Intro to Serverless SQL in Azure Synapse

Agenda

- Overview of the Serverless SQL pool
- Create the Synapse Workspace in Azure
- Walkthrough of the Synapse Workspace
- Analyse data using the Serverless SQL Pool
  - CSV, Parquet, JSON
  - Tips and tricks
- Next steps to try yourself

Brian Bønk Rueløkke

Principal & Enterprise arkitekt, Data & Analytics

Fellowmind

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

Microsoft MVP
Most Valuable Professional

Microsoft
FastTrack Recognized
Solution Architect
Power BI
2022 >>

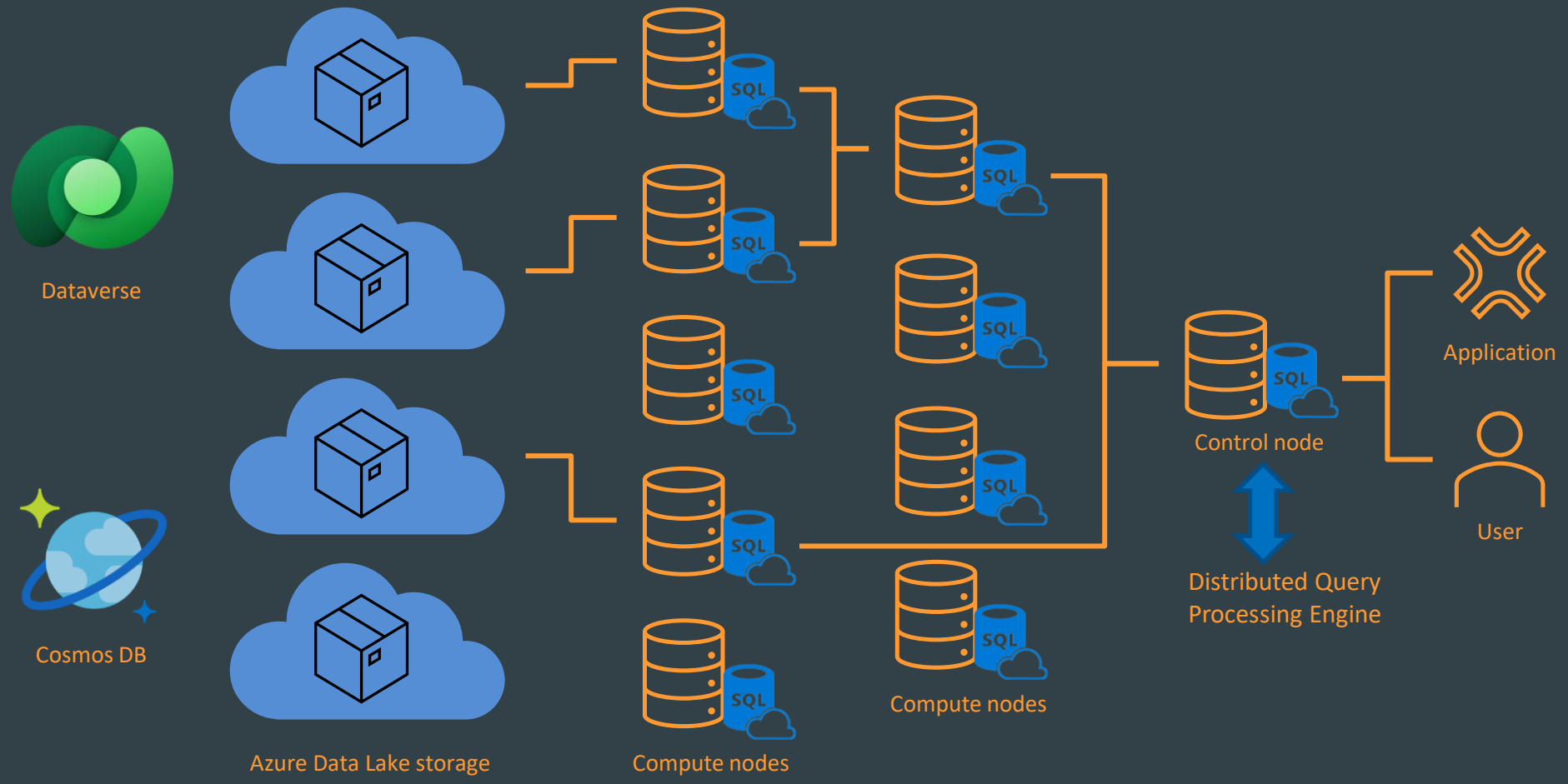
Microsoft
Certified Trainer
Data Platform
2018 >>

The Serverless SQL Pool

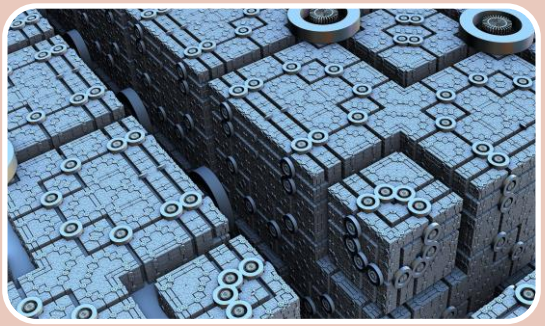
- A part of every Synapse Analytics Workspace
- Enables you to query data from Azure Data Lake, Cosmos DB and Dataverse
- Using familiar T-SQL syntax to query data in place without the need to move or copy the data

- What about the cost?
- Pay-per-use setup – approx. 5$ pr TB processed data

The Serverless SQL Pool - architecture



Benefits



Basic discovery and exploration

Logical Data Warehouse

Simple data transformations

Benefits

Data Engineers can explore the lake, transform and prepare data using this service, and simplify their data transformation pipelines.

