scieneers
DRIVEN BY DATA

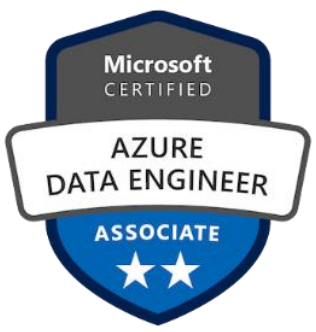
Azure Data Factory 2022

ADF & Synapse Pipelines - Intro & Unterschiede
Was gibt es Neues in ADF?

Stefan Kirner

Stefan Kirner

- PASS Chapter Lead Karlsruhe ski@sqlpass.de & Beirat
- Director Business Intelligence scieneers GmbH
- Twitter: @KirnerKa



Agenda

scieneers
DRIVEN BY DATA

Part 1 Azure Data Factory vs. Synapse Data Pipelines

- Intro
- Features Azure Data Factory only / Synapse Pipelines only
- Lieber Azure Data Factory vs. Synapse Pipelines wenn…

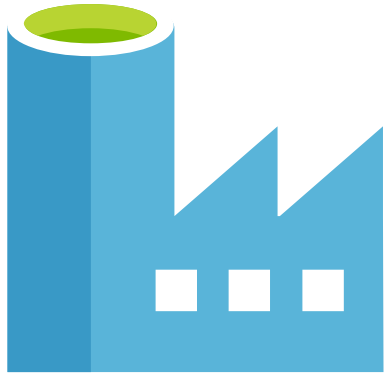
Part 2 Azure Data Factory - neue Top Features 2021/2022

- Flowlets und Data Libraries
- Activity & Data Flow updates
- Application Lifecycle Management
- SAP CDC is coming

scieneers
DRIVEN BY DATA

Azure Data Factory vs. Synapse Pipelines

Was ist was und wann benutze ich welches?



scieneers
DRIVEN BY DATA

Was ist Azure Data Factory?

Ganz einfach...

WHAT IS AZURE DATA FACTORY

READ THE DOCS

A CLOUD-BASED DATA INTEGRATION SERVICE THAT ORCHESTRATES DATA MOVEMENT & TRANSFORMATION BETWEEN DIVERSE DATA SOURCES ⟷ CLOUD COMPUTE RESOURCES AT SCALE

DATA INTEGRATION CHALLENGE

WHAT PROBLEM DOES IT SOLVE?

BIG DATA: DATABASES, FILES, DATA STREAMS, DEVICES

DIVERSE DATA SOURCES, DURATIONS

DATA INGEST → DATA TRANSFORMATION → STORED DATA, CLEANED & ENRICHED → ACTIONABLE INSIGHTS

DASHBOARDS, REPORTS

DATA OPS TEAM — DATA MOVEMENT

DATA ANALYSTS

HELP! We need to collect all this data!

WE NEED DATA WITH RELEVANT CONTEXT TO DO OUR ANALYSIS

HOW CAN WE ORCHESTRATE (DATA MOVEMENT) AND OPERATIONALIZE (WORKFLOW)

AZURE DATA FACTORY

- Raw, unorganized data stored in relational, non-relational, and other storage systems
- We need actionable business insights but raw data lacks the correlating context (by itself) for analysis
- Big data needs a scalable service than can ORCHESTRATE the data movement and OPERATIONALIZE the data processing workflows!

THE SCENARIO YOU RECOGNIZE

GAME DEV COMPANY

- HAVE: Petabytes of game log data in the cloud
- WANT: Actionable insight into my customers
  - gaming preferences
  - player demographics
  - user behaviors
- USES
  - develop new features
  - improve player experience
  - upsell, cross-sell and drive business growth

DATA ANALYST

WHAT I NEED...

- reference data from our on-prem services
  - customer info
  - marketing campaign
  - game metadata
- combined with gameplay data logs from cloud
  - game play behaviors
  - success, failure rates etc.

WHAT I MUST DO

- collect data from sources
- process data to transform it
- publish data into warehouse for analysts

AUTOMATE THIS WORKFLOW (TRIGGER IT BY AN EVENT OR MANUALLY — MONITOR & MANAGE DAILY SCHEDULES)

HOW DOES ADF WORK?

CAN YOU TELL ME WHAT ADF DOES AT A HIGH LEVEL?

LET'S BREAK IT DOWN

AZURE DATA FACTORY IS A SET OF INTERCONNECTED SYSTEMS ... THAT GIVE YOU AN END-TO-END PLATFORM ... FOR YOUR DATA ENGINEERING!

DESIGN WIZARD OR CODE + COMPUTE

CONNECT & COLLECT

1. Build an information production system to connect all data sources and adapt ingest to their diverse intervals and speeds! (CONNECTORS)
2. Collect all data in a centralized location to facilitate processing (e.g. transformation) next (COPY ACTIVITY)

TRANSFORM & ENRICH

3. Transform the collected data using actions to aggregate, filter, clean etc. Use code-free UI based mapping data transformation graph creation! (OR) use compute to transform data by hand. (MAPPING DATA FLOW)

CI/CD & PUBLISH

4. Manage data pipeline ops using Azure DevOps and GitHub! Incrementally develop and deploy your E-T-L process then publish processed data into Azure Data Warehouse, Azure SQL Database, Azure Cosmos DB – or your BI analytics engine! (PIPELINE OPS)

MONITOR & ALERT

5. Once your data pipeline is created and deployed, monitor it for the success/failure rates! (AZURE MONITOR)

Use Azure Monitor's API, PowerShell, Health panels on Azure Portal!

LET'S TALK DATA INTEGRATION PATTERNS

The most common pattern is ETL

- E**XTRACT** = connect to sources, copy data to central store
- T**RANSFORM** = process the data for analysis, data transformation at scale
- L**OAD** = move the data to data warehouse or analytics engines for business insights

Another pattern is ELT where raw (native) data is itself loaded (stored) before the transformation phase ....

AZURE DATA FACTORY HAS CODE-FREE ETL AS A SERVICE!

You always have the option to do hand-coded transformation using Azure compute ... but mapping data flows provide a UI-based wizard to simplify your pipeline setup ..

5 ASPECTS IT COVERS!

1. **INGEST DATA**
   - Copy data on-prem or in cloud!
   - 90+ native connectors
   - Serverless, auto-scaling!
2. **CONTROL FLOW**
   - Design (via UI) or Create (via SDK) your data pipelines
   - Utilize workflow constructs (loops, parameters, variables...) for efficiency
3. **DATA FLOW**
   - Code-free data transforms (execute automagically on Spark)
   - Scale out with Azure Integration Runtimes, generate data flows with SDK (code) or Designer (UI)
4. **SCHEDULE OPS**
   - Build/Maintain ops schedules for your data pipelines
   - Options include: Wall clock, Event based, Tumbling windows, chaining
5. **MONITOR OPS**
   - View active executions and pipeline history – details at activity/pipeline
   - Establish alerts for key events or progress notifications

7 THINGS TO KNOW ABOUT AZURE DATA FACTORY

1. ENTERPRISE READY — Data Integration At cloud scale!
2. ENTERPRISE DATA READY — 90+ connectors! It just works.
3. CODE-FREE TRANSFORMATION — UI driven mapping data flows
4. RUN CODE ON ANY AZURE COMPUTE — For hands-on data transformations
5. MANY SSIS PACKAGES RUN ON AZURE — (Reboot on-prem in ADF in 3 steps)
6. ADF CAN MAKE DATA OPS SEAMLESS — source control, automated deploy, simple templates
7. SECURE DATA INTEGRATION — Managed virtual networks protect against data exfiltration, simplify your networking!

