FABRIC CAPACITIES

DIVIDE & CONQUER

Sam Debruyn

Data Saturday Holland
October 2024

Thank you sponsors!















Rate Data Saturday Holland

SUPPORT CASPER

1 review = 1 €

Towards beating pancreatic cancer

Who am I?

Sam Debruyn

📍 Heist-op-den-Berg, BE

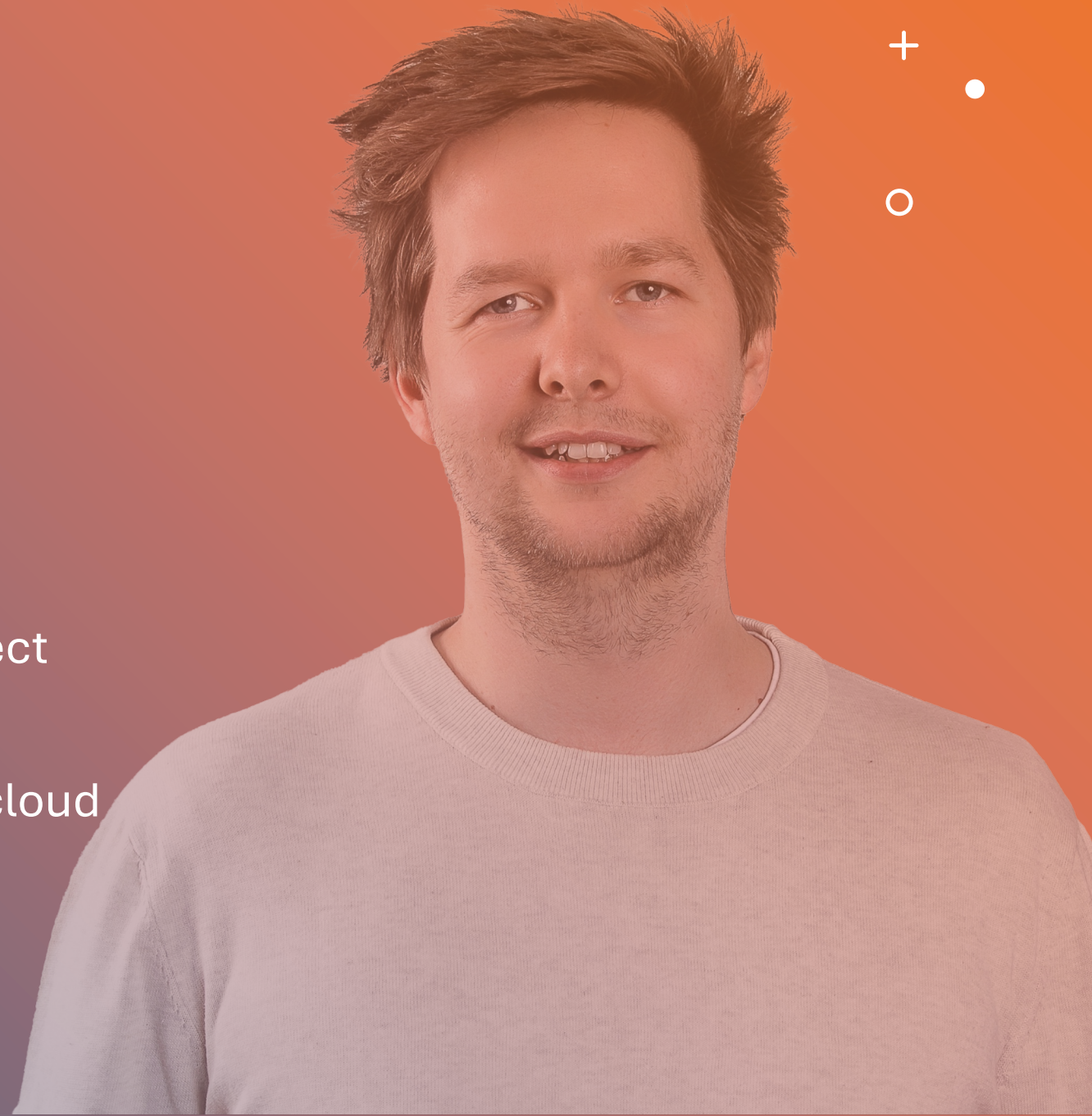
💼 Consultant / Data & Cloud Architect

5️⃣ years in data

🔟 years in software / architecture / cloud

🫶 Fabric, Azure, modern data stack

Microsoft® MVP Most Valuable Professional



What we'll talk about

Intro
Why multiple Workspaces
Capacities
Throttling
Avoiding throttling
Conclusion

SLIDES AVAILABLE AT THE END

Setting the stage…

Fabric medallion architecture example



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.

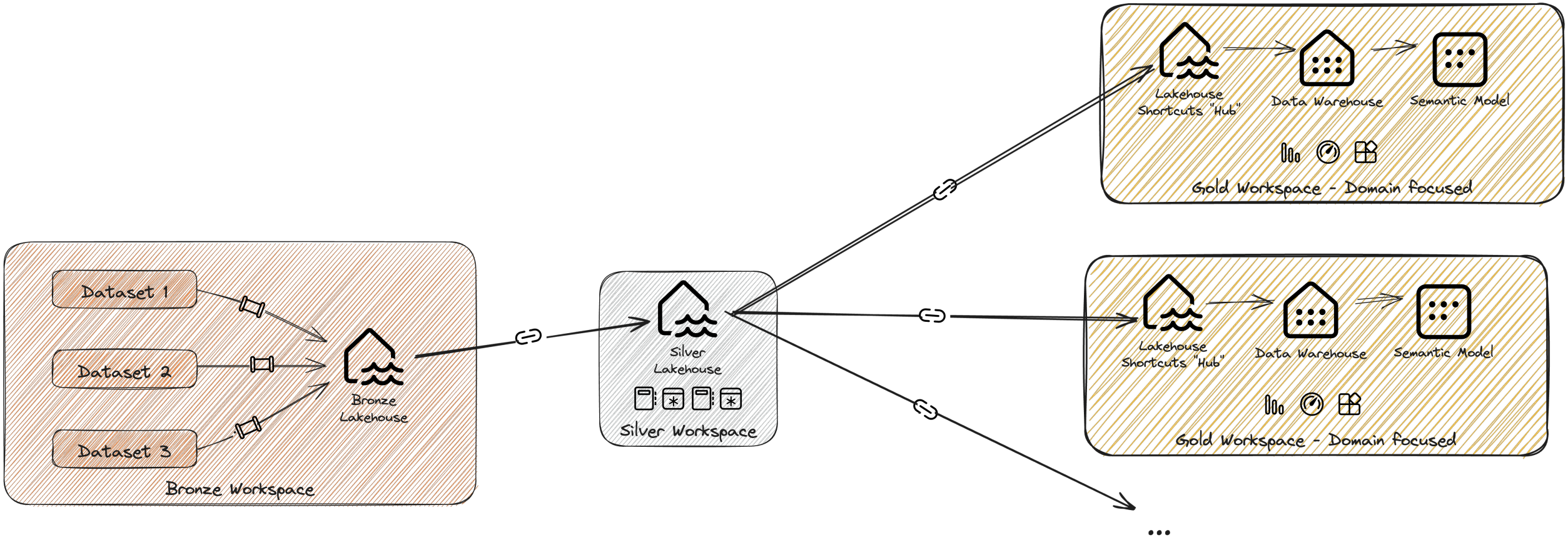


The 3 Layers of the Medallion Architecture

Curated/gold

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

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.

