

scieneers
DRIVEN BY DATA

Data Factory in Fabric

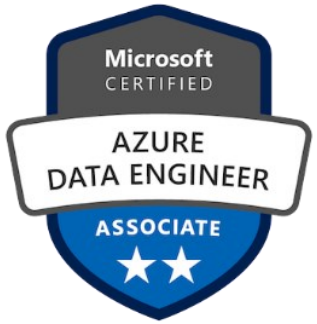
the good, the bad and the ugly

Stefan Kirner



- > 20 years experience using Microsoft Data Platform
- Data Monster Chapter Lead Karlsruhe & Beirat
- Director Business Intelligence scieneers GmbH
- Twitter: @KirnerKa
- LinkedIn: https://www.linkedin.com/in/stefan-kirner/













Agenda

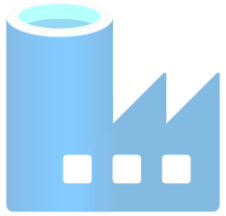
- Intro – Data Factory in Fabric
- Features vs. Azure Version
- Application Lifecycle Management
- Cost Comparison
- Demo
- Summary

Intro – Data Factory in Fabric



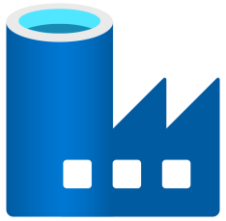
Evolution Azure Data Factory v1 to Fabric Data Factory

ADF v1



Data
Ingestion

ADF v2



+ Transform
+ Integration
Runtimes

Data
Factory in
Synapse



+ Spark Notebooks
Integration

Data
Factory
In Fabric



+ SaaS
+ Data Platform
Design
Integration

Microsoft Fabric in a nutshell

scieneers
DRIVEN BY DATA

Technology

End-to-End-
Analyse- and
Dataplattform

- Data Engineering
- Data Science
- Datawarehouse
- Dashboarding

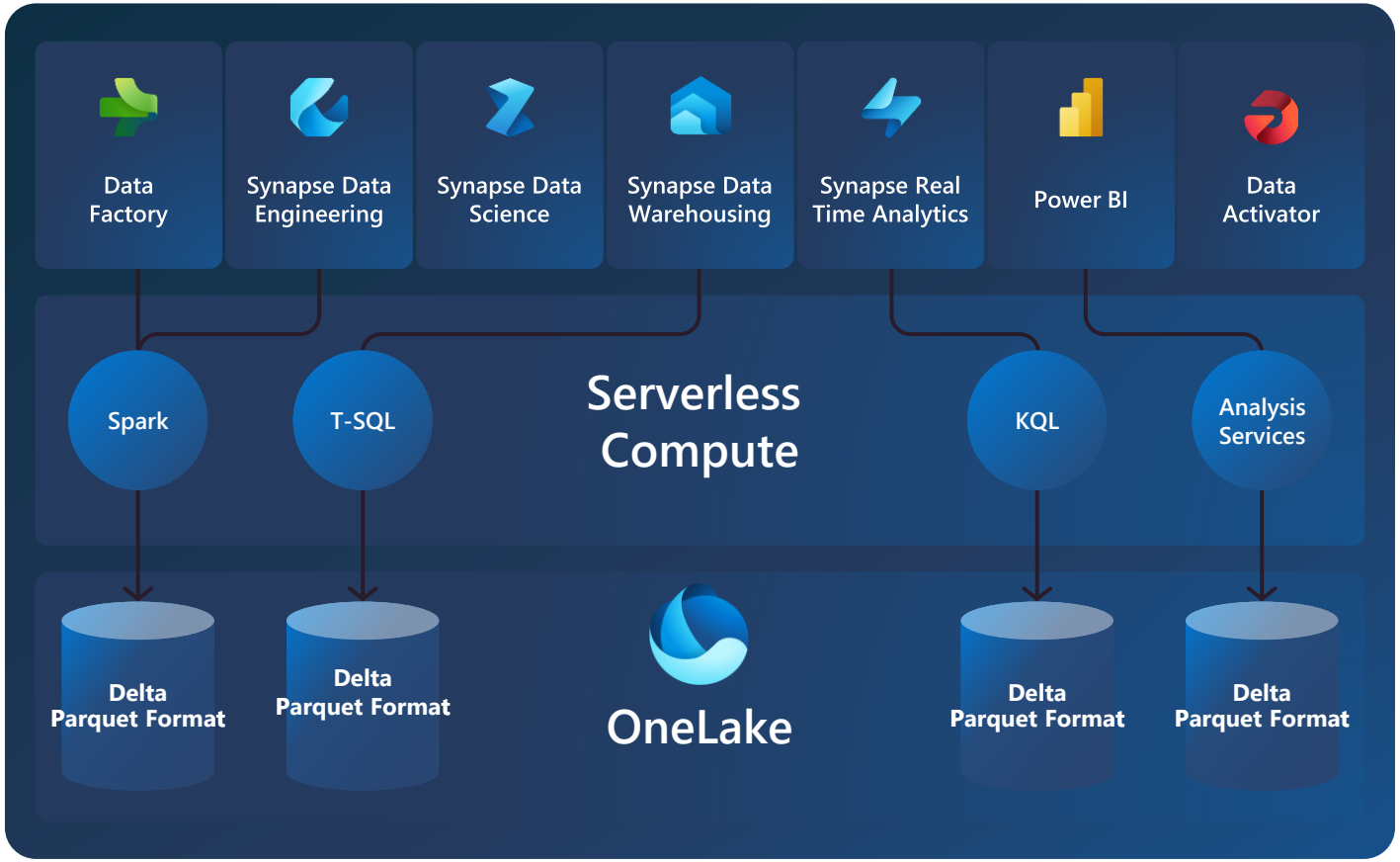
Business

Software as
a Service

- Monthly
Fixprice
- Savings by
Reservations
- Pausable
(off-times)
- Up/Downscale
anytime

Parts of Fabric and how they interact

- Unites any tooling for analytics in one GUI
- One storage for all different compute engines and tools
- Separation of compute and storage
- One central security model possible

