Kamil Nowiński

PASS
SQLSATURDAY
SLOVENIA | 08 DEC 2018

Azure Data Factory: Data Flows – first blood

Dragon source: https://www.flickr.com/photos/wili/2628869994/in/gallery-41926029@N05-72157622307278981/

Thank you to our AWESOME sponsors!

Microsoft

gen-i

Xnet
vaš zanesljiv IT partner

IN5I6hT
KNOW
YOUR
NUMBERS

redgate
ingeniously simple

SentryOne™

RevDeBug

About me

Kamil Nowinski

Microsoft CERTIFIED
Solutions Associate
SQL Server 2012

MCSE Data Platform

MVP Microsoft® Most Valuable Professional

Microsoft Data Platform **MVP**
Speaker, blogger, data enthusiast

Senior Data Engineer at ASOS (www.asos.com)
13+ yrs experience as DEV/DBA

Member of the Data Community PL (SQLDay)

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

redgate FRIENDS

(Almost) BRAND NEW BLOG

INTERVIEWS

SQLPlayer

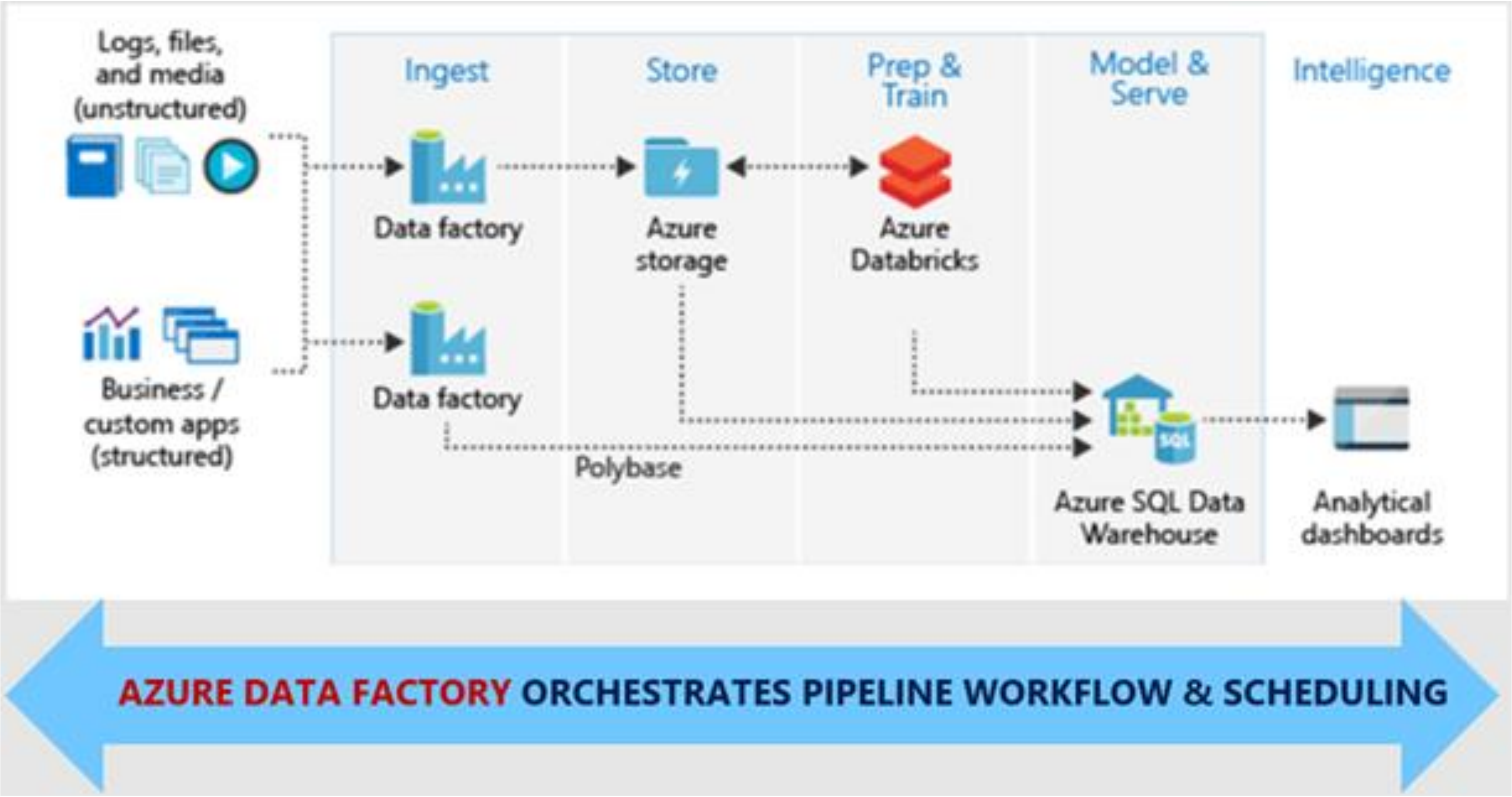
Play with data & have fun!

www.SQLPlayer.net

PODCAST – Interviews with…

www.SQLPlayer.net

What the Azure Data Factory is?



Access all your data:

- 75+ connectors & growing
- Azure IR available in 20 regions
- Hybrid connectivity using self-hosted IR: on-prem & VNet

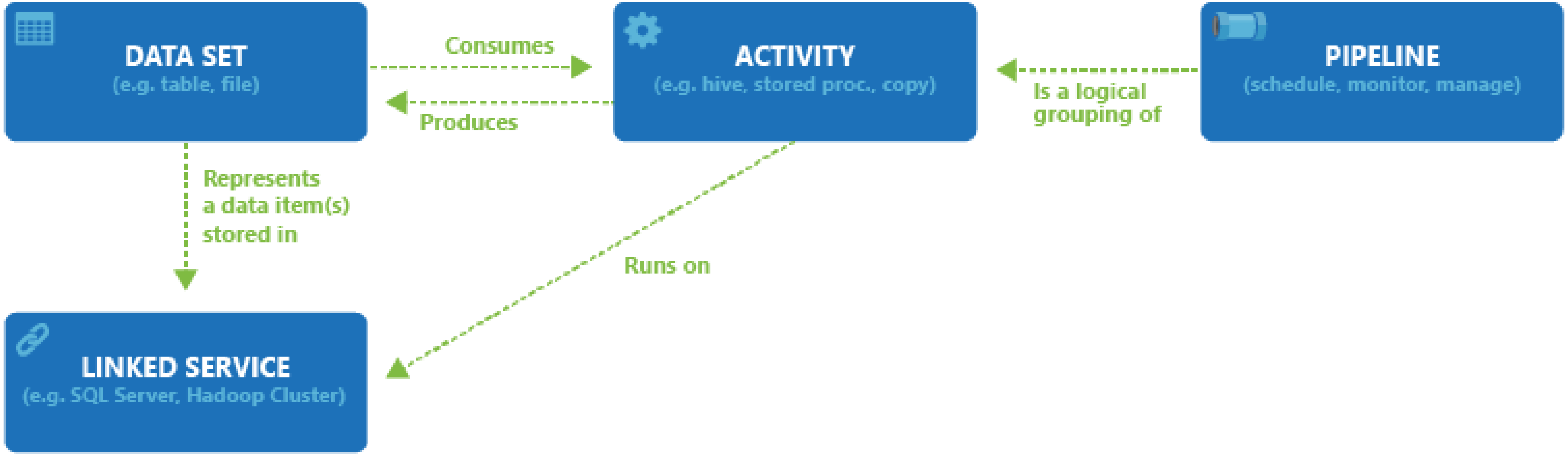
Azure (13)	Database (24)		File Storage (5)	NoSQL (3)	Services and Apps (28)		Generic (4)
Blob Storage	Amazon Redshift	Netezza	Amazon S3	Cassandra	Amazon MWS	Office 365 *	HTTP
Cosmos DB (MongoDB API) *	DB2	Oracle	File System	Couchbase	CDS for Apps	Paypal	OData
Cosmos DB (SQL API)	Drill	Phoenix	FTP	MongoDB	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



DATA SET
(e.g. table, file)

Consumes

Produces

ACTIVITY
(e.g. hive, stored proc., copy)

