MEDALLION FOR DATA MESH

EXPLORING WORKSPACE, CAPACITY, AND DOMAIN DESIGN

Sam Debruyn

Data Bash

October 2024

Who am I?

Sam Debruyn

📍 Heist-op-den-Berg, BE

💼 Consultant / Data & Cloud Architect

5️⃣ years in data

🔟 years in software / architecture / cloud

🫶 Fabric, Azure, modern data stack

MVP Microsoft® Most Valuable Professional

What we'll talk about

Data Mesh

Medallion

Workspaces & Capacities

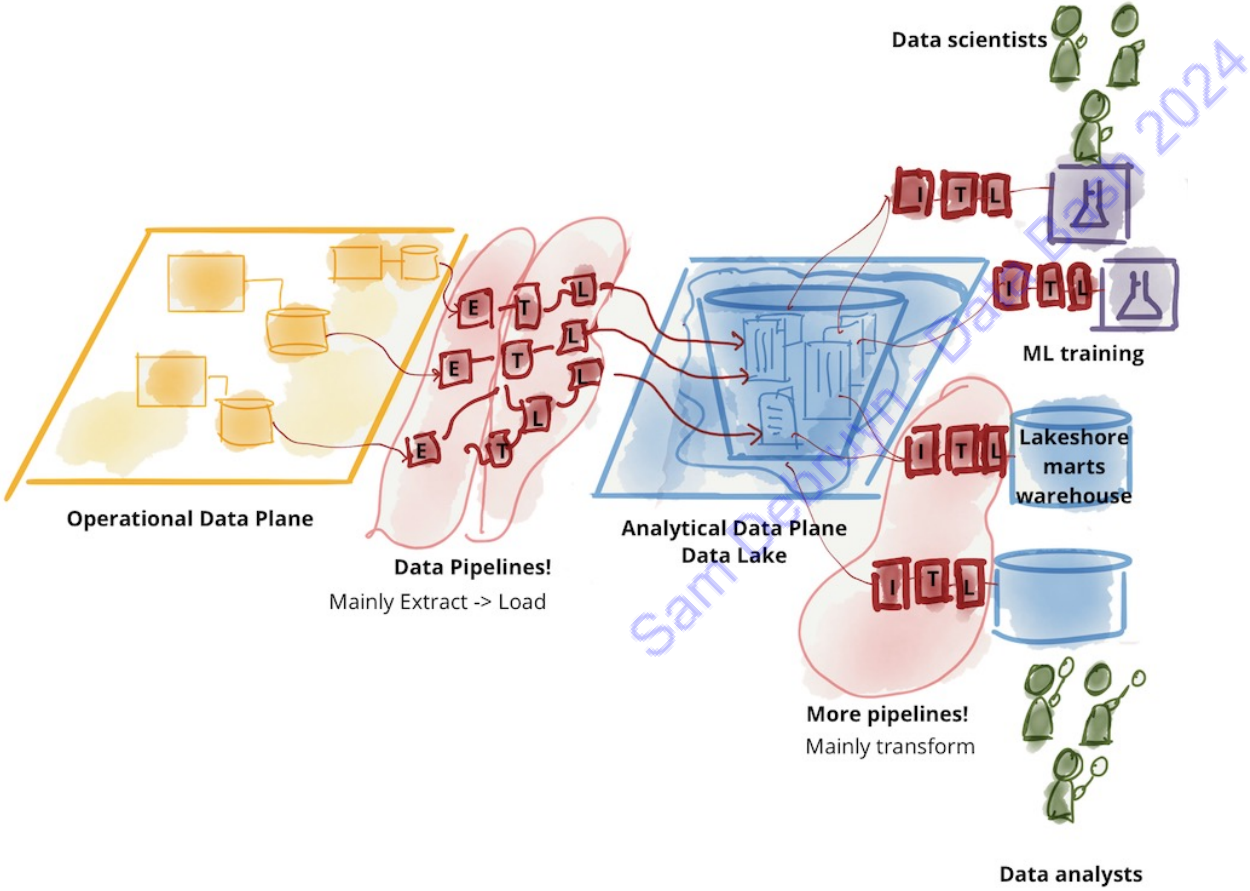
Medallion & Data Mesh on Fabric

Capacity design for scalability

Access control & Domains

Data mesh

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

The 4 Principles of the Data Mesh

DOMAIN-ORIENTED
DECENTRALIZED DATA OWNERSHIP

DATA PRODUCT THINKING

The 4 Principles of the Data Mesh

DOMAIN-ORIENTED DECENTRALIZED DATA OWNERSHIP

DATA PRODUCT THINKING

