Azure Data Factory: Mapping Data Flow – first blood

PASS SQLSATURDAY MADRID | 28 SEP 2019

#SQLSatMadrid

Kamil Nowinski

Principal Microsoft Consultant

altius

MVP Microsoft® Most Valuable Professional

@NowinskiK

BIG Thanks to SQLSatMadrid sponsors

Platinum

plain concepts

KABEL

aura

channeladvisor

Gold

SolidQ

encamina
PIENSA EN COLORES

東京
TOKIOTA

alesonITC
data thinking

zaltor
Software Distribution

Suricatta

Silver

devscope

SentryOne
SQL Player
Play with data & have fun!

Azu a Facto

Venue

Universidad Politécnica de Madrid
ETSI SISTEMAS INFORMÁTICOS

Global

Microsoft

altius

PASS
SQLSATURDAY
ORGANIZER

@NowinskiK

Kamil Nowiński

Microsoft Data Platform **MVP**

Speaker, blogger, data enthusiast

Principal Microsoft Consultant at Altius (www.altiusdata.com)

15+ yrs experience as DEV/BI/(DBA)

Member of the Data Community PL

Project member of „SCD Merge Wizard”

Founder of blog SQLPlayer (www.SQLplayer.net)

SQL Server Certificates:

MCITP, MCP, MCTS, MCSA, MCSE Data Platform,

MCSE Data Management & Analytics

Moreover: Bicycle, Running, Digital photography

@NowinskiK, @SQLPlayer

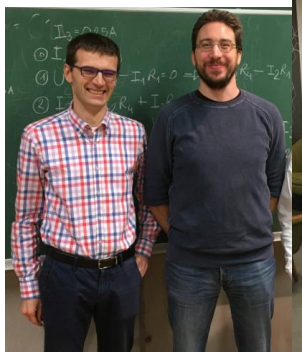
BLOG & Interviews

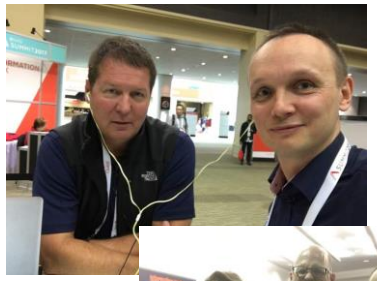
www.SQLPlayer.net

PODCAST – interviews with...

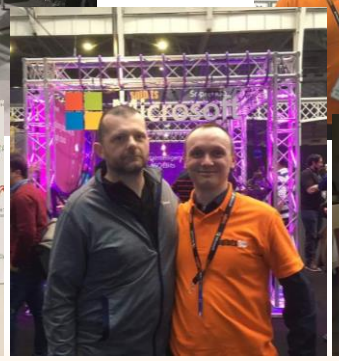


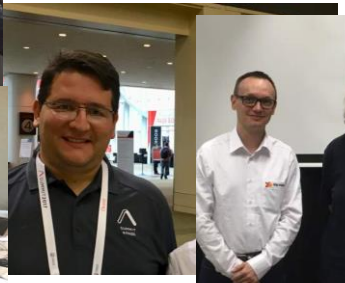




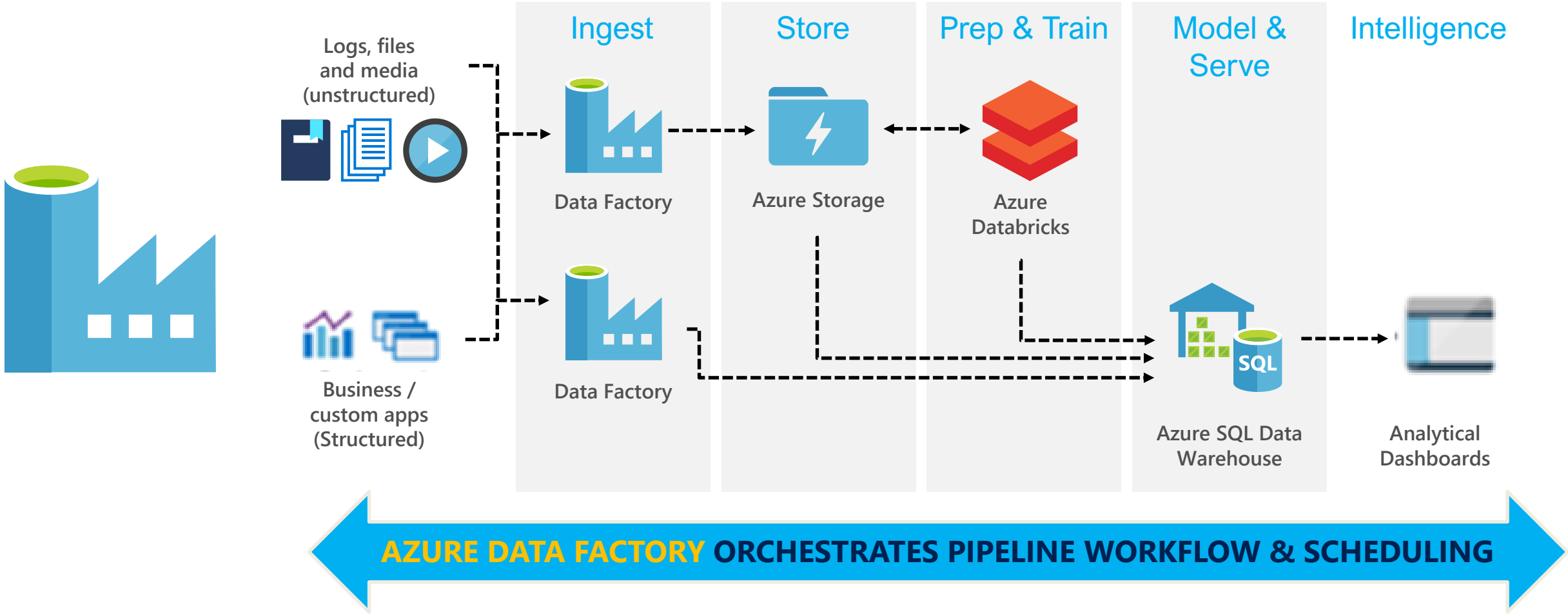







Scan me

What the Azure Data Factory is?

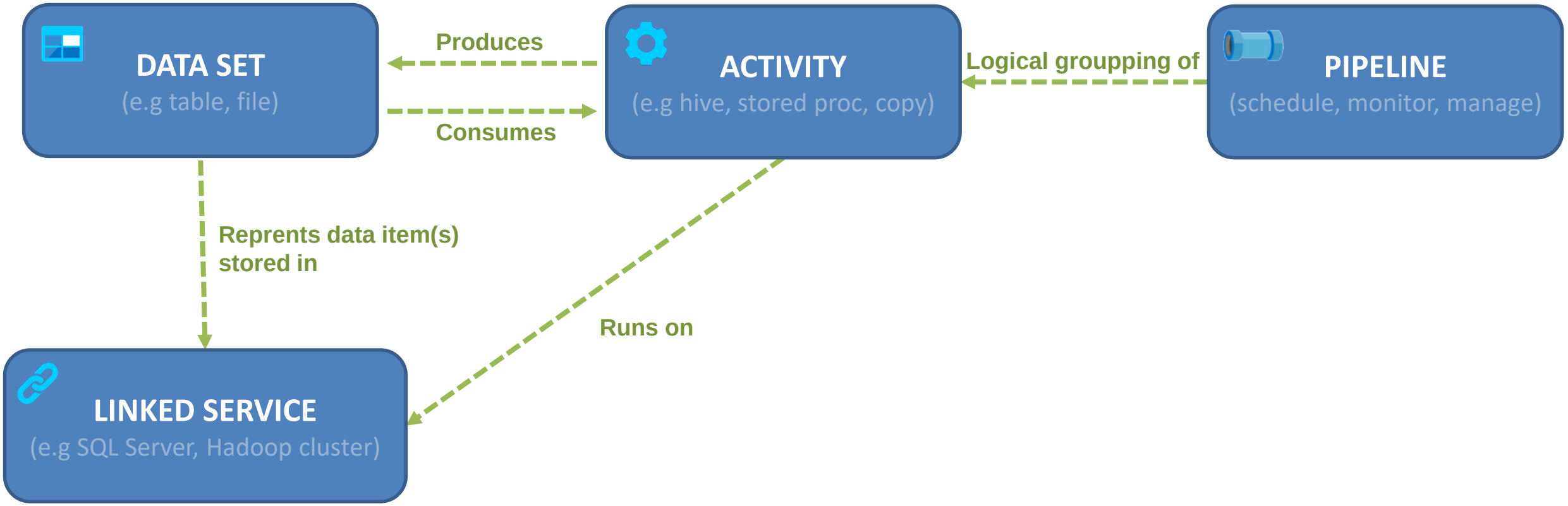


Access all your data

- 80+ connectors & growing
- Azure IR available in 21 regions
- Hybrid connectivity using self-hosted IR: on-prem & VNet

