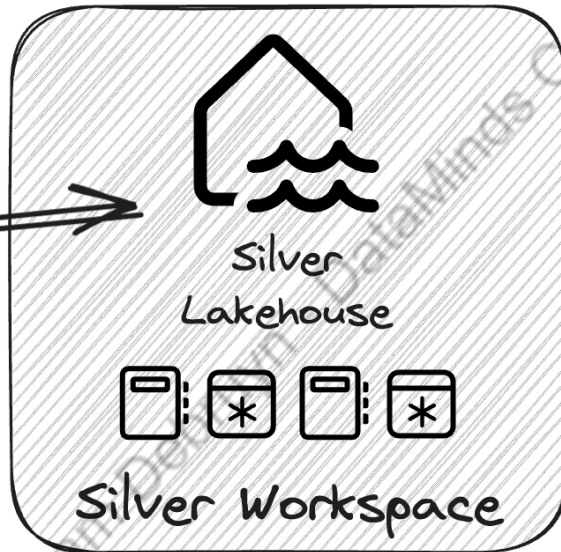


# Silver

Sam Debruyn - DataMinds Connect 2024




# Silver



Data is linked from Bronze Workspace using Shortcuts

Managed by central data team with specialized data engineers

Common tools: Spark Jobs, Notebooks, dbt, ...

 **NEW/HINT:** use Schema Shortcuts to not have to create a separate Shortcut per table

# Silver layer layout

Some teams tend to prefer data vault here



Other approaches:

- Replicate layout from bronze
- Wide tables
- ...



# Gold

Sam Debruyn - DataMinds Connect 2024



# Gold



Data is linked from Silver  
Workspace using Shortcuts

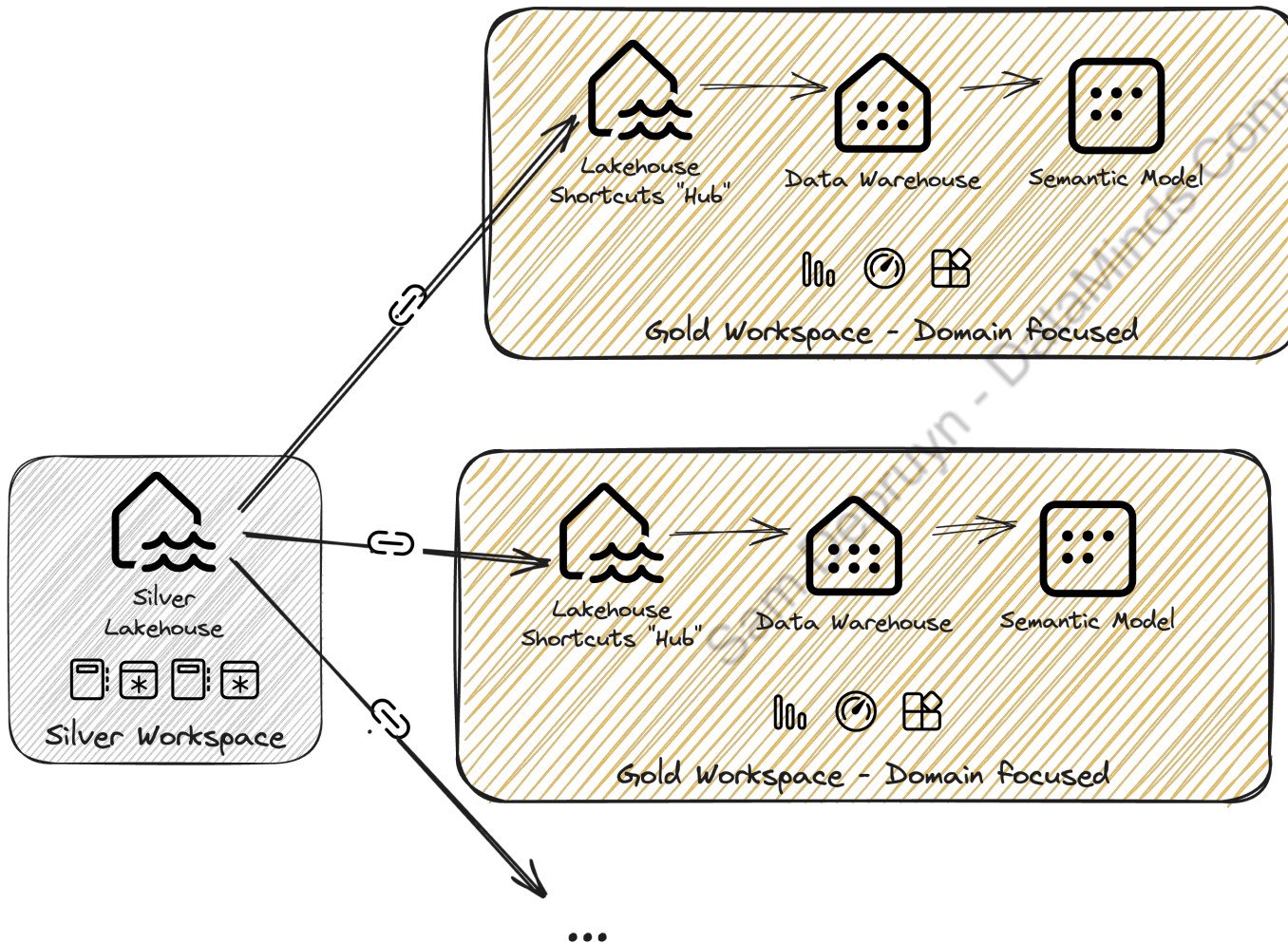
Data modelling in SQL in a Data  
Warehouse