SELF-SERVICE ANALYTICS

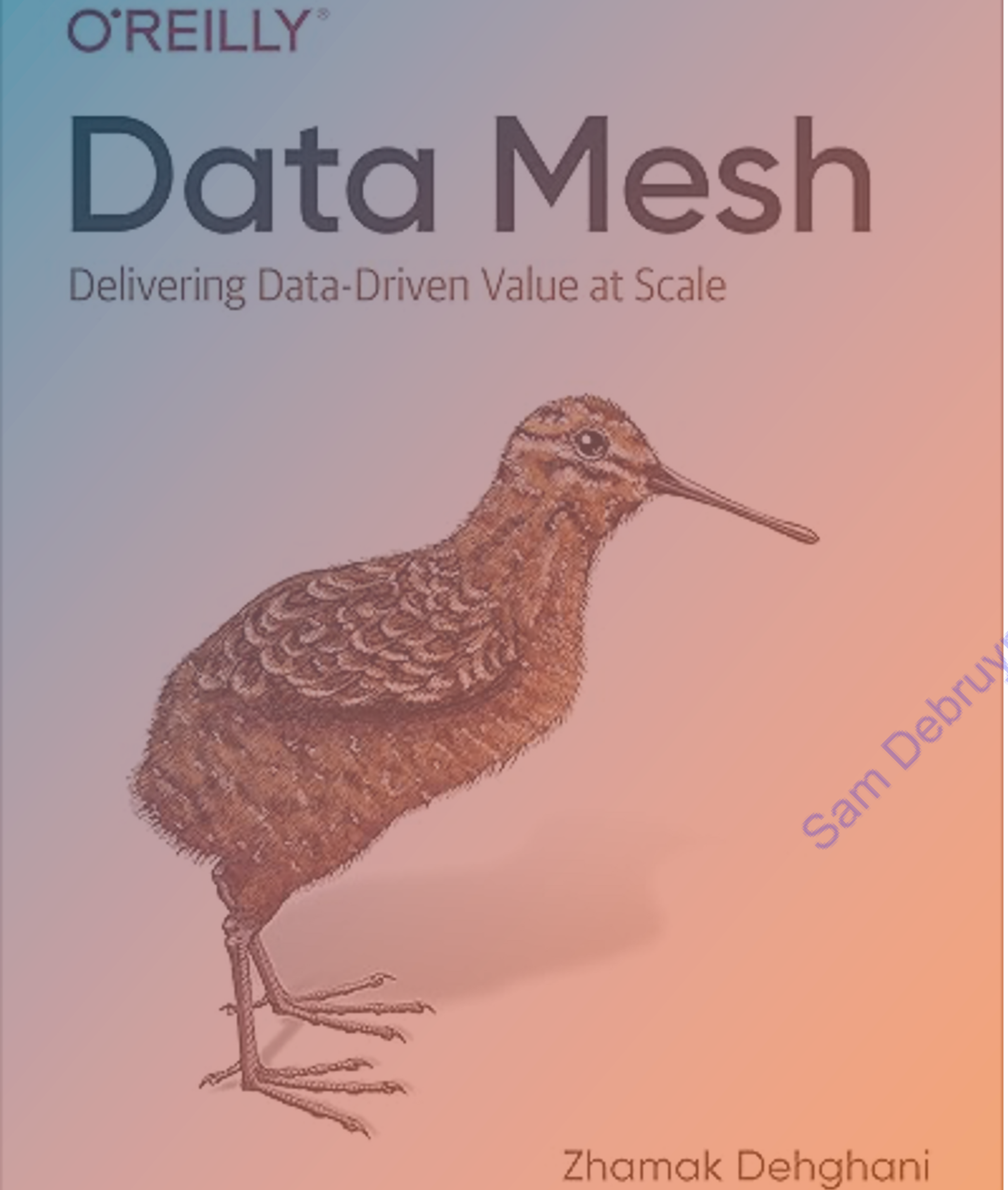
The 4 Principles of the Data Mesh

DOMAIN-ORIENTED DECENTRALIZED DATA OWNERSHIP

DATA PRODUCT THINKING

SELF-SERVICE ANALYTICS

FEDERATED GOVERNANCE



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

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

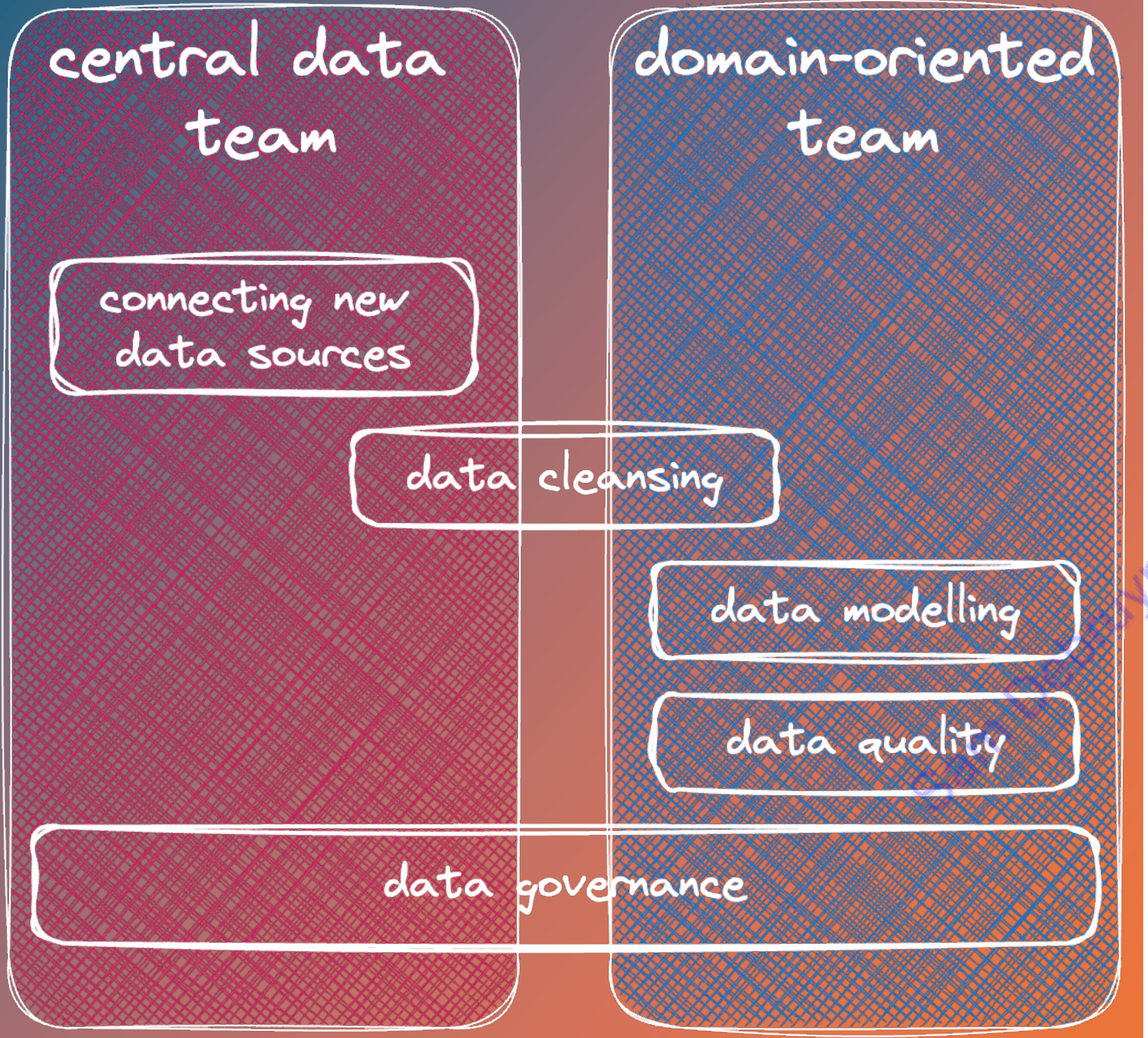
Cleansed/silver

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.



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

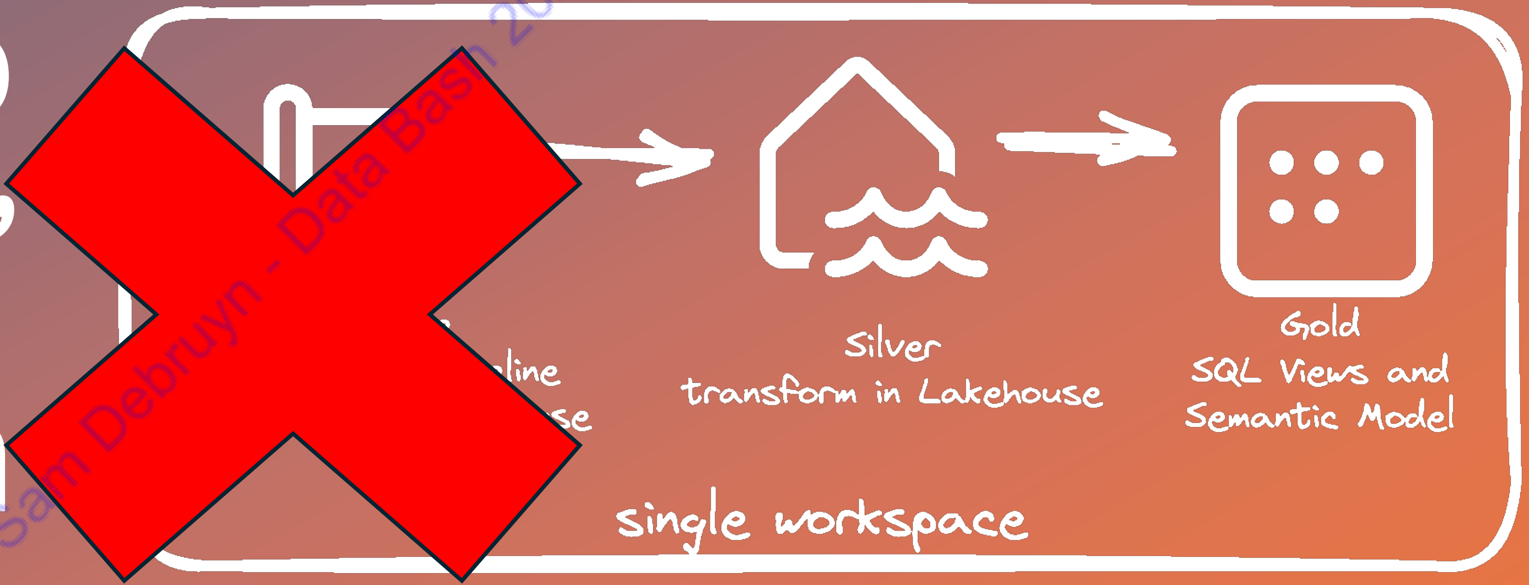
You often start from the data products in the gold layer and work your way back.

So… how do we bring these concepts together?

Data mesh and medallion with Fabric

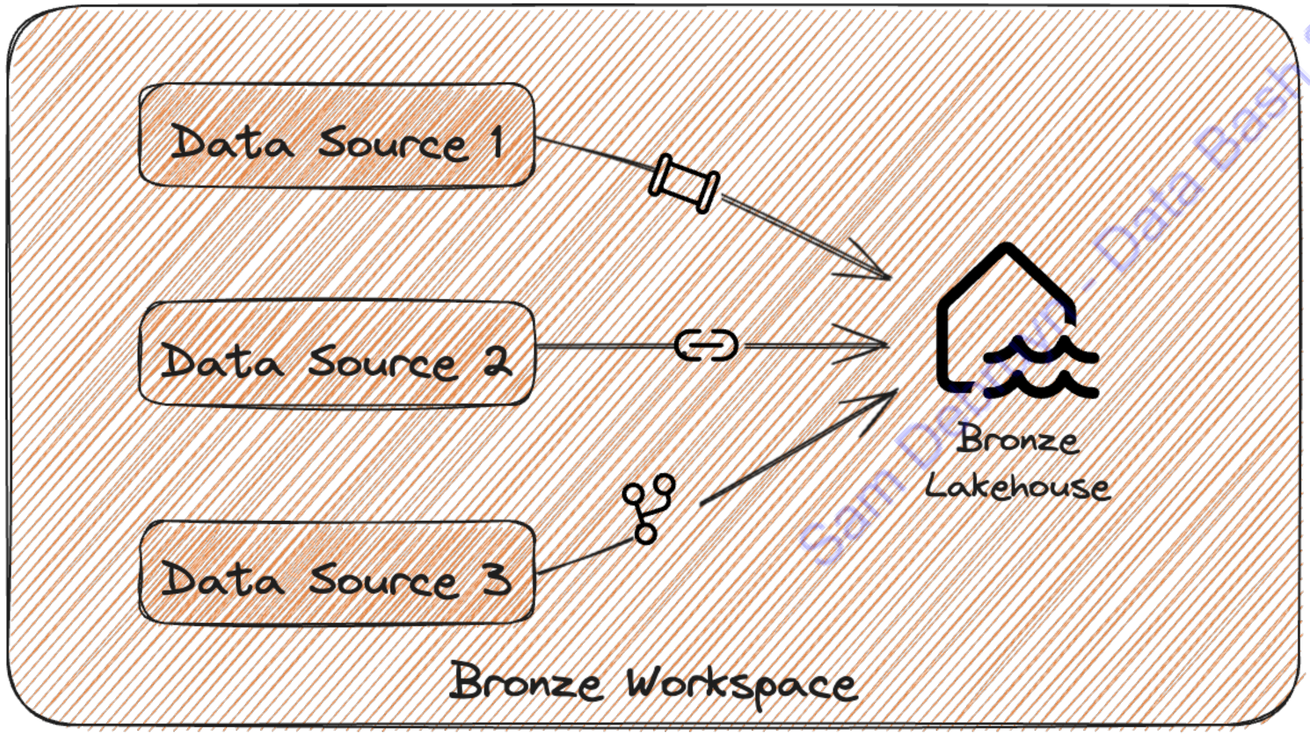
Let’s look at Workspace design for medallion and data mesh in Microsoft Fabric

