Regionalgruppe Rhein-Main

PASS DEUTSCHLAND e.V.

Data flows in Azure Data Factory – endlich auch hier Transformationen!

Stefan Kirner

03.04.2019

About me

Stefan Kirner

- PASS Chapter Lead Karlsruhe ski@sqlpass.de
- Teamlead BI Solutions @inovex
- MCSE for Data Management & Analytics & Cloud Infrastructure
- Microsoft P-TSP Data Platform
- Twitter: @KirnerKa

PASS Essentials

Uwe Ricken: SQL Server Security – Sicherheit in Microsoft SQL Server

2. Mai 2019

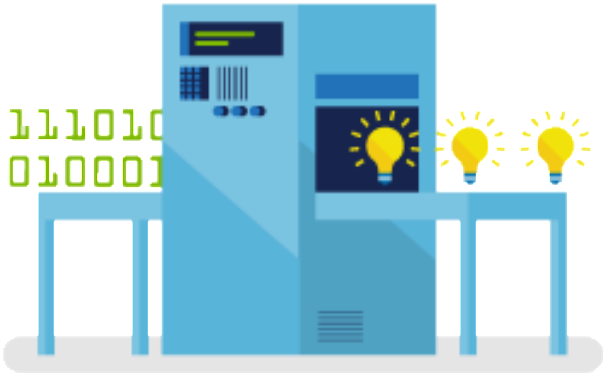
Karlsruhe, inovex GmbH

https://www.sqlpass.de/events/sql-server-security-2019-05-02/

- Richtige Konfiguration des Dienstkontos von Microsoft SQL Server und SQL Server Agent
- Die Besonderheiten von sysadmin
- Danger und Beauty von „xp_cmdshell“
- […]
- Alle Sicherheitsaspekte von Microsoft SQL Server werden in diesem Workshop ausführlich behandelt und sollen auf eigenen Laptops mit umfangreichen Demos und Übungen selbst ausprobiert werden.

Agenda

- Intro Data Factory v2
- Control Flow & Triggers
- Data Flow
- Q+A



Intro
Data Factory v2

Data Factory

Managed Data Integration Service

New Pipeline Model

Rich pipeline orchestration
Triggers – on-demand, schedule, event

Data Movement as a Service

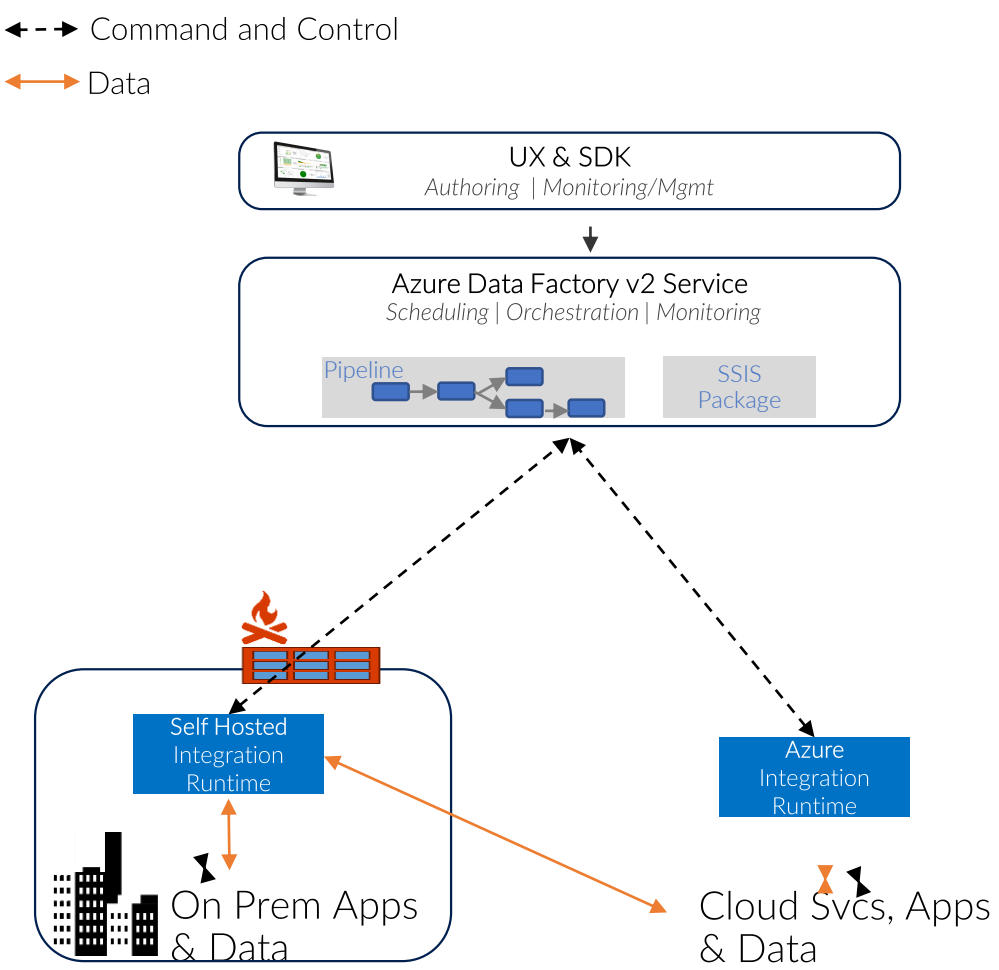
Cloud, Hybrid
30 connectors provided
Data flow as NEW Data Transformation Layer

SSIS Package Execution

In a managed cloud environment
Use familiar tools, SSMS & SSDT

Author & Monitor

Programmability (Python, .NET, Powershell, etc)
Visual Tools for Control Flow and NEW: Data Flow



Data Factory

A data integration account.