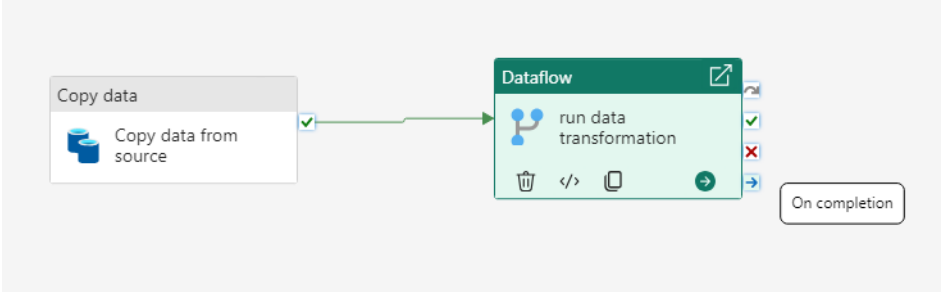


Feature comparison



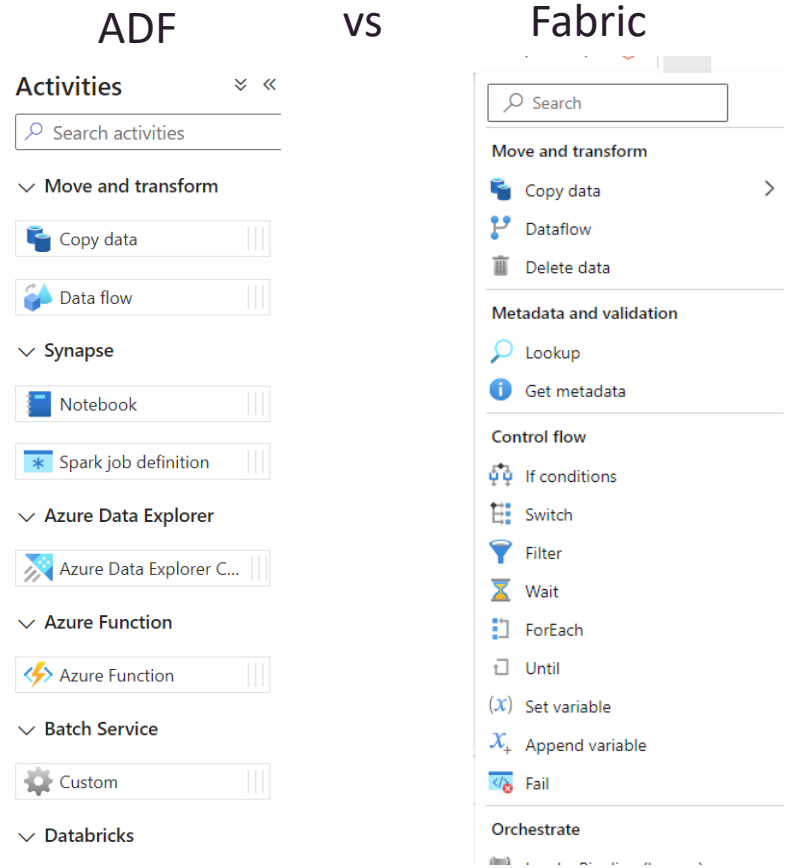
Pipelines to structure data loading activities by running order and constraints

- Continues to live in both versions
- Orchestration and triggering of ETL/ELT processes
- Also the well-known and often used copy activity
- Differences in support of connectivity and some activities



Data Pipeline Activities continuity between ADF and Fabric version

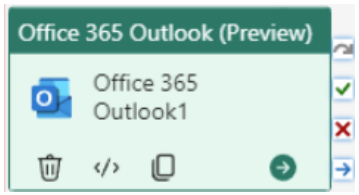
- 90% of activities in ADF also supported in Fabric
- Not supported pipeline activities in Fabric version:
  - Validation
  - Run SSIS package
  - Run Mapping Data Flow
- Different
  - Exec Databricks Notebooks supported in Fabric
    - Missing options for Python / JAR
    - Spark Jobs available as separate component

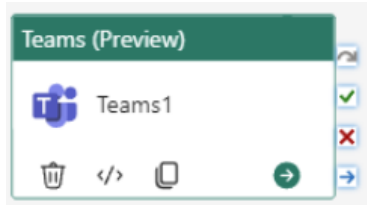


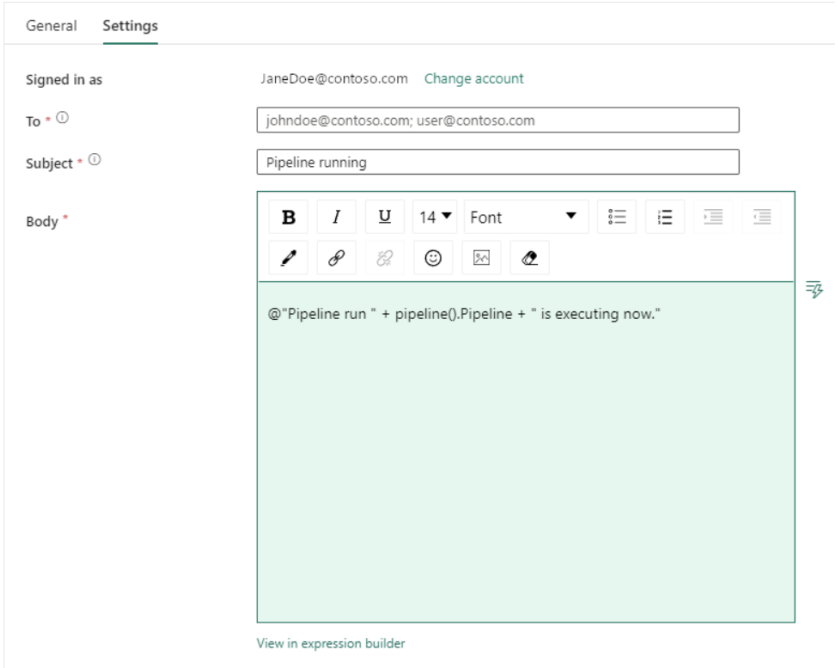
https://learn.microsoft.com/en-us/fabric/data-factory/activity-parity

Activities only in Fabric Data Factory

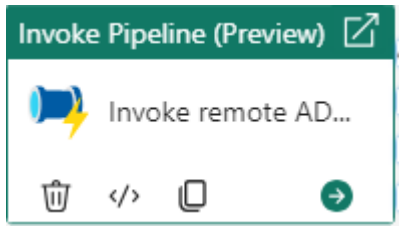
- Send emails using O365 and messages to Teams
  - Good for pushing data bound notifications or about runtime states
  - tbd is using personal accounts for this, still kind of buggy
- Invoke ADF & Synapse remote pipelines
- Data Flow Gen 1 & 2
- Refresh Semantic Model