Data is separated by domain

Workspace also serves as access  
control boundary

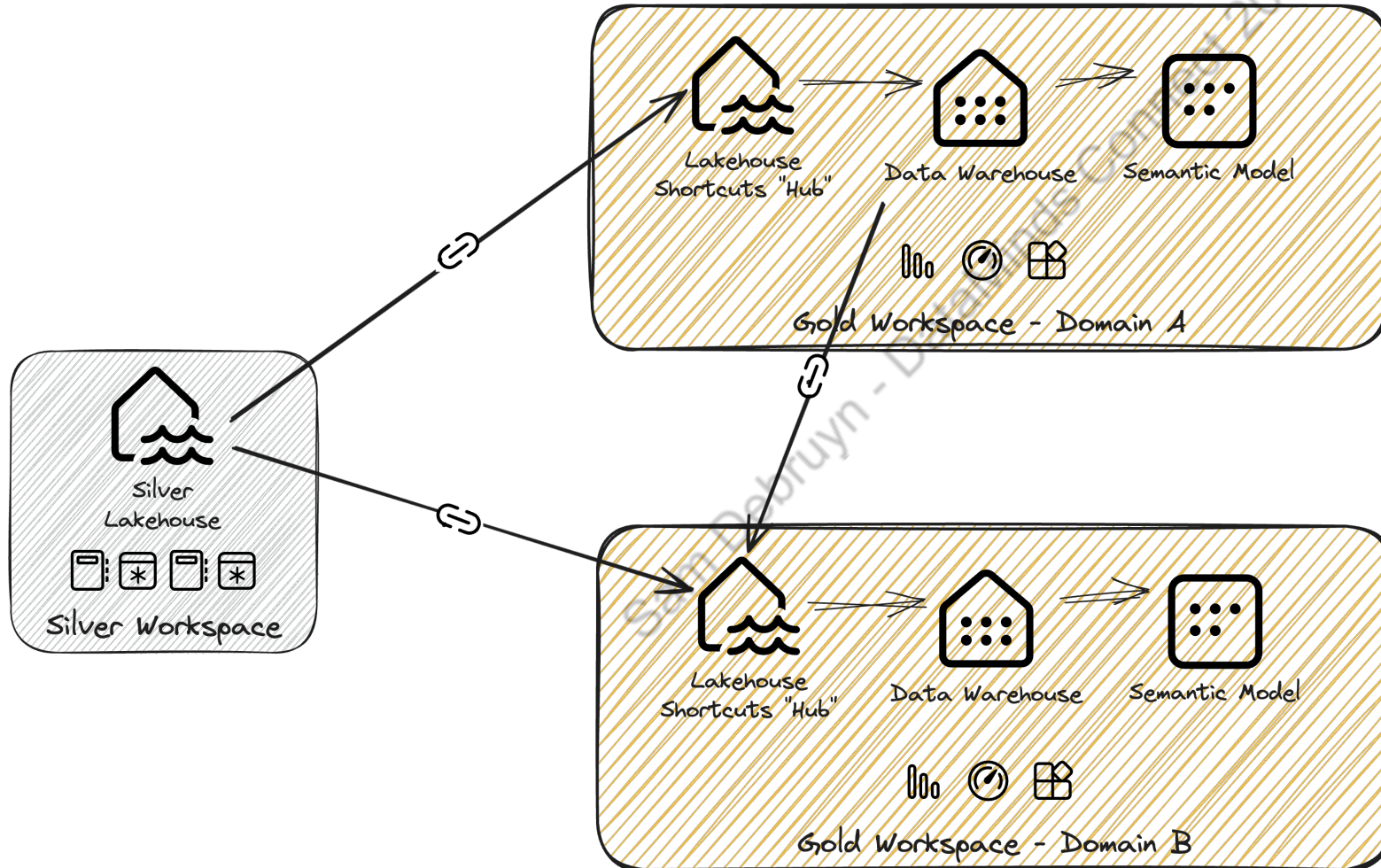
Decentralized data domain focused  
teams

Core skill: analytics engineering





# Gold



Shortcuts go to Lakehouses, from there you can use the SQL Analytics Endpoint version in the Data Warehouse.

If needed, you can still link data from one Gold Workspace to the other using Shortcuts.