Azure (13)	Database (24)		File Storage (5)	NoSQL (3)	Services and Apps (28)		Generic (4)
Blob Storage	Amazon Redshift	Netezza	Amazon S3	Cassandra	Amazon MWS	Office 365 *	HTTP
Cosmos DB (MongoDB API) *	DB2	Oracle	File System	Couchbase	CDS for Apps	Paypal	OData
Cosmos DB (SQL API)	Drill	Phoenix	FTP	MongoDB	Concur	QuickBooks	ODBC
Data Lake Storage Gen1	Google BigQuery	PostgreSQL	HDFS		Dynamics 365	Salesforce	REST *
Data Lake Storage Gen2	Greenplum	Presto	SFTP		Dynamics CRM	Salesforce Marketing Cloud	
DB for MySQL	HBase	SAP BW			GE Historian	Salesforce Service Cloud	
DB for PostgreSQL	Hive	SAP HANA			Google AdWords	SAP C4C	
File Storage	Impala	Spark			HubSpot	SAP ECC	
Kusto *	Informix	SQL Server			Jira	ServiceNow	
Search Index	MariaDB	Sybase			Magento	Shopify	
SQL DB	Microsoft Access	Teradata			Marketo	Square	
SQL DW	MySQL	Vertica			Oracle Eloqua	Web table	
Table Storage					Oracle Responsys	Xero	
					Oracle Service Cloud	Zoho	
	Supported as Source and Sink						
	Supported as Source only						
	Supported as Sink only						

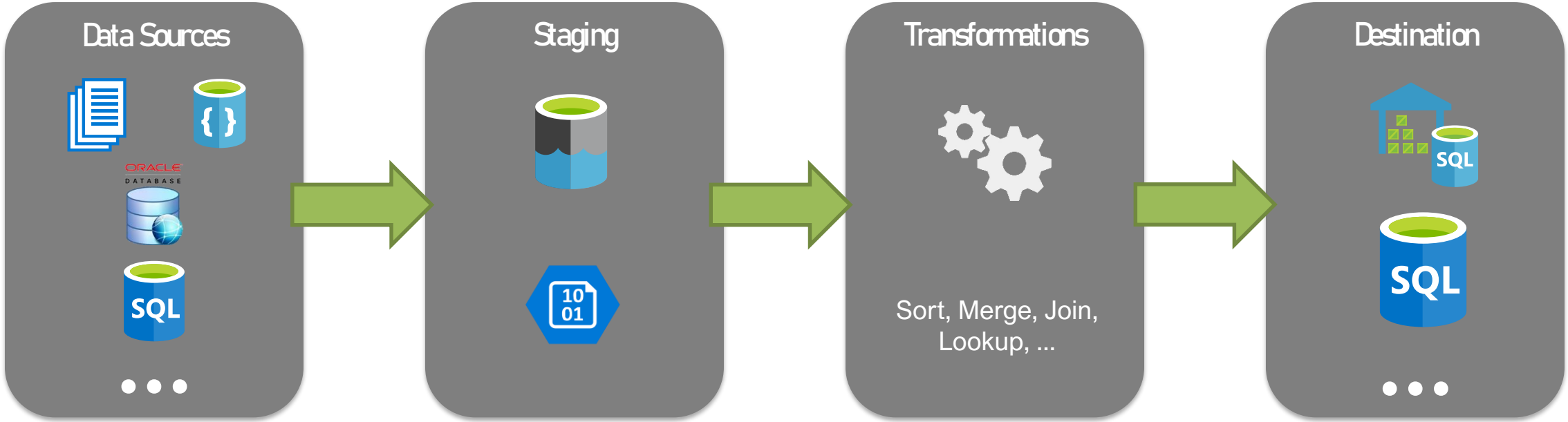
ADF Key Concepts



Visual Data Transformations with

MAPPING DATA FLOW

What the hell (Mapping) Data Flows are?



- Explicit user action
- User places data source(s) on design surface, from toolbox
- Select explicit sources

- Implicit/Explicit
- Data Lake staging area as default
- User does not need to configure this manually
- Advanced feature to set staging area options
- File formats/types: (Parquet, JSON, txt, CSV, ...)

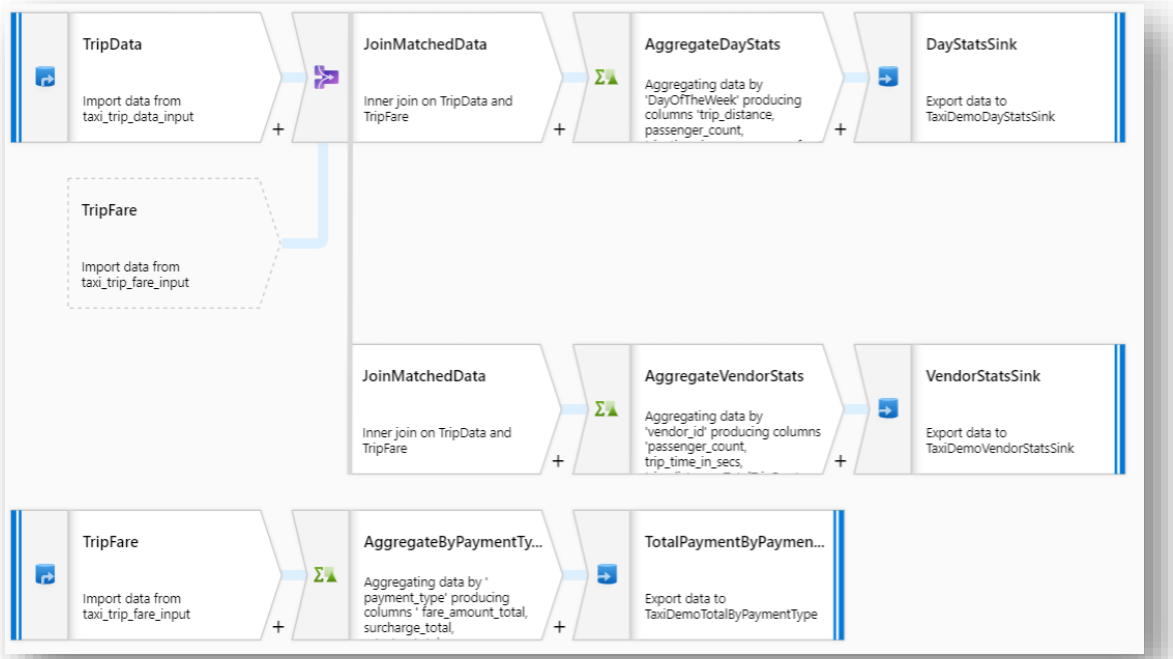
- Explicit user action
- User places transformations on design surface, from toolbox
- User must set properties for transformation steps and step connectors

- Explicit user action
- User chooses destination connector(s)
- User sets connector property options

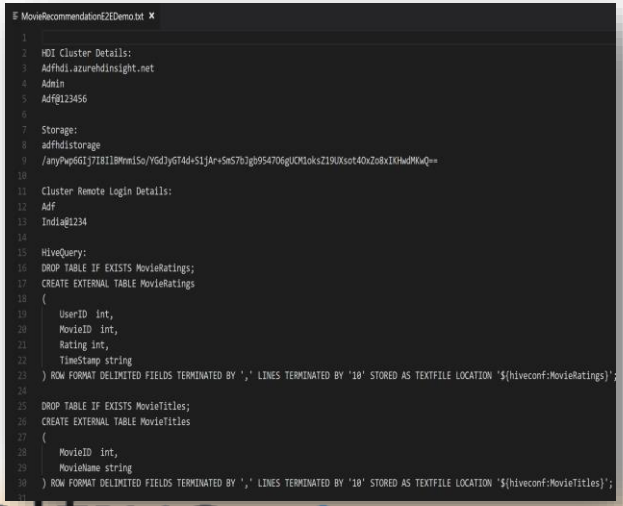
src: (Microsoft) ADF Data Flow Private Preview Overview