The "Hello World" version



Bronze

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

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

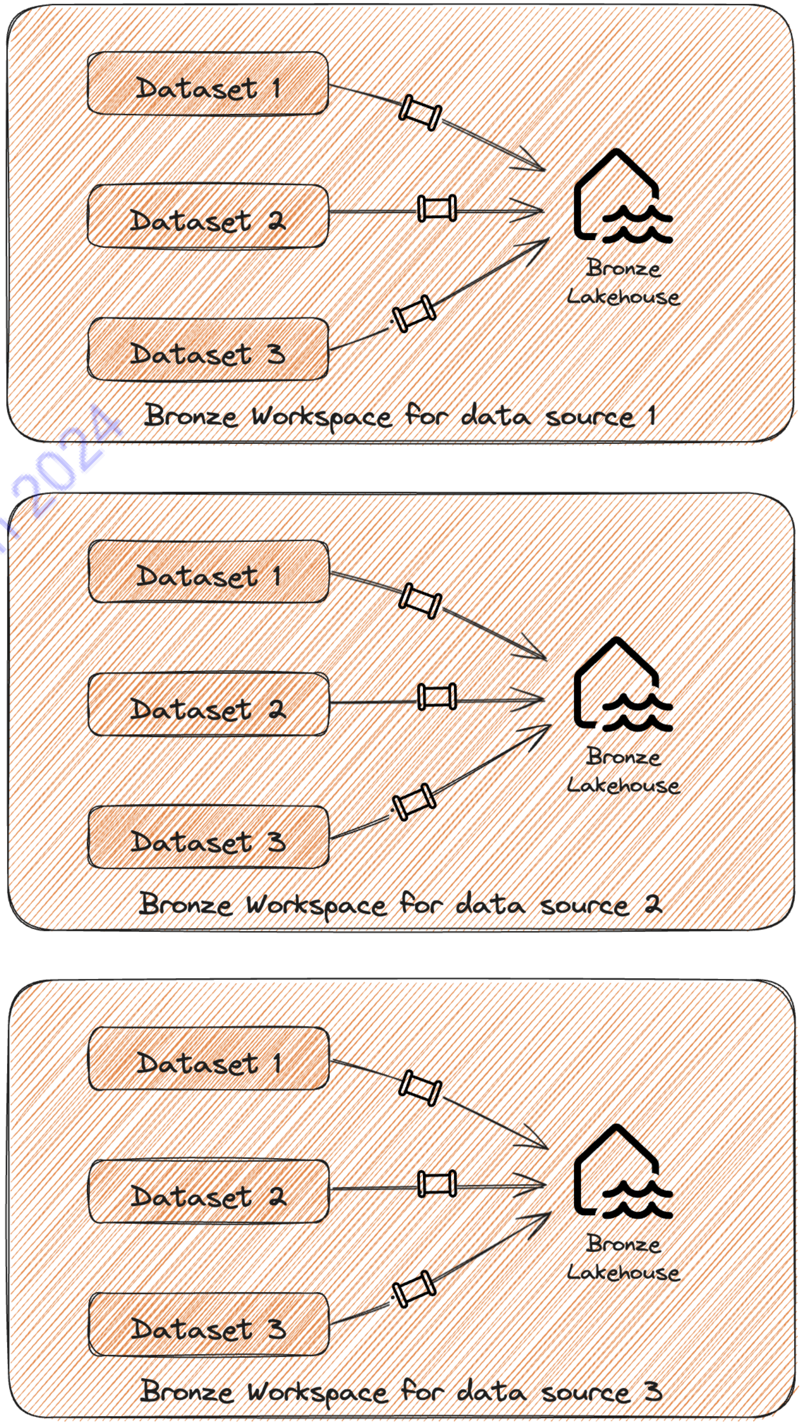
- Data Factory / Copy Activity
- **Copy Job**
- **Database Mirroring**
- Dataflow Gen2
- Notebooks / Spark Jobs
- ADLS APIs
- OneLake File Explorer
- DWH SQL APIs: COPY INTO / OPENROWSET

Bronze layer layout

Multiple options, depending on data platform size & complexity:

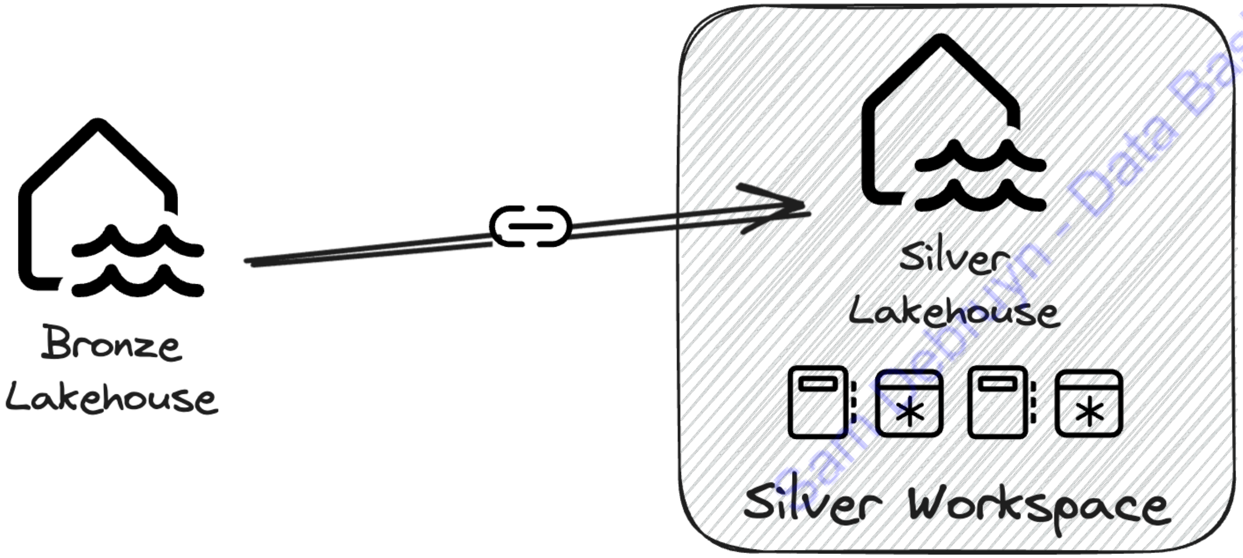
- Single Workspace, Single Lakehouse, 1 schema per source system
- Single Workspace, 1 Lakehouse per source system
- 1 Workspace per source system