Location of orchestration, service metadata

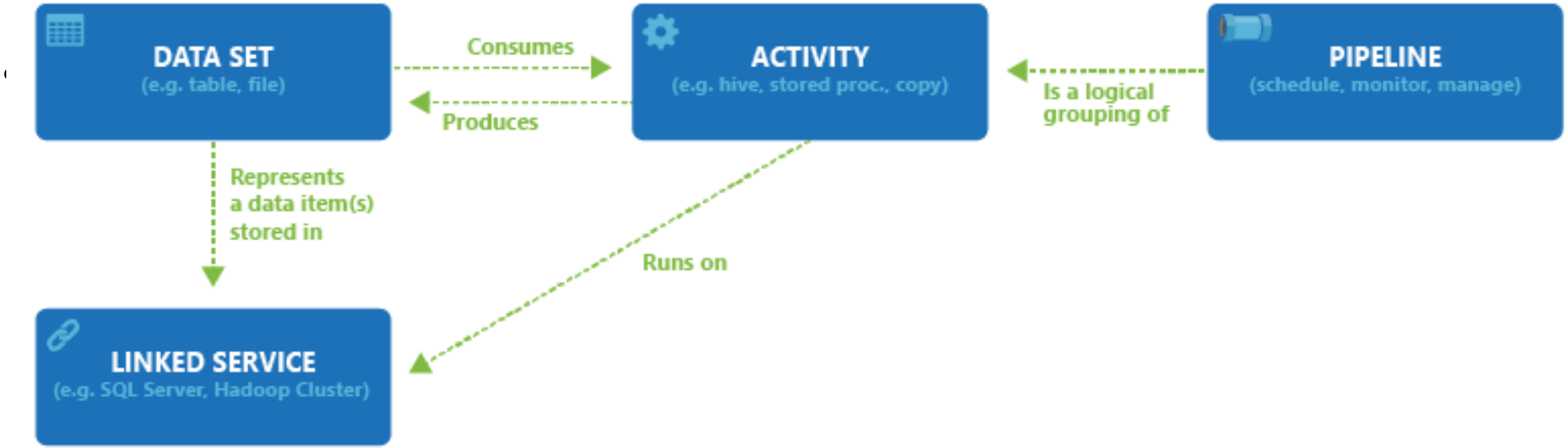
Integration Runtime (IR)

ADF's execution engine

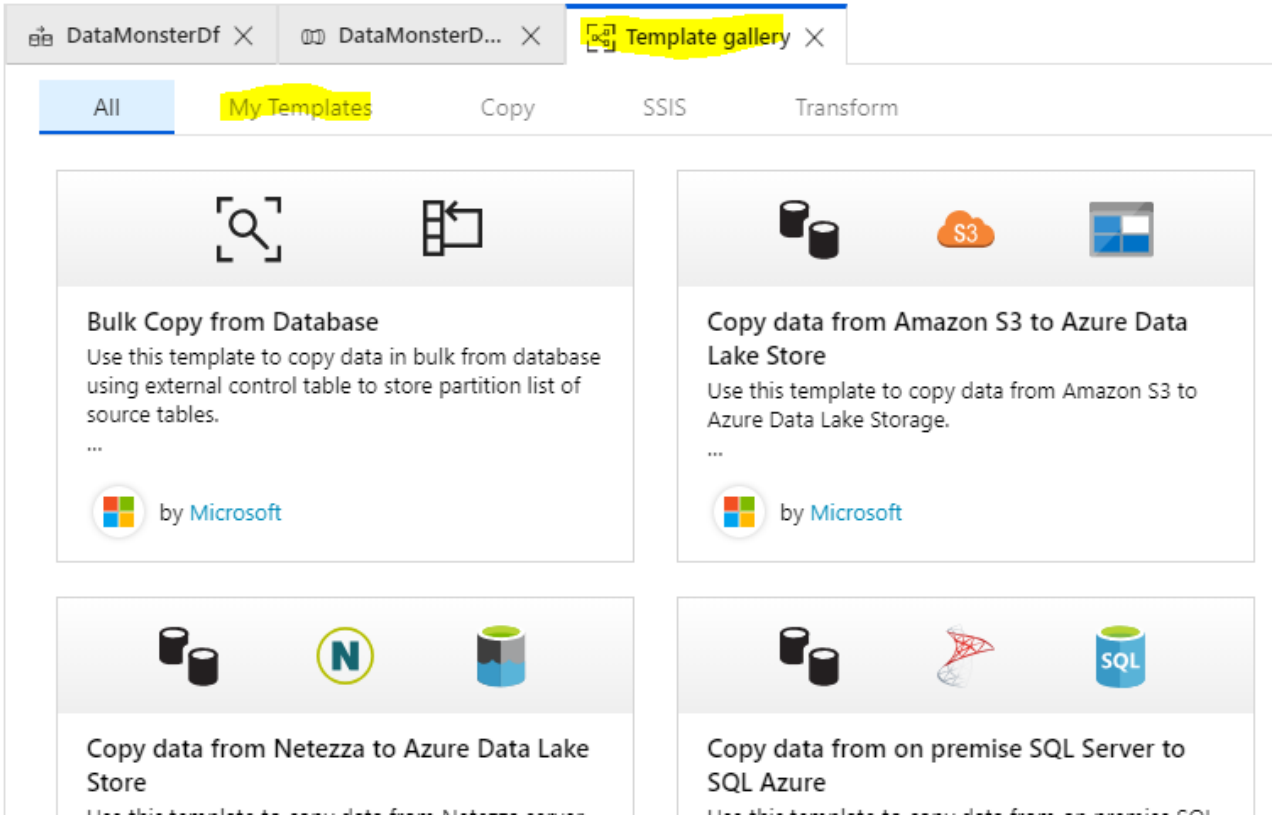
Three core capabilities:

- data movement
- pipeline activity execution
- SSIS package execution

Data Factory Essentials

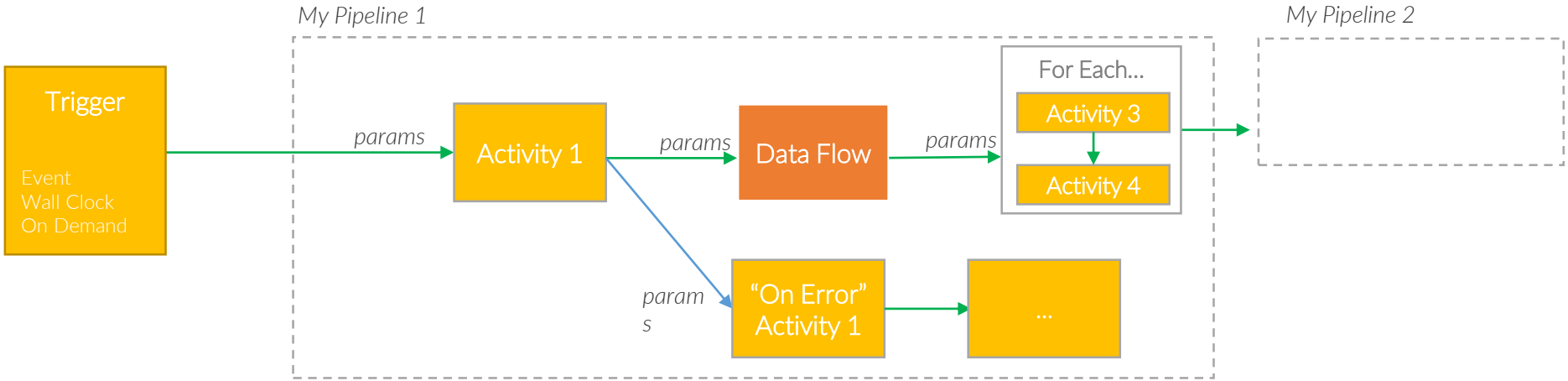


New: Kickstart using Templates



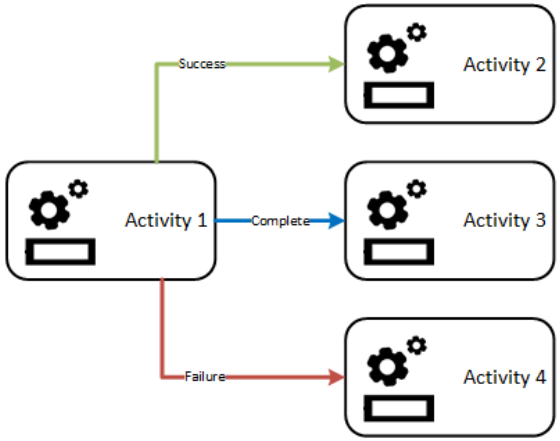
Control Flow & Triggers

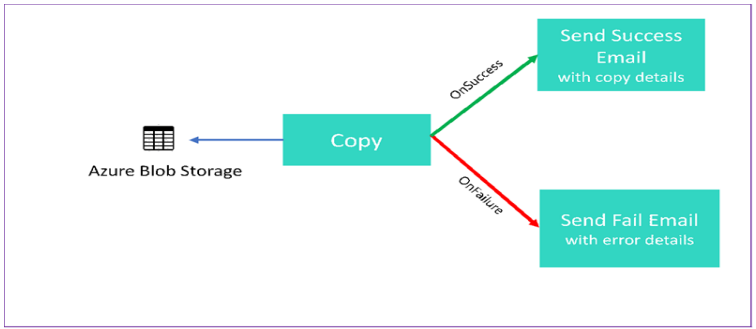
ADFv2 Pipelines



Activities

Concepts





Branching

Dependencies of activities in a pipeline

Possible constraints:

- On success
- On failure
- On completion

Also custom 'if' conditions will be available for branching based expressions

https://www.purplefrogsystems.com/paul/2017/09/whats-new-in-azure-data-factory-version-2-adfv2/

Triggers

How do pipelines get started

- on-demand
- Wall-clock Schedule
- Tumbling Window (aka time-slices in v1)
- Event on Blob Store

