# Understanding Fabric Capacities

Benni De Jagere



Slides





# 17th edition SQLDay Conference



**12-14** May 2025, WROCŁAW + ONLINE

Platinum sponsors





**Gold sponsors** 











Silver sponsors















# Benni De Jagere

**Senior Program Manager | Fabric Customer Advisory Team ( FabricCAT )** 







in

@BenniDeJagere





sessionize /bennidejagere



**#SayNoToPieCharts** 





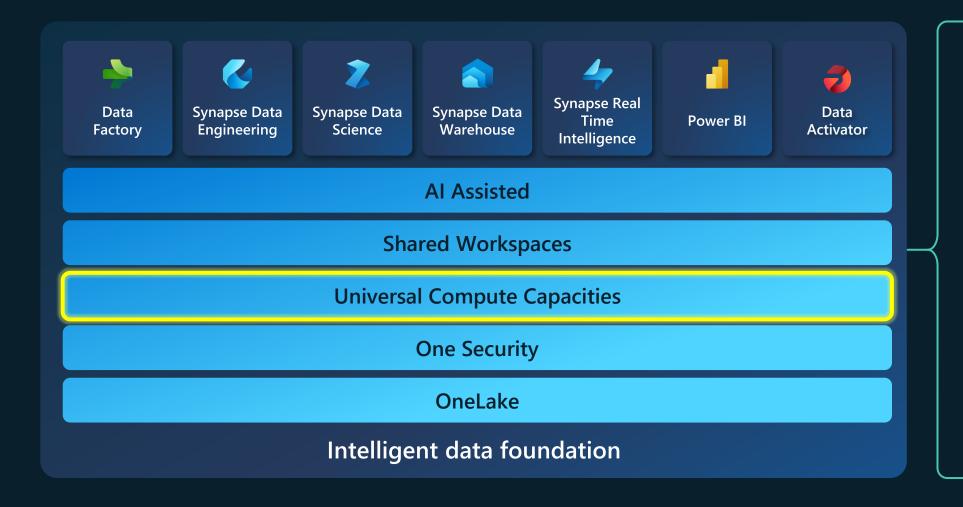




## Fabric Capacities Introduction



### Microsoft Fabric



### Single...

Onboarding and trials

Sign-on

Navigation model

UX model

Workspace organization

Collaboration experience

Data Lake

Storage format

Data copy for all engines

Security model

CI/CD

Monitoring hub

Governance & Capacity Metrics

Data Hub

### Capacities are a shared resource

#### **Shared across workloads**

A single capacity is providing the compute power for all Fabric workloads.

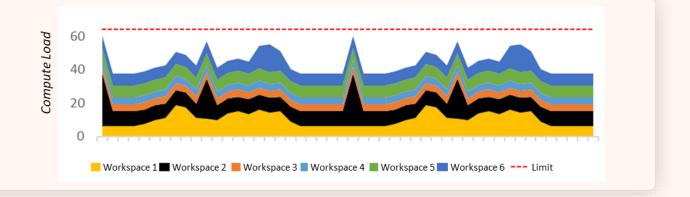
There is no need to allocate compute for each workload separately.



#### **Shared Across Projects**

A single capacity typically supports dozens of separate projects simultaneously, each managed in its own workspace.

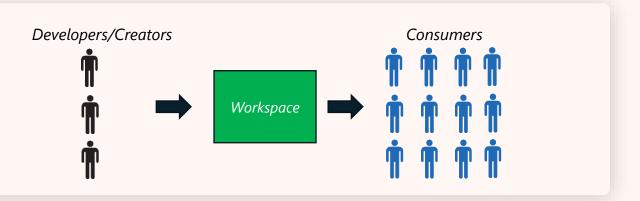
It is rare to have a capacity dedicated to a single project



#### **Shared across users**

For each project, many developers will share a workspace where collaborative development and consumption at scale is managed.

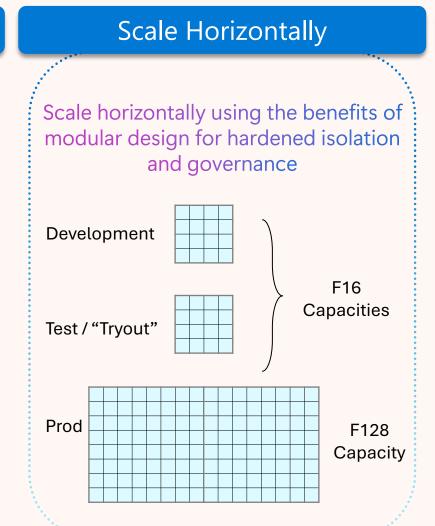
Each creator can provision any artifact and run any job without the need for any pre-approval or planning

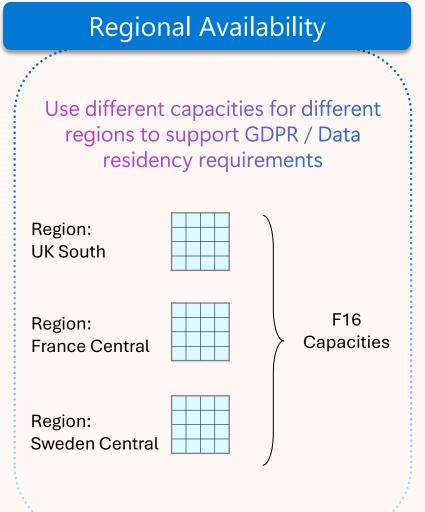


## Capacities are flexible building blocks for growth

Capacities can be configured in endless ways to meet scale, usage and governance requirements while tuning to minimize TCO and performance goals

## Scale Vertically Increased capacity size provides more throughput 8 CU's F8 Capacity F16 Capacity 16 CU's F64 Capacity 64 CU's 2048





## Provisioning and Deploying Capacities

### Purchased in Azure

- **Purchased** either as a PAYG or RI resource
- Provisioned with a certain amount of compute units, analogous to CPU cores.
- The more capacity units are provisioned, the more load the capacity can support
  - Multiply SKU size by 30s to match platform evaluation in metrics app
- Capacities are priced at a fixed hourly rate, based on capacity units provisioned
- The RI commitment (1-year reserved instance) enjoys a 41% discount

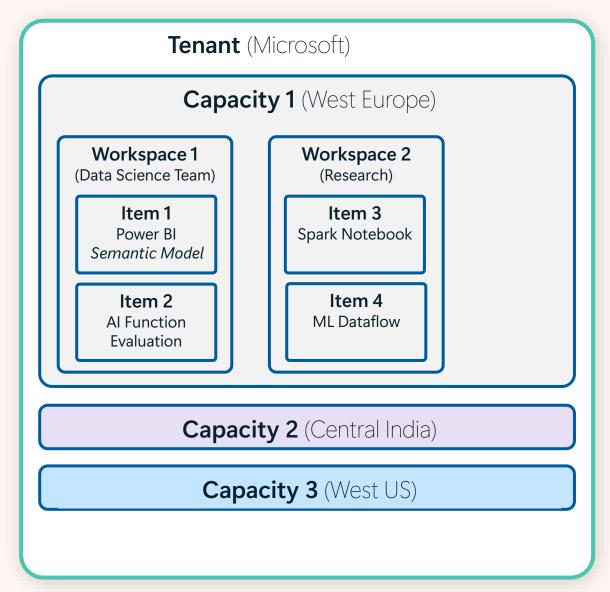
#### Universal Compute Capacities SKU Sizing

SKU	Capacity Units (CU)	CU's (per 30s)	Power BI SKU	Power BI V-cores
F2	2	60	-	0.25
F4	4	120	-	0.5
F8	8	240	A1	1
F16	16	480	A2	2
F32	32	960	A3	4
F64	64	1920	P1	8
F128	128	3840	P2	16
F256	256	7680	Р3	32
F512	512	15360	P4	64
F1024	1024	30720	P5	128
F2048	2048	61440	-	256

## Provisioning and Deploying Capacities

### Deployed to Regions

- Each capacity resides in a specific region of the buyers' choice where both the data & compute reside
- Workspaces are assigned to a capacity that provides the compute and storage for all the workspace artifacts
- Multiple capacities can be purchased, deployed and managed by different owners residing in a single tenant allowing each business unit to pay for their own consumption





Bursting and Smoothing

## Smoothing intro and benefits

Load stabilization

Smoothing helps capacities self-stabilize by flattening large spikey loads into a smooth load profile, eliminating temporal spikes

Eliminates Scheduling contention

Large/scheduled Jobs usage (not execution) are smoothed over 24 hours, eliminating the need to decide the timing and order of job execution

Surge protection

Interactive operations smoothed over several minutes, preventing a single user with a very demanding query from hogging the entire capacity



### What is Bursting?

### Job acceleration

Bursting provides extra compute resources to jobs and queries to accelerate their completion

### Go beyond

The extra resources of bursting allow jobs to utilize far more resources than "face value"

Instead of running a job on 64 CU and completing in 60 seconds, bursting could use 256 CUs to complete the job in 15 seconds.

Same amount of work, just completed faster

### No hassle, No overload

**Bursting is automatic** when the system reasons it can accelerate the job by applying extra resources. No settings are required.

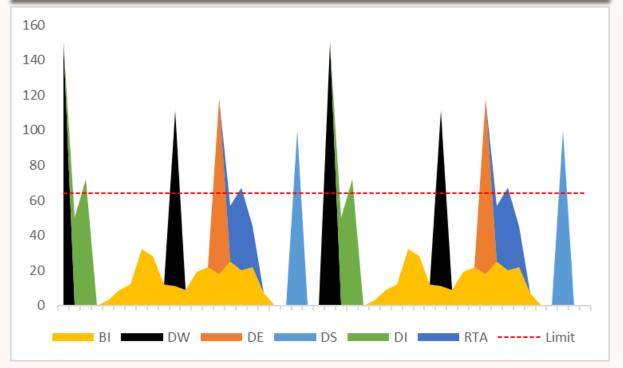
Bursting prevents an overload as the *smoothing* mechanism will always flatten the resource burst

### Bursting and smoothing | before and after

Looking at an example of a 64 CU capacity, running multiple workloads over a couple of days...

