Abstract; to make sure you're in the right room

Have you tried to find the most effective way to ingest and process your data?
Were you just as confused as I was after reading all the marketing materials?
Would you like to gain more insight into maximising your capacity?

These are questions I've had. And ones I regularly try to help customers with.

In this session, I'll help you learn the differences between Lakehouse, SQLDB, and Warehouse performance, processing speed and cost. I'll also demonstrate how the Realtime stack can serve as a viable alternative for ingestion and processing flat files in terms of speed, scalability and ease of use.
After this session, you'll walk away with real-life insights from Fabric and knowledge of which buttons not to press.

If you've attended the previous instalment of this session, you'll know I won't do my testing with AdventureWorks. It will be a lot of (messy) data to give Fabric a run for its capacity units!

C2-Vertrouwelijk


Everything is proceeding as I have foreseen.

Loadtesting Fabric, part II

The sequel

Reitse Eskens

C2-Vertrouwelijk

When stuff goes off the rails



What happened previously



The main findings

Lakehouses were very good

Warehouses were very slow and expensive

Dataflows Gen2 only with FastCopy



F2 or F64

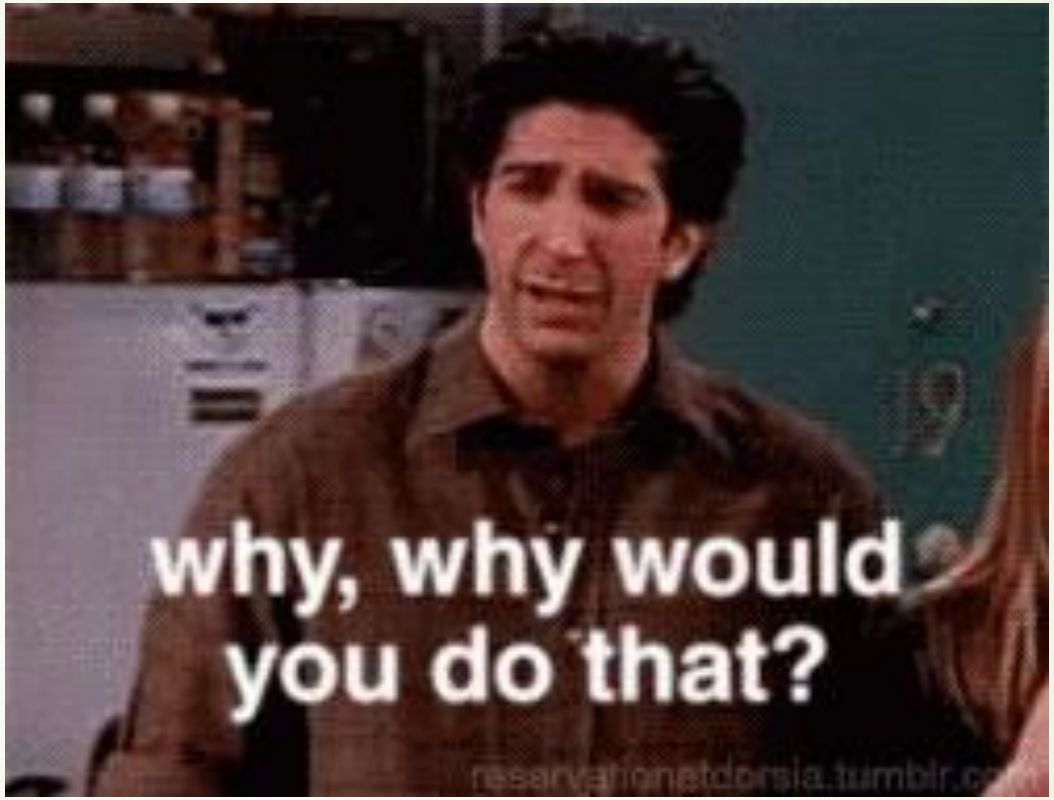
- With a lot of data, big differences in processing
  - Mostly time
  - Sometimes capacity exhaustion

Let’s do it again!

C2-Vertrouwelijk

Break… I mean loadtesting Fabric, again

Why



Because reasons



WHEN WILL IT BREAK



HOW WILL IT BREAK



WHAT'S THE MOST EXCITEMENT FOR YOUR EURO

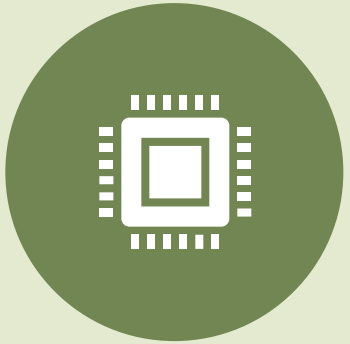
No happy flows



(AZURE) SQL DB IS QUITE EASY

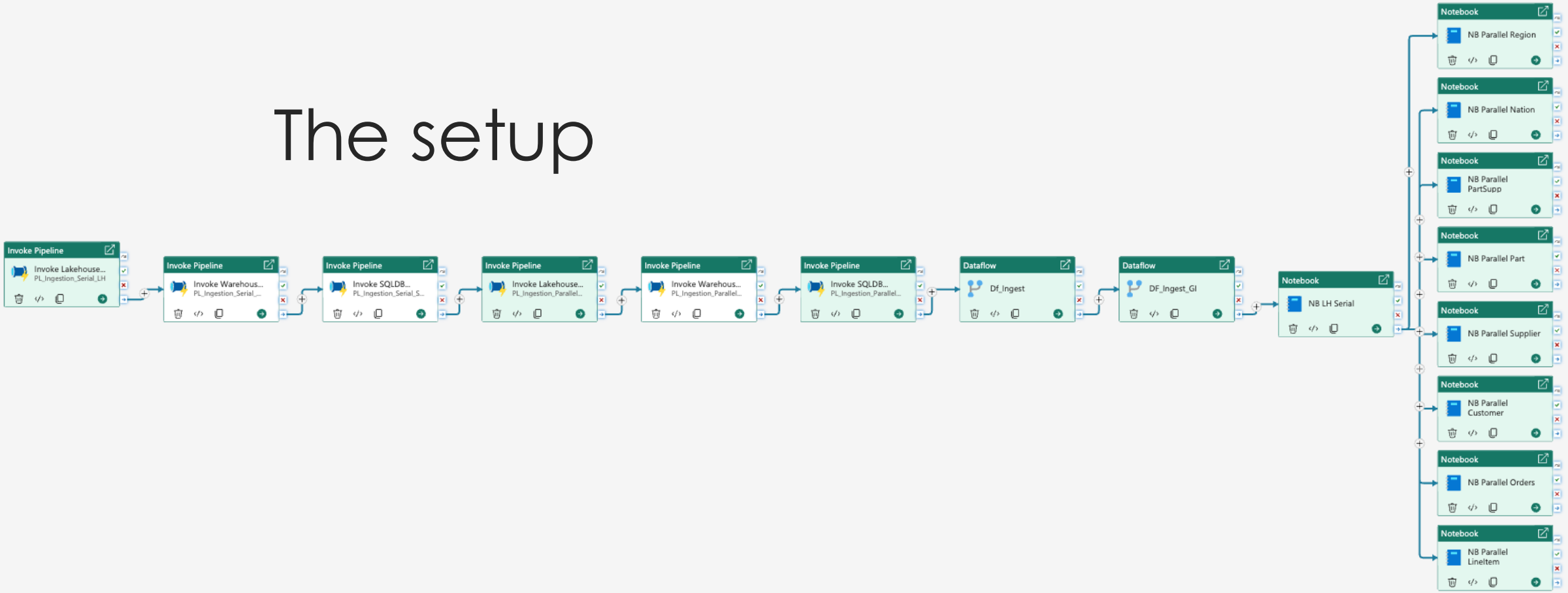


ADVENTUREWORKS DATABASE IS TOO SMALL



FREE API'S ARE USUALLY VERY SIMPLE

The setup


Invoke Pipeline
Invoke Lakehouse...
PL_Ingestion_Serial_LH
Invoke Pipeline
Invoke Warehous...
PL_Ingestion_Serial_...
Invoke Pipeline
Invoke SQLDB...
PL_Ingestion_Serial_S...
Invoke Pipeline
Invoke Lakehouse...
PL_Ingestion_Parallel...
Invoke Pipeline
Invoke Warehous...
PL_Ingestion_Parallel...
Invoke Pipeline
Invoke SQLDB...
PL_Ingestion_Parallel...
Dataflow
Df_Ingest
Dataflow
DF_Ingest_GI
Notebook
NB LH Serial
Notebook
NB Parallel Region
Notebook
NB Parallel Nation
Notebook
NB Parallel PartSupp
Notebook
NB Parallel Part
Notebook
NB Parallel Supplier
Notebook
NB Parallel Customer
Notebook
NB Parallel Orders
Notebook
NB Parallel Lineitem