# Gold

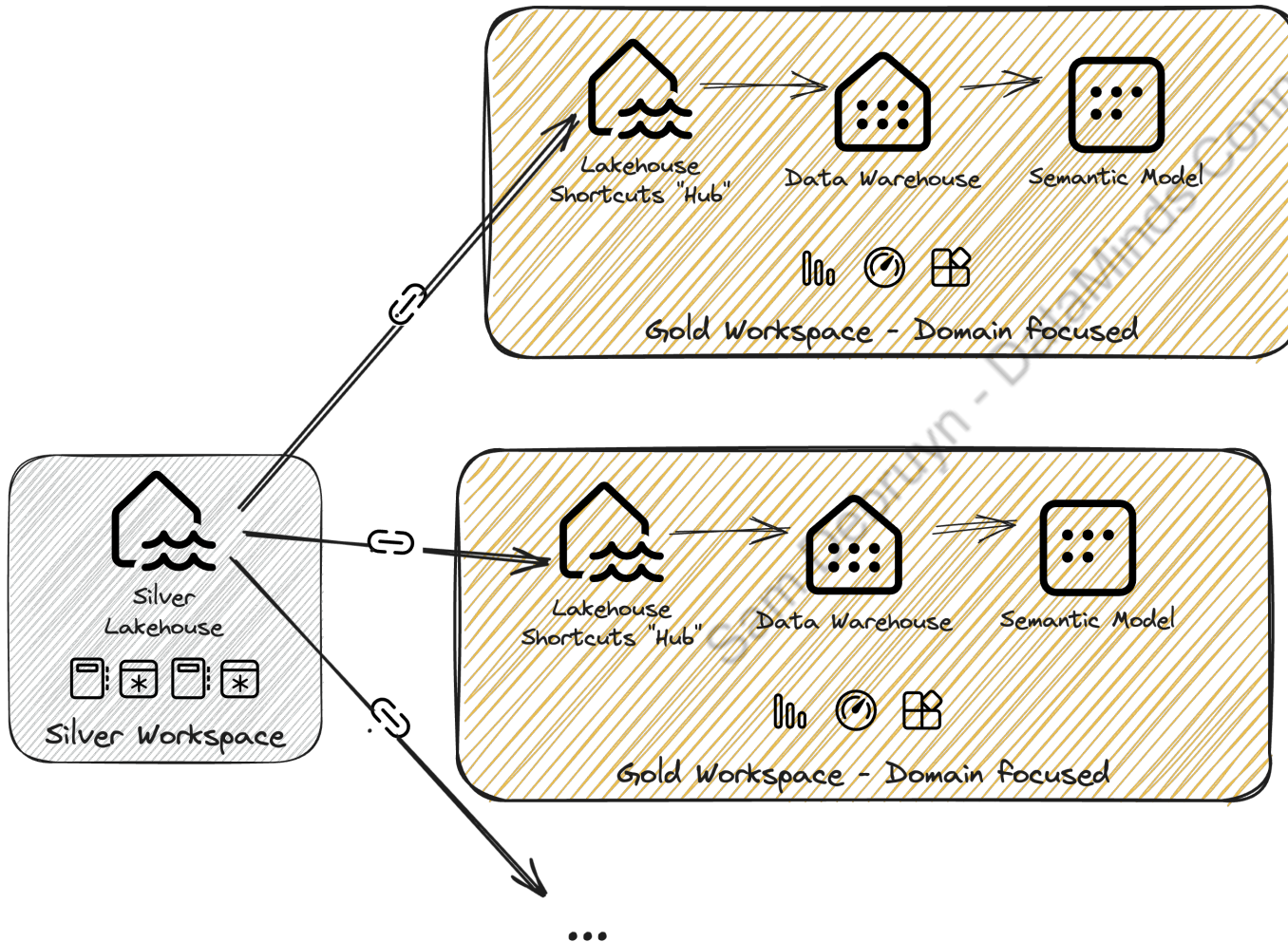


## Why the split by domain?

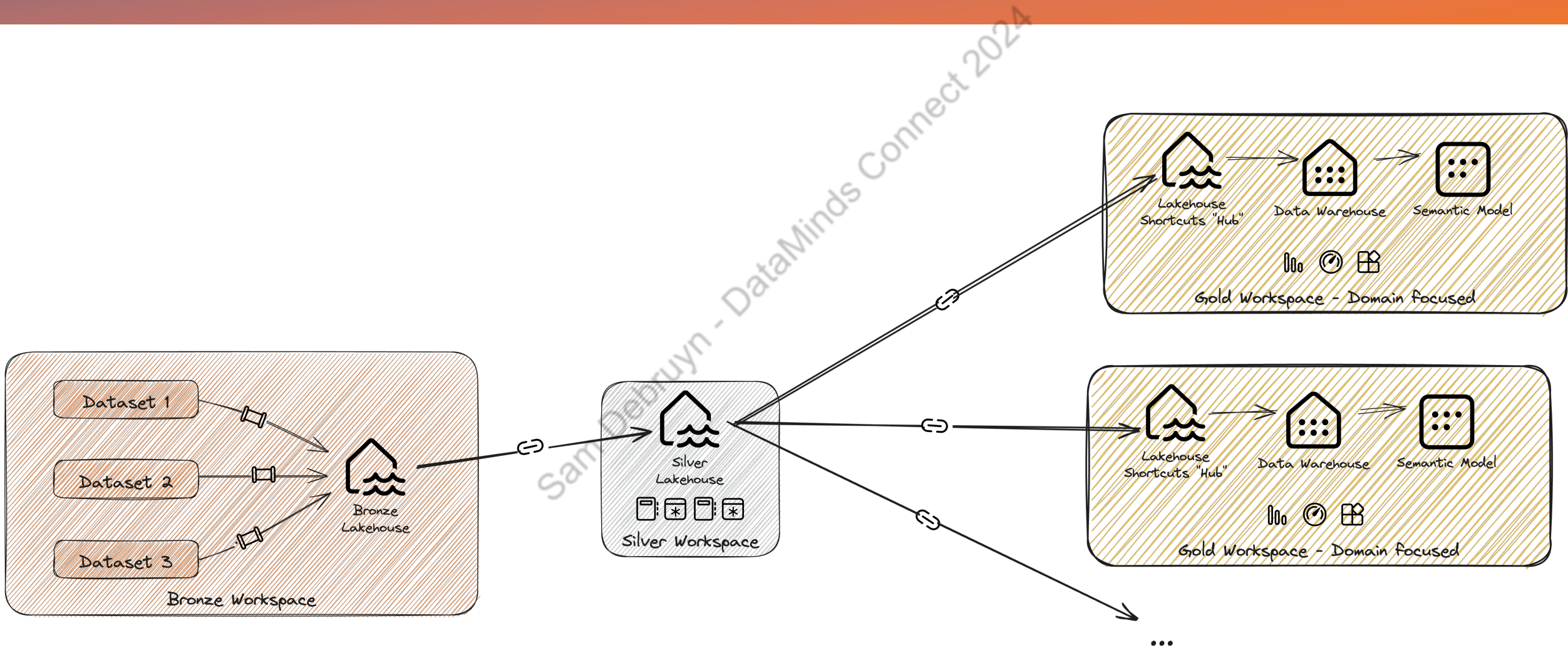
Clearly indicates the ownership of the data. E.g. HR team owns data on staffing, sales team owns data on sales numbers, ...

