

Star Schema ALL The Things! But why?

Benni De Jagere



Slides



Big thank you to our great sponsors



backstage



INSPARI
a valantic company



Viz.fo

Raffle prizes from the sponsors





Benni De Jagere

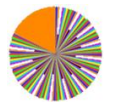
Senior Program Manager | Fabric Customer Advisory Team (FabricCAT)



dataMinds



sessionize



Fabric CAT

.be Member

@BenniDeJagere

/bennidejagere

/bennidejagere

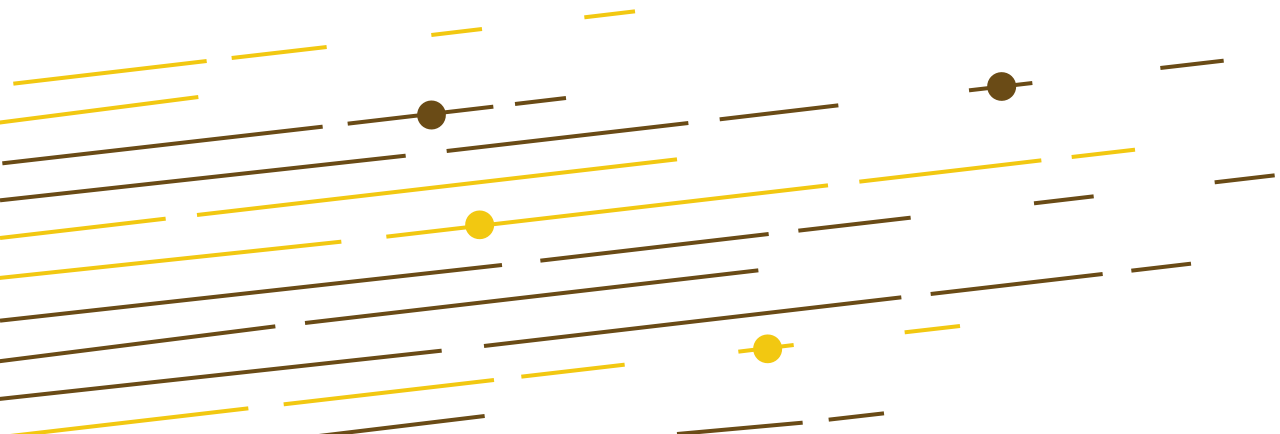
/bennidejagere

#SayNoToPieCharts



What spurred the idea for this session?

Spoiler Alert: It was (yet) another X discussion

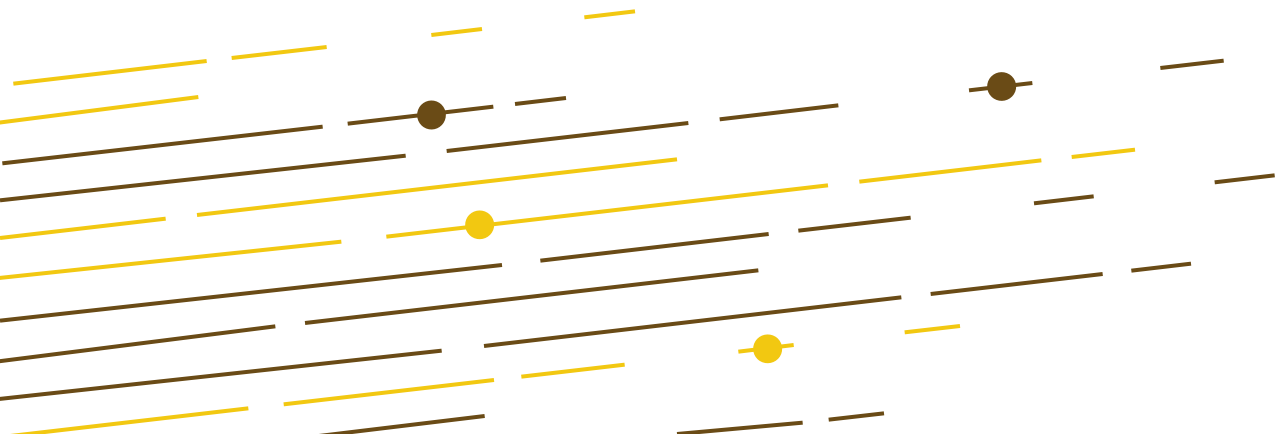


It all started with an X, how did it end up like this?

It was only an X, it was only an X

- *"You should never do [xyz]"*
- *"You always need to [xyz]"*
- *"I won't even touch a model if it's not [xyz]"*
- But why?
- [Kurt Buhler - the Goblin behind the Model](#)

Session Objectives



Session Objectives

- Star Schema ALL the things! (For Power BI)
- Convince you to be critical of best practices
- Take you through my thought process
 - Hang on tight! 🎢

The Data & Architecture

The Data

www.citibikenyc.com/system-data

Public Open Data

Starts June 2013

Information about every trip

Longer than 60 seconds

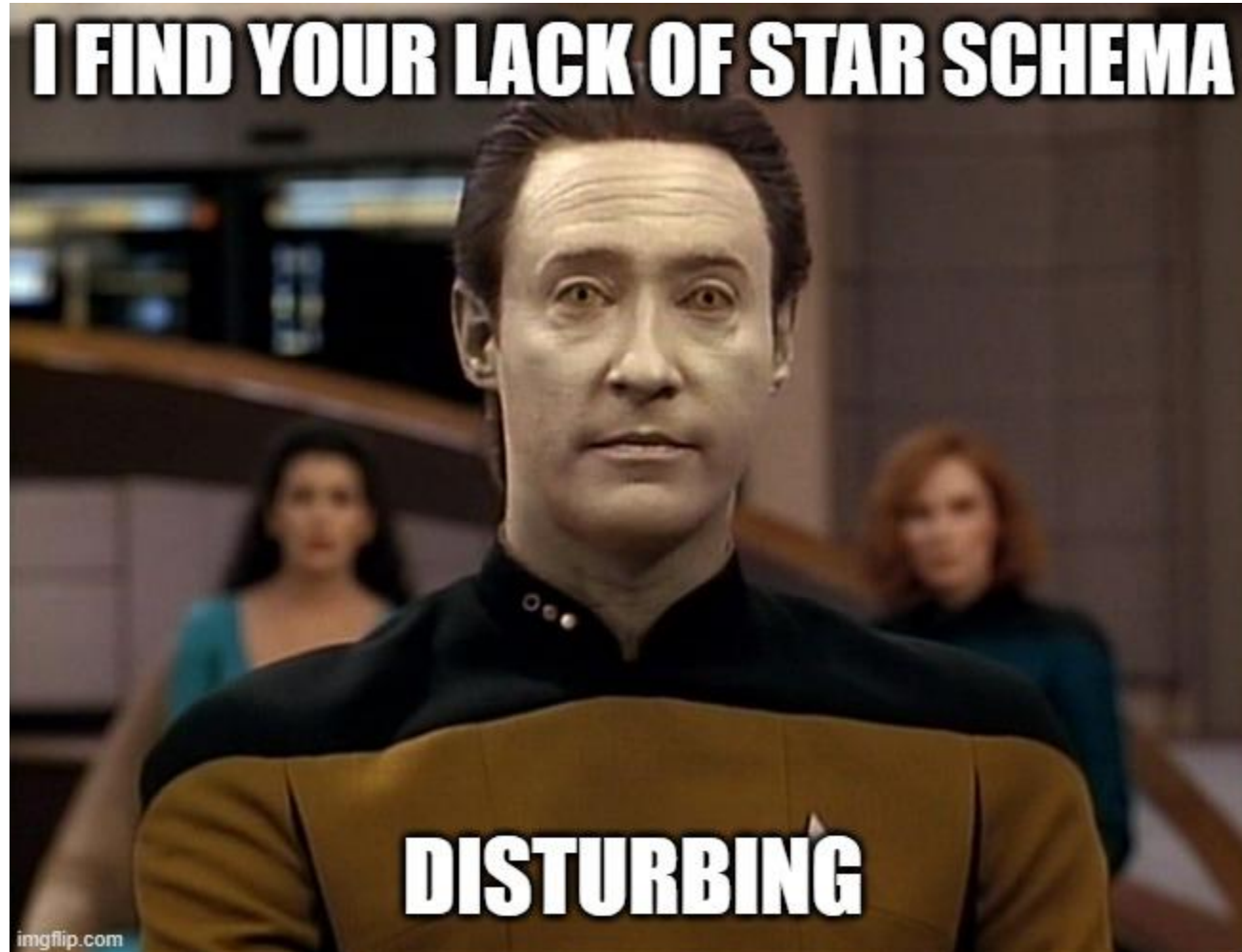
Only 'actual trips'

