

Azure Data Factory: Data Flows – first blood

Kamil Nowiński

About me

Kamil Nowinski



Microsoft
CERTIFIED
Solutions Associate
SQL Server 2012



Microsoft Data Platform **MVP**

Speaker, blogger, data enthusiast

Senior Data Engineer at ASOS (www.asos.com)

15+ yrs experience as DEV/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

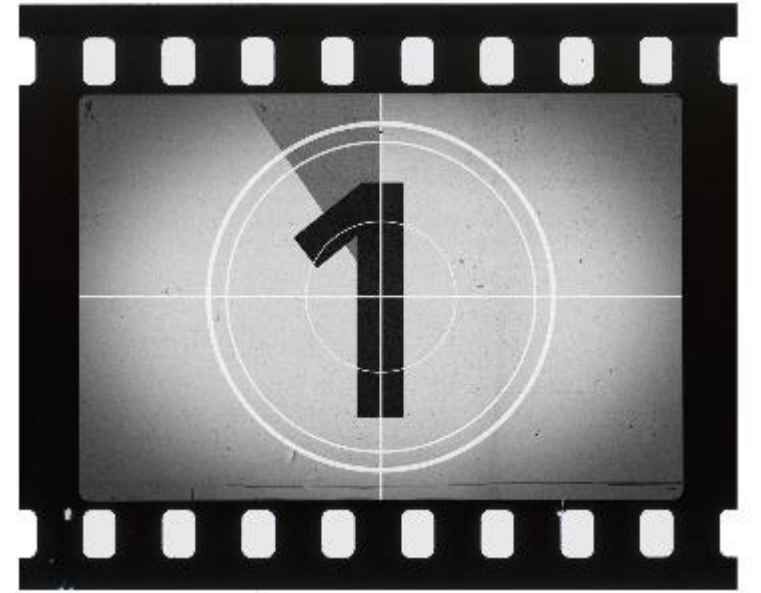
@NowinskiK, @SQLPlayer

BLOG & Interviews



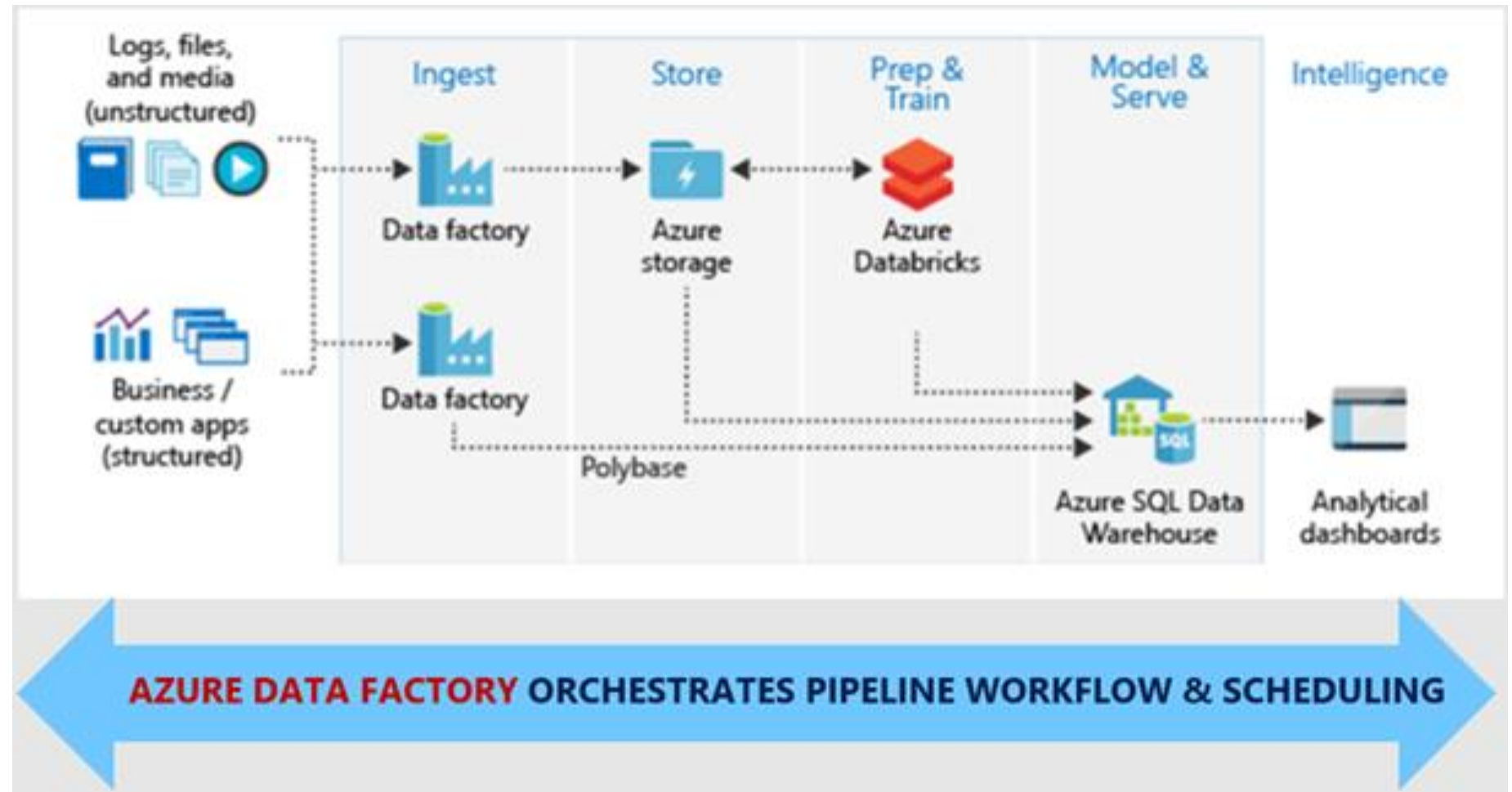
www.SQLPlayer.net

#AskSQLFamily - Trailer



<https://sqlplayer.net/2019/01/seventeen-months-of-podcasting-recap/>

What the Azure Data Factory is?