Code-free Data Transformation at Scale

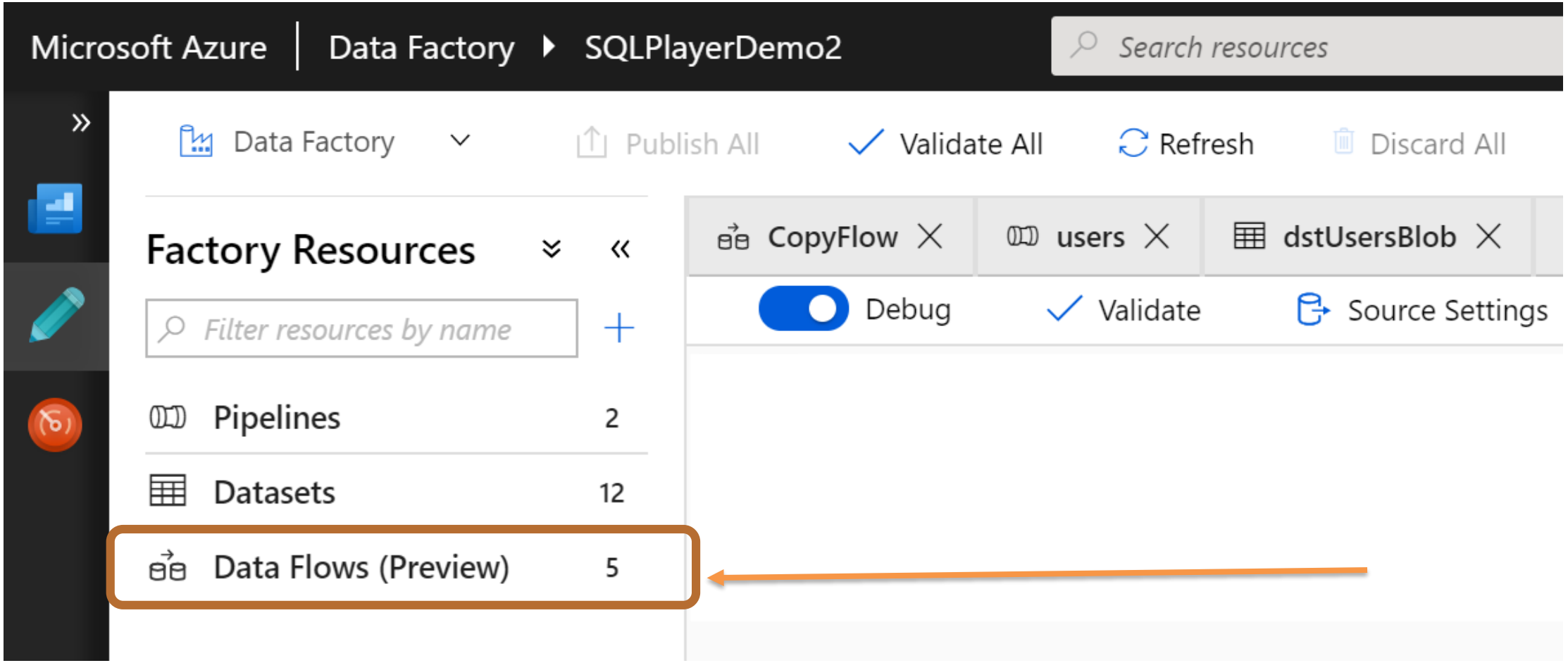
- Does not require understanding of Spark, Big Data Execution Engines, Clusters, Scala …
- Focus on building business logic and data transformation
  - Data cleansing
  - Aggregation
  - Data conversions
  - Data prep
  - Data exploration
  - ETL Data Loading into DW



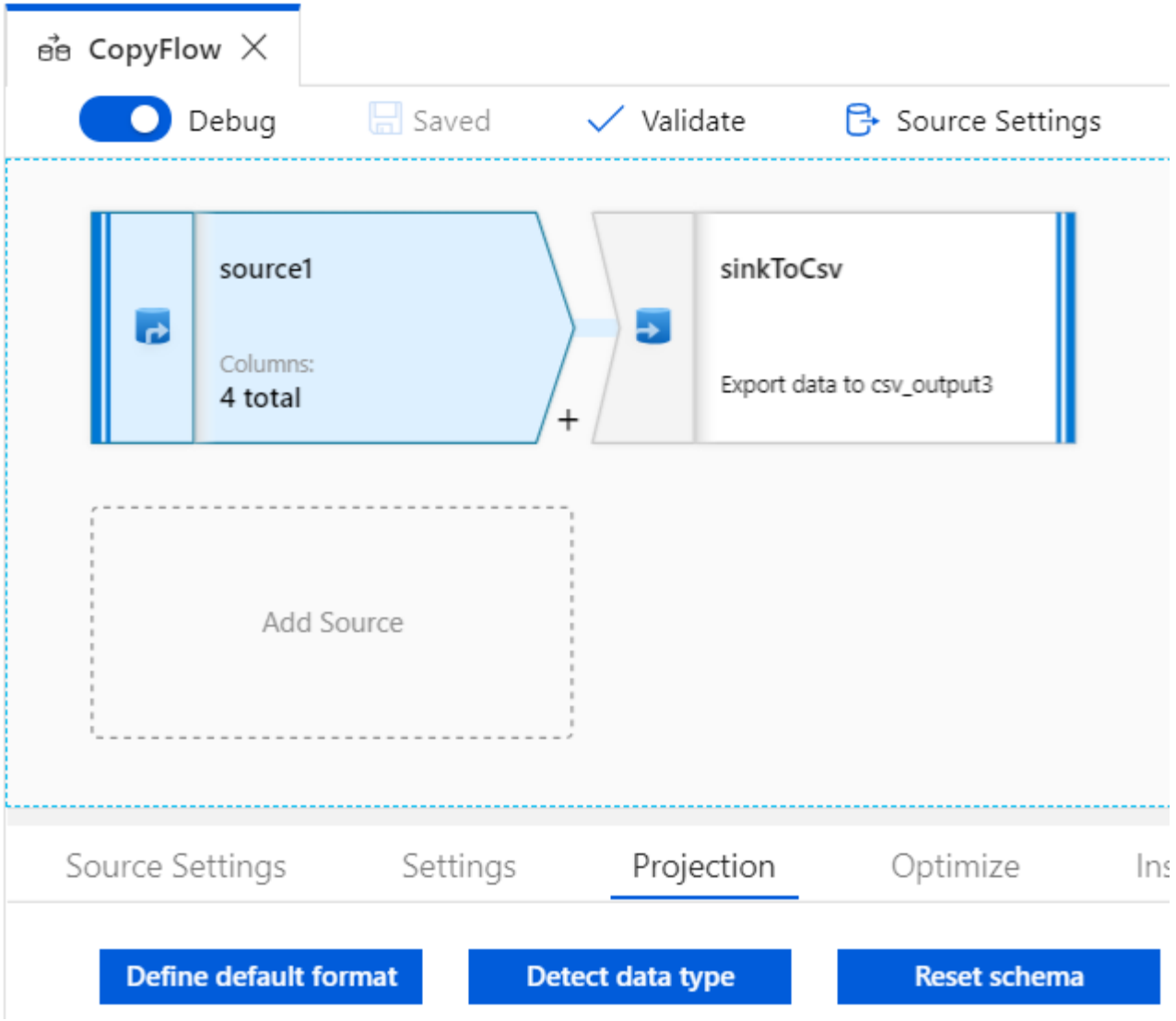
… not



Authoring of Azure Data Factory (v2) – what's new?