Masterdata

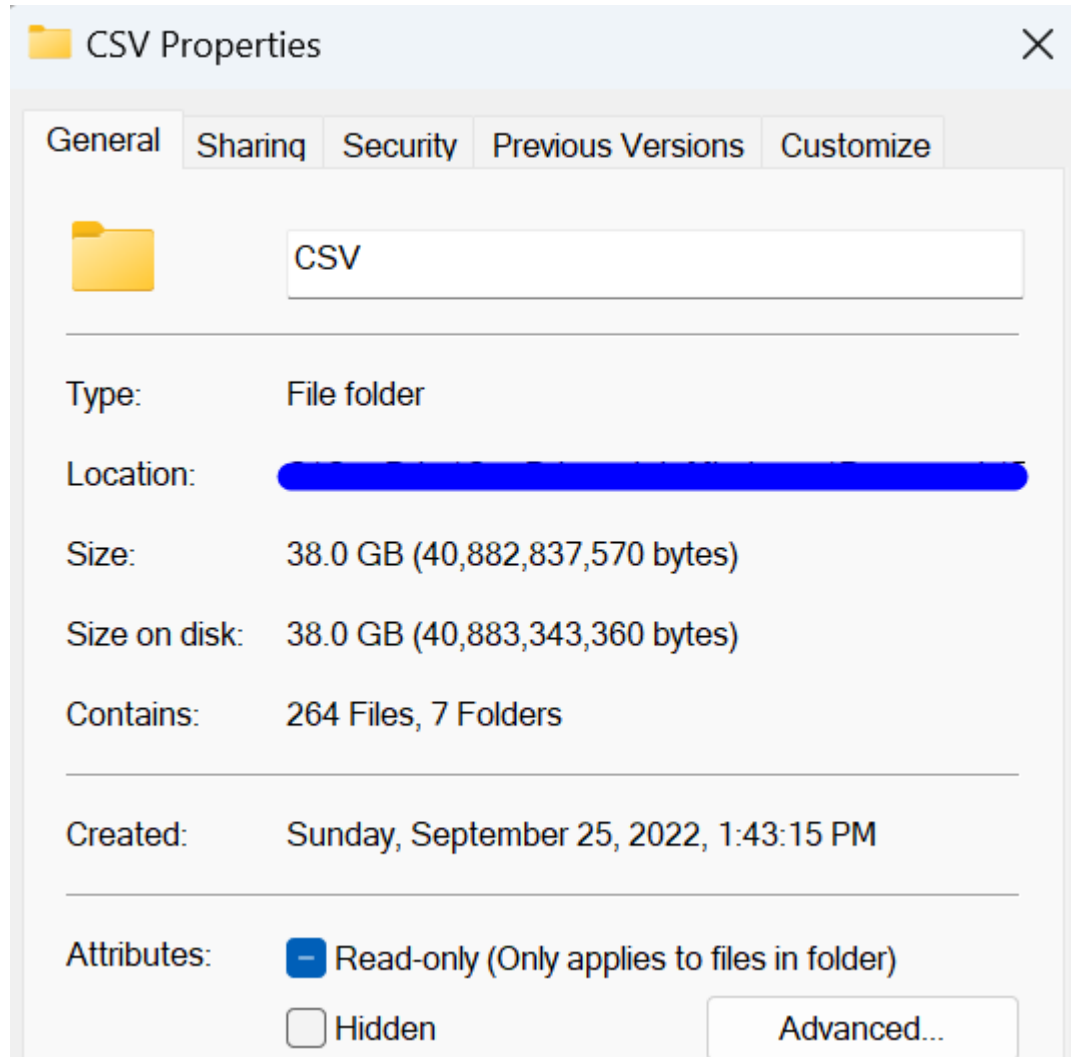


<https://i0.wp.com/thenypost.files.wordpress.com/2013/12/citibike1.jpg>

The Data

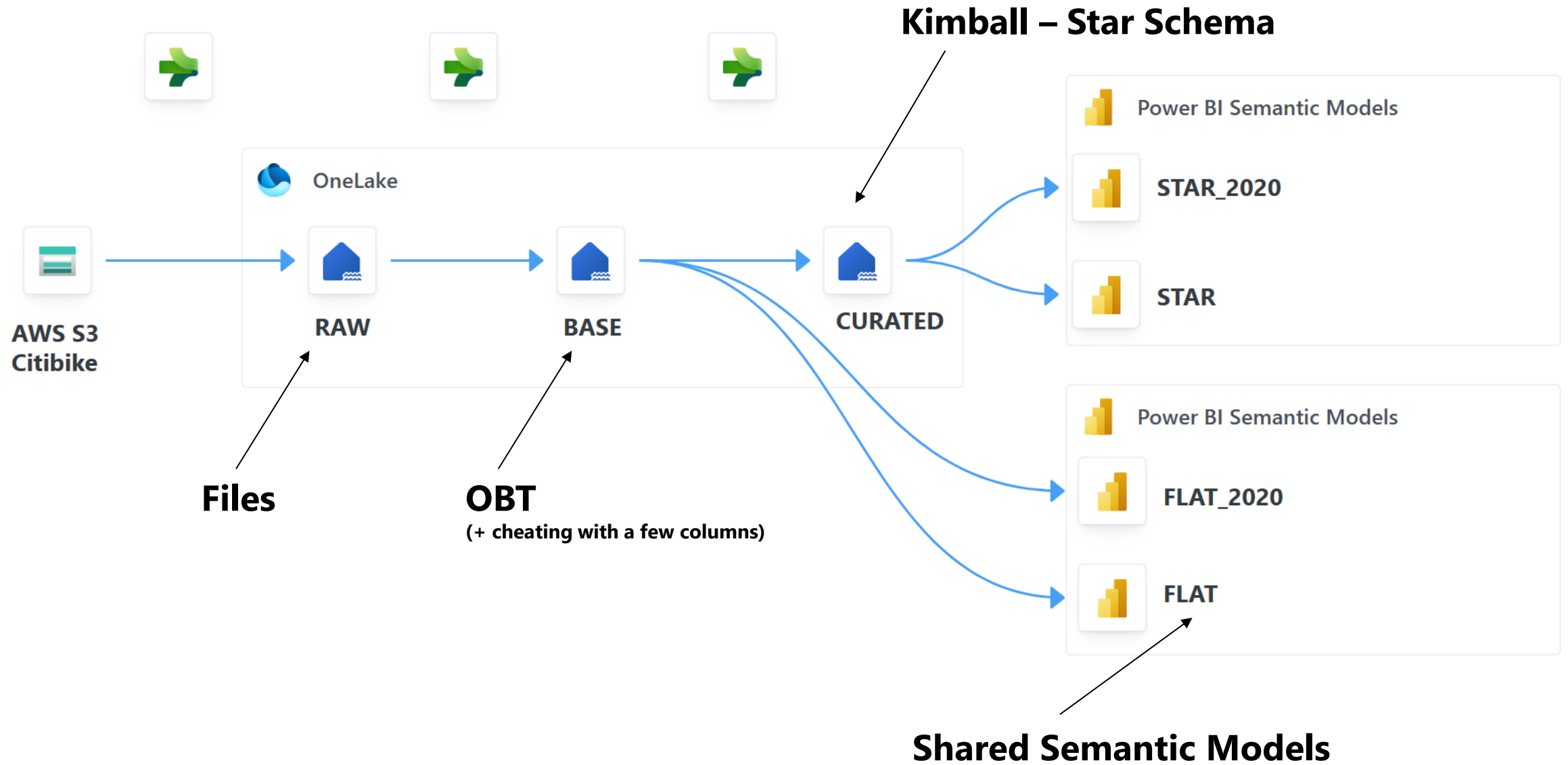


The Data



	Lakehouse Name	Table Name	Num_Files	Num_Rowgroups	Num_Rows	Delta_Size_MB	Last OPTIMIZE Timestamp	Last VACUUM Timestamp
0	NYCCitibike_CURATED	Trips_FA	42	75	212794868	6205	None	None
	Lakehouse Name	Table Name	Num_Files	Num_Rowgroups	Num_Rows	Delta_Size_MB	Last OPTIMIZE Timestamp	Last VACUUM Timestamp
0	NYCCitibike_BASE	Trips	24	73	212794868	6205	None	None
3	NYCCitibike_CURATED	UserType_DI	1	1	3	0	None	None
4	NYCCitibike_CURATED	Gender_DI	1	1	59	0	None	None
5	NYCCitibike_CURATED	TripType_DI	1	1	3	0	None	None
6	NYCCitibike_CURATED	MemberType_DI	1	1	3	0	None	None
7	NYCCitibike_CURATED	Region_DI	1	1	8	0	None	None
8	NYCCitibike_CURATED	Batch_DI	1	1	10	0	None	None
9	NYCCitibike_CURATED	FileType_DI	1	1	3	0	None	None
10	NYCCitibike_CURATED	Station_DI	1	1	3991	0	None	None
11	NYCCitibike_CURATED	RideType_DI	1	1	4	0	None	None
12	NYCCitibike_CURATED	Bike_DI	1	1	35553	0	None	None
13	NYCCitibike_CURATED	TripsXL_FA	0	0	0	0	None	None
14	NYCCitibike_CURATED	TripsXXL_FA	0	0	0	0	None	None
15	NYCCitibike_CURATED	Date_DI	1	1	7304	0	None	None

The Architecture



The 'Metrics'

aka "What do I care about?"

The Metrics

Refresh Time

Model Size

(Re)Usability

DAX Complexity

Performance

Cost

Mystery

The Tools

The Tools

Performance Analyzer Pane

DAX Studio

VertiPaq Analyzer

Tabular Editor 2

SSMS Profiler

Visualize Your Refresh

The Models

Remember when I said 'No Shortcuts?'

Let's take a shortcut