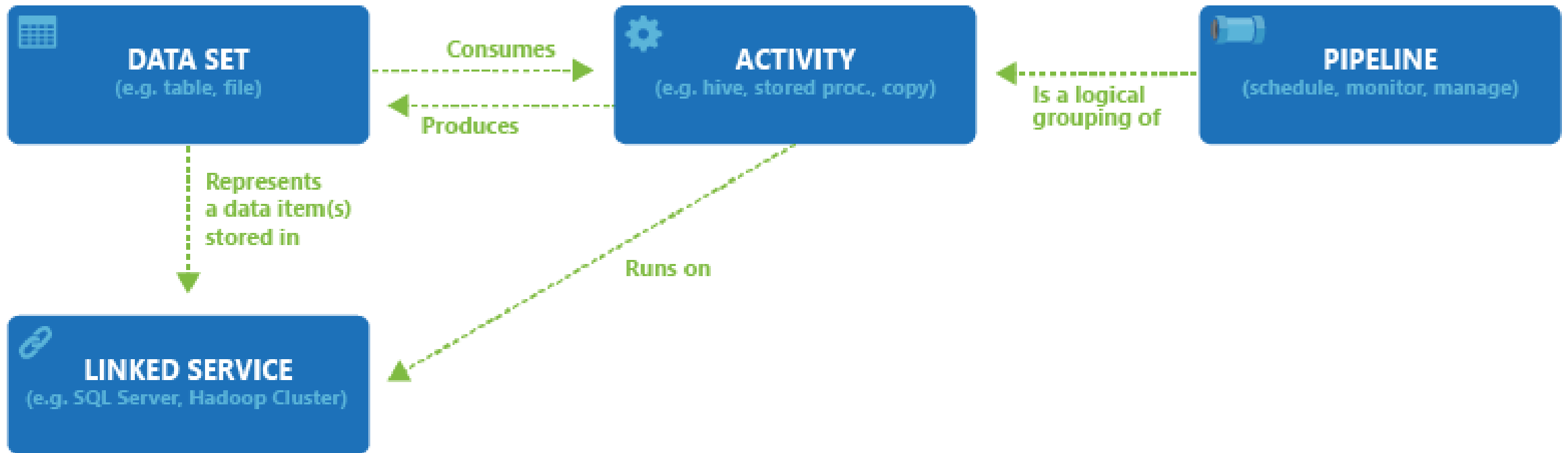


Access all your data

- 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



How to create Azure Data Factory with Data Flow?

Home > Data factories > New data factory

Data factories

ASOS.com Ltd

+ Add Edit columns More

Filter by name...

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

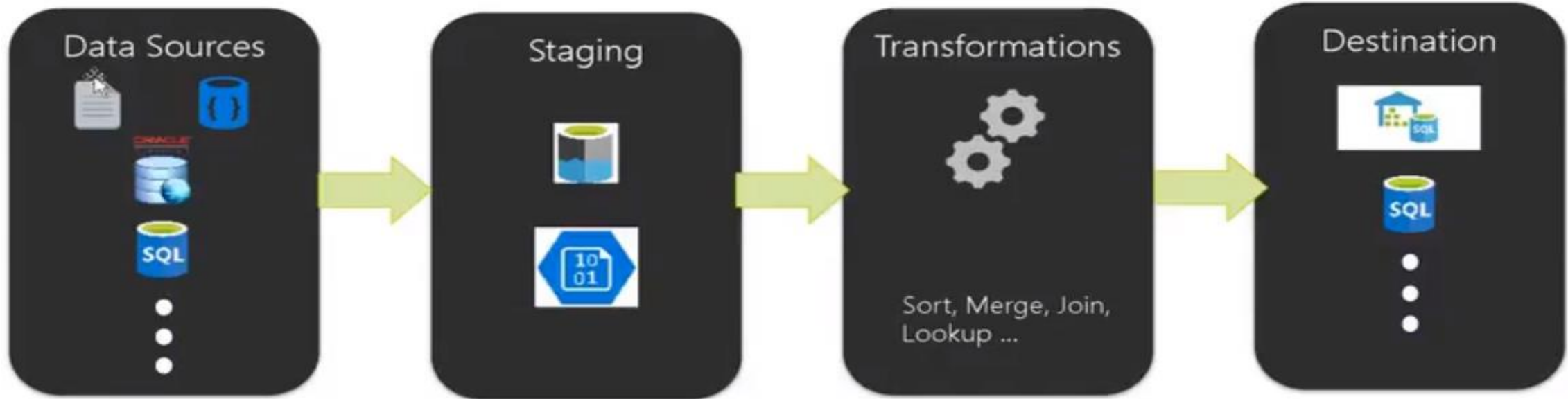
Include data flow sample ☐

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: adfdflowext@microsoft.com

Visual Data Transformations with

DATA FLOW

What the hell Data Flows are?