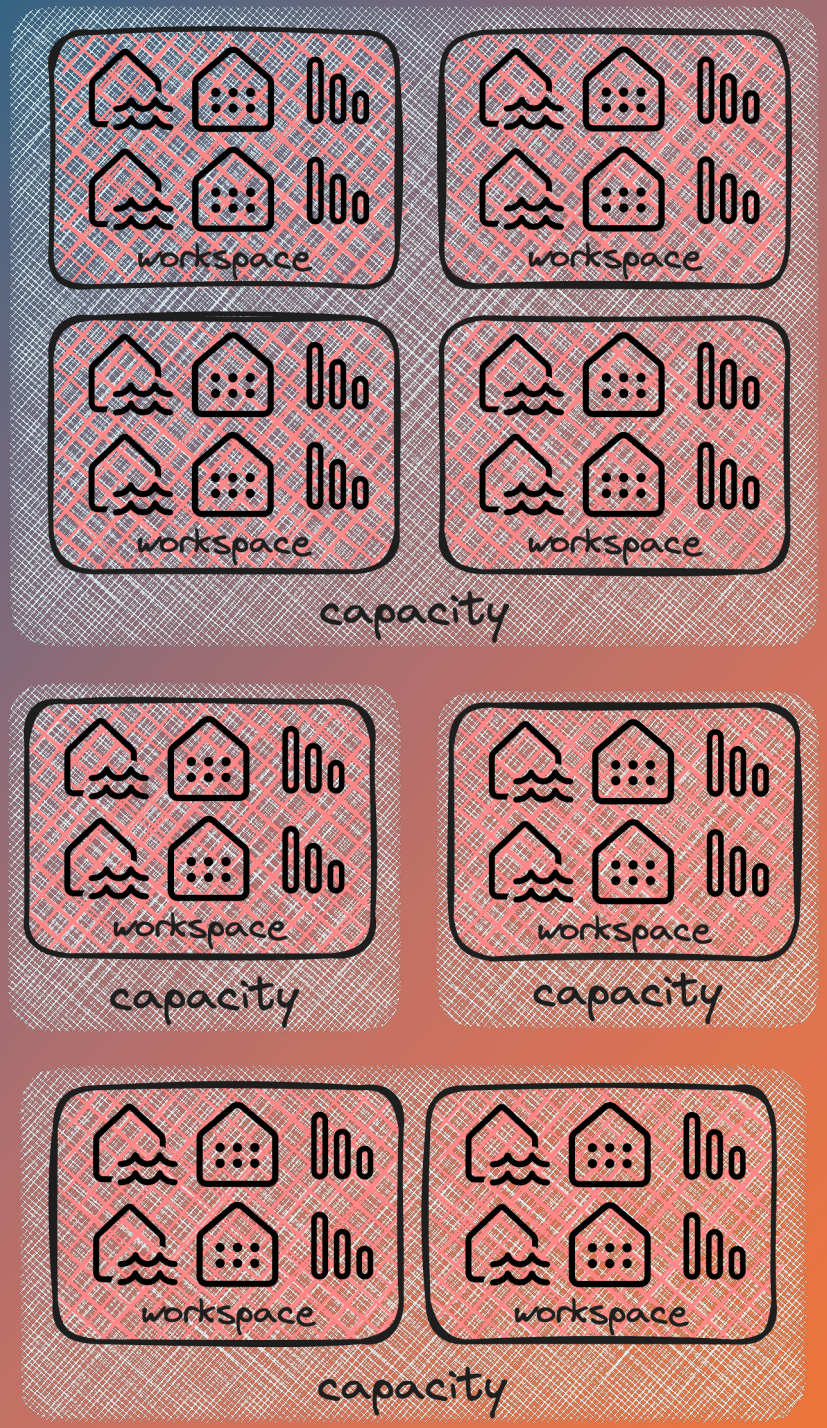
While you can create all lakehouses in a single Fabric workspace, we recommend that you create each lakehouse in its own, separate Fabric workspace. This approach provides you with more control and better governance at the zone level.

Easy to extend

This can be extended for

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

Workspaces & Capacities



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

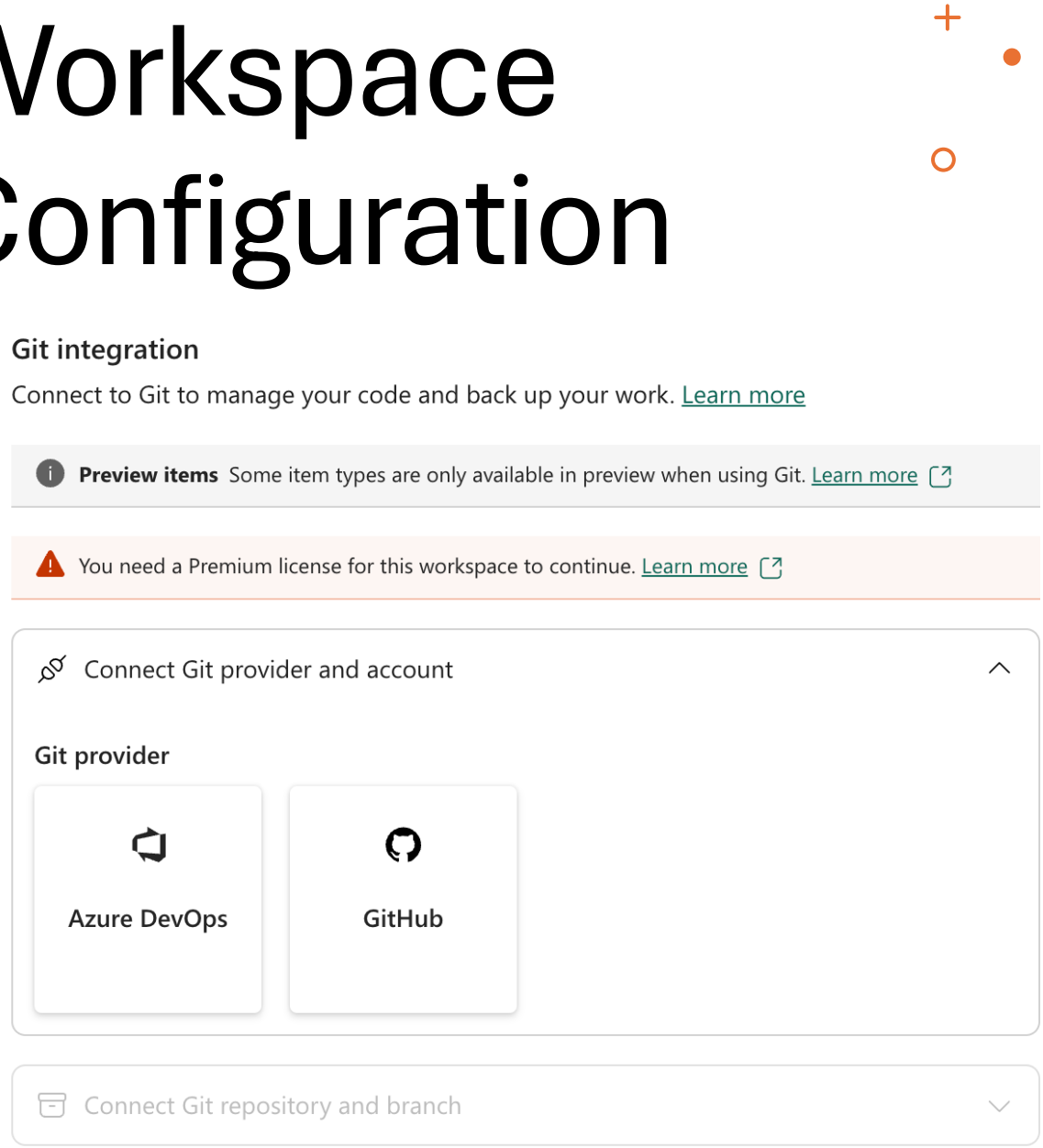
Why should you create separate Workspaces?