Data should be transformed as far upstream as possible, and as far downstream as necessary.

Matthew Roche, 2021

(The purple haired sword aficionado in a feline themed team)

<https://ssbipolar.com/2021/05/31/roches-maxim>

From a previous session

 1_NYC_Citibike_BASE.pbix	4,397,349 KB
 2_NYC_Citibike_DataTypes.pbix	3,775,468 KB
 3_NYC_Citibike_AutoDateTime.pbix	2,553,543 KB
 4_NYC_Citibike_UnusedColumns.pbix	1,761,946 KB
 5_NYC_Citibike_StarSchema.pbix	837,947 KB
 6_NYC_Citibike_Report_v1 (Calculated Column).pbix	1,023,519 KB
 7_NYC_Citibike_Report_v2(NewCards).pbix	837,363 KB
 7_NYC_Citibike_Report_v2.pbix	837,355 KB
 8_NYC_Citibike_Report_v3_UnusedRows.pbix	199,357 KB

The Shortcut

- PowerQuery transformations didn't scale
 - Led to timeouts, capacity pressure, ..
 - DAX Calculated Columns/Tables scaled even less
-
- Could you get it to work well?
 - Yes, but it would require time, resources, and skill

Let's Compare!

Refresh Time

How to measure

- Use Profiler to run a trace
- Save it as 'Trace XML file'
- Leverage Phil Seamark - '[Visualize your refresh](#)'
- Compare results and notes

Job Trace Reporting

Star Schema – 2020 only

Select a Request ID

All

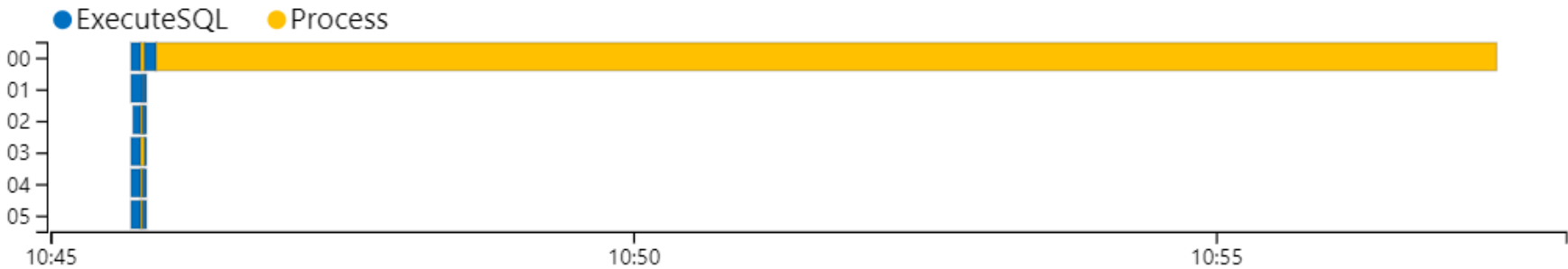
▼

532K

Total CPU Time

11 mins 43 sec

Duration



ObjectName	Rows Read	Duration Measure (Seconds)	Rows per second
TripType	3	6	0.50
FileType	3	1	3.00
MemberType	3	1	3.00
UserType	3	1	3.00
RideType	4	1	4.00
Region	8	1	8.00
Batch	10	1	10.00
Gender	59	1	59.00
StopStation	3,991	6	665.17
DateStart	7,304	6	1,217.33
StartStation	3,991	1	3,991.00
Total	20,074,475	703	28,555.44