The setup



ONE MASTER PIPELINE



PARALLEL AND SERIAL DATA INGESTION



LAKEHOUSE, WAREHOUSE, FABRIC SQL DB AND EVENTHOUSE



COPY ACTIVITY, NOTEBOOKS, DATAFLOWS GEN2, EVENTSTREAM

The data



500ISH GIGABYTE OF TPC-H DATA



SMALL FILES (FEW KB) TO LARGE ONES (75 GB)



ALL CSV, AS GOOD AS IT GETS

Metrics



FABRIC SQL DB DASHBOARD



FABRIC CAPACITY METRICS APP



FABRIC MONITORING



COST CALCULATION

Metrics



FABRIC SQL DB DASHBOARD



FABRIC UNIFIED ADMIN MONITORING



FABRIC MONITORING



COST CALCULATION

FUAM is your best friend



Open Source and free to use



A lot of details



Worth your investment in time (and CU(s))



https://github.com/microsoft/fabric-toolbox/tree/main/monitoring/fabric-unified-admin-monitoring

Reitse Eskens

Engineer | Architect

Axians Business Analytics



Reitse.Eskens@axians.com

/in/reitseeskens

https://sqlreitse.com

@2meterDBA

Microsoft® Most Valuable Professional

Microsoft CERTIFIED TRAINER MCT

axians

Fabric Featured Partner Microsoft

C2-Vertrouwelijk

Storage options

C2-Vertrouwelijk

Lakehouse

C2-Vertrouwelijk

The Lakehouse (not the one made by Lego)



SPARK



HIGH CODE



CONFIGURABLE

Warehouse

C2-Vertrouwelijk

Warehouse (you know where to find it)

SQL ‘shell’ around spark

Low code and high code approaches

Full T-SQL (still) not supported

Eventhouse

C2-Vertrouwelijk

Eventhouse (not just a party place)

KQL

Can synchronise to OneLake

Blazing fast, at a cost

Fabric SQL Database

C2-Vertrouwelijk

Fabric SQL (not your average SQL)

Azure SQL made available in Fabric

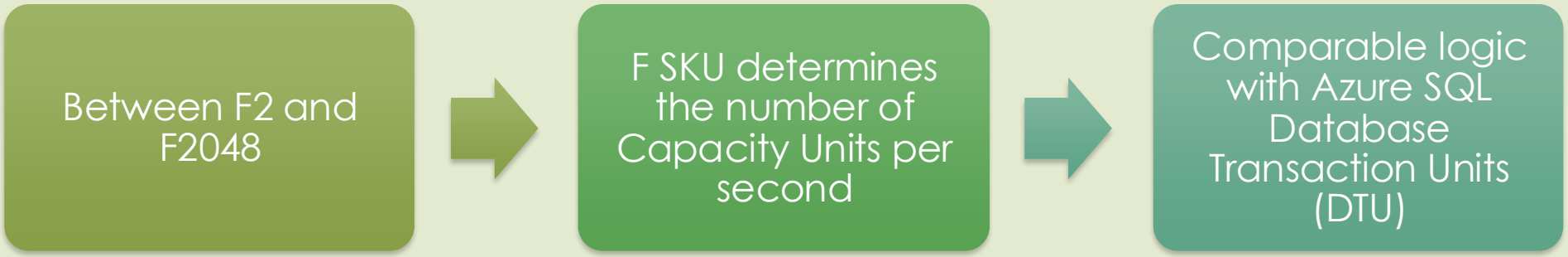
0 configuration

It's a different story

Capacity and Cost

C2-Vertrouwelijk

Capacities


Between F2 and F2048
F SKU determines the number of Capacity Units per second
Comparable logic with Azure SQL Database Transaction Units (DTU)

Fabric Capacity and cost

- F2 – F2048
- Higher = more power
- Every step = X2
- Reservations can be applied on yearly basis

Fabric Capacity and cost

F2 = 60 CU(s) per 30 Seconds

F2 = 60 * 2 * 60 *24 = 172.800 per day

F2 costs 272, 97 euro each month, 9 euro each day

1 CU = 9/172.800 = 0.00005 euro

60 CU(s) = 0,003 euro

Fabric Capacity and cost

F64 = 1920 CU(s) per 30 Seconds

F64 = 1920 * 2 * 60 *24 = 5.529.600 per day

F64 costs 8.734,93 euro each month, 291 euro each day

1 CU = 291/5.529.600 = 0.00005 euro

1920 CU(s) = 0,101 euro

More power?



Ingestion methods

C2-Vertrouwelijk

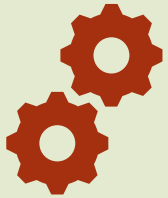
Ingestion (are you hungry too?)

- Pipeline
- Dataflow
- Notebook
- Eventstream

Pipeline



Think Azure Data Factory



Orchestration, simple transformations and data ingestion



Use it wisely

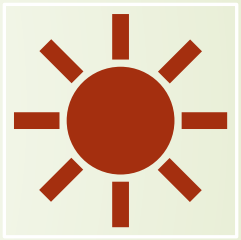
Dataflow Gen2



Low code solution



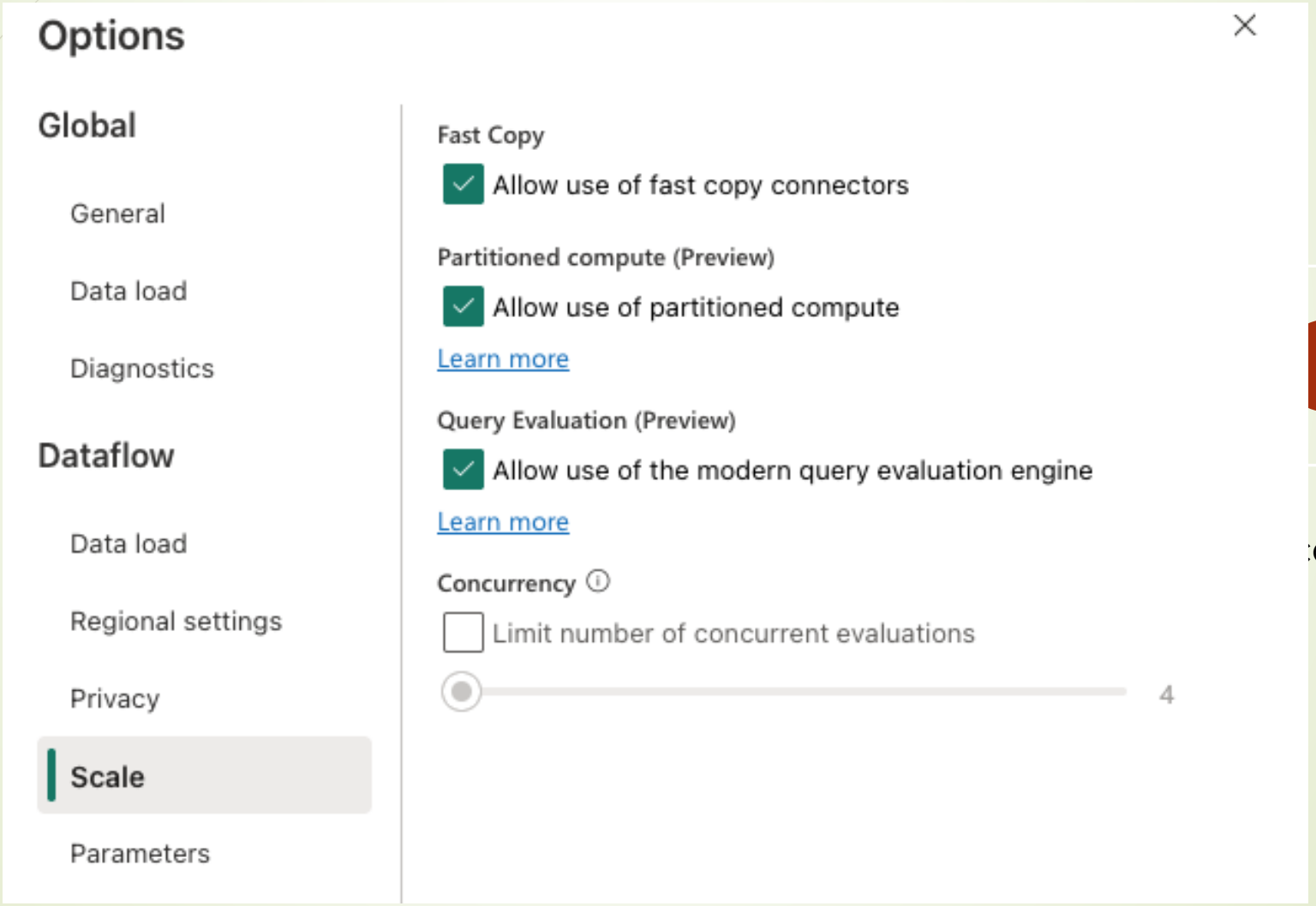
The new CI/CD is faster and more efficient