Bronze layer layout



Silver

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

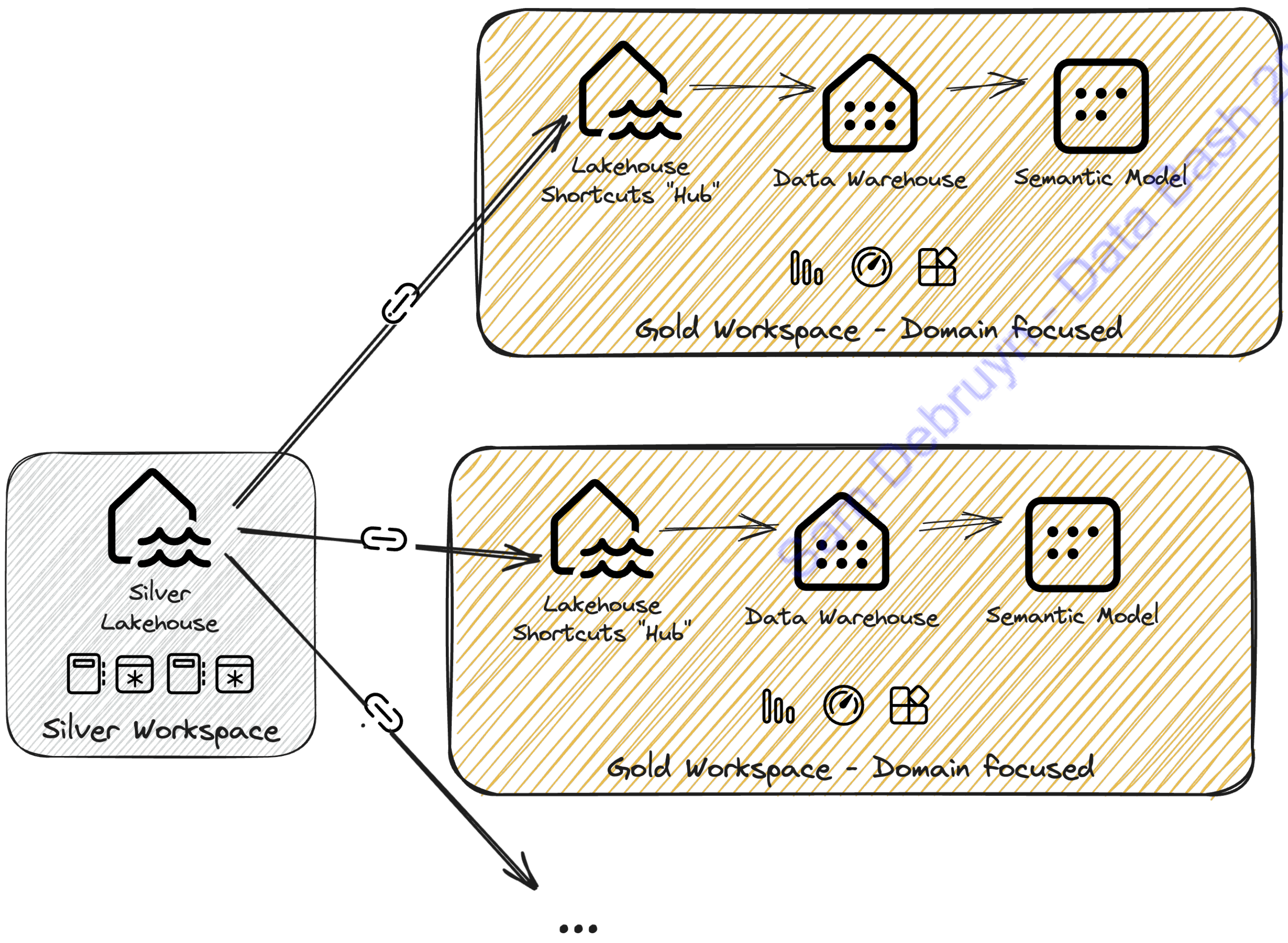
Some teams tend to prefer data vault here

Other approaches:

- Replicate layout from bronze
- Wide tables
- …

Gold

Gold


Lakehouse Shortcuts "Hub"
Data Warehouse
Semantic Model
Gold Workspace - Domain focused
Lakehouse Shortcuts "Hub"
Data Warehouse
Semantic Model
Gold Workspace - Domain focused
Silver Lakehouse
Silver Workspace
...