Job Trace Reporting

Flat Table– 2020 only

Select a Request ID

All

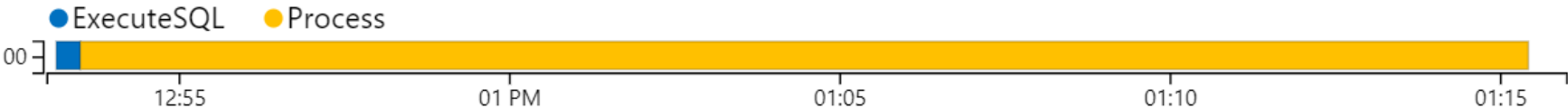
▼

1M

Total CPU Time

22 mins 32 sec

Duration



ObjectName	Rows Read	Duration Measure (Seconds)	Rows per second
Trips	19,843,659	1337	14,841.93
Total	19,843,659	1337	14,841.93

Houston, we have a problem

- Flat Table was not able to refresh through 'Refresh Now' in UI
 - Memory Footprint exceeded P1 allocation
 - So I cheated 😊
-
- When increased to P2, refresh timed out after 5 hours
 - Incremental Refresh was configured for both models

Did you know?

- By default, Power BI creates an Attribute Hierarchy
 - Adds Model Size, Refresh Time
- Mostly used for MDX / Excel PivotTable
- Can be disabled for columns that are not:
 - Visible
 - Used in Sort By Column
 - Used in Hierarchies

<https://blog.crossjoin.co.uk/2018/07/02/isavailableinmdx-ssas-tabular/>

<https://data-mozart.com/hidden-little-gem-that-can-save-your-power-bi-life/>

Model Size

Model Size (2020)

OBT (Flat)

 00_NYCCitibike_FLAT_2020 (PBI Service)

Total Size in Memory	Last Data Refresh		Analysis Date
2.07 GB ⓘ	3/17/2024 2:15:44 PM +01:00		3/17/2024 2:24:01 PM +01:00
Compatibility	Tables	Columns	Server
1567	1	21	powerbi://api.powerbi.com/v1.0/myorg/BDJ_NYCCitibike_StarSchemaAllTheThings


Star Schema

 00_NYCCitibike_STAR_2020 (PBI Service)

Total Size in Memory	Last Data Refresh		Analysis Date
299.46 MB ⓘ	3/17/2024 11:57:27 AM +01:00		3/17/2024 11:58:22 AM +01:00
Compatibility	Tables	Columns	Server
1567	16	133	powerbi://api.powerbi.com/v1.0/myorg/BDJ_NYCCitibike_StarSchemaAllTheThings


Model Size (Full)

OBT (Flat)

 00_NYCCitibike_FLAT (PBI Service)

Total Size in Memory	Last Data Refresh		Analysis Date
18.49 GB ⓘ	3/17/2024 5:49:58 PM +01:00		3/17/2024 10:14:04 PM +01:00
Compatibility	Tables	Columns	Server
1567	4	42	powerbi://api.powerbi.com/v1.0/myorg/BDJ_NYCCitibike_StarSchemaAllTheThings

Star Schema

 00_NYCCitibike_STAR (PBI Service)

Total Size in Memory	Last Data Refresh		Analysis Date
3.09 GB ⓘ	3/18/2024 12:11:00 AM +01:00		3/18/2024 6:14:07 AM +01:00
Compatibility	Tables	Columns	Server
1567	16	133	powerbi://api.powerbi.com/v1.0/myorg/BDJ_NYCCitibike_StarSchemaAllTheThings

Let's talk about relationships..

(Why GUIDs and Business Keys do not work)

- Relationships need to be materialized
- We want to fit as much as possible into Memory (speed+ +)
 - Cardinality and Data Type impact this
- Business Keys can change over time
- How do you want your model to evolve?

<https://www.sqlbi.com/articles/costs-of-relationships-in-dax/>

<https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/natural-durable-supernatural-key/>

<https://data-marc.com/2023/05/17/the-hidden-impact-of-keys-in-your-power-bi-data-model/>

<https://exceleratorbi.com.au/replace-guids-with-a-surrogate-key-for-better-performance/>

Large Model Storage Format

- Default Segment Size goes from 1M to 8M
- Keep in mind you can no longer download .pbix

<https://learn.microsoft.com/en-us/power-bi/enterprise/service-premium-large-models>

<https://learn.microsoft.com/en-us/power-bi/enterprise/service-premium-large-models#default-segment-size>

<https://www.sqlbi.com/tv/explaining-segment-size-in-power-bi-premium-unplugged-29/>

<https://www.sqlbi.com/blog/marco/2021/06/29/choosing-azure-analysis-services-or-power-bi-premium-for-large-datasets/>

(Re)Usability

(Re)Usability

- Which column do I use?
- Hello, Auto Date/Time!
- Need more columns for Time Analysis
- Solution needed for base columns for Measures
 - Added Duration, Customer Age to Table
- Any logic I add to the Model, will be hard to reuse
- Also the space for a discussion about Implicit vs. Explicit measures

<https://data-mozart.com/understanding-explicit-vs-implicit-measures-in-power-bi/>

Performance + DAX Complexity

Trips Taken

20M

Time Taken

435M

Dur AVG

21.91

Start Station	Trips Started
1 Ave & E 68 St	100,753
West St & Chambers St	99,364
W 21 St & 6 Ave	99,191
12 Ave & W 40 St	97,415
Broadway & W 60 St	91,855

Stop Station	Trips Ended
West St & Chambers St	101,768
W 21 St & 6 Ave	100,314
1 Ave & E 68 St	100,260
12 Ave & W 40 St	99,334
E 17 St & Broadway	93,967