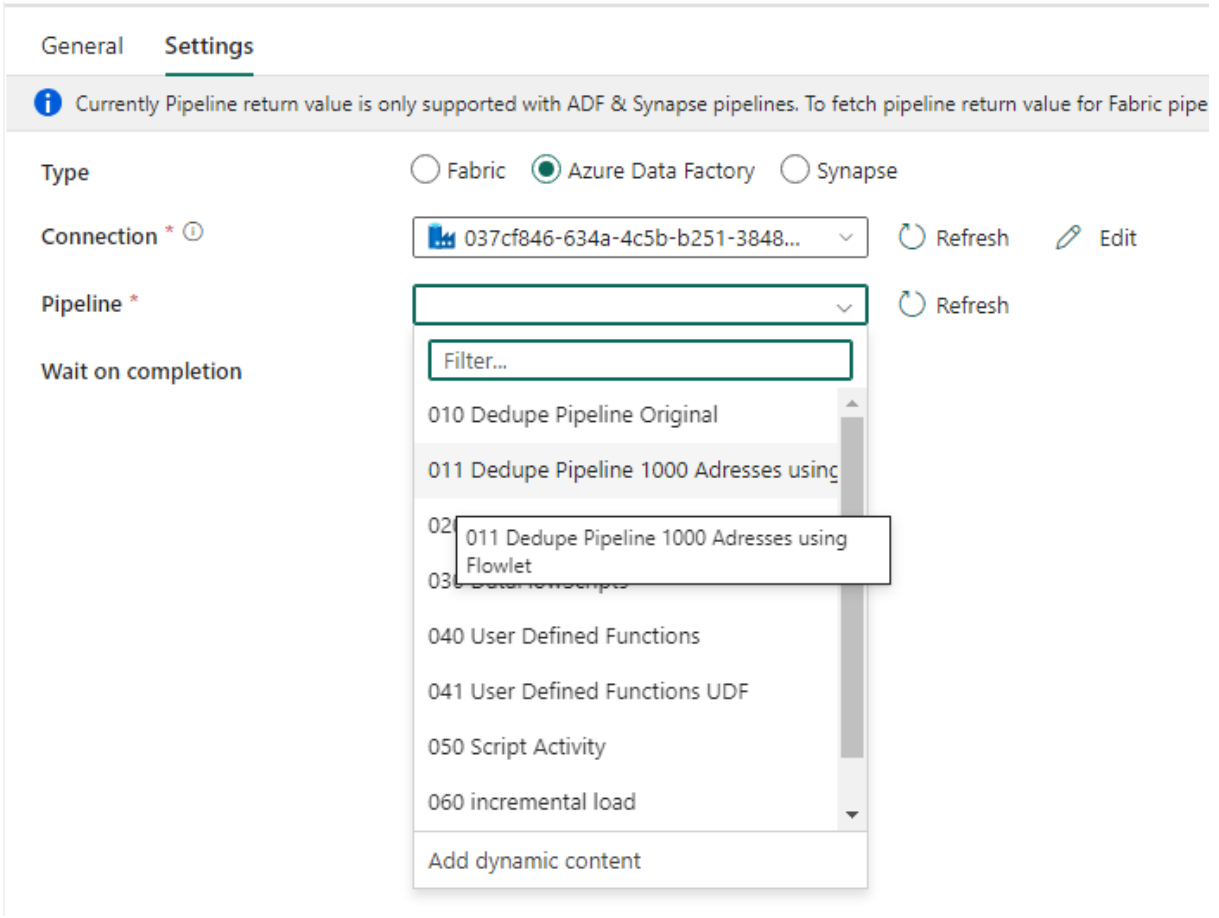




How to invoke remote ADF & Synapse pipelines







Data Factory and alternative ways for getting data into Fabric

scieneers
DRIVEN BY DATA

Ingest Data

Store data in Lakehouse or Warehouse

- Fabric ADF:
  - Copy Activity
  - Data Flow
- Notebook / Spark Job
- Eventstream
- Database Mirroring
  - Snowflake, CosmosDB, Azure SQL
- Pro:
  - Data historization independent of source system capabilities
  - Easier Data Validation and Error handling

Shortcuts

Set reference to access data – no import

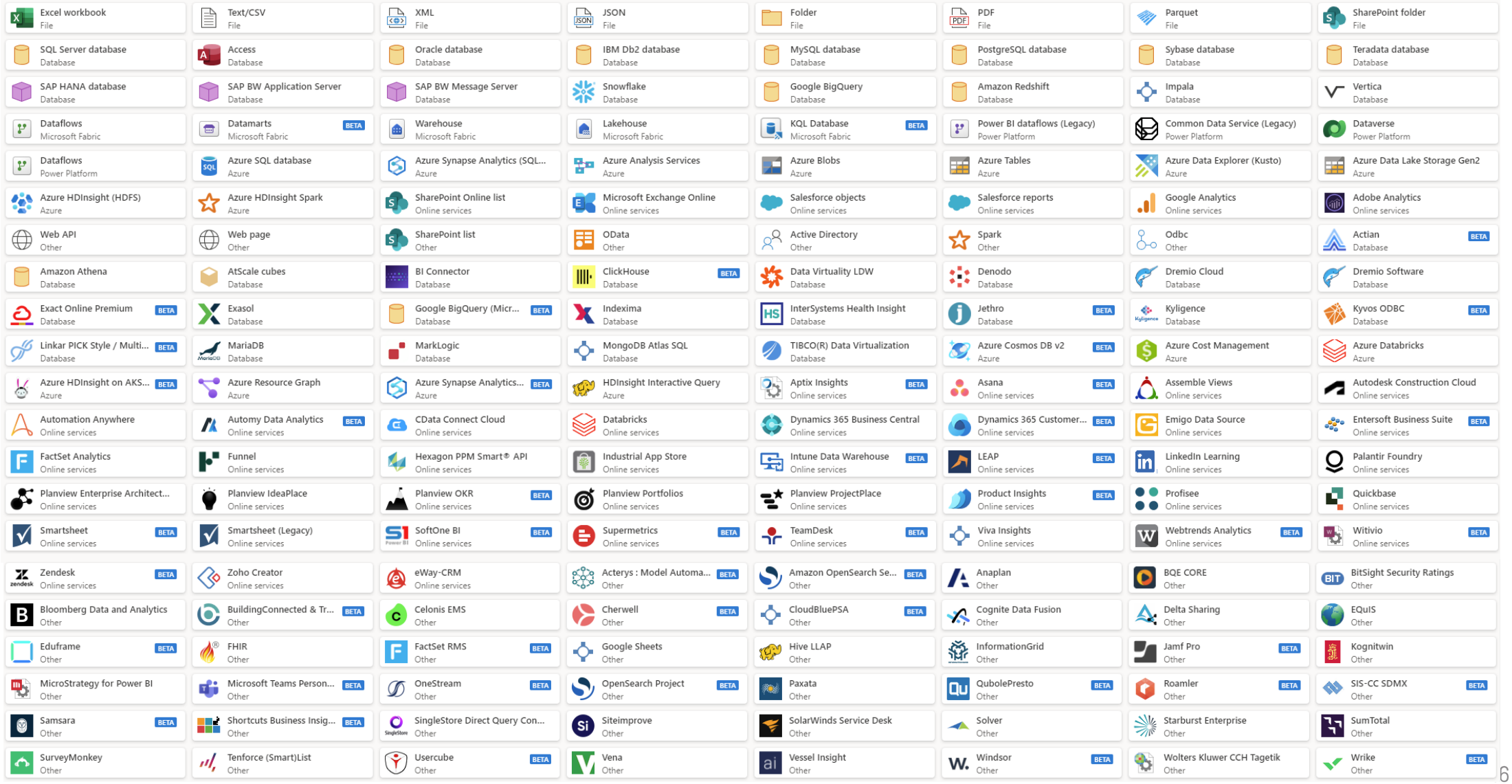
- External Shortcut
  - AWS, Azure, Google…
- Internal Shortcut
  - Fabric Lakehouse, Warehouse, KQL tables
- Pro:
  - Realtime integration - Data is always up to date
  - Do not have to handle incremental / update logic yourself

New activity in Fabric: Data Flows Gen2

- Data Destinations: Data can be persisted after extraction and transformation
- GUI based data transformations, produces reusable M code
- Replaces Mapping, Wrapping and Data Flow Gen1
- Save as draft could avoid some headaches
- Better scalability due to storing data instead of keeping anything in memory
- Reusing M queries from Power BI Data Flow Gen1 and Power BI possible

