Analytics at scale with Power BI and Azure Synapse

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Data-Marc.com

Dave Ruijter

Solution Architect Data & Analytics
Blue Rocket Data Consulting & Solutions

✉ dave@blue-rocket.it

@DaveRuijter

linkedin.com/in/DaveRuijter

ModernData.ai

BLUE ROCKET DATA CONSULTING & SOLUTIONS

Data-Marc.com

Marc Lelijveld

Data & Analytics consultant
Macaw Netherlands

Microsoft Power Platform
FASTTRACK RECOGNIZED SOLUTION ARCHITECT
2022

MVP Microsoft® Most Valuable Professional

Microsoft MCT

✉ marc.lelijveld@outlook.com

@MarcLelijveld

in linkedin.com/in/MarcLelijveld

Data-Marc.com

BLUE ROCKET DATA CONSULTING & SOLUTIONS

Data-Marc.com

Agenda

- Challenges
  - Data Platform
  - Power BI
- Deep Dive
  - Data Platform
  - Power BI
- Hybrid Tables & Incremental Refresh
- Refresh Challenges & Orchestration
- Scaling

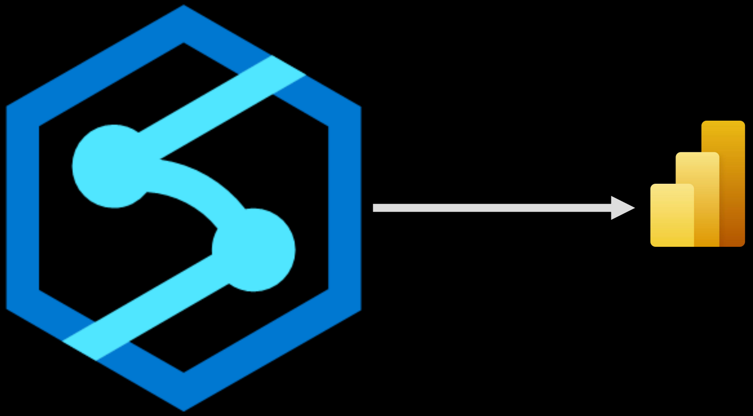
After this session

Design and implement	Orchestrate	Performance	Cost
Better design and implement complex data models, including hybrid tables, aggregations, and combined storage modes (import, DirectQuery , dual).	Orchestrate the end-to-end data processing, with a pipeline chain from data ingest in the data lake house to the incremental Power BI dataset refresh.	Use different techniques to identify performance bottlenecks in your solutions and how to solve those ("does it fold"?).	Implement a cost-efficient solution, that still meets the scalability demands.

Solution challenges

two separate worlds

Data Platform

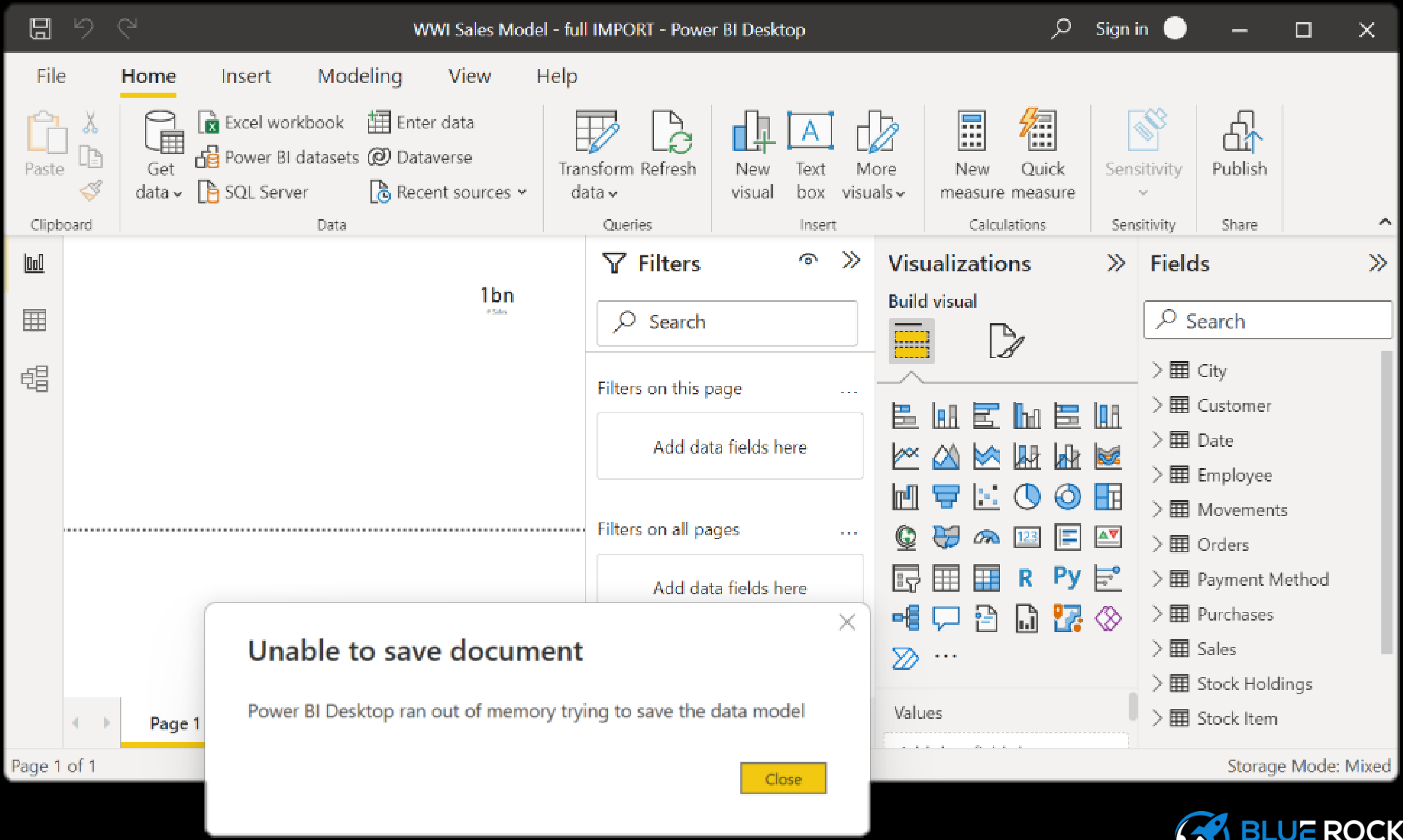


Power BI

Data Platform Solution Challenges

- Power BI report can't handle the volume of data
- Showing near real-time data in Power BI report

Power BI report can't handle the volume



Unable to save document

Power BI Desktop ran out of memory trying to save the data model

Close

BLUE ROCKET DATA CONSULTING & SOLUTIONS

Data-Marc.com

Should I put everything on DirectQuery instead?

DirectQuery limitations

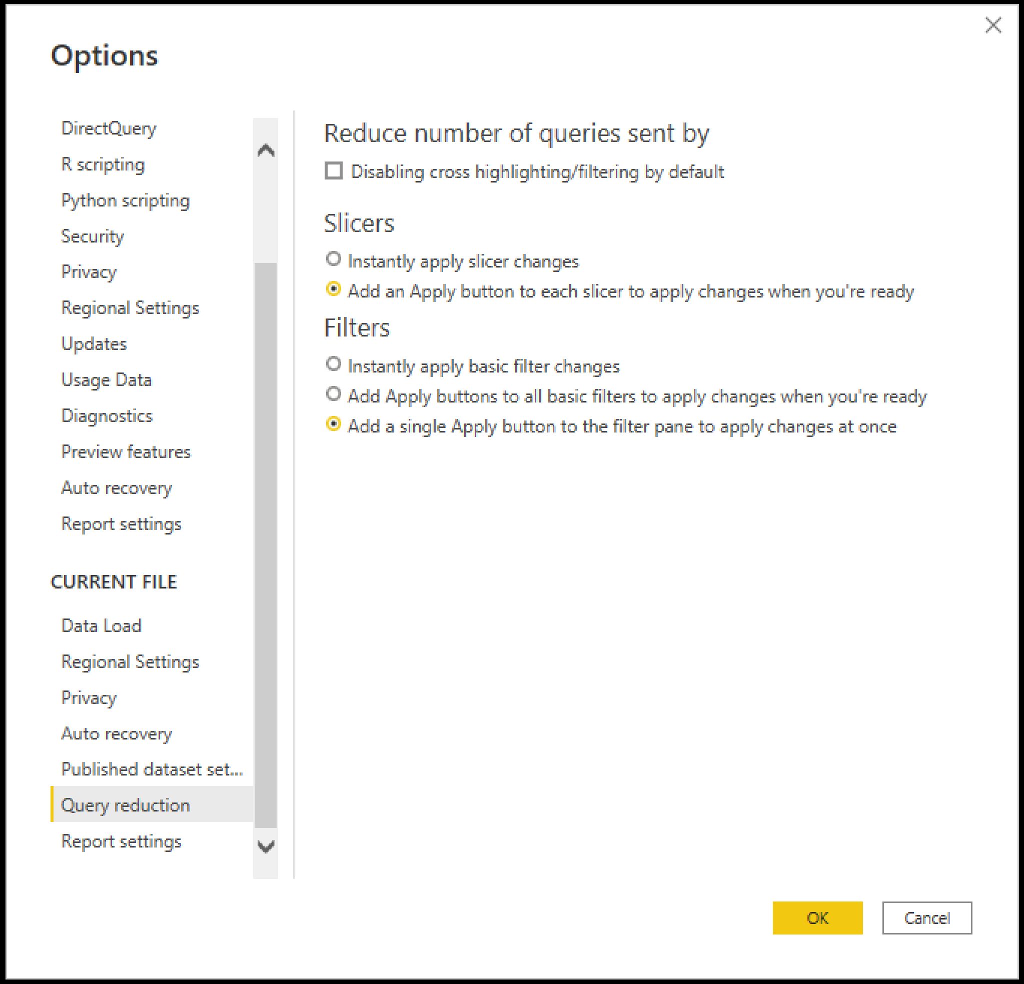
- Limited Power Query capabilities
- DirectQuery != streaming / live! Front-end still caches data
- No built-in date hierarchy (year/quarter/month/day)
- Lowest granularity data is seconds (no milliseconds)
- No parent-child support in DAX with *PATH()*
- Slower end user performance
- 1M row per query
- DAX limitations, only simple calculations possible

DirectQuery query reduction

Consider requesting to click **Apply** before queries are executed to the source

Applies to

- Slicers
- Filters (filter pane)



Reduce number of queries sent by

□ Disabling cross highlighting/filtering by default

Slicers

○ Instantly apply slicer changes

● Add an Apply button to each slicer to apply changes when you're ready

Filters

○ Instantly apply basic filter changes

○ Add Apply buttons to all basic filters to apply changes when you're ready

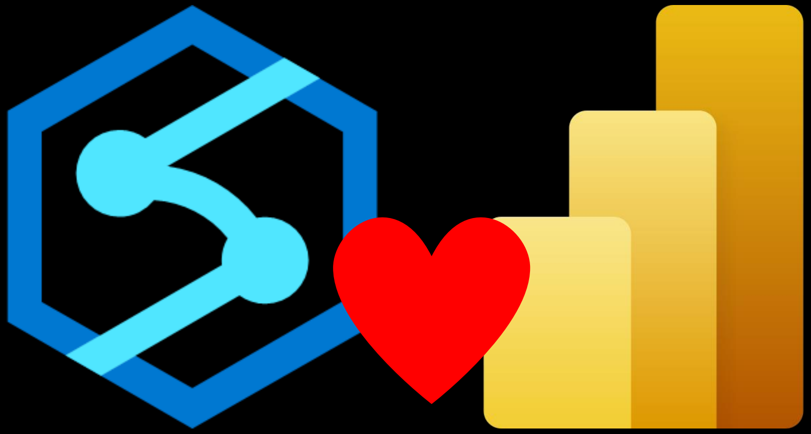
● Add a single Apply button to the filter pane to apply changes at once

Showing near real-time data in Power BI

- Refresh takes to long
- Poor end-user performance on DirectQuery
- Streaming datasets only allow one table
- Potentially queries are not foldable, therefore incremental refresh does not work (depending on source)

Data Platform Solution ~~Challenges~~

- Optimize Power BI model
- Use Aggregations
- Use Hybrid Table
- End-2-end orchestration