Date

Last1Select

No filters applied

Hour

All

Region

All

Start Station

All

Stop Station

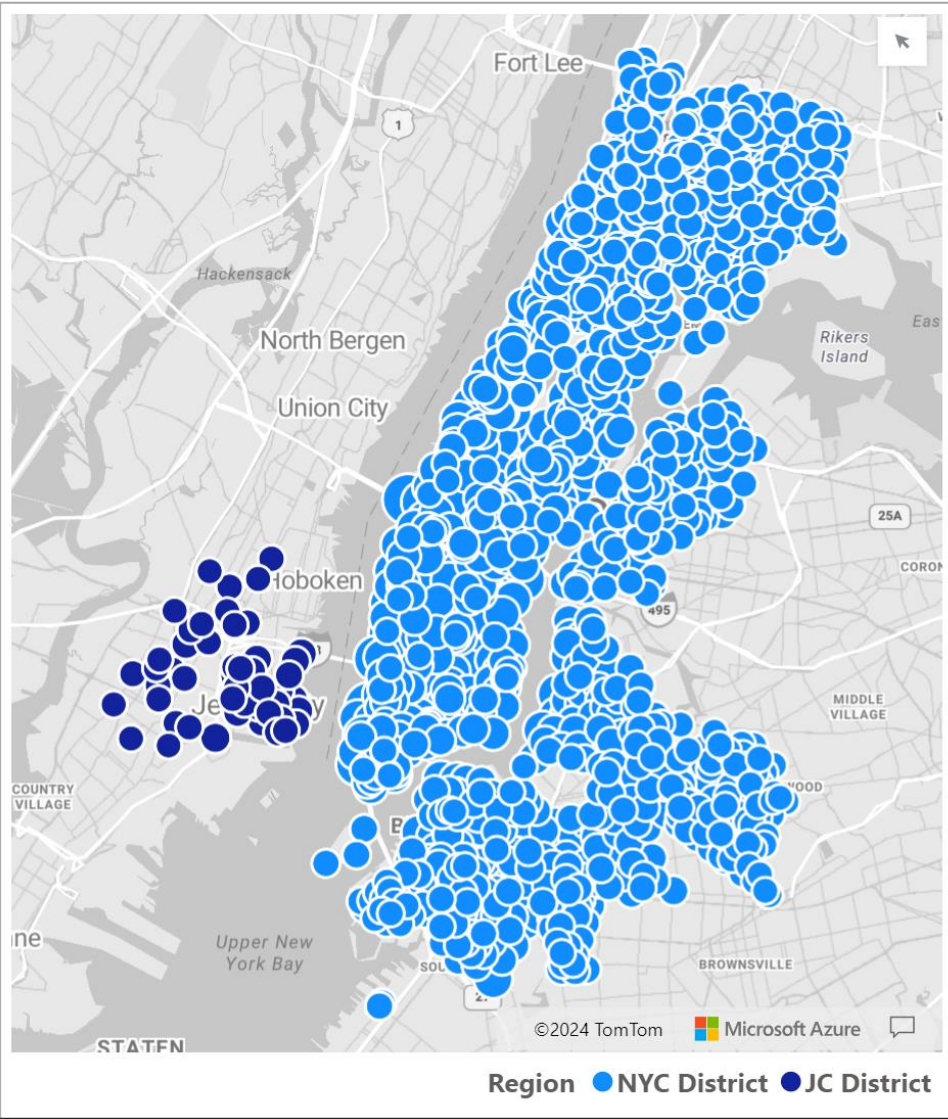
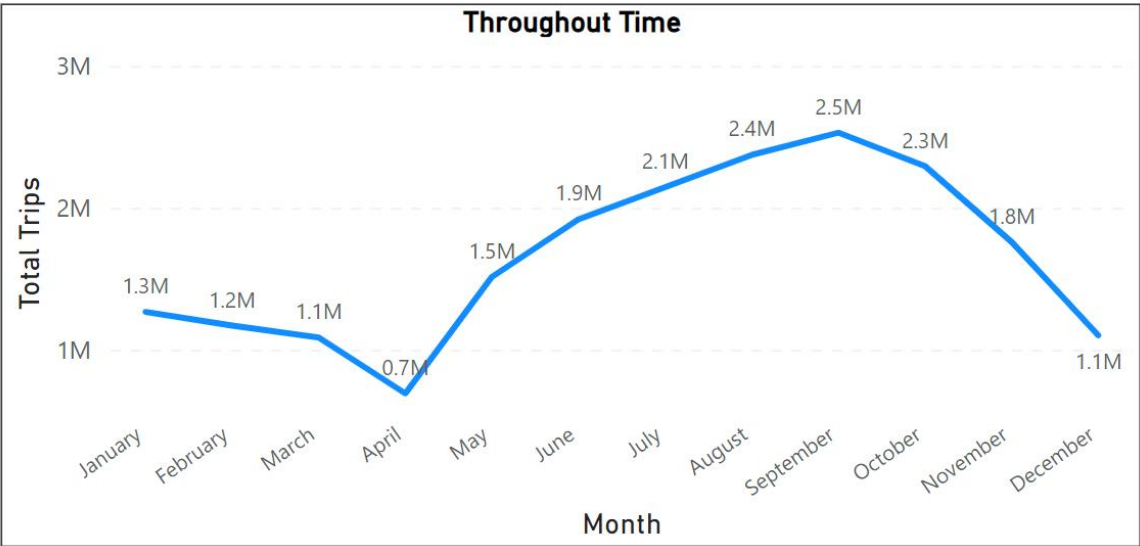
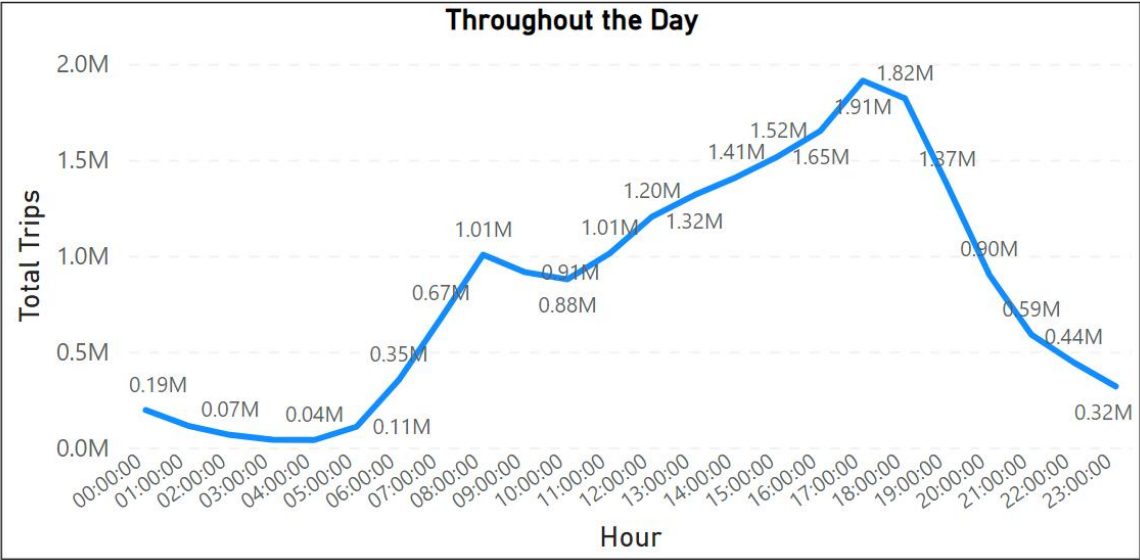
All

User Type

All

Ride Type

All



Star Schema (2020)