Simple Copy Flow



Mapping Data Flow: Components = Actions *

Multiple inputs/outputs

- New Branch
- Join
- Conditional Split
- Union
- Lookup

Schema modifier

- Derived Column
- Aggregate
- Surrogate Key
- Pivot
- Unpivot
- Window

Row modifier

- Exists
- Select
- Filter
- Sort

Custom

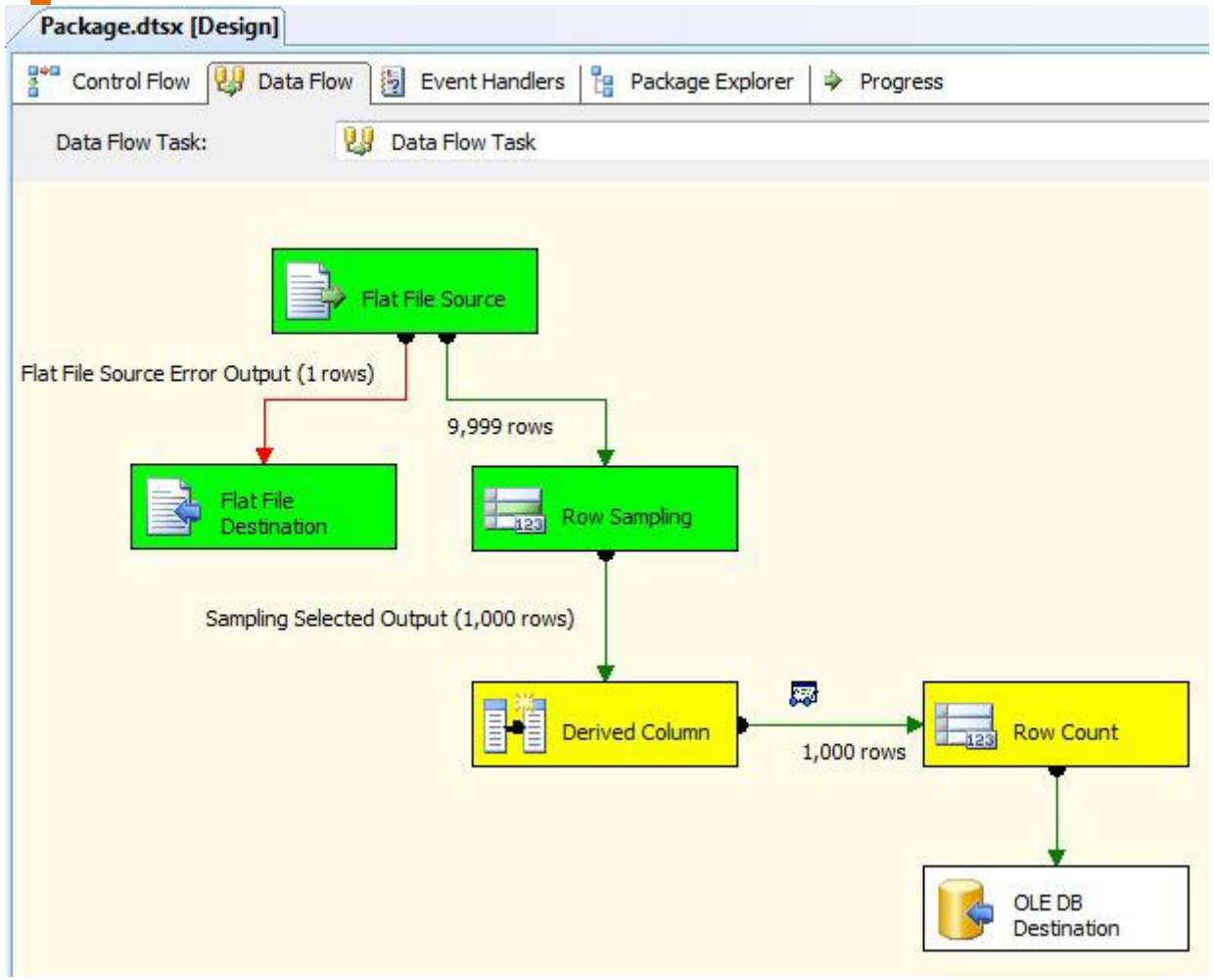
- Extend

Destination

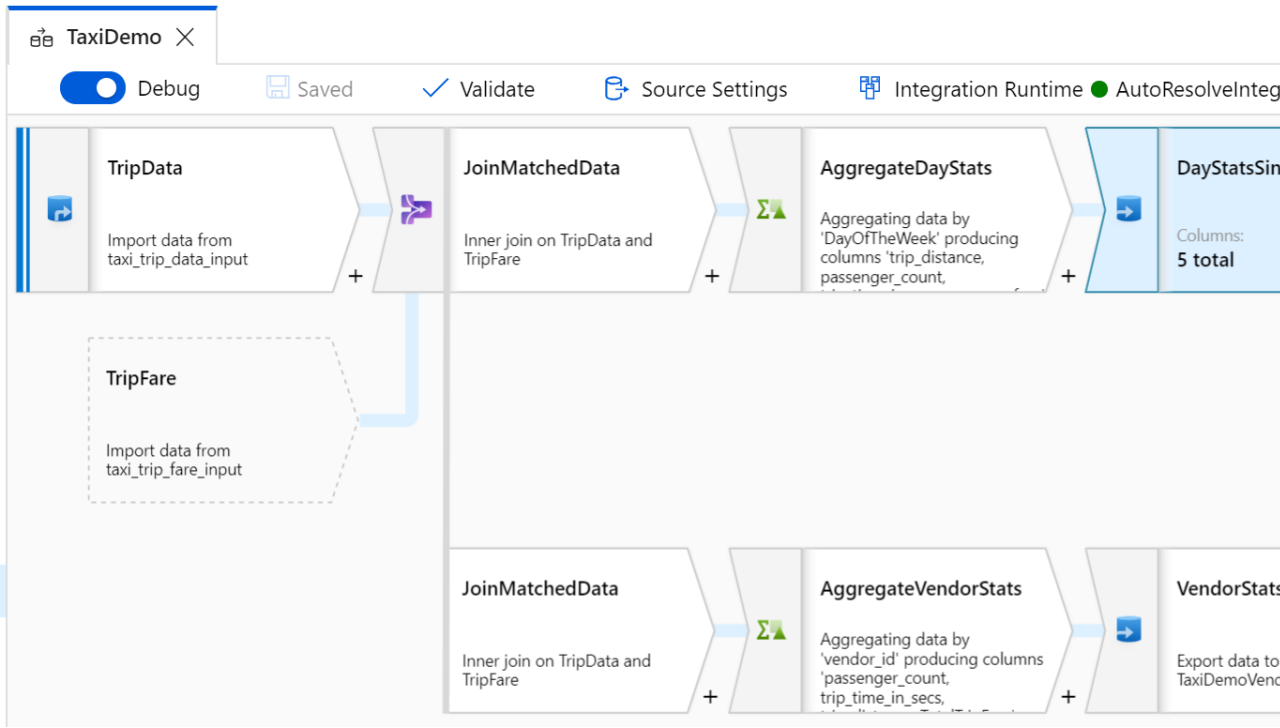
- Sink

* With some small exceptions