Available Connections for Data Flows Gen2

scieneers
DRIVEN BY DATA

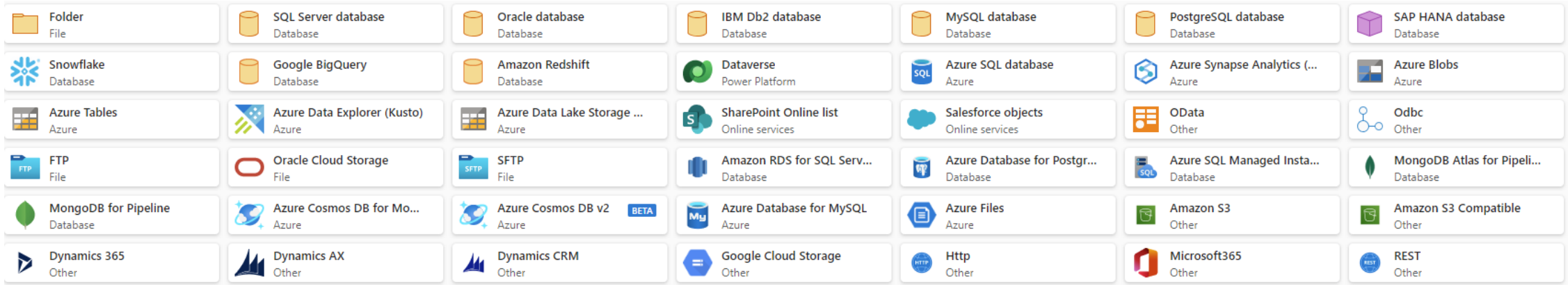


PowerBI.com, 18.9.2024

Available Connections for Copy Activity in Fabric

scieneers
DRIVEN BY DATA

Documentation outdated and inconsistent – always check in Fabric



PowerBI.com, 18.9.2024

Data Ingestion: comparing Data Flows vs. Copy Activity

scieneers
DRIVEN BY DATA

Data Flows	Copy Activity

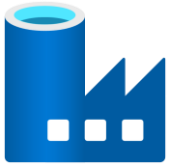
• Can do much in one single step:
  • extract, transform & store data
  • combine different data sources and reference data
• Connections to 300+ source systems, much more than Copy activity
	• Best for large datasets
	• Better for metadata-driven approaches
	  • Good for ETL/ELT Automation
	  • 1:1 copy to staging pattern
	• Best integration in control flow
	  • like looping, parametrization..
	  • triggering external actions like Databricks NBs, Azure ML or Stored Procedures for transformation
	• Deployment pipelines and git integration supported

„Data Flows“ in ADF vs. Power BI vs. Fabric

ADF Mapping Data Flows	Power BI Data Flows (Gen 1)	Fabric Data Flows Gen2
• Persist data in different sinks • Spark based scale out • Non-reusable scala code generated at runtime • Pay by runtime, could be expensive	• Persist data only Power BI Model • No scale out, scale up due to In-memory engine • Reusable M-Code as result • Fixed price, pay by license	• Persist data in different sinks • No scale out, scale up due to In-memory engine • Reusable M-Code as result • Fixed price, pay by license

Artefacts naming – not too much changes

scieneers
DRIVEN BY DATA



Azure Data Factory



Data Factory in Fabric

Pipeline vs. Data Pipeline

Linked Service vs. Connection

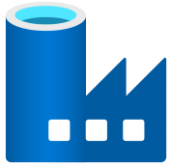
Dataset vs. N.A.

Publish vs. Run, Save

Self-hosted Integration Runtime (SHIR) vs. On-premises Data Gateway

https://learn.microsoft.com/en-us/fabric/data-factory/compare-fabric-data-factory-and-azure-data-factory

Connectivity to your on-premises source data



Self-hosted Integration Runtime (SHIR)

- Client Agent which establishes encrypted connectivity between on-prem and Azure services
- Used by ADF & Synapse Analytics



On-premises Data Gateway

- Client Agent which establishes encrypted connectivity between on-prem and Microsoft cloud services
- Used by Power BI, Power Apps, Logic Apps, Azure Analysis Services, Power Automate & Fabric

Impact using SHIR vs Data Gateway, just plug & play?

scieneers
DRIVEN BY DATA

- Data Factory used for Ingestion / Copying source data in Fabric
- Any connectivity provider for Fabric DF had to be adapted to use On-premises Data Gateway
- **Not all** connectivity options supported in ADF are also available for Fabric Copy Activity at this point in time (e.g. SAP), but most for Data Flows Gen2
- Even more confusion: Fabric Pipelines and Data Flows Gen2 each support different sources
- Make sure that your data source is really supported
- Workaround: additional ADF / Synapse Links etc

How to automate execution – triggers ADF vs. Fabric DF

Different types of triggers available – scheduled and events

Type *

Schedule

Filter...

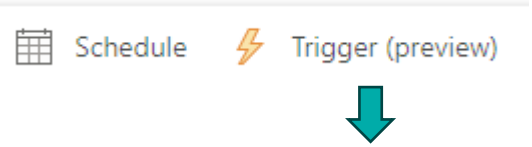
Schedule

Tumbling window

Storage events

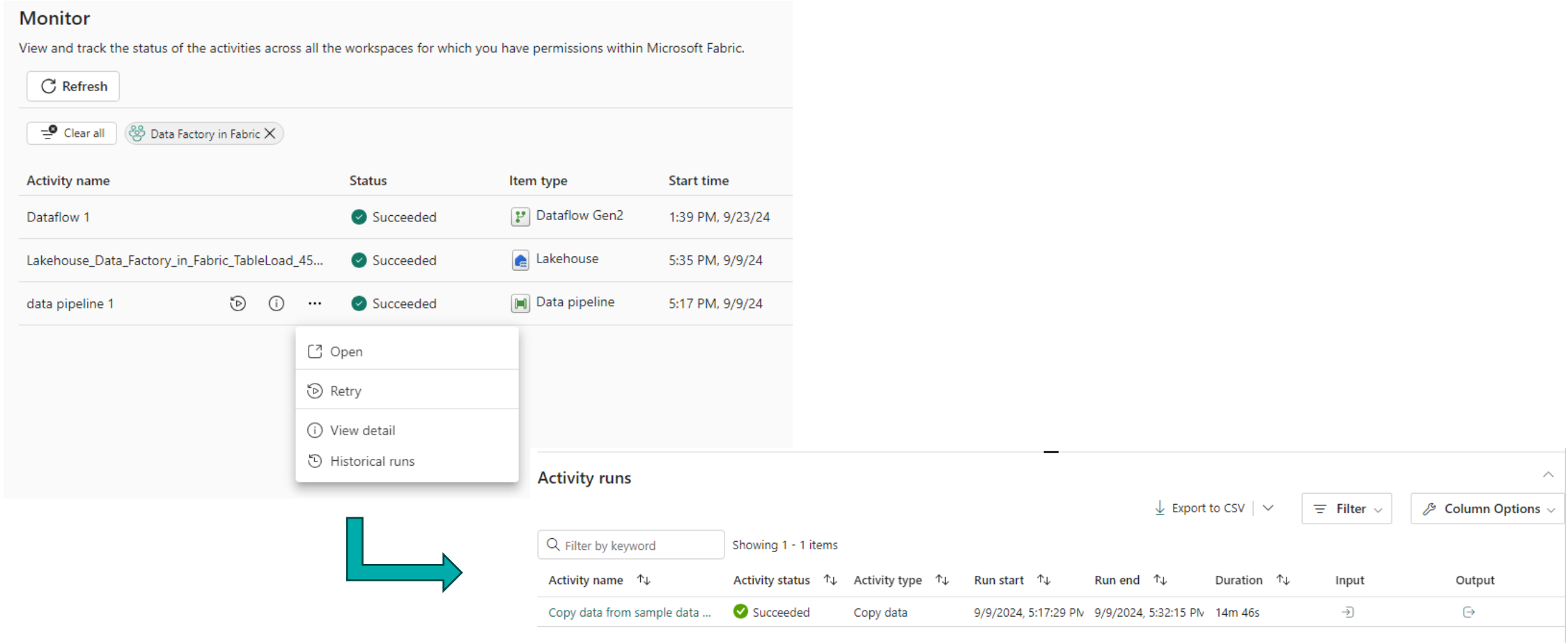
Custom events

Scheduled and events in preview





Check state of your processes using the central Monitoring Hub in Fabric



What about other runtimes supported in ADF?