Performance analyzer ... >>

Start recording Refresh visuals Stop

Clear Export

Name	Duration (ms) ↓
Recording started (3/18/2024 2:17:10 PM)	-
Changed page	-
+ Hour Slicer	444
+ Hour Slicer	442
+ Throughout the Day	2961
+ Time Calculation Slicer	439
+ Start Station Slicer	439
+ Stop Station Slicer	438
+ Throughout Time	2493
+ User Type Slicer	436
+ Ride Type Slicer	436
+ Button	520
+ Button	521
+ Trips Taken	2365
+ Trips per Station End	2941
+ Trips per Start Station	2510
+ Map per Start Station	2906

Flat Table (2020)

Performance analyzer ... >>

Start recording Refresh visuals Stop

Clear Export

Name	Duration (ms) ↓
Recording started (3/18/2024 2:15:24 PM)	-
Changed page	-
+ Hour Slicer	371
+ Hour Slicer	370
+ Throughout the Day	2752
+ Time Calculation Slicer	367
+ Start Station Slicer	366
+ Stop Station Slicer	366
+ Throughout Time	2910
+ User Type Slicer	364
+ Ride Type Slicer	364
+ Button	444
+ Button	444
+ Trips Taken	2924
+ Trips per Station End	2640
+ Trips per Start Station	2780
+ Map per Start Station	3034

Trips Taken

213M

Time Taken

4bn

Dur AVG

16.90

Start Station	Trips Started
W 21 St & 6 Ave	1,141,210
West St & Chambers St	1,047,019
E 17 St & Broadway	1,042,618
Pershing Square North	1,019,837
Broadway & E 14 St	937,139

Stop Station	Trips Ended
W 21 St & 6 Ave	1,148,298
West St & Chambers St	1,086,118
E 17 St & Broadway	1,085,619
Pershing Square North	976,465
Broadway & E 14 St	935,530

Date

Last

1

Select

No filters applied

Hour

All

Region

All

Start Station

All

Stop Station

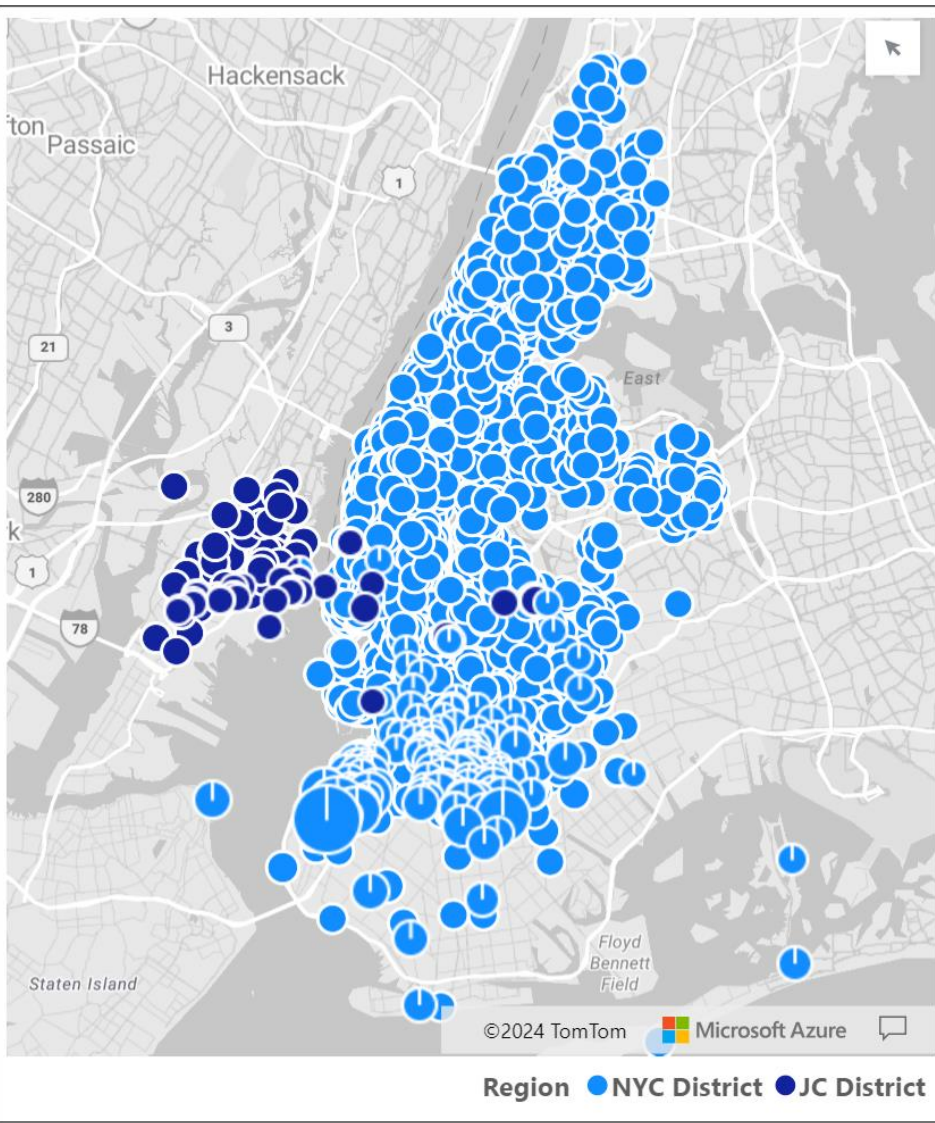
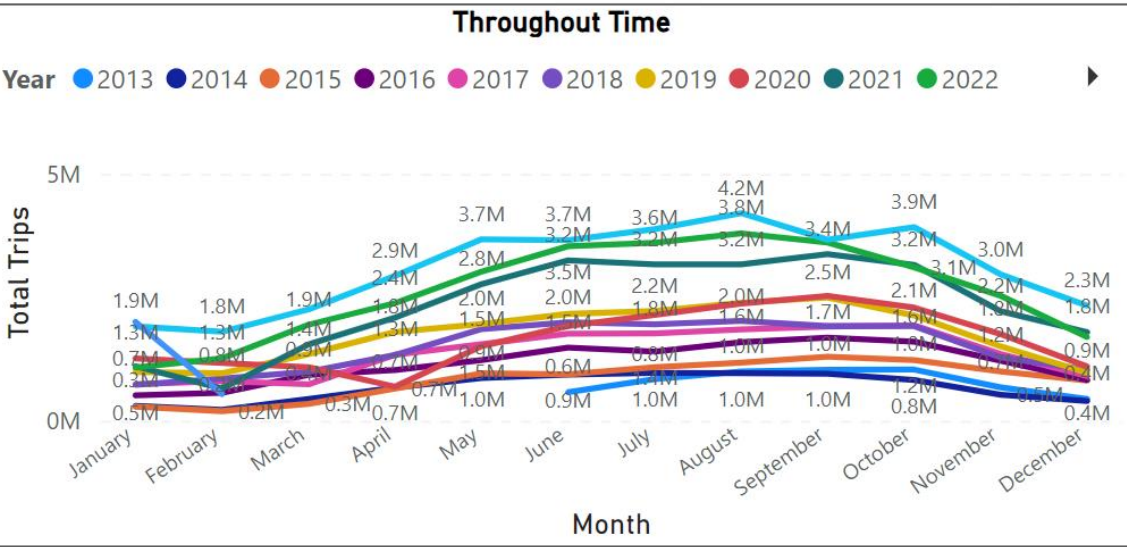
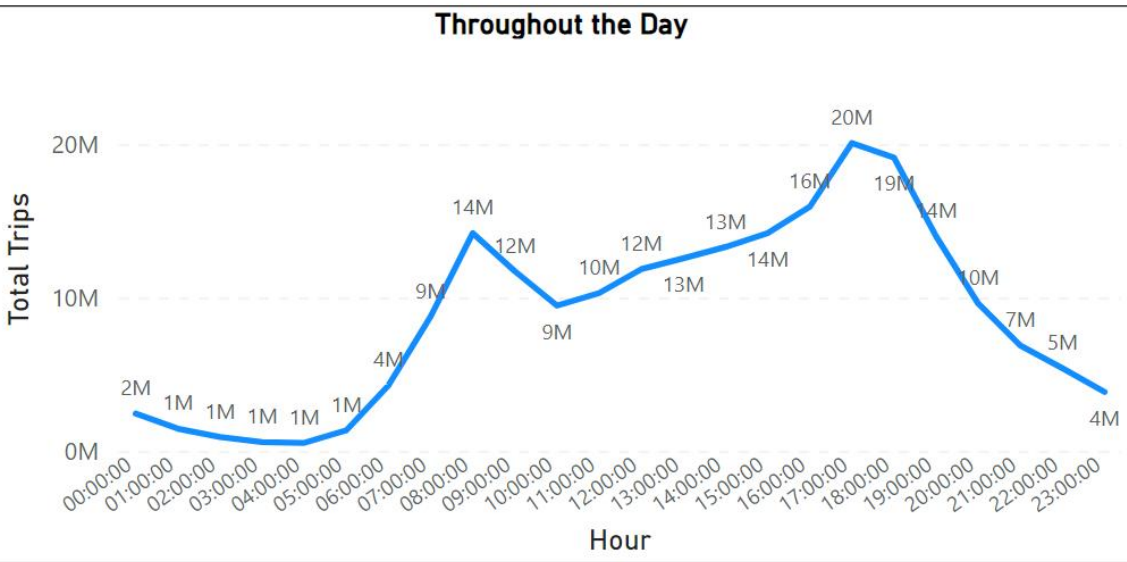
All

