CETAS vs Views in Azure Synapse

A look under the engine cover

Bi Factory

DATA KNOWLEDGE ADVISOR

Brian Bønk Rueløkke

Principal & Enterprise arkitekt, Data & AI

Fellowmind

https://brianbonk.dk
https://linkedin.com/in/brianbonk

Microsoft

FastTrack Recognized
Solution Architect
Power BI
2022 >>

Microsoft

Certified Trainer

2018 >>

The Open Group Open Certified

TOGAF® 9
Certified

Agenda

THE SETUP

DEMO

POP THE HOOD

DEMO

THE SETUP

- NYC Taxi data from https://www.nyc.gov
- Data from Jan 2009 to Aug 2022
- Stored as parquet files in ADLSg2
- 1.685.246.650 trips / rows

THE DATALOAD

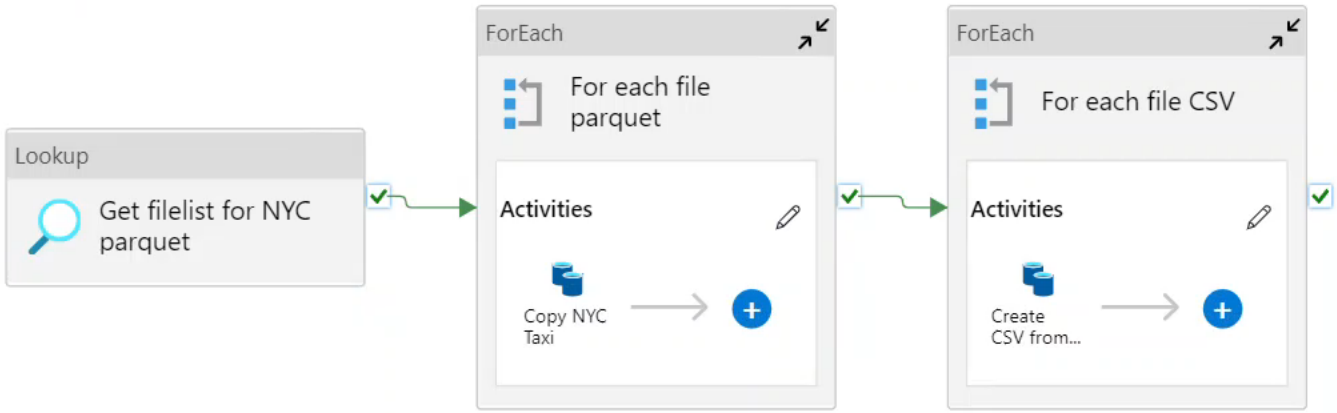
DEMO

Activities

Search activities

- Move & transform
- Azure Data Explorer
- Azure Function
- Batch Service
- Databricks
- Data Lake Analytics
- General
- HDInsight
- Iteration & conditionals
- Machine Learning
- Power Query

Validate | Debug | Add trigger



Parameters | Variables | Settings | Output

+ New

THE SERVERLESS SQL POOL

DEMO

Synapse live | Validate all | Publish all 1

Data

Workspace | Linked

Filter resources by name

- Azure Data Lake Storage Gen2 3
  - syn-howtobeadataengineer-serve (...
  - (Attached Containers)
  - BrianBoenkDataLake (brianboenk)
    - nyctaxi
    - share
    - synapsefs

setup | nyctaxi

Run | Undo | Publish | Query plan | Connect to: Built-in | Use database: Chester

```
1  -- create database
2  create database Chester
3
4  --use new database
5  use Chester;
6
7  --create schema
8  create schema nyctaxi
9
10 -- create credentials for containers in storage account
11 CREATE DATABASE SCOPED CREDENTIAL NYCData
12 WITH IDENTITY='SHARED ACCESS SIGNATURE',
13 SECRET = 'sv=2021-06-08&ss=bfqt&srt=sco&sp=rwdlacupyx&se=2099-11-21T19:24:43Z&st=2021-11-21T11:24:43Z&spr=https&sig=qv3FPPcUOf3tPbs9LuMWXos2sA%2F3phLjIJe9oilqzVU%3D'
14 GO
15
16 -- create the external data source
17 CREATE EXTERNAL DATA SOURCE ChesterDataSource WITH (
18     LOCATION = 'https://brianboenk.dfs.core.windows.net',
19     CREDENTIAL = NYCData
20 );
21
22 -- create the external file format
23 CREATE EXTERNAL FILE FORMAT SynapseParquetFormat
24 WITH (
25         FORMAT_TYPE = PARQUET
26       );
27
28 -- create nyctaxi.allparquet view
29 create or alter view nyctaxi.allparquet as
30 SELECT
31     *
32 FROM
33     OPENROWSET(
34         BULK 'https://dslhowtobeadataengineer.dfs.core.windows.net/nyctaxi/parquet/yellow_tripdata_*.parquet'
```

Results | Messages

No results to show

Your query yielded no displayable results

00:00:00 Query executed successfully.