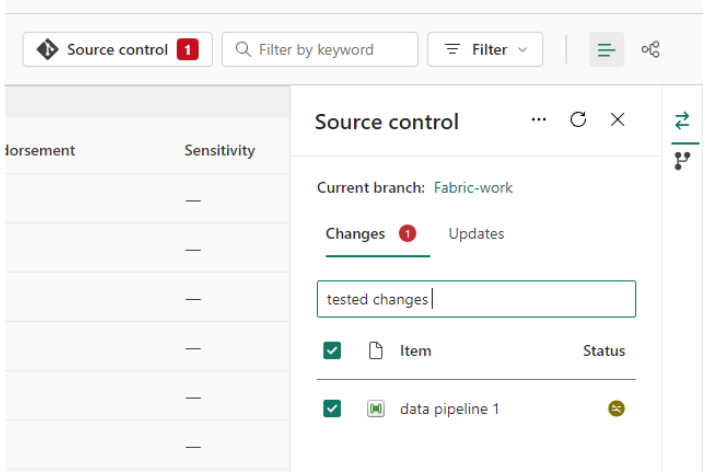
- No concept „integration runtimes“ in Fabric
- ADF SSIS integration runtimes?
  - Not supported yet and not on the roadmap for Fabric
- ADF Runtime for airflow?
  - Support in Fabric in preview as „ Apache Airflow Job“

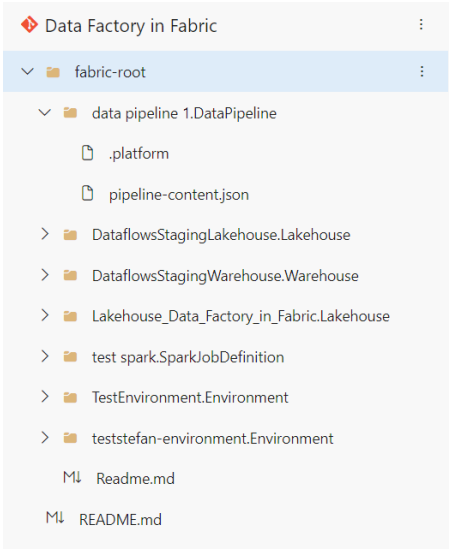
Application Lifecylce Management



Teamwork in Fabric Development

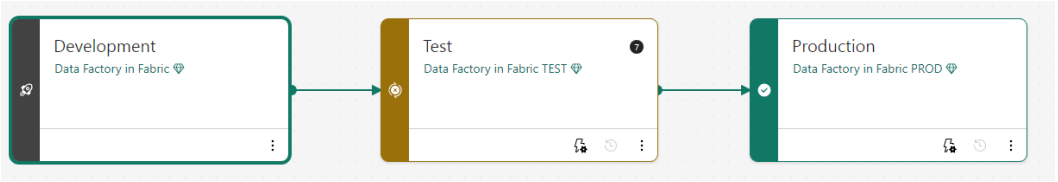
- w/o a repo save commits to the live version (for anyone)
- Make sure nothing gets lost by using a git repository
  - No auto-commit when saving, go to workspace, source control and commit or get updates
- For parallel working with notebooks the lowest Fabric capacities are way too limited

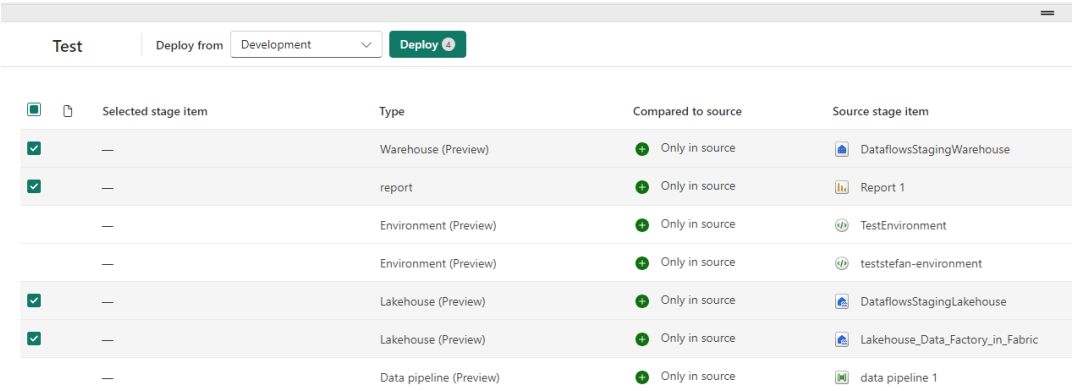


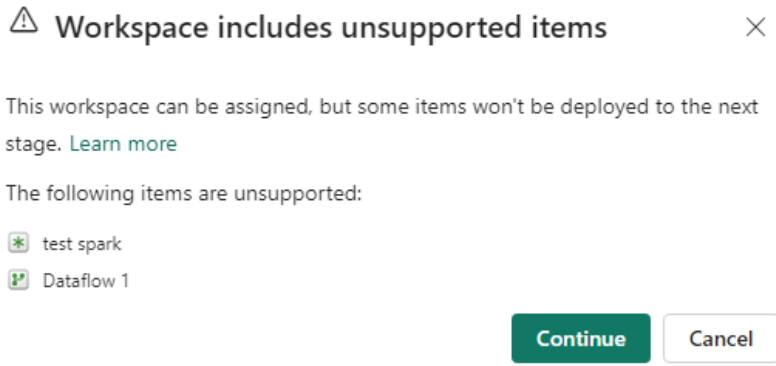


Deployments

- GUI-based Deployment pipelines in Fabric available
- 3 workspaces dev, test, prod to be assigned
- Selection of single items possible
- Still not all items supported including Dataflows Gen2 und Spark Jobs
- Difficult to customize e.g. creating folders / copy of config files
- Only deploy from Stage 2 Stage not from a git repo branch
- Just one for all?
- Still buggy…

