Azure Synapse Analytics

August 2021 Camilla Gaardsted

SuperUsers A/S

Relationel Databases in Azure (SQL databases)

MS SQL 4 options in Azure:

- Azure SQL Database
- VM with MS SQL Server 2017+ standard/enterprise
- Managed instance
- **Azure Synapse Analytics (Data warehouse)**

Alternativer:

- Azure database for MySQL, PostgreSQL, etc
- VM with Oracle, MySQL, etc

Azure Synapse Analytics

Hed tidligere Azure SQL Data Warehouse

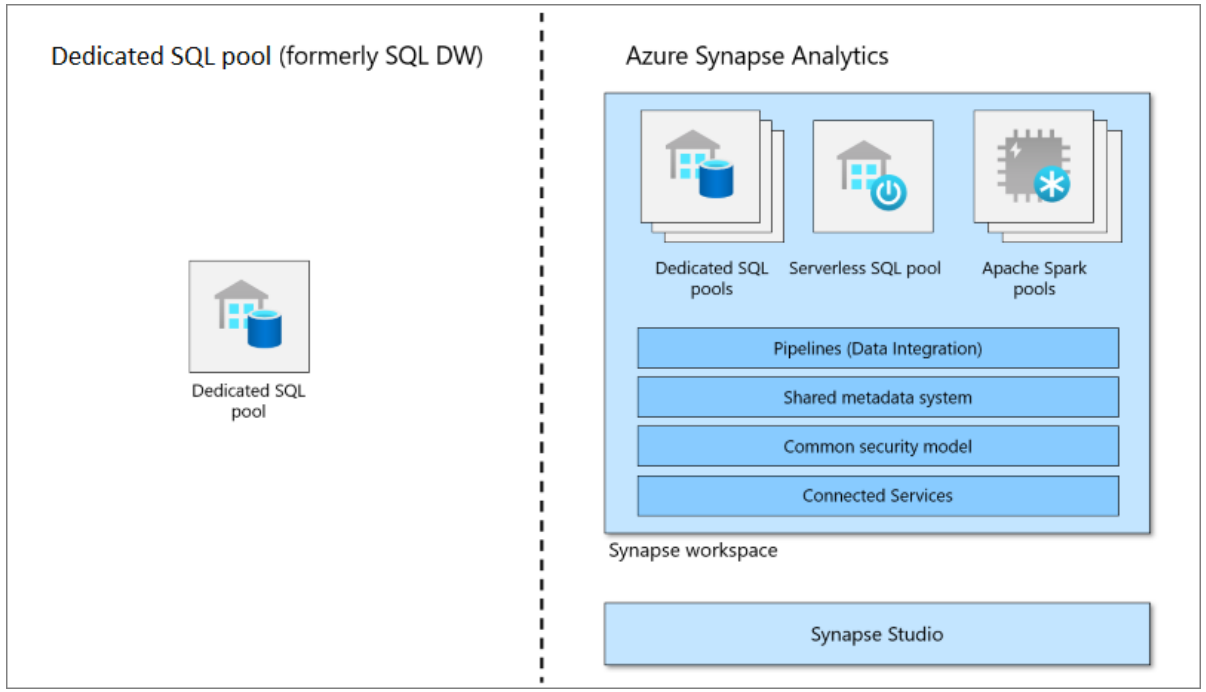
Nu en hel service som består af:

Azure Synapse SQL

Azure Synapse Pipelines

Azure Synapse Link

Apache Spark pool



Azure Synapse Workspace

Workspace name must be unique to form address:

- <workspace name>.sql.azuresynapse.net

Requires a primary storage account (Spark)

- Data Lake (Gen 2) with a container

Filesystem is the container in the storage account

Storage blob data contributor role for workspace

- Optional for current user

Firewall rules at workspace level

- Allow all (default)