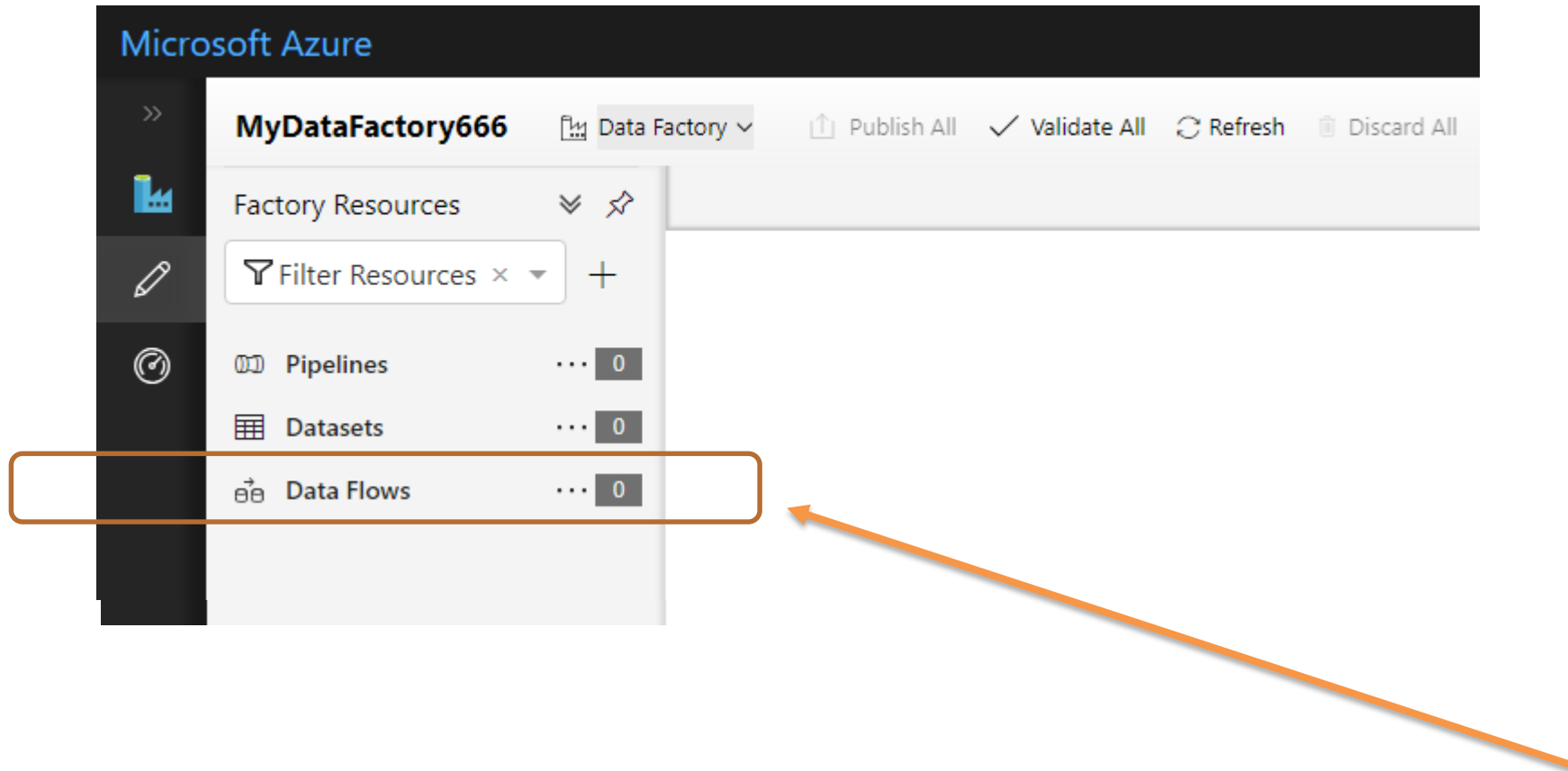
- Explicit user action
- User places data source(s) on design surface, from toolbox
- Select explicit sources