Is a logical grouping of

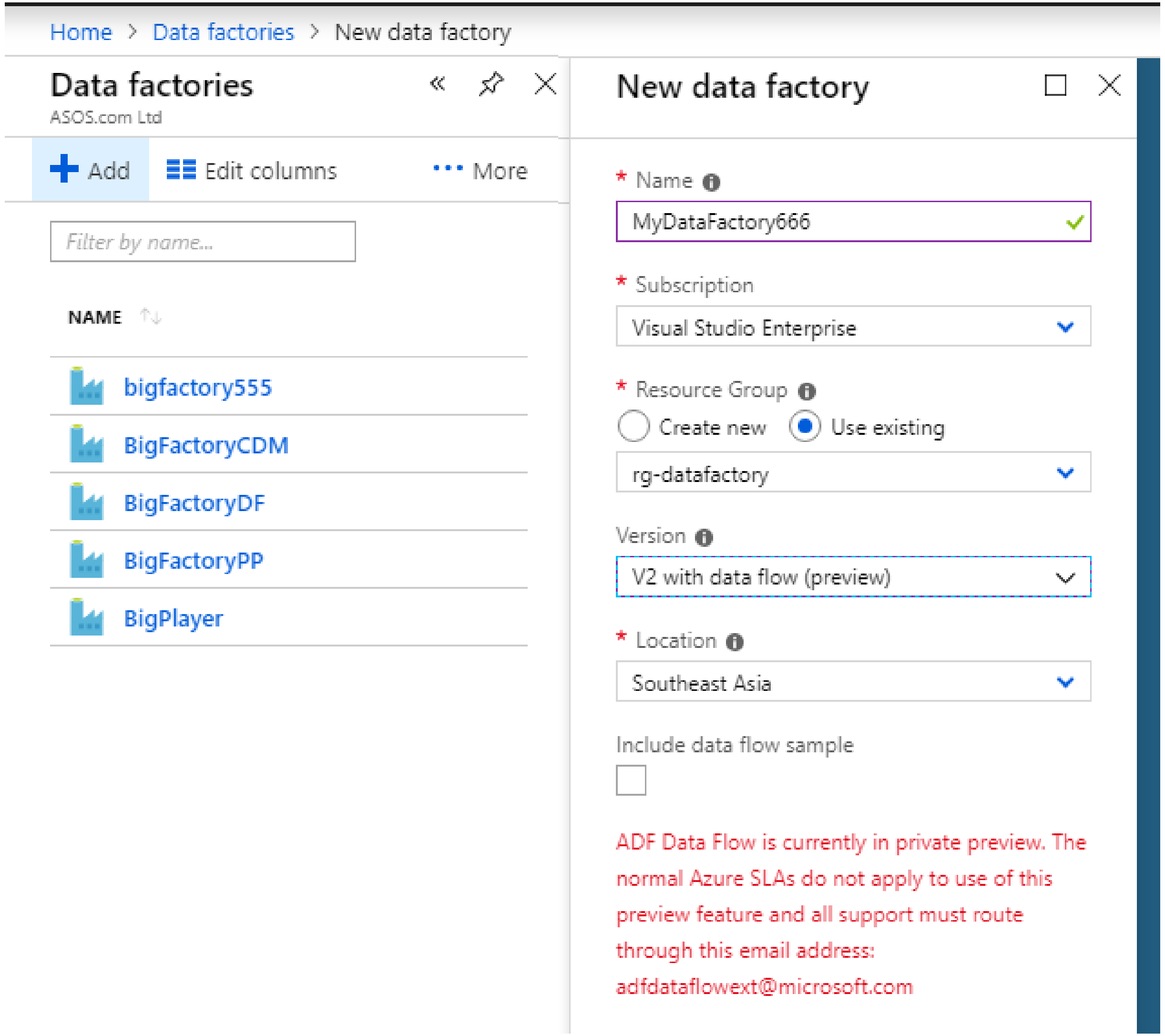
PIPELINE
(schedule, monitor, manage)

Represents a data item(s) stored in

Runs on

LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

How to create Azure Data Factory with Data Flow?



Home > Data factories > New data factory

Data factories

ASOS.com Ltd

NAME
bigfactory555
BigFactoryCDM
BigFactoryDF
BigFactoryPP
BigPlayer

New data factory

* Name: MyDataFactory666 ✓

* Subscription: Visual Studio Enterprise

* Resource Group: ○ Create new ● Use existing

rg-datafactory

Version: V2 with data flow (preview)

* Location: Southeast Asia

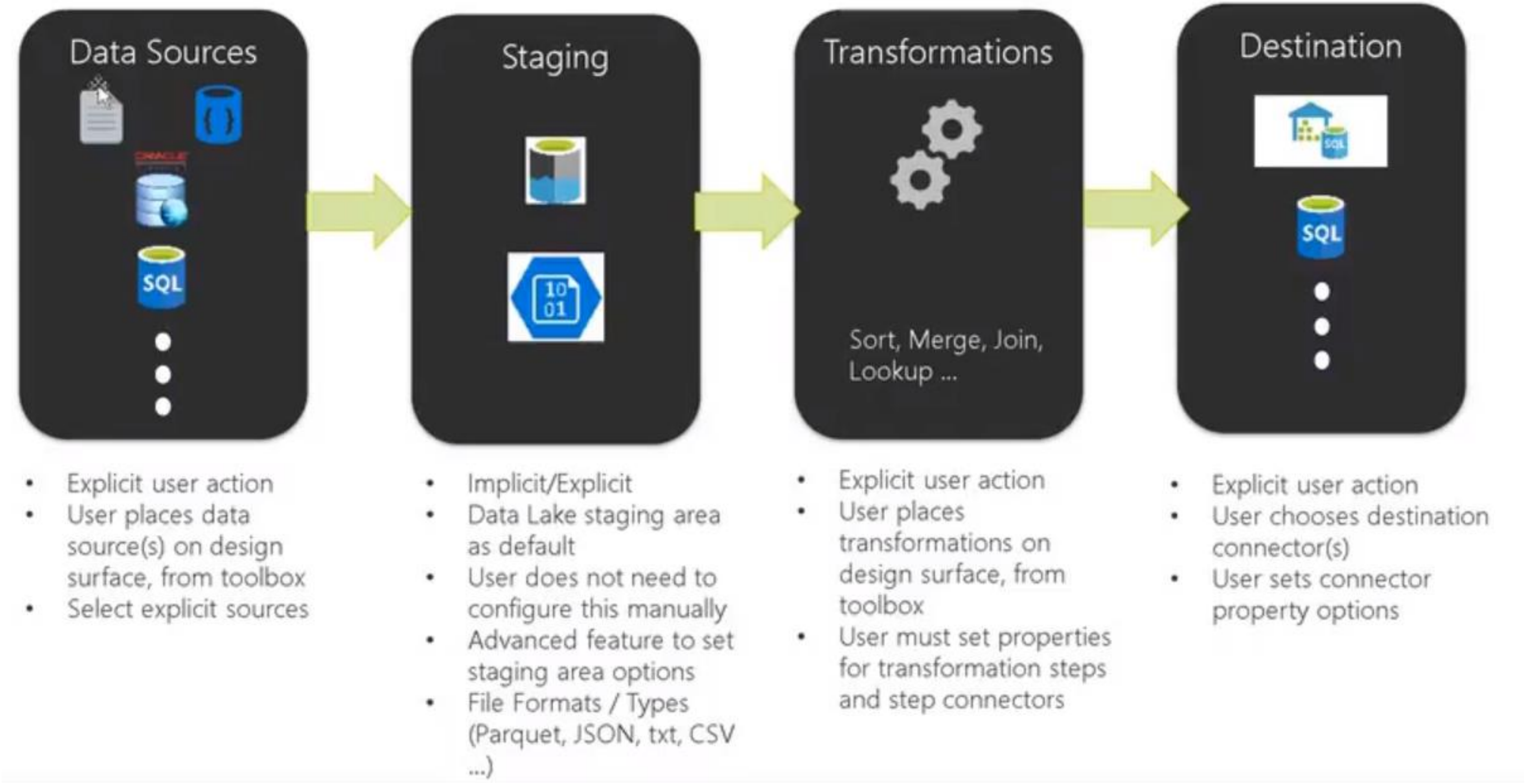
Include data flow sample □

ADF Data Flow is currently in private preview. The normal Azure SLAs do not apply to use of this preview feature and all support must route through this email address: adfdataflowext@microsoft.com

PASS SQLSATURDAY SLOVENIA | 08 DEC 2018

Visual Data Transformations with Data Flow

What the hell Data Flows are?