Rule name	Start IP	End IP
allowAll	0.0.0.0	255.255.255.255

Azure Synapse Workspace

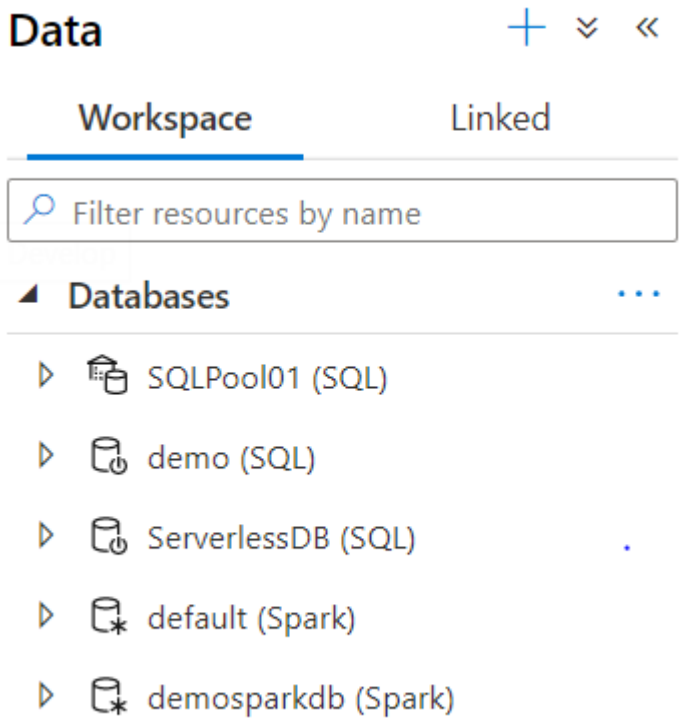
SQL Admin username

SQL Active Directory admin (auto created when AD user creates it)

Azure Synapse Analytics - Pools

- Serverless SQL pool
- Dedicated SQL pool
- Apache Spark pool

All 3 can contain databases



Serverless SQL pool

A workspace has a Built-in serverless pool

- <workspace name>-ondemand.sql.azuresynapse.net

Query data directly in the primary data lake without loading it

- OPENROWSET (script in Synapse Studio)

Metadata objects e.g views and external tables

Storage account authentication via (database) credentials unless they are public

Pay per use model for queries you run

Tables in Spark databases are automatically visible, and they can be queried by serverless SQL pool.

Serverless pool - security

Permissions to read external data from datalake

- Permission to read from datalake (Azure)
- Permission to OPENROWSET (SQL)

Azure permissions to a storage account via

- AD user
- SAS
- Managed Identity for the Synapse workspace
- Anonymous access

Permission is given by a SQL (database) credential

OPENROWSET function

OPENROWSET reads content of file(s) in a remote data source and returns the content as a set of rows

The datasource is an Azure Storage account

Is referenced in the FROM clause as a table

Storage path supports wildcards

Serverless pool

OPENROWSET to read from

- csv
- Parquet
- Json
- delta lake
- Cosmosdb container analytical store

Mapping between parquet datatypes and sql datatypes

Serverless Pool - OPENROWSET

```
OPENROWSET BULK 'unstructured_data_path’
, [DATA_SOURCE = <data source name>]
, FORMAT= 'PARQUET' | 'DELTA’ | 'CSV’
```

Read data from blob storage/datalake file or folder with wildcards

Data source definition may contain a database scoped credential

Parquet data storage format

- Column oriented data storage format
- Efficient data compression

Column names and data types are automatically read from Parquet files in Synapse.

Type mappings

Synapse – filename function

```
SELECT
    r.filepath() AS filepath
    ,r.filepath(1) AS [year]
    ,r.filepath(2) AS [month]
    ,COUNT_BIG(*) AS [rows]
FROM OPENROWSET(
        BULK 'csv/taxi/yellow_tripdata_*-*.csv',
        DATA_SOURCE = 'SqlOnDemandDemo',
        FORMAT = 'CSV',
        PARSER_VERSION = '2.0',
        FIRSTROW = 2
    )
WITH (
    vendor_id INT
) AS [r]
WHERE
    r.filepath(1) IN ('2017')
    AND r.filepath(2) IN ('10', '11', '12')
GROUP BY
    r.filepath()
    ,r.filepath(1)
    ,r.filepath(2)
ORDER BY
    filepath;
```