- 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

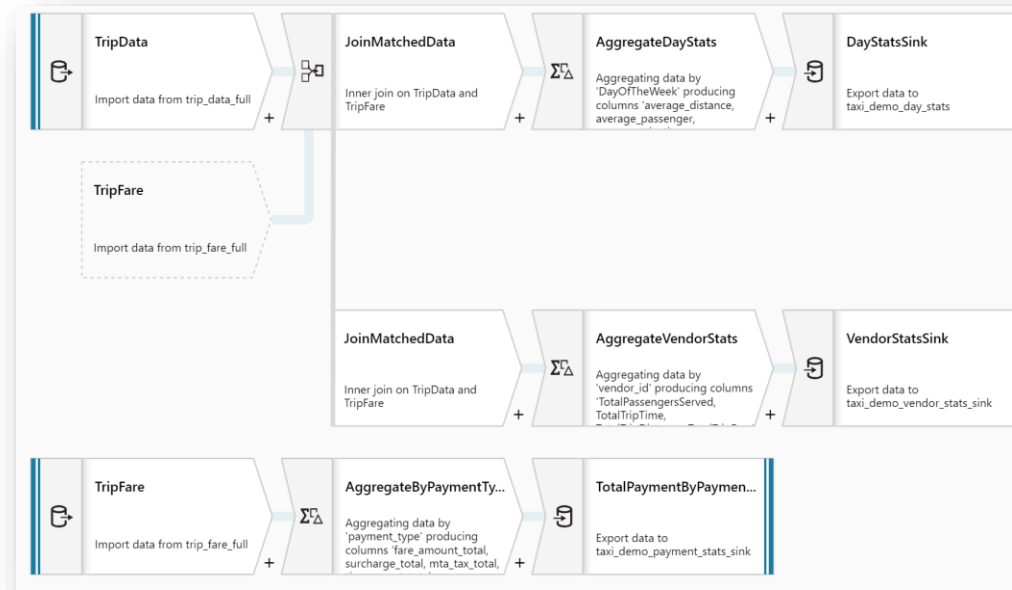
- Explicit user action
- User chooses destination connector(s)
- User sets connector property options

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



... not

```
1 E MovieRecommenderE2EDemo.bt
2
3 HDI Cluster Details:
4 Adfhd.azurehdinsight.net
5 Admin
6 Adf@123456
7
8 Storage:
9 adfhdstorage
10 /any?w=661771811BwmiSo/YGdJyG7d+s1JAr+SnS7b3g954706gK0KsksZ19Uosot40x28x104wdMwQ==
11
12 Cluster Remote Login Details:
13 Adf
14 India@1234
15
16 HiveQuery:
17 DROP TABLE IF EXISTS MovieRatings;
18 CREATE EXTERNAL TABLE MovieRatings
19 (
20   UserID int,
21   MovieID int,
22   Rating int,
23   Timestamp string
24 ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '${hiveconf:MovieRatings}';
25
26 DROP TABLE IF EXISTS MovieTitles;
27 CREATE EXTERNAL TABLE MovieTitles
28 (
29   MovieID int,
30   MovieName string
31 ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '${hiveconf:MovieTitles}';
```

Authoring of Azure Data Factory (v2)

Microsoft Azure

Search resources

BigPlayer Data Factory

Publish All Validate All Refresh Discard All ARM Template

Factory Resources

Filter Resources

Pipelines 2

Datasets 9

- Badges
- BadgesBlob
- BadgesBlobWithHeader
- BadgesStatsByName
- BadgesStatsByNameBlob
- Crimes_BlobCsv
- Src_Users
- Users_BlobCsv
- UsersTest

Data Flows 3

- StackOverflow
 - badgesGroupByName
 - badgesGroupByName2
 - users

users

Debug Validate

sourceUsers

Import data from Users_BlobCsv

Select1

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

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'