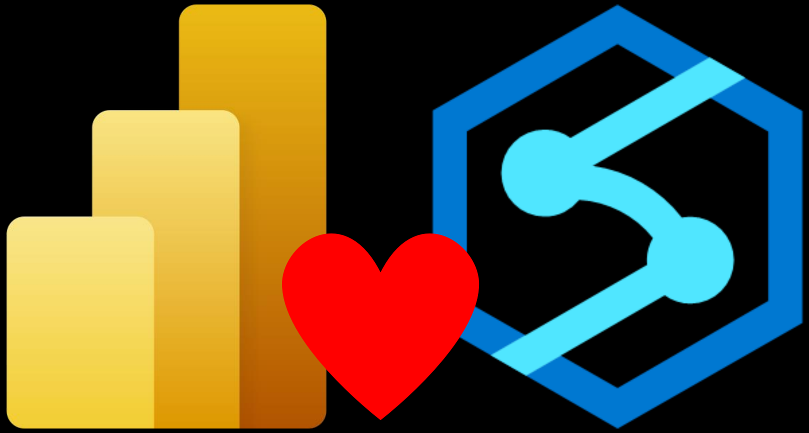


Power BI Solution Challenges

- Loading data from challenging sources (flat files, APIs)
- Some sources are manually maintained (like mapping tables)
- Data stored on decentralized storages, like SharePoint pages.
- Data from source systems are exported, rather than connected to analytical systems
- Store historical data in Power BI for trend analysis

Power BI Solution ~~Challenges~~

- Use Synapse Analytics to ingest data
  - Easily connect to various types of data sources
- Use Data Lakehouse: Bronze, Silver & Gold layers
  - Easily store historical data



Better together

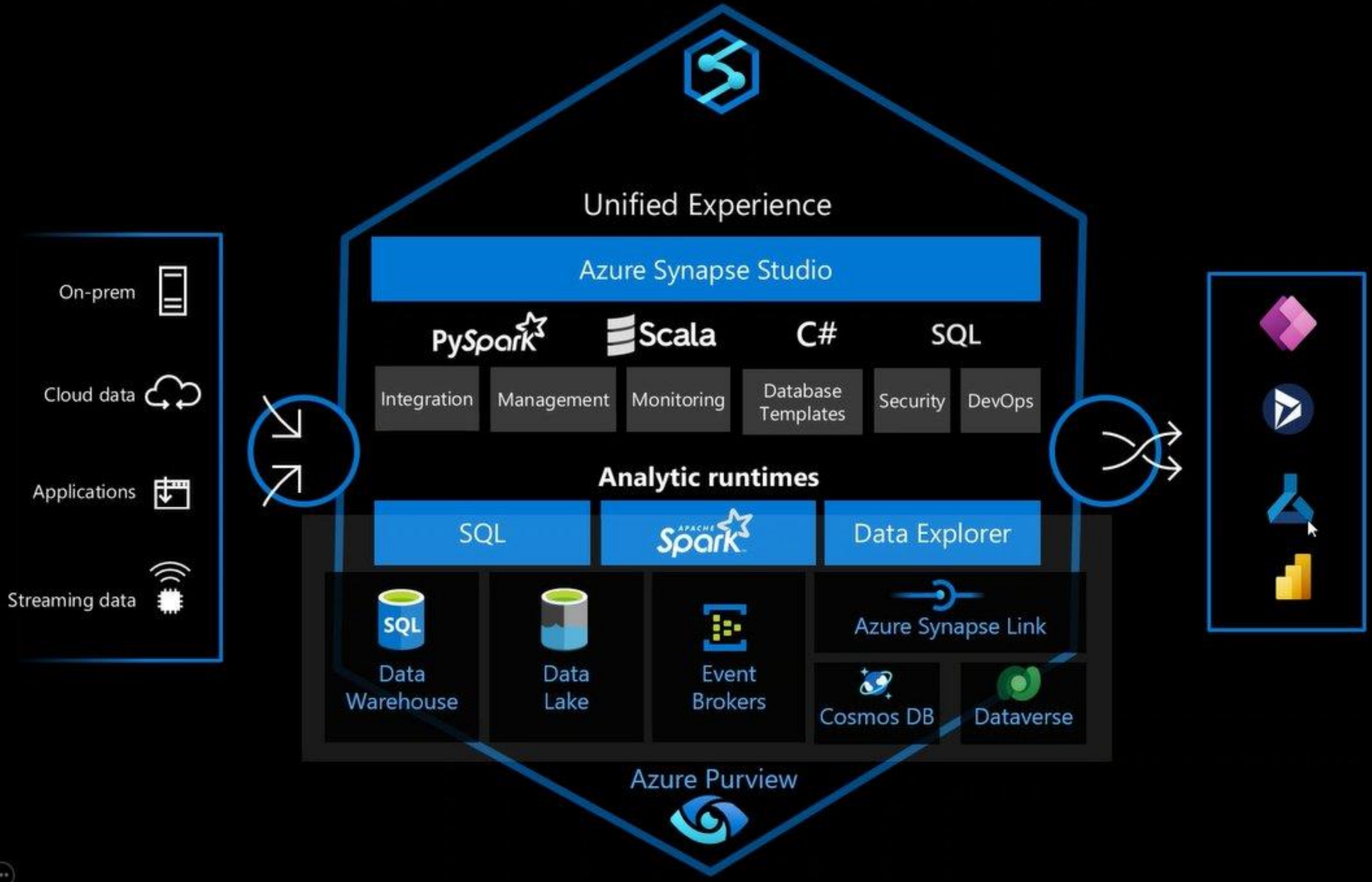
Data platform deep dive

Improvement areas

- Ingesting data from APIs using Synapse Analytics
- Store (historical!) data in the delta lakehouse architecture

Azure Synapse Analytics

The first unified, cloud native platform for converged analytics



On-prem

Cloud data

Applications

Streaming data

Unified Experience

Azure Synapse Studio

PySpark Scala C# SQL

Integration | Management | Monitoring | Database Templates | Security | DevOps

Analytic runtimes

SQL | Spark | Data Explorer

Data Warehouse

Data Lake

Event Brokers

Azure Synapse Link

Cosmos DB

Dataverse

Azure Purview

Ingesting data from APIs using Synapse Analytics

- Pipeline (same as ADF)
- Data Flow (same as ADF) *not to be confused with dataflows in PBI*
- Spark Notebook (4 languages available)
- Wrangling Dataflows (Same as ADF)

Store (historical!) data in the lake

- The Layered approach: Bronze, Silver & Gold

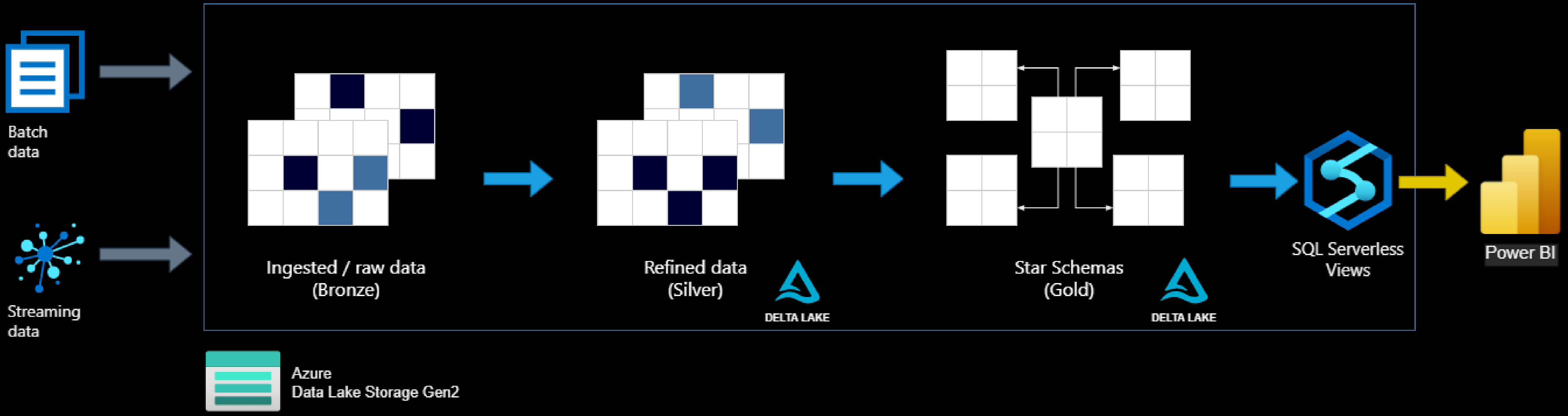
- Keep original raw data, build up history -> **bronze**
- Cleanse and refine data, standard file format -> **silver**
- Aggregate, prepare, transform, merge, make start schema -> **gold**

Store (historical!) data in the lake

- The Layered approach:
  Bronze, Silver & Gold
- Keep original raw data, build up history -> **bronze**
- Cleanse and refine data, standard file format -> **silver**
- Aggregate, prepare, transform, merge, make start schema -> **gold**

Bronze	**Silver**	**Gold**
Raw data All history, system replayable	Apply metadata Protect data (GDPR) Current & historical view	Implement business rules Fit for purpose

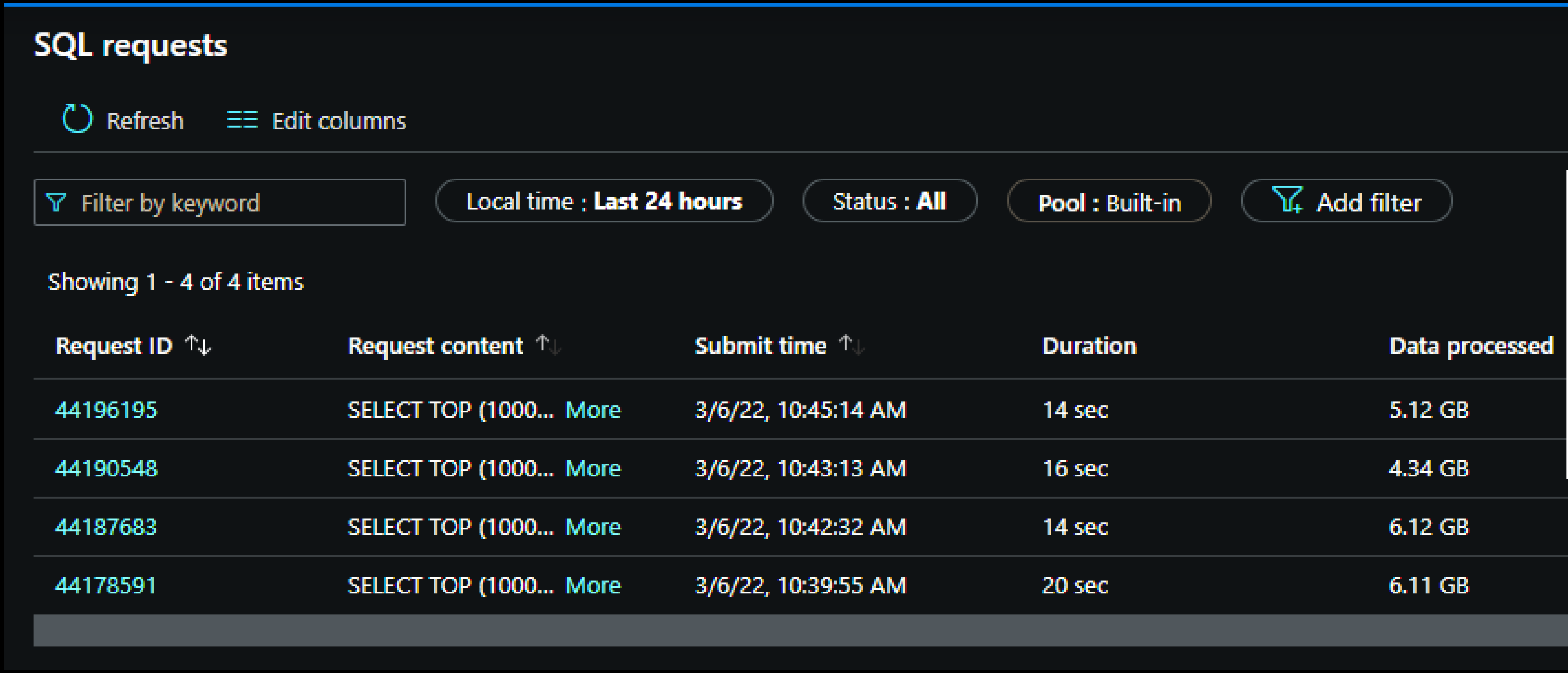
Using Delta Lakehouse with layered approach