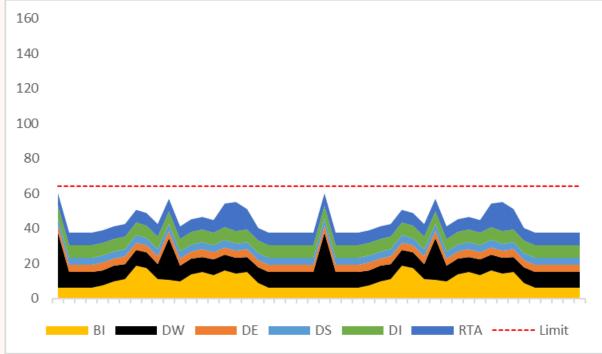
### **Before Smoothing**

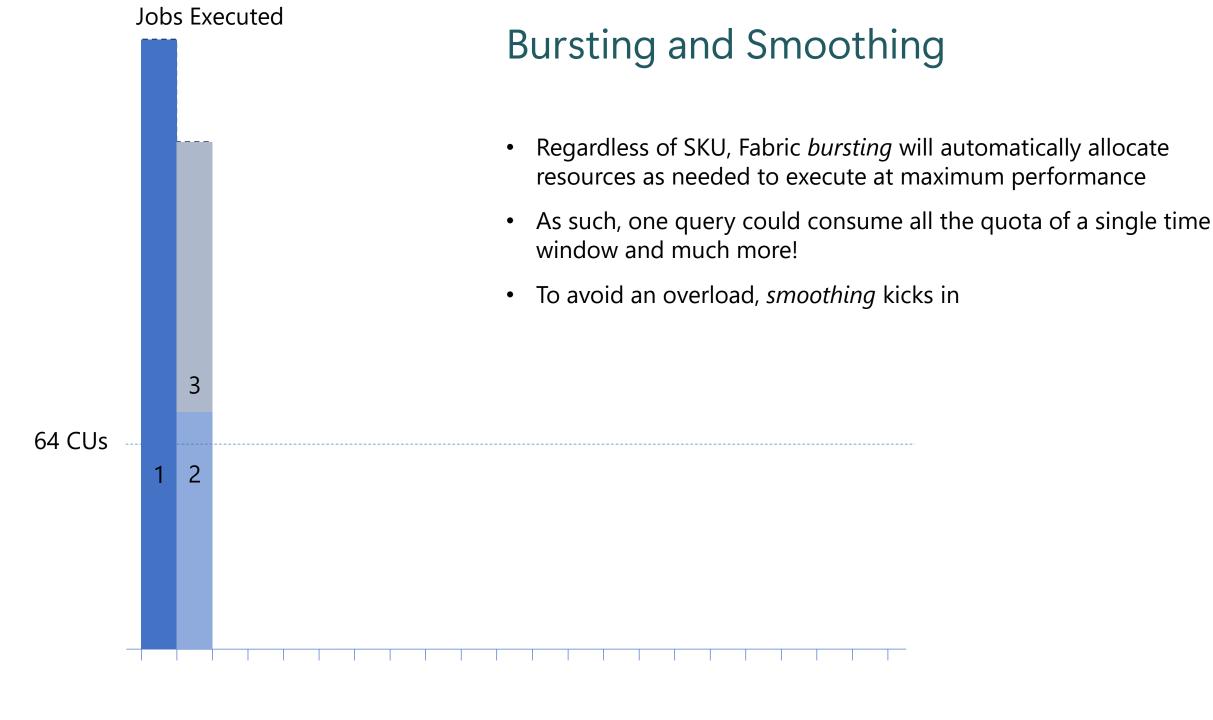
- Actual load as executed on the capacity before smoothing
- Bursting accelerates jobs execution by resource boosting
- The capacity could be overloaded 25% of the time
- Some of the overloads are more than 2x the limit
- There are periods of no/low usage