Data is linked from Silver Workspace using Shortcuts

Data modelling in SQL in a Data Warehouse

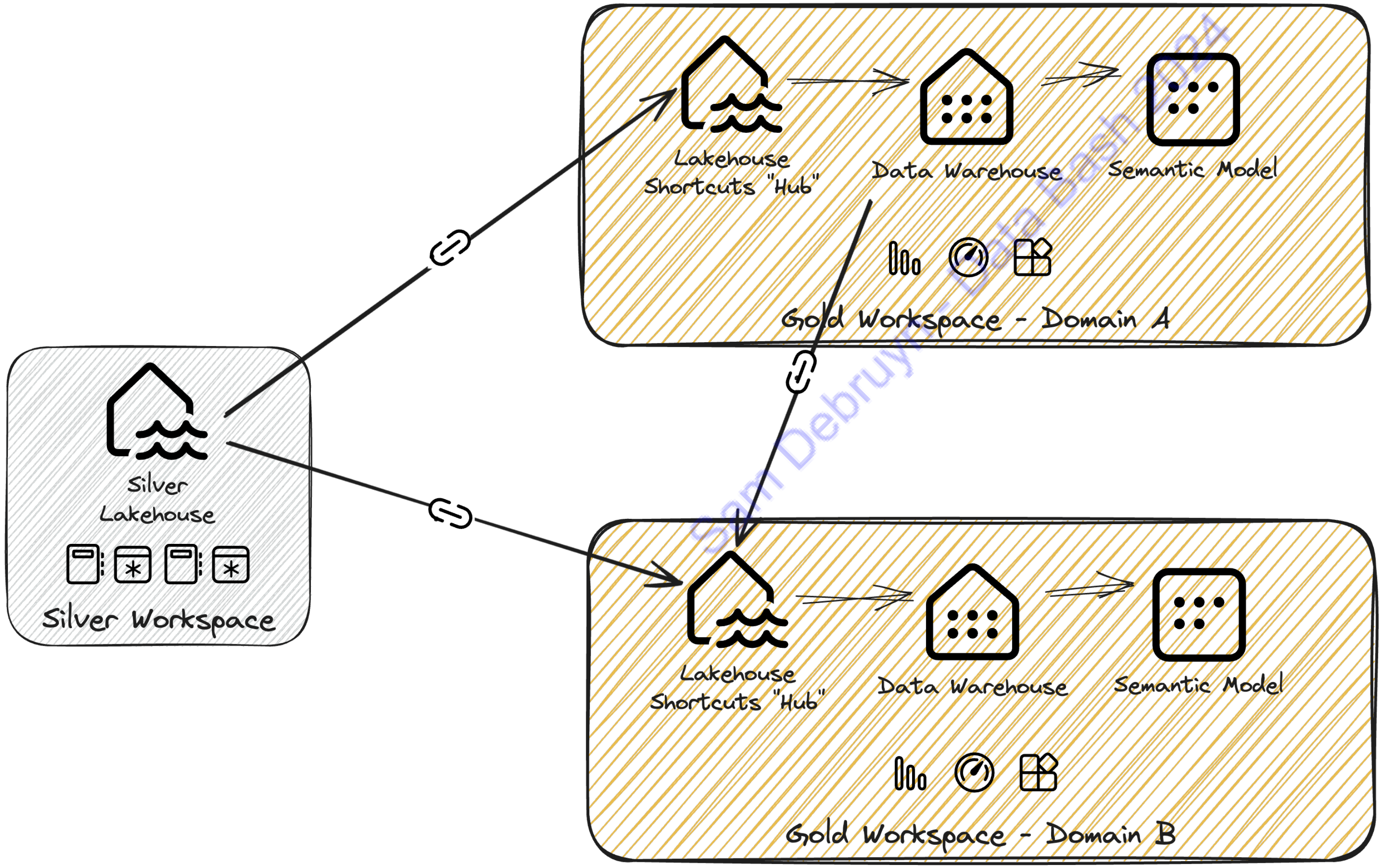
Data is separated by domain

Workspace also serves as access control boundary

Decentralized data domain focused teams

Core skill: analytics engineering

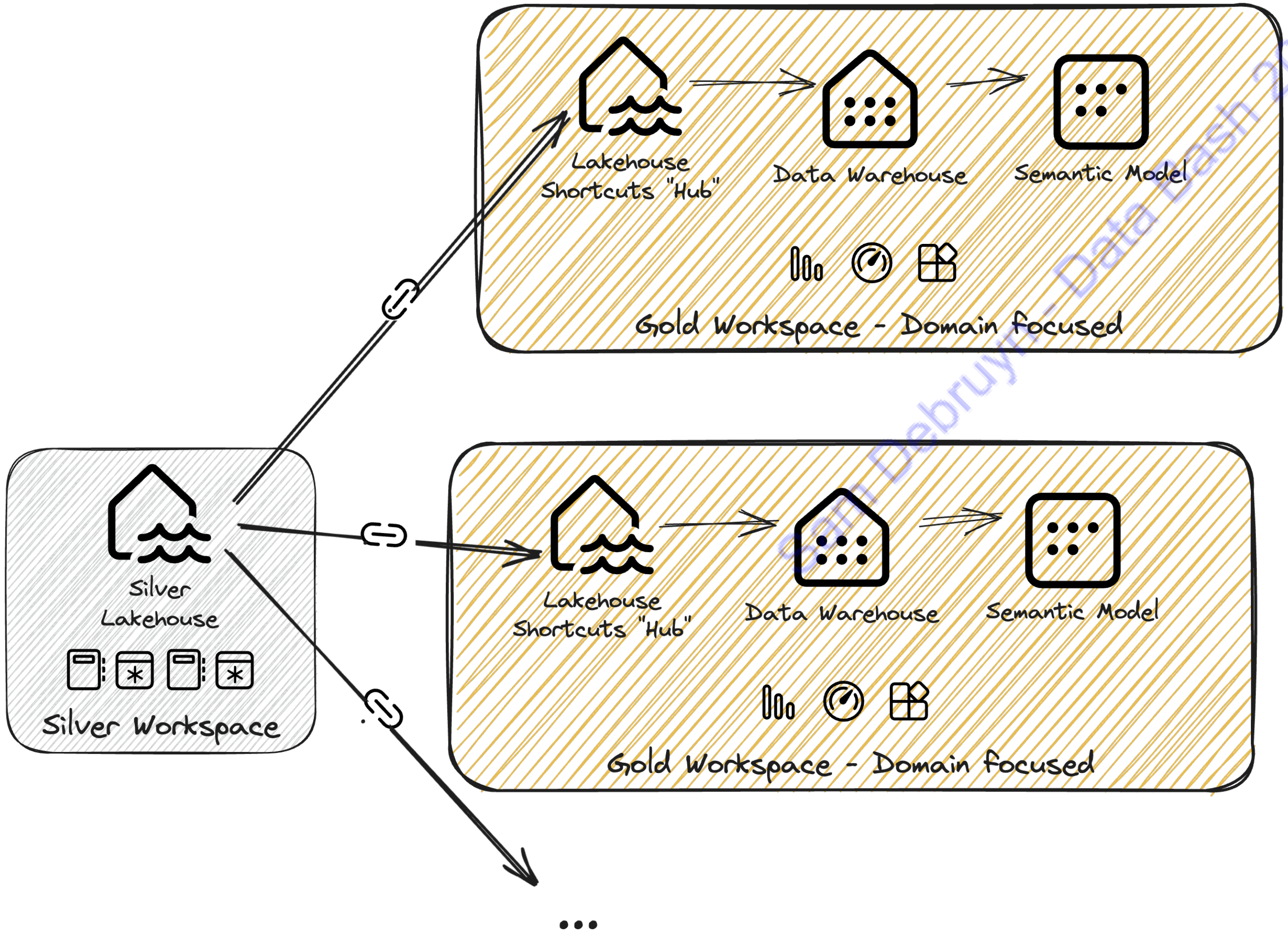
Gold



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


Lakehouse Shortcuts "Hub"
Data Warehouse
Semantic Model
Gold Workspace - Domain focused
Lakehouse Shortcuts "Hub"
Data Warehouse
Semantic Model
Gold Workspace - Domain focused
Silver Lakehouse
Silver Workspace
...

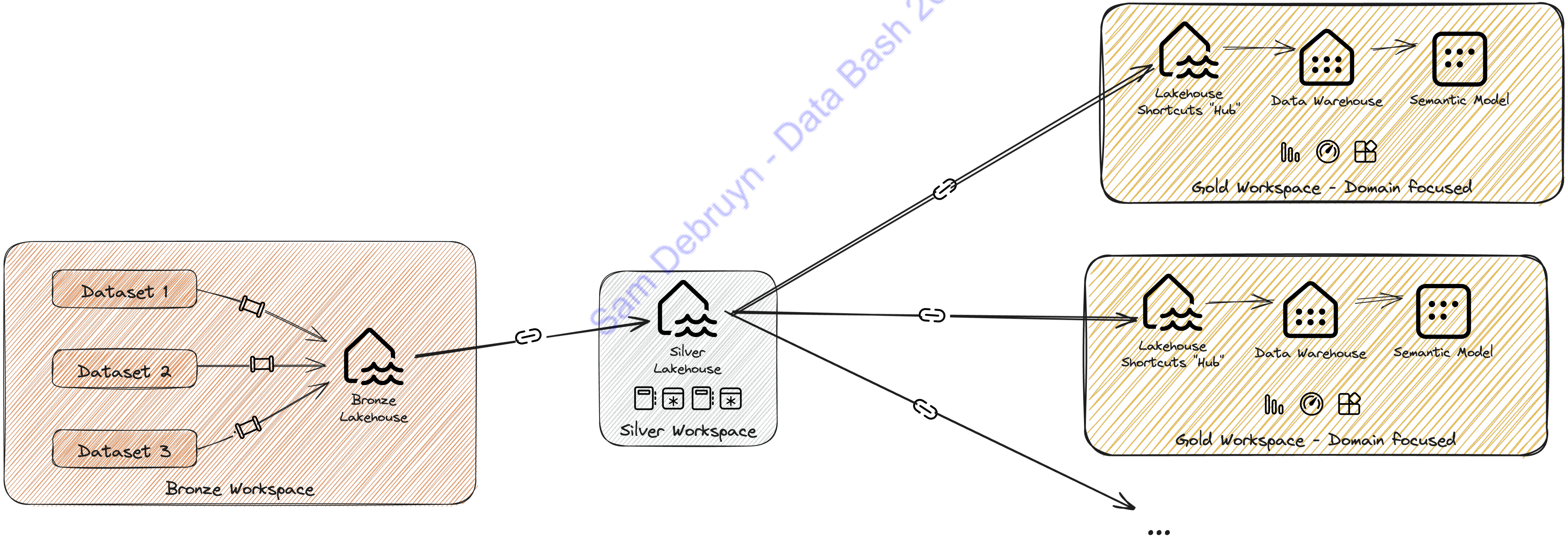
Why the split by domain?

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

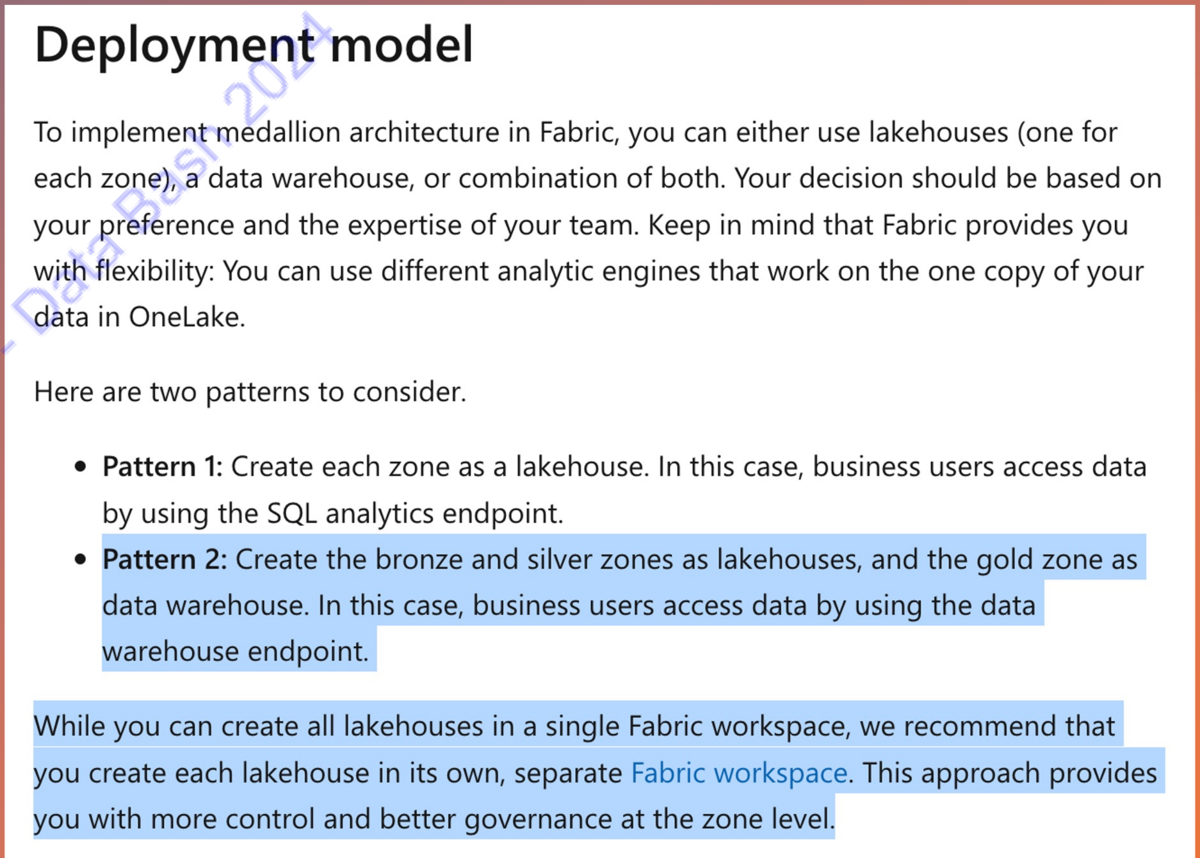
Creates a producer-consumer relationship

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.