Workspace owners are both responsible and accountable for the data they produce

Creates a producer-consumer relationship



# Overview: entire platform (example)





# Did I invent this?



No, this is also how Microsoft recommends it

## Deployment model

To implement medallion architecture in Fabric, you can either use lakehouses (one for each zone), a data warehouse, or combination of both. Your decision should be based on your preference and the expertise of your team. Keep in mind that Fabric provides you with flexibility: You can use different analytic engines that work on the one copy of your data in OneLake.

Here are two patterns to consider.

- **Pattern 1:** Create each zone as a lakehouse. In this case, business users access data by using the SQL analytics endpoint.
- **Pattern 2:** Create the bronze and silver zones as lakehouses, and the gold zone as data warehouse. In this case, business users access data by using the data warehouse endpoint.

While you can create all lakehouses in a single Fabric workspace, we recommend that you create each lakehouse in its own, separate **Fabric workspace**. This approach provides you with more control and better governance at the zone level.

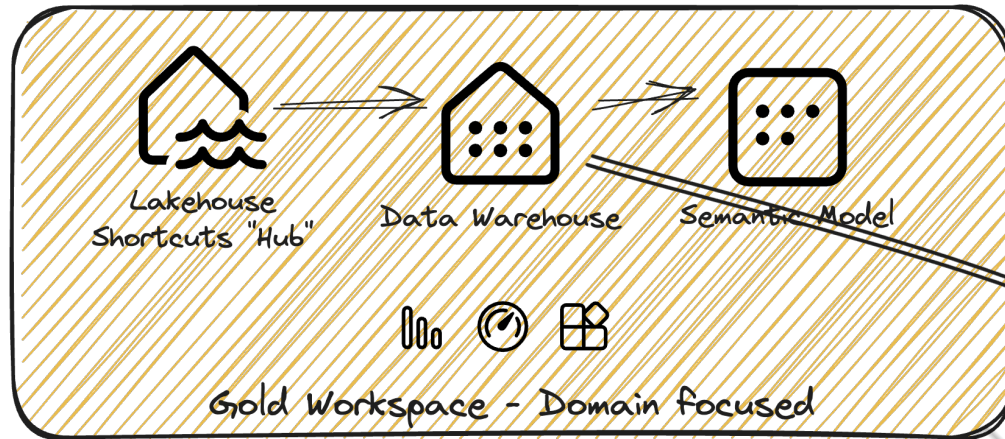


Easy to  
extend

This can be extended for

- Real-time data
- Multiple environments / shared environments
- Data sharing
- AI
- ...

# Platinum



What about Advanced Analytics? Or specific use-cases not fitting into regular Gold Workspaces?

Hyper-specialized Workspaces can be conceived similarly to Azure resource groups / "project folders"

Sam Debruyne - DataMinds Connect 2024



# Workspaces & Capacities

Sam Debruyn - DataMinds Compete 2024





+

•

○

# Why should you create separate Workspaces?

## **Workspace configuration**

Some settings on the Workspace level might be different for different workloads.

→ Different workloads might require different configurations

Sam Debruyn - DataMind Connect 2024

+

•

○

Why should  
you create  
separate  
Workspaces?

## **Access control**

- Admin
- Member
- Contributor
- Viewer

Sam Debruyn - DataMinds.com - 2024

+

•

○

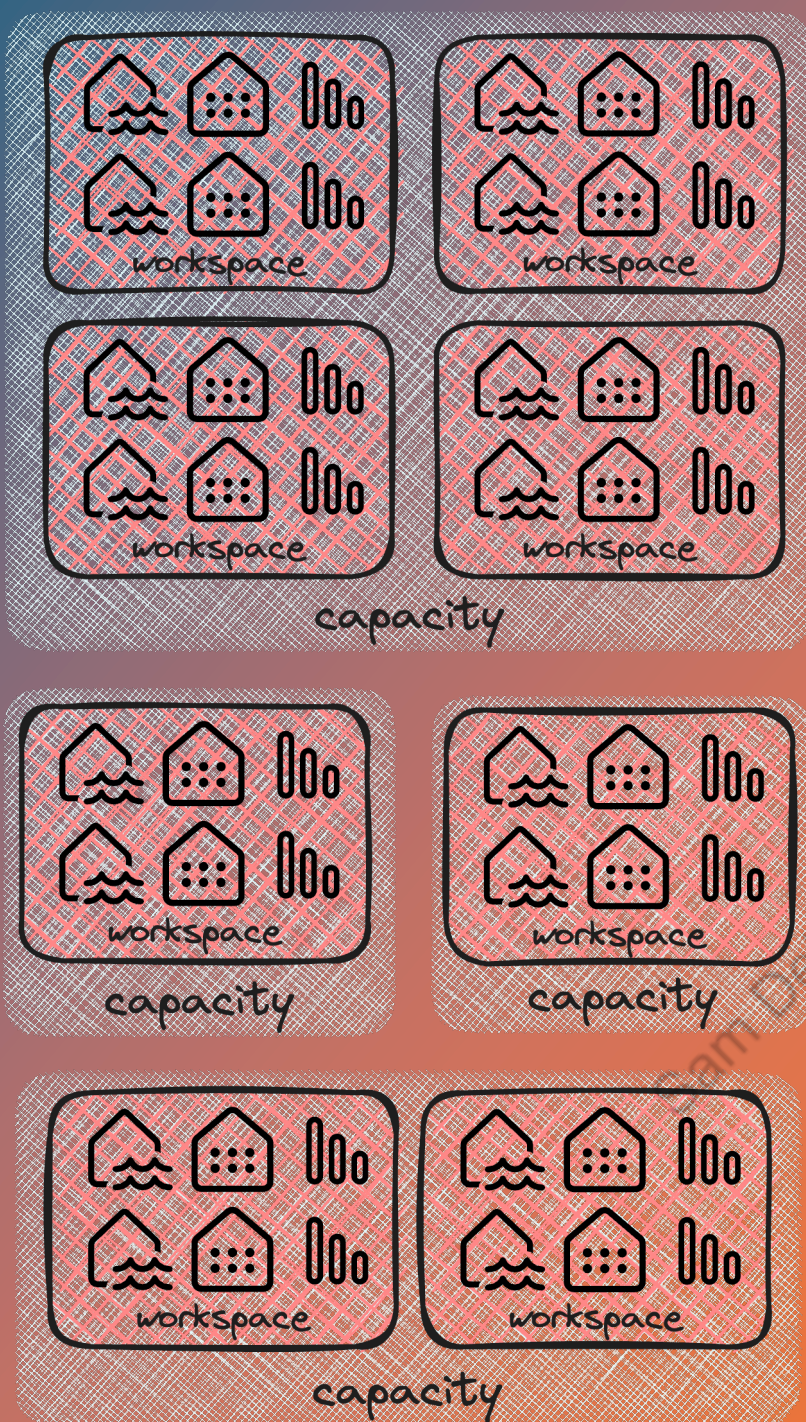
Why should  
you create  
separate  
Workspaces?