THE COPARISON

DEMO

POP THE HOOD

DEMO

Documentation from Microsoft

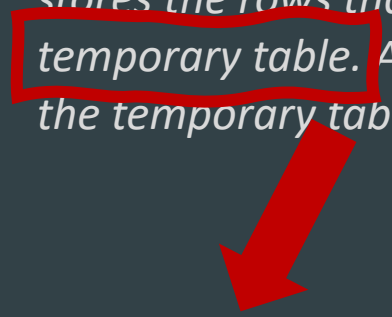
Maximum size of the result set	Up to 400 GB shared between concurrent queries.
Maximum concurrency	Not limited and depends on the query complexity and amount of data scanned. One serverless SQL pool can concurrently handle 1,000 active sessions that are executing lightweight queries. The numbers will drop if the queries are more complex or scan a larger amount of data.

Serverless SQL pool self-help - Azure Synapse Analytics | Microsoft Learn

External table’s performance

In ad-hoc query scenarios, such as SELECT FROM EXTERNAL TABLE, PolyBase stores the rows that are retrieved from the external data source in a temporary table. After the query completes, PolyBase removes and deletes the temporary table. No permanent data is stored in SQL tables.

CREATE EXTERNAL TABLE (Transact-SQL) - SQL Server | Microsoft Learn



TempDB?

TempDB?

DEMO

The TempDB

Results Messages

	file_id	file_guid	type	type_desc	data_space_id	name	physical_name	max_size_gb	growth
1	1	*NULL*	0	ROWS	1	tempdev	C:\WFRoot\DB.4\Fabric\work\Appl…	106	4096
2	2	*NULL*	1	LOG	0	templog	C:\WFRoot\DB.4\Fabric\work\Appl…	16	8192
3	3	3e22194b-26f4-4421-a833-762c…	0	ROWS	1	tempdev2	C:\WFRoot\DB.4\Fabric\work\Appl…	106	4096
4	4	a3022e02-3b2c-40f4-a392-27e9…	0	ROWS	1	tempdev3	C:\WFRoot\DB.4\Fabric\work\Appl…	106	4096
5	5	97f123e0-f69b-429a-bd8e-42e0…	0	ROWS	1	tempdev4	C:\WFRoot\DB.4\Fabric\work\Appl…	106	4096

The maximum number of concurrent PolyBase queries is 32

CREATE EXTERNAL FILE FORMAT (Transact-SQL) - SQL Server | Microsoft Learn

Considerations

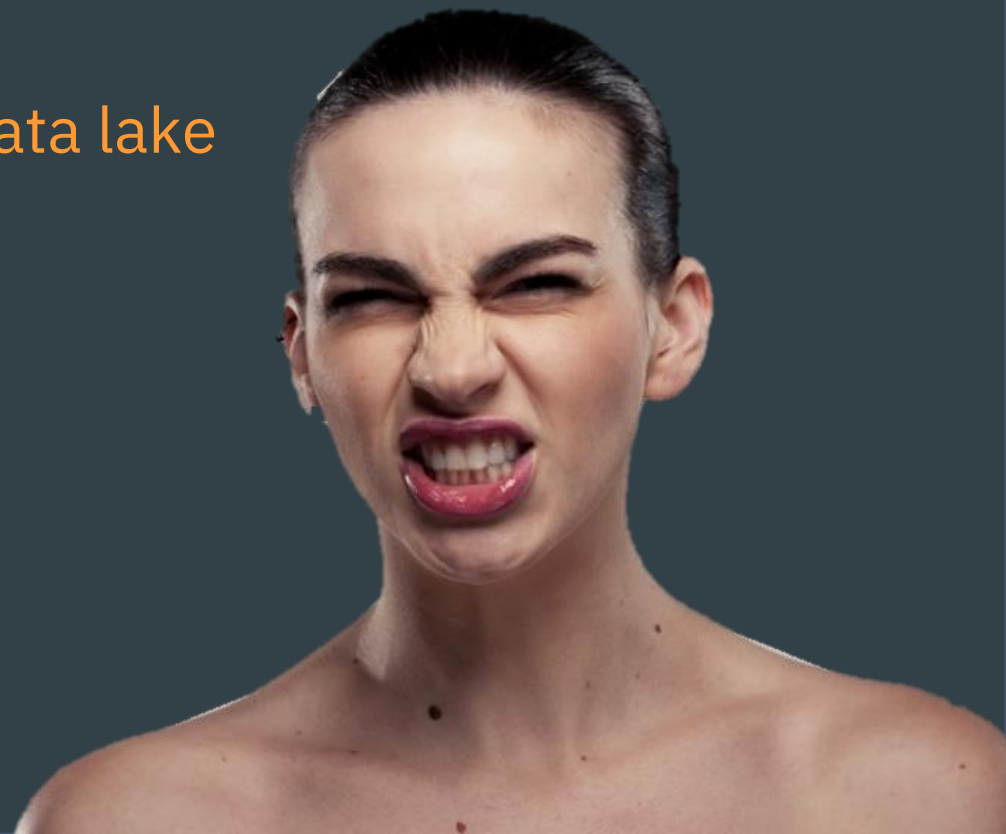
SELECT …. FROM <external_table> WHERE <field> = <value>