Use the cool options

C2-Vertrouwelijk

Dataflow Gen2



ool options

C2-Vertrouwelijk

Notebook



HIGH CODE

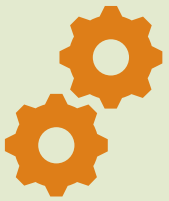


DEFAULT PYSPARK, OTHER LANGUAGES AVAILABLE



DO WHATEVER YOU WANT

Eventstream



Managed process to ingest into Realtime



Configuration depends on your source data

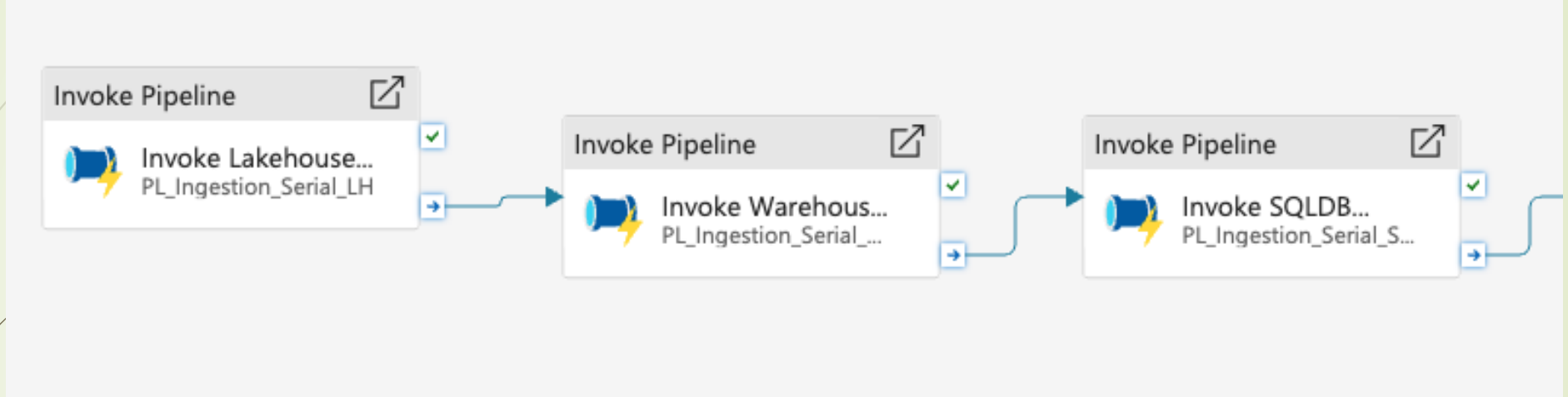


Use functions inside the Eventhouse to model and prepare

Let's go!



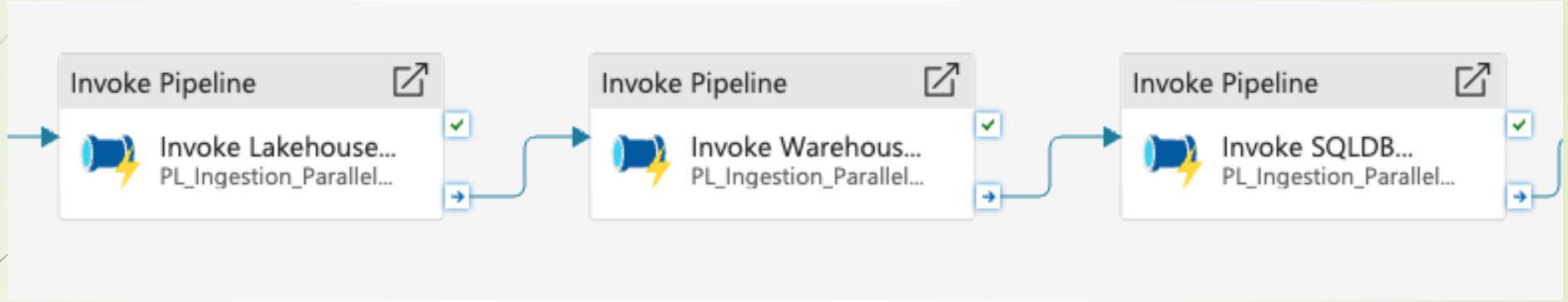
Pipeline to destination (serial)



Destination	Time	Capacity Units	Cost in Euro's
Lakehouse	1h 42m	499.320	24,96
Warehouse	2h 3m	563.400	28,17
Fabric SQL DB	4h 54m	1.376.100	68,80

C2-Vertrouwelijk

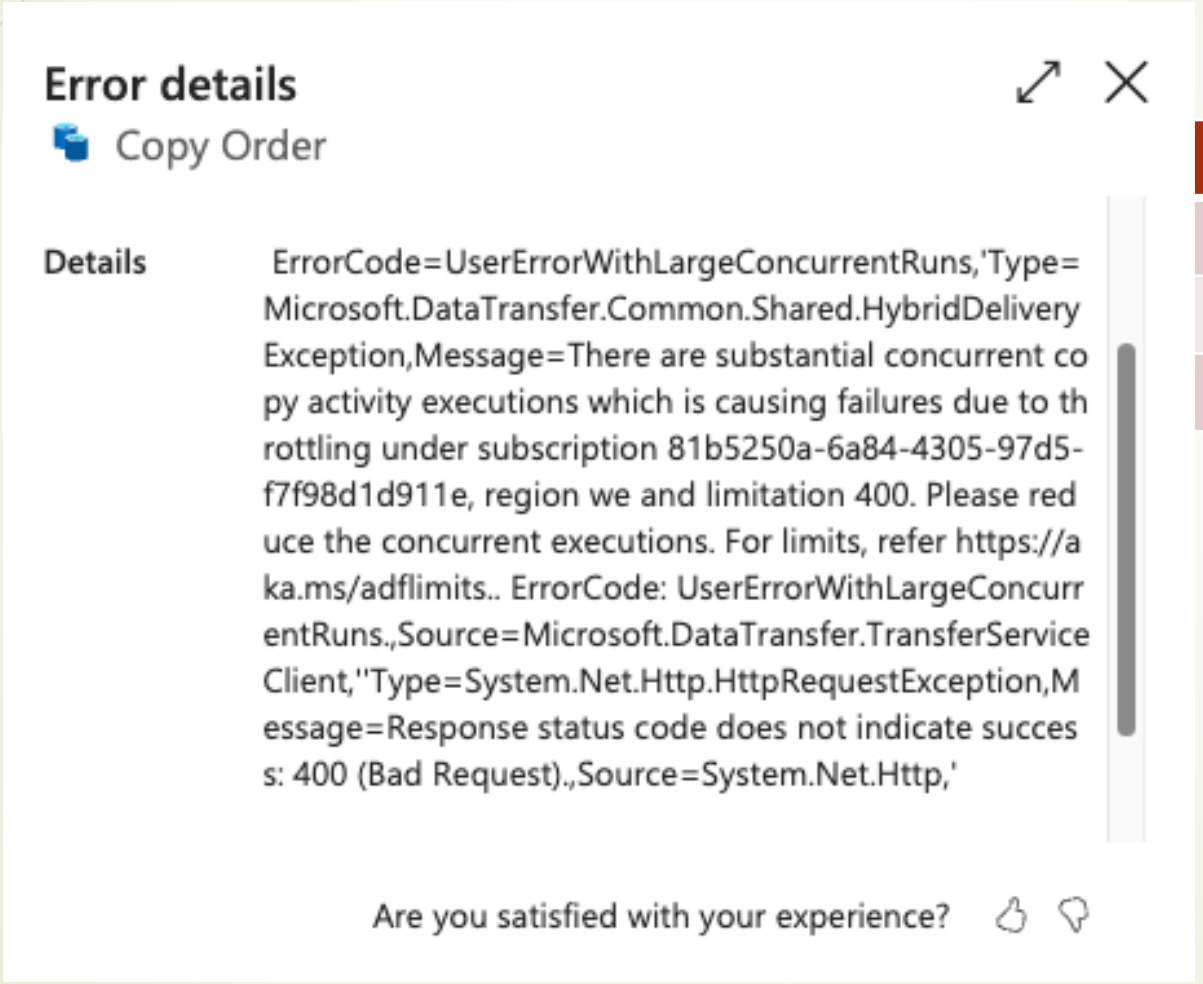
Pipeline to destination (parallel)