Workspace configuration

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

→ Different workloads might require different configurations

Workspace Configuration



Git integration

Store Fabric items in GitHub or Azure DevOps

Workspace Configuration

Workspace identities

Create and manage a workspace identity that users can use to authenticate to data sources.
Learn more

Identity details

Name	**dbt-fabric-conf24**
ID	**eb87bed5-ef9f-411c-b996-3d4d7008eee3**
Role	**Workspace Contributor**
State	**Active**

Authorized users

Name ↓	Permissions
Sam Debruyn	Can edit members
dbt-fabric-conf24	Can edit members
grp-dbt-fabric-conf24	Can edit members
dbt-fabric-conf24	Can use identity

Workspace Identity

Securely connect to resources like Storage Accounts over private networking

Workspace Configuration



Network security

Managed private endpoints let people securely connect to an Azure resource or Private Link service. Learn more

Managed private endpoints

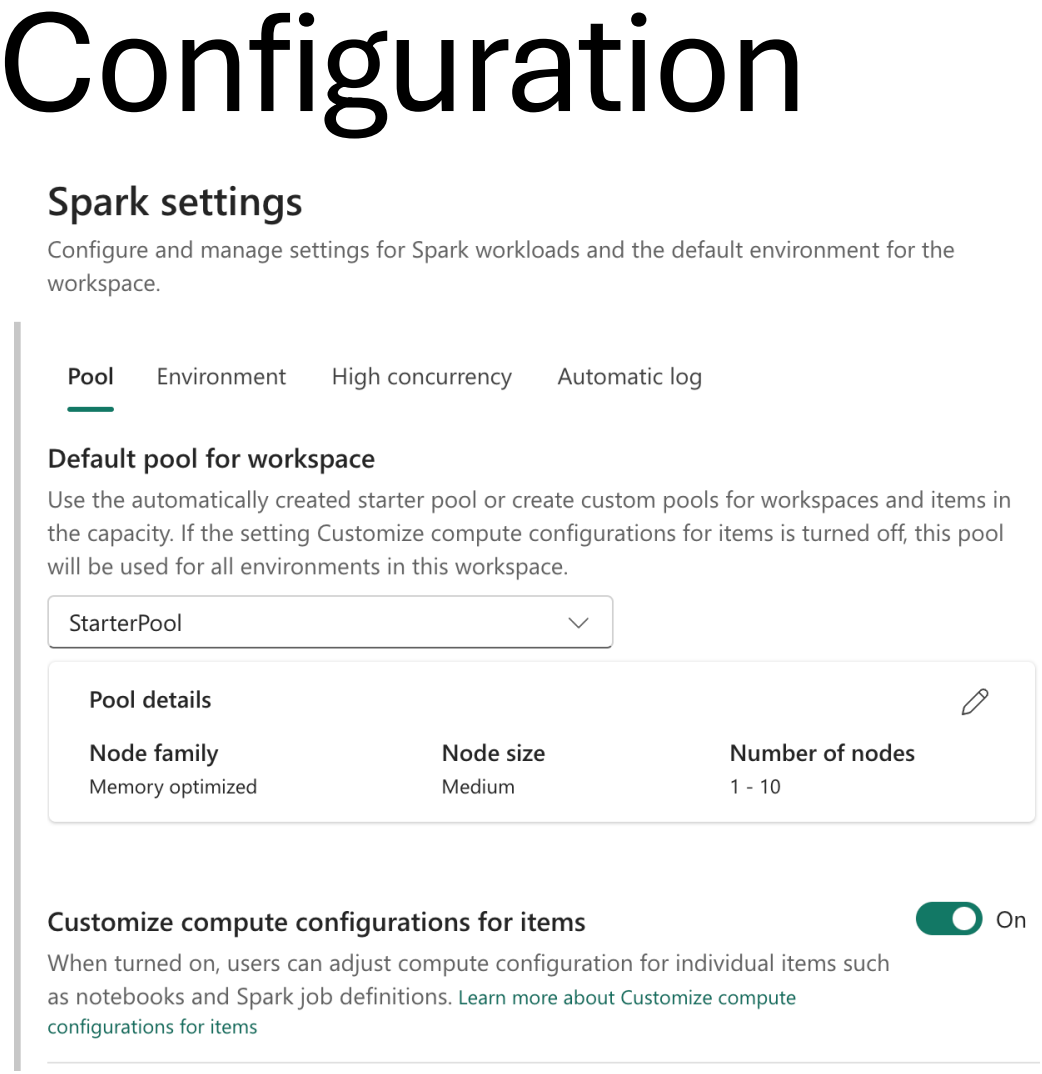
Managed private endpoints are currently available for notebooks and Spark job definitions. Learn more