6 INTEGRATION RUNTIME (IR)

INTEGRATED RUNTIME (IR) is COMPUTE INFRASTRUCTURE USED BY ADF TO PROVIDE FULLY MANAGED

- DATA FLOWS (transformation)
- DATA MOVEMENT (movement)
- ACTIVITY DISPATCH (route to service)
- SSIS PACKAGE EXECUTION (in managed compute)

3 IR TYPES

IR Types	CAPABILITY	N/W
AZURE - SSIS	SSIS package execution	(public or private)
SELF HOSTED	Activity Dispatch + Data Movement	public or private
AZURE	Activity Dispatch + Data Movement + Data Flows	public or private

- SELF-HOSTED
- AZURE-SSIS
- AZURE

PICK WHAT WORKS FOR YOUR DATA NEEDS ... AND YOUR NETWORKING REQUIREMENTS!

POTENTIAL TO HAVE HYBRID DATA INTEGRATION SOLUTIONS

5 ADDITIONAL CONCEPTS

1 CONTROL FLOW

REFERS TO THE ORCHESTRATION OF PIPELINE ACTIVITIES INCLUDING:

- CHAINING ACTIVITIES
- BRANCHING ACTIVITIES
- CUSTOM STATE PASSING
- DEFINING PARAMETERS
- PASSING ARGUMENTS (for pipeline run)
- LOOPING CONTAINERS

Think of ADF as the conductor for an orchestra of activities! (KNOWS THE BIG PICTURE)

2 PIPELINE RUN

AN INSTANCE OF PIPELINE EXECUTION

(arguments) A1 A2 A3 (parameters)

INSTANTIATE PIPELINE BY PASSING ARGUMENTS (values) TO PARAMETERS (placeholders) DEFINED IN IT.

PASS ARGUMENTS manually OR BY DEFINING triggers

3 TRIGGERS

TRIGGER EVENT — DIFFERENT TRIGGER TYPES FOR VARIOUS EVENTS!

UNIT OF PROCESSING THAT DETERMINES WHEN TO KICK OFF PIPELINE RUNS.

4 PARAMETERS

"READ-ONLY" KEY/VALUE PAIRS POPULATED FROM RUN CONTEXT AT EXECUTION

BOTH 'DATASET' AND 'LINKED SERVICE' ARE STRONGLY-TYPED, REUSABLE, REFERENCEABLE PARAMETER ENTITIES!!

ACTIVITY REFERENCES DATASET TO CONSUME DATA PROPERTIES, AND 'LINKED SERVICE' FOR THE CONNECTION INFO (to data store for copy, compute resource for flow)

5 VARIABLES

ARE USED INSIDE PIPELINES TO STORE TEMPORARY VALUES, STATE.

USED WITH PARAMETERS TO PASS VALUES BETWEEN ACTIVITIES, DATA FLOWS, PIPELINES.

2 ACTIVITIES

REPRESENT A SINGLE PROCESSING STEP IN THE PIPELINE e.g.

- COPY ACTIVITY: copy data from one store (source) to other (destination)
- HIVE ACTIVITY: run a Hive query on an HDInsight cluster

THREE TYPES OF ACTIVITY SUPPORTED BY ADF

- Data Movement
- Data Transformation
- Control

5 MAPPING DATA FLOWS

ADF EXECUTES LOGIC ON SPARK CLUSTER FOR YOU (automagically!) — NO NEED TO MANAGE OR MAINTAIN YOUR OWN CLUSTER!

- CREATE & MANAGE DATA TRANSFORMATION GRAPHS THAT WORK ON ANY SIZE DATA
- BUILD UP REUSABLE LIBRARY OF DATA TRANSFORMATION ROUTINES!
- EXECUTE TRANSFORM ROUTINES IN PIPELINES (AS ACTIVITY) TO SCALE!

HEY PRESTO!

4 LINKED SERVICES

TWO USES

- REPRESENT DATA SOURCE (data sources for INGEST)
- REPRESENT COMPUTE RESOURCE (compute resources for TRANSFORM)

- REPRESENT CONNECTION INFORMATION TO LINK ADF TO EXTERNAL SERVICES
- WORKS CLOSELY WITH DATASETS. THIS SPECIFIES HOW TO CONNECT & DATASET DEFINES WHAT DATA LOOKS LIKE...

3 DATASETS

DATA STORE — DATA SET (view)

- REPRESENTS DATA STRUCTURE GIVING SELECT VIEW INTO DATA STORE
- POINTS TO THE SPECIFIC DATA SUBSET TO USE IN ACTIVITY...
- .. FOR INPUT & OUTPUT

1 PIPELINES

A1 A2 A3 — PIPELINE STAGE

- A PIPELINE IS A LOGICAL GROUPING OF ACTIVITIES THAT PERFORM ONE UNIT OF WORK
- AN ADF INSTANCE CAN HAVE ONE OR MORE PIPELINES!

Example: Pipeline that ingests data from Azure Blob (A1), runs Hive query on HDInsight to partition it (A2) and moves result to next store (A3)

Benefit: Manage correlated activities as a single entity!

Execution: Activities may be chained (run sequentially) or be independent (run in parallel) within pipeline!

HERE ARE THE KEY TERMS AND CONCEPTS YOU NEED TO KNOW TO USE ADF!

AND A FEW KEY TERMS

- CONTROL FLOW
- PIPELINE RUN
- TRIGGERS
- PARAMETERS
- VARIABLES

ADF TOOLKIT

LET'S TALK ABOUT COMPONENTS OF AZURE DATA FACTORY

6 MAIN CONCEPTS

- PIPELINES
- ACTIVITIES
- DATASETS
- LINKED SERVICES
- DATA FLOWS
- INTEGRATION RUNTIMES

Azure Data Factory Parts

Microsoft: „Code-free ETL as a service“

Ingest

- Multi-cloud and on-premise hybrid copy data
- 100+ native connectors
- Serverless and auto-scale
- Use wizard for quick copy jobs

Control Flow

- Design code-free data pipelines
- Generate pipelines via SDK
- Utilize workflow constructs: loops, branches, conditional execution, variables, parameters, …

Data Flow

- Code-free data transformations that execute in Spark
- Scale-out with Azure Integration Runtimes
- Generate data flows via SDK
- Designers for data engineers and data analysts

Schedule

- Build and maintain operational schedules for your data pipelines
- Wall clock, event-based, tumbling windows, chained