Destination	Time	Capacity Units	Cost in Euros
Lakehouse	5h 42m	355.680	17,78
Warehouse	1h 23m	116.100 (ERROR)	5,80
Fabric SQL DB	1h 18m	867.780(ERROR)	43,39

C2-Vertrouwelijk

Pipeline to destination (parallel)

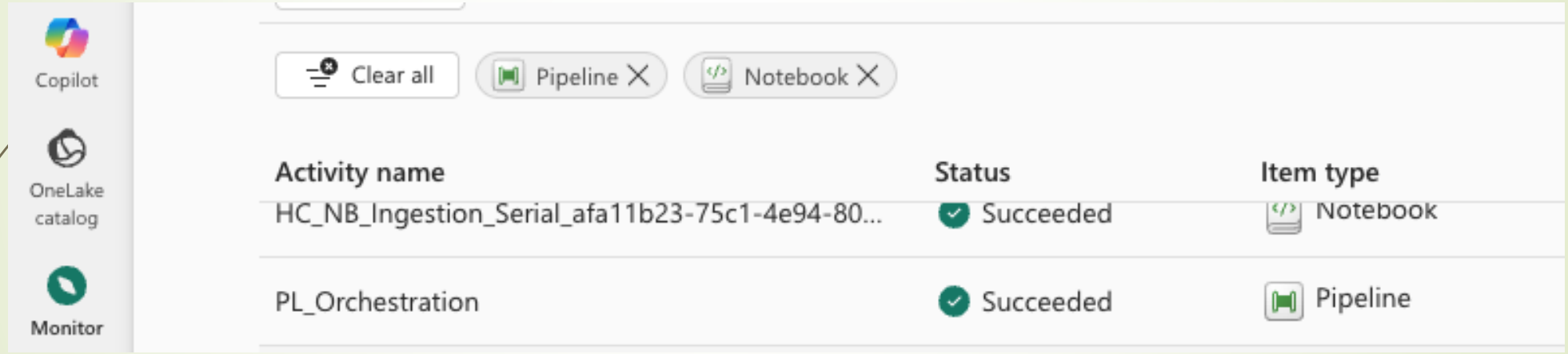


	Capacity Units
	355.680
	116.100 (ERROR)
	867.780(ERROR)

C2-Vertrouwelijk

Why the 400?

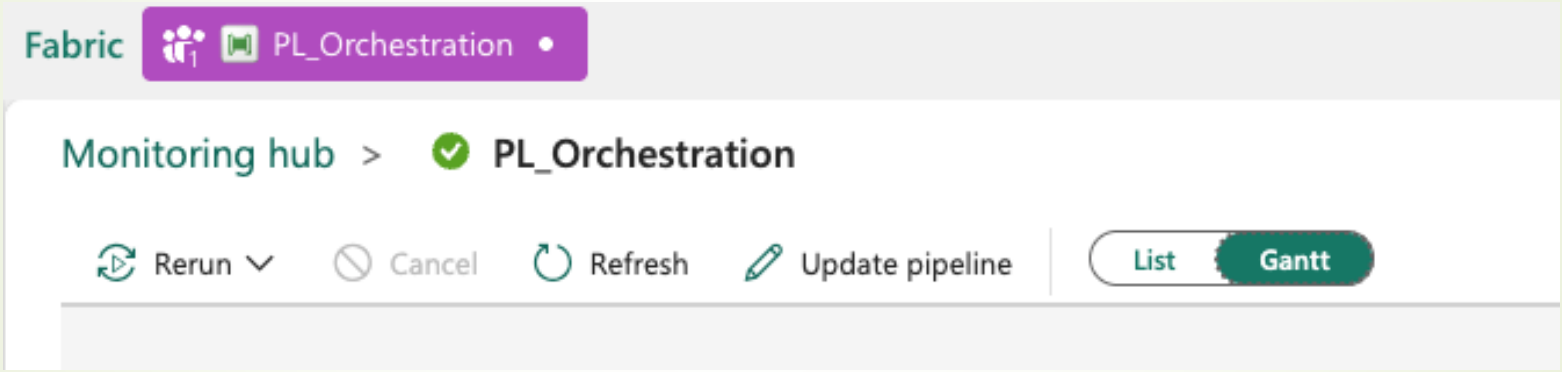
- Let's go into the monitoring



C2-Vertrouwelijk

Why the 400?