Data Sources

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

Staging

- 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 ...)

Transformations

Sort, Merge, Join, Lookup ...

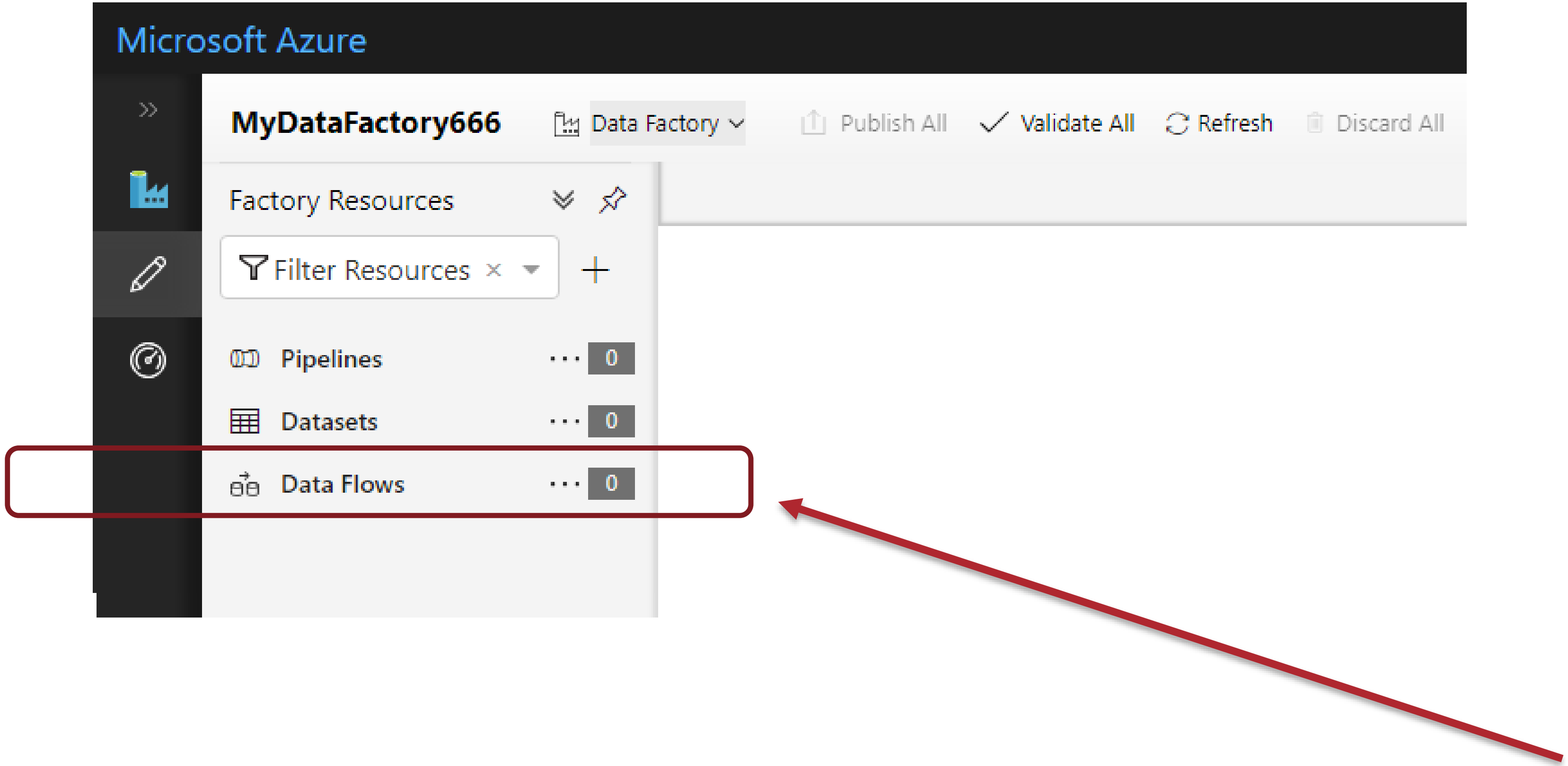
- Explicit user action
- User places transformations on design surface, from toolbox
- User must set properties for transformation steps and step connectors

Destination

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

src: (Microsoft) ADF Data Flow Private Preview Overview

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

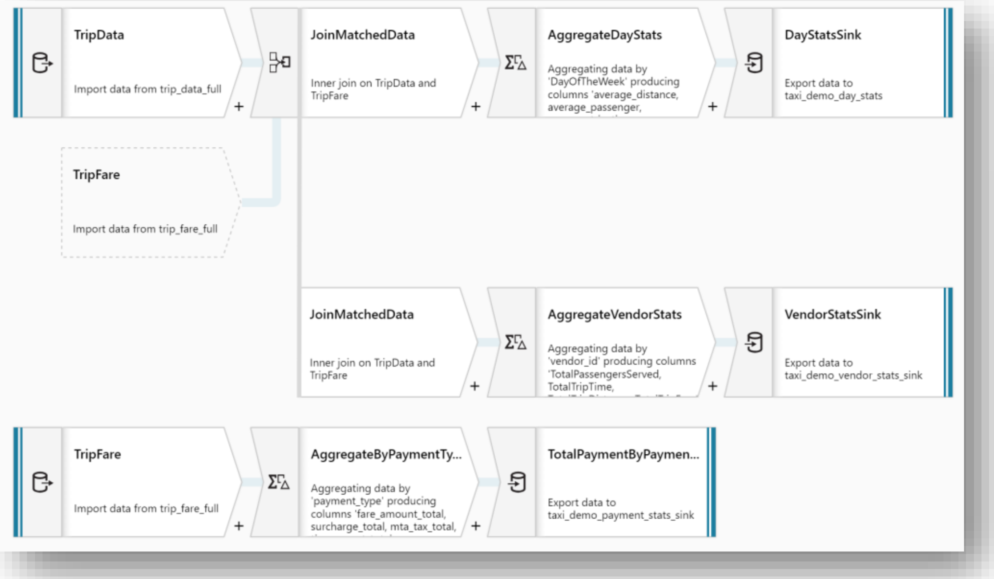


Code-free Data Transformation At Scale

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



TripData — Import data from trip_data_full

JoinMatchedData — Inner join on TripData and TripFare

AggregateDayStats — Aggregating data by 'DayOfTheWeek' producing columns 'average_distance, average_passenger,

DayStatsSink — Export data to taxi_demo_day_stats

TripFare — Import data from trip_fare_full

JoinMatchedData — Inner join on TripData and TripFare

AggregateVendorStats — Aggregating data by 'vendor_id' producing columns 'TotalPassengersServed, TotalTripTime,

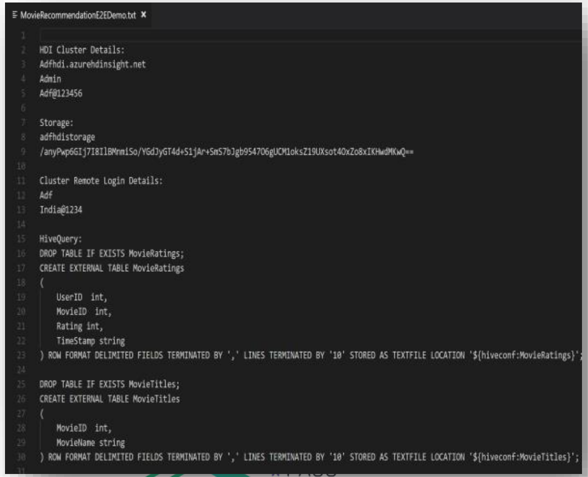
VendorStatsSink — Export data to taxi_demo_vendor_stats_sink

TripFare — Import data from trip_fare_full

AggregateByPaymentTy... — Aggregating data by 'payment_type' producing columns 'fare_amount_total, surcharge_total, mta_tax_total,