**Capacity  
Management**

Sam Debruyn - DataMinds Connect 2024



# Fabric concepts: Workspaces & Capacities



## Capacity

- pool of Capacity Units
- matches a certain amount of compute power
- to be spread amongst one or more Workspaces

## Workspace

- logical grouping of items
- Lakehouses, Warehouses, Reports, KQL, ...
- possible access control boundary

---

# Capacities

Used for everything which should be “billed” in Fabric

SKU indicates the amount of available Capacity Units

F2: 2 Capacity Units (CU's)

F8: 8 Capacity Units (CU's)



Saini Debajyoti DataMind Connect 2024



---

# Capacity SKUs



Actual billing is done in Capacity Unit Seconds (CUs)

Note difference between **CUs** (Capacity Unit Seconds) and Capacity Units (**CU's**)

Amount of available CUs is SKU x seconds.

F2: 2 CU, base budget per second is 2 CUs

F8: 8 CU, base budget per second is 8 CUs

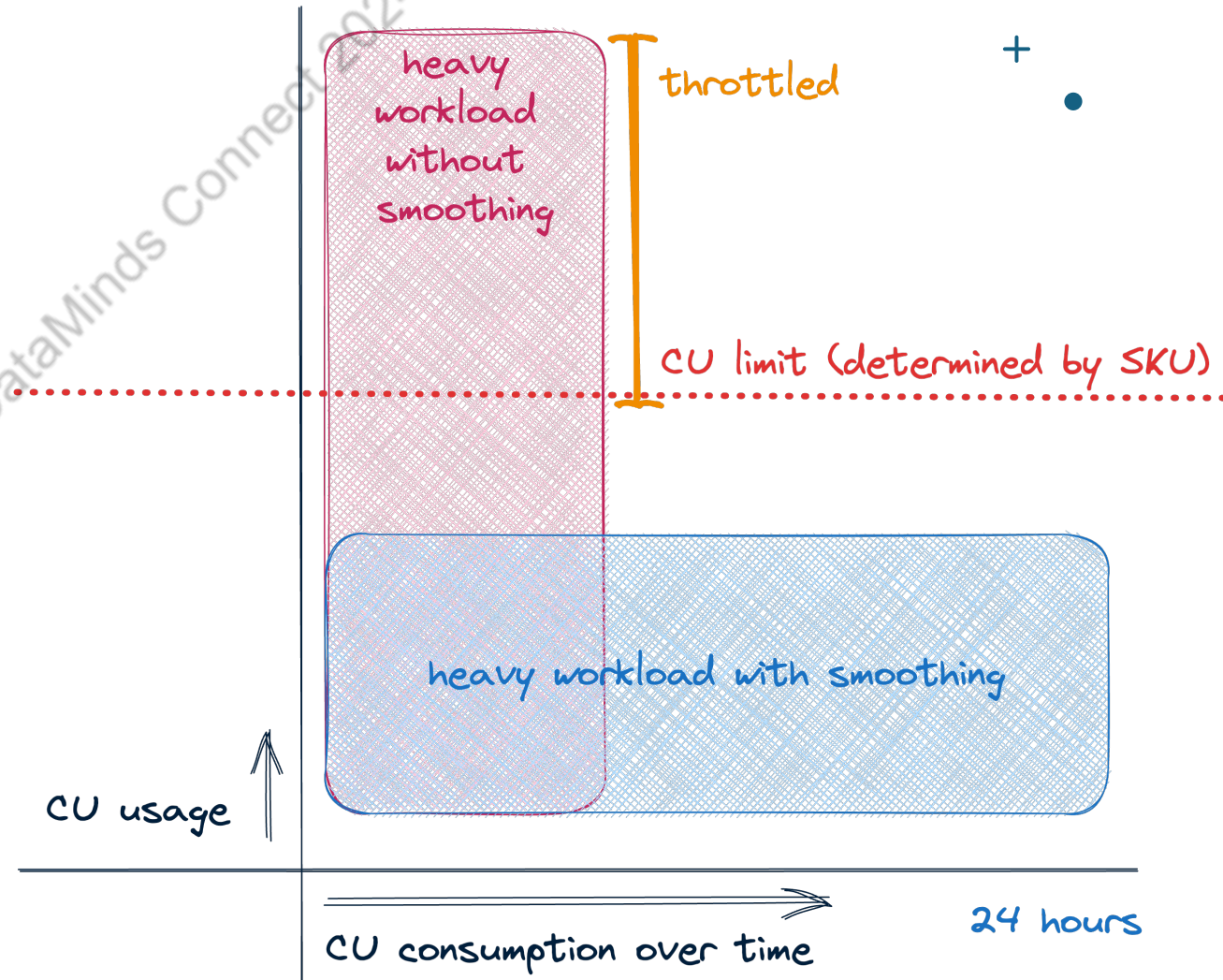
# Bursting & smoothing

Workload billing is spread out over time

Interactive workloads: 5 to 60 minutes (e.g. Power BI)

Background workloads: up to 24 hours (e.g. Spark Job)