+ Create Refresh Delete

Network security

Securely connect to Azure resources over Private Link

Workspace Configuration



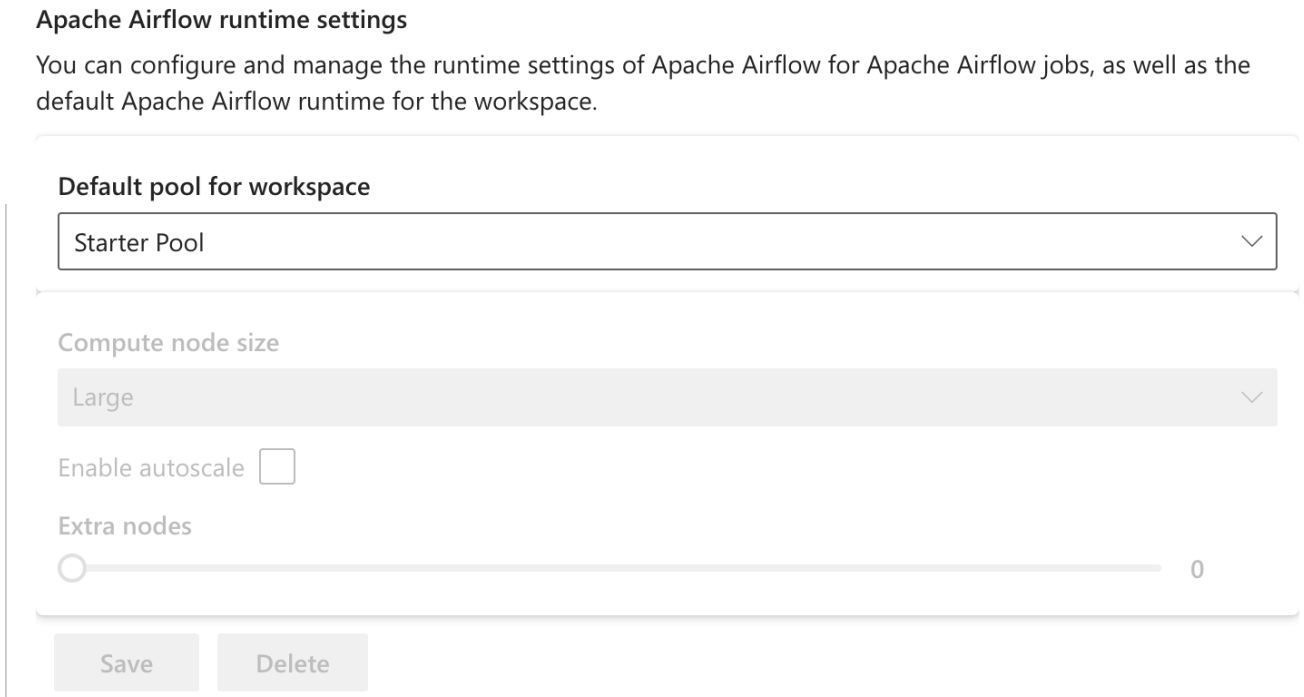
Spark settings

Configure Spark pool nodes

- Direct impact on Capacity usage
- Can be overridden per Notebook / Spark Job using Environments

Workspace Configuration





Apache Airflow runtime settings

Configure Airflow performance

- Direct impact on Capacity usage

Why should you create separate Workspaces?

Capacity Management

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

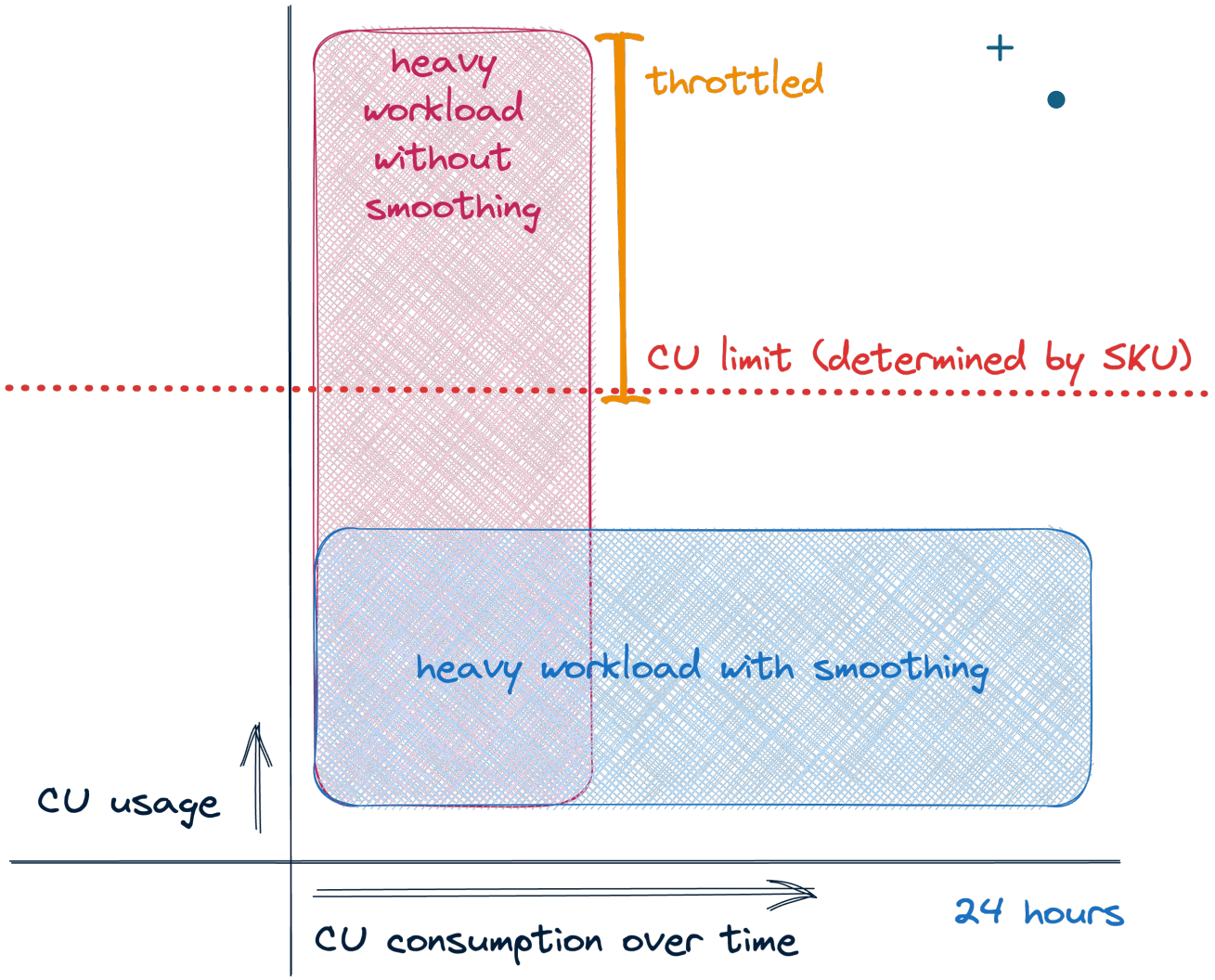
Bursting & smoothing

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



Example background

Dataflow Gen2 refresh

Scheduled Notebook

SQL queries

Copilot

Data Pipelines

OneLake read/write

Basically everything which is not a Power BI interaction

Example interactive

Interacting with a Power BI Dashboard

DirectQuery

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*
…	…	…	…	…

Bursting & smoothing

Bursting allows you to have peak usage:

- No need to schedule jobs sequentially
- Everything runs faster

Smoothing simplifies the billing and brings flexibility:

- Optimize your usage, avoid paying for what you don’t use





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

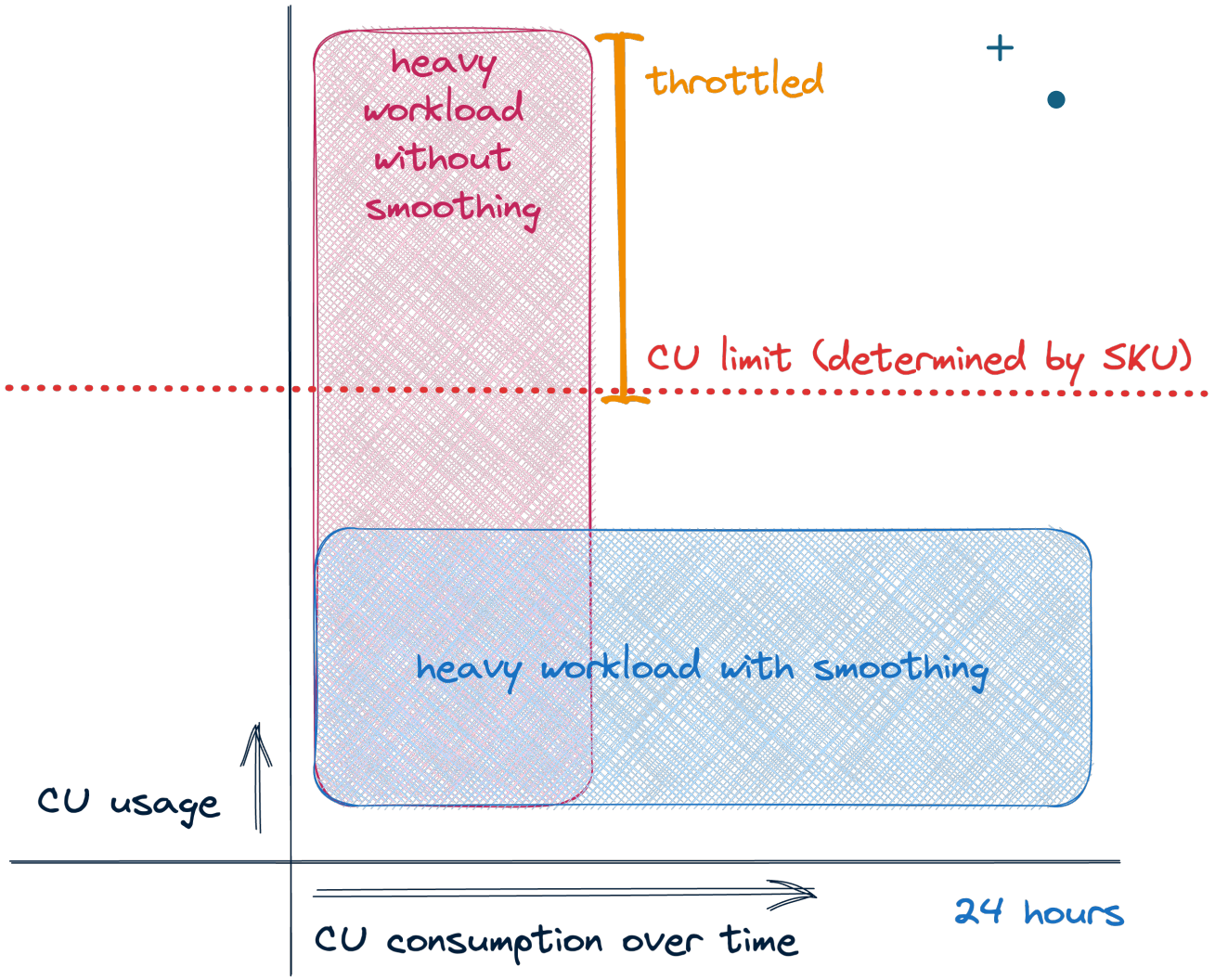
- Capacities have regions
  - Not all features are available in every region
    - Availability Zones (supported regions)
  - Compliance requirements

Bursting & smoothing

⚠ ⚠ ⚠ **Bursting & smoothing might allow a single user to use up the entire Capacity for 24 hours**

→ other users or production workloads might be throttled

Consider which teams will be using a Capacity through their Workspaces



Throttling

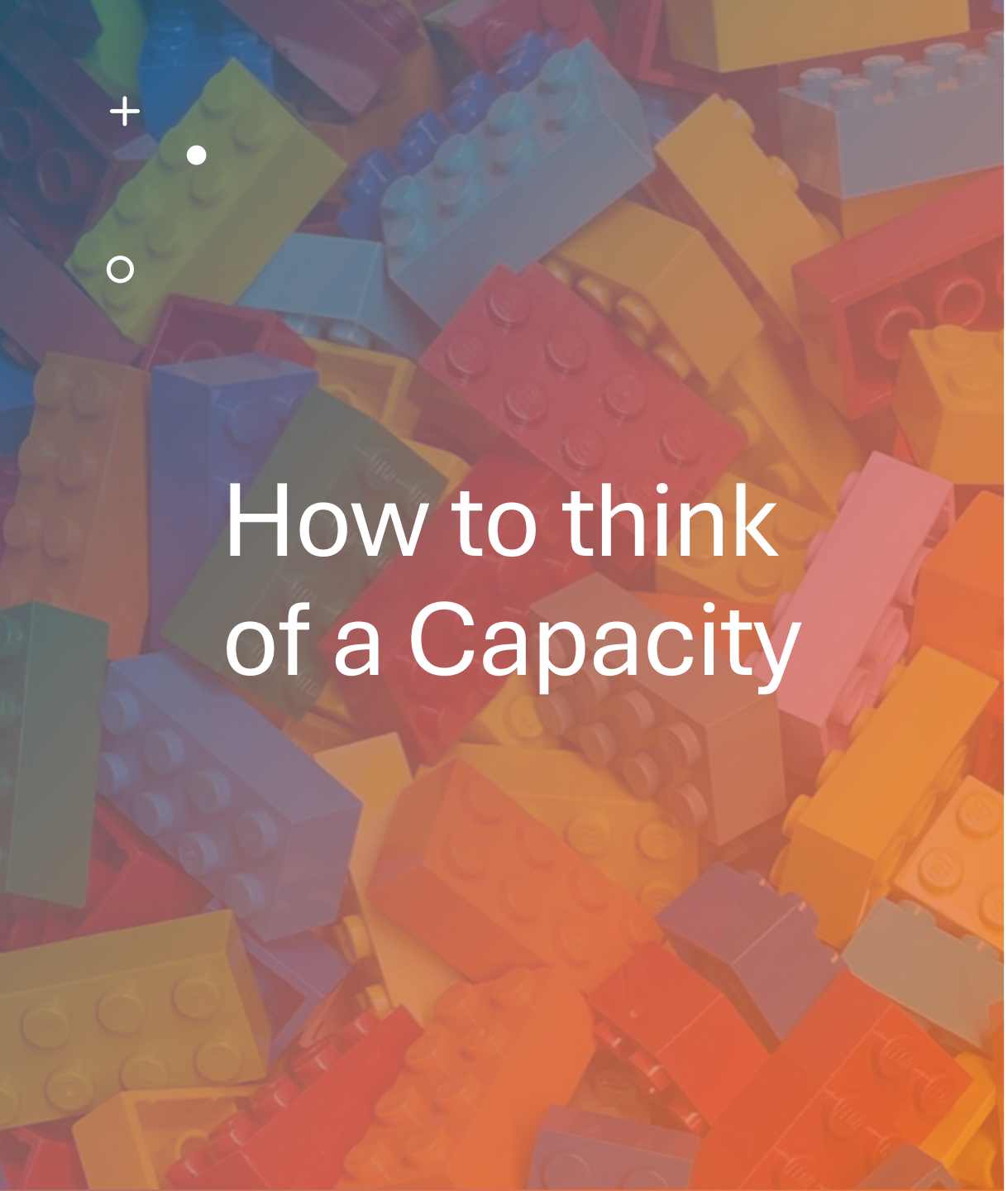
Overage: CUs consumed over what was available for your operation

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



How to think of a Capacity

Imagine you have a box of Legos.

Every second, you're given **yellow** Lego bricks.

The Lego bricks are magical bricks which turn **green** after you've kept them for **10 minutes**.

To interact with Power BI, you can only use the **yellow** Lego bricks to build something.

For everything else, you can always combine both yellow and green bricks, but you'll first always use as much **green** bricks as possible before you start using the **yellow** ones.

2 things can happen:

- You run out of bricks (**throttling**)
- Your box with bricks is filled up (max amount of CUs available)

Your Fabric SKU determines 2 things:

- The amount of **yellow** bricks you're given every second
- The size of the box



What do when throttled

Option 1: wait



What do when throttled

Option 2: pause and resume Capacity ⏸️▶️

F Capacity only!

The overage is instantly paid off.

⚠️ The overage will be billed extra on your next Azure monthly bill. This is on top of your F SKU.