User Type

All

Ride Type

All



Clear all slicers

Apply all slicers

Star Schema (Full)

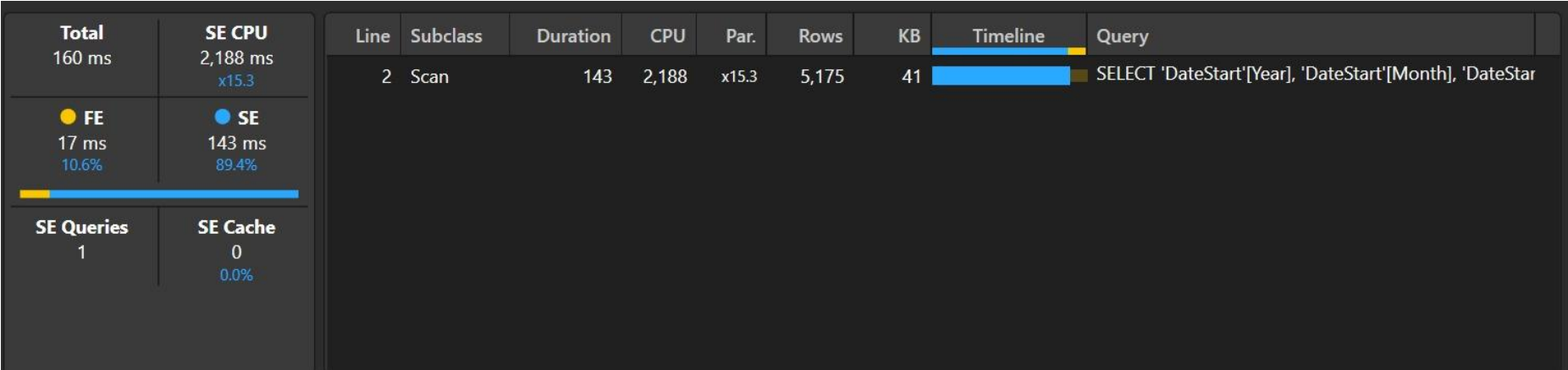
Performance analyzer		... >>
▶ Start recording	↻ Refresh visuals	⊞ Stop
◇ Clear		📄 Export
Name	Duration (ms) ↓	
🕒 Recording started (3/18/2024 2:22:00 PM)	-	
📄 Changed page	-	
+ Hour Slicer	339	
+ Hour Slicer	338	
+ Throughout the Day	2918	
+ Time Calculation Slicer	336	
+ Start Station Slicer	335	
+ Stop Station Slicer	334	
+ Throughout Time	2914	
+ User Type Slicer	333	
+ Ride Type Slicer	332	
+ Button	424	
+ Button	424	
+ Trips Taken	2503	
+ Trips per Station End	3406	
+ Trips per Start Station	3369	
+ Map per Start Station	3521	

Flat Table (Full)

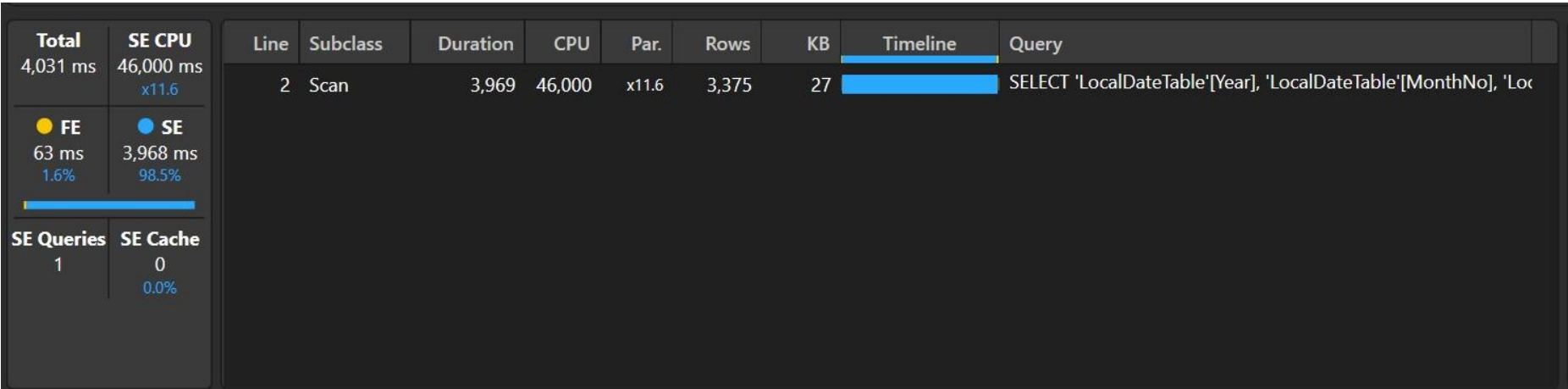
Performance analyzer		... >>
▶ Start recording	↻ Refresh visuals	⊞ Stop
◇ Clear		📄 Export
Name	Duration (ms) ↓	
🕒 Recording started (3/18/2024 2:23:27 PM)	-	
📄 Changed page	-	
+ Hour Slicer	470	
+ Hour Slicer	467	
+ Throughout the Day	7430	
+ Time Calculation Slicer	462	
+ Start Station Slicer	460	
+ Stop Station Slicer	459	
+ Throughout Time	7712	
+ User Type Slicer	447	
+ Ride Type Slicer	446	
+ Button	249	
+ Button	249	
+ Trips Taken	5066	
+ Trips per Station End	8653	
+ Trips per Start Station	5729	
+ Map per Start Station	198907	