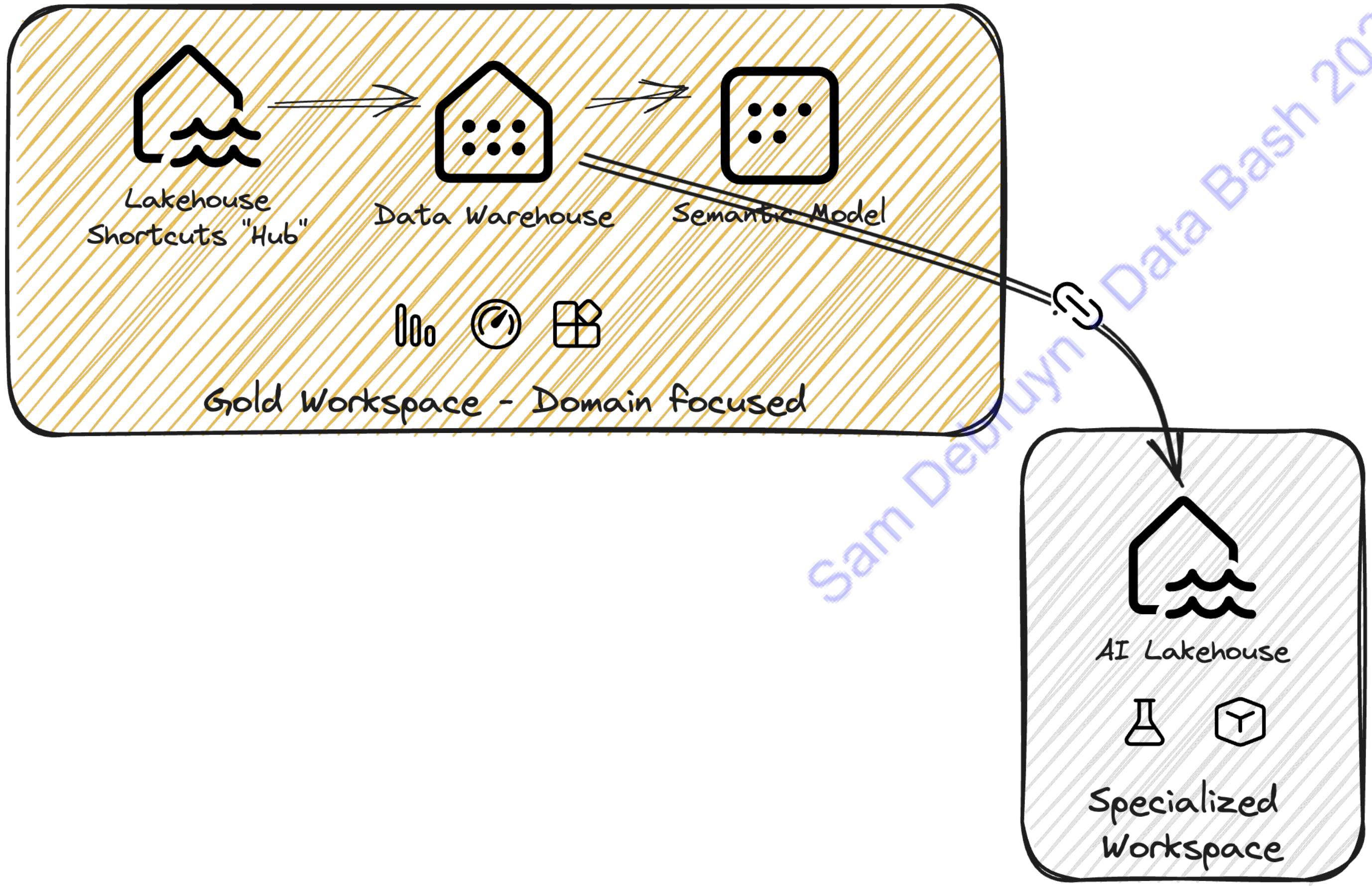
While you can create all lakehouses in a single Fabric workspace, we recommend that you create each lakehouse in its own, separate Fabric workspace. This approach provides you with more control and better governance at the zone level.

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"

Workspaces & Capacities

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

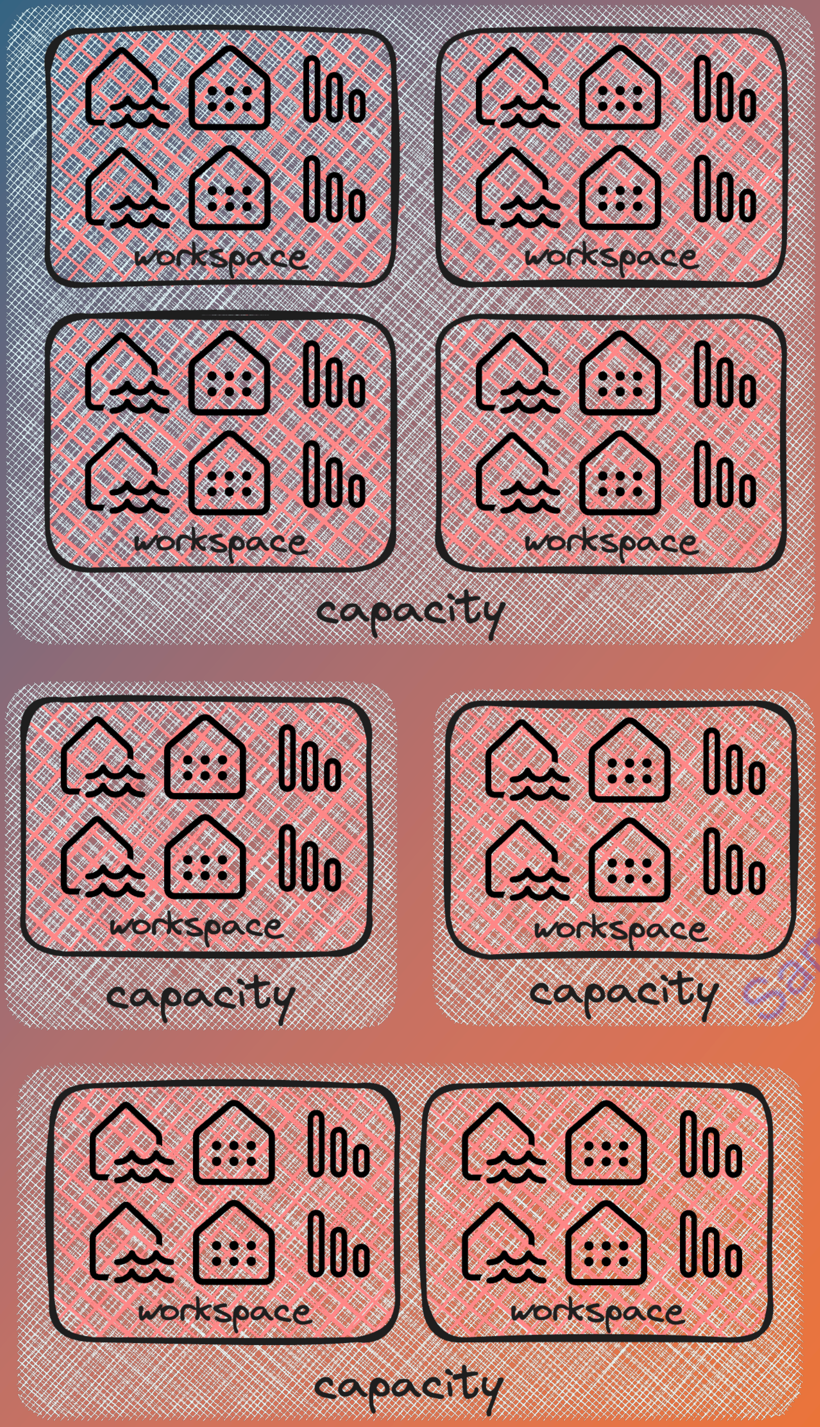
Why should you create separate Workspaces?

Access control

- Admin
- Member
- Contributor
- Viewer

Why should you create separate Workspaces?

Capacity Management



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)