TotalPaymentByPaymen... — Export data to taxi_demo_payment_stats_sink

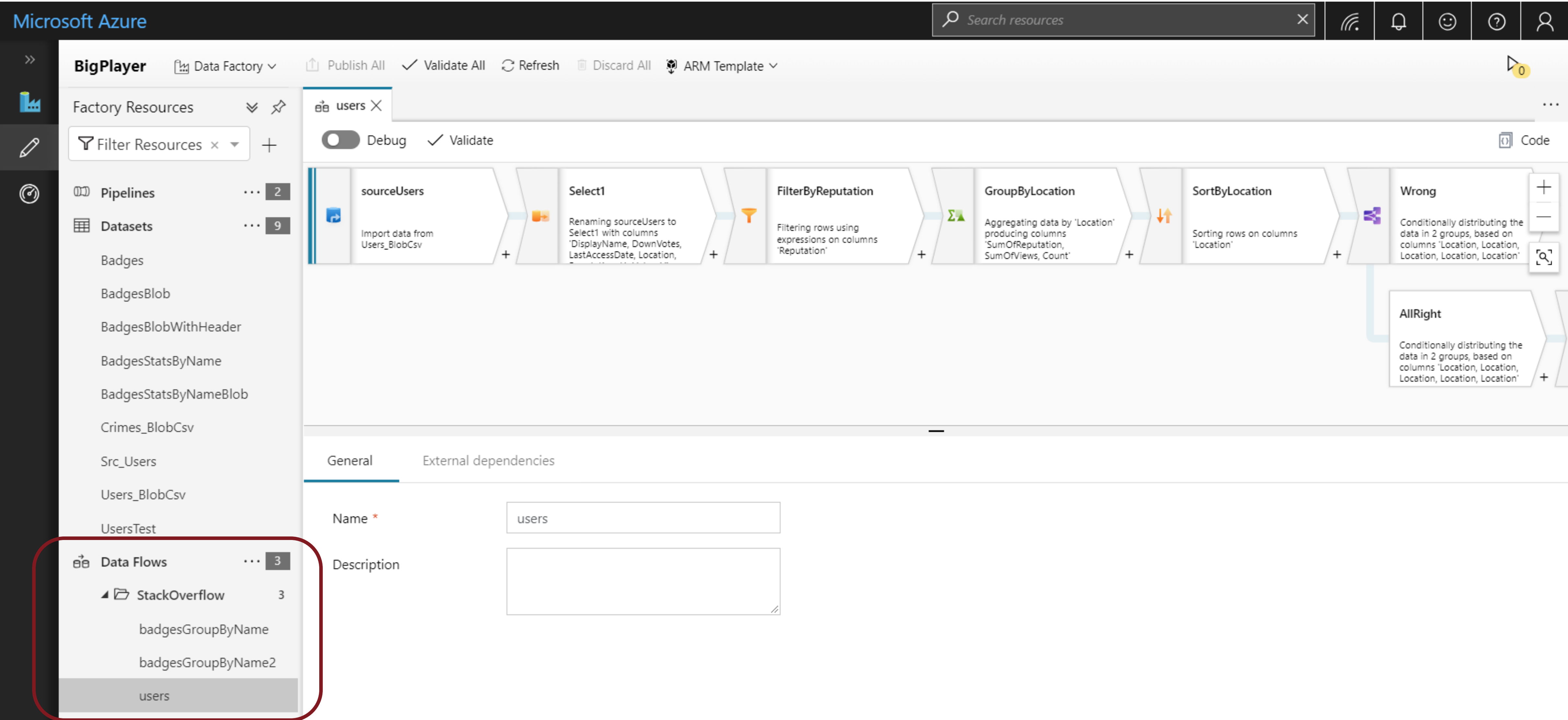
… not



MovieRecommendationE2EDemo.txt ×

```

HDI Cluster Details:
Adfhdi.azurehdinsight.net
Admin
Adf@123456

Storage:
adfhdistorage
/anyPwp6GIj7I8IlBMnmiSo/YGdJyGT4d+S1jAr+SmS7bJgb954706gUCM1oksZ19UXsot4OxZo8xIKHwdMKwQ==

Cluster Remote Login Details:
Adf
India@1234

HiveQuery:
DROP TABLE IF EXISTS MovieRatings;
CREATE EXTERNAL TABLE MovieRatings
(
    UserID  int,
    MovieID  int,
    Rating int,
    TimeStamp string
) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '10' STORED AS TEXTFILE LOCATION '${hiveconf:MovieRatings}';

DROP TABLE IF EXISTS MovieTitles;
CREATE EXTERNAL TABLE MovieTitles
(
    MovieID  int,
    MovieName string
) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '10' STORED AS TEXTFILE LOCATION '${hiveconf:MovieTitles}';
```

Authoring of Azure Data Factory (v2)



sourceUsers: Import data from Users_BlobCsv

Select1: Renaming sourceUsers to Select1 with columns 'DisplayName, DownVotes, LastAccessDate, Location,

FilterByReputation: Filtering rows using expressions on columns 'Reputation'

GroupByLocation: Aggregating data by 'Location' producing columns 'SumOfReputation, SumOfViews, Count'

SortByLocation: Sorting rows on columns 'Location'

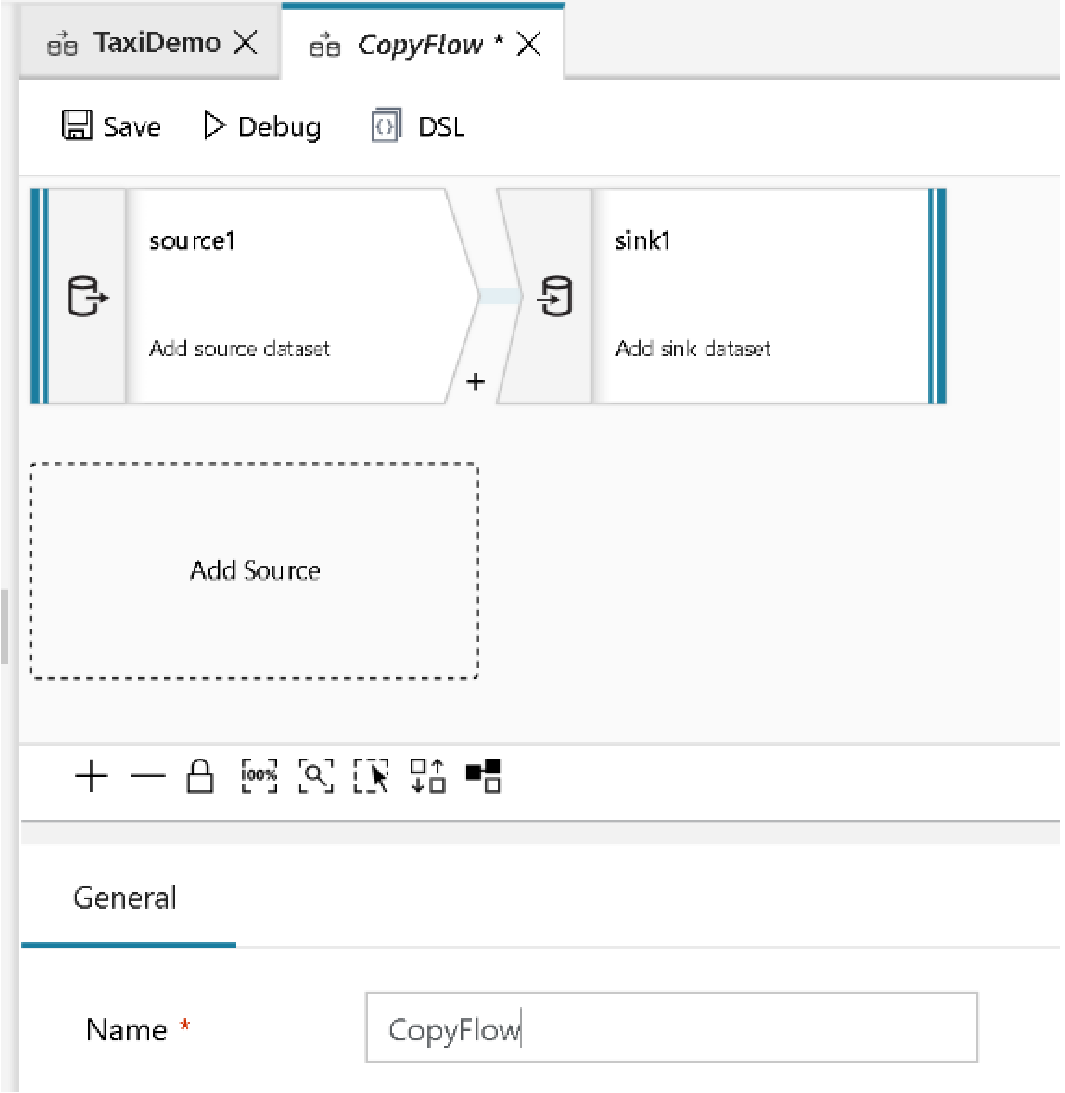
Wrong: Conditionally distributing the data in 2 groups, based on columns 'Location, Location, Location, Location, Location'

