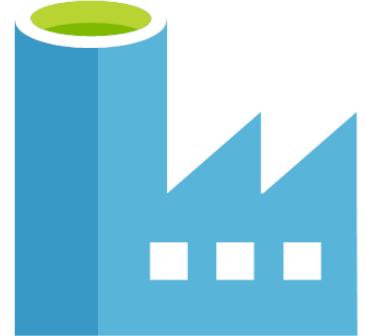




Azure Data Factory v2 with Mapping Data Flow (first blood)



Microsoft®
Most Valuable
Professional

Kamil Nowiński

Principal Microsoft Consultant

altius

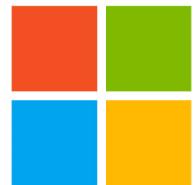


Azure Data Factory with Mapping Data Flow (first blood)

altius @NowinskiK



PLATINUM



Microsoft

GOLD



adatis



dbWatch
DATABASE CONTROL

Quest®



PYRAMID
ANALYTICS



ZAP

KNOW YOUR BUSINESS

BRONZE



coeo
Making SQL sense



DLM
Consultants



Beacon
INTELLIGENCE



redgate

DATA RELAY

@DataRelay_UK

#DataRelay

DataRelay.co.uk

Thank you to our sponsors. We couldn't do it without you!



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

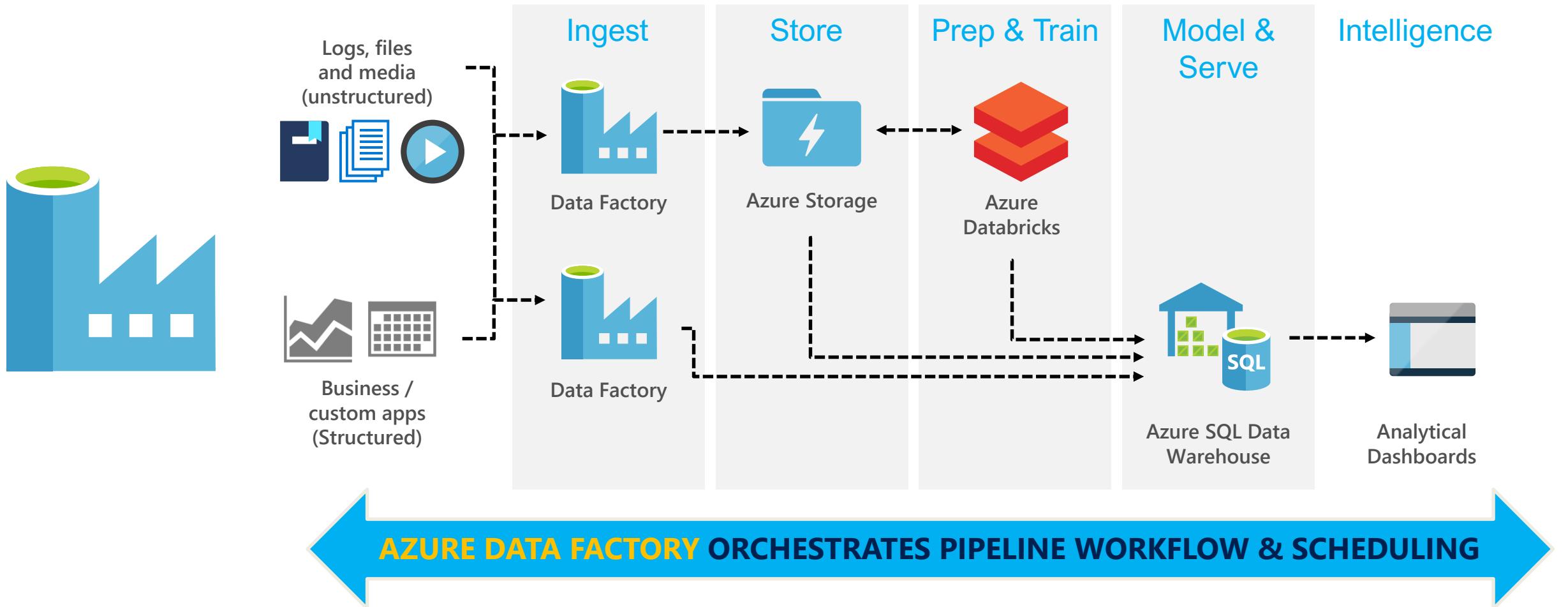


www.SQLPlayer.net

PODCAST – interviews with...



What the Azure Data Factory is?