Serverless pool – filepath function

Filepath function

- When called without a parameter, it returns the full file path that the row originates from. When DATA_SOURCE is used in OPENROWSET, it returns path relative to DATA_SOURCE.
- When called with a parameter, it returns part of the path that matches the wildcard on the position specified in the parameter. For example, parameter value 1 would return part of the path that matches the first wildcard

Azure Synapse Pool

Pool name

Performance level: Minimum DW100c

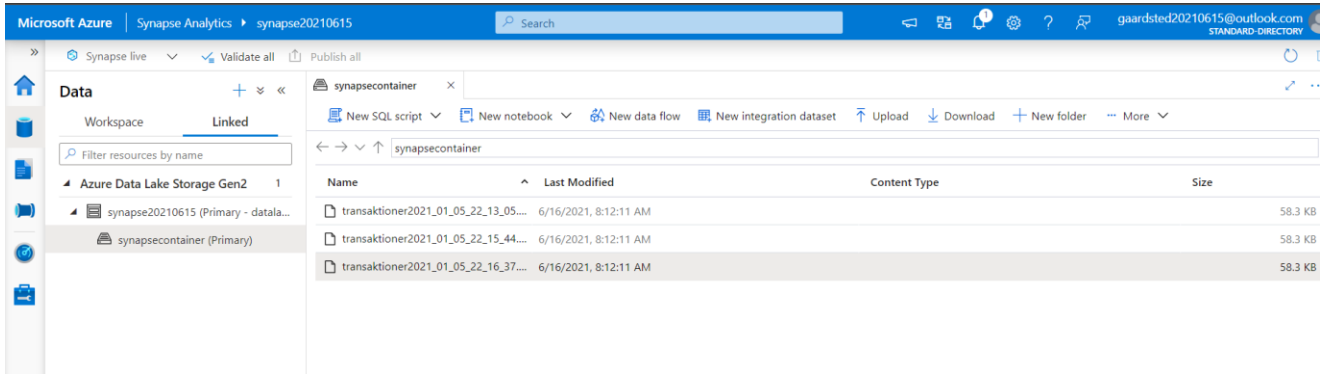
Scale up/down when needed

Pause to save money (**Very important!!!**)

Azure Synapse Studio

Via url from portal or https://web.azuresynapse.net

Generate OPENROWSET sql for datalake files



Azure Synapse - Spark pool

Node size
- Small (4 vCores/32 BG)
- Medium (8 vCores/64 GB)
- Large (16 vCores / 128 GB)
- Xlarge (32 vCores / 256 GB)
- XXLarge (64 vCores / 512 GB)

Number of nodes (3 to 200)
- Fixed number
- Autoscale (between min max)

Auto-pause after #Idle minutes

Apache Spark version

Environment packages

Spark configuration

Spark pool

Kør scripts i en Notebook

Mix Python, Scala, SQL, C#

Læs data fra Data Lake/dedicated pool

Behandl data

Gem data i Spark database

Skriv output til dedicated pool

Azure Synapse Analytics Load af data

Storage account anbefales

Once or via ADF each night/month etc

BCP / SQL Bulk copy

COPY INTO or

Polybase fastest and scalable way to load data

- Extract the source data into text files.
- Load the data into Azure Blob storage, Hadoop, or Azure Data Lake Store.
- Import the data into SQL Data Warehouse staging tables using PolyBase.
- Transform the data (optional).
- Insert the data into production tables.