Monitor

- View active executions and pipeline history
- Detail activity and data flow executions
- Establish alerts and notifications

https://docs.microsoft.com/en-us/azure/data-factory/introduction

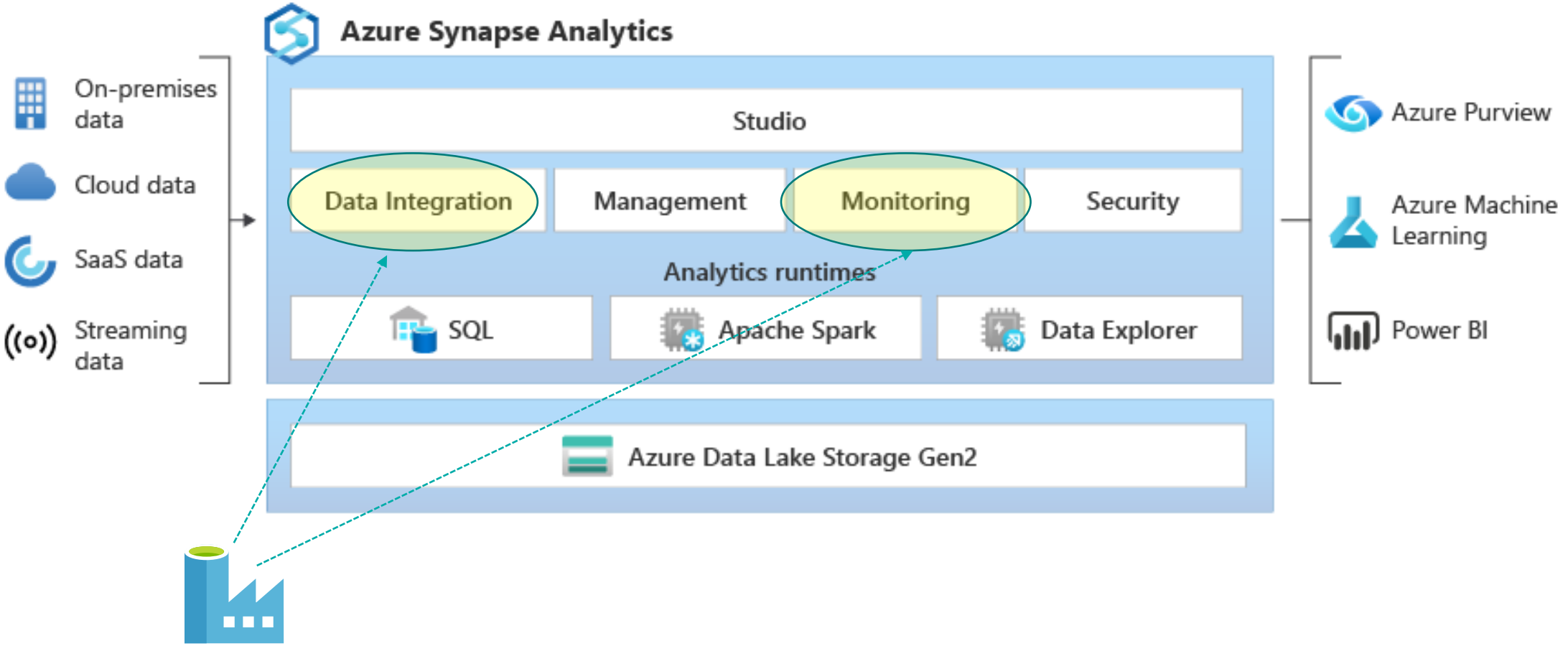
scieneers
DRIVEN BY DATA

Demo

Überflug Azure Data Factory

What is Azure Synapse Analytics

Microsoft: “Azure Synapse is an enterprise analytics service"



https://docs.microsoft.com/en-us/azure/synapse-analytics/overview-what-is

scieneers
DRIVEN BY DATA

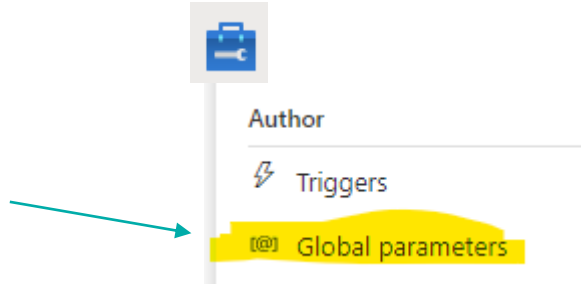
Demo

Überflug Synapse Pipelines

Azure Data Factory only stuff

Not available in Synapse

- Self-hosted Integration runtime sharing
- Support for Cross-region Integration Runtime (Data Flows)
- Power Query support aka Wrapping Data Flows
- Comfortable deployment parameters management
- *Update Snowflake source / destination*
- *Update: SSIS runtime, run SSIS packages -> in Preview for Synapse now*
- *Update: ADF Solution templates also available in Synapse*



https://docs.microsoft.com/en-us/azure/synapse-analytics/data-integration/concepts-data-factory-differences

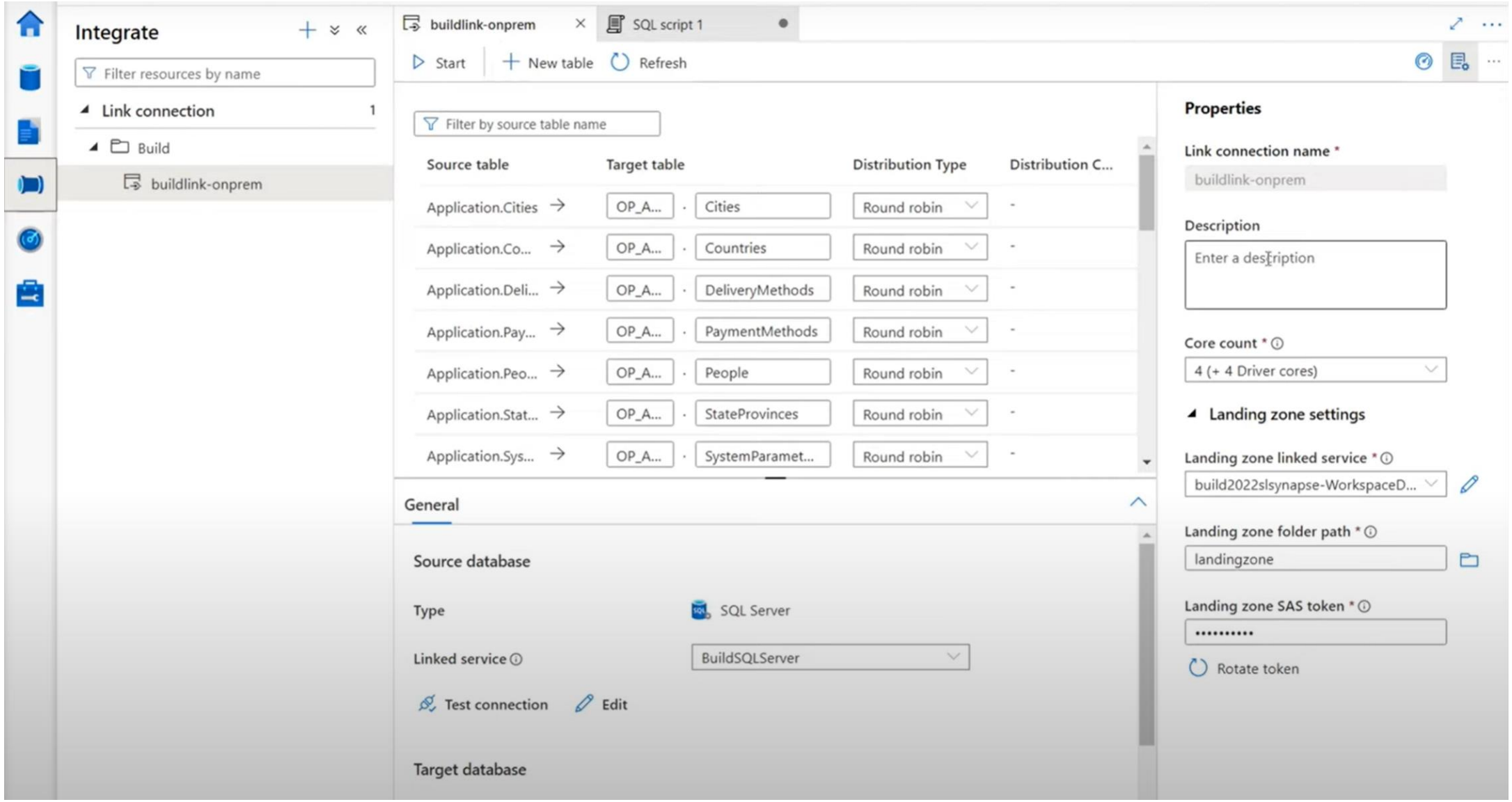
Synapse only stuff (not in ADF)