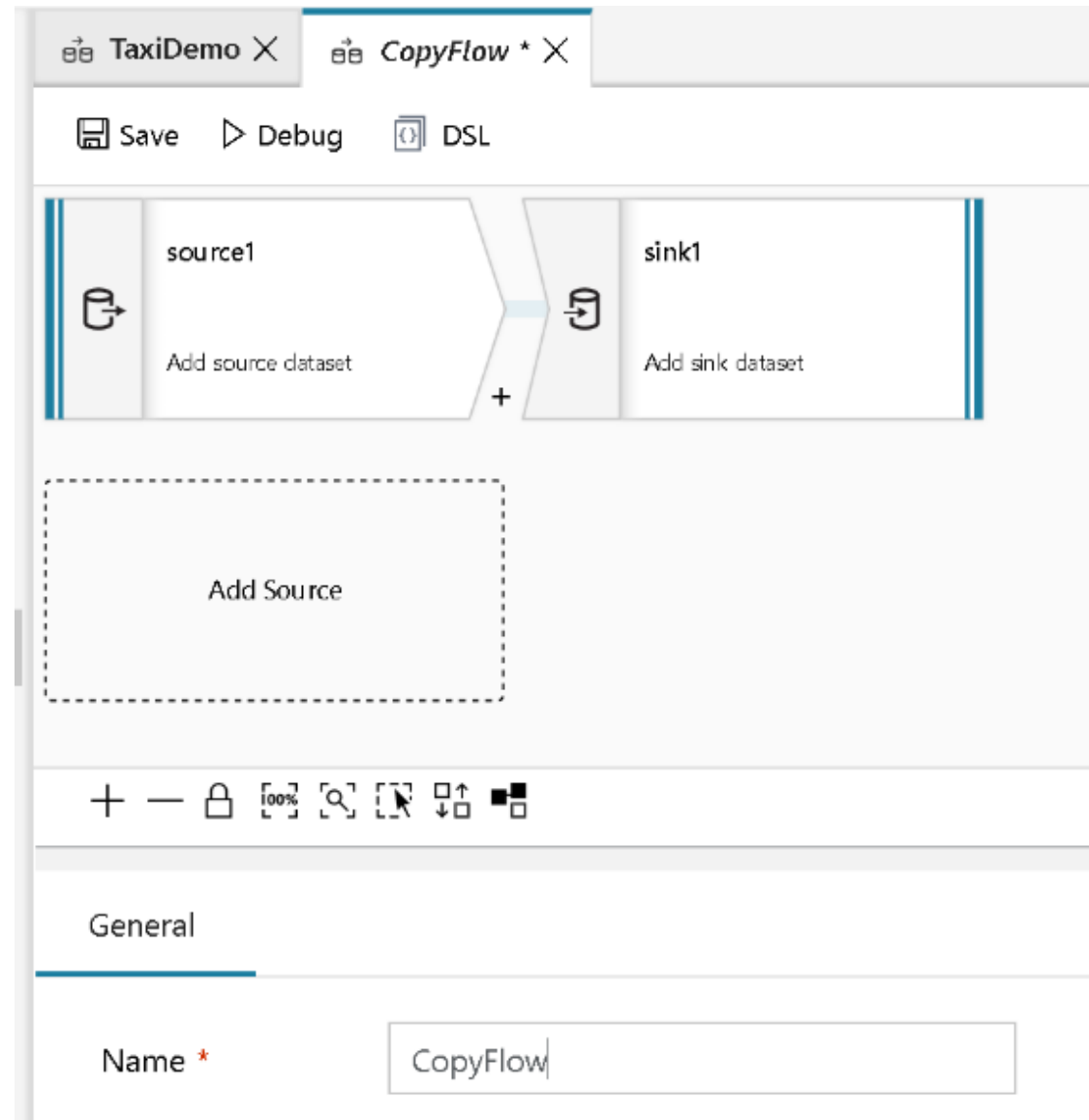
General External dependencies

Name *

users

Description

Simple Copy Flow



Guided experience to build data flows

The screenshot displays the Microsoft Azure Data Factory interface for a workspace named 'SQLPlayerDemo'. The left sidebar shows 'Factory Resources' with a search bar and a list of resources: Pipelines (5), Datasets (16), and Data Flows (Preview) (6). Under 'Data Flows (Preview)', the 'usersql' flow is selected, showing a sub-tree with 'Beta' (2), 'StackOverflow' (3), 'badgesGroupByName', 'badgesGroupByName2', and 'users'. The main canvas shows a data flow pipeline with the following steps: 'source1' (Columns: 13 total), 'Select1' (Renaming source1 to Select1 with columns 'Id, Age, DisplayName, DownVotes'), 'FilterByReputation' (Filtering rows using expressions on columns 'Reputation'), 'GroupByLocation' (Aggregating data by 'Location' producing columns 'Reputation, DownVotes, Views'), 'Sort1' (Sorting rows on columns 'Location'), 'Filter1' (Filtering rows using expressions on columns 'Location, Location'), and 