Azure Synapse Analytics - Polybase

ELT – Extract Load Transform

Formats: CSV, ORC, Parquet, Gzip, Snappy

Polybase with T-SQL

External table has a table schema

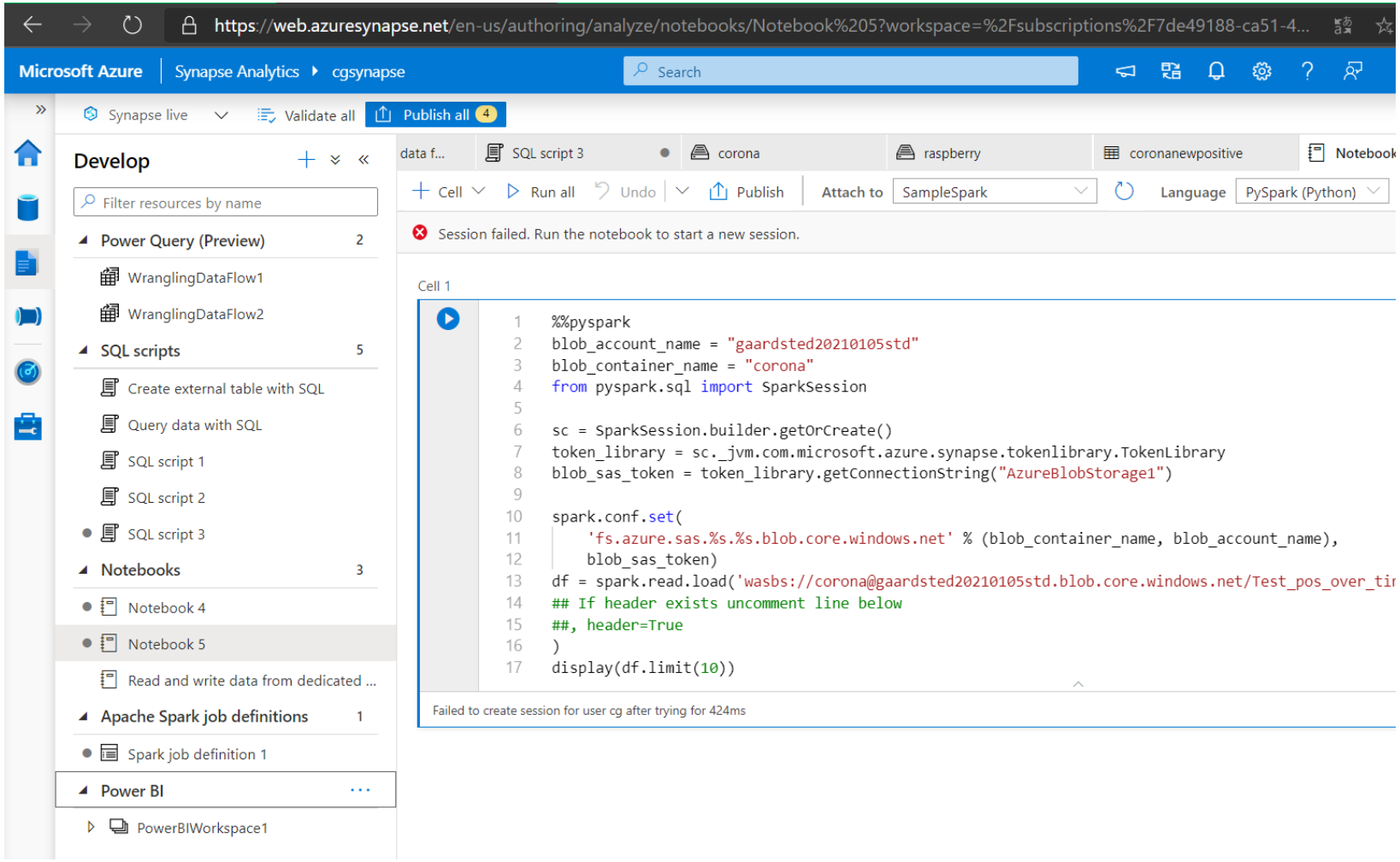
- Like a view it points to data
- Data is stored outside SQL pool

CETAS – Export a resultset

CREATE EXTERNAL TABLE AS SELECT (CETAS)

For dedicated SQL pool or serverless SQL pool

Azure Synapse Analytics - Develop



Azure Synapse Analytics - Backup

- A *data warehouse snapshot* creates a restore point you can leverage to recover or copy your data warehouse to a previous state
- A *data warehouse restore* is a new data warehouse that is created from a restore point of an existing or deleted data warehouse
- Snapshots of your data warehouse are taken throughout the day creating restore points that are available for seven days.
- Snapshots are not taken when a dedicated SQL pool is paused.
- Create manual snapshots before pausing.
- Dedicated SQL pool supports an eight-hour recovery point objective (RPO).
- You can restore your data warehouse in the primary region from any one of the snapshots taken in the past seven days.
- A geo-backup is created once per day to a paired data center (can be disabled to save cost)

Synapse – ADSL Gen2 storage account

An Azure Synapse workspace uses a default storage container for

- Storing the backing data files for Spark tables
- Execution logs for Spark jobs
- Managing libraries that you choose to install

Access to Synapse Workspace via RBAC

Access control is found under Manage in the Synapse Workspace

Azure AD users and groups (RBAC – role based access control)

Azure Synapse Analytics - Spark - dataframe

Dedicated pool - Tables

Table distribution

- Round-robin (default)
- Hash (distribution column)
- Replicated

Statistics

- Created automatically
- Not updated automatically (!)