Most operations are background operations



# Bursting & smoothing

+

SKU	CU's	Available CUs for interactive 10min workloads	Available CUs for background 24h workloads	Actual workload duration & consumption
F2	2	1.200	172.800	ASAP*
F4	4	2.400	345.600	ASAP*
F8	8	4.800	691.200	ASAP*
F16	16	9.600	1.382.400	ASAP*
F32	32	19.200	2.764.800	ASAP*
F64	64	38.400	5.529.600	ASAP*
F128	128	76.800	11.059.200	ASAP*
...	...	...	...	...



# Impact of SKU choice

Capacities determine feature availability

E.g. CoPilot, Power BI only F64 or higher

Capacities determine how features are available

Nodes and cores/node in Spark (2 vCores per CU – burst factor 3 | 0.25 nodes per CU)

- Compute nodes in Data Warehouse

---

# Capacity level settings

Capacities have regions

Not all features are available in every region

Availability Zones (supported regions)

Compliance requirements



Sam Debuyn DataMinds Connect 2024



---

# Throttling

+



**Overage:** CUs consumed over what was available for your operation

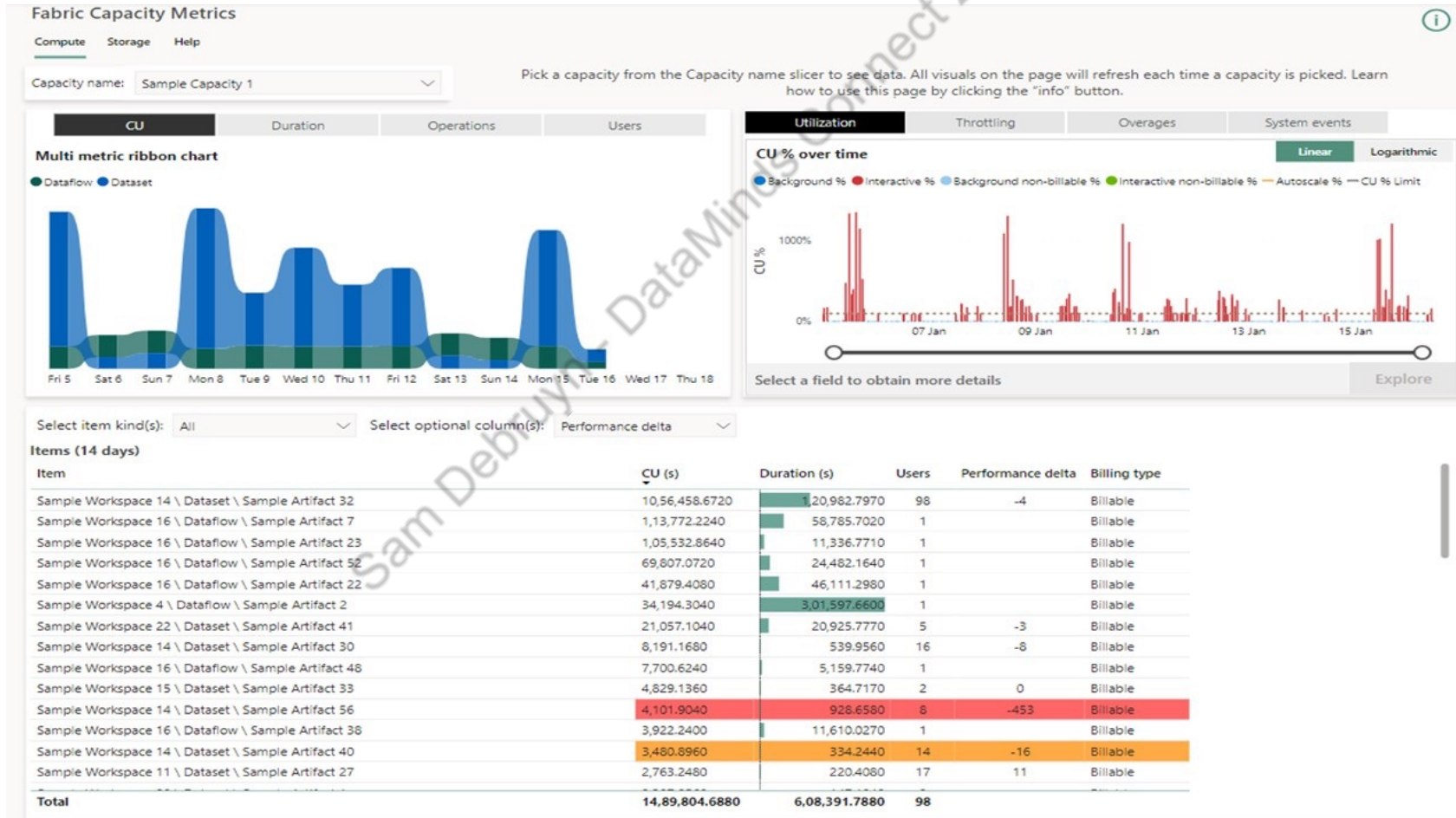
## **Throttling on interactive operations:**