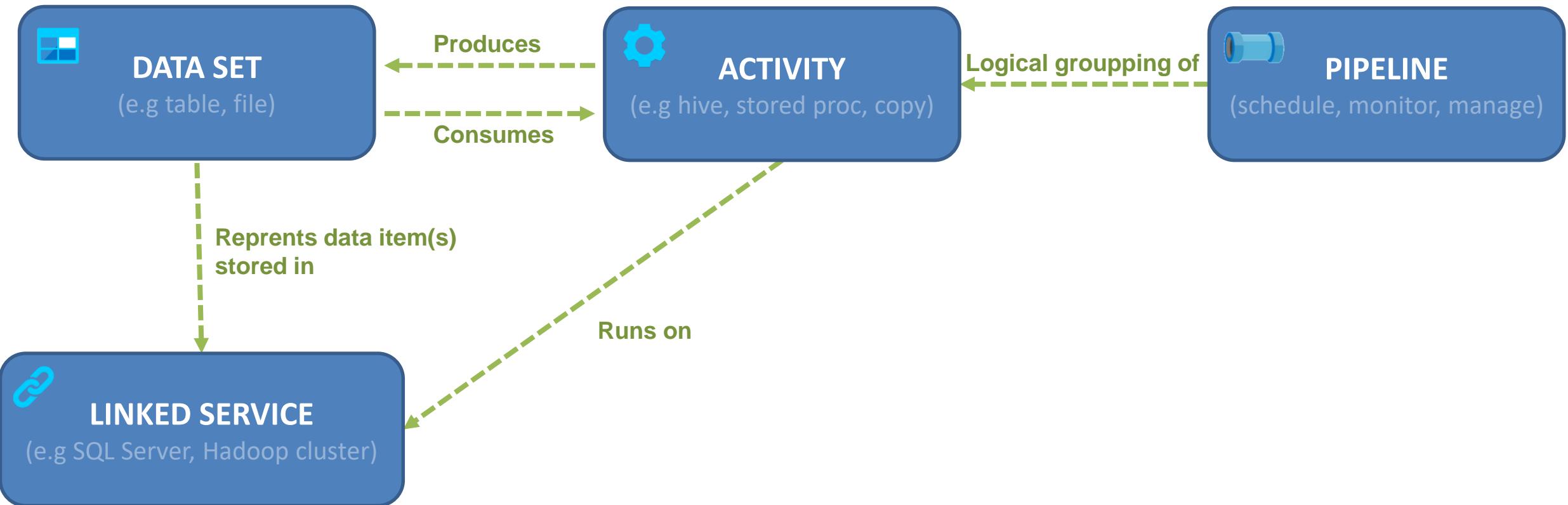


Access all your data

- 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	SFTP	Dynamics 365	Salesforce	REST *			
Data Lake Storage Gen2	Greenplum	Presto	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				
		Supported as Source and Sink								
		Supported as Source only								
		Supported as Sink only								