External tables loaded to Serverless has no index

Search predicate often is pushed down to the data lake

Joins on 100 rows pr table

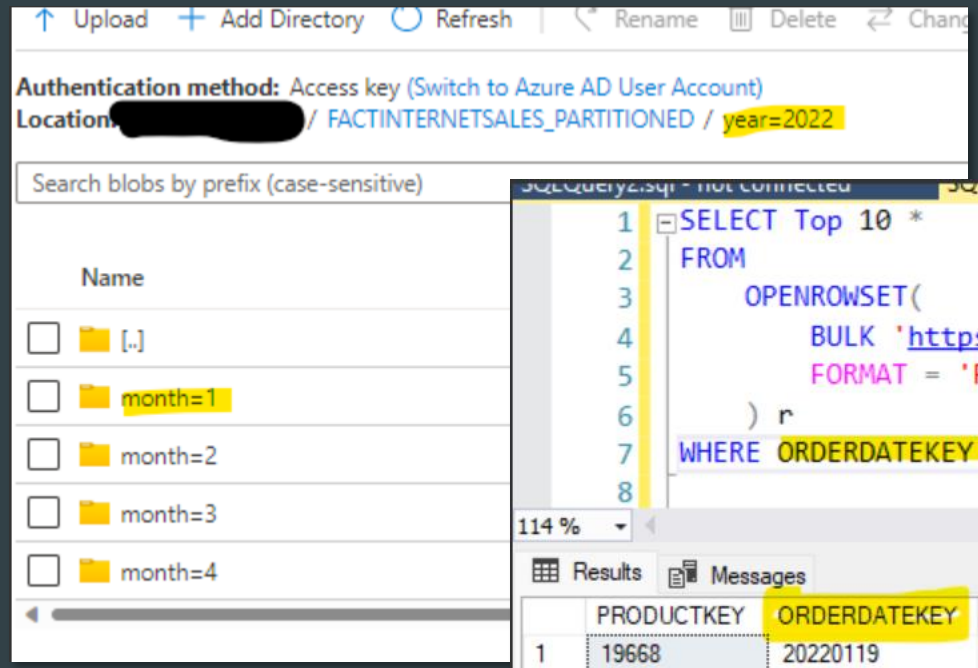
Joins on 10 mio rows pr table



This just in... (From Microsoft Nov 11, 2022)

..Queries process the entire dataset and don't benefit at all by partitions...

This just in... (From Microsoft Nov 11, 2022)




```
1  SELECT Top 10 *
2  FROM
3      OPENROWSET(
4          BULK 'https://[redacted].dfs.core.windows.net/[redacted]/FACTINTERNETSALES_PARTITIONE
5          FORMAT = 'PARQUET'
6      ) r
7  WHERE ORDERDATEKEY = 20220119
8
```

114 %

Results | Messages

	PRODUCTKEY	ORDERDATEKEY	DUEDATEKEY	SHIPDATEKEY	CUSTOMERKEY	PROMOTIONKEY	CURRENCYKEY	SALESTERRITORYKEY	SALESORDE
1	19668	20220119	20220120	20220121	40220	1649	14	5	SO4022020
2	19668	20220119	20220120	20220121	5402	1649	14	5	SO5402202
3	19668	20220119	20220120	20220121	1987	1649	14	5	SO1987202
4	19668	20220119	20220120	20220121	33491	1649	14	5	SO3349120
5	19668	20220119	20220120	20220121	45440	1649	14	5	SO4544020
6	19668	20220119	20220120	20220121	38780	1649	14	5	SO3878020
7	19668	20220119	20220120	20220121	32806	1649	14	5	SO3280620
8	19668	20220119	20220120	20220121	5394	1649	14	5	SO5394202
9	19668	20220119	20220120	20220121	32811	1649	14	5	SO3281120
10	19668	20220119	20220120	20220121	5392	1649	14	5	SO5392202

This just in... (From Microsoft Nov 11, 2022)