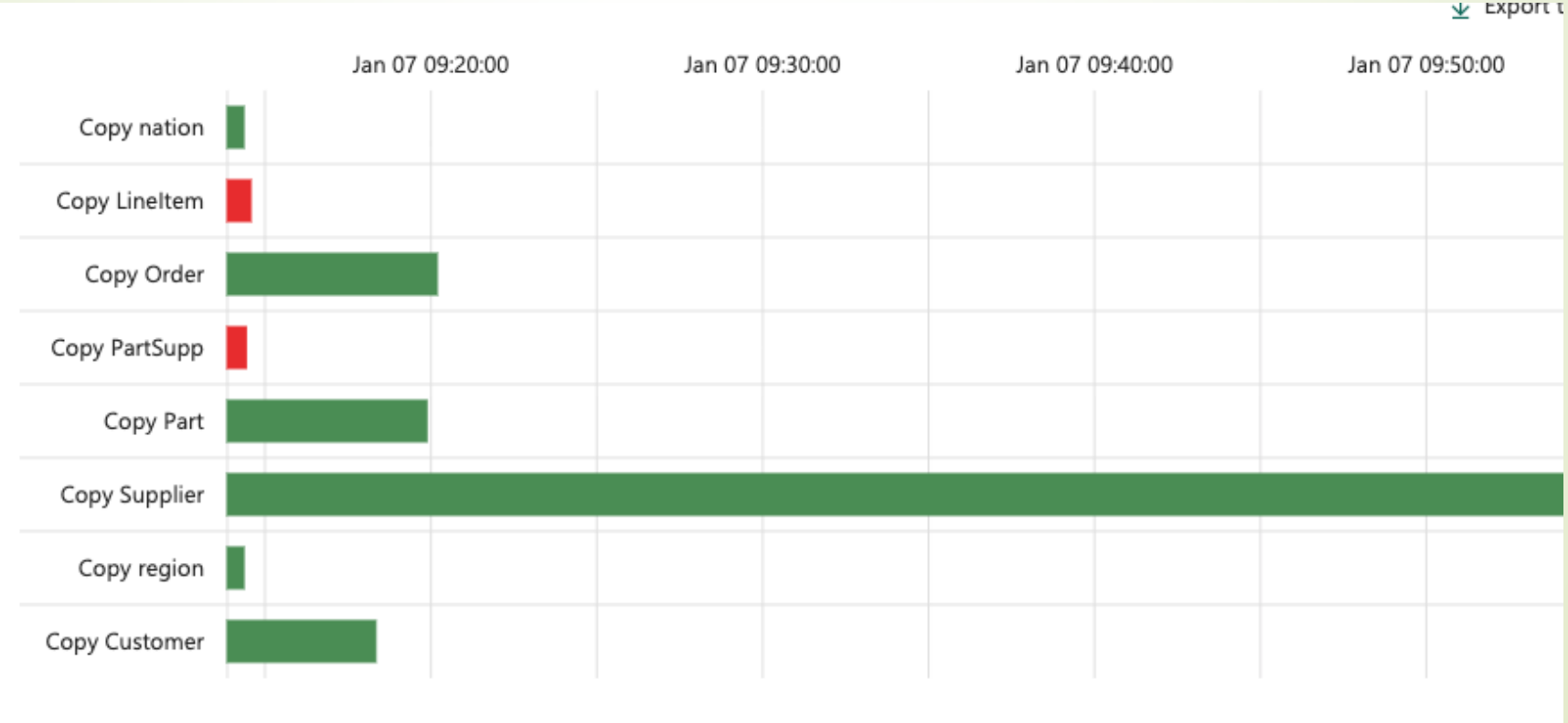
- Let's go into the monitoring



C2-Vertrouwelijk

Why the 400?

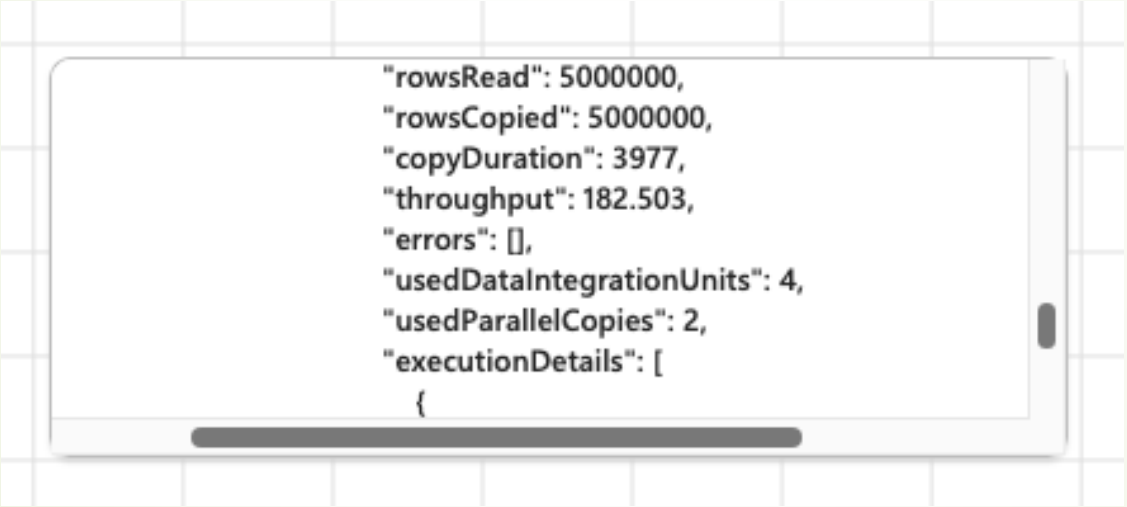
- Let's go into the monitoring



C2-Vertrouwelijk

Why the 400?

- Let's go into the monitoring



C2-Vertrouwelijk

What 400?

Table name	Used Data Integration Units	Running total
Nation	4	4
Order	256	260
Part	44	304
Supplier	4	308
Region	4	312
Customer	36	348
LineItem	256	604 Failed
PartSupp	228	832 Failed

C2-Vertrouwelijk

What do the docs say?

- https://learn.microsoft.com/en-us/azure/azure-resource-manager/management/azure-subscription-service-limits#azure-data-factory-limits

Concurrent Data Integration Units[1] consumption per subscription per Azure Integration Runtime region	Region group 1[2]: 6,000 Region group 2[2]: 3,000 Region group 3[2]: 1,500	Region group 1[2]: 6,000 Region group 2[2]: 3,000 Region group 3[2]: 1,500

C2-Vertrouwelijk

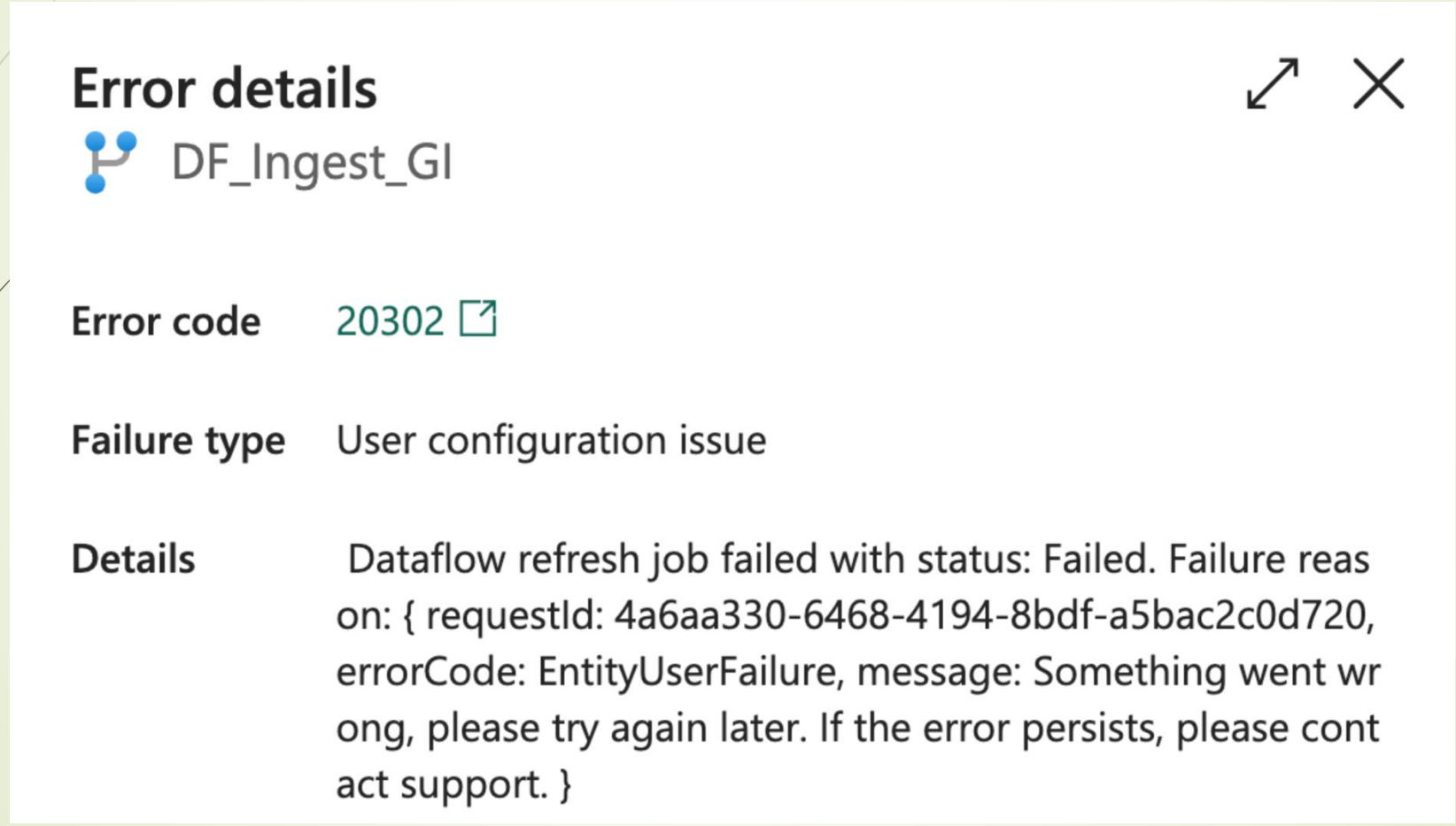
Dataflows to Lakehouse

- Error upon error
- Too many files, too much data

Dataflow Gen2	Dataflow Gen2 CI/CD
1.214.279 CU(s) (60,71 Euro)	64.469 CU(s) (3,22 Euro)
On average 4h runtime	2h runtime before error

C2-Vertrouwelijk

Dataflows to Lakehouse



C2-Vertrouwelijk


Nothing to see here.
Please disperse.