Azure Synapse Analytics - Tables

Table features not supported in dedicated SQL pool:

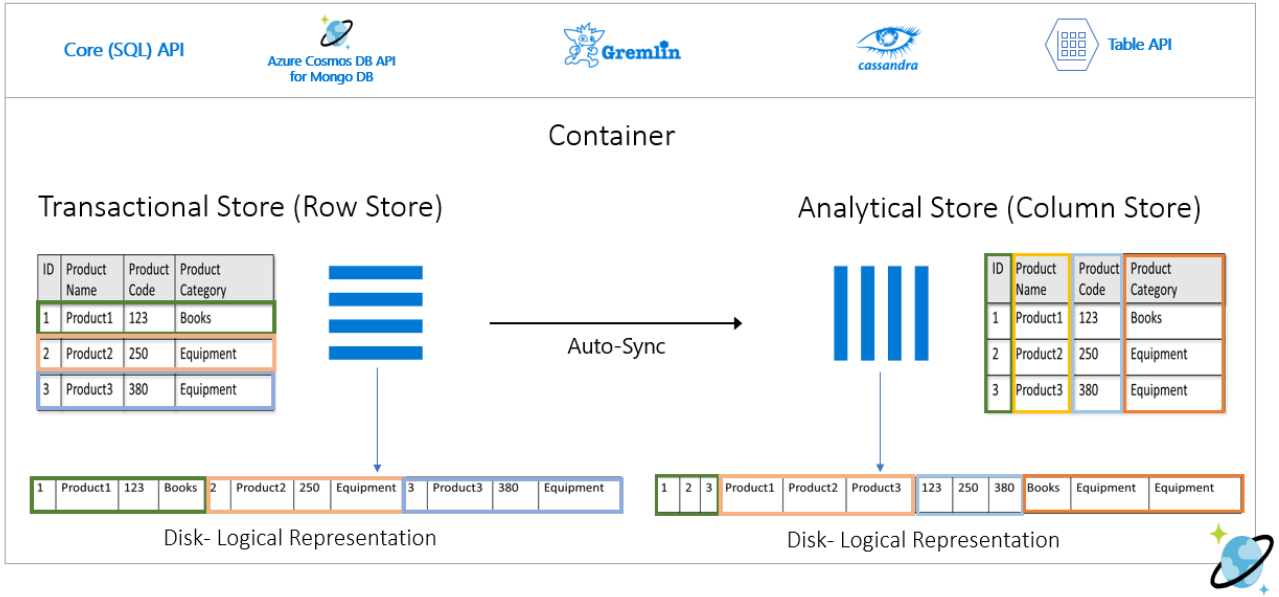
- Foreign key, Check Table Constraints
- Computed Columns
- Indexed Views
- Sequence
- Sparse Columns
- Surrogate Keys. Implement with Identity.
- Synonyms
- Triggers
- Unique Indexes
- User-Defined Types