AllRight: Conditionally distributing the data in 2 groups, based on columns 'Location, Location, Location, Location, Location'

Name * users

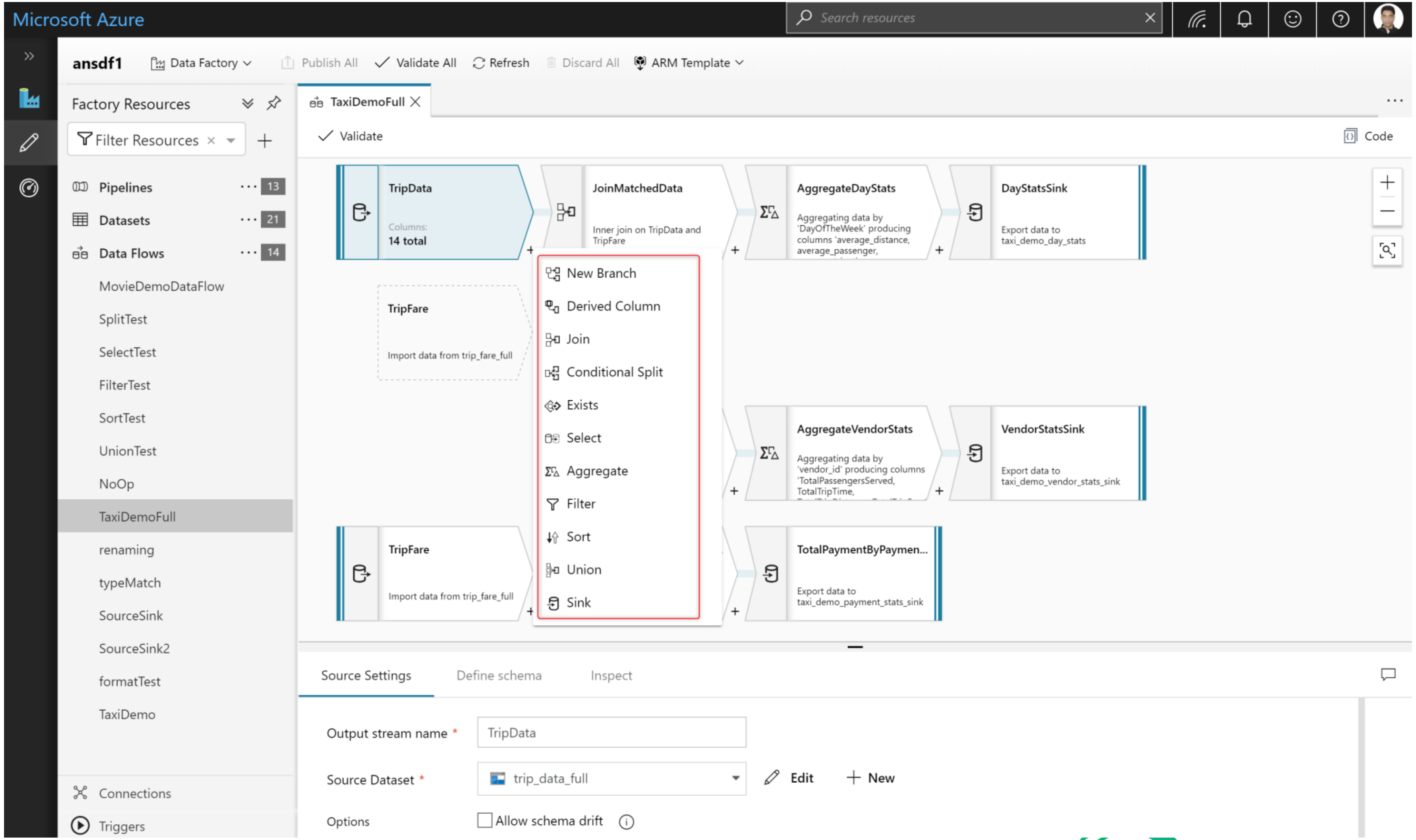
Description

Simple Copy Flow



Name * CopyFlow

Guided experience to build data flows



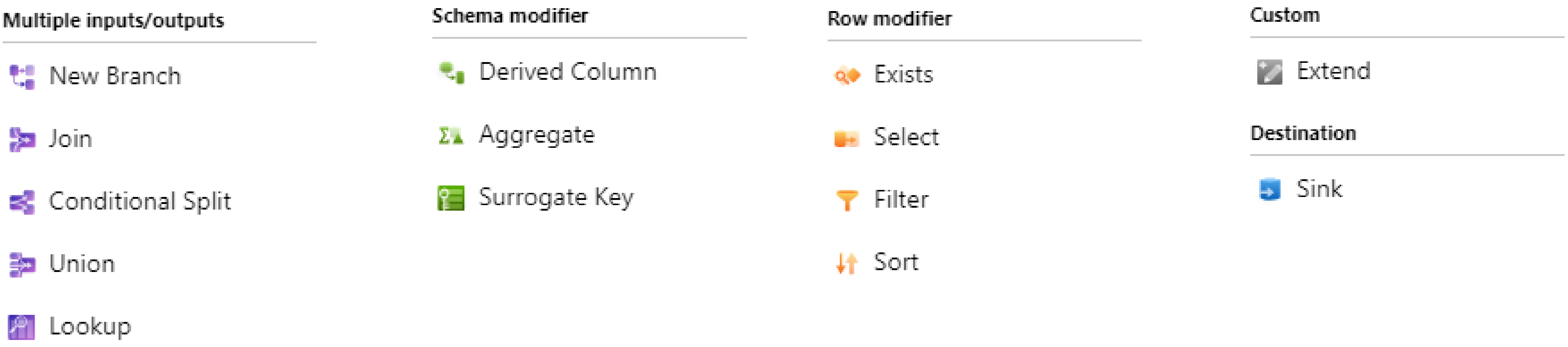
Output stream name * TripData

Source Dataset * trip_data_full

Options □ Allow schema drift

SQLSATURDAY
SLOVENIA | 08 DEC 2018

Data Flow: Components = Actions *



Multiple inputs/outputs

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

Schema modifier

- Derived Column
- Aggregate
- Surrogate Key

Row modifier

- Exists
- Select
- Filter
- Sort

Custom

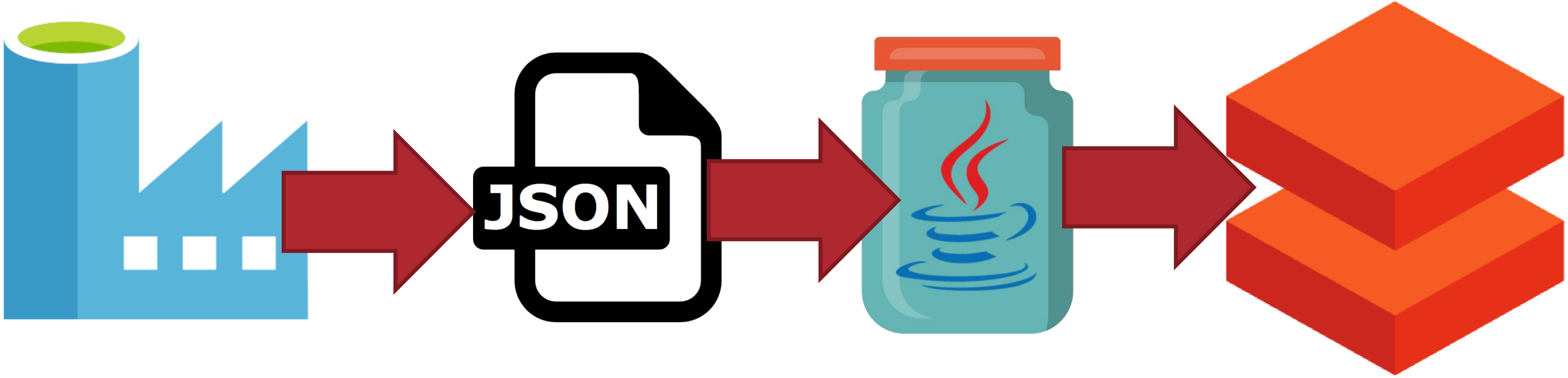
- Extend

Destination

- Sink

* With some small exceptions

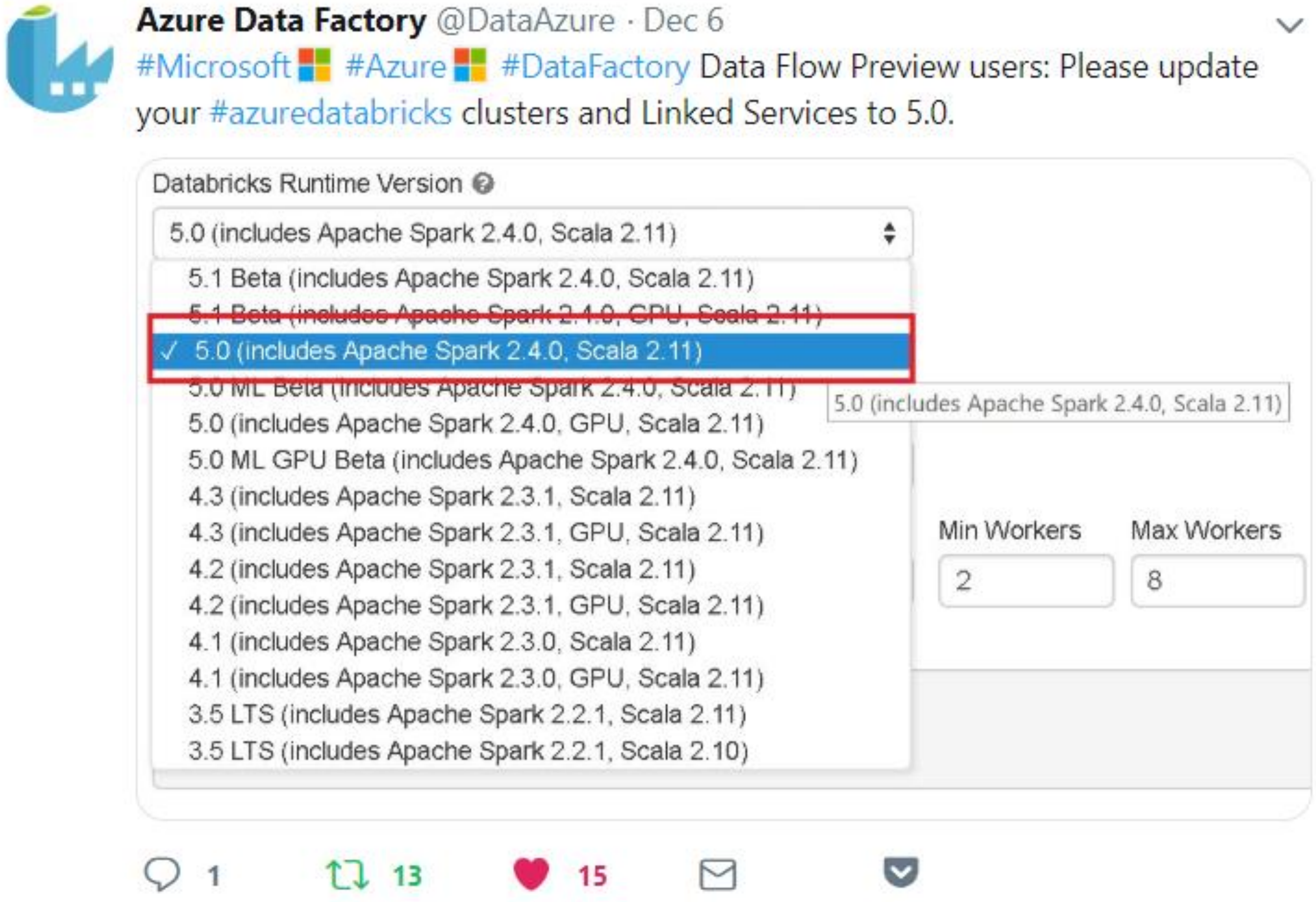
What is going on behind the scenes?



JAR

Azure Databricks

Azure Databricks version



Azure Data Factory @DataAzure · Dec 6