- 1) When no CUs are available for the next 10 minutes → 20 seconds delay on new interactive operations (does not impact ongoing operations)
- 2) When no CUs are available for the next hour → interactive operations are denied

## **Throttling on background operations:**

When no CUs are available for the next 24 hours → all operations are denied

# Capacity Metrics App



+

•

○

Why should  
you create  
separate  
Workspaces?

## Capacity Management

Workspace → 1 Capacity

So to be able to split workloads  
over Capacities, they first have to  
be split over Workspaces

+

•

○

# Why should you create separate Workspaces?

## **3) Access Control**

Managing access on the Workspace boundary is easy.

You can still share specific subsets of data using the Data Sharing feature.

Sam Debruyn - DataMinds Connect 2024

# How access can be managed in Fabric



**Workspace level  
roles: Admin,  
Member, Contributor,  
Viewer**



Item sharing: Read,  
Edit, Share



Data sharing: Read,  
ReadData, ReadAll



OneLake RBAC  
(preview)

Note: this will probably be improved with the introduction of OneSecurity





# Domains

Sam Debruyn - DataMinds Connect 2024



The problem:  
how to get an  
overview of tens  
(hundreds?) of  
Workspaces



Sam Debruyn - DataMinds Connect 2024

+

o



# Domains

Logically grouping together data in an organization by bundling Workspaces in Domains

Domains can have Subdomains

Managed by Domain Admins and Domain Contributors

Centralize or group certified datasets

# Domains: OneLake Data Hub



## OneLake data hub

Discover, manage, and use data from across your org. [Learn more about OneLake data hub](#)

Finance Domain

### Recommended ⓘ



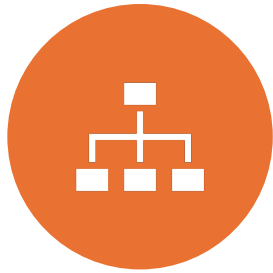
# Recap

Sam Debruyn - DataMinds Connect 2024





# RECAP: The 4 Principles of the Data Mesh



DOMAIN-ORIENTED  
DECENTRALIZED DATA  
OWNERSHIP



DATA PRODUCT  
THINKING



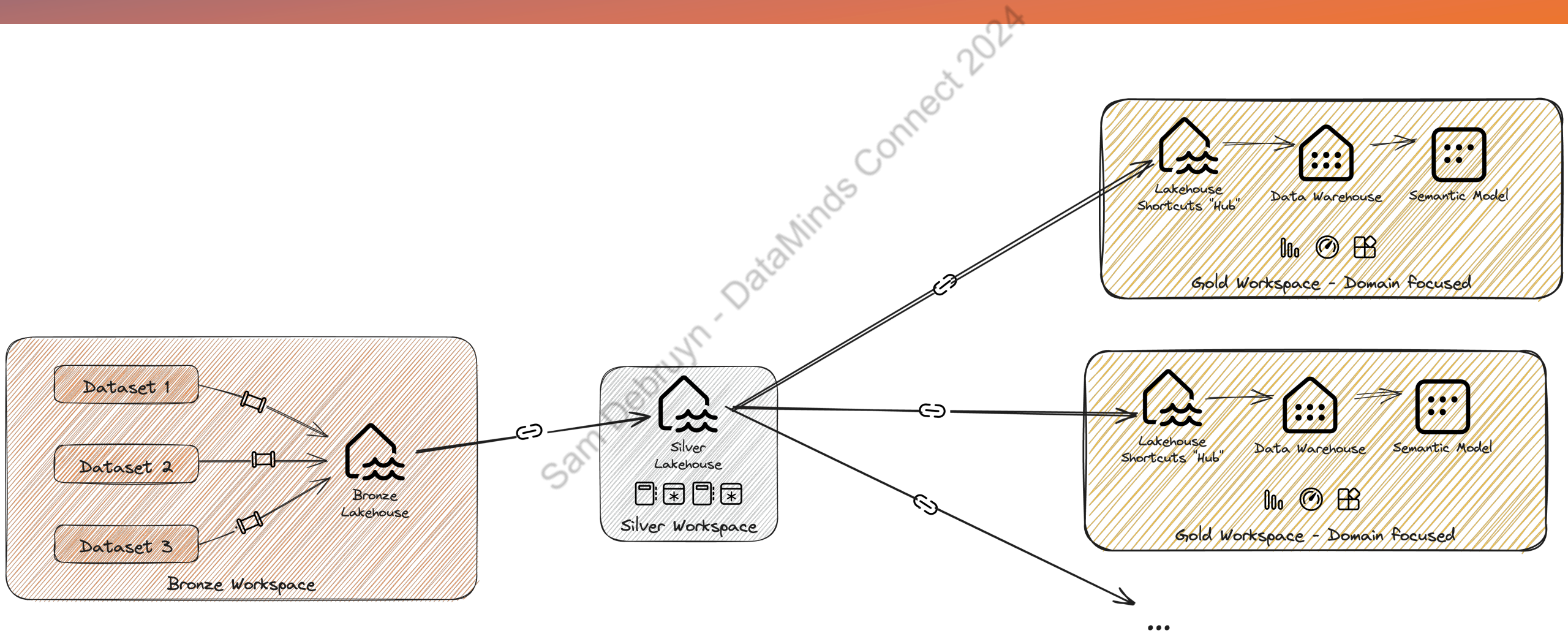
SELF-SERVICE  
ANALYTICS



FEDERATED  
GOVERNANCE

Sam Debruyn - DataMinds Connect 2024

# RECAP: Medallion layers: bronze, silver, gold



+

•

○

# Recap

Split Workspaces by  
type of workload and  
role in the data fabric

Single Capacities are  
good for trials, but we  
should avoid them for  
actual  
implementations