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

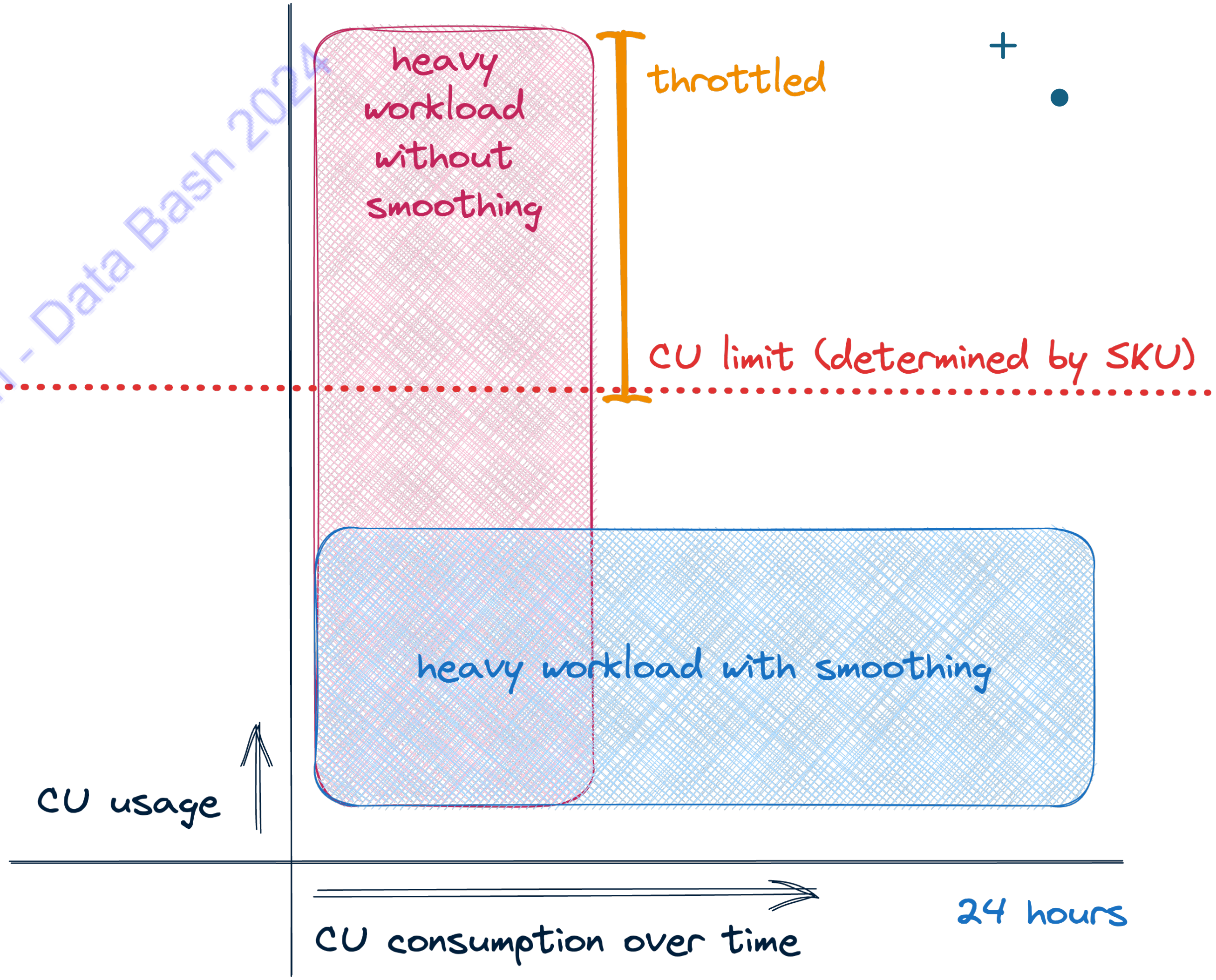
Bursting & smoothing

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


heavy workload without smoothing
throttled
CU limit (determined by SKU)
heavy workload with smoothing
CU usage
CU consumption over time
24 hours

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

Throttling

Overage: CUs consumed over what was available for your operation

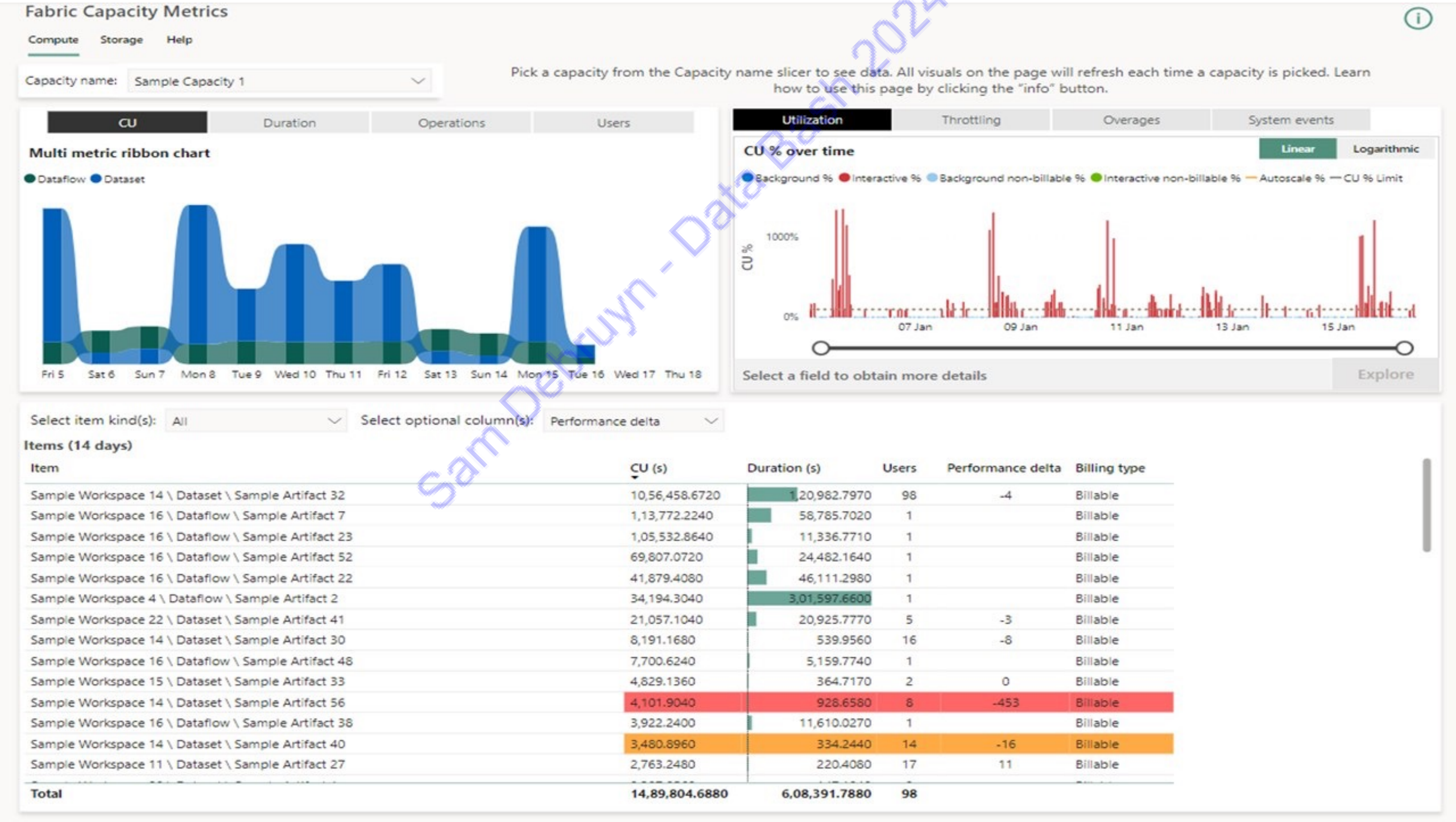
Throttling on interactive operations:

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



Fabric Capacity Metrics

Capacity name: Sample Capacity 1

Pick a capacity from the Capacity name slicer to see data. All visuals on the page will refresh each time a capacity is picked. Learn how to use this page by clicking the "info" button.

Select a field to obtain more details

Select item kind(s): All

Select optional column(s): Performance delta

Items (14 days)

Item	CU (s)	Duration (s)	Users	Performance delta	Billing type
Sample Workspace 14 \ Dataset \ Sample Artifact 32	10,56,458.6720	1,20,982.7970	98	-4	Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 7	1,13,772.2240	58,785.7020	1		Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 23	1,05,532.8640	11,336.7710	1		Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 52	69,807.0720	24,482.1640	1		Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 22	41,879.4080	46,111.2980	1		Billable
Sample Workspace 4 \ Dataflow \ Sample Artifact 2	34,194.3040	3,01,597.6600	1		Billable
Sample Workspace 22 \ Dataset \ Sample Artifact 41	21,057.1040	20,925.7770	5	-3	Billable
Sample Workspace 14 \ Dataset \ Sample Artifact 30	8,191.1680	539.9560	16	-8	Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 48	7,700.6240	5,159.7740	1		Billable
Sample Workspace 15 \ Dataset \ Sample Artifact 33	4,829.1360	364.7170	2	0	Billable
Sample Workspace 14 \ Dataset \ Sample Artifact 56	4,101.9040	928.6580	8	-453	Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 38	3,922.2400	11,610.0270	1		Billable
Sample Workspace 14 \ Dataset \ Sample Artifact 40	3,480.8960	334.2440	14	-16	Billable
Sample Workspace 11 \ Dataset \ Sample Artifact 27	2,763.2480	220.4080	17	11	Billable
Total	**14,89,804.6880**	**6,08,391.7880**	**98**		

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.