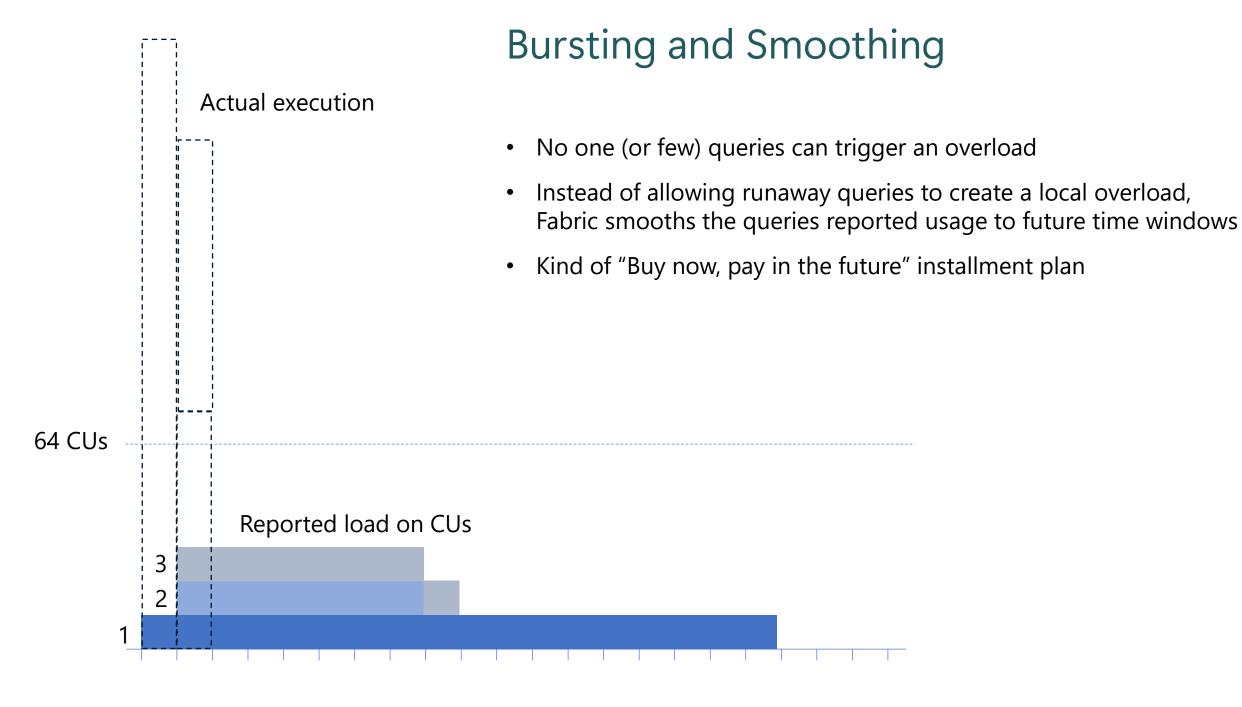


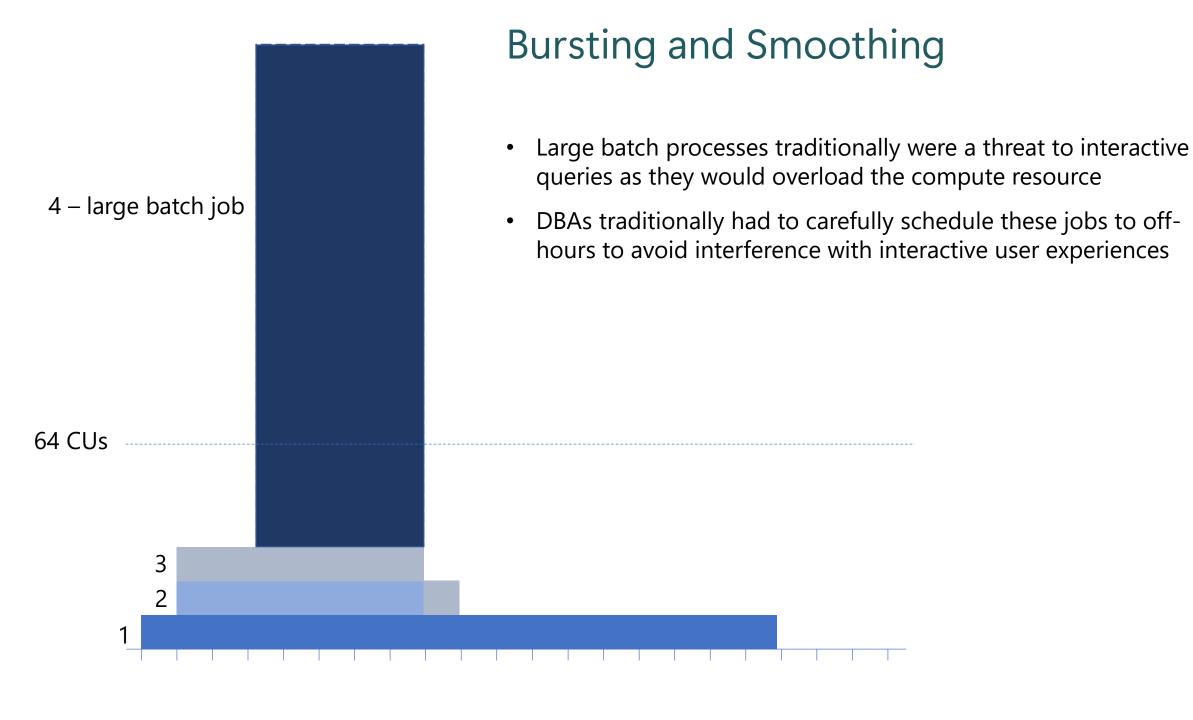
### After Smoothing

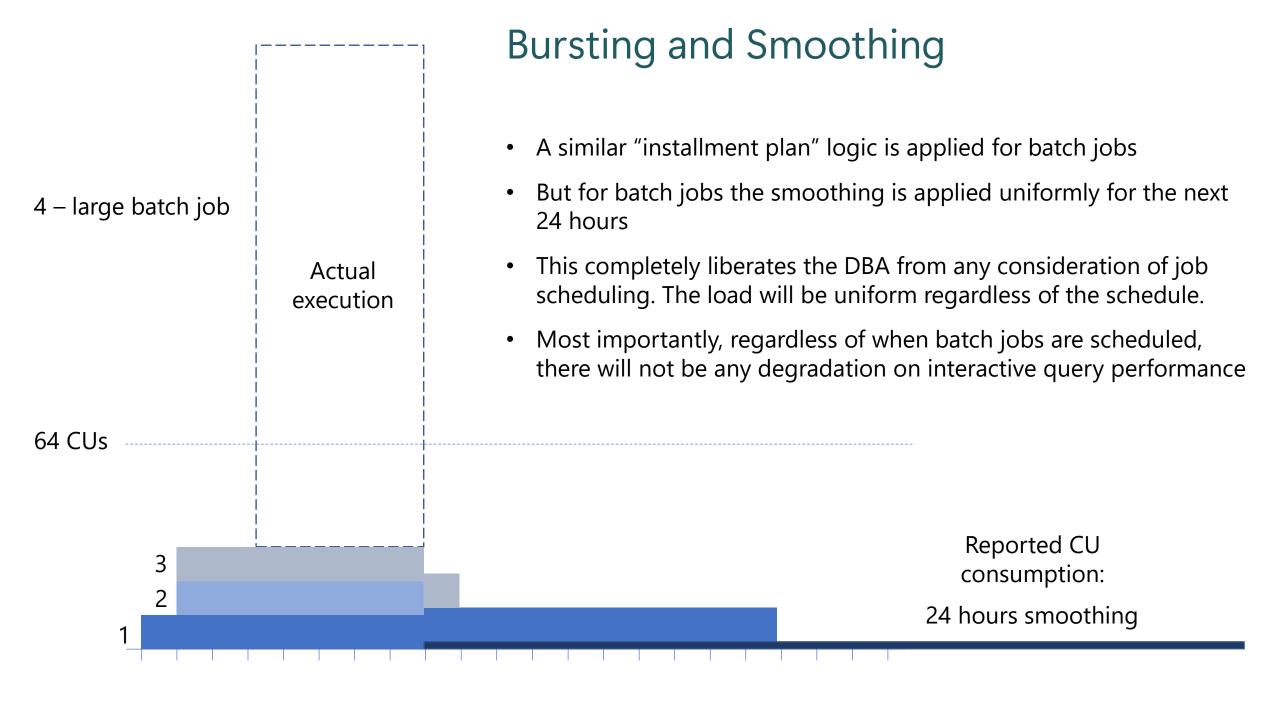
- Shows the reported load (not runtime execution) against the capacity limits
- There is NO overload, and consumption is more stable
- The smoothing of usage fills in gaps











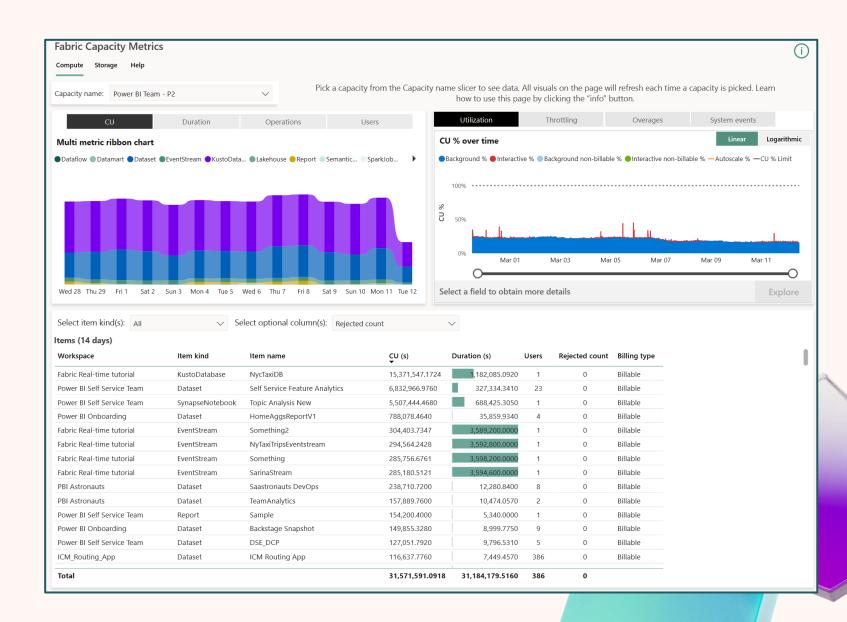


# Monitoring with Capacity Metrics

## **Capacity Metrics**

# Monitor Capacities and Plan capacity scale-up with confidence

- Tenant wide visibility into capacity usage for all Fabric experiences
- Identify resource usage trends and their impact to autoscale & throttling
- View preview workload usage alongside production workloads to make data-driven capacity sizing decisions

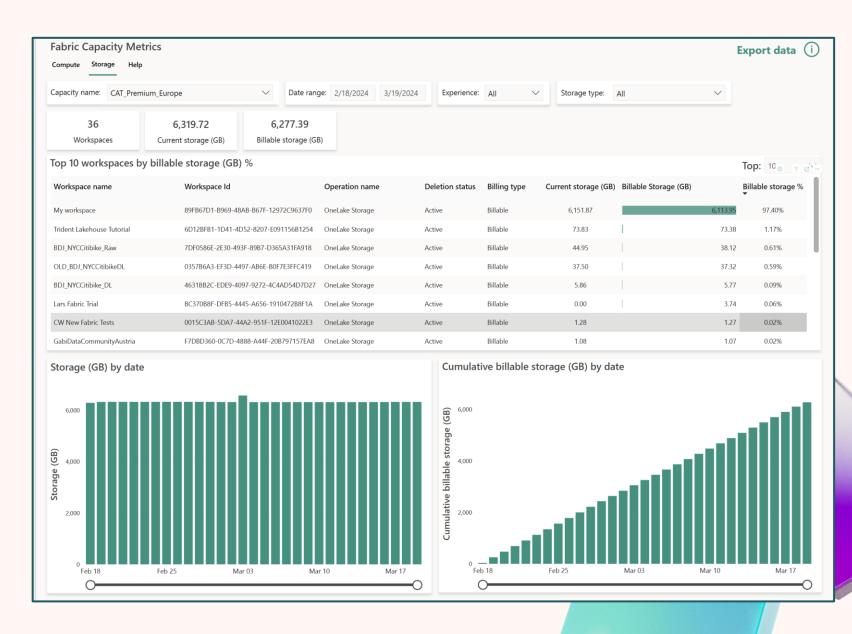


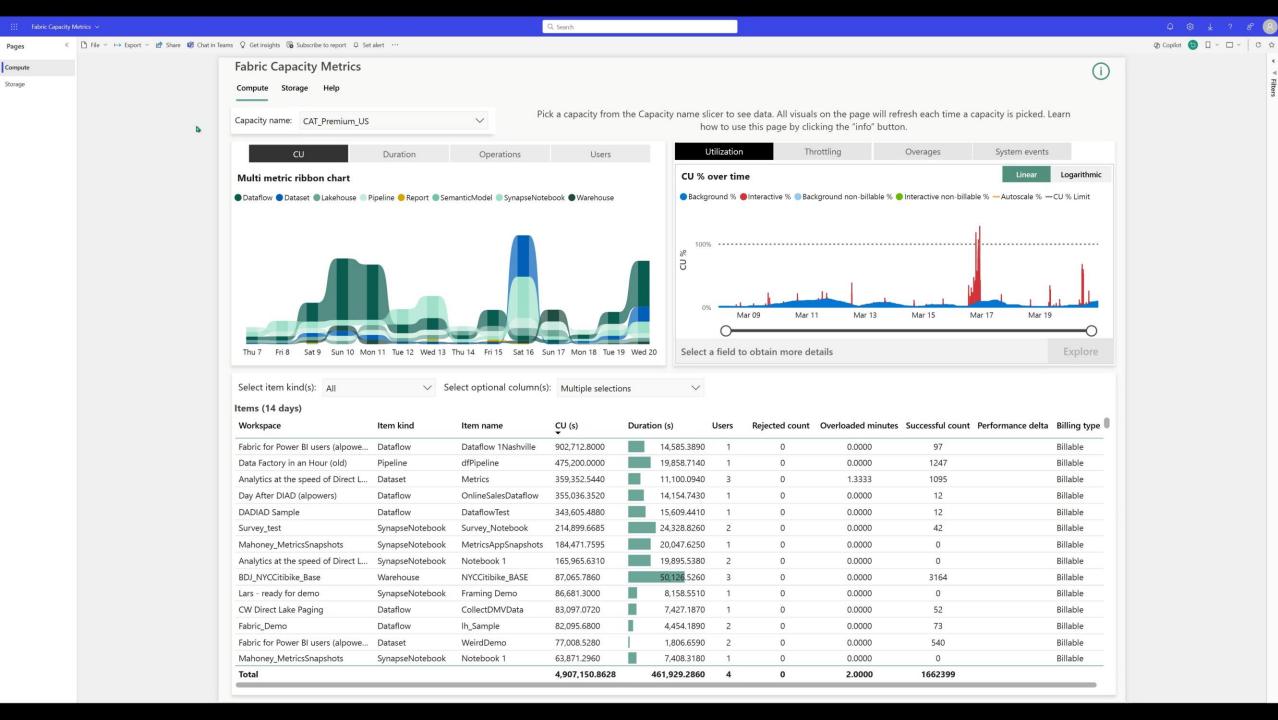
### Capacity Metrics

# Monitor OneLake consumption

Measure the trends of workspace storage consumption against capacity limits, by day or hour