Access control can be  
complex, start by  
managing access on  
the Workspace level

Bundle Workspaces in  
Domains

Sam Debruyn - DataMind Project 2024





# Sam's 5 golden rules for Workspace & Capacity design in Fabric

**DO NOT** mix different layers of the medallion architecture in a single Workspace.

**DO NOT** mix data from different domains in the same Gold Workspace.

**DO** assign every Workspace 2 things: a Capacity\* and a Domain.

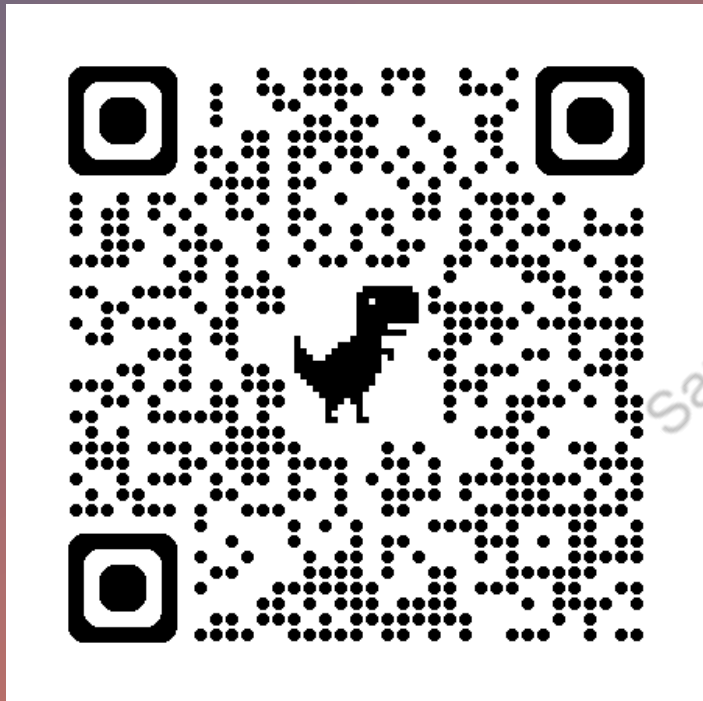
**DO** split Workspaces and their linked Capacities by workload

- Ingestion
- Processing/transformations
- Ad-hoc exploration & development
- Consumption

**DO** build for future extensibility, there is no known valid limit on the amount of Workspaces.

\*: Power BI Pro / Premium Per User Workspaces excluded

# Slides



Slides available in  
the DataMinds  
Connect app or at

[https://debruyn.dev/  
dmc24](https://debruyn.dev/dmc24)

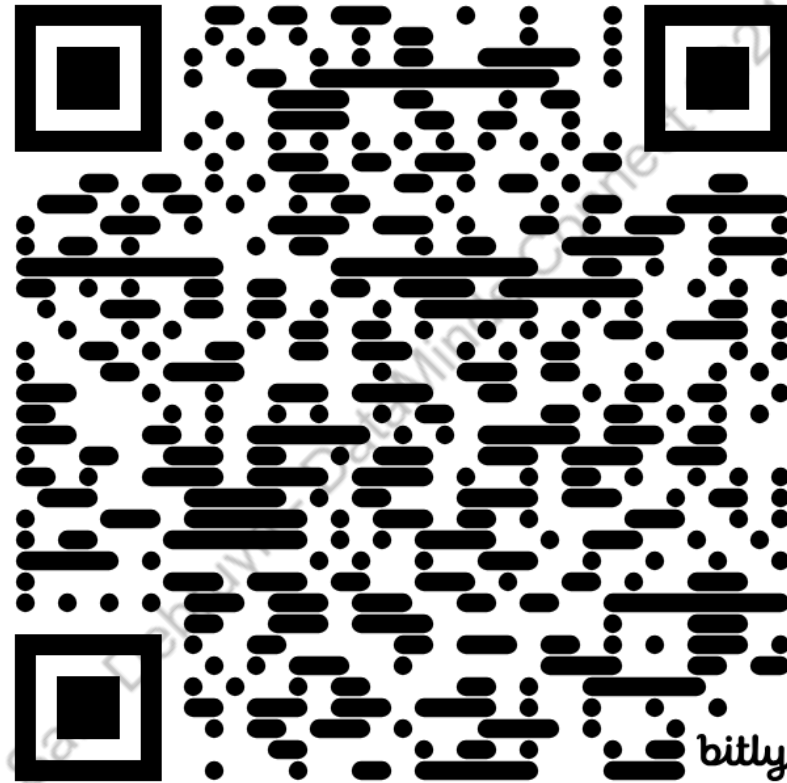




# Thank you, partners



# Session Feedback



[https://bit.ly/dMC2024\\_SessionFeedback](https://bit.ly/dMC2024_SessionFeedback)

# Questions?



Sam Debruyn - DataMinds Connect 2024

sam@debruyn.dev

<https://debruyn.dev>