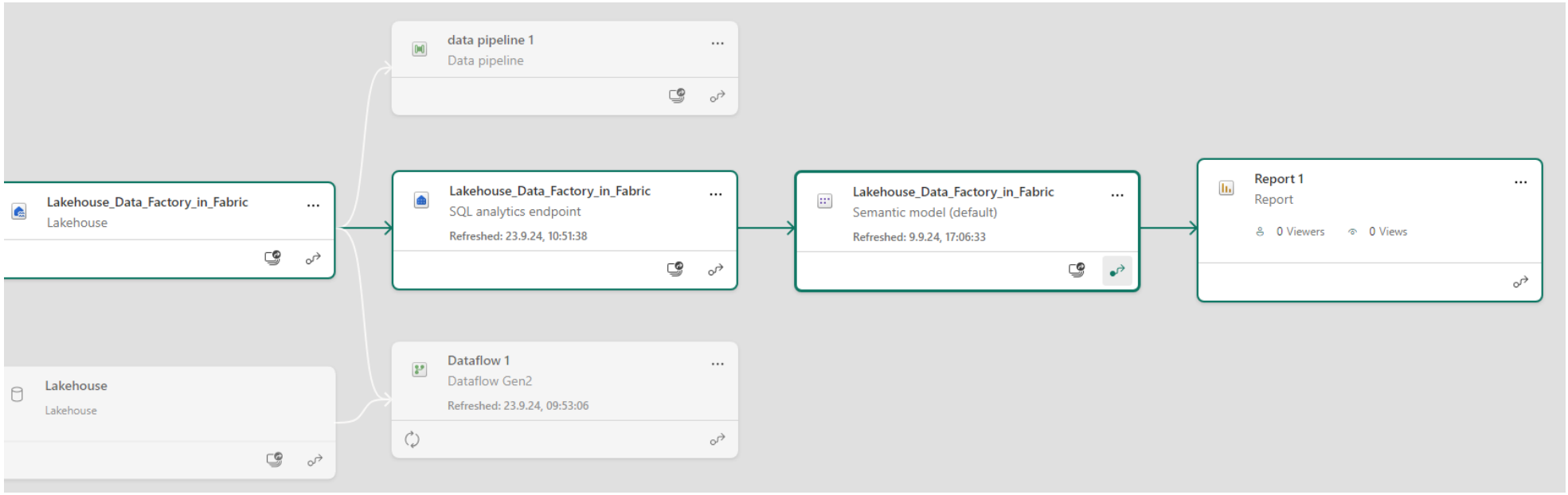






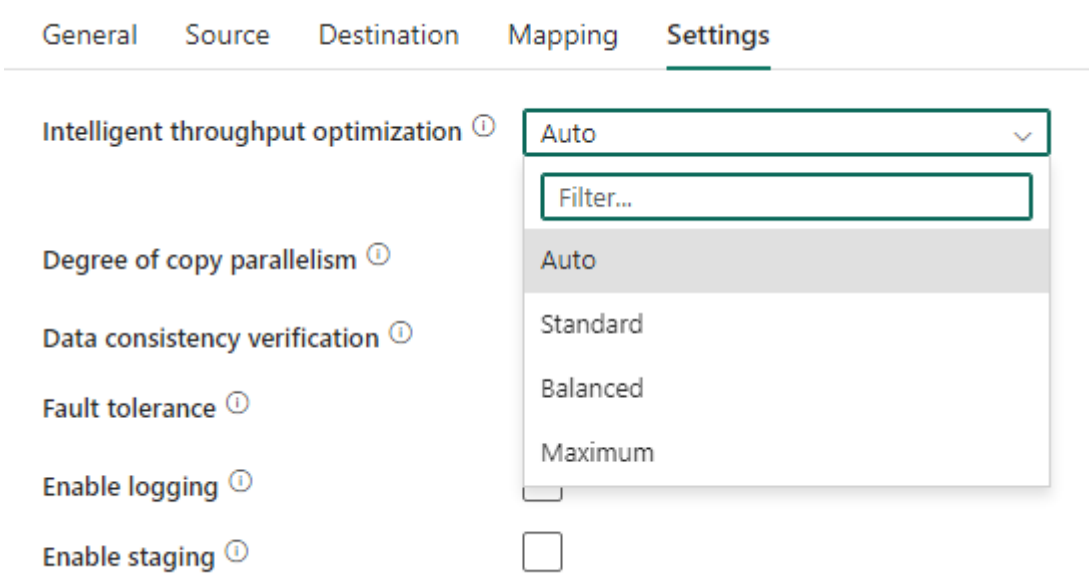
Now as this should be one product – is there also Data Lineage?

Yes! Maybe more detailed in future



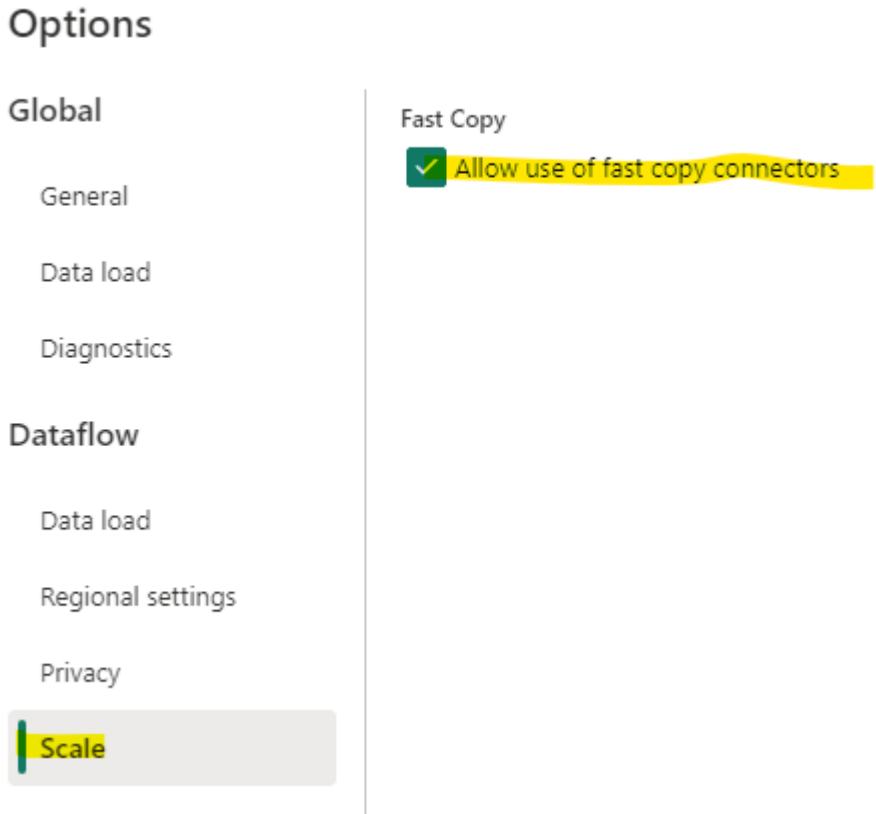
How to adjust performance for Copy Activity

- Option to adjust inside each activity
- No Data Integration Unit steps like in ADF (due to capacity limits)
- other options are equal including Staging option



How to adjust performance for Data Flows Gen2

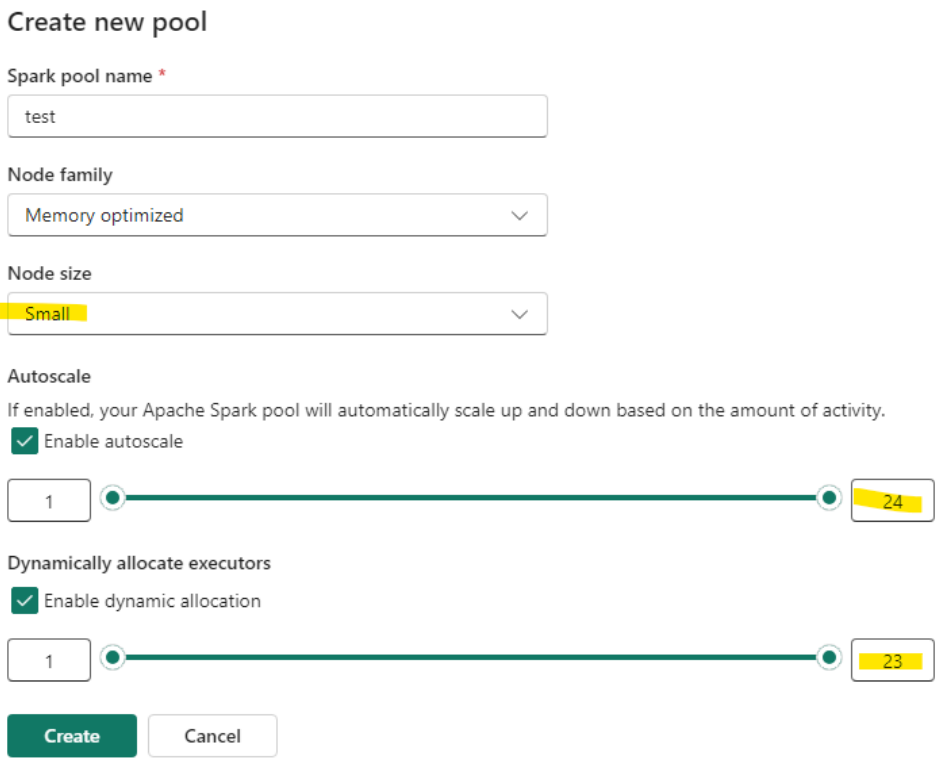
- Fast copy allowed available
- Use Staging option
- „Buffer“ data and reference