Throughout Time - Graph

Star Schema (Full)



Flat Table (Full)



Map per Start Station - Graph

Star Schema (Full)

Total 1,016 ms	SE CPU 13,813 ms x14.5	Line	Subclass	Duration	CPU	Par.	Rows	KB	Timeline	Query
FE 63 ms 6.2%	SE 953 ms 93.8%	2	Scan	0	0		2,805	11	<div></div>	SELECT 'StartStation'[StartStationLatitude], 'StartStation'[StartSt
		4	Scan	0	0		2,804	11	<div></div>	SELECT 'StartStation'[StartStationLatitude], 'StartStation'[S
		6	Scan	953	13,813	x14.5	2,994	36	<div></div>	SELECT 'Region'[Region], 'StartStation'[StartStationLatitude], 'St
		8	Scan	0	0		2,799	33	<div></div>	SELECT 'StartStation'[StartStationLatitude], 'StartStation'[S
SE Queries 4	SE Cache 0 0.0%									

Flat Table (Full)

Total	SE CPU	Line	Subclass	Duration	CPU	Par.	Rows	KB	Timeline	Query											
181,563 ms	50,859 ms x2.9	2	Scan	17,797	50,859	x2.9	0,382,926	477,725		SELECT 'Trips'[start_station_latitude], 'Trips'[start_station_l											
<div><div></div><div>FE</div></div> 163,766 ms 90.2%	<div><div></div><div>SE</div></div> 17,797 ms 9.8%																				
<div><div></div><div>SE Queries</div></div> 1											<div><div></div><div>SE Cache</div></div> 0 0.0%										







Cost

Cost

- Full Refresh for Flat Table exceeds P1 allowance
- Consistently, the Star Schema consumes less CPU
 - During Refresh
 - During Ad-hoc queries
 - During Reporting

Bringing it all together

Overview of Metrics

	Star Schema	Flat Table
Refresh Time		
Model Size		
(Re)Usability		
Performance		
DAX Complexity		
Cost		

What about Lucky Number Seven?

Correct Results

Have you heard about 'AutoExists'?

- Applies to SUMMARIZECOLUMNS only
- When using multiple Filters on a single table
- AutoExists will treat it as a single Filter
- Can lead to WRONG results!

<https://www.sqlbi.com/articles/the-importance-of-star-schemas-in-power-bi/>

<https://www.sqlbi.com/articles/understanding-dax-auto-exist/>

<https://www.sqlbi.com/tv/auto-exist-on-clusters-or-numbers-unplugged-22/>

Data

Year ▲	Developer	Language
2016	Alberto	C#
2017	Daniele	C#
2017	Alberto	DAX
2017	Marco	DAX
2017	Daniele	Python
2018	Daniele	C#
2018	Marco	C#
2018	Alberto	DAX
2018	Marco	DAX

Credits: <https://www.sqlbi.com/articles/understanding-dax-auto-exist/>

```
1 # Projects = COUNTROWS ( Projects )
2
3 # Projects All Time = CALCULATE (
4     [# Projects],
5     ALL ( Projects[Year] )
6 )
```

 COPY

 CONVENTIONS

#2



Year

☒ 2017
☐ 2018

Language

☐ C#
☒ DAX
☒ Python

3

Projects

5

Projects All Time

Year

☐ 2017
☒ 2018

Language

☐ C#
☒ DAX
☒ Python

2

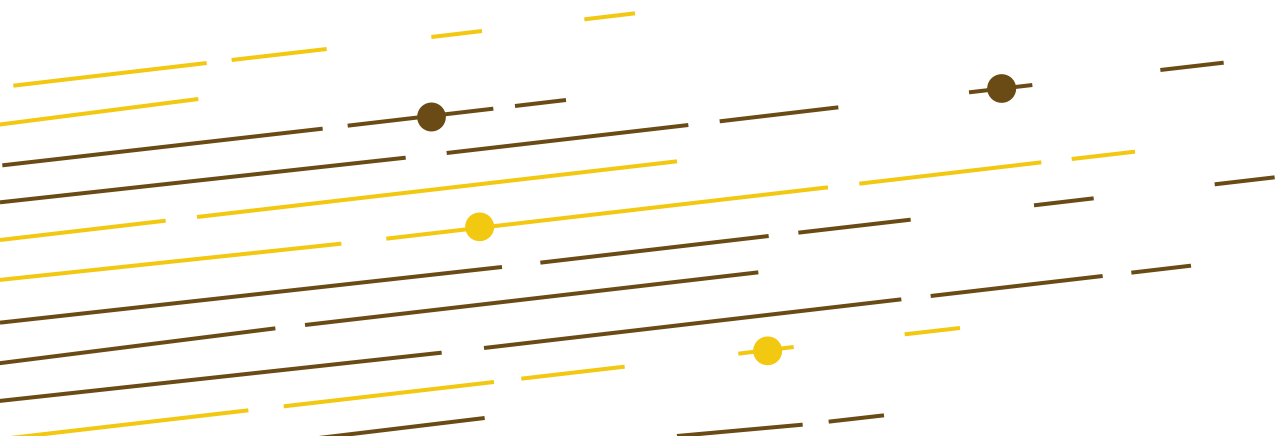
Projects

4

Projects All Time

Credits: <https://www.sqlbi.com/articles/understanding-dax-auto-exist/>

BONUS: What about a Galaxy?



Dealing with multiple Fact Tables

- Relationships between large tables do not scale well
 - Especially if they are considered Many to Many and Bi-Directional
 - Be cautious of surprising results
- Look for an approach with Conformed dimensions

Wrap Up

STAR SCHEMA

But make sure you understand why!



ALL THE MODELS

imgflip.com

Thanks, @KoVer!

Resources

- <https://learn.microsoft.com/en-us/power-bi/guidance/star-schema>
- <https://guyinacube.com/2021/02/24/why-power-bi-loves-a-star-schema/>
- <https://data-goblins.com/checklists>
- <https://www.sqlbi.com/articles/measuring-the-dictionary-size-of-a-column-correctly/>
- <https://www.sqlbi.com/articles/the-importance-of-star-schemas-in-power-bi/>
- <https://www.sqlbi.com/articles/power-bi-star-schema-or-single-table/>



Slides can be found at :

https://github.com/BenniDeJagere/Presentations/{Year}/{Date}_{Event}



Feedback





Thank you