#Microsoft #Azure #DataFactory Data Flow Preview users: Please update your #azuredatabricks clusters and Linked Services to 5.0.

Databricks Runtime Version

5.0 (includes Apache Spark 2.4.0, Scala 2.11)

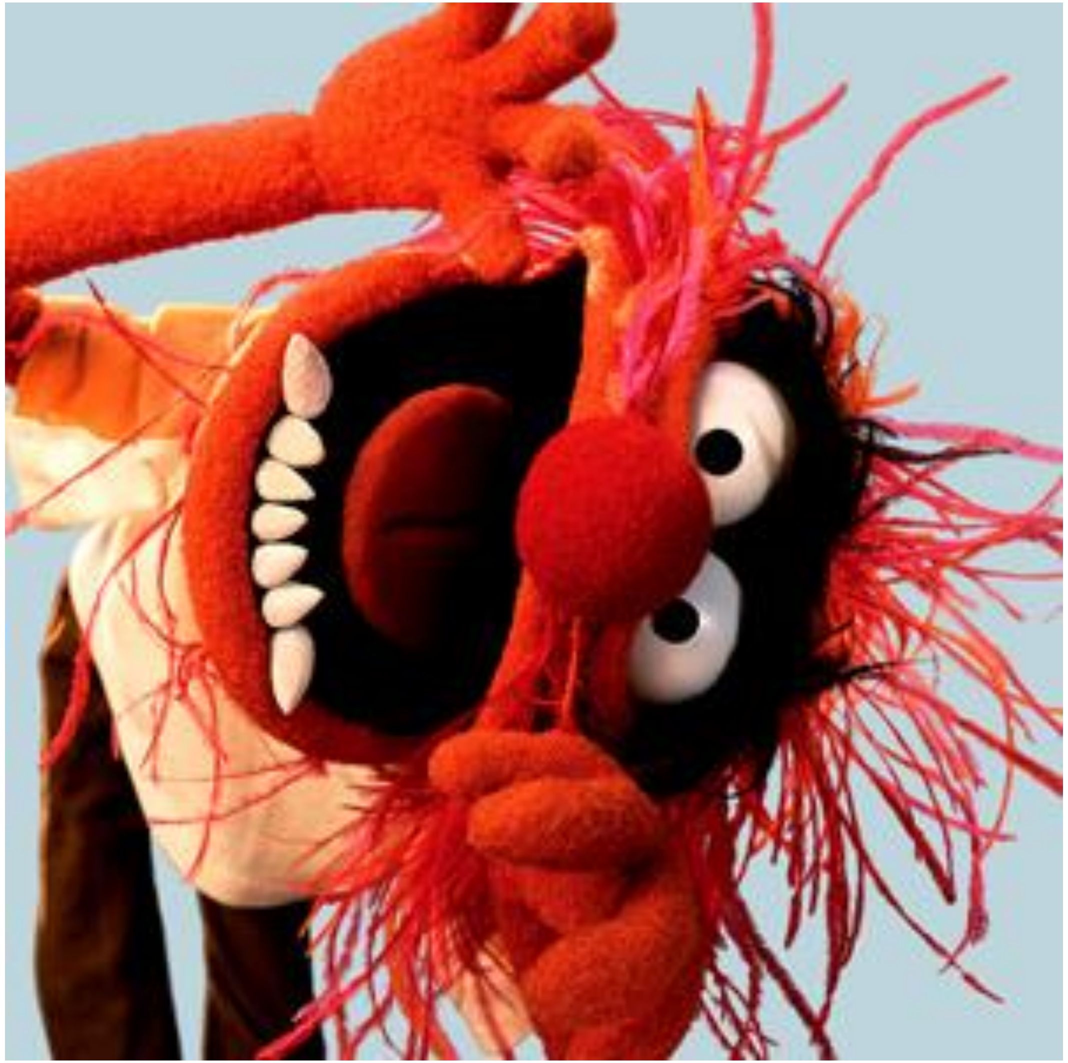
- 5.1 Beta (includes Apache Spark 2.4.0, Scala 2.11)
- 5.1 Beta (includes Apache Spark 2.4.0, GPU, Scala 2.11)
- ✓ 5.0 (includes Apache Spark 2.4.0, Scala 2.11)
- 5.0 ML Beta (includes Apache Spark 2.4.0, Scala 2.11)
- 5.0 (includes Apache Spark 2.4.0, GPU, Scala 2.11)
- 5.0 ML GPU Beta (includes Apache Spark 2.4.0, Scala 2.11)
- 4.3 (includes Apache Spark 2.3.1, Scala 2.11)
- 4.3 (includes Apache Spark 2.3.1, GPU, Scala 2.11)
- 4.2 (includes Apache Spark 2.3.1, Scala 2.11)
- 4.2 (includes Apache Spark 2.3.1, GPU, Scala 2.11)
- 4.1 (includes Apache Spark 2.3.0, Scala 2.11)
- 4.1 (includes Apache Spark 2.3.0, GPU, Scala 2.11)
- 3.5 LTS (includes Apache Spark 2.2.1, Scala 2.11)
- 3.5 LTS (includes Apache Spark 2.2.1, Scala 2.10)