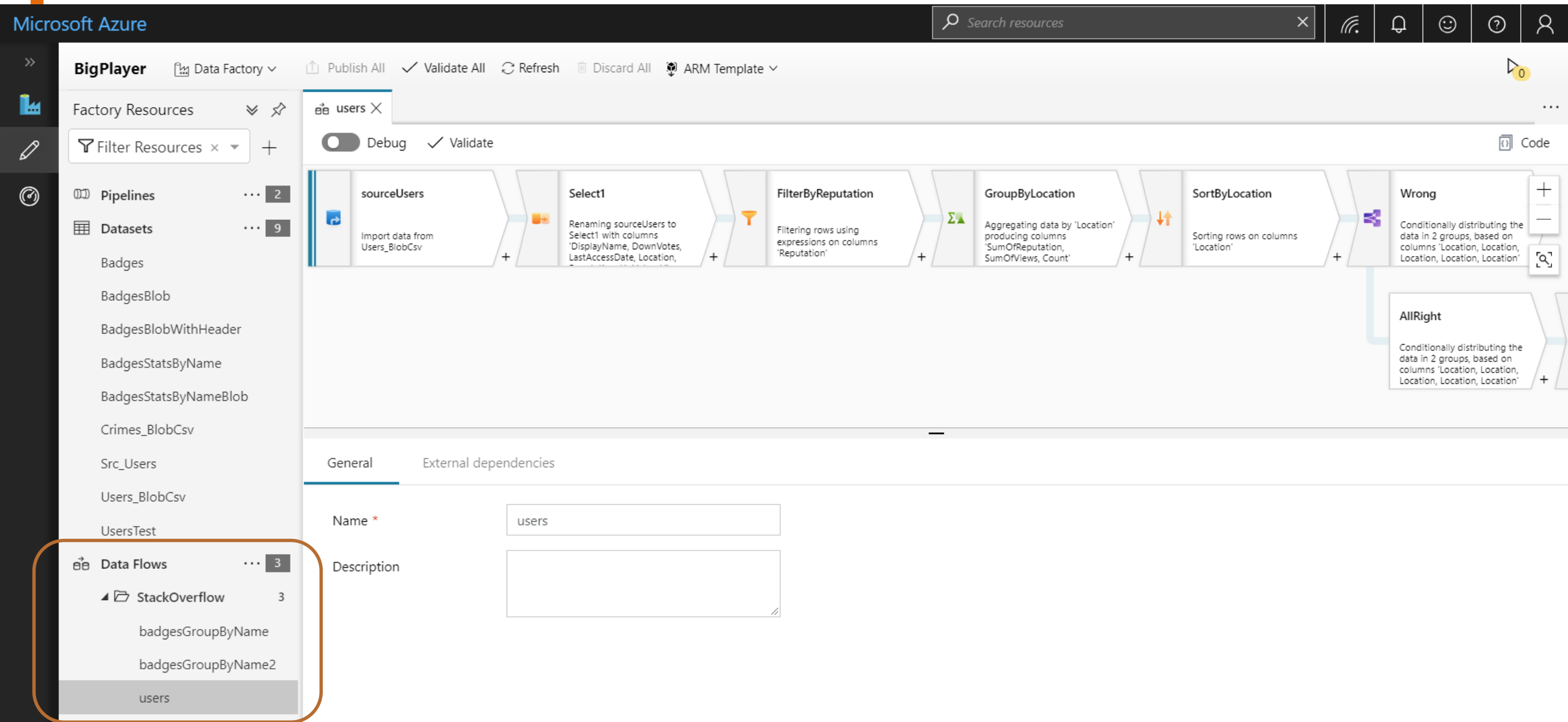
SSIS Data Flow VS ADF Mapping Data Flow



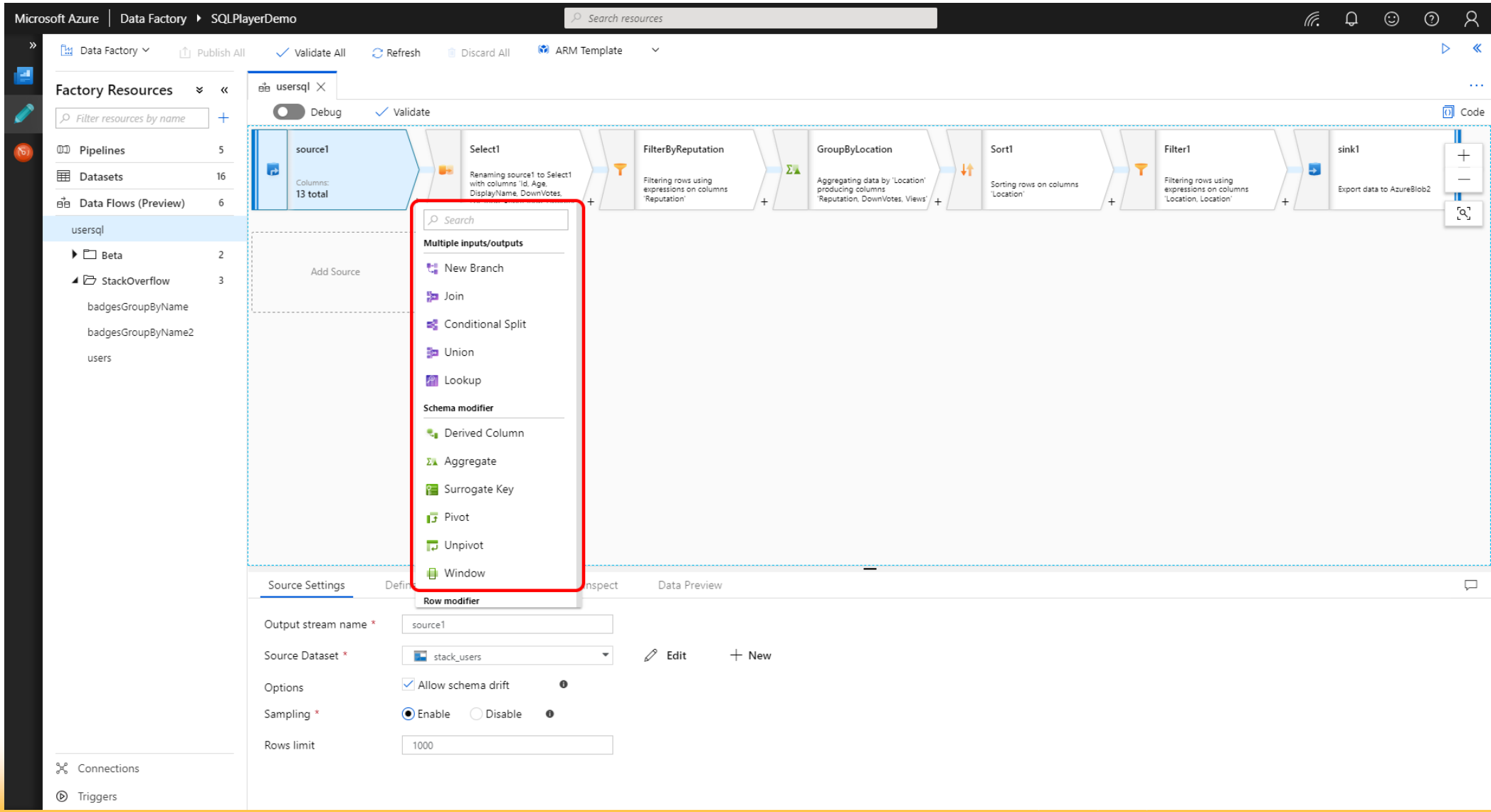
https://www.red-gate.com/simple-talk/sql/ssis/debugging-data-flow-in-sql-server-integration-services/



Authoring of Azure Data Factory (v2)

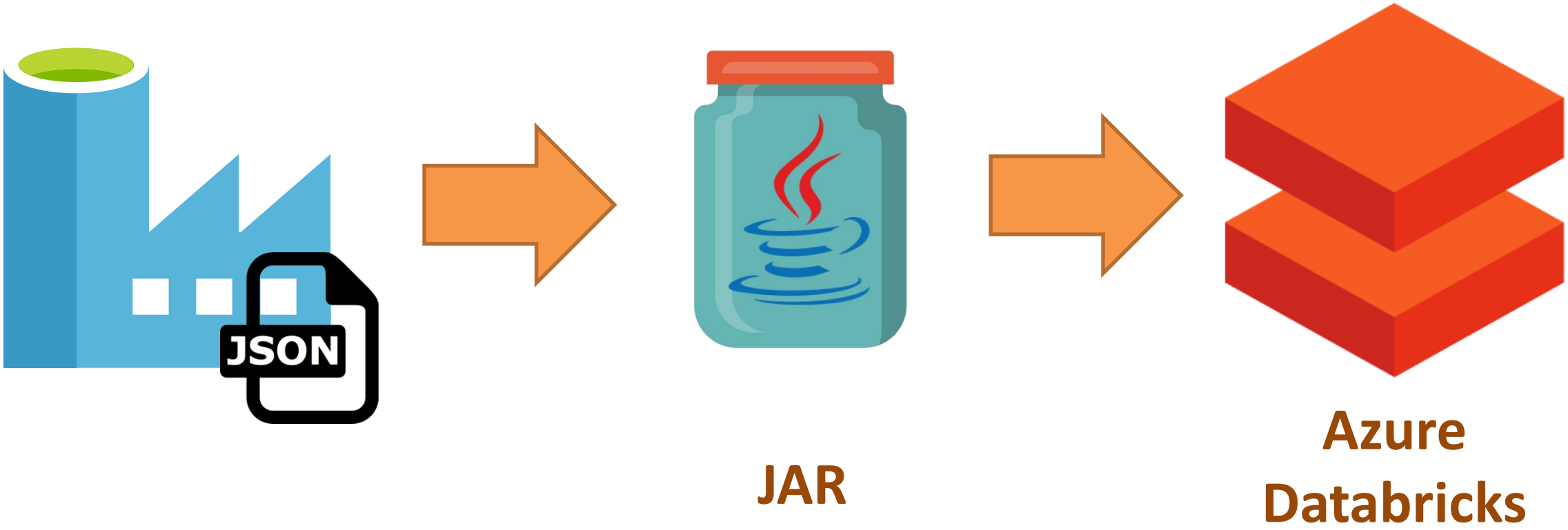


Guided experience to build data flows



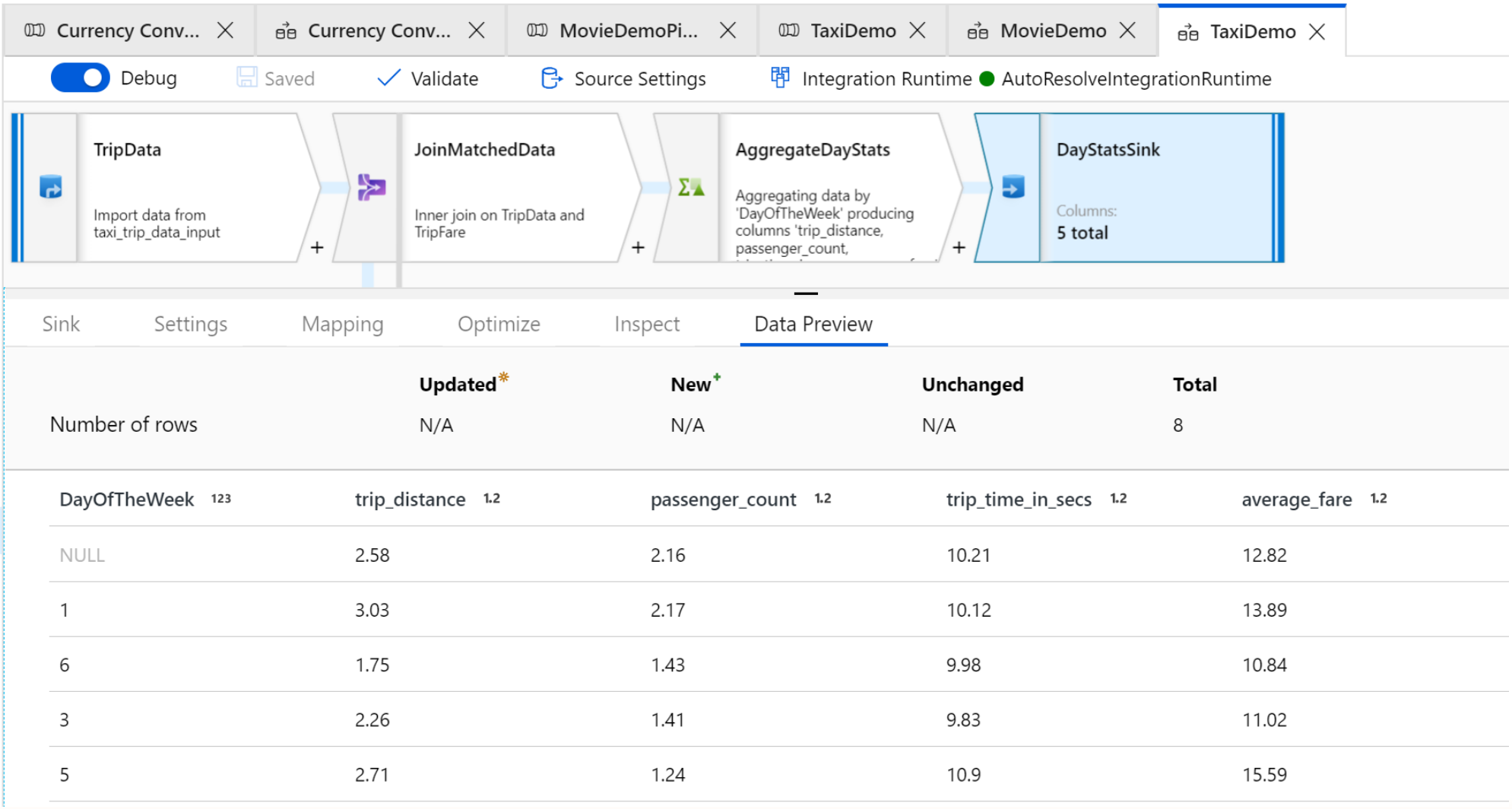
SQL Player
Play with data & have fun!