Demo Control Flow

Get all data of a system by metadata

Pipeline check metadata

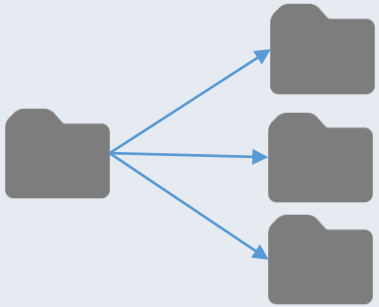
Activity: For each table in source

exec

exec

exec

Activity: Copy Data to Data Lake / Blob Store

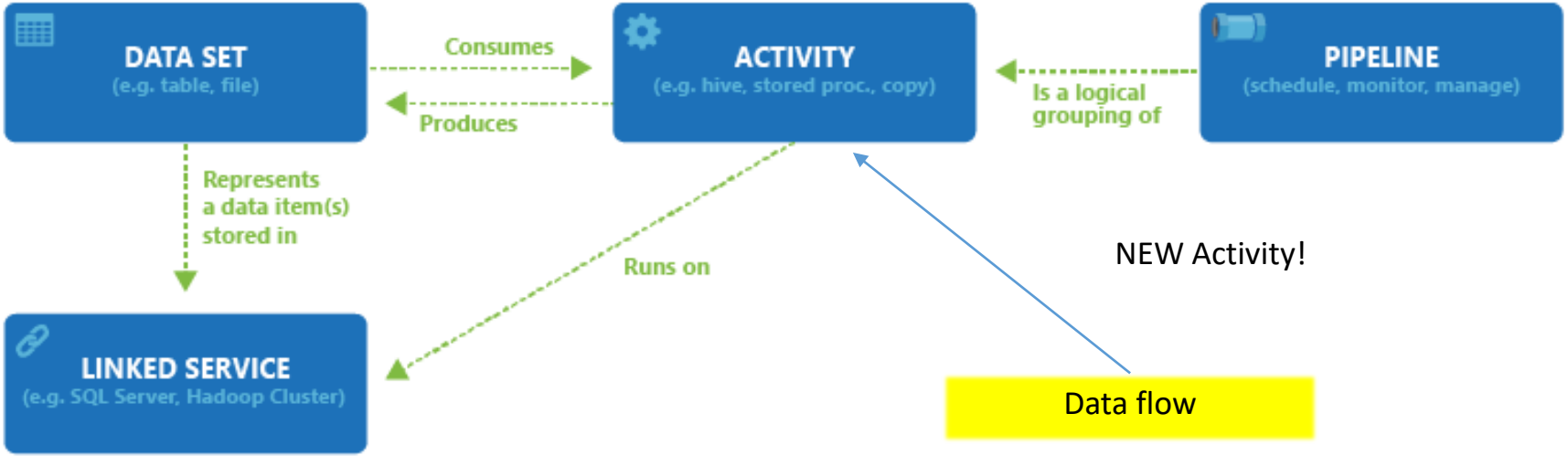


Data Flow

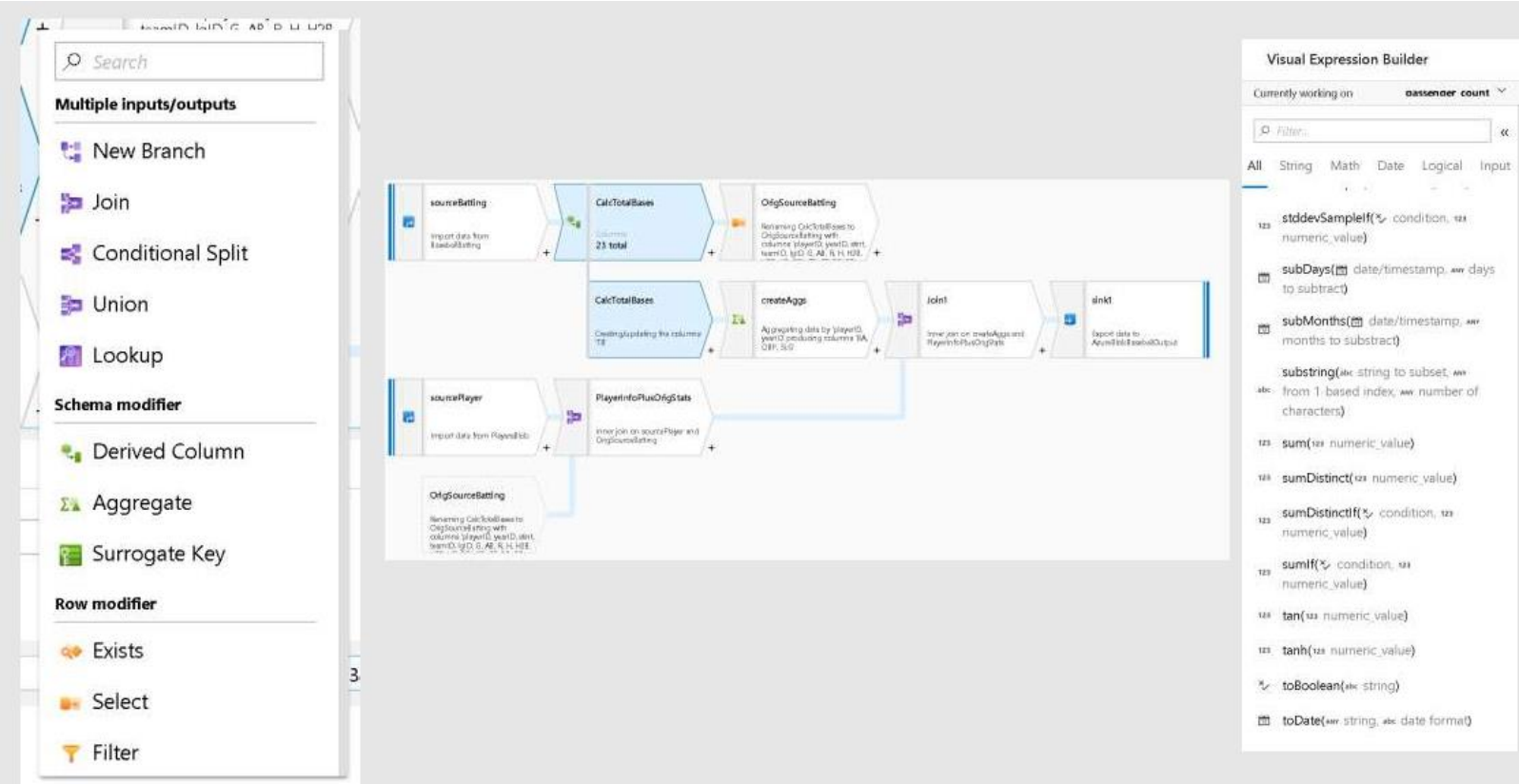
Big data transformations without coding

Data Factory Essentials

Artefacts in Data Factory

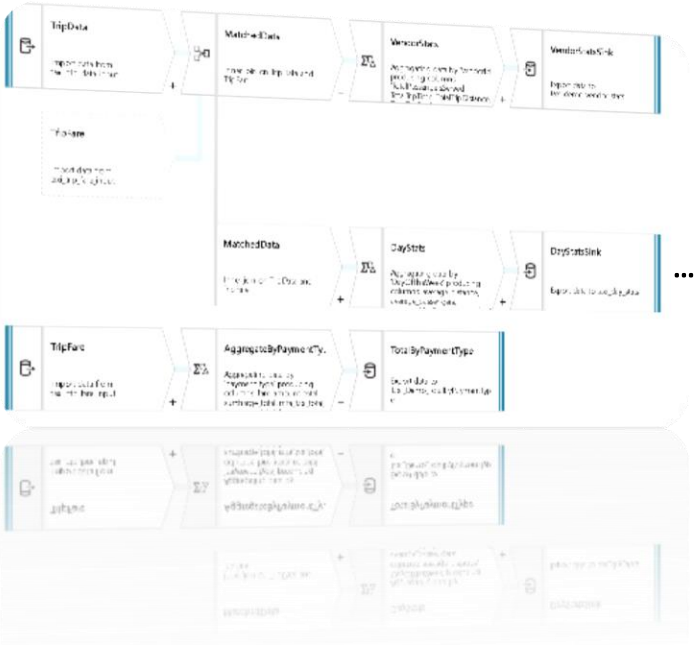


Visual / Mapping Data Flows



Code-free Data Transformation At Scale

- Does not require understanding of Spark, Big Data Execution Engines, Clusters, Scala, Python …
- Focus on building business logic and data transformation
  - Data cleansing
  - Aggregation
  - Data conversions
  - Data prep
  - Data exploration



… not …

Demo

Hey @KirnerKa, I‘m feelin so lonely
at the MS Data Platform

Lets invite all monster friends from

DUNGEONS & DRAGONS

Using

DATA FLOWS

AZURE DATA FACTORY