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

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

Domains

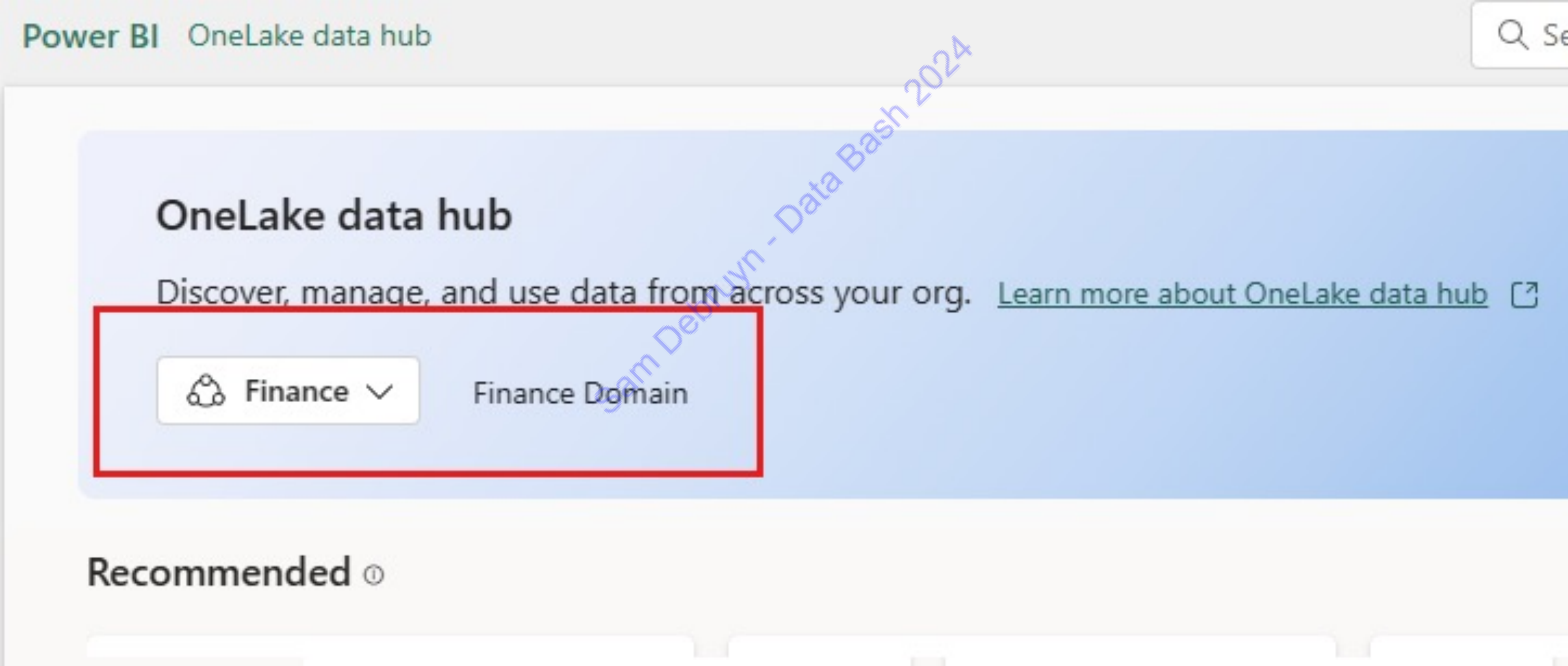
Logically grouping together data in an organization by bundling Workspaces in Domains

Domains can have Subdomains

Managed by Domain Admins and Domain Contributors

Centralize or group certified datasets

Domains: OneLake Data Hub



Power BI OneLake data hub

OneLake data hub

Discover, manage, and use data from across your org. Learn more about OneLake data hub

Recommended

Recap

RECAP: The 4 Principles of the Data Mesh

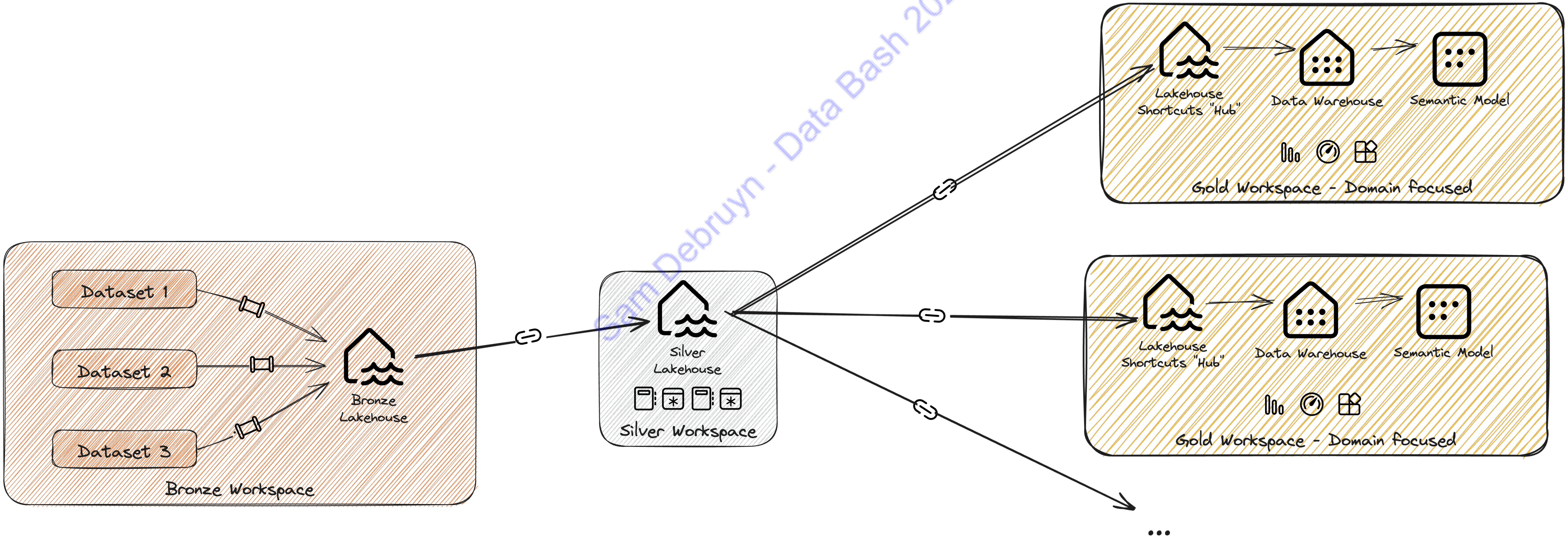
DOMAIN-ORIENTED DECENTRALIZED DATA OWNERSHIP

DATA PRODUCT THINKING

SELF-SERVICE ANALYTICS

FEDERATED GOVERNANCE

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’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.

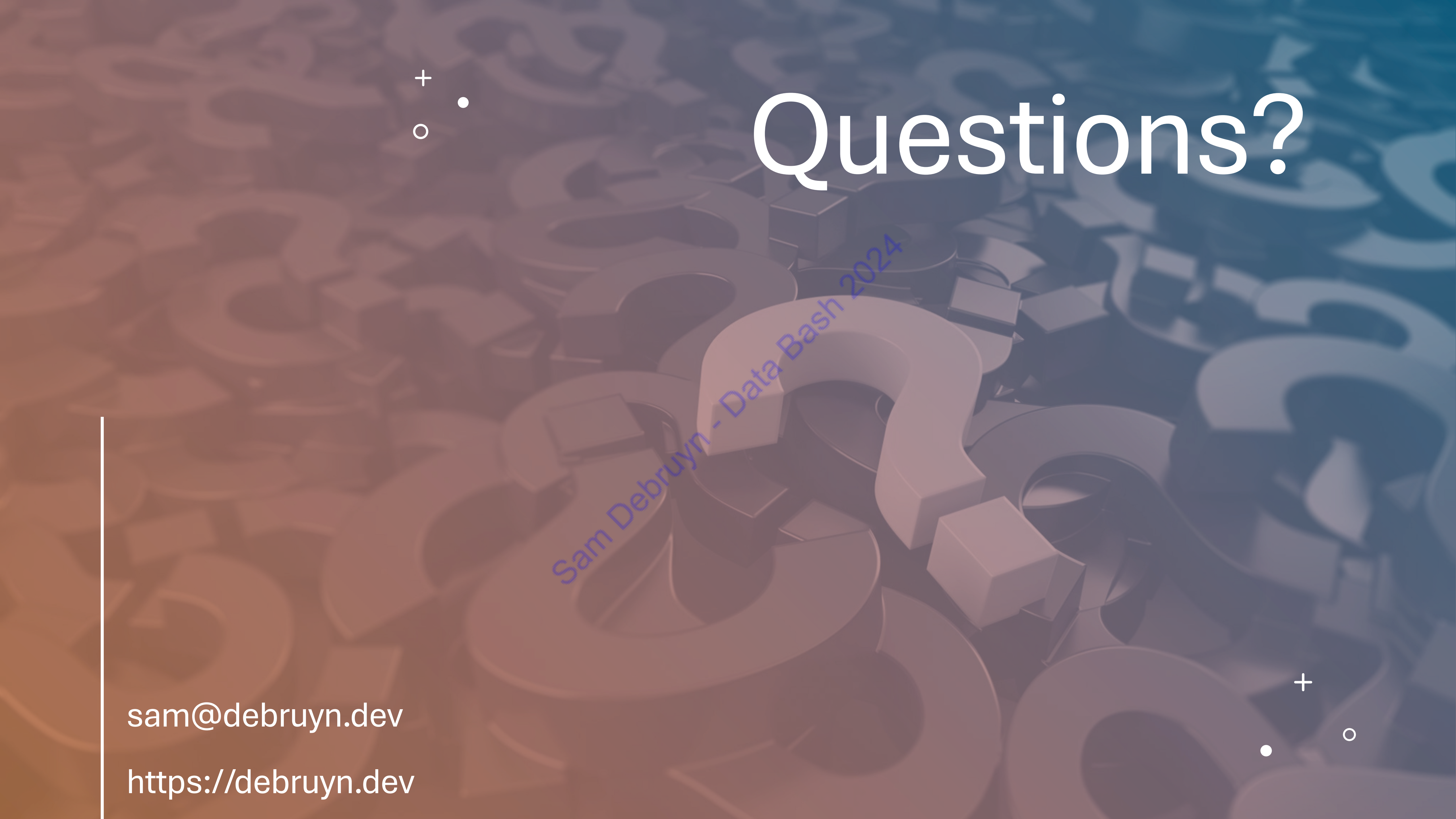
DO split Workspaces and their linked Capacities by workload

- 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

Questions?



sam@debruyn.dev

https://debruyn.dev