5.0 (includes Apache Spark 2.4.0, Scala 2.11)

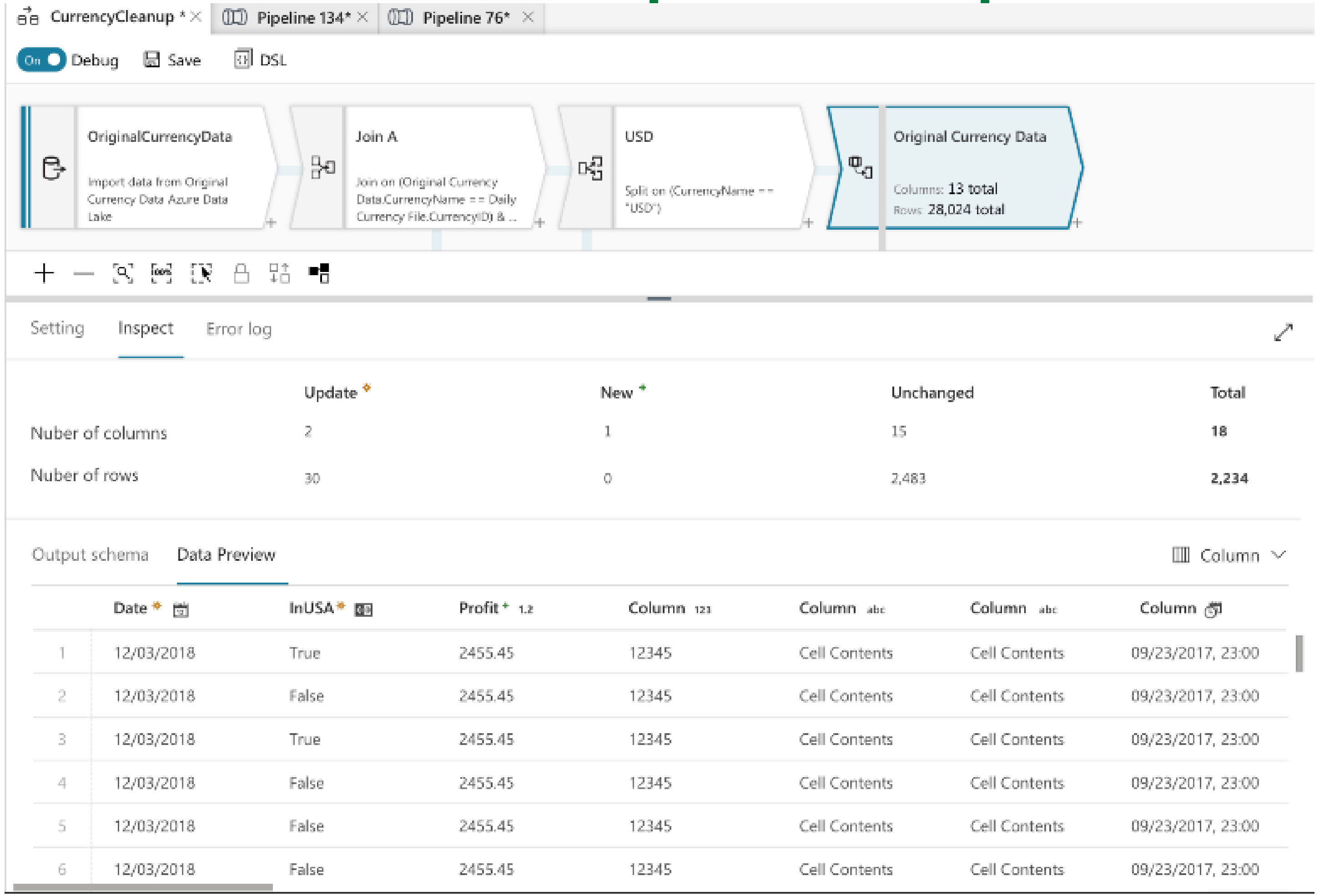
Min Workers: 2

Max Workers: 8

DEMO TIME



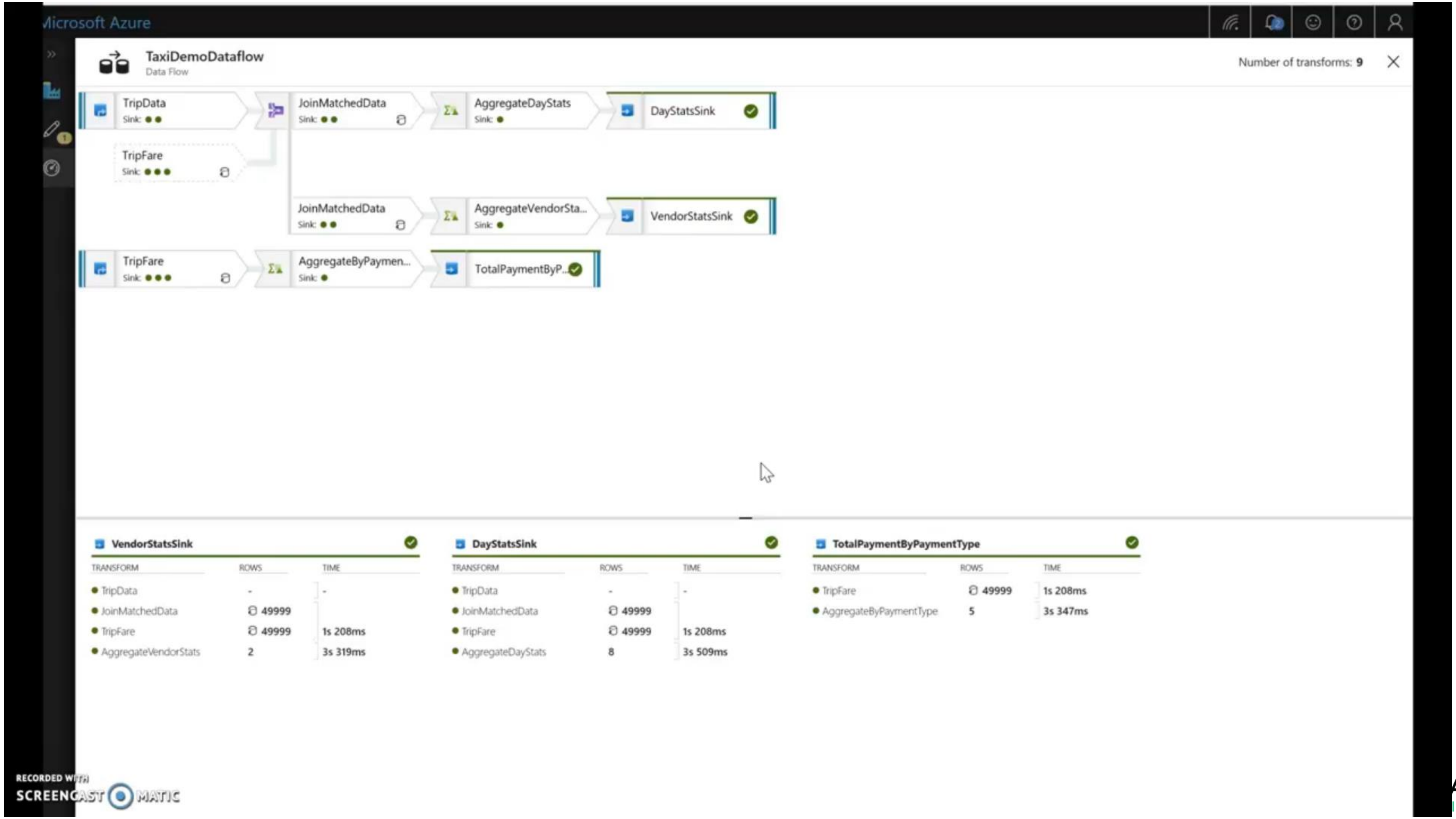
Debug mode provides row-level context and visible results in inspector pane