Reconcile costs with internal chargeback processes





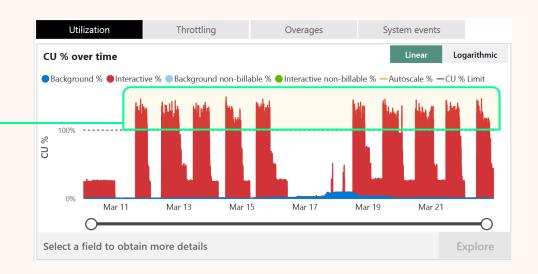


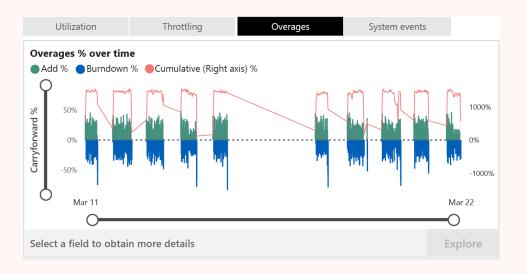
# Capacity Throttling Policies

## Throttling intro

- Throttling is the platform policy for managing consumption that exceeds throughput is provided by SKU choice
- When workloads exceed the throughput of a capacity a cumulative debt is tracked to be burned down
- Cumulative debt is used to determine throttling policies and is burned down when resources are free

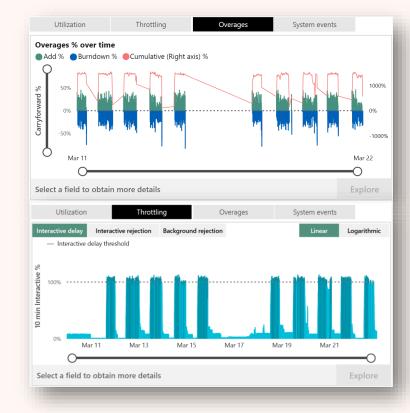
Overage Operation	Description
Overages - Added	<ul> <li>Timepoint when job requests exceed the throughput of a capacity, overages are was added to the cumulative buffer to burn down.</li> <li>This graph simplifies identification of the optimal timepoint to load timepoint drill to analyze the user operations that contributed to an overage.</li> </ul>
Overages - Burndown	<ul> <li>Overages being reconciled when future capacity is free to burn down</li> </ul>
Overages - Cumulative	<ul> <li>The total amount of queued work on the capacity to be burned down in the future when the capacity is not fully utilized</li> </ul>





## Capacity throttling evolution for Fabric

- For Fabric, throttling policies were refined to deliver multiple benefits
  - Reduced throttling for capacities that only experience occasional spikes
  - Added overage protection rejection policies prevent overloaded capacities from irrecoverable overload
  - Optimizations for long-running jobs: We're optimizing the platform for long-running jobs, so if a job exceeds capacity limits, it will run to completion and the overage will be burned down against future capacity



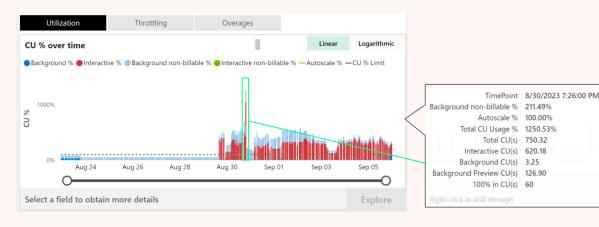
Smoothed Capacity - Future Use	Platform Policy	Customer Impact
<= 10m	Overage Protection	Jobs can consume 10 minutes of future capacity use without throttling
> 10m → <= 60m	Interactive Delay	User requested interactive type jobs will be throttled
> 60m → <= 24h	Interactive Rejection	User requested interactive type jobs will be rejected
> 24h	Background Rejection	User Scheduled background jobs will be rejected from execution



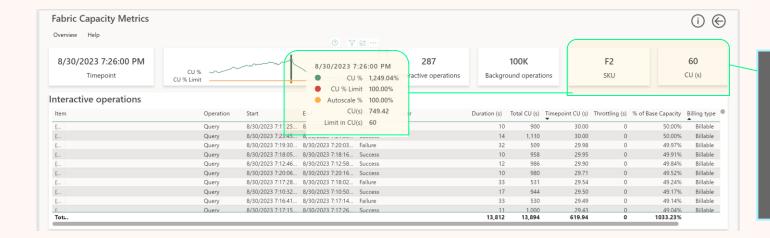
# Capacity Planning with Capacity Metrics

## Capacity planning case study - measurement

Start with a test or trial capacity to evaluate the load of specific Fabric Experiences i.e., Power BI Datasets, Spark Notebooks or a Datawarehouse



If usage is above the current capacity limits, choose the desired utilization rate to accommodate via capacity scale up



Load Capacity Metrics timepoint drill to analyze:

- Total CU's consumed : 749 CU(s)
- Capacity Size : (F2)
- CU(s) available on your capacity : 60 CU(s)

### Capacity planning case study – SKU selection

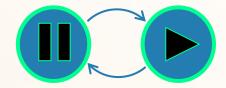
To accommodate a 749 CU(s) load the admin can purchase an F32 capacity providing 960 CU(s) of throughput

#### **Universal Compute Capacities SKU Sizing**

SKU	Capacity Units (CU)	CU's (per 30s)	Power BI SKU	Power BI V-cores
F2	2	60	-	0.25
F4	4	120	-	0.5
F8	8	240	A1	1
F16	16	480	A2	2
F32	32	960	A3	4
F64	64	1920	P1	8
F128	128	3840	P2	16
F256	256	7680	Р3	32
F512	512	15360	P4	64
F1024	1024	30720	P5	128
F2048	2048	61440	-	256



# Pausing and Resuming Capacities



### Pausing and Resuming Capacities

### Why pause capacities?

- 1) It helps manage compute costs.
- 2) It clears any debt that has accumulated. Use it to quickly resolve throttling.

#### What does it do?

Workloads stop execution within 10 minutes of Pause action

New requests are not allowed to Start

Smoothed usage will be reconciled

Note: OneLake storage costs continue to be billed while a capacity is paused



### **How Capacity Pause & Resume works**

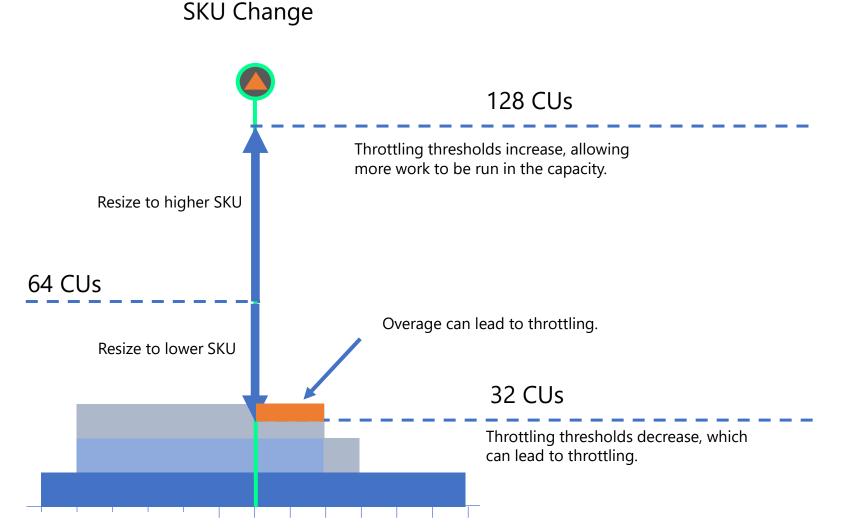
Smoothed usage is **reconciled**. Later, it can be **resumed**. When a capacity is **paused**... Pause event on Capacity The capacity starts Total smoothed usage is shown as with zero utilization or compute utilization on the timepoint smoothed usage. directly after the Pause event. PayGo Price applies to the overage. A billing event is sent for this 64 CUs consumed compute.

### **How Capacity Resize works**

When a capacity is **resized**...

The allowed CUs per timepoint increase or decrease.

This changes the throttling allowed limits based on the new SKU's CUs and the throttling windows.

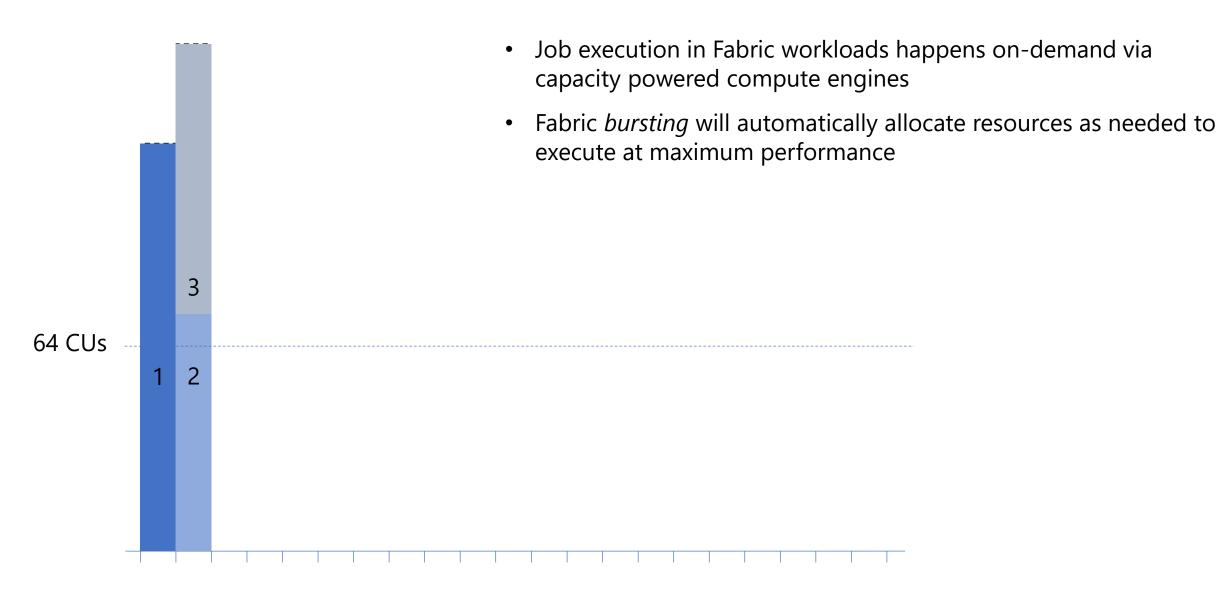


### **Key Insights**

- Sizing up will incur the cost of the new SKU
- Sizing down could lead to more throttling
- Review your Throttling Thresholds before sizing down your SKU.