What is going on behind the scenes?



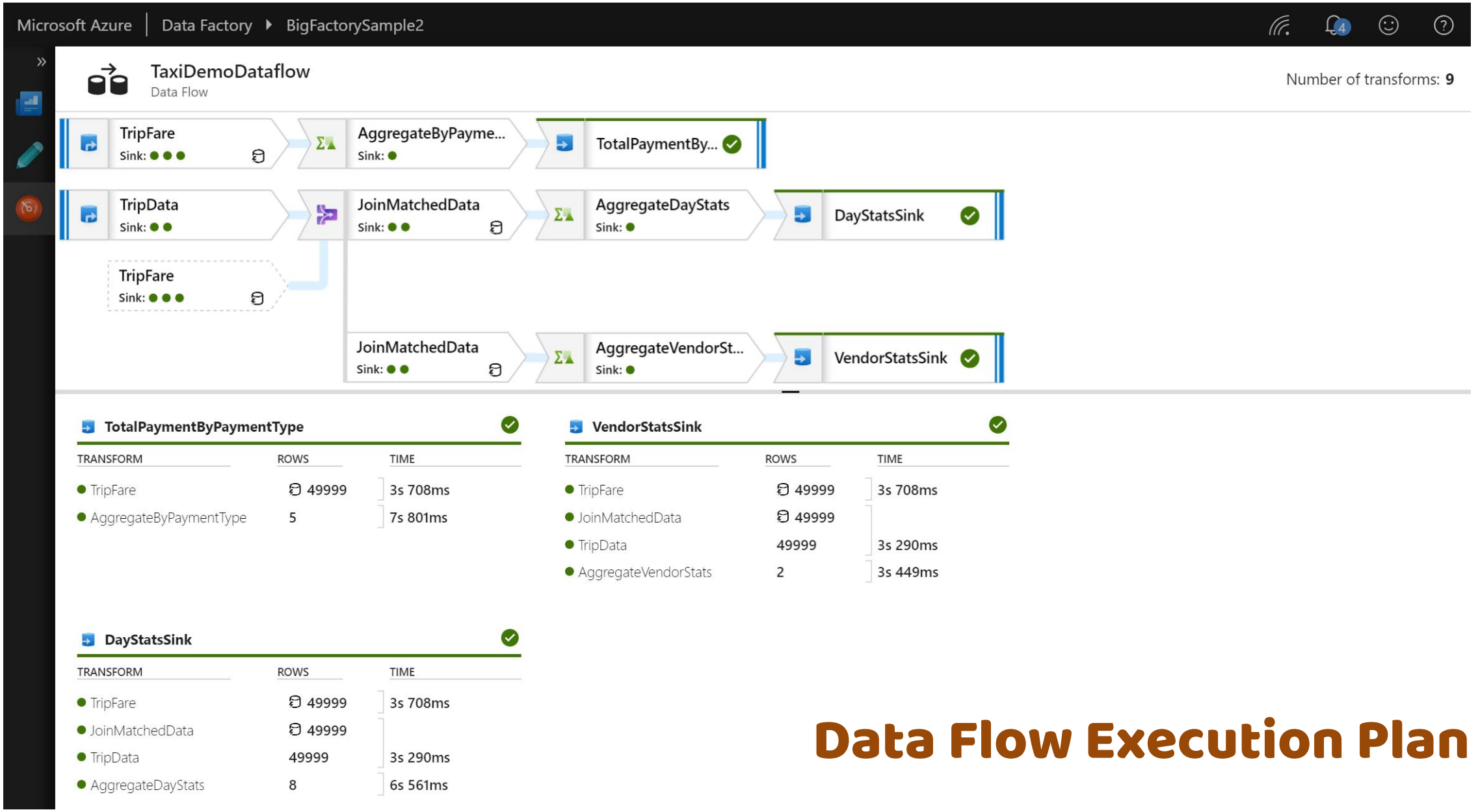
DEMO TIME

Data Preview in Debug mode

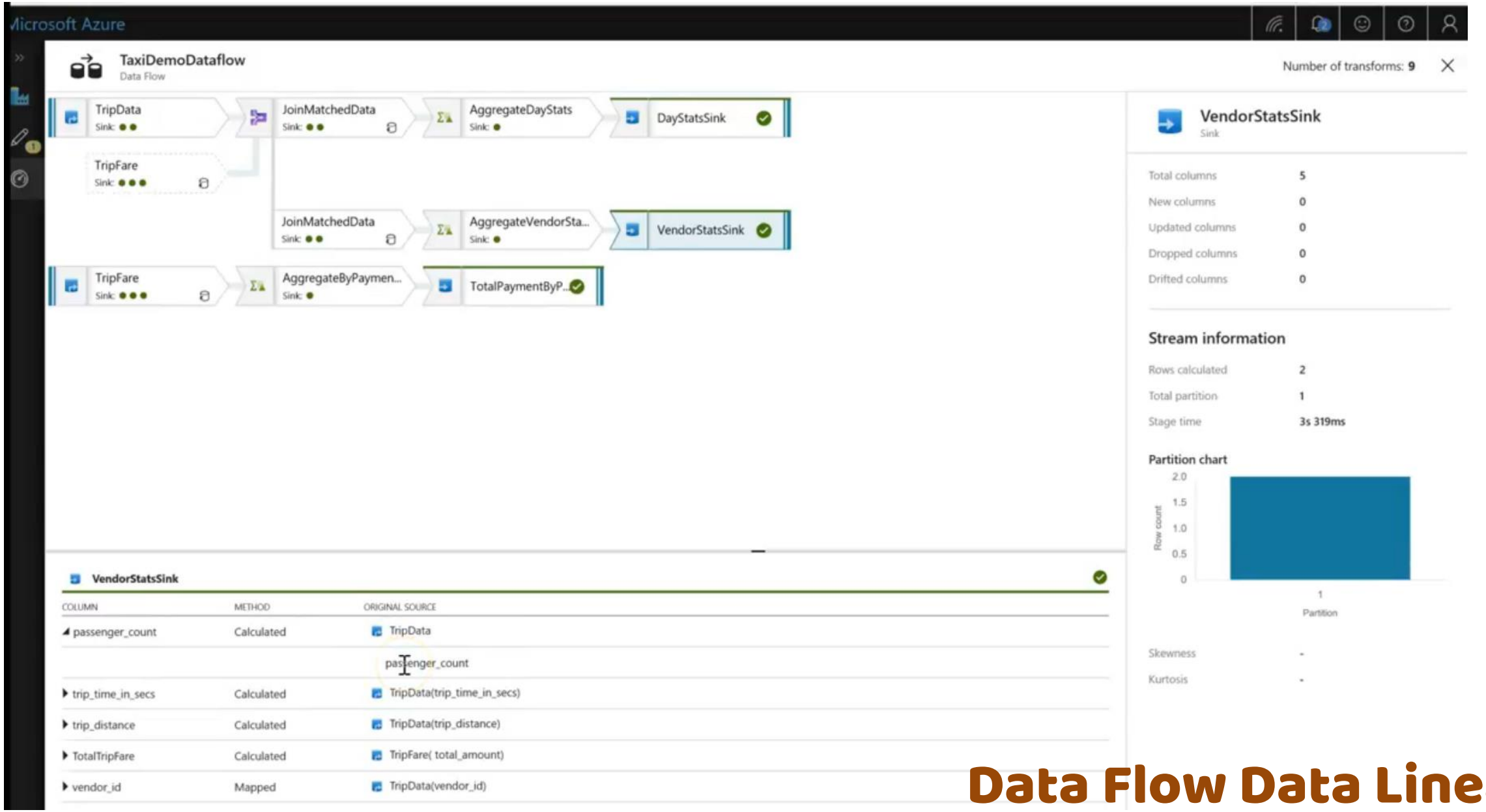


	Updated*	New+	Unchanged	Total
Number of rows	N/A	N/A	N/A	8

DayOfTheWeek 123	trip_distance 1.2	passenger_count 1.2	trip_time_in_secs 1.2	average_fare 1.2
NULL	2.58	2.16	10.21	12.82
1	3.03	2.17	10.12	13.89
6	1.75	1.43	9.98	10.84
3	2.26	1.41	9.83	11.02
5	2.71	1.24	10.9	15.59