Demo Data Platform

Lakehouse architecture

Warm-up time of Serverless SQL pools

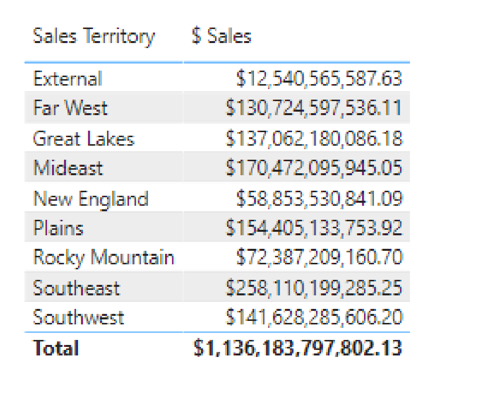


SQL requests

Local time : **Last 24 hours** Status : **All** Pool : Built-in

Showing 1 - 4 of 4 items

Request ID	Request content	Submit time	Duration	Data processed
44196195	SELECT TOP (1000... More	3/6/22, 10:45:14 AM	14 sec	5.12 GB
44190548	SELECT TOP (1000... More	3/6/22, 10:43:13 AM	16 sec	4.34 GB
44187683	SELECT TOP (1000... More	3/6/22, 10:42:32 AM	14 sec	6.12 GB
44178591	SELECT TOP (1000... More	3/6/22, 10:39:55 AM	20 sec	6.11 GB



Sales Territory	$ Sales
External	$12,540,565,587.63
Far West	$130,724,597,536.11
Great Lakes	$137,062,180,086.18
Mideast	$170,472,095,945.05
New England	$58,853,530,841.09
Plains	$154,405,133,753.92
Rocky Mountain	$72,387,209,160.70
Southeast	$258,110,199,285.25
Southwest	$141,628,285,606.20
Total	**$1,136,183,797,802.13**

Best practices for serverless SQL pools

- Azure AD Pass-through Authentication performance <= shared access signature credentials
- Colocate
- Same region
- Convert large CSV and JSON files to Parquet
- Try to optimize storage layout by using partitioning and keeping your files in the range between 100 MB and 10 GB
- Use appropriate data types (smallest, integer-based, varchar)
- Use filename and filepath functions to target specific partitions

Better together

Power BI deep dive

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

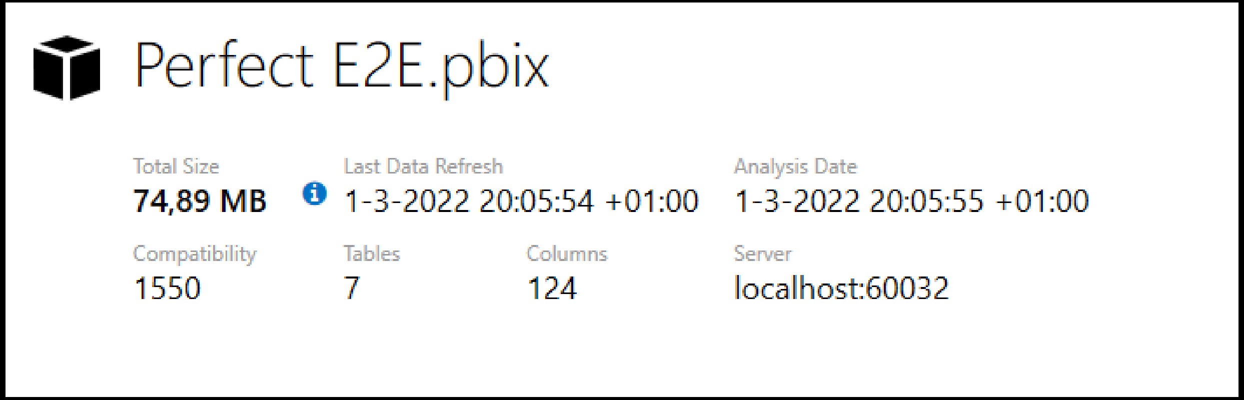
Data-Marc.com

Improvement areas

- Data model
- Query Folding
- Aggregations
- Storage modes
- Hybrid tables

But before we start changing our solution, let’s measure…

- Refresh durations
- Model Size
- Vertipaq Analyzer
- Performance Analyzer
- Query folding applied?



Perfect E2E.pbix

Total Size
74,89 MB

Last Data Refresh
1-3-2022 20:05:54 +01:00

Analysis Date
1-3-2022 20:05:55 +01:00

Compatibility
1550

Tables
7

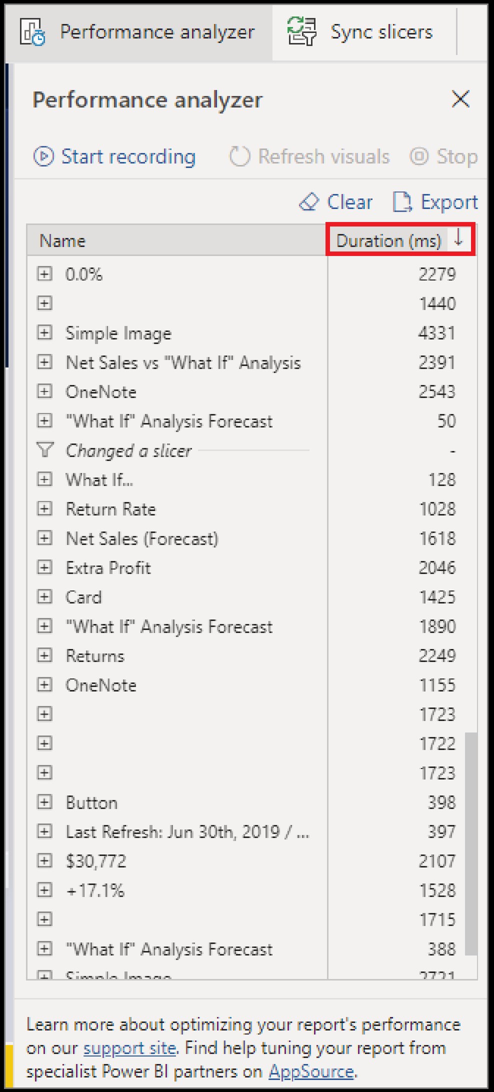
Columns
124

Server
localhost:60032

BLUE ROCKET DATA CONSULTING & SOLUTIONS

Data-Marc.com

Performance analyzer in Power BI Desktop



Performance analyzer

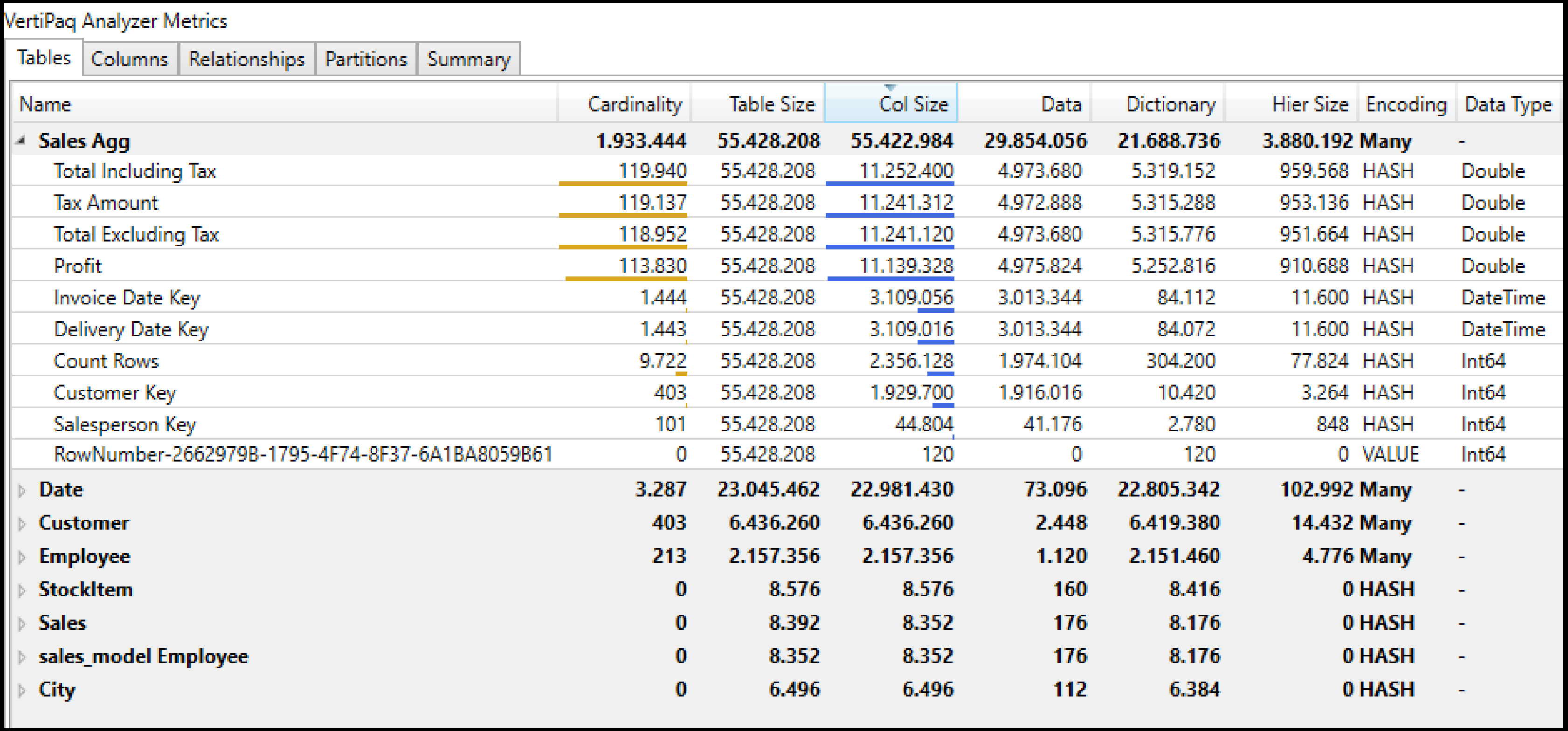
Name	Duration (ms)
0.0%	2279
	1440
Simple Image	4331
Net Sales vs "What If" Analysis	2391
OneNote	2543
"What If" Analysis Forecast	50
Changed a slicer	-
What If...	128
Return Rate	1028
Net Sales (Forecast)	1618
Extra Profit	2046
Card	1425
"What If" Analysis Forecast	1890
Returns	2249
OneNote	1155
	1723
	1722
	1723
Button	398
Last Refresh: Jun 30th, 2019 / ...	397
$30,772	2107
+17.1%	1528
	1715
"What If" Analysis Forecast	388
Simple Image	2721

Learn more about optimizing your report's performance on our support site. Find help tuning your report from specialist Power BI partners on AppSource.