Results | Messages

```
Statement ID: {71C1401F-4375-42FC-A189-7705AB5CF2F3} | Query hash: 0x7A3A4ECC0D0E0871 | Distributed request ID: {423CA2C2-0C81-4F50-8880-86ADB5EAC854}.
Total size of data scanned is 46 megabytes, total size of data moved is 1 megabytes, total size of data written is 0 megabytes.

(10 rows affected)

Completion time: 2022-11-10T11:29:57.0647663+01:00
```

Authentication method: Access key (Switch to Azure AD User Account)
Location [redacted] / FACTINTERNETSALES_PARTITIONED / year=2022 / month=1

Search blobs by prefix (case-sensitive) | Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size
[..]					
part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25.c000.s...	11/4/2022, 2:29:31 PM	Hot (Inferred)		Block blob	8.18 MiB

Synapse Serverless SQL Pool - Performance and cost optimization with partitioning - Microsoft Community Hub

This just in... (From Microsoft Nov 11, 2022)

But – we just have to it correctly…

This just in... (From Microsoft Nov 11, 2022)

```
11  SELECT Top 10 r.filepath(1) [Year],r.filepath(2) [Month], r.filename() [FileName],*
12  FROM
13      OPENROWSET(
14          BULK 'https://[redacted].dfs.core.windows.net/[redacted]/FACTINTERNETSALES_PARTITIONED/*/*/*.parquet',
15          FORMAT = 'PARQUET'
16      ) r
17  WHERE ORDERDATEKEY = 20220119
18      and r.filepath(1)='YEAR=2022' --(1) first level -> Year -> physical name of the folder
19      AND r.filepath(2)='MONTH=1' --(2) second level -> Month -> physical name of the folder
20
```

14 %

Results | Messages

	Year	Month	FileName	PRODUCTKEY	ORDERDATEKEY	DUEDATEKEY	SHIPDATEKEY	CUSTOMERKEY	PROMOTIONKEY	CURREN
1	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	5402	1649	14
2	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	1987	1649	14
3	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	32806	1649	14
4	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	40220	1649	14
5	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	33491	1649	14
6	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	5394	1649	14
7	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	5392	1649	14
8	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	45440	1649	14
9	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	38780	1649	14
10	year=2022	month=1	part-00173-f15597b8-9dec-4a0c-a269-4f0fa421aa25....	19668	20220119	20220120	20220121	32811	1649	14

This just in... (From Microsoft Nov 11, 2022)

```
11 SELECT Top 10 r.filepath(1) [Year],r.filepath(2) [Month], r.filename() [FileName],*
12 FROM
13     OPENROWSET(
14         BULK 'https://[redacted].dfs.core.windows.net/[redacted]/FACTINTERNETSALES_PARTITIONED/*/*/*.parquet',
15         FORMAT = 'PARQUET'
16     ) r
17 WHERE ORDERDATEKEY = 20220119
18     and r.filepath(1)='YEAR=2022' --(1) first level -> Year -> physical name of the folder
19     AND r.filepath(2)='MONTH=1' --(2) second level -> Month -> physical name of the folder
20
```

14 %

Results | Messages

Year	Month	FileName	PRODUCTKEY	ORDERDATEKEY	DUEDATEKEY	SHIPDATEKEY	CURREN

14 %

Results | Messages

```
Statement ID: {7FFC48D6-8161-494D-8FEB-15BB346AE3B8} | Query hash: 0x27877913FB387648 |  uted request ID:  AF444D04B}.
Total size of data scanned is 9 megabytes, total size of data moved is 1 megabytes, total  of data written
(10 rows affected)

Completion time: 2022-11-10T11:37:35.5970053+01:00
```

Synapse Serverless SQL Pool - Performance and cost optimization with partitioning - Microsoft Community Hub

This just in... (From Microsoft Nov 11, 2022)

But wait

All the time in the world

Maximum query duration	30 minutes.
Maximum size of the result set	Up to 400 GB shared between concurrent queries.
Maximum concurrency	Not limited and depends on the query complexity and amount of data scanned. One serverless SQL pool can concurrently handle 1,000 active sessions that are executing lightweight queries. The numbers will drop if the queries are more complex or scan a larger amount of data.

Serverless SQL pool self-help - Azure Synapse Analytics | Microsoft Learn

Key take aways

Work with data to fit the known TempDB configuration and file sizes

Know that you only have 100 GB for #-tables, CTE's and table variables

Refit the data pipelines to partition your files to fit the engine

Can you do incremental loading, then do it – also in Power BI

Move your joins to after Synapse serverless – Power BI does a very good job

Brian Bønk Rueløkke

Principal & Enterprise arkitekt, Data & AI

Fellowmind

https://brianbonk.dk
https://linkedin.com/in/brianbonk

Microsoft

FastTrack Recognized
Solution Architect
Power BI
2022 >>

Microsoft

Certified Trainer

2018 >>

The Open Group Open Certified

TOGAF® 9
Certified