Not available in Azure Data Factory

- Spark Notebooks & Clusters
- SQL Serverless / dedicated pool (stored procedures for ETL)
- Detailed Monitoring Spark Jobs for Data Flows
- Azure Synapse Link for SQL, Dataverse, Cosmos DB

https://docs.microsoft.com/en-us/azure/synapse-analytics/data-integration/concepts-data-factory-differences

Kurzer Blick: Azure Synapse Link for SQL



Announcing Synapse Link for SQL: https://www.youtube.com/watch?v=pgusZy34-Ek

Better use Azure Data Factory for

scieneers
DRIVEN BY DATA

- Small projects (copy a bunch of tables every day)
- To keep a small footprint in infrastructure
- If having only permissions in dedicated Azure Resource Group
- As orchestration tooling for Databricks projects
- Migration from M-Code (Power BI)

Better use Synapse Data Integration if

scieneers
DRIVEN BY DATA

- You plan to do data exploration on a data lake using SQL or connect Power BI
- Data platform projects including many data sources
- Spark Notebooks used as part of data integration
- Want to use the newest features
- Want to sync masses of different tables via Azure Synapse Link .. (having money for dedicated SQL pools)

scieneers
DRIVEN BY DATA

Neues in der Azure Data Factory

- Flowlets und Data Libraries
- Activity & Data Flow updates
- Application Lifecycle Management
- SAP CDC is coming

https://docs.microsoft.com/en-us/azure/data-factory/whats-new

scieneers
DRIVEN BY DATA

Flowlets

In Mapping Data Flows

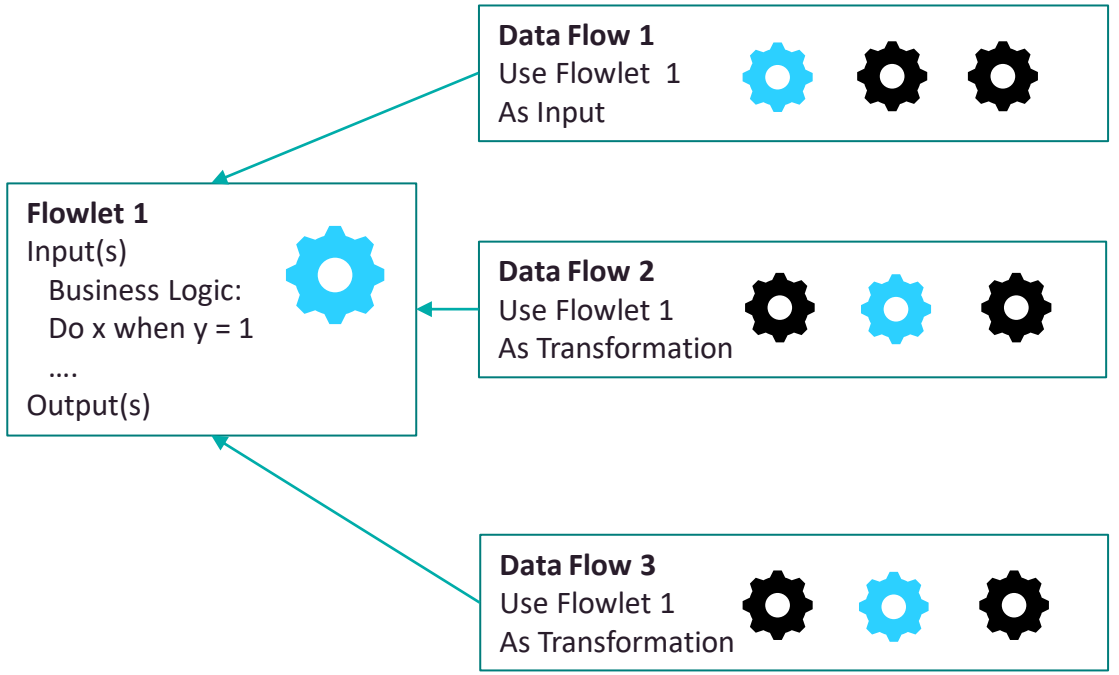
Flowlets in Mapping Data Flows

Wiederverwendbarer „Container“ von Aktivitäten

- **D**on’t **R**epeat **Y**ourself Prinzip (DRY) :

Durch Kapselung Duplikation von Code / Logik vermeiden

- Kann from scratch oder aus bestehenden Mapping Data Flows erstellt werden
- Einsatz als Quelle oder Transformation möglich
- Datentypen und Felder möglichst generisch und klein halten
- Schnittstellen-Vertrag zwischen Flowlets und den Data Flows die es verwenden