- DAX Query
- Visual Display
- Other
  - Preparing queries
  - Waiting for other visuals to complete
  - Other background processes

Vertipaq analyzer

See where your data volume is



VertiPaq Analyzer Metrics

Name	Cardinality	Table Size	Col Size	Data	Dictionary	Hier Size	Encoding	Data Type
Sales Agg	**1.933.444**	**55.428.208**	**55.422.984**	**29.854.056**	**21.688.736**	**3.880.192**	**Many**	-
Total Including Tax	119.940	55.428.208	11.252.400	4.973.680	5.319.152	959.568	HASH	Double
Tax Amount	119.137	55.428.208	11.241.312	4.972.888	5.315.288	953.136	HASH	Double
Total Excluding Tax	118.952	55.428.208	11.241.120	4.973.680	5.315.776	951.664	HASH	Double
Profit	113.830	55.428.208	11.139.328	4.975.824	5.252.816	910.688	HASH	Double
Invoice Date Key	1.444	55.428.208	3.109.056	3.013.344	84.112	11.600	HASH	DateTime
Delivery Date Key	1.443	55.428.208	3.109.016	3.013.344	84.072	11.600	HASH	DateTime
Count Rows	9.722	55.428.208	2.356.128	1.974.104	304.200	77.824	HASH	Int64
Customer Key	403	55.428.208	1.929.700	1.916.016	10.420	3.264	HASH	Int64
Salesperson Key	101	55.428.208	44.804	41.176	2.780	848	HASH	Int64
RowNumber-2662979B-1795-4F74-8F37-6A1BA8059B61	0	55.428.208	120	0	120	0	VALUE	Int64
Date	**3.287**	**23.045.462**	**22.981.430**	**73.096**	**22.805.342**	**102.992**	**Many**	-
Customer	**403**	**6.436.260**	**6.436.260**	**2.448**	**6.419.380**	**14.432**	**Many**	-
Employee	**213**	**2.157.356**	**2.157.356**	**1.120**	**2.151.460**	**4.776**	**Many**	-
StockItem	**0**	**8.576**	**8.576**	**160**	**8.416**	**0**	**HASH**	-
Sales	**0**	**8.392**	**8.352**	**176**	**8.176**	**0**	**HASH**	-
sales_model Employee	**0**	**8.352**	**8.352**	**176**	**8.176**	**0**	**HASH**	-
City	**0**	**6.496**	**6.496**	**112**	**6.384**	**0**	**HASH**	-

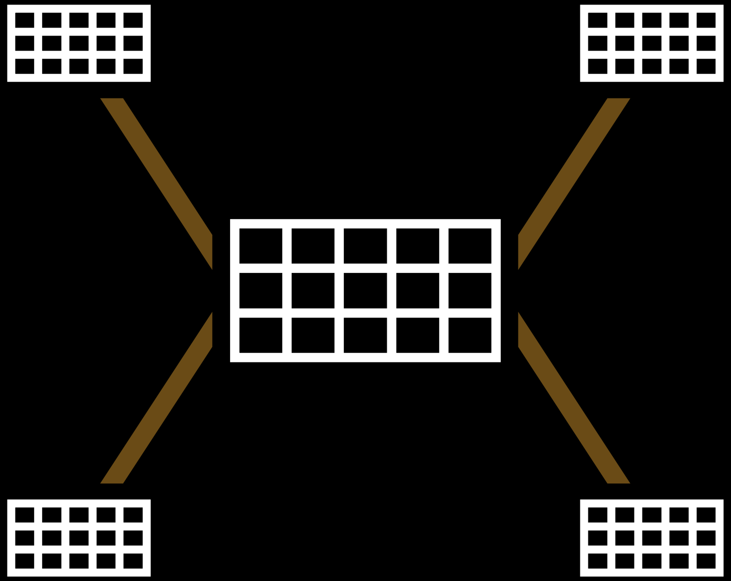
Star schema all the things!

Facts

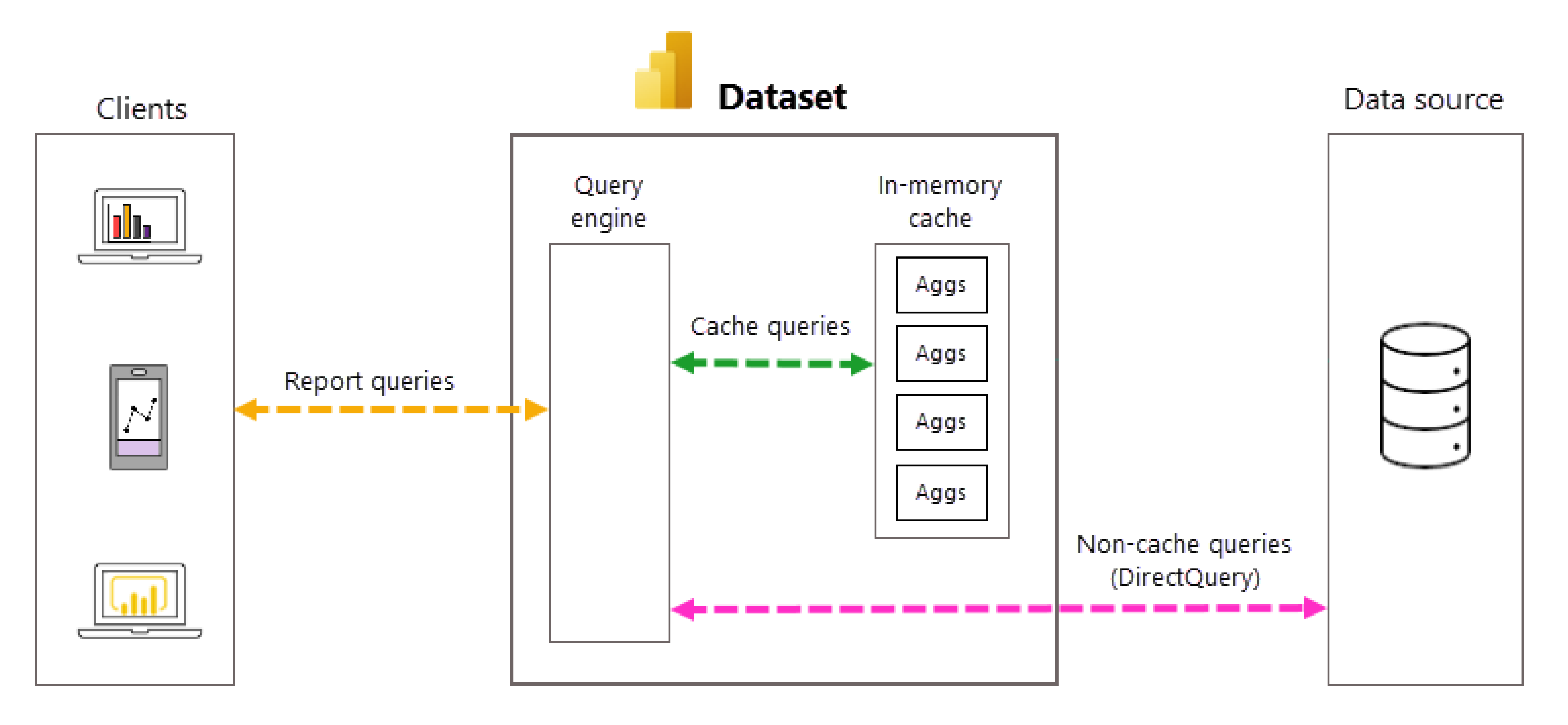
- Contains numerical information about a business process or items to be aggregated
- Aggregations provide totals, averages, etc.
  Power BI implements these using Measures
- Usefulness limited without context
  Context is provided by dimensions that slice the data

Dimensions

- Contains descriptive information that define how a fact should roll up.
- Examples: Date, Month, Customer, Geography, Product, Payment type.
- Without dimensions there is no context.
- Also called: Lookup table on steroids.



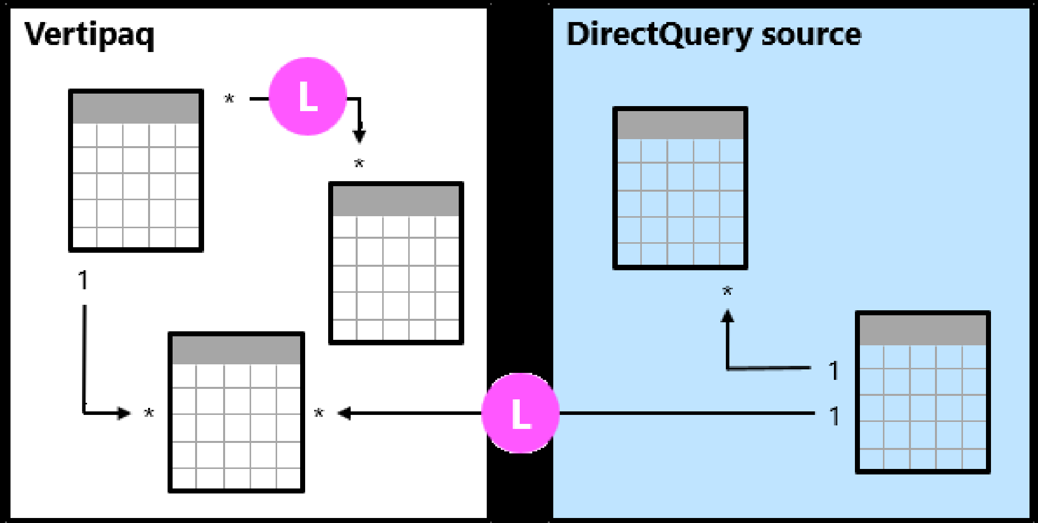
Aggregations



Relationships and storage modes

A model relationship is *limited* when there's no guaranteed "one" side. It can be the case for three reasons:

- The relationship uses a Many-to-many cardinality type (even if one or both columns contain unique values)
- The storage mode combination is Import and DirectQuery
- The relationship is cross source group