Azure Cosmos DB analytical store

Only supported for SQL API and MongoDB

Enable at container level for a new container

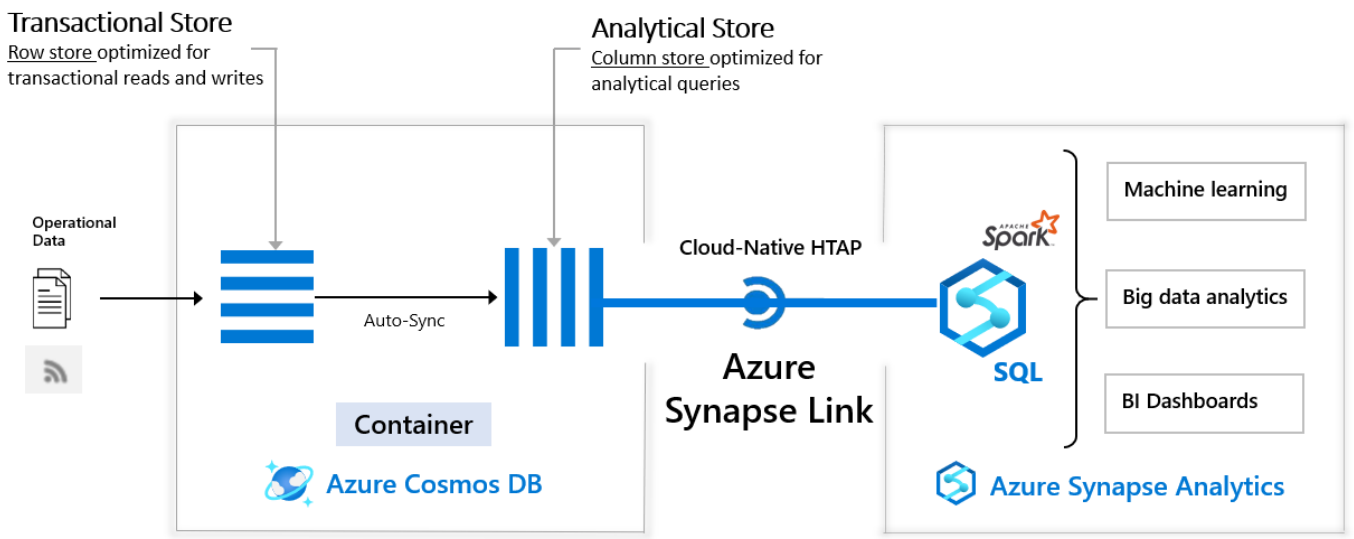
Data is synced and saved twice!



Azure Synapse Link for Cosmos DB

Run near real-time analytics over operational data in Azure Cosmos DB

- Spark pool
- Serverless pool



TTL for records in Analytical store

Synapse Spark Pool - Read data from analytical store

Synapse – Managed virtual network

Synapse – Managed private endpoints

Requires a Managed virtual network

Azure AD – conditional access

Synapse Analytics - TDE

Encrypting data at rest (datafiles, translog + backup)

Transparent Data Encryption (TDE)

Default OFF (modsat for Azure SQL Database)

Service managed key / BYOK in Azure key vault

Synapse – Spark - TokenLibrary

- Apache Spark can reference the linked services from Synapse via the TokenLibrary
- TokenLibrary can also fetch secrets from Azure Key Vault
  - Specify key vault as a connection string
  - Synapse workspace managed identity needs Get secrets permission on vault

Dedicated pool – Workload group

- System defined workload groups and roles
  - Dynamic groups
  - Static groups
- Smalrc is the default workload group for all queries
  - Performance gets worse when scaling up!!!

To solve this either:

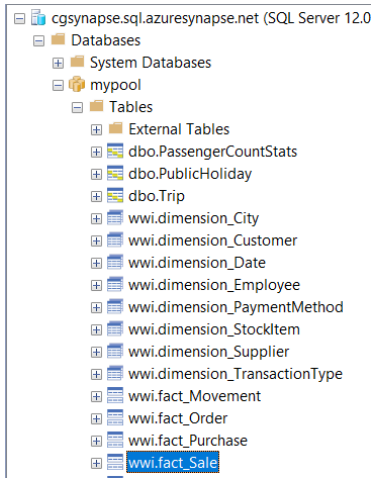
- Assign a user/role to a system defined workload group
- Create a custom workload group and assign a user/role to it

Slowly changing dimensions

- Type 1 SCD
- Type 2 SCD

Azure Synapse – Table distribution

- Round robin
- Replicated
- Hash



Synapse - External tables

Spark Pool – Databaser og tabeller

Default database findes fra start

Opret selv flere databaser

Database tabeller gemmes i filer i storage account container

Databaser kan ses og læses fra serverless pool (rettigheder?)