scieneers
DRIVEN BY DATA

Demo

Flowlets

scieneers
DRIVEN BY DATA

Hands-On

Flowlets

scieneers
DRIVEN BY DATA

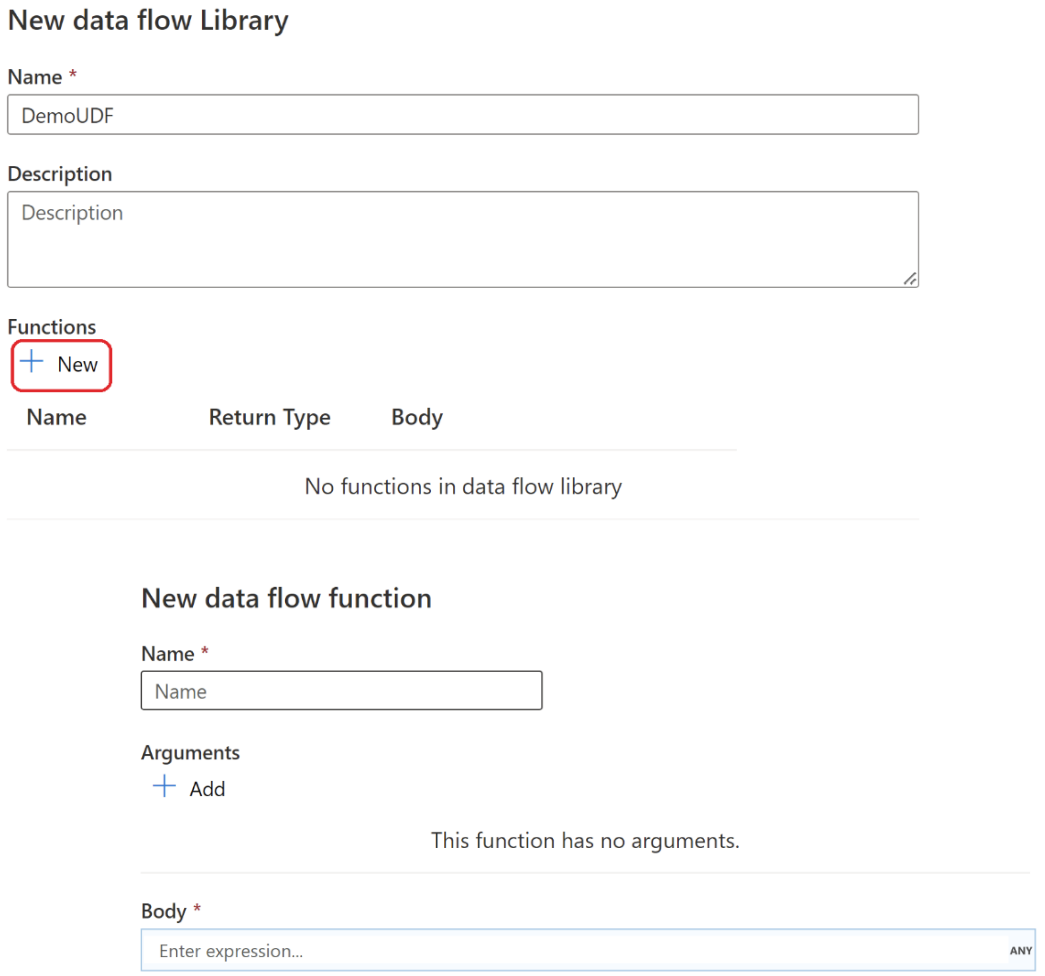
Data Flow Libraries

User Defined Functions

Data Flow Libraries – User defined functions

Noch mehr Kapselung – aber auf einer anderen Ebene

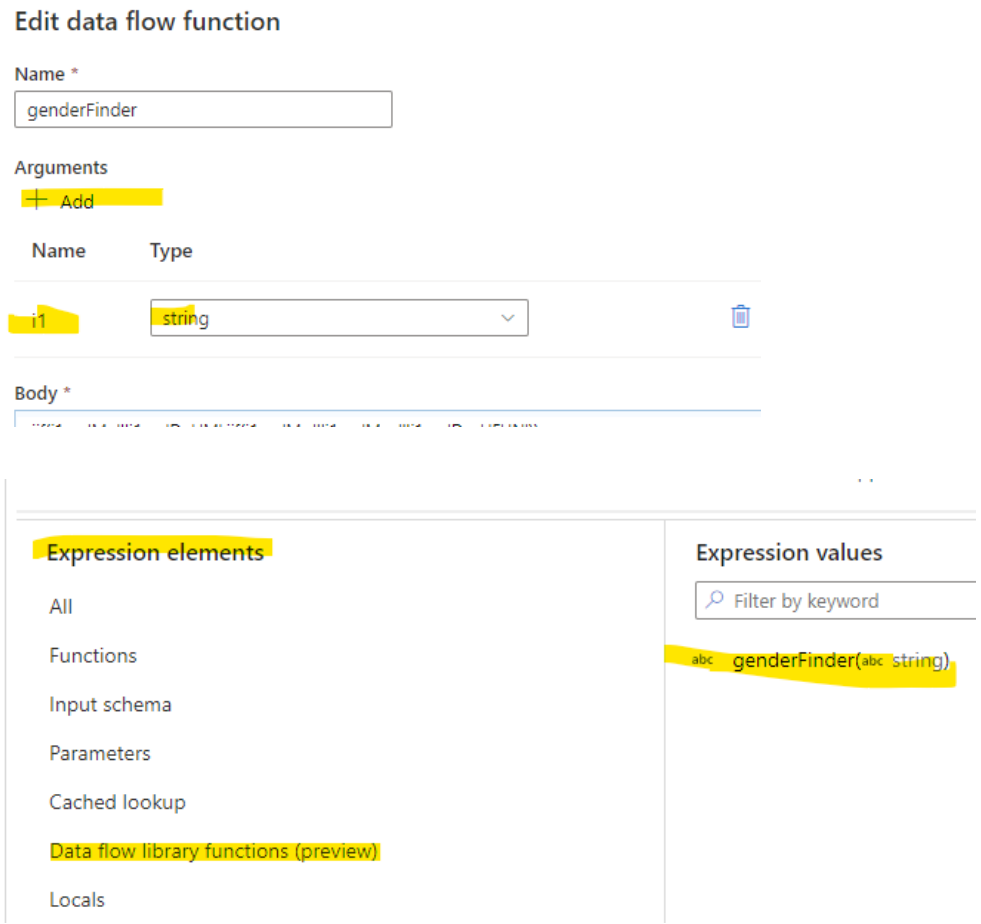
- Anlegen von Funktionen in der Expression Language die in allen Mapping Data Flows verwendet werden können
- Parametrisierung möglich
- Wiederverwendbarkeit von komplizierten Codeschnippseln
- DRY Prinzip



Data Flow Libraries – User defined functions

Definition von User Defined Functions mit Parametern

- Parameter als Arguments hinzufügen
- Namen sind vorgegeben als i1, i2 ...
- Im bekannten Expression Editor entwickeln
- Einbinden dann im Data Flow im Expression Editor z.B. bei Derived Column unter Data flow library functions



scieneers
DRIVEN BY DATA

Demo

Data Flow Libraries – User Defined Functions

scieneers
DRIVEN BY DATA

Activity & Data Flow updates

Script Activity

Besser mit Datenbanken können

- Kann einiges mehr als Lookup und Stored Procedure Activity (siehe rechts)
- Ein oder mehrere Statements
- Auch Rückgabe von Result sets sowie Abfrage-Meldungen