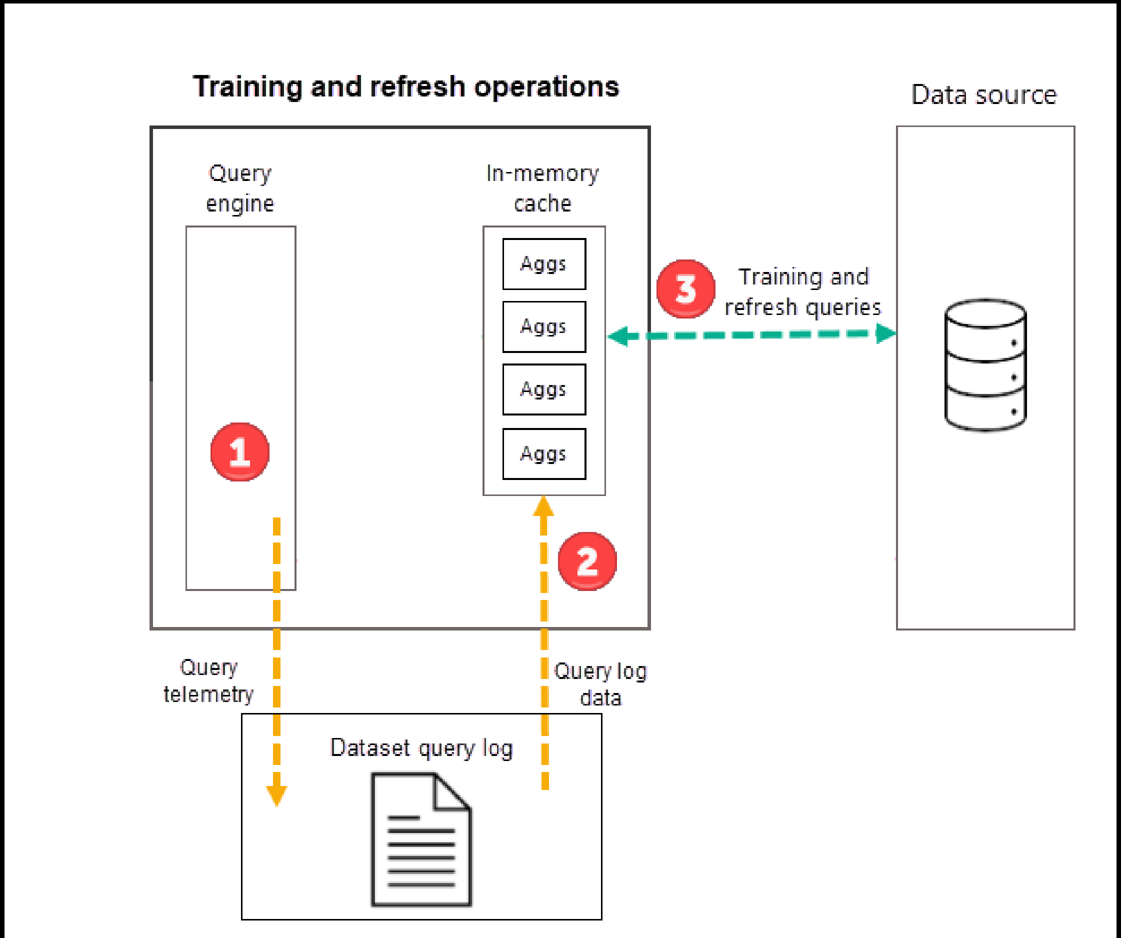
Automatic Aggregations

Power BI Premium per User, Premium Capacity and Embedded datasets

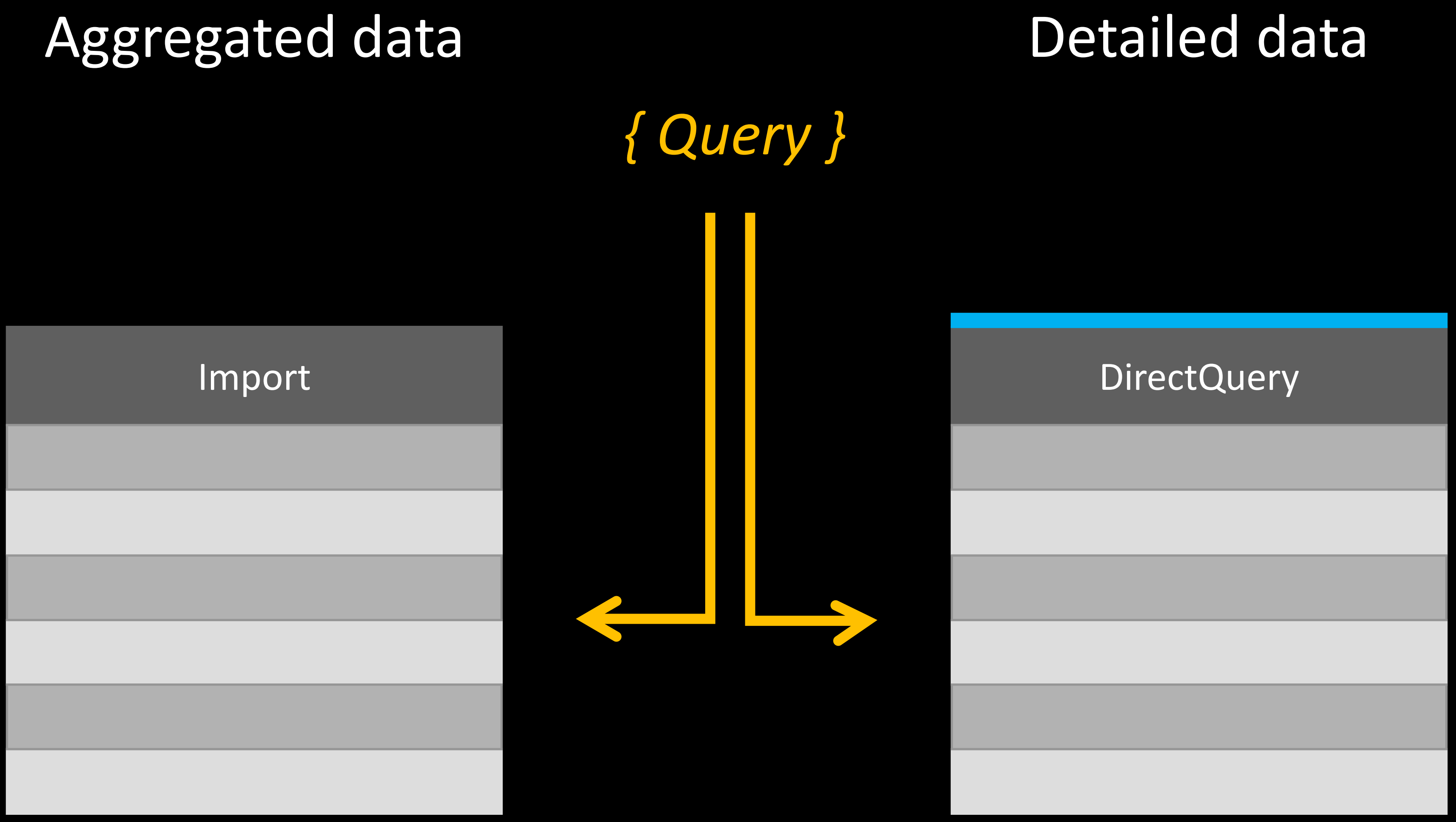
Automatic aggregations based on Query logs (7 days)

Supported sources:

- Azure SQL Database
- Azure Synapse Dedicated SQL pool
- Google BigQuery
- Snowflake



Demo Aggregations



Aggregated data

Detailed data

{ Query }

Import

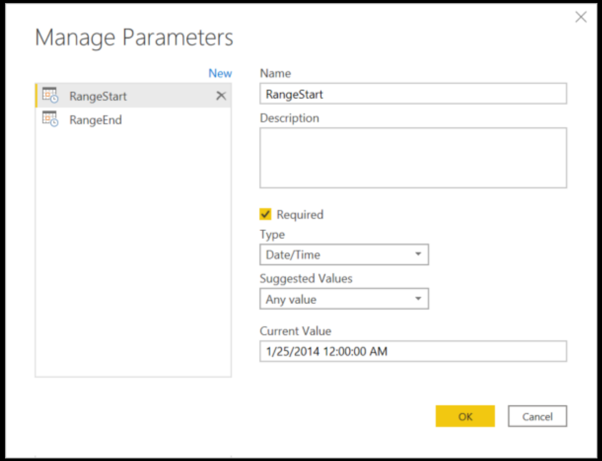
DirectQuery

Hybrid tables & incremental refresh

Incremental refresh

- Incremental refresh is supported for Power BI Premium, Premium per user, **Power BI Pro**, and Power BI Embedded datasets.
- Getting the latest data in **real time** with DirectQuery is **only supported for Power BI Premium**, Premium per user, and Power BI Embedded datasets.

Incremental refresh config



Manage Parameters

New

RangeStart

RangeEnd

Name: RangeStart

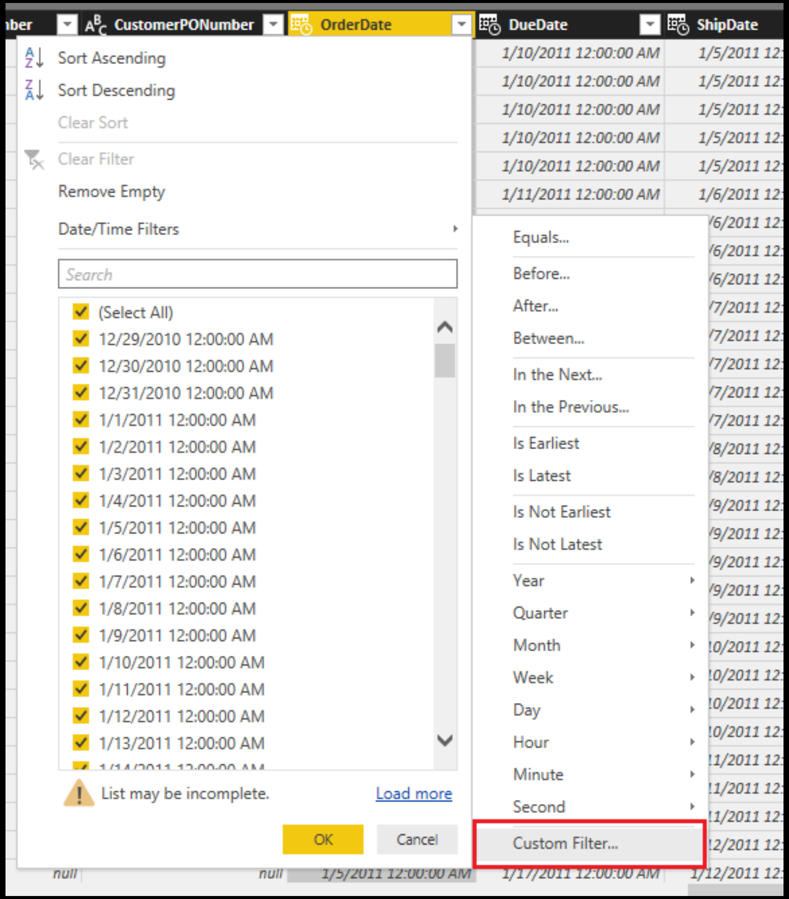
Description:

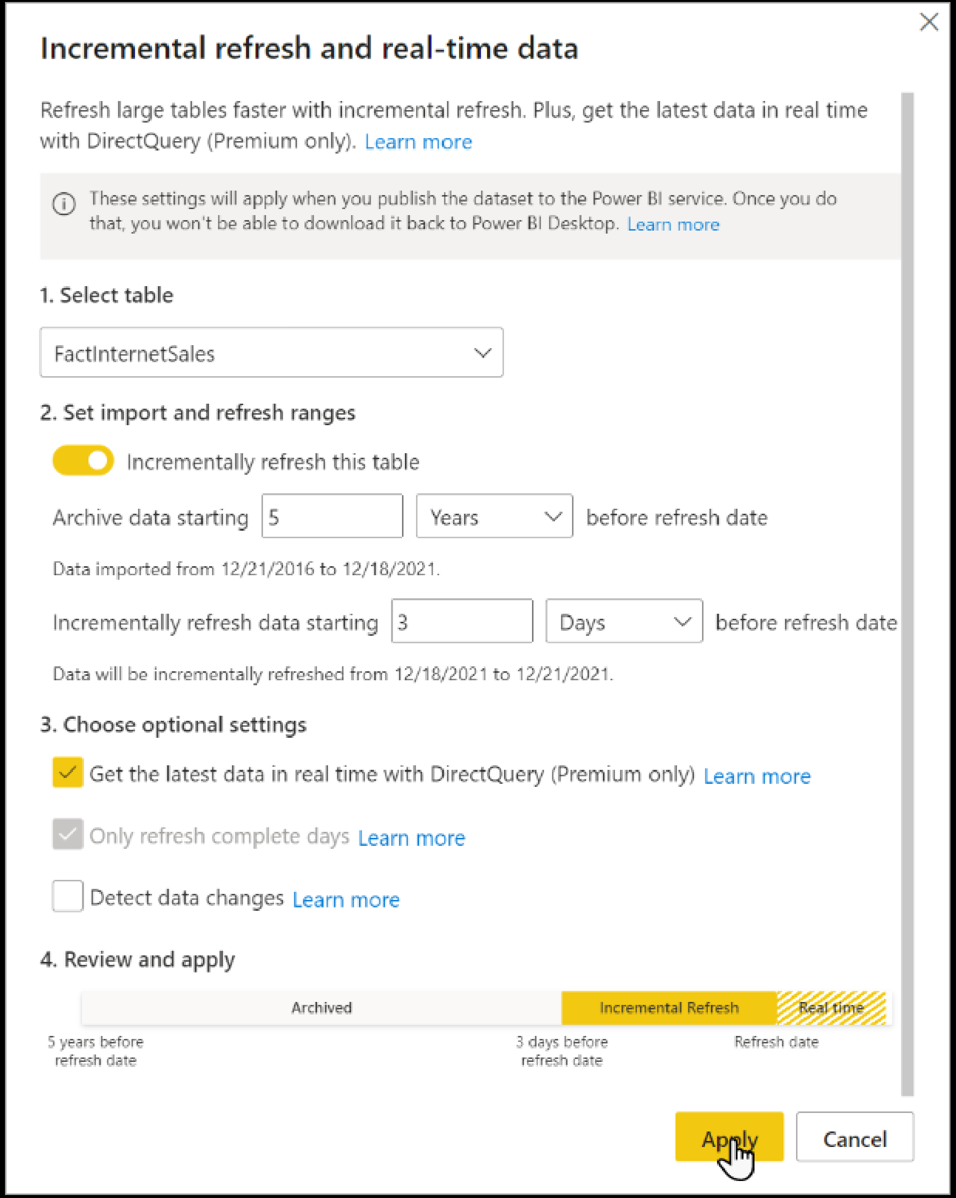
☑ Required

Type: Date/Time

Suggested Values: Any value

Current Value: 1/25/2014 12:00:00 AM





Incremental refresh and real-time data

Refresh large tables faster with incremental refresh. Plus, get the latest data in real time with DirectQuery (Premium only). Learn more

ⓘ These settings will apply when you publish the dataset to the Power BI service. Once you do that, you won't be able to download it back to Power BI Desktop. Learn more

1. Select table

FactInternetSales

2. Set import and refresh ranges

● Incrementally refresh this table

Archive data starting 5 Years before refresh date

Data imported from 12/21/2016 to 12/18/2021.

Incrementally refresh data starting 3 Days before refresh date

Data will be incrementally refreshed from 12/18/2021 to 12/21/2021.