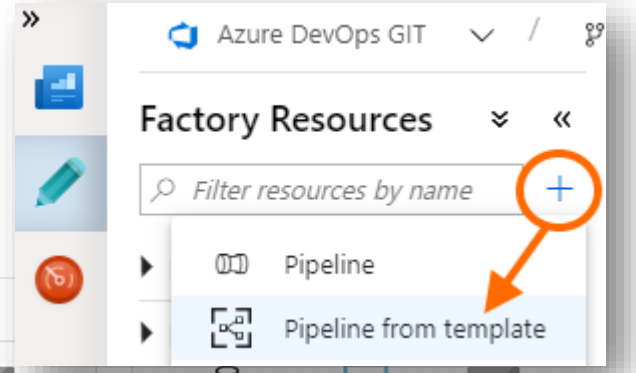


Data Flow Execution Plan



Data Flow Data Lineage

ADF Template Gallery



Template gallery

Filter Reset all filter

Search templates

Categories
- Copy
- Data Flow
- SSIS
- Transform

Create by
- Microsoft
- My templates

Tag
All

Services used
All

Import template

Bulk Copy from Database to Azure Data Explorer
Use this template to copy large amount of data in bulk from database like SQL Server, Google BigQuery, etc to Azure Data Explorer (ADX), using...
by Microsoft

Bulk Copy from Database
Use this template to copy data in bulk from database using external control table to store partition list of source tables.
...
by Microsoft

Copy data from Google BigQuery to Azure Data Lake Store
Use this template to copy data from Google BigQuery to Azure Data Lake Storage.
...
by Microsoft

Copy data from HDFS to Azure Data Lake Store
Use this template to copy data from HDFS (Hadoop Distributed File System) to Azure Data Lake Storage.
...
by Microsoft

Copy data from Netezza to Azure Data Lake Store
Use this template to copy data from Netezza server to Azure Data Lake Storage.
...
by Microsoft

Copy data from on premise SQL Server to SQL Azure
Use this template to copy data from on premise SQL Server to SQL Azure.
...
by Microsoft

Copy data from on premise SQL Server to SQL Data Warehouse
Use this template to copy data from on premise SQL Server to SQL Data Warehouse.
...
by Microsoft

Copy data from Oracle to SQL Data Warehouse
Use this template to copy data from Oracle server to SQL Data Warehouse.
...
by Microsoft

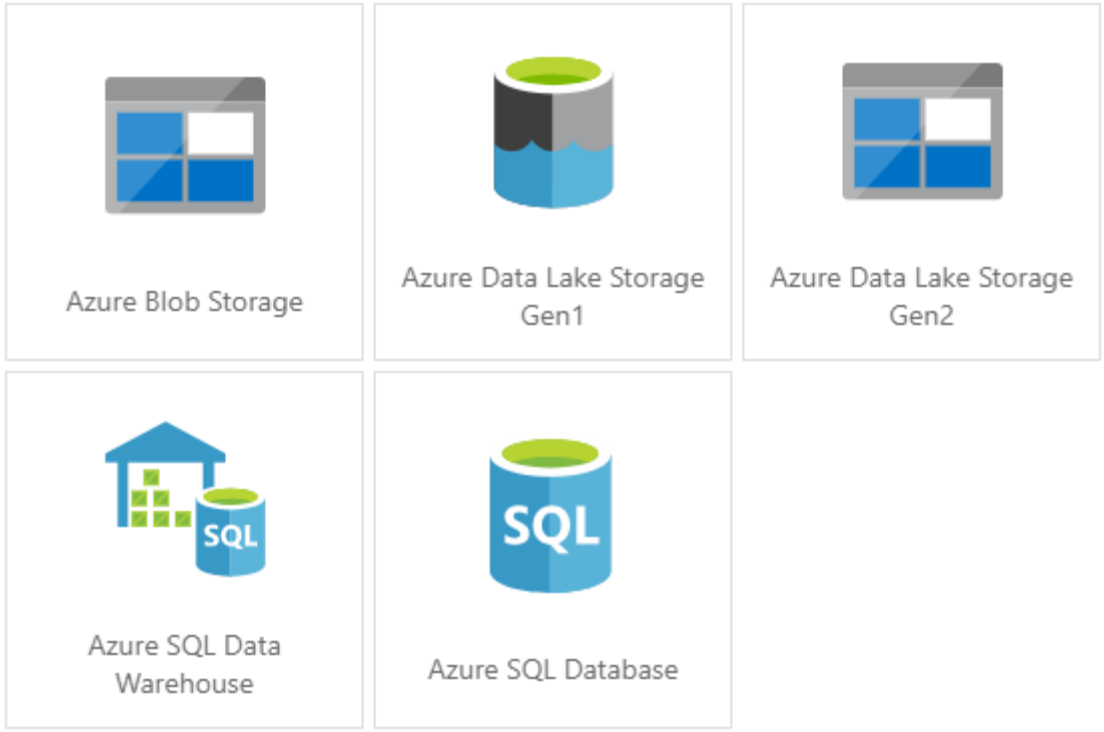
Copy delta data from AWS S3 to Azure Data Lake Storage Gen2

Copy multiple files containers between File Stores

Copy new files only by LastModifiedDate
Use this template to copy new and changed files only by using LastModifiedDate.

Data Flow Search Log Analytics
This is a sample that takes the U-SQL SearchLog analytics example and turns it into an ADF Data Flow:

Mapping Data Flow – Source & Sink

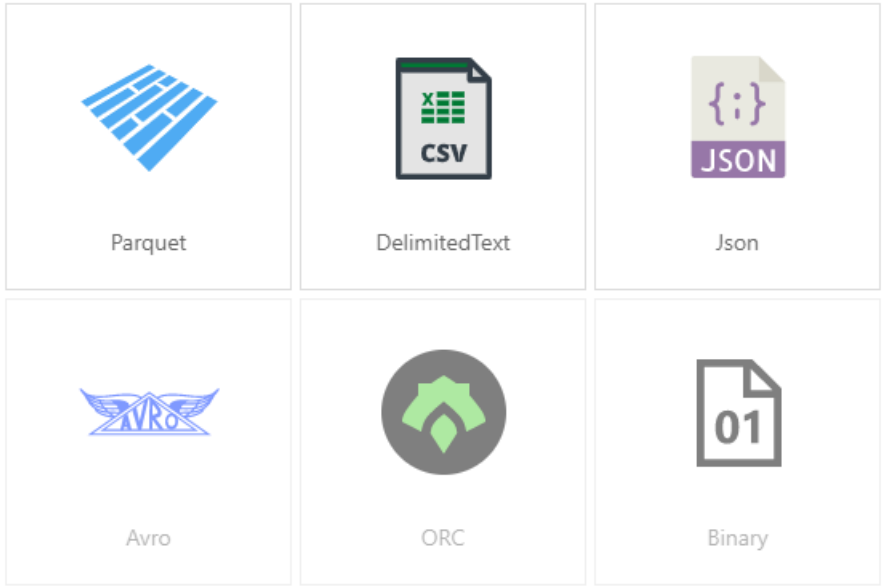


Mapping Data Flow – Source & Sink capabilities

- New capabilities for Source transformations:
  - wildcards, file sets,
  - move file / Delete file,
  - auto-detect types,
  - schema validation
  - query statement

- New capabilities for Sink transformations:
  - output to single file,
  - clear folder,
  - truncate table / recreate table,
  - naming patterns