	Script Activity	**Lookup Activity**	**SProc Activity**
Supported data source	Database (SQL family, Snowflake, Oracle)	all data sources	SQL family
Supported Operations	Read / Modify	Read*	Modify
Multiple query support	Yes	No**	Yes***
Query parameter support	Input / Output	Not supported	Input
Output Result set support	One or more	One	No
Output query logs (PRINT)	Yes	No	No
Integrated CICD (ADF)	Yes	Yes	No

scieneers
DRIVEN BY DATA

Demo

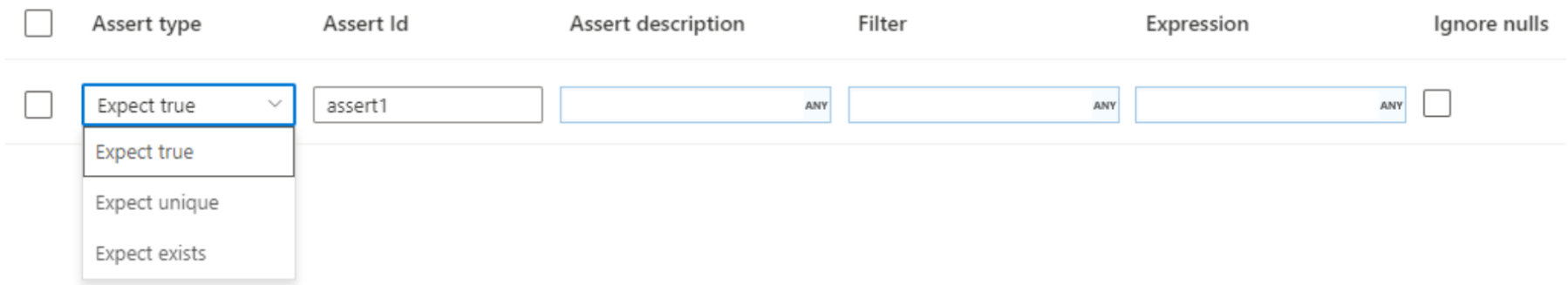
Script Task

Assert Transformation

scieneers
DRIVEN BY DATA

Datenqualität inhaltlich checken im Data Flow

- Erstellen von Regeln die auf die Spalten im Flow angewendet werden
- Checken auf
  - Wertebereiche (Expect true)
  - Eindeutigkeit (Expect unique)
  - Vergleich zwischen Streams (Expect exists)
- Markieren der Zeilen oder Abbruch möglich



scieneers
DRIVEN BY DATA

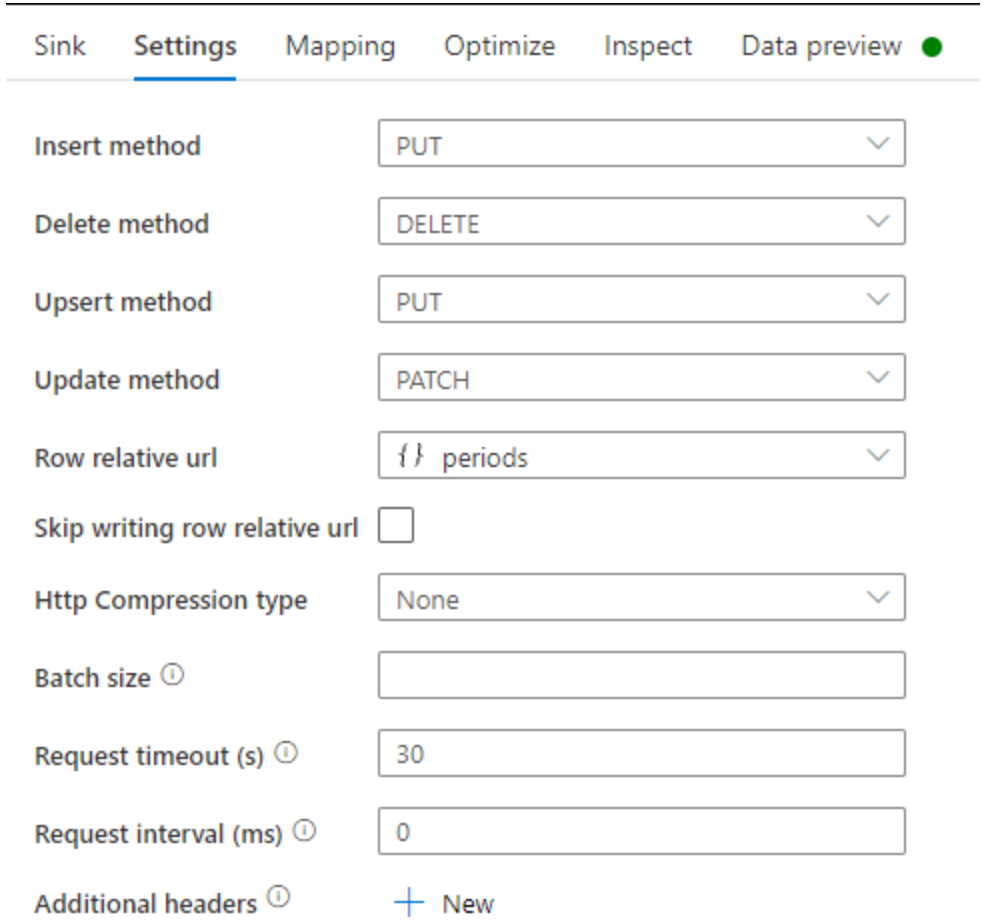
Demo

Assert Transformation

Rest Connector für Mapping Data Flows

Als Source und Sink einsetzbar

- requestMethod HTTP, Get & Post erlaubt
- integration datasets und inline datasets
- Relativ breiter Pagination Support
- Bei Verwendung als Sink ein Alter Row davorschalten um die Action zu bestimmen

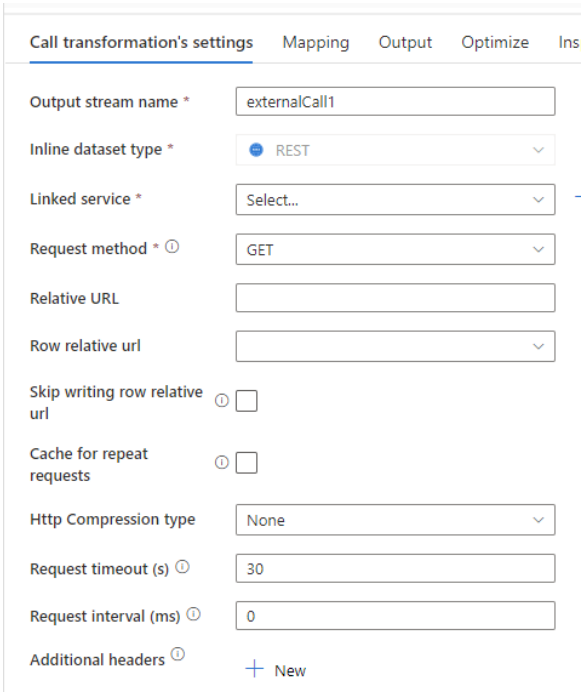


https://docs.microsoft.com/en-us/azure/data-factory/connector-rest?tabs=data-factory#mapping-data-flow-properties

External Call Transformation

Auch innerhalb eines Data Flows direkt Rest Services nutzen

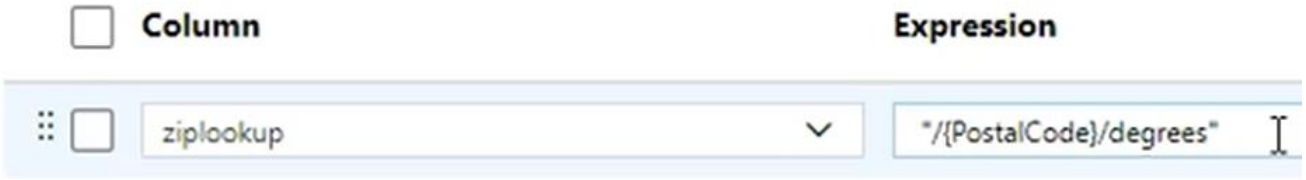
- Einbinden externer Services im Zeilenmodus, z.B. Wetterdaten pro GPS Koordinate der Zeilen im Data Flow
- Wird noch erweitert auf Stored Procedures