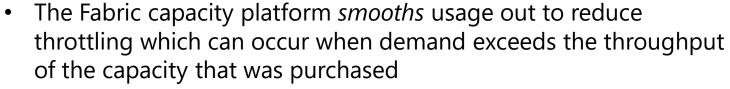
## **Bursting and Smoothing**

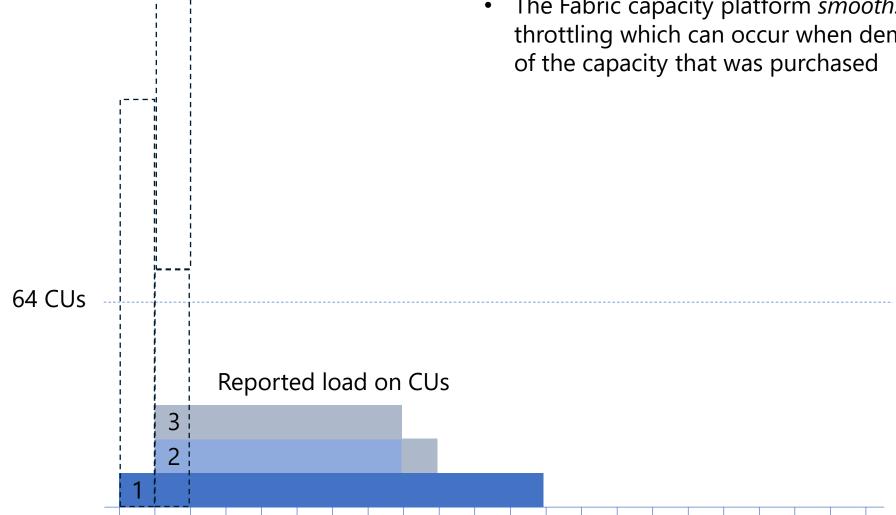


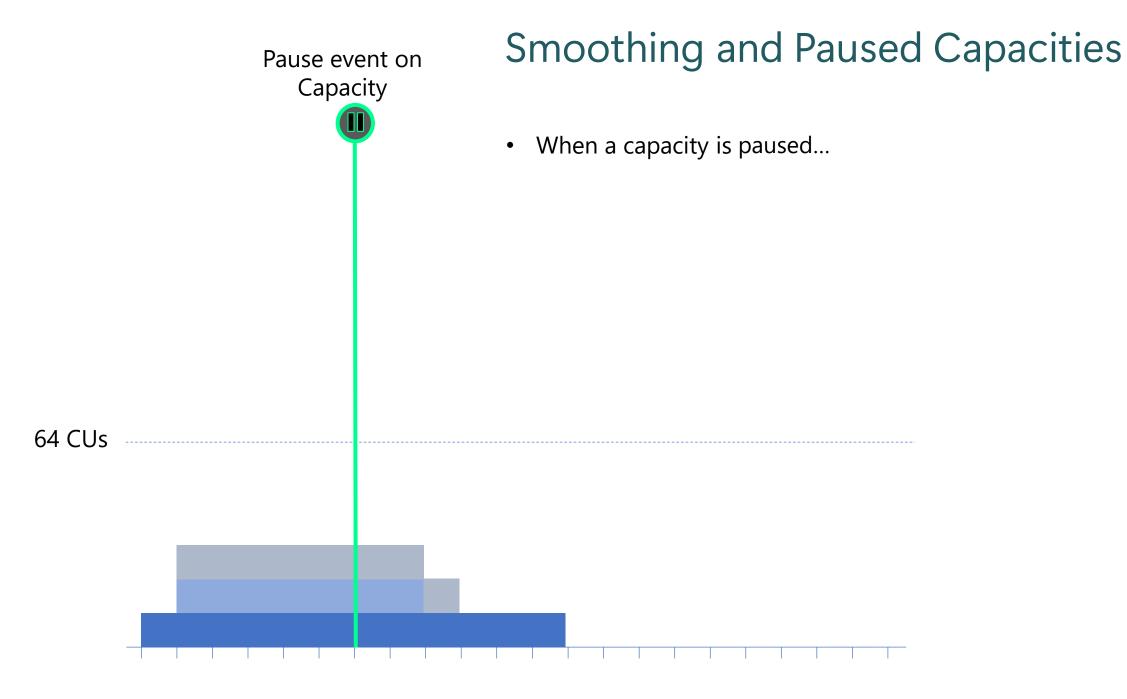


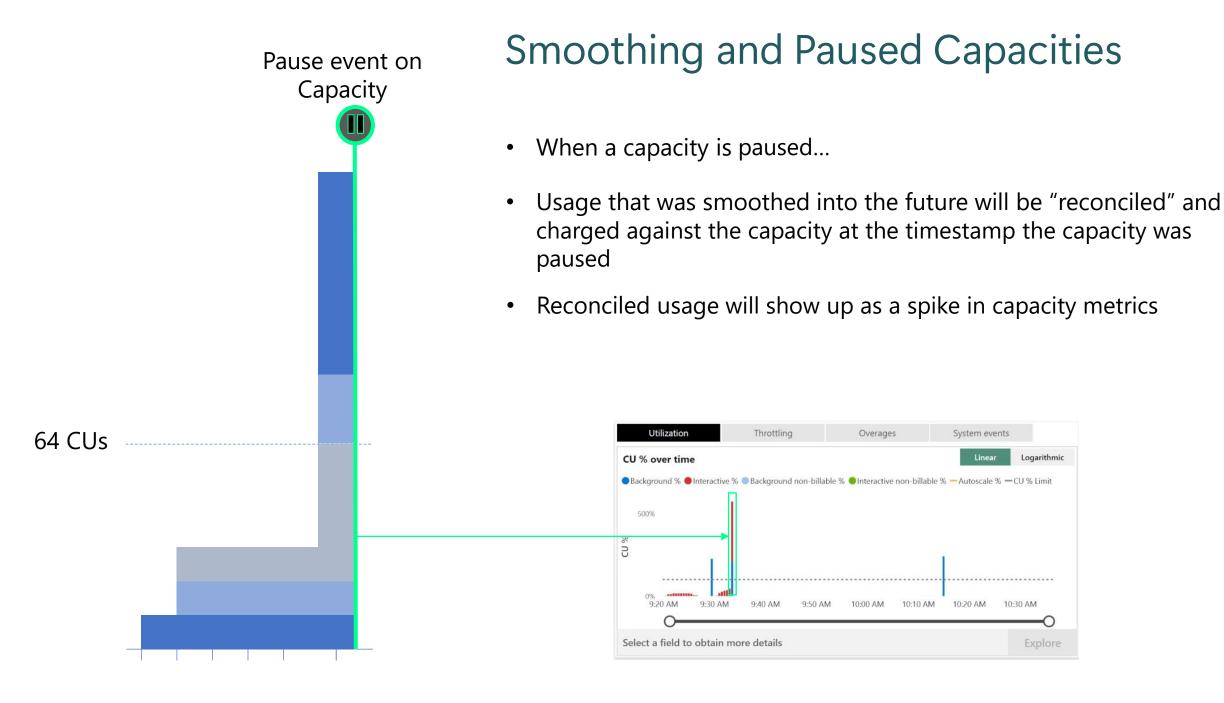


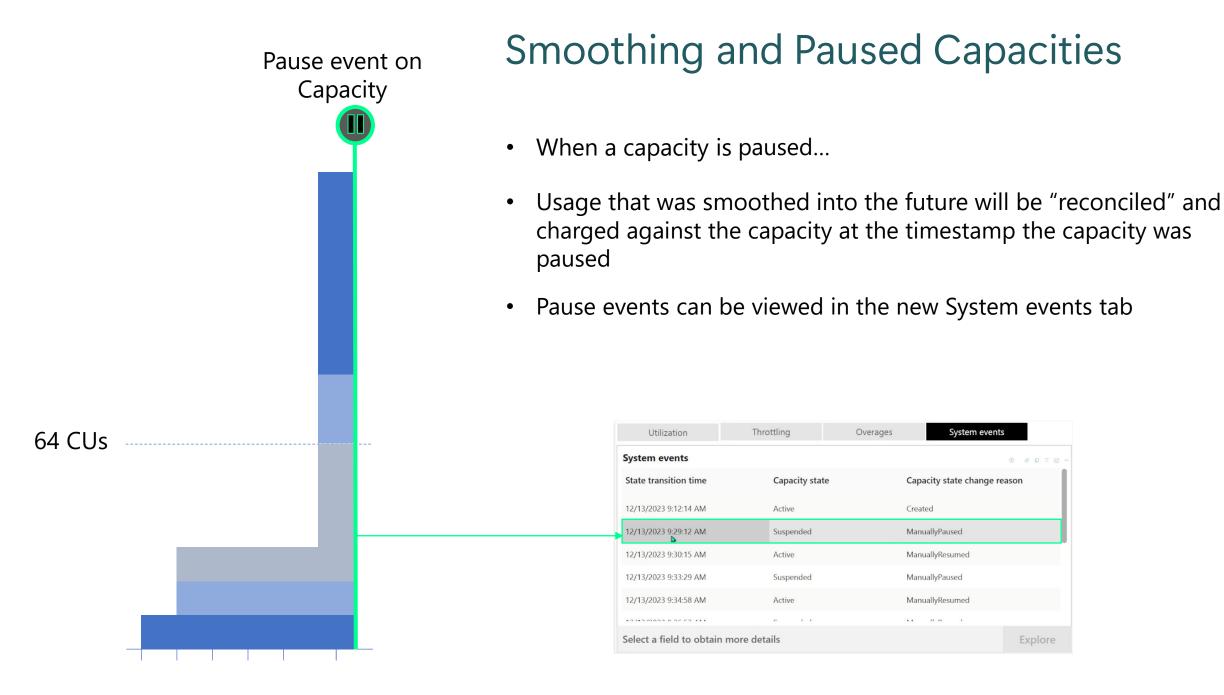
## **Bursting and Smoothing**







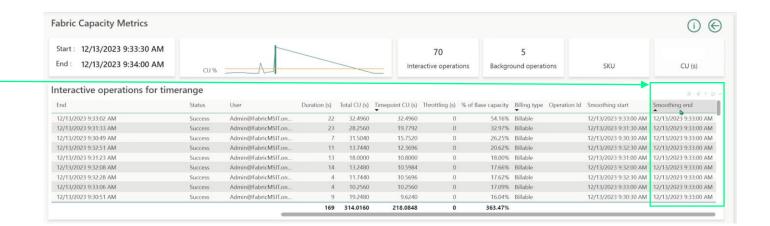




# Pause event on Capacity 64 CUs

## Smoothing and Paused Capacities

- When a capacity is paused...
- Usage that was smoothed into the future will be "reconciled" and charged against the capacity at the timestamp the capacity was paused
- Pause events timestamp is shown in the smoothing end field in timepoint drill views

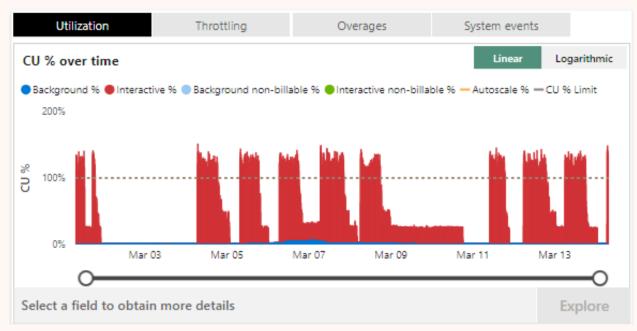


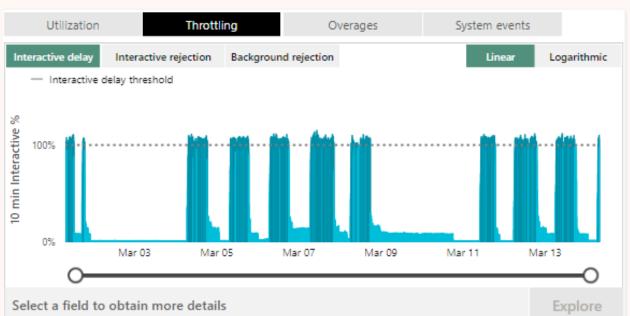
## Microsoft Fabric COMMUNITY CONFERENCE



Bonus: Tips and Tricks for capacity management and monitoring

## My capacity is being throttled! What can I do?





Over 100% utilization doesn't always result in throttling

No penalty until you hit 100% on one of the throttling tabs

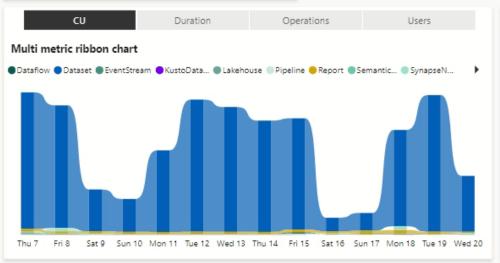
Note: For F SKU, if throttled, you can pause/resume to pay now and clear the carry forward, but that is not a long-term solution

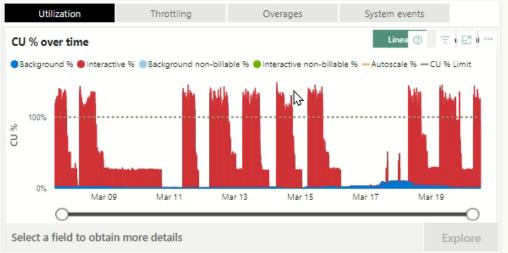
## Metrics **App Tips**

**Fabric Capacity Metrics** Compute Storage

Capacity name: CAT Premium Europe

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.



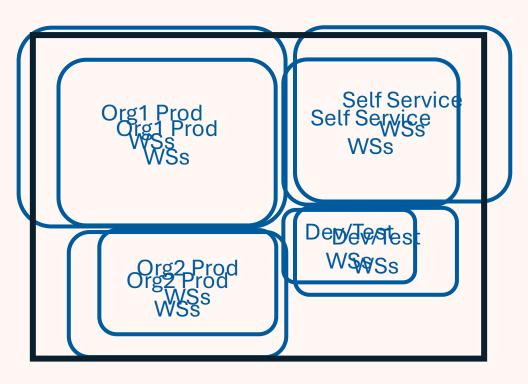


Select optional column(s): Rejected count Select item kind(s): All Items (14 days) Workspace CU (s) Item kind Item name Duration (s) Users 32,451,086,7200 LoadTest Dataset Phils test dataset\_YR\_Max 2,302,896,3680 2 0