- http://en.wikipedia.org/wiki/Image:Newdndlogo.jpg https://www.flickr.com/photos/65092514@N08/16702792783 https://www.flickr.com/photos/benimoto/2774962417

Source Data ~1300 monster colleagues

Name	CriticallyRate	Size	TypeSubtype	Hi
Aarakocra	<1	Medium	monstrous humanoid	1
Aballin	4	Large	ooze	3
Abeil, Queen	12	Medium	monstrous humanoid	14
Abeil, Soldier	6	Large	monstrous humanoid	6
Abeil, Vassal	2	Medium	monstrous humanoid	1
Aboleth	7	Huge	aberration(aquatic)	8
Aboleth Mage, 10th-level Wizard	17	Huge	aberration(aquatic)	18
Abrian	1	Medium	magical beast(extraplanar)	2
Abyssal Ant Swarm	16	Medium	aberration(extraplanar, swarm)	20
Achaierai	5	Large	outsider(evil, extraplanar, lawful)	6
Ahuizotl	6	Large	aberration(aquatic)	7
Alaghi	4	Medium	monstrous humanoid	9
Allip	3	Medium	undead(incorporeal)	4
Androsphinx	9	Large	magical beast	12
Angel Of Decay	15	Large	undead	26
Angel, Astral Deva	14	Medium	outsider(angel, extraplanar, good)	12
Angel, Monadic Deva	12	Medium	outsider(extraplanar, good)	10