Setting for scale in Data Flow Gen2

How to adjust performance for Spark workloads (Notebooks)

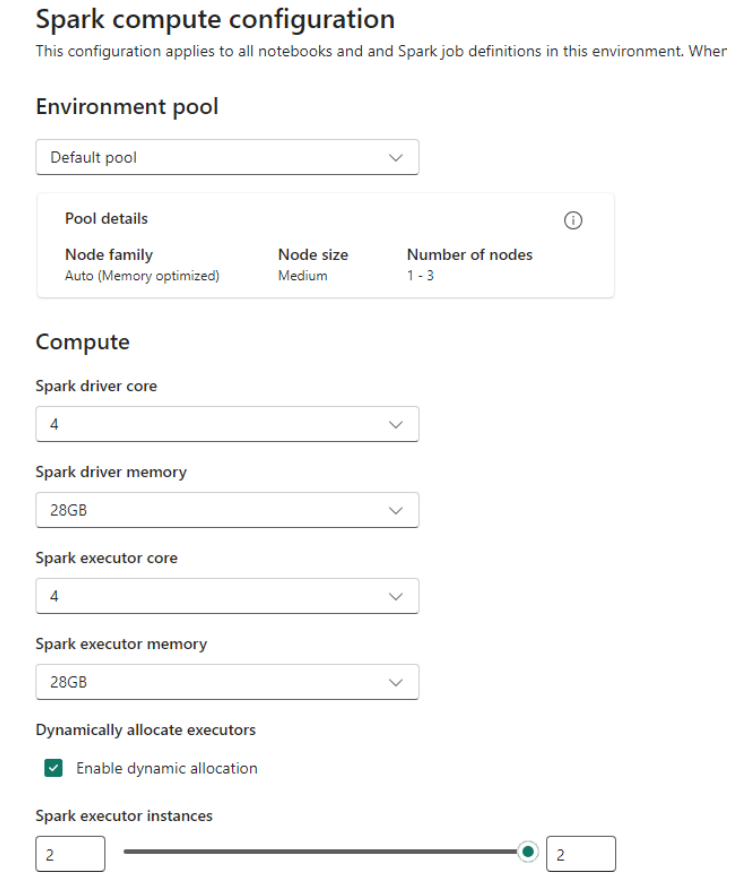
- Available Nodes depend on capacity and chosen Node Size:
  - 6 max for F4 (any node size)
  - 24 max for F16 small nodes
- Use option „high concurrency for Notebooks“
- Only Memory optimized Node Family available (Sep 24)



Spark Pool options in Workspace Settings for F16 capacity

How to adjust performance for Spark workloads (Notebooks)

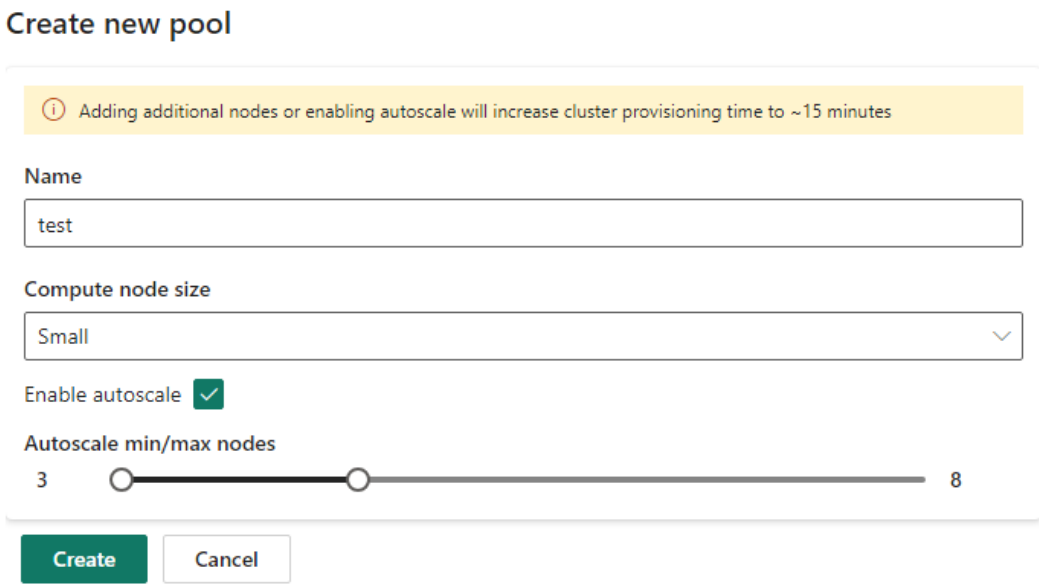
- More detailed compute options for the pool when creating a new environment



Compute and other options in a an environment

How to adjust performance for Apache Airflow Runtime

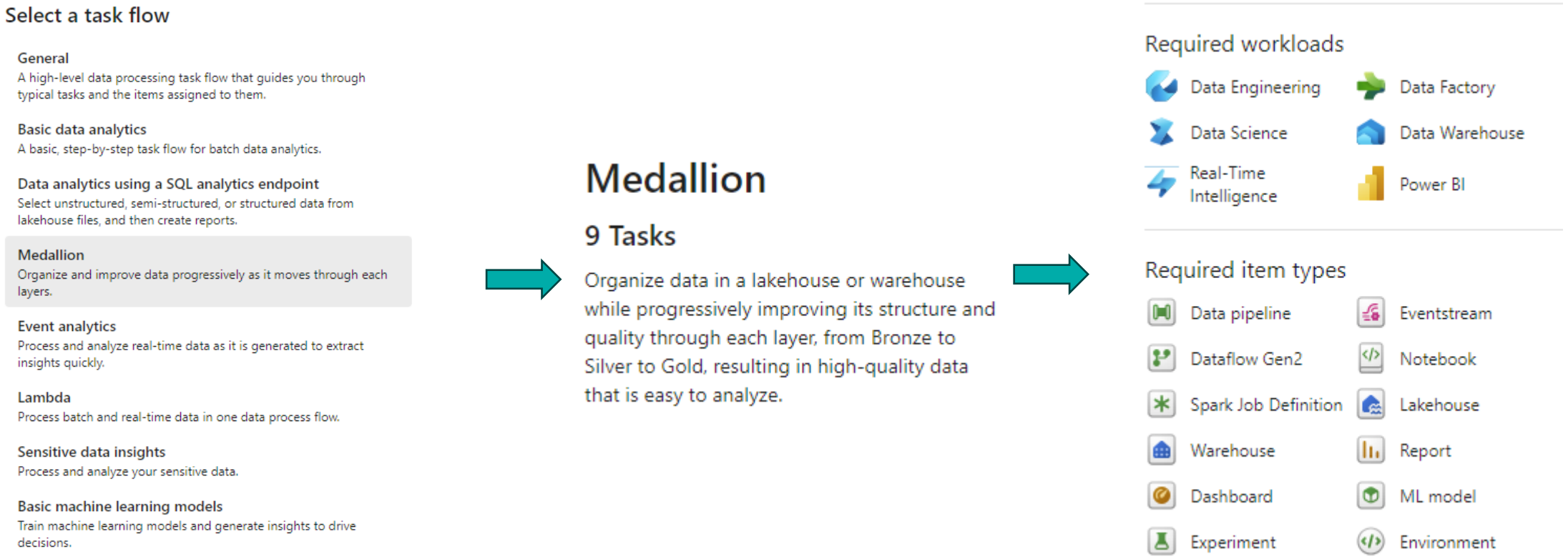
- Available Nodes depend on capacity and chosen Node Size:
  - 20 max for F4 and F16 for any node size?