External Call Transformation

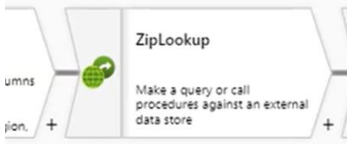
Am Beispiel Adressvalidierung

PostalCode abc
98011
98011
75201
85004
H1Y 2H5
98004



External Call Transformation

Konfiguration



Call transformation's settings | Mapping | Output | Optimize | In

Output stream name * ZipLookup

Inline dataset type * REST

Linked service * RestService6

Request method * GET

Relative URL

Row relative url abc ziplookup

Skip writing row relative url

Include header in output

Http Compression type None

Request timeout (s) 30

Call transformation's settings | Mapping | Output | Optimize | Inspect | Data preview

Options

Skip duplicate input columns

Skip duplicate output columns

Input columns * Auto mapping | Reset | Add mapping | Delete

createURL's column | Name as

abc ziplookup → ziplookup

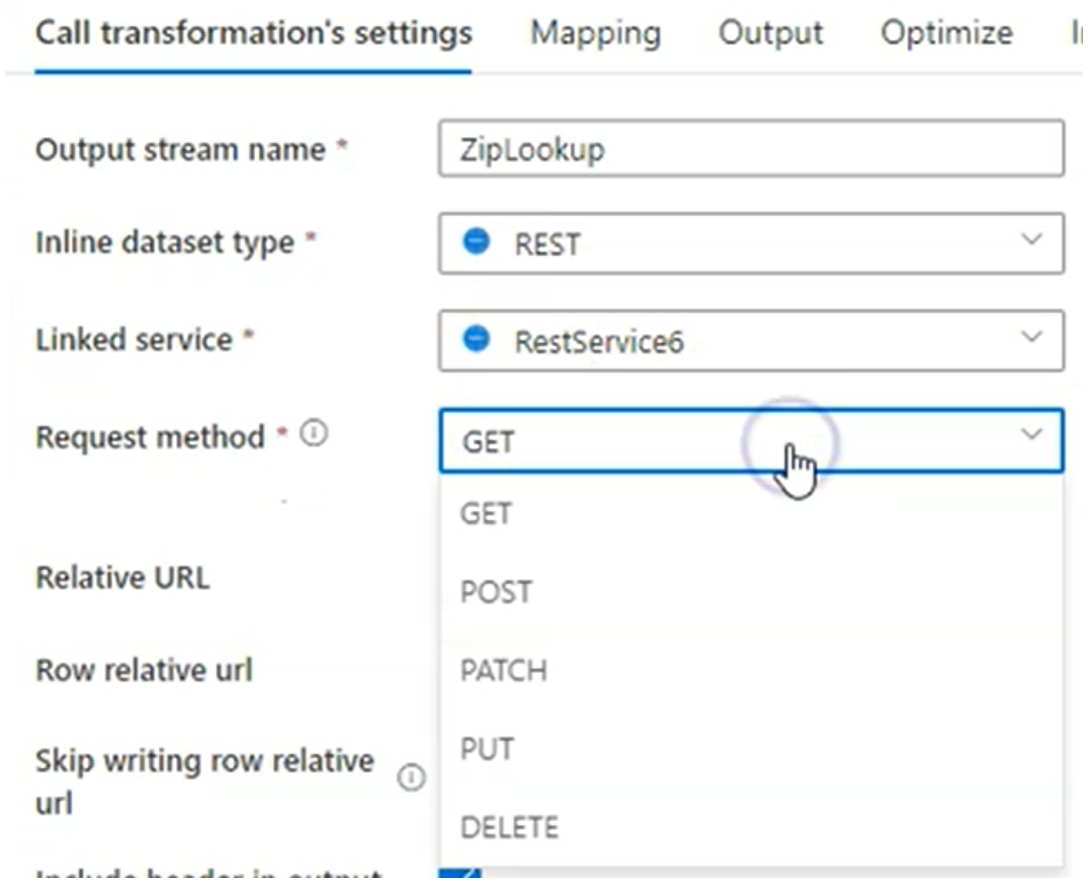
Call transformation's settings | Mapping | Output | Optimize | Inspect | Da

Import projection | Schema options

Response body definition (acceptable_city_names as string[], area_codes as string... { }

External Call Transformation

Unterstützte Rest Request Methoden



Achtung wird auch beim Preview ausgeführt (z.B. Delete)

Power Query

Aka Wrangling Data Flows

- Umbenannt, jetzt als „Power Query“ in ADF
- Wird Stück für Stück vervollständigt, Umstellung von Power BI Power Query bedeutet ja Spark hier im Backend
- Fühlt sich für mich trotzdem immer wie ein Fremdkörper an weil auch das Handling von Datasets anders ist usw.