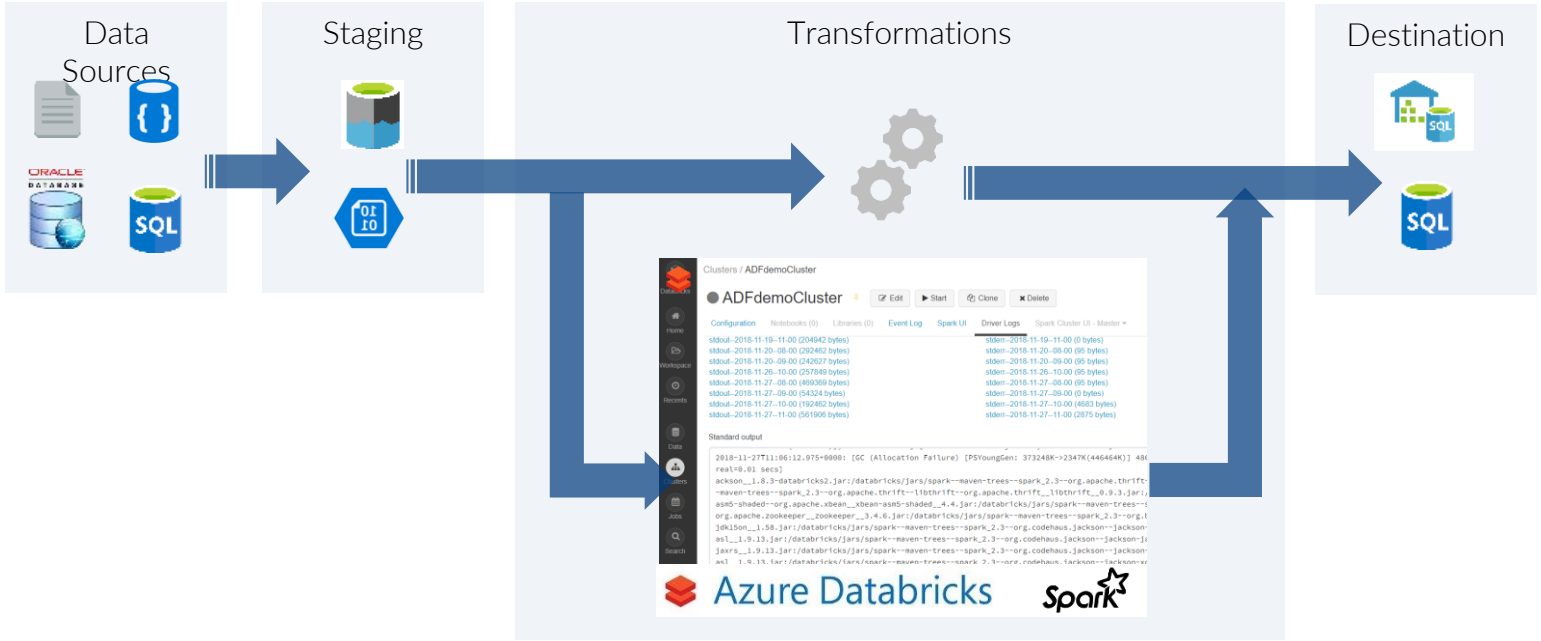
DEMO

AZURE DATA FACTORY
DATA FLOWS

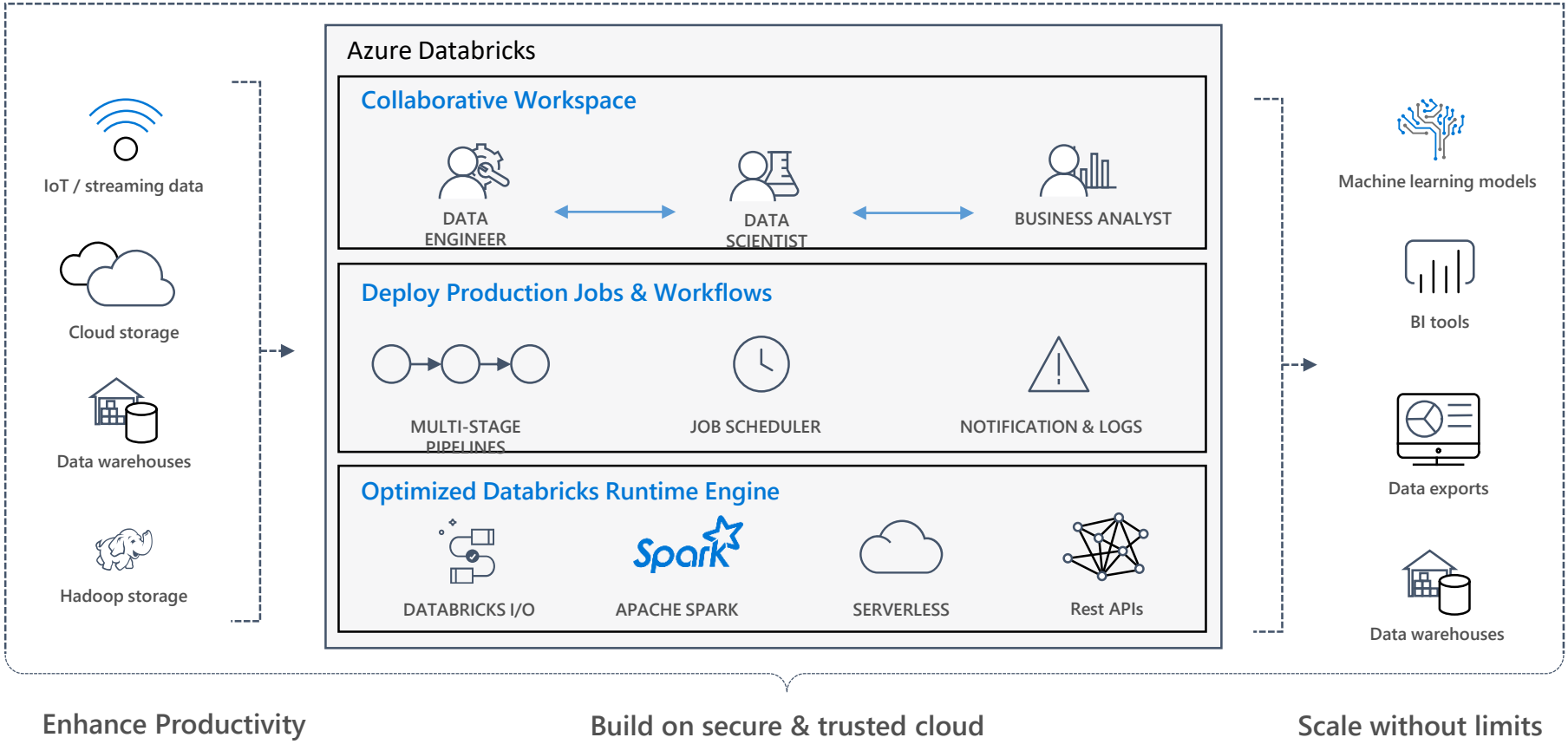
Visual Data Flow Key Tenets

- Visual “Data Flow Builder” / “Data Mapping”
- Extensible through scripting and expressions
- Data Flow can be embedded into ISV / SaaS apps
  - Embed UI
  - Embed Parameterize Data Flows