What do when throttled

Option 3: scale up

F Capacity only!

The overage is smoothed out as usual.

Only scale down again when overage is paid off.



What do when throttled

Option 4: scale out

Move Workspace(s) to a different Capacity without overages.

➡️ Throttled Capacity can smooth out CUs and will become available after overage is gone.

➡️ New Capacity instantly available



What do when throttled

Option 5: autoscaling

P Capacity only!

Enable autoscaling to temporarily have more CUs available

To do *before* being throttled ideally…



What do when throttled

Option 6: Shortcut

Shortcut to your data and use it from a different Workspace + Capacity

Capacity consuming CUs is the one that is reading the data, not the one that is hosting the data in its Workspace



Optimizing CUs usage and cost

SKU	Capacity unit (CU)	Pay-as-you-go	Reservation
F2	2	**€287.647**/month	**€171.048**/month ~41% savings
F4	4	**€575.293**/month	**€342.095**/month ~41% savings
F8	8	**€1,150.585**/month	**€684.190**/month ~41% savings
F16	16	**€2,301.169**/month	**€1,368.379**/month ~41% savings
F32	32	**€4,602.338**/month	**€2,736.758**/month ~41% savings

Optimizing CUs usage and cost

Reservations <> Capacities!

Reservation: pool of available CU's to be used in one or more Capacities

E.g.

Reservation of 64 CU's:

- 1x F64
- 2x F32
- 2x F16 + 1x F32
- 4x F4 + 2x F8 + 2x F16

Optimizing CUs usage and cost

Reservation <> max amount of CU’s

💡 **HINT**:

- Reserve what you’re using on a daily base
- Combine with PAYG for infrequent usage
- Create Logic App to auto-pause PAYG Capacity
- Resume PAYG when needed
- E.g. can even be through Fabric API from a Notebook in the Capacity using reserved CU’s

Optimizing CUs usage and cost

Require Power BI Premium?

PPU is still available!

Move Power BI reports to separate Workspace(s) + activate PPU on these Workspace(s)

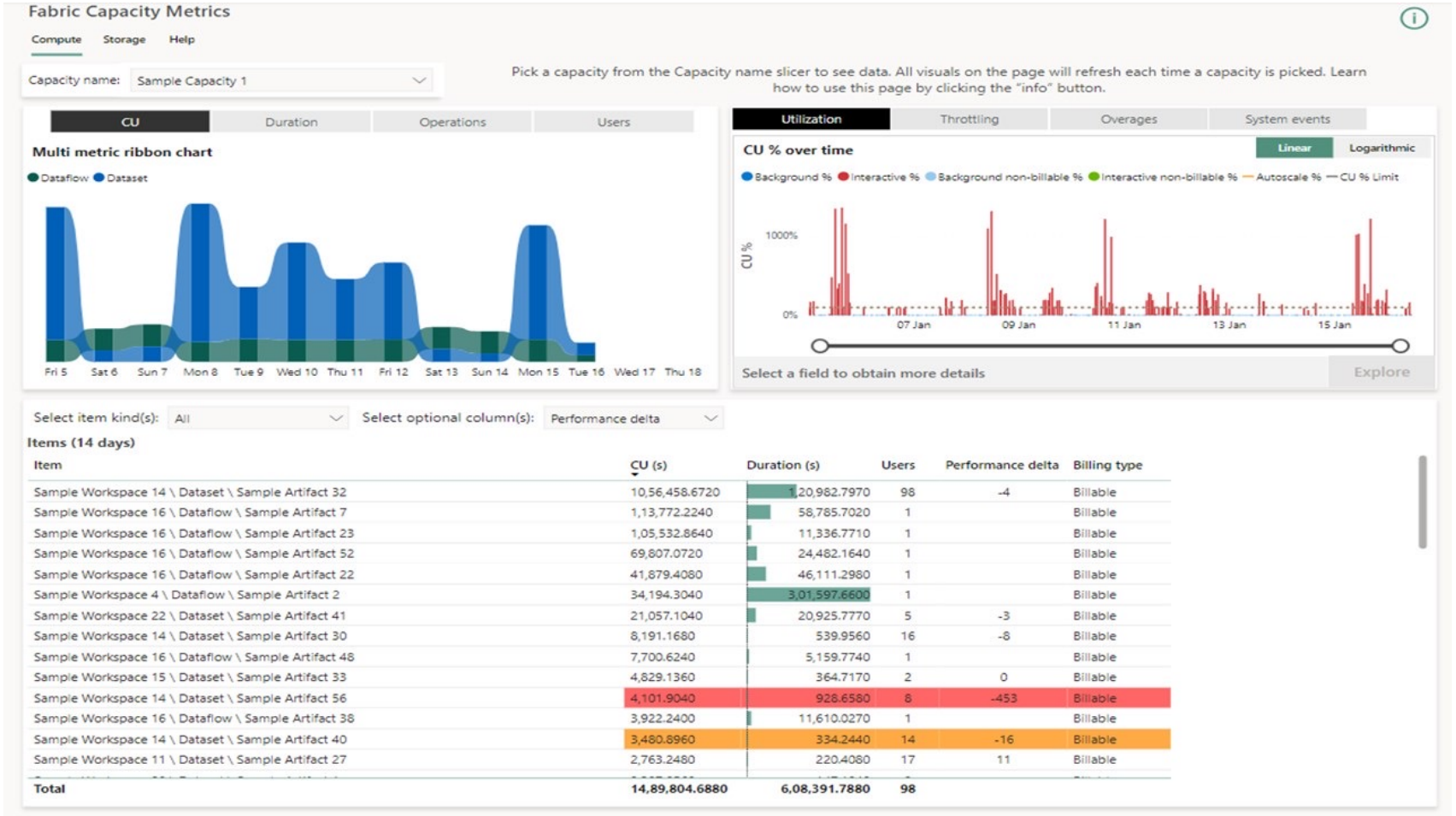


More ideas

Use data from Capacity Metrics App or time-based triggers to:

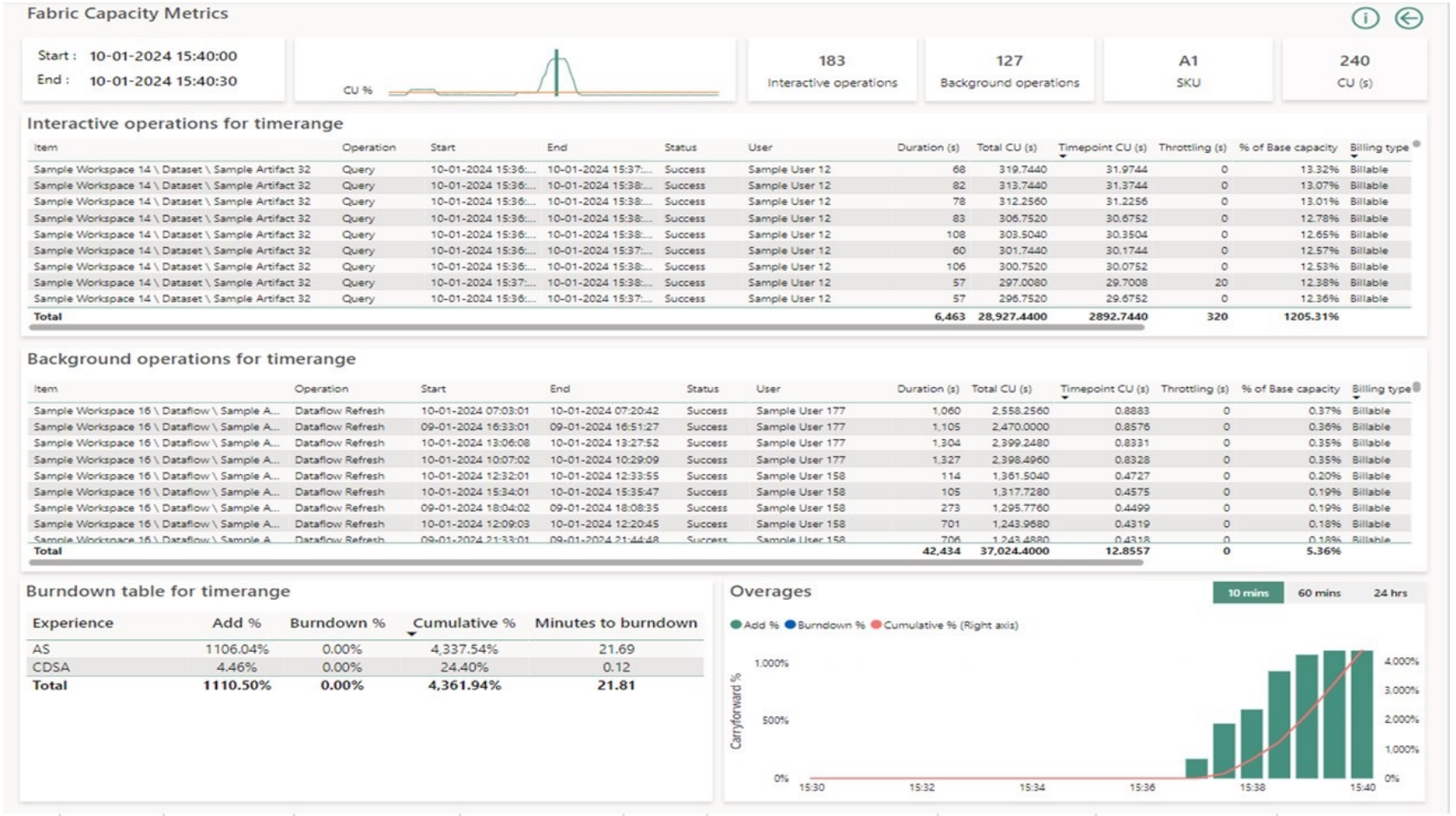
- Automatically scale up/down Capacity
- Automatically move Workspaces to different Capacities
- Automatically pause/resume Capacities
- Automatically create new Capacities

Capacity Metrics App

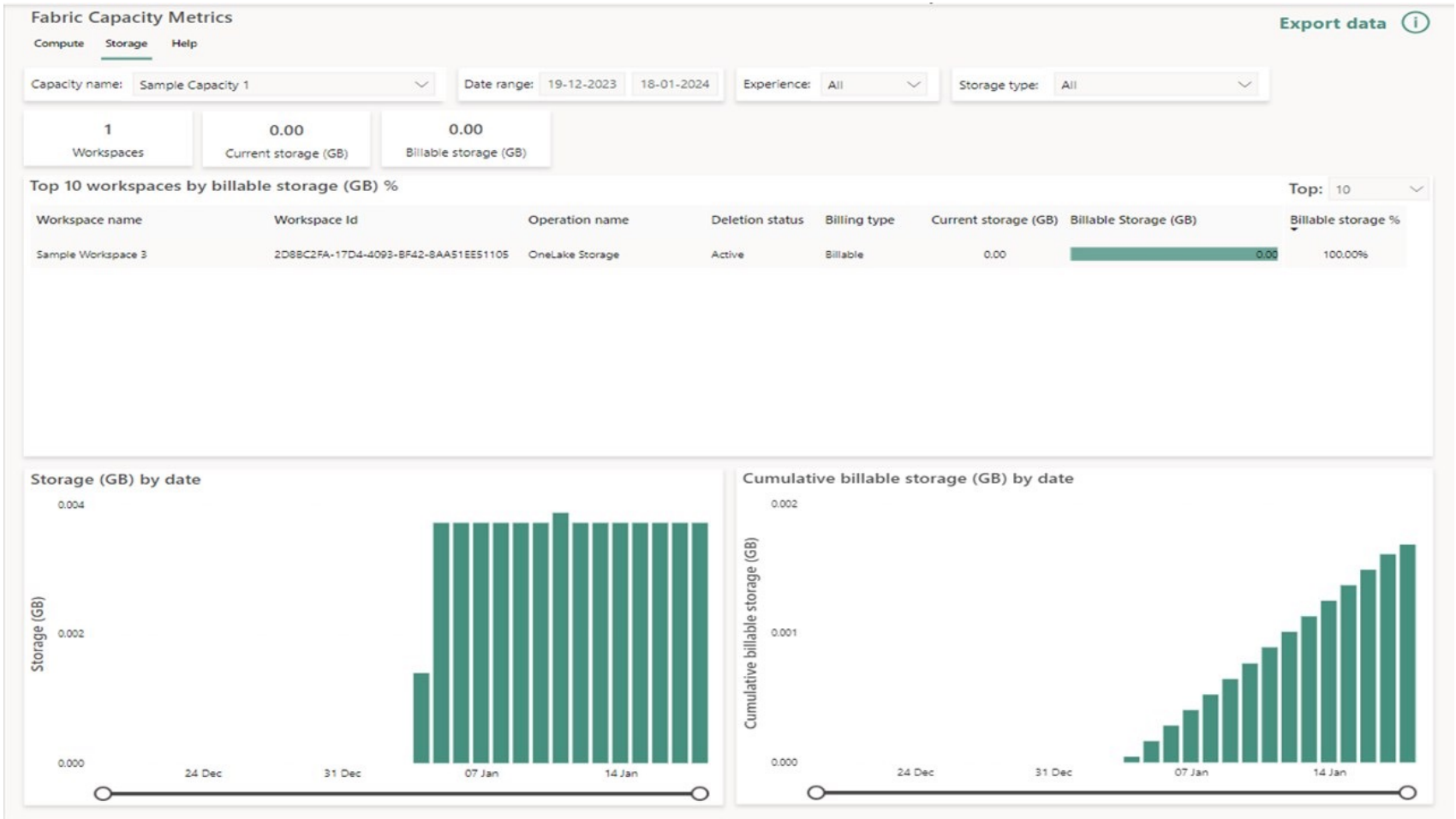


Item	CU (s)	Duration (s)	Users	Performance delta	Billing type
Sample Workspace 14 \ Dataset \ Sample Artifact 32	10,56,458.6720	1,20,982.7970	98	-4	Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 7	1,13,772.2240	58,785.7020	1		Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 23	1,05,532.8640	11,336.7710	1		Billable
Sample Workspace 16 \ Dataflow \ Sample Artifact 52	69,607.0720	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	926.6560	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**		

Capacity Metrics App



Capacity Metrics App



Announced at FabConEurope...