Data Scientists can quickly reason about the contents and structure of the data in the lake, thanks to features such as OPENROWSET and automatic schema inference.

Data Analysts can explore data and Spark external tables created by Data Scientists or Data Engineers using familiar T-SQL language or their favourite tools, which can connect to serverless SQL pool.

BI Professionals can quickly create Power BI reports on top of data in the lake and Spark tables.

T-SQL support

- DDL: Databases, Schemas, Views and Stored Procedures -> No Tables, Materialized Views and Functions!
- Security: Logins and users, Credentials, Grant/Deny on object level, AAD integration
- SELECT: (almost) full SELECT support – a minor subset of features are not (yet) available

DEMO

Create Synapse Workspace

The initial setup

Azure Synapse Analytics

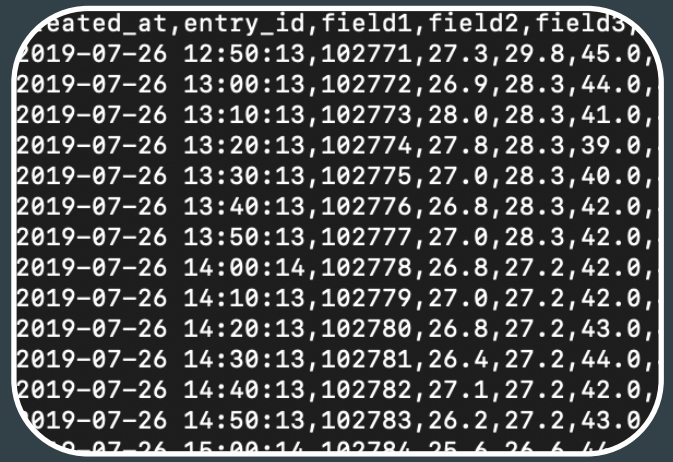
Azure Blob Storage

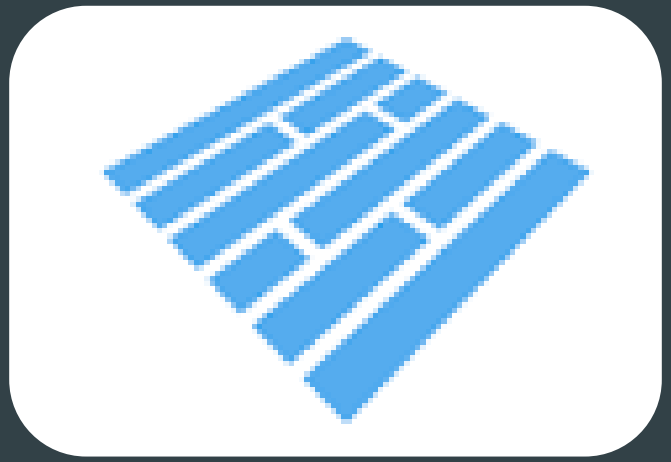
Serverless SQL endpoint

DEMO

A quick walkthrough of the Synapse workspace

Query data using Synapse SQL Serverless





CSV Parquet JSON

DEMO

Query CSV, JSON and Parquet data

Source: How to use OPENROWSET in serverless SQL pool - Azure Synapse Analytics | Microsoft Docs

CSV PARSER VERSIONS

CSV parser version 1.0 specifics:

- Following options aren't supported: HEADER_ROW.
- Default terminators are \r\n, \n and \r.
- If you specify \n (newline) as the row terminator, it will be automatically prefixed with a \r (carriage return) character, which results in a row terminator of \r\n.

CSV parser version 2.0 specifics:

- Not all data types are supported.
- Maximum character column length is 8000.
- Maximum row size limit is 8 MB.
- Following options aren't supported: DATA_COMPRESSION.
- Quoted empty string ("") is interpreted as empty string.
- DATEFORMAT SET option is not honored.
- Supported format for DATE data type: YYYY-MM-DD
- Supported format for TIME data type: HH:MM:SS[.fractional seconds]
- Supported format for DATETIME2 data type: YYYY-MM-DD HH:MM:SS[.fractional seconds]
- Default terminators are \r\n and \n.