- A graphical UI for building data transformation routines on Spark
- Built for resiliency and operationalized environments

ADF Data Flow Overview



Overview Azure Databricks

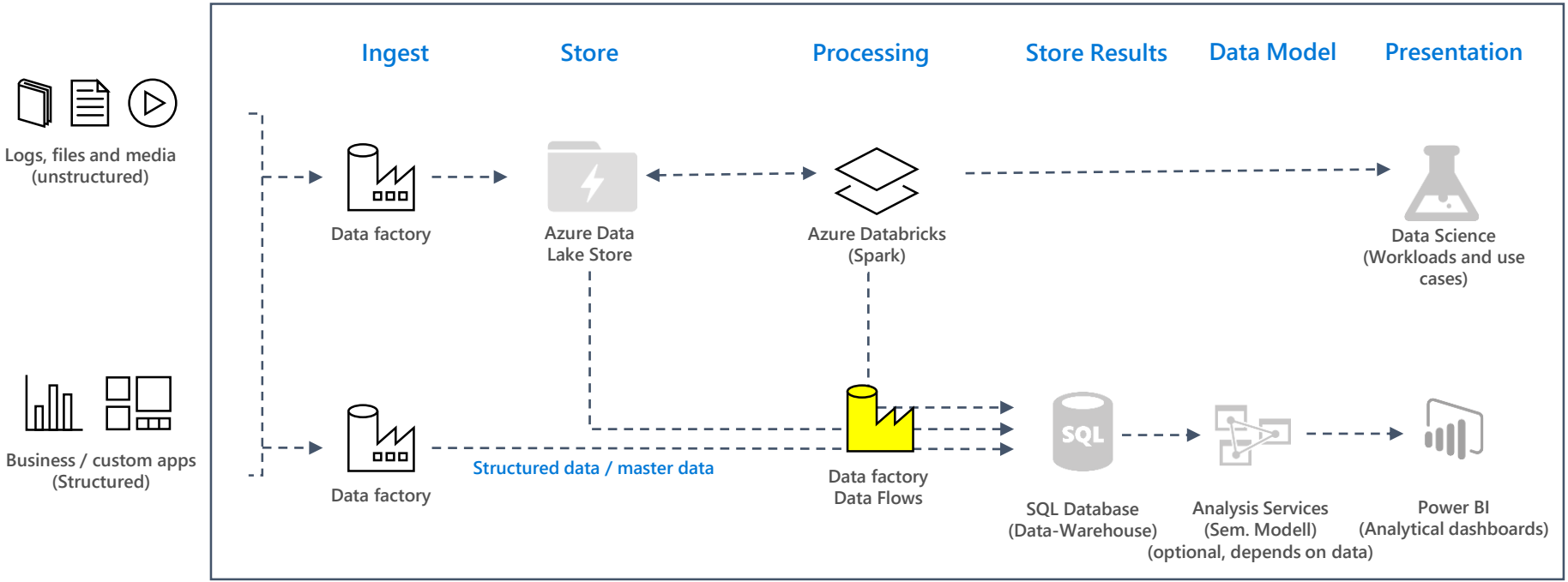


Enhance Productivity

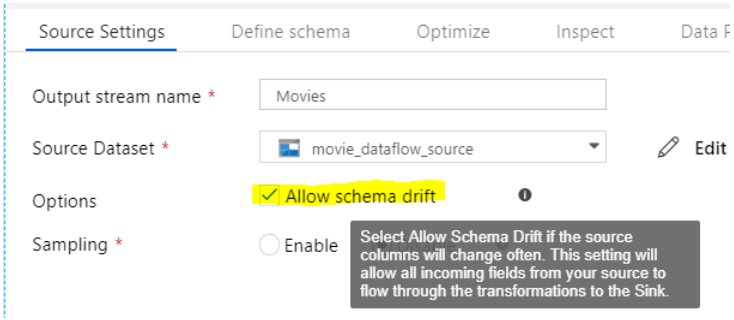
Build on secure & trusted cloud

Scale without limits

In context of Microsofts “Modern DWH”