Billing type Rejected count Billable CSA PT Report Life expectancy 789,600.0000 14,755.2200 0 Billable 322,122,5920 2 0 BDJ\_NYCCitibike\_StarSchemaAllT... Dataset 00\_NYCCitibike\_FLAT 7,810.0520 Billable 253,104.3680 0 GabiDataCommunityAustria wwilakehouse 12,047.0100 Billable Dataset BDJ\_NYCCitibike\_StarSchemaAllT... Dataset 00\_NYCCitibike\_STAR 133,060,0640 26,935.4120 2 0 Billable 127,017.4050 12,927.6500 2 0 # ankit copilot demo SynapseNotebook Notebook 1 Billable 93,539,2000 # ankit copilot demo Report Power BI Session Service 5,580,0000 0 Billable GabiDataCommunityAustria Lakehouse wwilakehouse 73,768.2224 2.4700 0 Billable 9,025.5060 BDJ\_NYCCitibike\_StarSchemaAllT... Dataset 00\_NYCCitibike\_STAR (Full) 72,103.2960 2 0 Billable DbrowneFabricTest 41,743,9430 4,891.2610 0 SynapseNotebook Notebook 1 Billable 37.969.7683 2 BDJ\_NYCCitibike\_DL NYCCitibike 2,660,9720 0 Billable Lakehouse 35.391.4400 Dataflow CopilotDemoDataflowGen2 1,929,9490 2 0 # ankit copilot demo Billable PBI Monitor - PBICAT Dataflow PBI - Activity Monitor - Dataflow 32,553.7440 1,815.4310 0 Billable BDJ NYCCitibike Raw Pipeline nine BASE NYCCitibike 28 800.0000 2 233,1790 Billable Total 34,896,258.5263 6,028,289.7330 0

Note: Didn't even go to Timepoint Detail page. Useful for usernames and individual operation time/CU

## When Capacity Units Run Out Option 1 – Optimize



WSs = Workspaces

Capacity

### Approach

 Work with content creators to follow best practices and reduce CU consumption

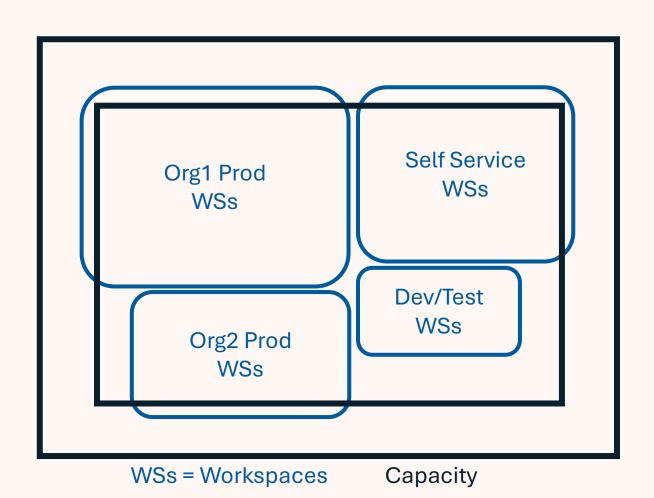
#### Pros

- Avoids increased cost
- Learning carries over to future content

#### Cons

• Can be difficult/time consuming

## When Capacity Units Run Out Option 2 – Scale Up



Options to add compute

- Move to a bigger P SKU or RI F SKU
- Turn on autoscale (P SKU)
- Manual/Dynamic change size (F SKU)

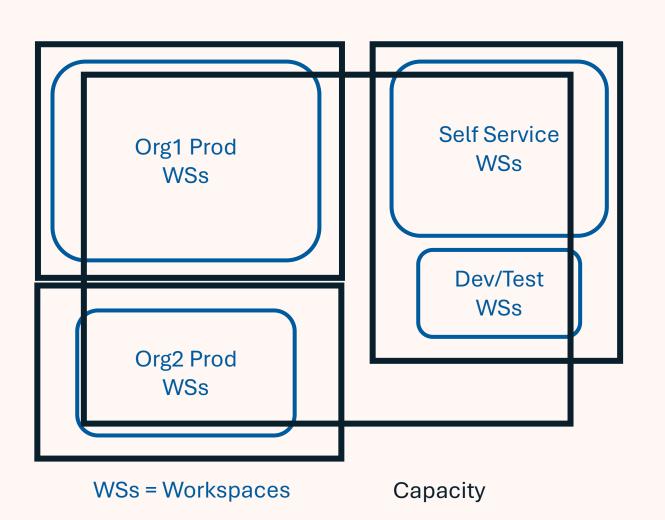
#### Pros

- Add CUs for all items
- Easy

#### Cons

- Cost
- Bad actors (items with unintentionally high CU burn) can still be a problem

## When Capacity Units Run Out Option 3 – Scale Out



### **Options**

 Create multiple smaller P or F SKUs based on organization, type of work, etc.