CETAS

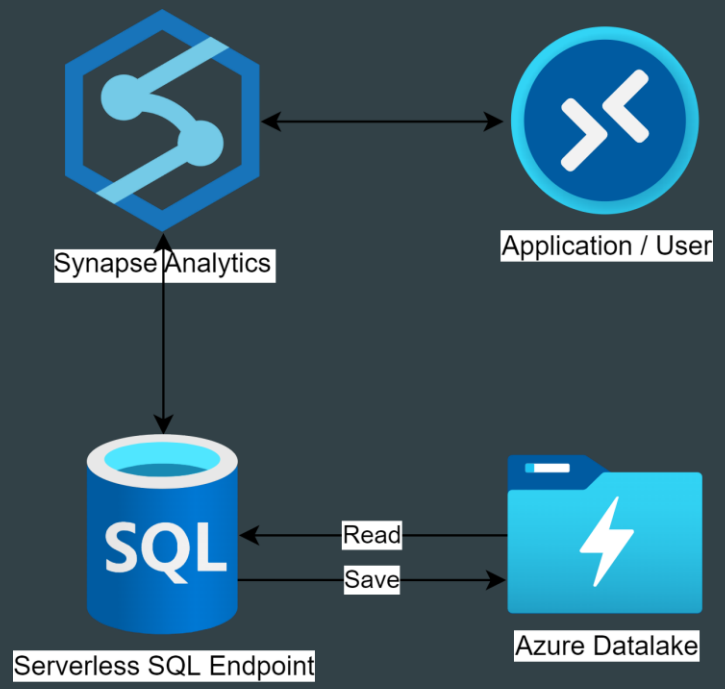
Synapse SQL also provides the ability to save a coded view from the DataLake to a physical file.

This process is called a CETAS – **C**reate **E**xternal **T**able **AS**.

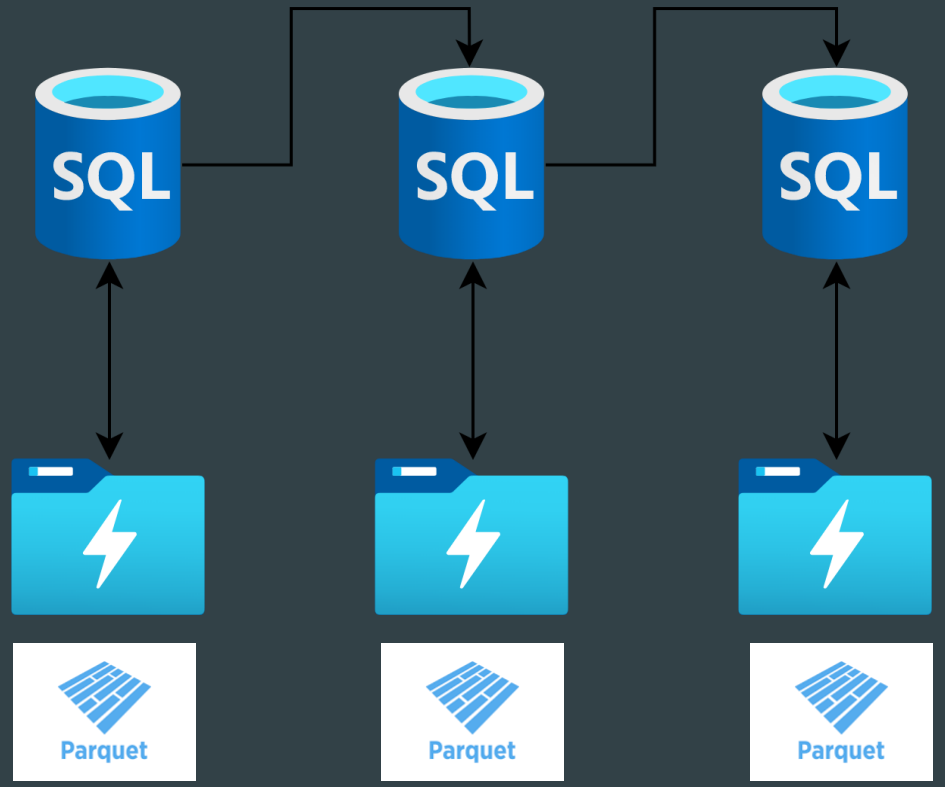
It saves a Parquet file in the DataLake in a specified folder.

The usage of this could be to provide precalculated dataset to an application or end user with very little wait time. The precalculation happens in the creation of the CETAS.

CETAS



Virtual Data Warehouse



Tips & Tricks

- If your CSV file is UTF-8 formatted, then use a UTF-8 collation when creating the database – this helps and removes and lot frustration
- Try to use PARSER VERSION 2 when reading CSV – it is much faster, but demands specific formatting in the CSV file – read more in the addendum slides.

NEXT STEP

Try it out for yourself

https://github.com/brianbonk/public/tree/master/Speaks/2023

Demo data and scripts

- All demo data and scripts will be provided.
- You need to change the “address” to the storage account, to point to your own account.
- Data is free to use and comes from the public Taxi data endpoint from NY Taxi website (Parquet and CSV) and my own blog (JSON).

Want more?

Then these two videos could be the next step into the Synapse world

Supercharge Power BI with Azure Synapse Analytics (with Mathias Halkjær) – Havens Consulting

https://www.youtube.com/live/CgSuTHKJH60

Azure Synapse Serverless SQL Pools - CETAS vs Views vs Parquet – Data Toboggan

https://www.youtube.com/watch?v=f3uQCbJO16U