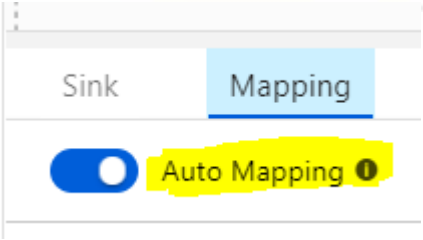


Nice Feature: Handling Schema Changes



Data flow will accept both columns (here in derived columns)

Use Auto Mapping
in Sink



- https://www.youtube.com/watch?v=vSTn_aGq3C8 Video zum Feature von Mark Kromer

Pricing

- You pay for the data flow cluster execution and debugging time per vCore-hour
- minimum cluster size to run a data flow is 8 vCores
- Execution and debugging charges are prorated **by the minute** and rounded up
- preview discount until GA
- **€0,092** per vCore-hour (General Purpose)
- With 8 Cores: **€ 0,73 per hour** | € 17,66 per day | ~€ 530 per month

  max. costs when running 24x7

- https://azure.microsoft.com/en-us/pricing/details/data-factory/data-pipeline/ am 3.4.2019

Conclusion: Data Flow in ADF

Is cool because…

- **visual design** for fast learning & understanding
- ETL using spark technology in the background which **could** scale (but does not have to in any case)
- Azure Databricks as **elastic processing engine** for different workloads, tools and user groups
- Integration in Control Flow enables modelling **dependencies** and **cost-efficient** orchestration of Azure resources

Data Flow Limited Preview Support & SLAs

- Azure SLAs are NA for preview services (private or public preview) until GA of the service.
- Limited Preview Support
  - Handled directly with the Azure Engineering team via adfdataflowext@microsoft.com. Turn-around time on fixing issues during private preview will depend upon access to customer data sources and customer Databricks clusters for RCA and debugging.
- Public Preview Support
  - Normal Azure customer service channels

Q+A

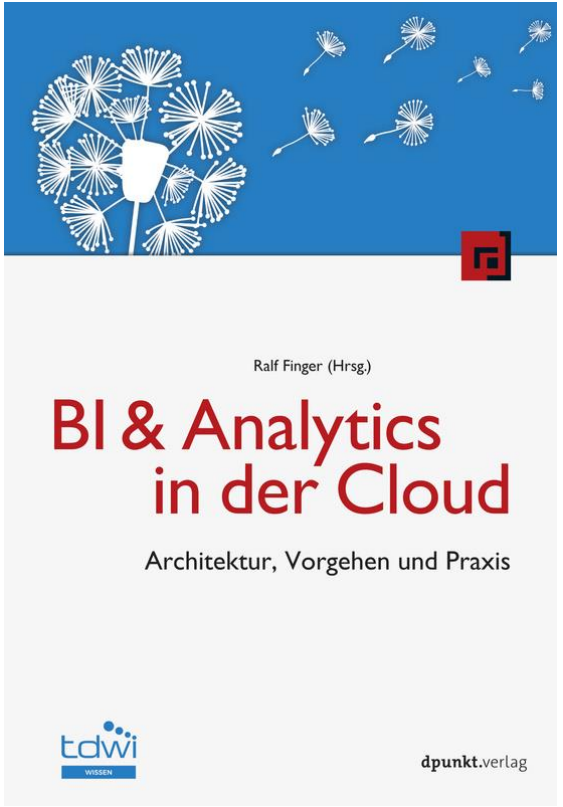
Links and further informations

1. Microsoft documentation:
https://docs.microsoft.com/en-us/azure/data-factory/
2. Azure Data Factory – data flows preview documentation
https://github.com/kromerm/adfdataflowdocs
3. Cool screencasts about data flows
https://github.com/kromerm/adfdataflowdocs/tree/master/videos
4. Another good blogpost about ADF Data Flows
https://visualbi.com/blogs/microsoft/azure/azure-data-factory-data-flow-activity/
5. Comparison ADF Data Flows vs. SSIS vs. T-SQL
https://sqlplayer.net/2018/12/azure-data-factory-v2-and-its-available-components-in-data-flows/

Buchempfehlung: BI und Analytics in der Cloud

Architektur, Vorgehen und Praxis

- Neu-Erscheinung im Oktober 2018, 262 Seiten
- Sammlung von Autorenbeiträgen zum Thema
- Kapitel: „Mehrwerte von Cloud Services in hybriden Data-Warehouses“ von Stefan Kirner



https://www.dpunkt.de/buecher/13158/9783864905919-bi-%26-analytics-in-der-cloud.html

inovex

inovex ist ein IT-Projekthaus
mit dem Schwerpunkt „Digitale Transformation“:

Product Ownership · Datenprodukte
Web · Apps · Smart Devices · BI
Big Data · Data Science · Search
Replatforming · Cloud · DevOps
Data Center Automation & Hosting
Trainings · Coachings

inovex gibt es in Karlsruhe · Pforzheim · Stuttgart · München · Köln · Hamburg

Und natürlich unter www.inovex.de

Wir nutzen Technologien,
um unsere Kunden glücklich zu machen.
Und uns selbst.