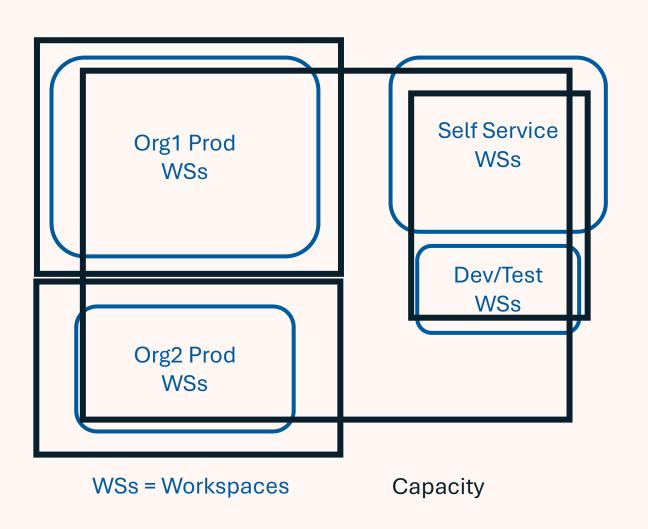
#### Pros

- Easy
- Provides some isolation from bad actors (items with unintentionally high CU burn)
- Flexibility in capacity settings/governance

#### Cons

- Cost
- High CU items have increased chance of throttling

## When Capacity Units Run Out Option 4 – Isolate



## Approach

Provide isolated capacity for key items built by experienced developers

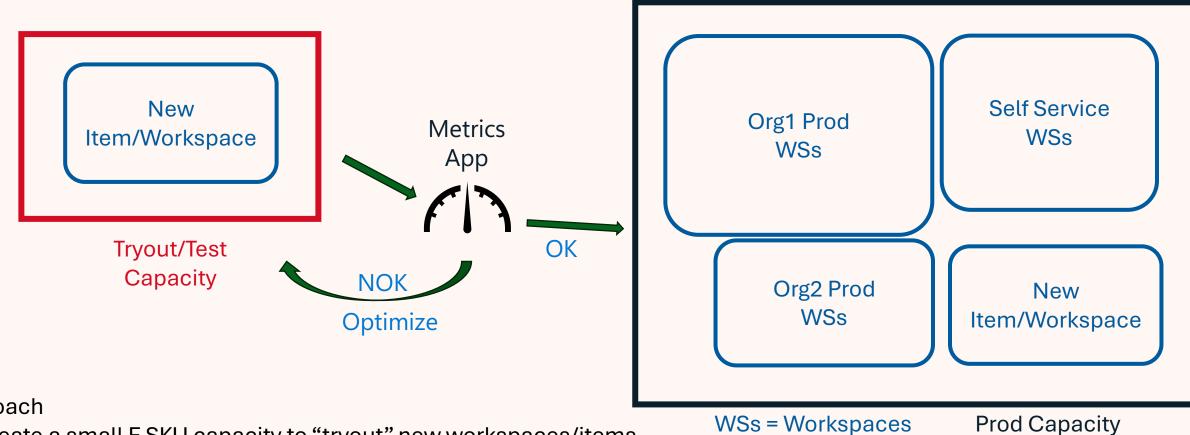
#### Pros

- Easy
- Provides isolation from items built by inexperienced developers and/or rapid unplanned usage growth
- Flexibility in capacity settings/governance

#### Cons

- Cost
- May lead to frustration of lower priority content developers/consumers

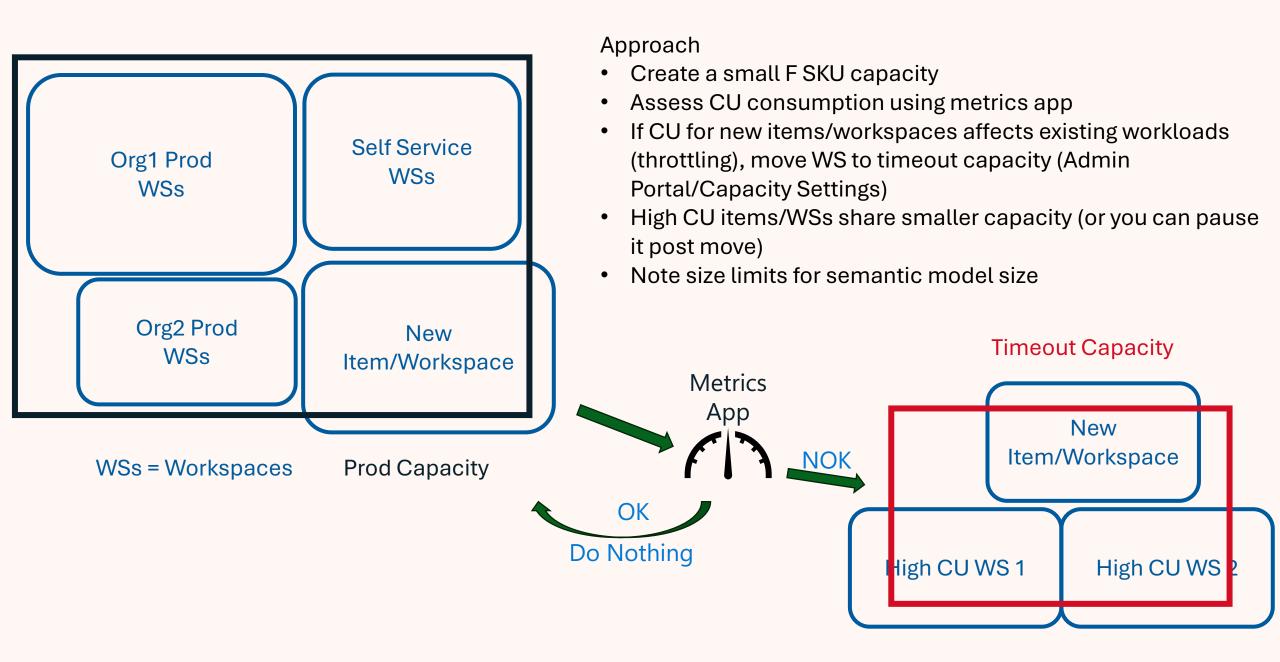
## Isolation Strategy #4a – Tryout Capacity



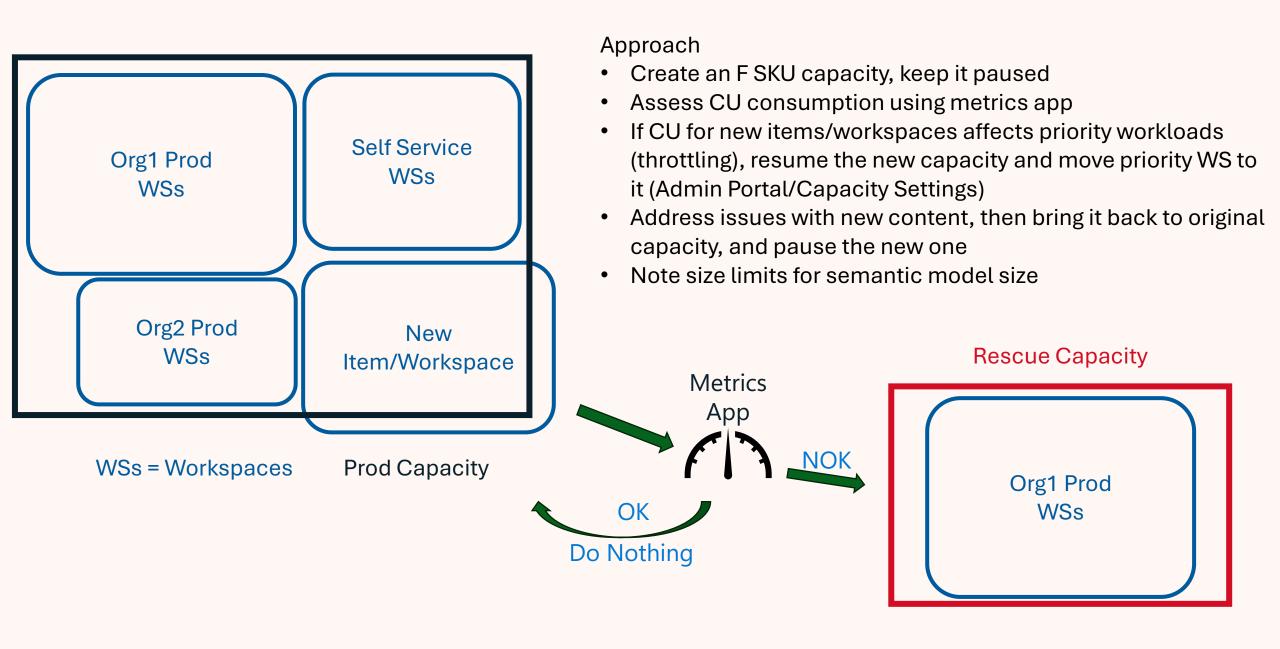
### Approach

- Create a small F SKU capacity to "tryout" new workspaces/items
- Assess CU consumption using metrics app
- If acceptable, move to prod capacity
- If not, optimize
- Pause tryout capacity when not in use, if possible
- Note size limits for semantic model size