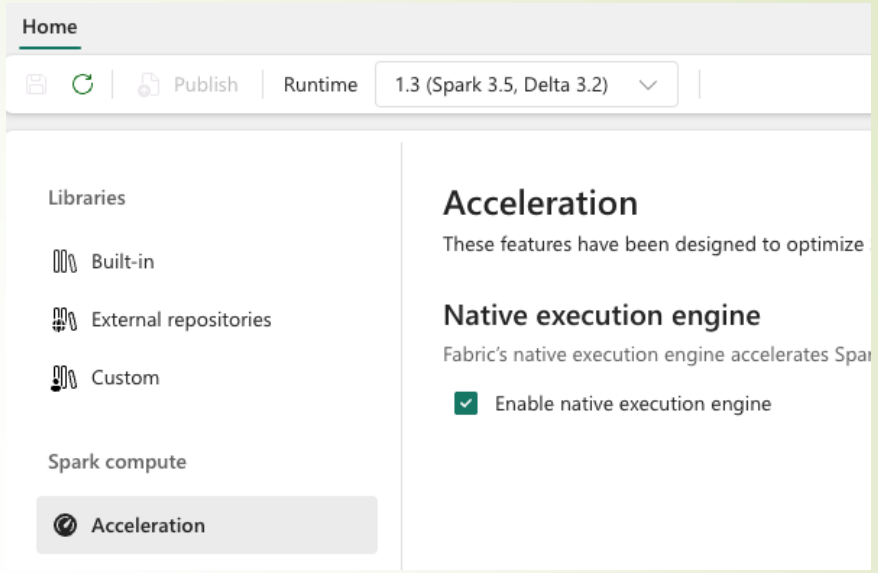
Notebooks and environments

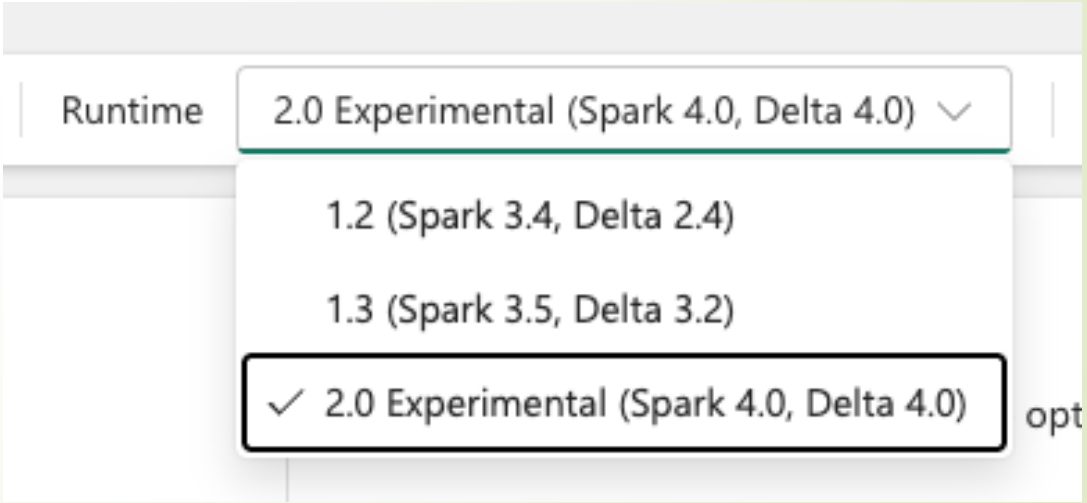
C2-Vertrouwelijk

Notebook environments

- Default (when you do nothing)
- Native Execution Engine (create your own environment)
- Fabric Runtime 2.0 (in preview, create your own environment

Native Execution Engine and Fabric 2.0





C2-Vertrouwelijk

Notebook to Lakehouse

Serial Time	Serial CU(s)	Parallel Time	Parallel CU(s)
24 minutes	107.627	26 minutes	??

C2-Vertrouwelijk

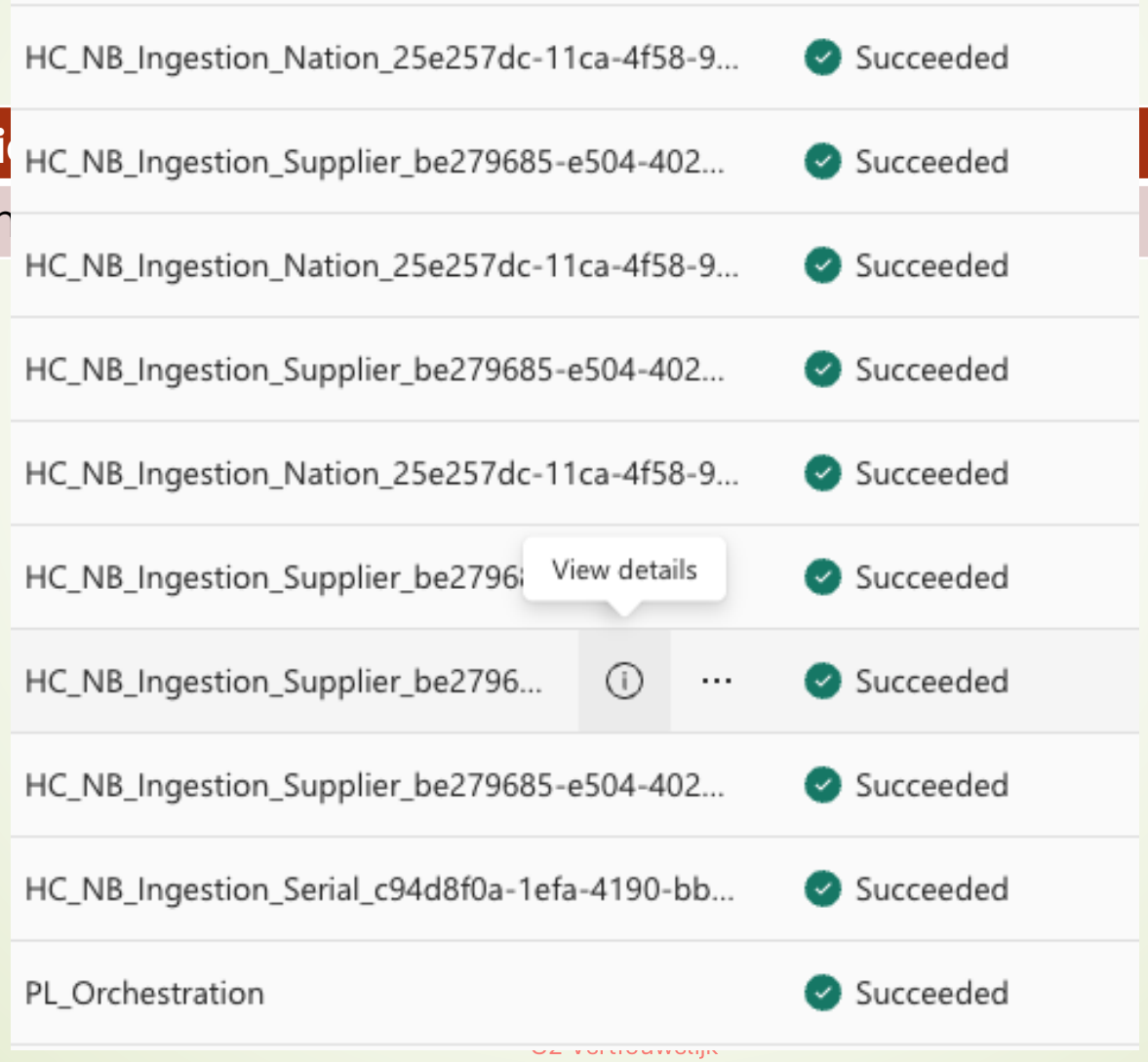
Notebook to Lakehouse

Item Name	Item State	FUAM Deleted State	Avg CU (s) 14d	Avg CU (s) 30d	U(s)
NB_Ingestion_Nation	Active	False	15,177	15,177	
NB_Ingestion_PartSupp	Active	False	1,263	1,263	
NB_Ingestion_Customer	Active	False	48,093	48,093	
NB_Ingestion_Supplier	Active	False	40,477	40,477	
NB_Ingestion_Serial	Active	False	107,627	107,627	
NB_Ingestion_Region	Active	False	34,620	34,620	
NB_Ingestion_Part	Active	False	4,186	4,186	
NB_Ingestion_Orders	Active	False	29,711	29,711	