scieneers
DRIVEN BY DATA

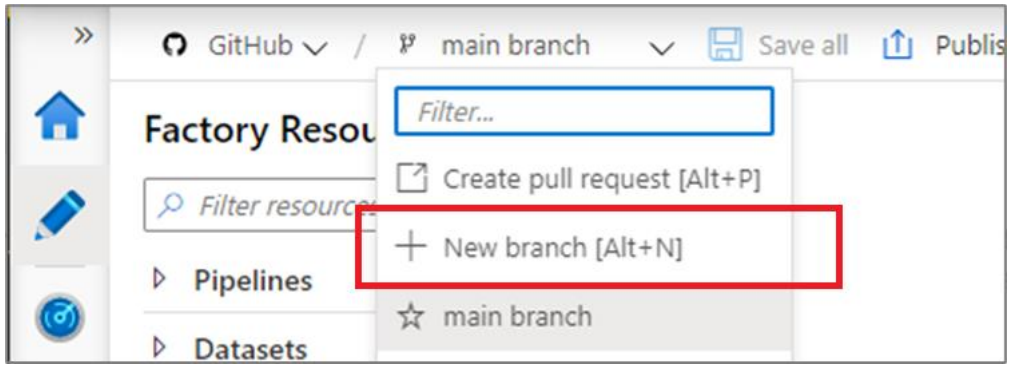
Application Lifecycle Management News

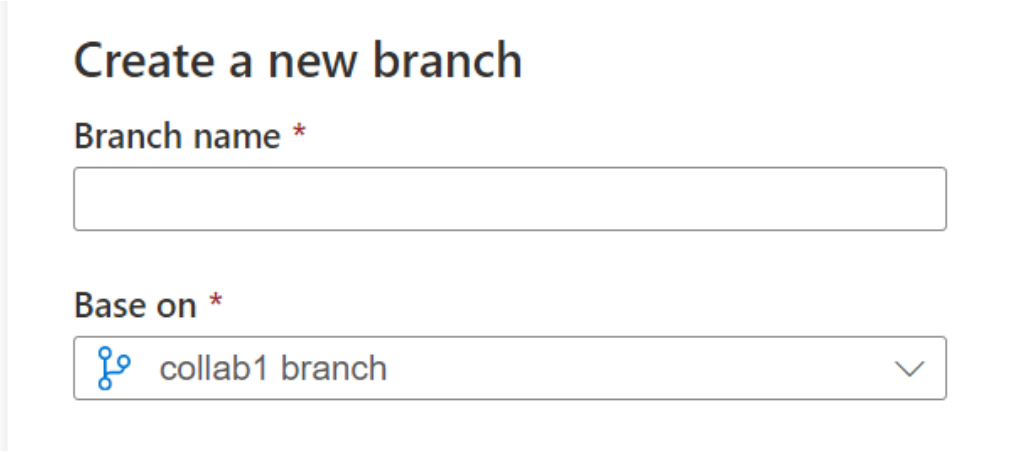
Build to survive

Azure Data Factory Github Integration

Limits entfallen & Branch of Branch

- Jetzt auch mehr als 1000 Data Factory Resourcen pro resource type möglich (gemeint sind datasets, pipelines & co.)
- Hilft bei der Arbeit mit Code Versioning & Deployments auch in Zusammenhang mit der GitHub API
- Es kann jetzt auch ein Branch basierend auf einem anderen Branch erstellt werden



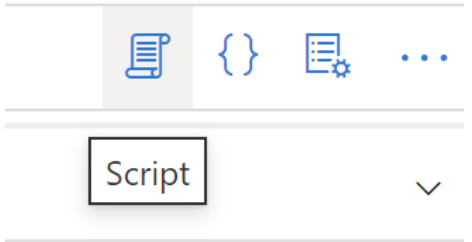


https://docs.microsoft.com/en-us/azure/data-factory/source-control#version-control

Intro Data Flow Script (DFS)

Programmiersprache hinter den Mapping Data Flows

- Data Flow Script auch Teil des Json Codes (ganz unten)
- Entsteht aber beides automatisch durch die grafische Oberfläche
- Editierbar
- Inhalte übertragbar in andere Data Flows / Projekte
- Verwenden in Powershell cmdlets
- Einfaches Find & Replace im Text
- Programmatisch Data Flows erzeugen aus Metadaten
- Einfacheres Finden von Abweichungen zwischen Versionen – vereinfacht auch in Json seit der Einführung von scriptline
- Viele Beispiele von MS oder Blog-Artikeln können einfach integriert werden




```
source(output(
        medallion as string,
        hack_license as string,
        vendor_id as string,
        rate_code as string,
        store_and_fwd_flag as string,
        pickup_datetime as string,
        dropoff_datetime as string,
        passenger_count as short,
        trip_time_in_secs as long,
        trip_distance as double,
        pickup_longitude as double,
        pickup_latitude as double,
        dropoff_longitude as double,
        dropoff_latitude as double
    ),
    allowSchemaDrift: false,
    validateSchema: false,
    ignoreNoFilesFound: false) ~> TripData
source(output(
        medallion as string,
        { hack_license} as string,
        { vendor_id} as string,
```

https://docs.microsoft.com/en-us/azure/data-factory/data-flow-script

Intro Data Flow Script (DFS)

Scriptline in Json Code Definition

- JSON Definition in ADF für Data Flow früher: c

```
        }
    ],
    "script": "source(output(\n\t\tmedallion as string,\n\t\thack_license as string,\n\t\tvendor_id as string,\n\t\trate_code as string,\n\t\tstore_and_fwd_fl
}
```

- JSON Definition in ADF für Data Flow seit Einführung scriptline:

```
],
"scriptLines": [
    "source(output(",
    "          {Emp ID} as string,",
    "          {Name Prefix} as string,",
    "          {First Name} as string,",
    "          {Middle Initial} as string,",
    "          {Last Name} as string,",
    "          Gender as string,",
    "          {E Mail} as string,",
    "          {Father's Name} as string,",
    "          {Mother's Name} as string,",
    "          {Mother's Maiden Name} as string,",
    "          {Date of Birth} as string,",
    "          {Time of Birth} as string,",
```

https://techcommunity.microsoft.com/t5/azure-data-factory-blog/adf-mapping-data-flows-adds-scriptlines-and-link-service/ba-p/3249929#M589

scieneers
DRIVEN BY DATA

Demo

Data Flow Script

scieneers
DRIVEN BY DATA

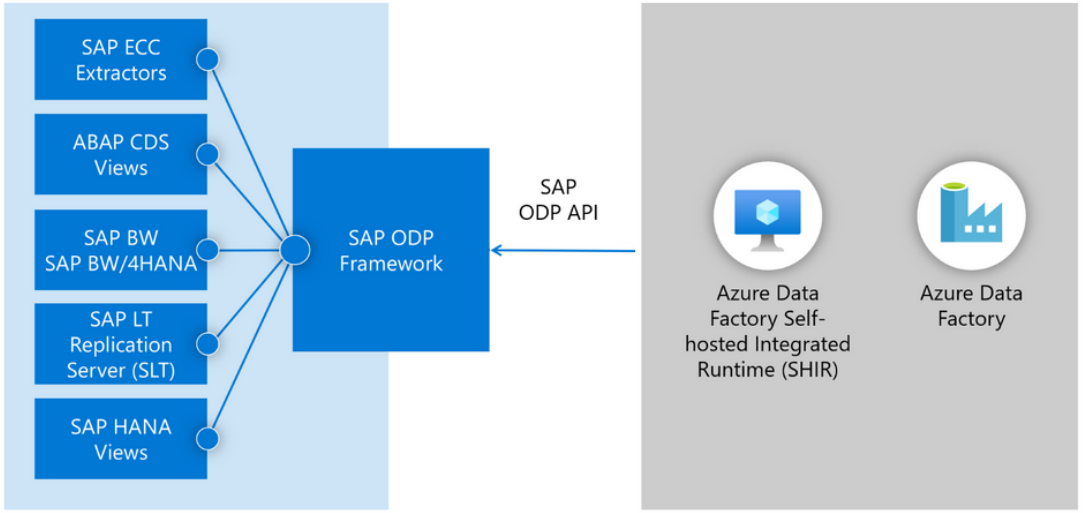
SAP CDC is coming!

I want it all (not) – incremental sourcing

SAP CDC

Verwenden von SAP eigener Change Tracking Mechanismen für Delta Loads

- new SAP CDC connector leveraging SAP ODP framework
- can connect to all SAP systems that support ODP, such as R/3, ECC, S/4HANA, BW, and BW/4HANA
- directly at the application layer or indirectly using SAP Landscape Transformation (SLT) replication server as a proxy
- fully or incrementally extract SAP data
- physical tables & logical objects such as ABAP Core Data Services (CDS) views
- will be released for public preview on June 30, 2022



https://techcommunity.microsoft.com/t5/azure-data-factory-blog/announcing-the-upcoming-preview-of-sap-cdc-in-azure-data-factory/ba-p/3420904

scieneers
DRIVEN BY DATA

Hands-On

Auf geht's Leute!