3. Choose optional settings

☑ Get the latest data in real time with DirectQuery (Premium only) Learn more

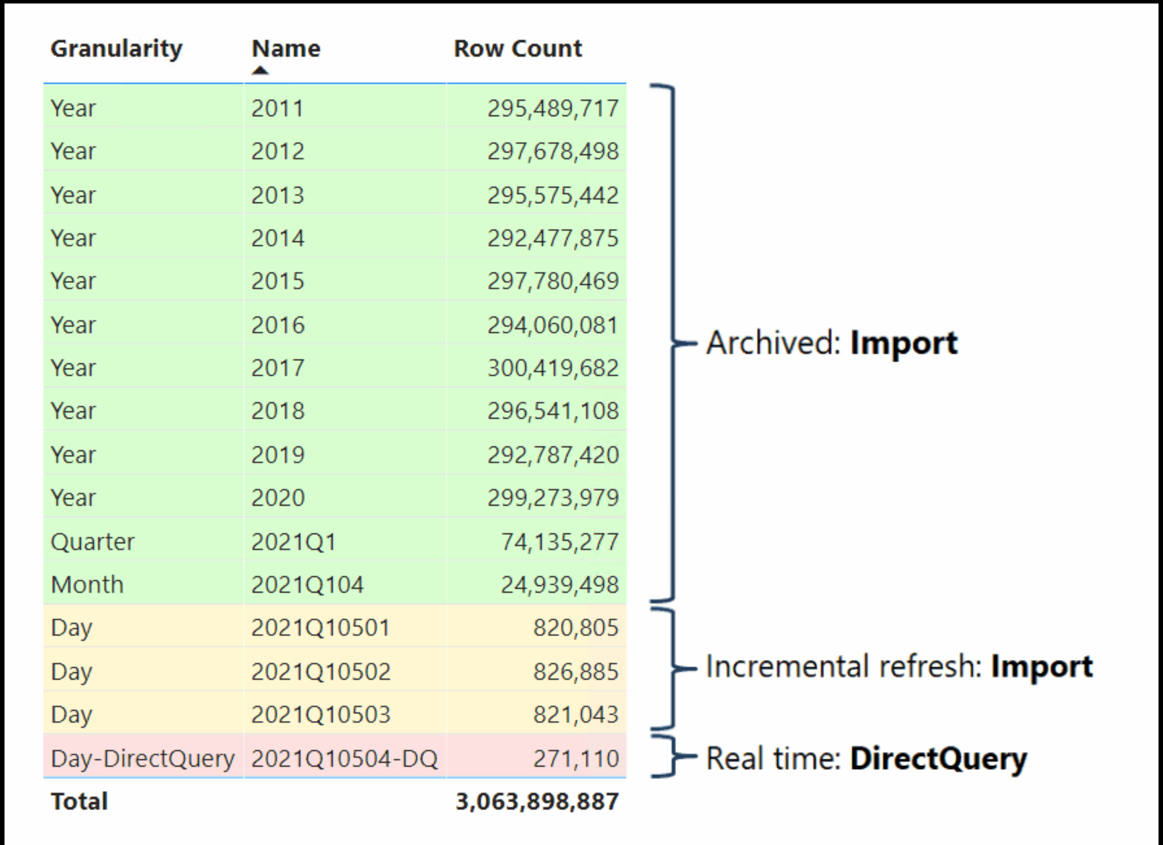
☑ Only refresh complete days Learn more

☐ Detect data changes Learn more

4. Review and apply

Hybrid tables

- Live / Realtime data in Power BI
- Combines different storage modes on partition level in a single table
- Goes hand-in-hand with Incremental Refresh



Granularity	Name	Row Count
Year	2011	295,489,717
Year	2012	297,678,498
Year	2013	295,575,442
Year	2014	292,477,875
Year	2015	297,780,469
Year	2016	294,060,081
Year	2017	300,419,682
Year	2018	296,541,108
Year	2019	292,787,420
Year	2020	299,273,979
Quarter	2021Q1	74,135,277
Month	2021Q104	24,939,498
Day	2021Q10501	820,805
Day	2021Q10502	826,885
Day	2021Q10503	821,043
Day-DirectQuery	2021Q10504-DQ	271,110
Total		3,063,898,887

Archived: **Import**

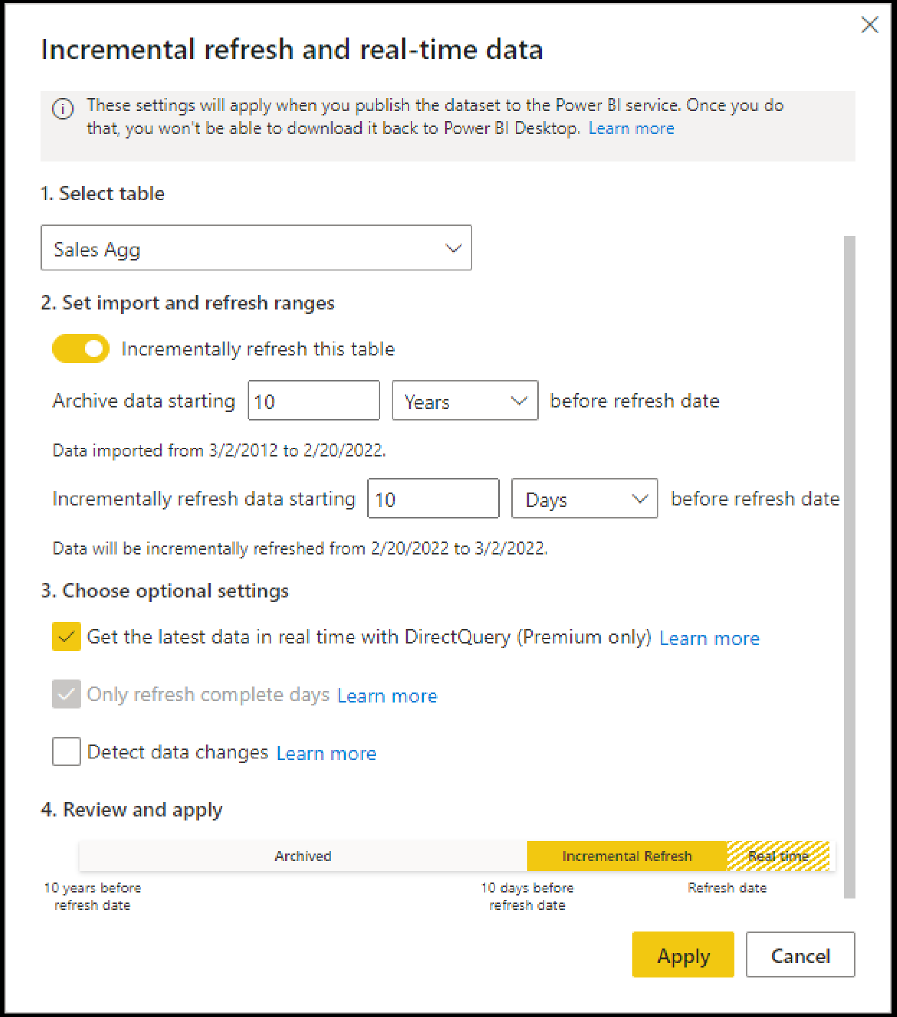
Incremental refresh: **Import**

Real time: **DirectQuery**

Hybrid tables

- Implementation with Incremental Refresh
- Customizable via 3rd party tooling like Tabular Editor

>> Limitation: Only 1 DQ partition per table allowed at the moment.



Incremental refresh and real-time data

ⓘ These settings will apply when you publish the dataset to the Power BI service. Once you do that, you won't be able to download it back to Power BI Desktop. Learn more

1. Select table

Sales Agg

2. Set import and refresh ranges

● Incrementally refresh this table

Archive data starting 10 Years before refresh date

Data imported from 3/2/2012 to 2/20/2022.

Incrementally refresh data starting 10 Days before refresh date

Data will be incrementally refreshed from 2/20/2022 to 3/2/2022.

3. Choose optional settings

☑ Get the latest data in real time with DirectQuery (Premium only) Learn more

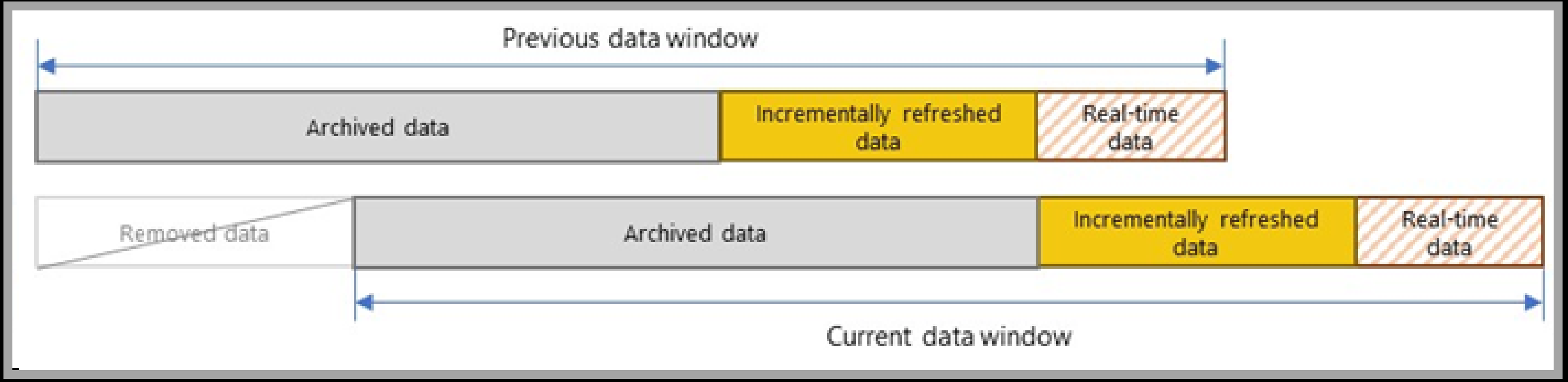
☑ Only refresh complete days Learn more

☐ Detect data changes Learn more

4. Review and apply

Hybrid tables – what challenge does it solve?

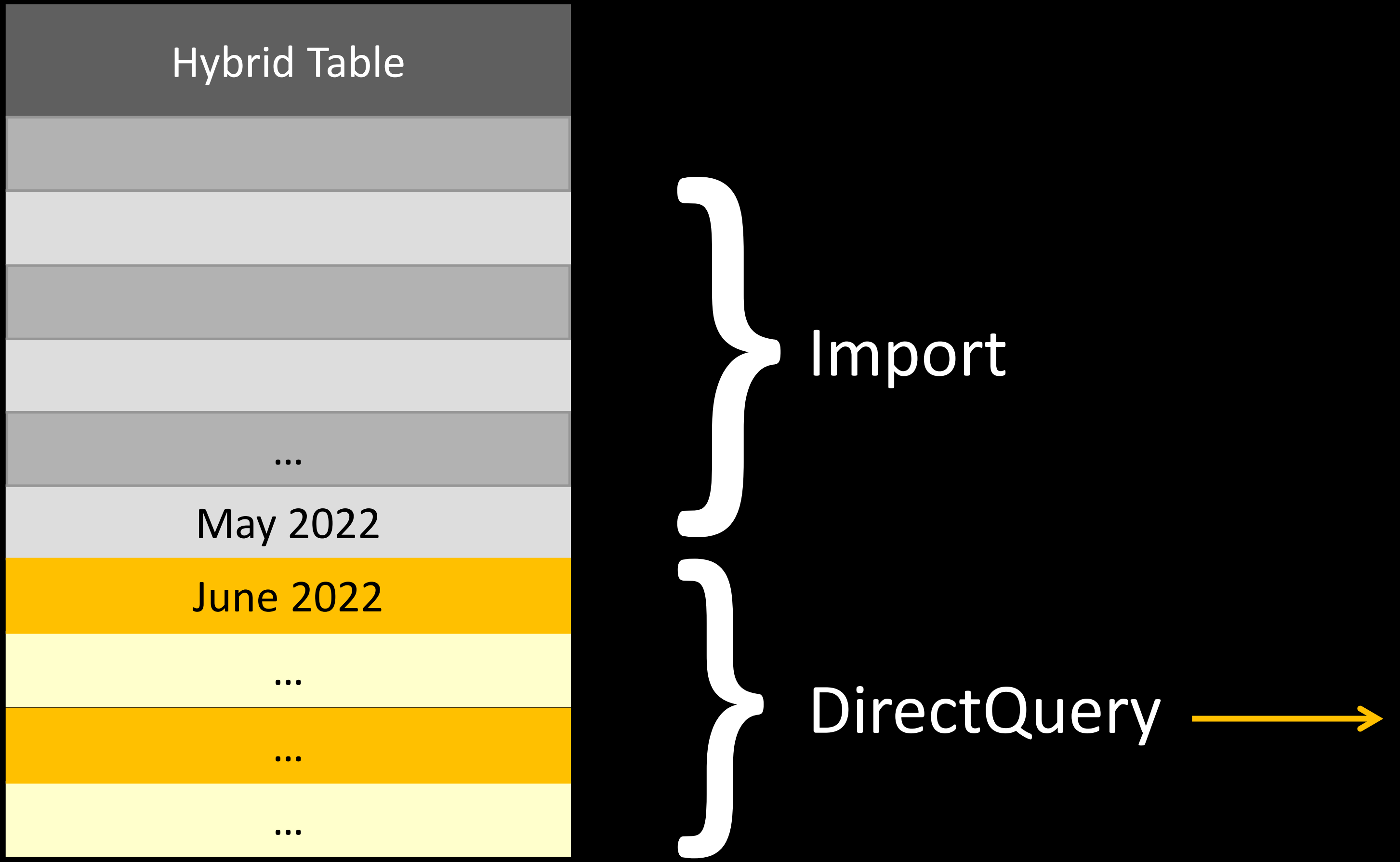
- Realtime scenarios without full tables on DQ mode
- No complex refresh mechanisms needed with partition refresh and queries over XMLA
- No more multiple tables and complex DAX to combine to achieve the same goal