Coming soon: Surge Protection

Enables throttling on background jobs

➡️ background operations get rejected before they consume all the CUs for the next 24 hours

➡️ interactive operations remain available

Useful for Capacities dedicated to interactive usage (Power BI)

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)

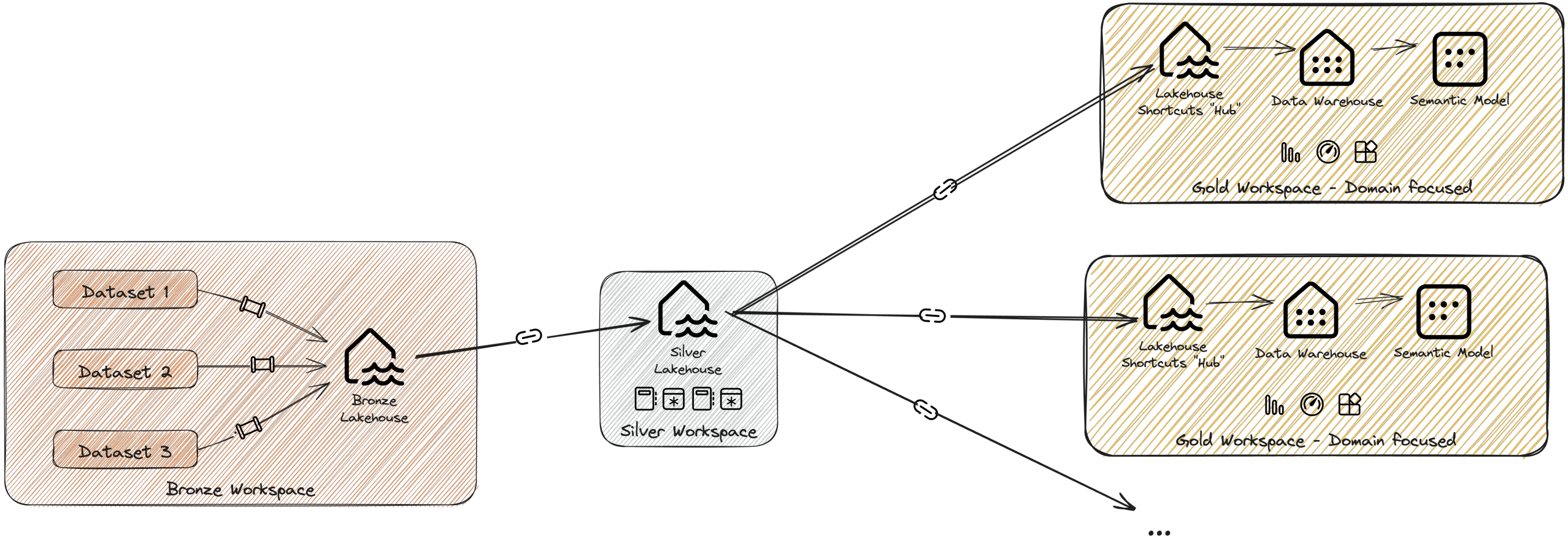
Note: this will probably be improved with the introduction of OneSecurity

Quiz



Recap

RECAP: Medallion layers: bronze, silver, gold



Recap

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

Recap

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

Recap

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



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

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



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

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



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

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

*: Power BI Pro / Premium Per User Workspaces excluded



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

DO split Workspaces and their linked Capacities by workload

- Ingestion
- Processing/transformations (optionally combined with ingestion)
- Ad-hoc exploration & development
- Consumption



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

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

Slides

Slides available at

https://debruyn.dev/dsh24

Thank you sponsors!















Rate Data Saturday Holland

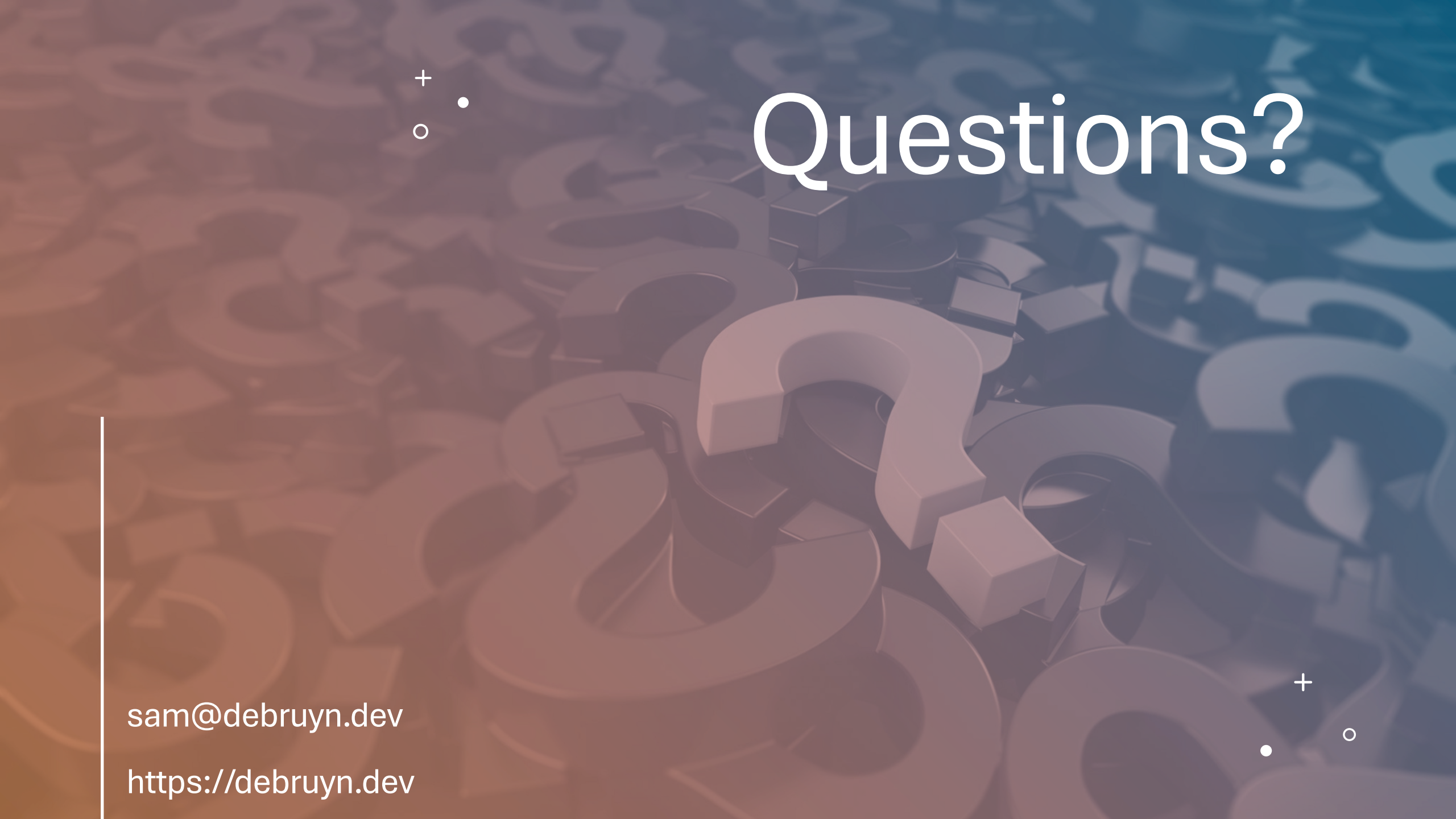
SUPPORT CASPER

1 review = 1 €

Towards beating pancreatic cancer

Rate this session





Questions?

sam@debruyn.dev

https://debruyn.dev