## Isolation Strategy #4b – Timeout Capacity



## Isolation Strategy #4c – Rescue Capacity

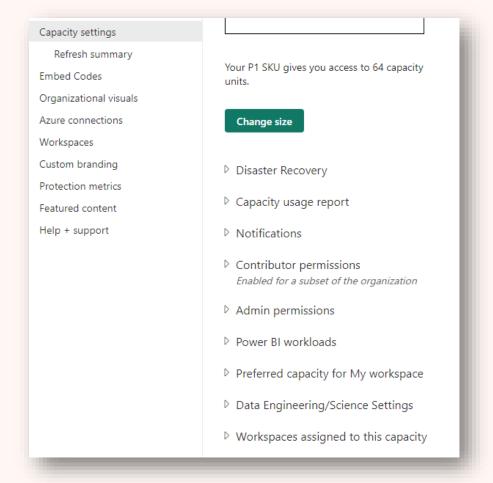


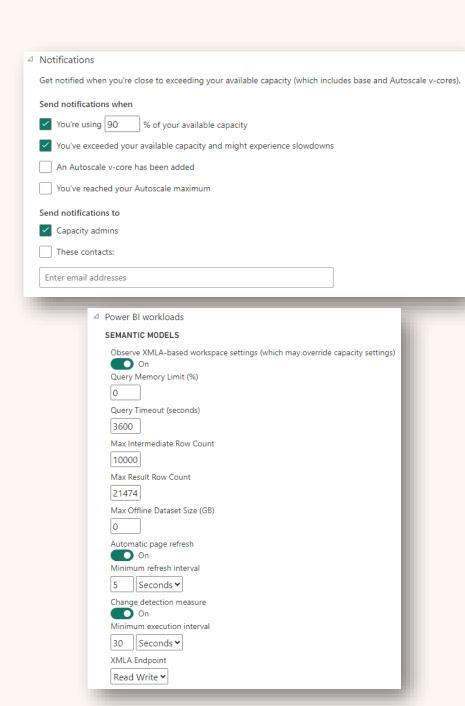
## Recommendations for Cost/CU Savings

- Invest in education, knowledge/best practice sharing, COE, etc. for creators and consumers (proactive optimization)
- Avoid data/report sprawl (leverage certified/promoted models, OneLake shortcuts, etc.)
- Leverage a multi-capacity strategy (isolate, tryout, timeout, etc.)
- Right size your capacities and leverage F SKUs for pause/resume/resize, or reserved instances for discounts
  - Consider a combo of RI and PAYGO (for predictable surge activity)
- Choose the right tool for the job and stay up to date on Fabric feature releases
  - High concurrency mode for notebooks

## Leverage the capacity settings in the UI

- Notifications on CU overuse
- Power BI workloads settings (e.g., query limits, page refresh)





## DIY Autoscale – Fabric Notebook (Bret Myers)

## Set SKU Ranges and Values

```
# Parameters to be passed in from pipeline.
minSku = 'F2' # min sku size we can scale down to
maxSku = 'F128' # max sku size we can scale up to

utilizationTolerance = 90 # Percentage of CU used to st
capacityName = 'fabricbamdemo' #capacity name to be mon
subscriptionId = '### wearing to the state of the state
```

## Get credentials

```
tenantId = mssparkutils.credentials.getSecret('keyVaultEndpoint', 'secretName_tenantId')
clientId = mssparkutils.credentials.getSecret('keyVaultEndpoint', 'secretName_clientId')
secret = mssparkutils.credentials.getSecret('keyVaultEndpoint', 'secretName_clientSecret')

api_pbi = 'https://analysis.windows.net/powerbi/api/.default'
api_azuremgmt = 'https://management.core.windows.net/.default'
```

#### Not all code shown

#### $\underline{FabricTools/CapacityAutoScale\ at\ main\cdot bretamyers/FabricTools\cdot GitHub}$

## Query metrics app model

```
from azure.identity import ClientSecretCredential
     import requests, json, math
     from pyspark.sql.functions import explode
     auth = ClientSecretCredential(tenant id=tenantId, client id=clientId, client secret=secret)
     access token = auth.get token(api pbi)
     header = {'Authorization': f'Bearer {access token.token}', 'Content-type': 'application/json'}
10
     body = {
        "queries":
12
13
            "query": f"""
14
             DEFINE
15
               MPARAMETER 'CapacityID' = "{capacityId}"
16
17
               VAR DS0FilterTable =
18
                 FILTER(
19
                   KEEPFILTERS(VALUES('TimePoints'[TimePoint])),
                    'TimePoints'[TimePoint] >= NOW() - 1
20
21
22
23
               VAR __DS0FilterTable2 = TREATAS({{"{capacityId}"}}, 'Capacities'[capacityId])
24
25
               VAR __DS@Core =
26
                 SELECTCOLUMNS (
                   KEEPFILTERS(
28
                     FILTER(
29
                       KEEPFILTERS(
                         SUMMARIZECOLUMNS(
30
31
                            'Capacities'[capacityId],
                            'Ttame'[Billable tune]
3.2
```

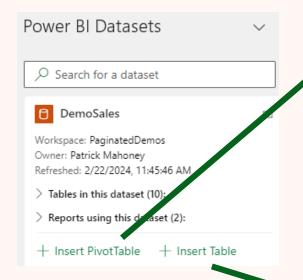
## Change SKU Size

## Most Common Capacity Issues (Power BI)

<b>Bad Practice</b>	Recommendations/Typical Resolution
Model issues (M:M, bi-di, snowflake, etc.) and/or inefficient DAX	Follow best practices (e.g., BPA), star schema
Too many visuals	Multi card, small multiples, Deneb, PowerPoint background, etc.
Big single visual (i.e., matrix with lots of rows, columns, and/or measures)	Improve report design (e.g., drillthrough, apply all Slicers, report page tooltip), field parameters, calc group guardrails, etc.
Complex RLS	Remodel to enable simple filter like Table[Email] = USERPRINCIPALNAME()
Very high concurrency	Optimize reports, DAX, etc. (big multiplier) Consider QSO, data subsets
Direct Query	Switch to import or Direct Lake, if possible. Aggregations, hybrid tables, etc.
Analyze in Excel	Automate downstream analytics with a Power BI report instead, subscriptions, DAX connected table, slicers/measures first, etc.
Excessive refresh	Don't "break the fold", incremental refresh, reduce frequency, optimize M code

## Save Those CUs – Getting Data Into Excel

## Analyze in Excel



#### **Connected Table**

#### Key Takeaways

- How you build it matters
  - Filters & measures first!
- This shows durations but it's CU that matters (test your use cases/models)
- Opt for DAX Connected Tables
  - Create pivot table from that, if needed

## X Rows, Measures, Filter

StartTime	Type	Duration	User	Database	Query
11:49:30	MDX	2,328ms	Power Bl	DemoSales	SELECT {[Measures].[To
11:49:26	MDX	0ms	Power Bl	DemoSales	SELECT {AddCalculated
11:49:23	MDX	0ms	Power Bl	DemoSales	SELECT {AddCalculated
11:49:17	MDX	1,875ms	Power Bl	DemoSales	SELECT {[Measures].[To
11:49:03	MDX	4,469ms	Power Bl	DemoSales	SELECT {[Measures].[To
11:48:54	MDX	3,938ms	Power Bl	DemoSales	SELECT {[Measures].[To

## Filter, measures, rows

StartTime	Туре	Duration	User	Database	Query
10:06:13	MDX	1,625ms	Power Bl	DemoSales	SELECT {[Measu
10:06:03	MDX	781ms	Power Bl	DemoSales	SELECT {[Measu
10:05:49	MDX	109ms	Power Bl	DemoSales	SELECT {[Measu
10:05:46	MDX	312ms	Power Bl	DemoSales	SELECT {[Measu
10:05:43	MDX	234ms	Power Bl	DemoSales	SELECT FROM [N
10:05:14	MDX	0ms	Power Bl	DemoSales	SELECT {AddCale

#### Refresh (same for both)

StartTime	Type	Duration	User	Database	Query
11:50:30	MDX	2,234ms	Power Bl	DemoSales	SELECT {[Measures].[To

## X Rows, Measure, Filter

StartTime	Type	Duration	User	Database	Query
01:28:50	DAX	31ms	Power Bl	DemoSales	DEFINE VAR _C
01:28:41	DAX	1,516ms	Power Bl	DemoSales	DEFINE VAR _C
01:28:40	DAX	16ms	Power Bl	DemoSales	DEFINE VARC
01:28:34	DAX	156ms	Power Bl	DemoSales	DEFINE VAR _C
01:28:33	DAX	16ms	Power Bl	DemoSales	DEFINE VARC
01:28:31	DAX	0ms	Power Bl	DemoSales	DEFINE VAR _C
01:28:30	DAX	141ms	Power Bl	DemoSales	DEFINE VARC
01:28:15	DAX	2,047ms	Power Bl	DemoSales	DEFINE VAR _C
01:28:11	DAX	1,797ms	Power Bl	DemoSales	DEFINE VAR _C
01:28:08	DAX	594ms	Power Bl	DemoSales	DEFINE VAR _C
01:27:56	DAX	281ms	Power Bl	DemoSales	DEFINE VARC
01:27:50	DAX	16ms	Power Bl	DemoSales	DEFINE VARC

## Filter, measures, rows

StartTime	Type	Duration	User	Database	Query
09:14:20	DAX	16ms	Power Bl	DemoSales	DEFINE VAR _DS0Filte
09:14:07	DAX	1,000ms	Power Bl	DemoSales	DEFINE VAR _DS0Filte
09:14:02	DAX	1,188ms	Power Bl	DemoSales	DEFINE VAR _DS0Filte
09:13:59	DAX	594ms	Power Bl	DemoSales	DEFINE VAR _DS0Filte
09:13:51	DAX	531ms	Power Bl	DemoSales	DEFINE VAR _DS0Filte
09:13:50	DAX	0ms	Power Bl	DemoSales	DEFINE VAR DS0Cor

## Refresh (same for both)

StartTime	Type	Duration	User	Database	Query
11:54:49	DAX	1,969ms	Power Bl	DemoSales	DEFINE VARDS0FilterTable = TREATA



# Slides