C2-Vertrouwelijk

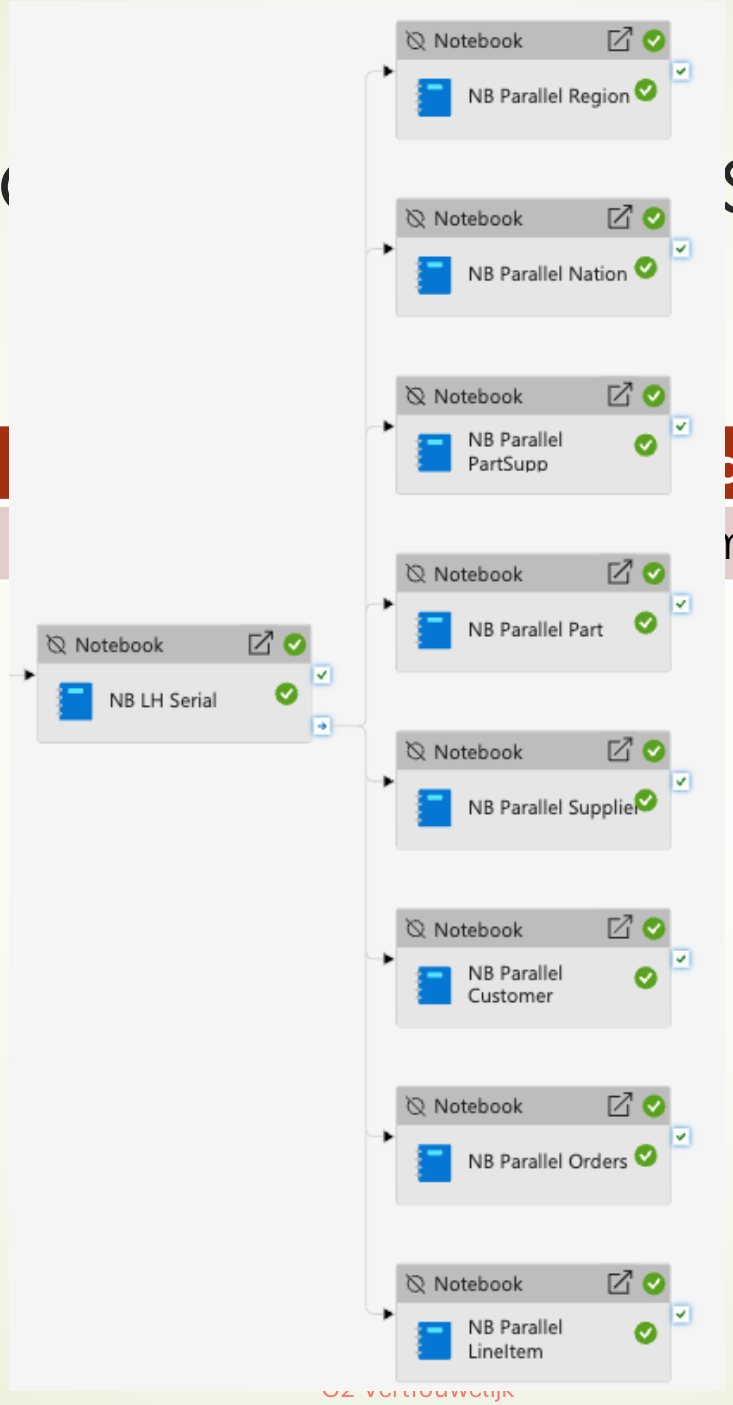
Notebook to Lakehouse

Seri		Parallel CU(s)
24 n		??



Notebo se

Serial Time	allel Time	Parallel CU(s)
24 minutes	inutes	??



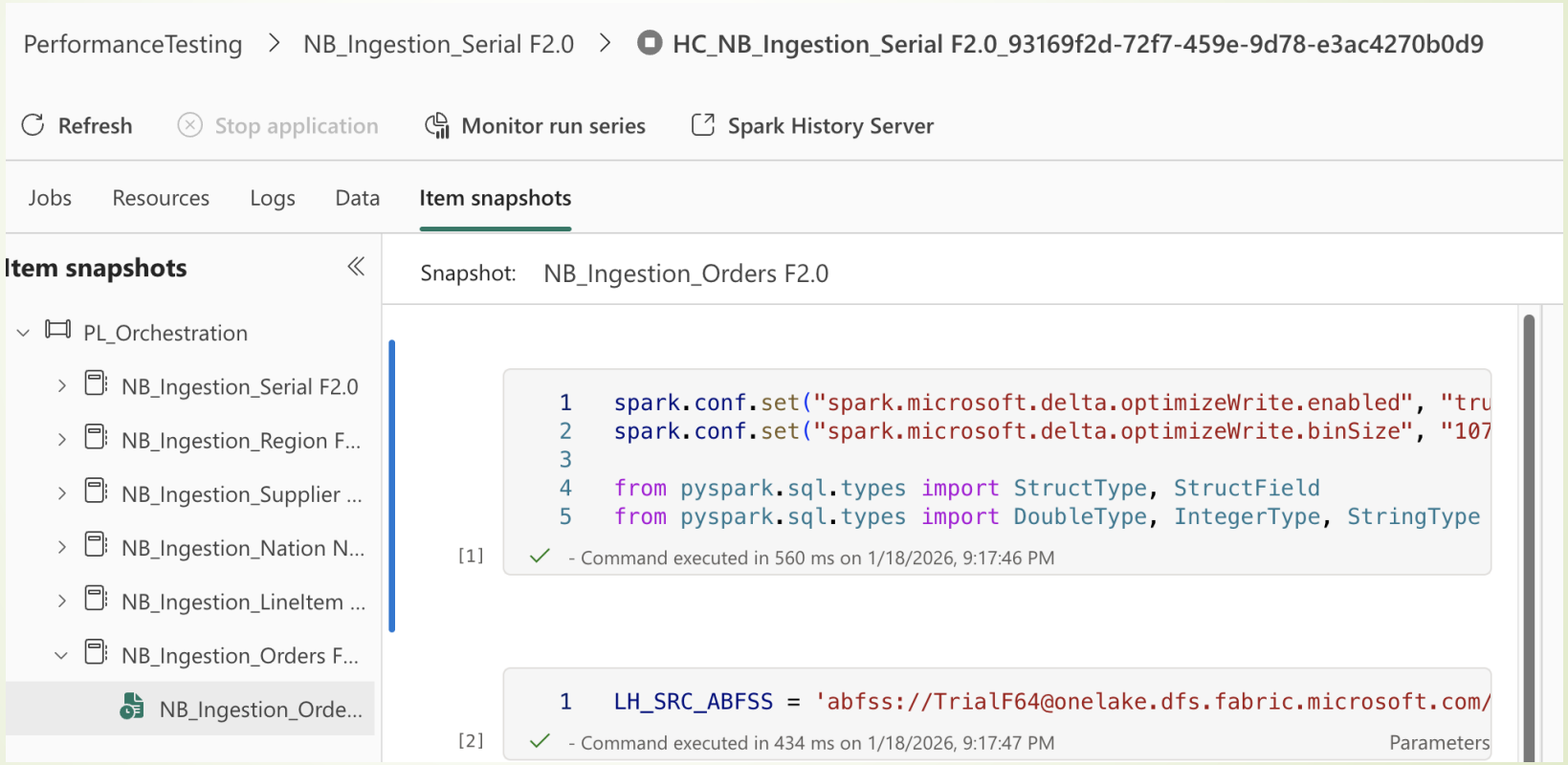
Why the missing numbers?