Mapping Data Flow – DataSet File Formats



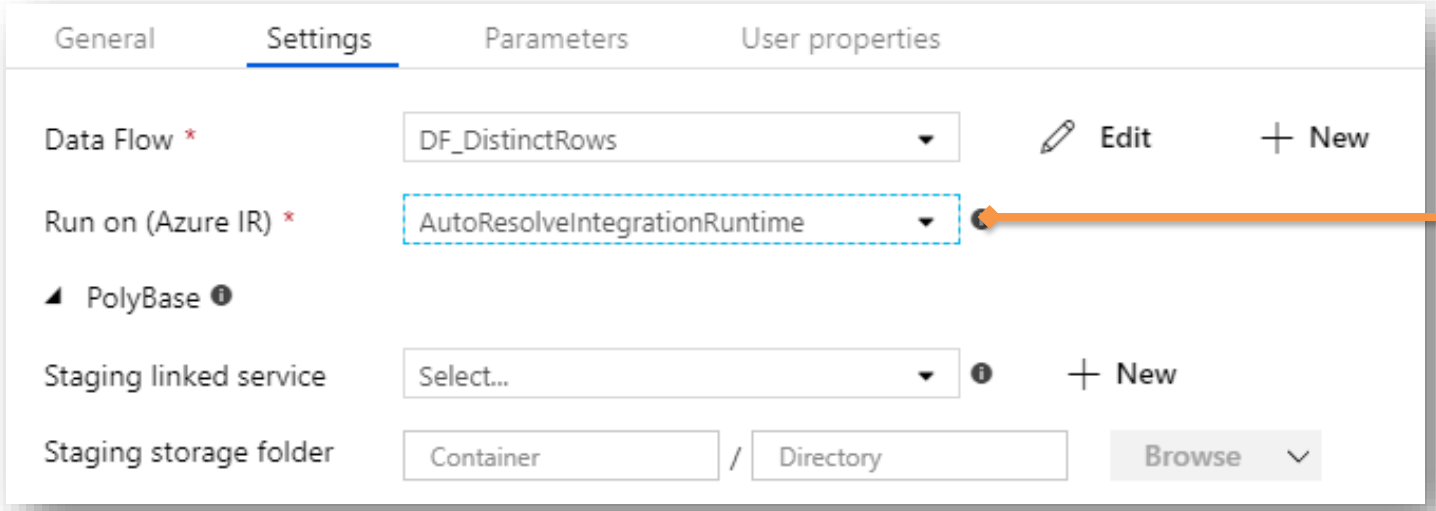
Available NOW

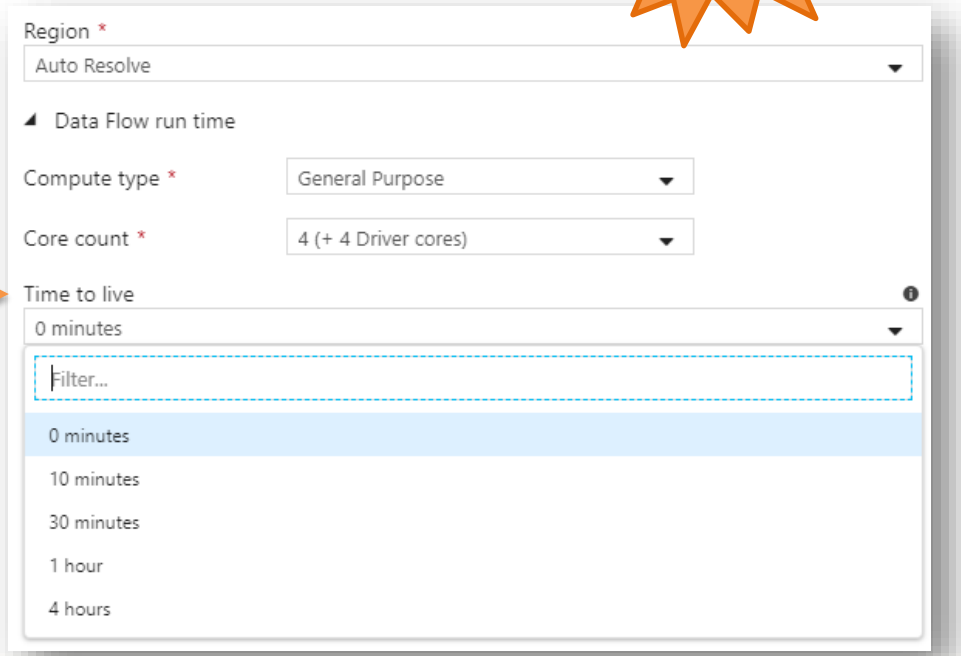
Available SOON

Mapping Data Flow – Execution Settings

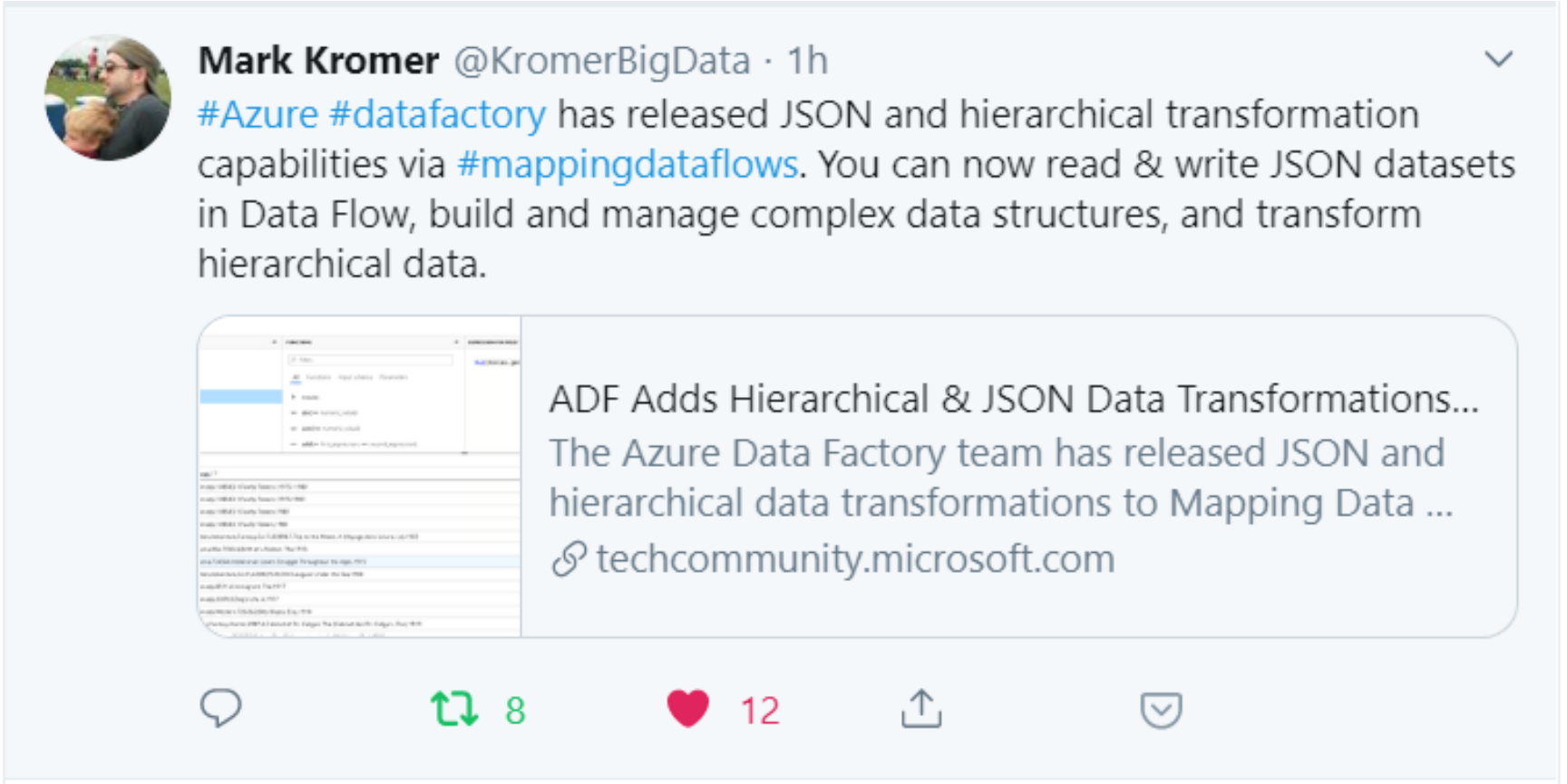
- The Execute Data Flow transformation:
  - Support **parameterized datasets**
  - Control **size of cluster** for specific Azure IR
  - Define **TTL (Time-To-Live)** to Azure IR to reduce data flow activity time

NEW!





Latest updates? Go Twitter!



SSIS vs ADF activities vs T-SQL

Activity	Description	SSIS equivalent	SQL Server equivalent
New branch	Create a new flow branch with the same data	Multicast (+icon)	`SELECT INTO` `SELECT OUTPUT`
Join	Join data from two streams based on a condition	Merge join	`INNER | LEFT | RIGHT JOIN,` `CROSS | FULL OUTER JOIN`

Conditional Split	Route data into different streams based on conditions	Conditional Split	`SELECT INTO WHERE condition1`
			`SELECT INTO WHERE condition2`
			`CASE ... WHEN`

Union	Collect data from multiple streams	Union All	`SELECT col1a UNION (ALL)` `SELECT col1b`
Lookup	Lookup additional data from another stream	Lookup	`LEFT | RIGHT JOIN`
Derived Column	Compute new columns based on the existing once	Derived Column	`SELECT Column1 * 1.09 as NewColumn`

Aggregate	Calculate aggregation on the stream	Aggregate	`SELECT Year(DateOfBirth) as Year,`
			`MIN(), MAX(), AVG()`
			`GROUP BY Year(DateOfBirth)`

http://bit.ly/ADFDFvsSSIS

http://bit.ly/ADFDF-CheatSheet

Resources

http://sqlplayer.net/ADF

Q&A

Muchas gracias! Thank you!

kamil@nowinski.net

@NowinskiK @SQLPlayer

SQLPlayer.net

https://github.com/NowinskiK/CommunityEvents

Kamil Nowinski

Microsoft Data Platform MVP

MCSE Data Platform & MCSE Data Management and Analytics