Hybrid tables – Keep in mind that…

- Premium feature
- DAX restrictions for DirectQuery apply
- Limited Power Query capabilities (due to DQ)
- Requires Large Dataset Format (storage) in workspace
- Performance hit on upstream data sources

Demo Hybrid Tables
Latest data in real time



Hybrid Table

…

May 2022

June 2022

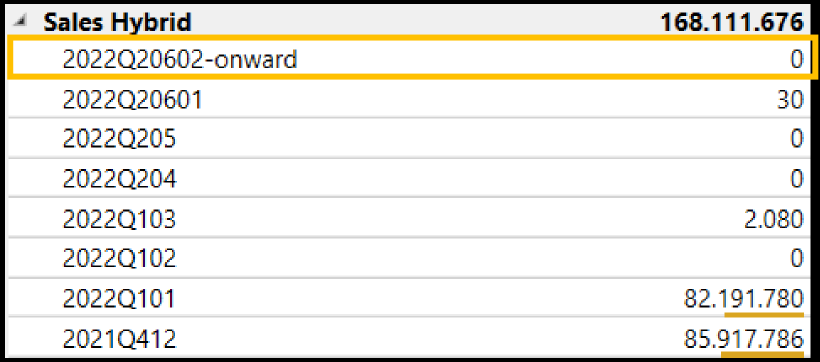
…

…

…

Import

DirectQuery



Sales Hybrid	168.111.676
2022Q20602-onward	0
2022Q20601	30
2022Q205	0
2022Q204	0
2022Q103	2.080
2022Q102	0
2022Q101	82.191.780
2021Q412	85.917.786

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Data-Marc.com

Refresh challenges

Refresh options

- Scheduled in the service
- Manual trigger
- Power Automate
- PowerShell
- API

Effective refreshing

Considerations

- Refreshing the entire model takes too long with high load on sources
- Can we only refresh certain tables?
- Can we only refresh certain partitions?
- Can we use DQ tables/partitions (Hybrid Tables)

What do we need?

- Async refresh API
- XMLA Endpoints

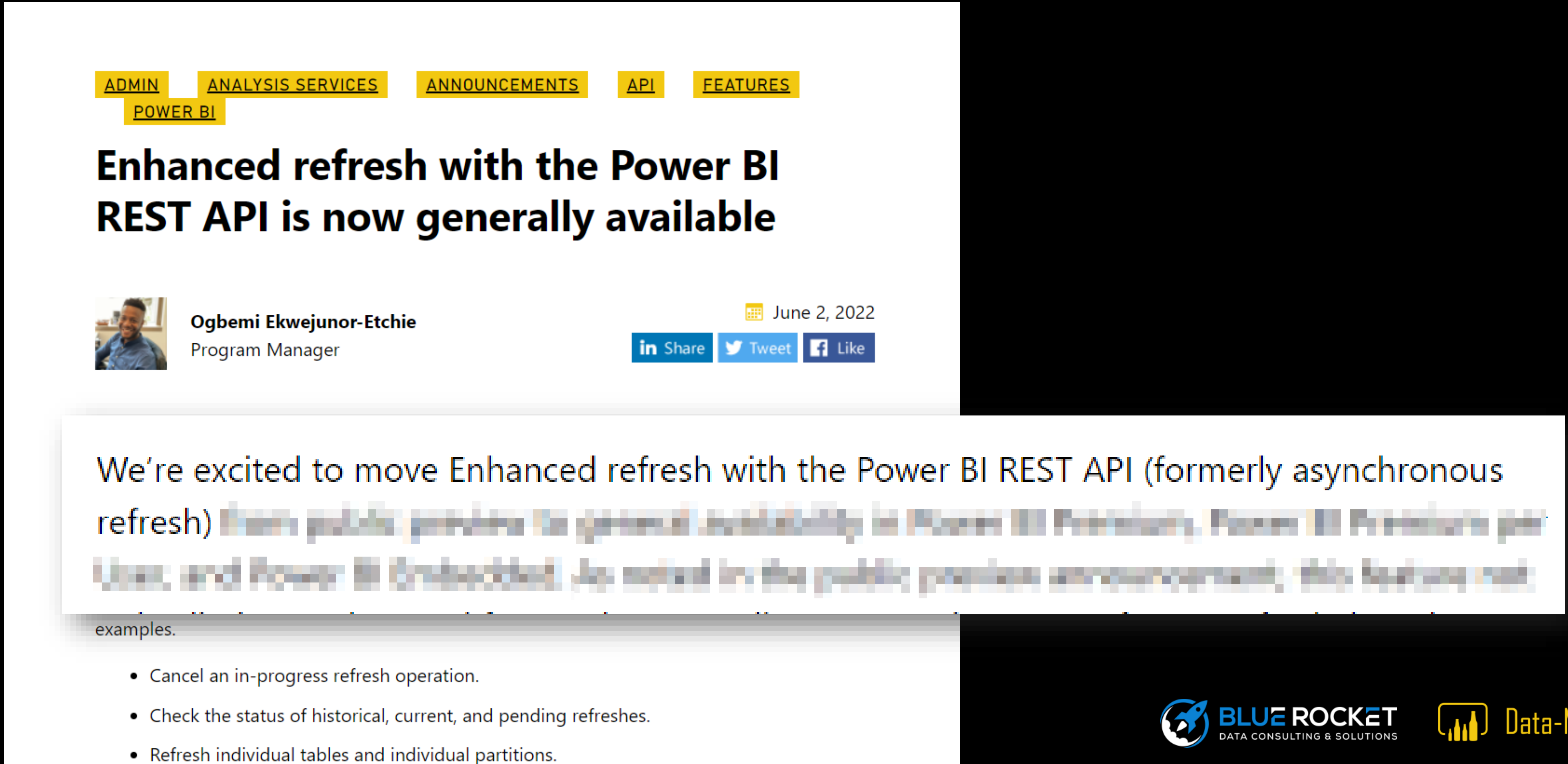
Async refresh API

- Specify the objects to refresh

POST
https://api.powerbi.com/v1.0/myorg/groups/f089354e-8366-4e18-aea3-4cb4a3a50b48/datasets/cfafbeb1-8037-4d0c-896e-a46fb27ff229/refreshes

```
{
   "type": "Full",
   "commitMode": "transactional",
   "maxParallelism": 2,
   "retryCount": 2,
   "objects": [
      {
         "table": "DimCustomer",
         "partition": "DimCustomer"
      },
      {
         "table": "DimDate"
      }
   ]
}
```

Wait a sec…



ADMIN ANALYSIS SERVICES ANNOUNCEMENTS API FEATURES POWER BI

Enhanced refresh with the Power BI REST API is now generally available

Ogbemi Ekwejunor-Etchie
Program Manager

June 2, 2022

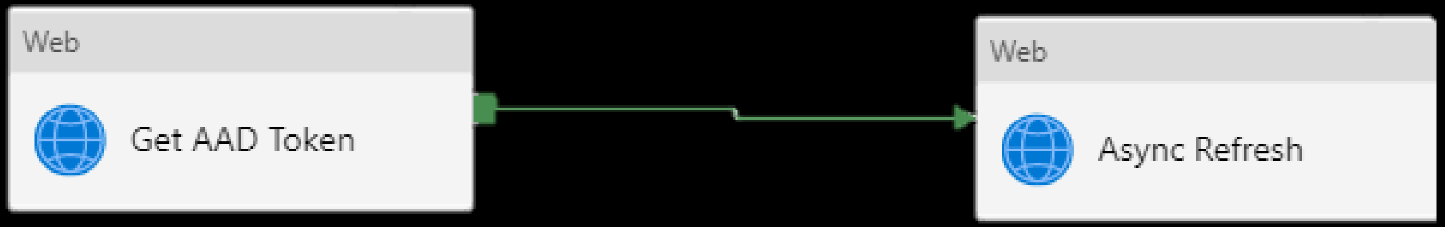
We're excited to move Enhanced refresh with the Power BI REST API (formerly asynchronous refresh) [redacted]

examples.

- Cancel an in-progress refresh operation.
- Check the status of historical, current, and pending refreshes.
- Refresh individual tables and individual partitions.

Central E2E orchestration

- Combine pipelines from Data Platform with Power BI
- Lowest latency between source and report
- Consider including backup operations for Power BI
- Incremental loading where possible

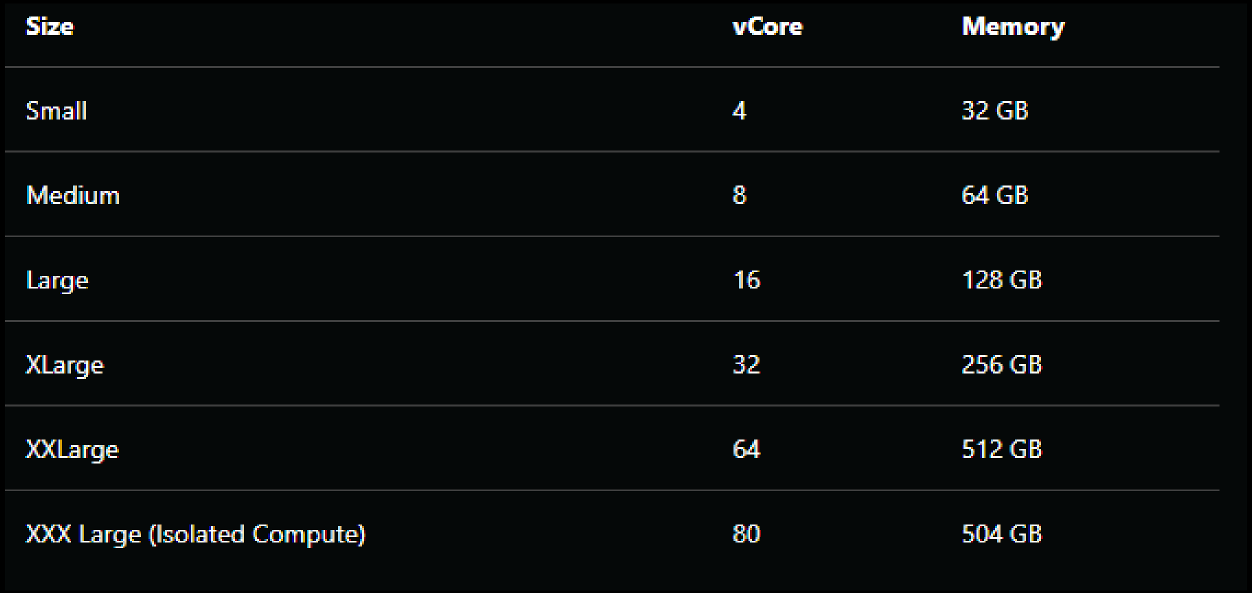


Demo end-2-end orchestration

Scaling

Scaling data platform

- Spark Cluster:
  - Use multiple cluster configs
  - Autoscale -> *It can take 1 to 5 minutes for a scaling operation to complete*
  - Dynamic allocation of executors
  - Automatic pause



Size	vCore	Memory
Small	4	32 GB
Medium	8	64 GB
Large	16	128 GB
XLarge	32	256 GB
XXLarge	64	512 GB
XXX Large (Isolated Compute)	80	504 GB