Apache Airflow Pool options for F16 capacity

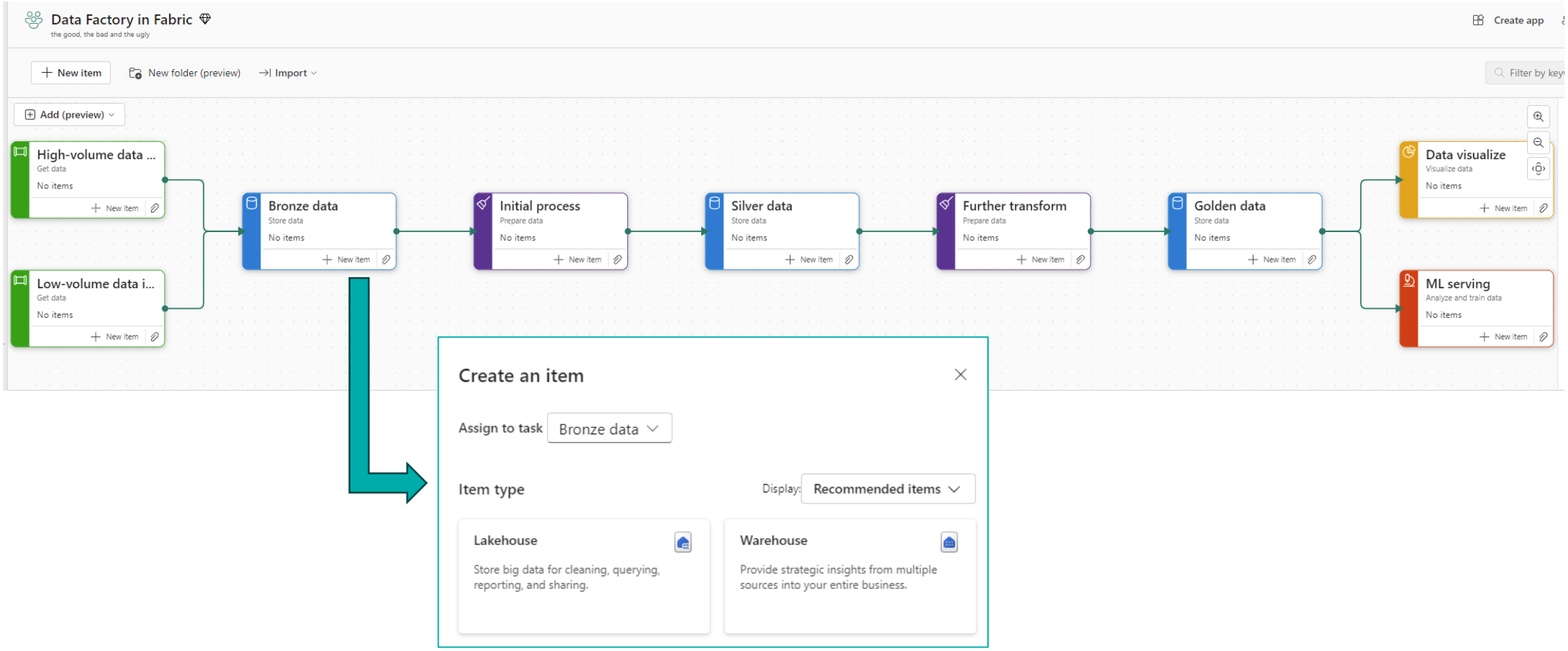
Task flows

Structure from the beginning



Task flows

Leads to this



Task flows

- Optional, not a must
- Nice overview of your items per layer and their last run state
- Not a replacement for control flow in pipelines
- No „run“ button
- Adjustable to your needs and pattern
- Good kickstart for newcomers to learn about architecture patterns

data pipeline 1	Data pipeline	High-volume data ingest	Stefan Kirner	—	—
Dataflow 1	Dataflow Gen2	Low-volume data ingest	Stefan Kirner	—	N/A
Lakehouse_Data_Factory_in_Fabric	Lakehouse	Bronze data	Stefan Kirner	—	—
Lakehouse_Data_Factory_in_Fabric	Semantic model (default)	Bronze data	Data Factory in Fabric	9.9.24, 17:06:33	N/A
Lakehouse_Data_Factory_in_Fabric	SQL analytics endpoint	Bronze data	Data Factory in Fabric	9.9.24, 17:35:04	N/A
test spark	Spark Job Definition	Initial process	Stefan Kirner	—	—

Integration and Migration scenarios

- No overall „Migration Assistant“ available
- Data Pipelines – most activities supported
- ADF OneLake/Lakehouse Connector to push data to Fabric
- Mapping Data Flows
  - Convert to Notebook / Spark Job using tooling*
  - MS Learn description
  - CAT offers help
- Invoke Remote Pipeline
- Mount existing Azure Data Factory into Fabric for managing and monitoring of all pipelines in Fabric GUI
- Reuse Data Flows Gen1 Queries in Gen2 by Export/Import in PQT Files or copy/paste

https://github.com/sethiaarun/mapping-data-flow-to-spark

Costs



From activity based to fixed price by capacity

ADF costs driven by consumption and # of activities run

- Could be cheaper for projects with less activies and runtimes
- But very complicated to predict and reproduce
- Additional costs for processing engines could occur
  - Databricks Cluster…

Fabric Data Factory costs depend on the chosen capacity

- fixed price for Fabric capacity
- Include Fabric based processing
- You could pay more for DF function due to higher Fabric capacity driven by other needs
  - E.g. less data integration but tons of users on the reports
- Optimization by pausing / scaling / reservations

Sample #1 Monthly statistics

Azure Data Factory Pipeline with 50 activities running 1x per month	Fabric Data Factory
• **Costs <100€*** • Calculation details: 200h DTU, 100h Pipeline, 50h external Pipeline, Data Flow 8Cores 50h • No external processing engine used	• **Costs 285€*** • F2 capacity is minimum requirement, but really slow and no parallel processing • all time up • Reduced: • Pausing, saving ½ time ~142€ • Reservation: 171€

*West Europe pricing

Sample #2 hourly production overview

Azure Data Factory Pipeline with 10 activities running 24x per day	Fabric Data Factory
• **Costs 1028€***	• **Costs 570€***

• Calculation details:
½ hour runtime 24x per day= 12h
360h per month
1440h DTU, 720h Pipeline, 360h external Pipeline, Data Flow 8Cores 360h
• No external processing engine used
	• F4 capacity could work
	• all time up
	• Reduced:
	  • Pausing, saving ½ time 285€
	  • Reservation: 343€
	• Next capacity level would be F8 with 1138€ / 686€ reserved

*West Europe pricing

Demo



Summary



The good, the bad and the ugly

Good

- Data Factory extended with know options and features
- SaaS model simplifies entry
- Fixed Price Model for certainty
- Data Lineage
- Dynamic cost saving by pause / scaling options

Bad

- High price entry point for small, focussed data movement projects
- Lowest capacities not usable for Spark Notebooks

Ugly

- Missing Deployments and git support for Data Flows Gen2
- GUI bugs
- Missing stableness
- Copy activity supporting less data sources in Fabric

scieneers
DRIVEN BY DATA

Thanks for your attention, I appreciate your feedback!

SQL KONFERENZ 2024