'sink1' (Export data to AzureBlob2). A context menu is open over the 'source1' step, listing options: 'Multiple inputs/outputs', 'New Branch', 'Join', 'Conditional Split', 'Union', 'Lookup', 'Schema modifier' (with sub-options: 'Derived Column', 'Aggregate', 'Surrogate Key', 'Pivot', 'Unpivot', 'Window'), and 'Row modifier'. The bottom panel shows the 'Source Settings' for 'source1', including 'Output stream name' (source1), 'Source Dataset' (stack_users), 'Options' (Allow schema drift checked), 'Sampling' (Enable selected), and 'Rows limit' (1000).

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

Setting Inspect Error log






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

Output schema Data Preview







	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: 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

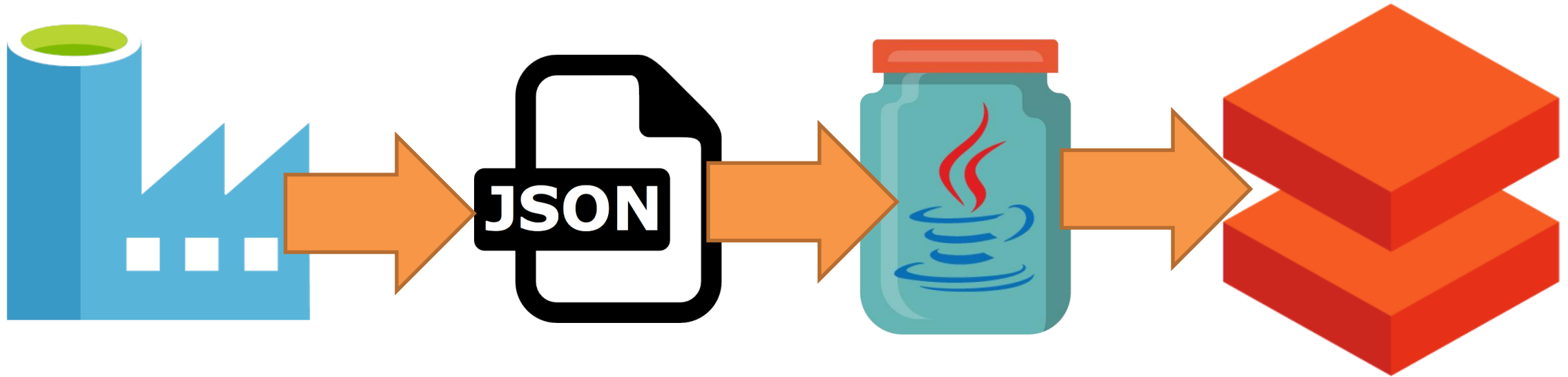
-  Extend

Destination

-  Sink

* With some small exceptions

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, CPU, 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)