High concurrency

It's amalgamated into the first notebook

You, yes You!, need to dive into the metrics of that specific notebook to get to the details

There they are!



C2-Vertrouwelijk


PERFECT
JUST PERFECT
makeameme.org

Differences between Default, Native Execution Engine and Runtime 2.0

File	Default	Native Execution	Fabric 2.0
Serial	22:30	22:37	**21:15**
region	01:19	00:56	**00:45**
Nation	01:15	**00:56**	03:00
PartSupp	**08:52**	09:30	09:45
Part	13:05	15:16	**07:05**
Supplier	01:53	**01:22**	02:10
Customer	04:38	**02:31**	04:27
Orders	**09:25**	12:20	10:00
LineItem	18:50	14:11	**10:57**

C2-Vertrouwelijk

Differences between Default, Native Execution Engine and Runtime 2.0

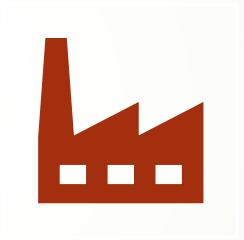
File	Default	Native Execution	Fabric 2.0
Serial	49.837	85.712	65.250
region	12.542		
Nation	3.062		
PartSupp	13.411		
Part	3.261		21.626
Supplier	40.477	2.123	6.693
Customer	4.839	10.370	6.512
Orders	29.711		13.843
LineItem	12.272		

C2-Vertrouwelijk

Eventstream to Eventhouse



Limited ingestion because of max file size that can be processed



Nation, Region, Supplier



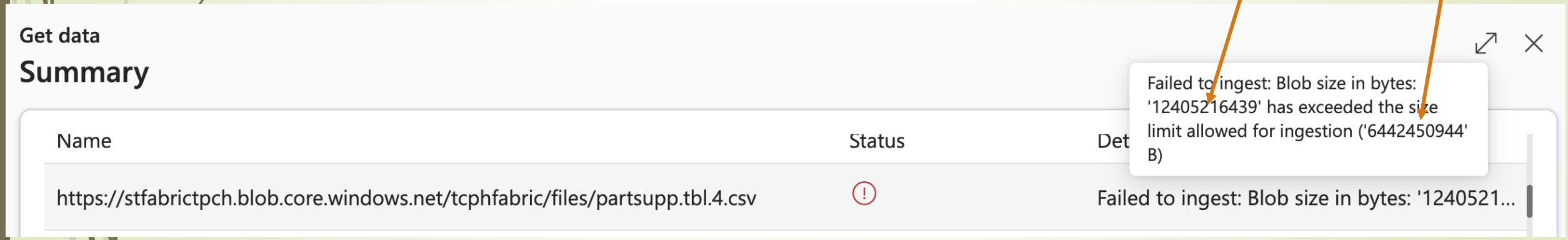
1505 CU(s)

C2-Vertrouwelijk

Eventstream to Eventhouse

11,5 GB

768 MB



C2-Vertrouwelijk

Tying it all together



Storage cost?

Storage	CU cost	Euro's
Lakehouse	8.758	0,43
Warehouse	10.626	0,53
Eventhouse	4.105	0,20
Fabric SQL DB	64.715	3,23

C2-Vertrouwelijk

Ingestion cost

Type	CU cost	Euro's
Pipelines	3.778.380	188,91
Notebooks	169.412	8,47
Event Streams	4.649	2,32
Dataflows Gen2	1.637.230	81,86
Dataflows Gen2 CI/CD	51.560	2,57

NOTE! The results are accurate at the time of measurement. Do not take these as absolute truths but use them as guidance for your own testing!

Numbers shown are averages, outliers will influence the outcomes

C2-Vertrouwelijk

Personal storage preferences

1. Lakehouse
2. Eventhouse
3. Warehouse
4. Fabric SQL DB (metadata only)

C2-Vertrouwelijk

Is there a best solution

No, but there is a best fit

Expertise, budget, workforce

It always depends

C2-Vertrouwelijk

Pro Tips



USE FUAM FOR MONITORING



TEST, TEST AND TEST AGAIN. NEW RELEASES CAN GO EITHER WAY

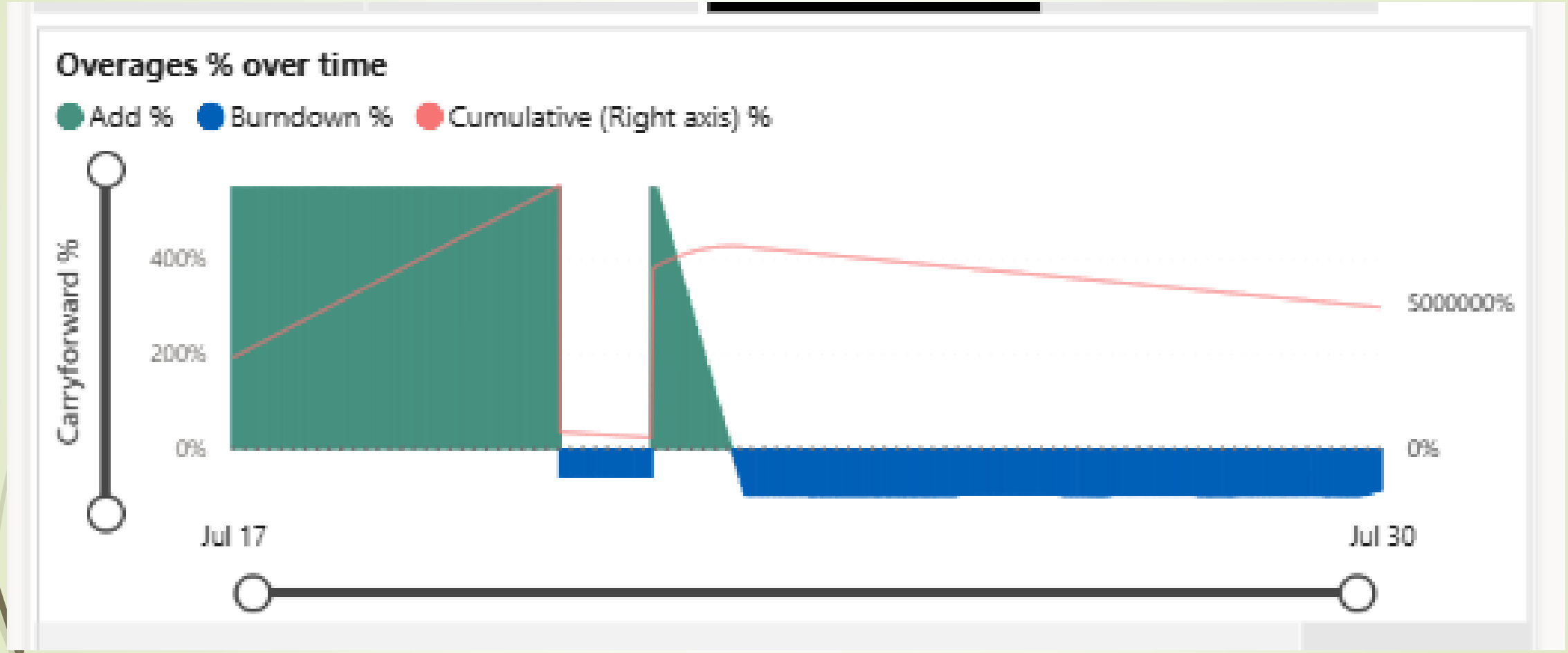


THE CHEAPEST SOLUTION MAY NOT BE THE BEST (FOR YOU)

When stuff goes off the rails



> 5.000.000% Overage with Always-On in RTI



C2-Vertrouwelijk





THANK YOU!