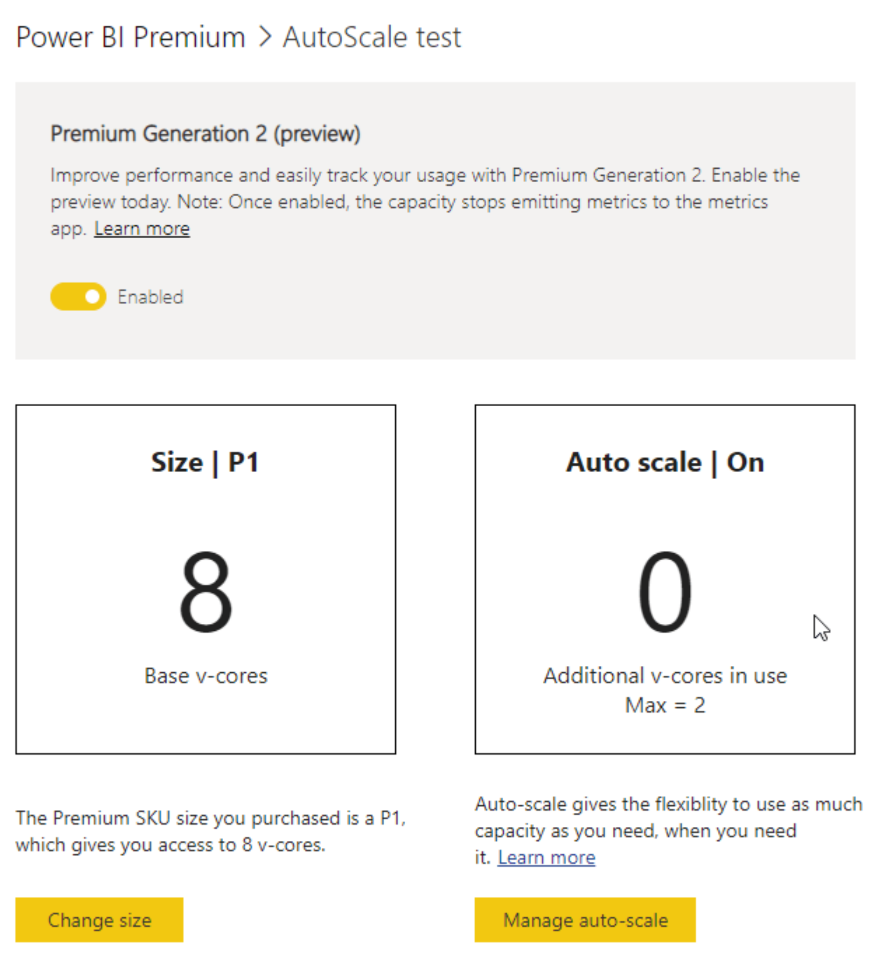
Auto-scale (gen2 only)

Auto-scale adds:

- Additional vCores
- Applies for at least 24h

Configured through

- Max. number of scalable vCores
- Azure subscription – Pay as you go



Power BI Premium > AutoScale test

Premium Generation 2 (preview)

Improve performance and easily track your usage with Premium Generation 2. Enable the preview today. Note: Once enabled, the capacity stops emitting metrics to the metrics app. Learn more

Enabled

Size | P1

8

Base v-cores

The Premium SKU size you purchased is a P1, which gives you access to 8 v-cores.

Auto scale | On

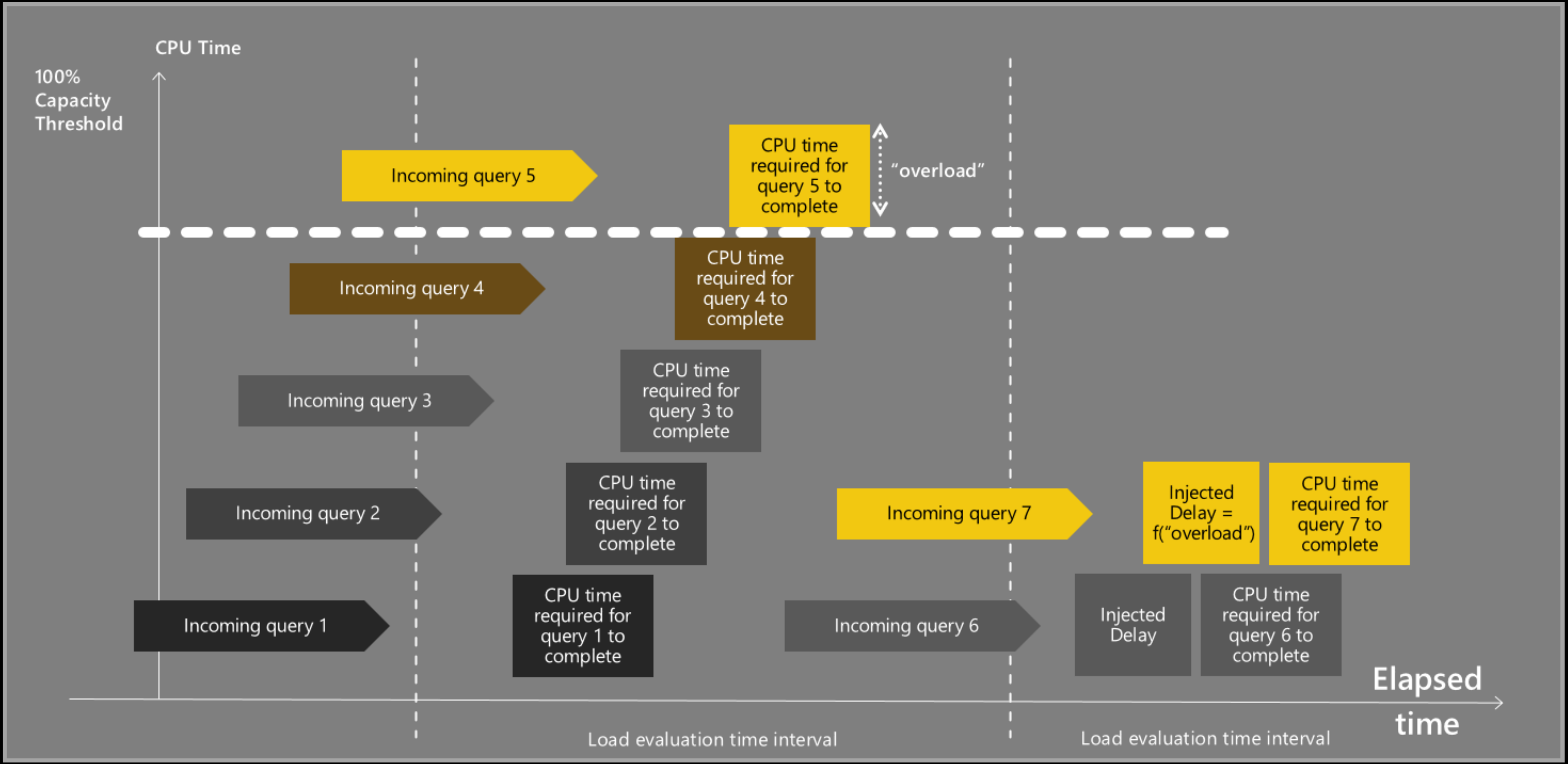
0

Additional v-cores in use
Max = 2

Auto-scale gives the flexiblity to use as much capacity as you need, when you need it. Learn more

BLUE ROCKET DATA CONSULTING & SOLUTIONS Data-Marc.com

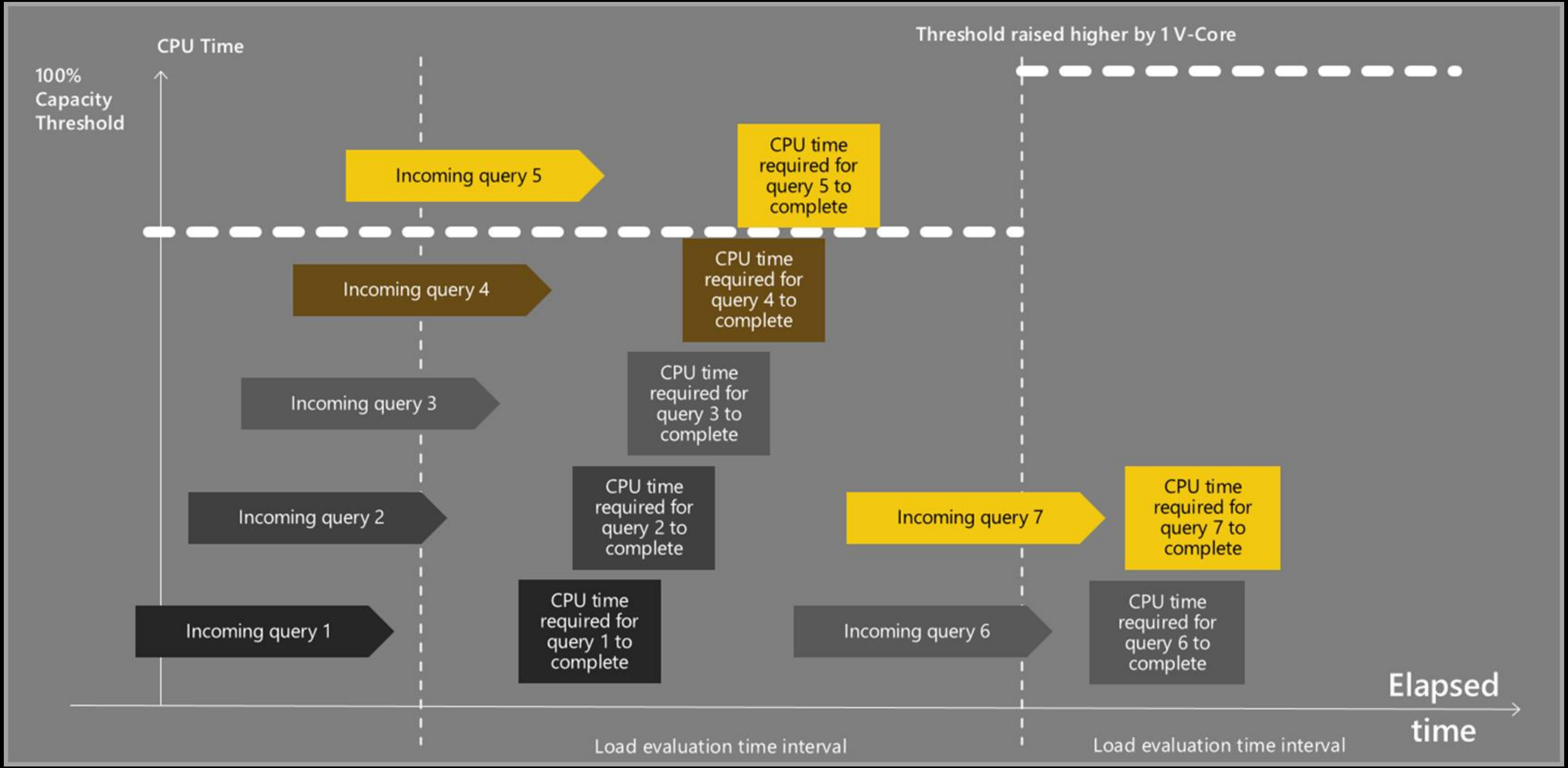
Power BI Premium – Auto scale


CPU Time
100%
Capacity
Threshold
Incoming query 5
CPU time required for query 5 to complete
"overload"
Incoming query 4
CPU time required for query 4 to complete
Incoming query 3
CPU time required for query 3 to complete
Incoming query 2
CPU time required for query 2 to complete
Incoming query 7
Injected Delay = f("overload")
CPU time required for query 7 to complete
Incoming query 1
CPU time required for query 1 to complete
Incoming query 6
Injected Delay
CPU time required for query 6 to complete
Elapsed time
Load evaluation time interval
Load evaluation time interval

BLUE ROCKET
DATA CONSULTING & SOLUTIONS

Data-Marc.com

Power BI Premium – Auto scale



CPU Time
100% Capacity Threshold
Threshold raised higher by 1 V-Core
Incoming query 5
CPU time required for query 5 to complete
Incoming query 4
CPU time required for query 4 to complete
Incoming query 3
CPU time required for query 3 to complete
Incoming query 2
CPU time required for query 2 to complete
Incoming query 7
CPU time required for query 7 to complete
Incoming query 1
CPU time required for query 1 to complete
Incoming query 6
CPU time required for query 6 to complete
Elapsed time
Load evaluation time interval
Load evaluation time interval

Wrap-up

Design and implement	Orchestrate	Performance	Cost
Better design and implement complex data models, including hybrid tables, aggregations, and combined storage modes (import, DirectQuery , dual).	Orchestrate the end-to-end data processing, with a pipeline chain from data ingest in the data lake house to the incremental Power BI dataset refresh.	Use different techniques to identify performance bottlenecks in your solutions and how to solve those ("does it fold"?).	Implement a cost-efficient solution, that still meets the scalability demands.

Thanks for attending!

Please fill in the evaluation

Evals.datagrillen.com