	Update *	New *	Unchanged	Total
Nuber of columns	2	1	15	**18**
Nuber of rows	30	0	2,483	**2,234**

	Date *	InUSA *	Profit * 1.2	Column 123	Column abc	Column abc	Column
1	12/03/2018	True	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
2	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
3	12/03/2018	True	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
4	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
5	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
6	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00

Data Flow Execution Plan



VendorStatsSink ✓

TRANSFORM	ROWS	TIME
● TripData	-	-
● JoinMatchedData	49999	
● TripFare	49999	1s 208ms
● AggregateVendorStats	2	3s 319ms

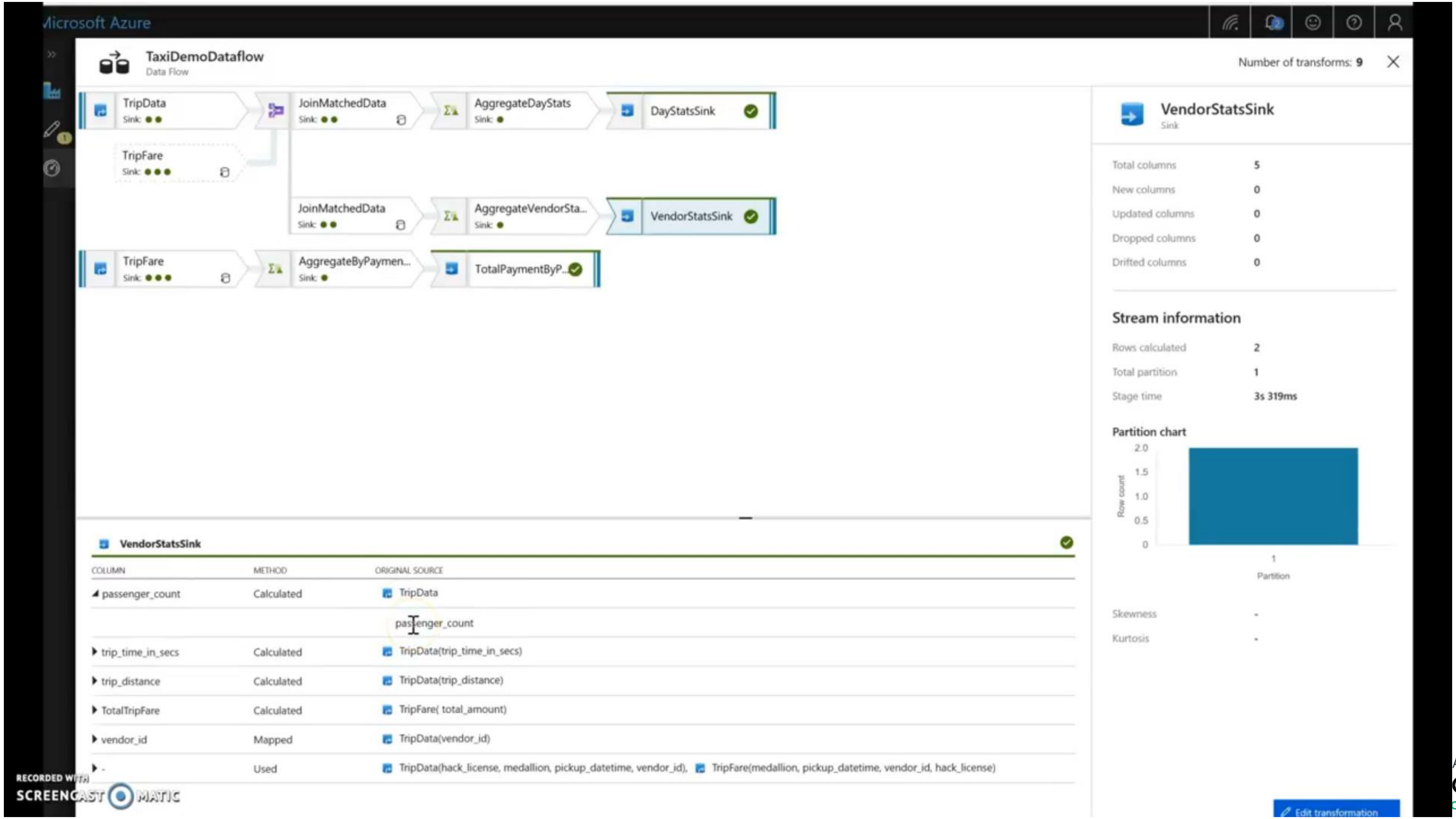
DayStatsSink ✓

TRANSFORM	ROWS	TIME
● TripData	-	-
● JoinMatchedData	49999	
● TripFare	49999	1s 208ms
● AggregateDayStats	8	3s 509ms

TotalPaymentByPaymentType ✓

TRANSFORM	ROWS	TIME
● TripFare	49999	1s 208ms
● AggregateByPaymentType	5	3s 347ms

Data Flow Data Lineage



VendorStatsSink
Sink

Total columns	5
New columns	0
Updated columns	0
Dropped columns	0
Drifted columns	0

Stream information

Rows calculated	2
Total partition	1
Stage time	3s 319ms

Partition chart

Skewness	-
Kurtosis	-

VendorStatsSink

COLUMN	METHOD	ORIGINAL SOURCE
◢ passenger_count	Calculated	TripData
		passenger_count
▸ trip_time_in_secs	Calculated	TripData(trip_time_in_secs)
▸ trip_distance	Calculated	TripData(trip_distance)
▸ TotalTripFare	Calculated	TripFare(total_amount)
▸ vendor_id	Mapped	TripData(vendor_id)
▸ -	Used	TripData(hack_license, medallion, pickup_datetime, vendor_id), TripFare(medallion, pickup_datetime, vendor_id, hack_license)

Resources

- Microsoft Azure Data Factory – Tutorials & API Reference
- Azure Data Factory Overview
- Azure Data Factory – Data integration service
- ADF Data Flow’s documentation
- ADF Data Flow’s videos
- SQLPlayer blog:
    - Azure Data Factory v2 and its available components in Data Flows
    - Follow this tag on SQLPlayer blog: ADFDF

Q&A

Thank you! Hvala! Dziękuję!

kamil@nowinski.net

@NowinskiK @SQLPlayer

SQLPlayer.net

Kamil Nowinski

Microsoft Data Platform MVP

MCSE Data Platform & MCSE Data Management and Analytics

Thank you to our AWESOME sponsors!

Microsoft

gen-i

Xnet
vaš zanesljiv IT partner

IN5I6hT
KNOW
YOUR
NUMBERS

redgate
ingeniously simple

SentryOne™

RevDeBug