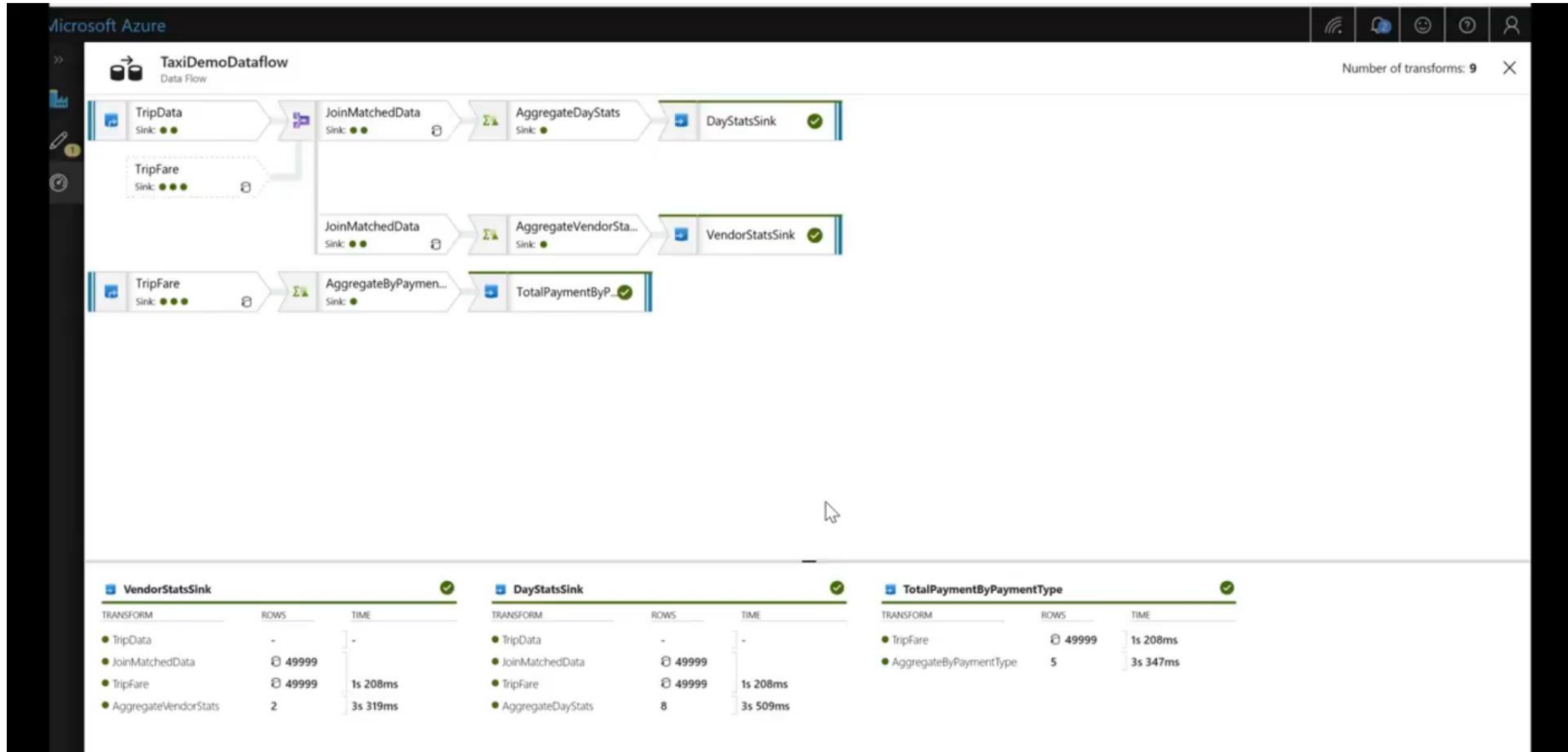
5.0 (includes Apache Spark 2.4.0, Scala 2.11)

Min Workers: 2 Max Workers: 8



DEMO TIME

Data Flow Execution Plan



Data Flow Data Lineage

Microsoft Azure

TaxiDemoDataflow
Data Flow

Number of transforms: 9

VendorStatsSink
Sink

Column	Method	Original Source
passenger_count	Calculated	TripData
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)

Stream information

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

Partition chart

Row count

1 Partition

Skewness

Kurtosis

Edit transformation

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



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