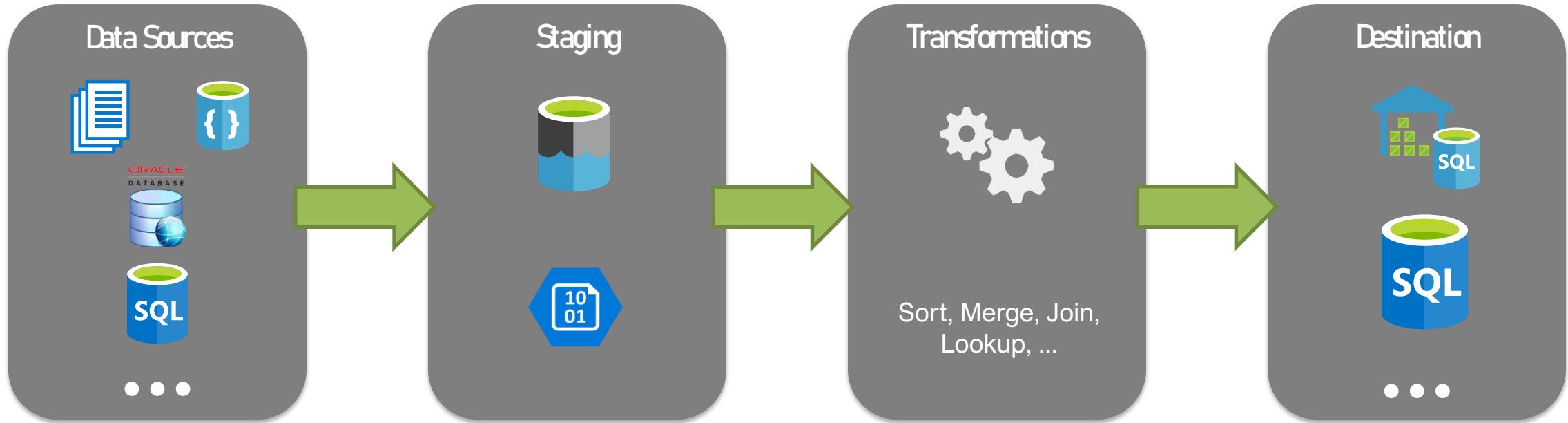
ADF Key Concepts



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, ...)

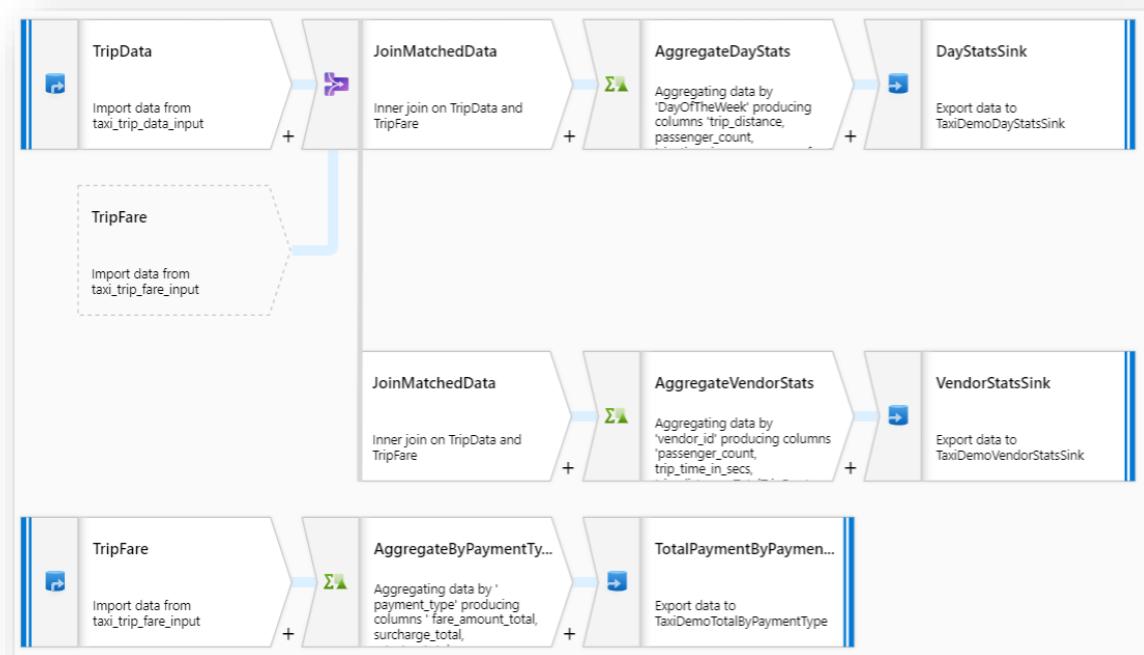
- Explicit user action
- User places transformations on design surface, from toolbox
- User must set properties for transformation steps and step connectors

- 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



Azure Data Factory with Mapping Data Flow (first blood)

... not

```
File MovieRecommendation4EDemo.txt
1
2 HDFS Cluster Details:
3 Adfhd1.azurehdinsight.net
4 Admin
5 Adfg123456
6
7 Storage:
8 adfhdstorage
9 /anyPw6G1j7z8tjBWhm1so/YGdyG74d-S1JAr+sN7bJgb954705gUChLokzI9UXct4OxZo8xIKHdMKw==_
10
11 Cluster Remote Login Details:
12 Adf
13 India@1234
14
15 HiveQuery:
16 DROP TABLE IF EXISTS MovieRatings;
17 CREATE EXTERNAL TABLE MovieRatings
18 (
19   UserID int,
20   MovieID int,
21   Rating int,
22  TimeStamp string
23 ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '10' STORED AS TEXTFILE LOCATION '${hiveconf:MovieRatings}';
24
25 DROP TABLE IF EXISTS MovieTitles;
26 CREATE EXTERNAL TABLE MovieTitles
27 (
28   MovieID int,
29   MovieName string
30 ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '10' STORED AS TEXTFILE LOCATION '${hiveconf:MovieTitles}';
```

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

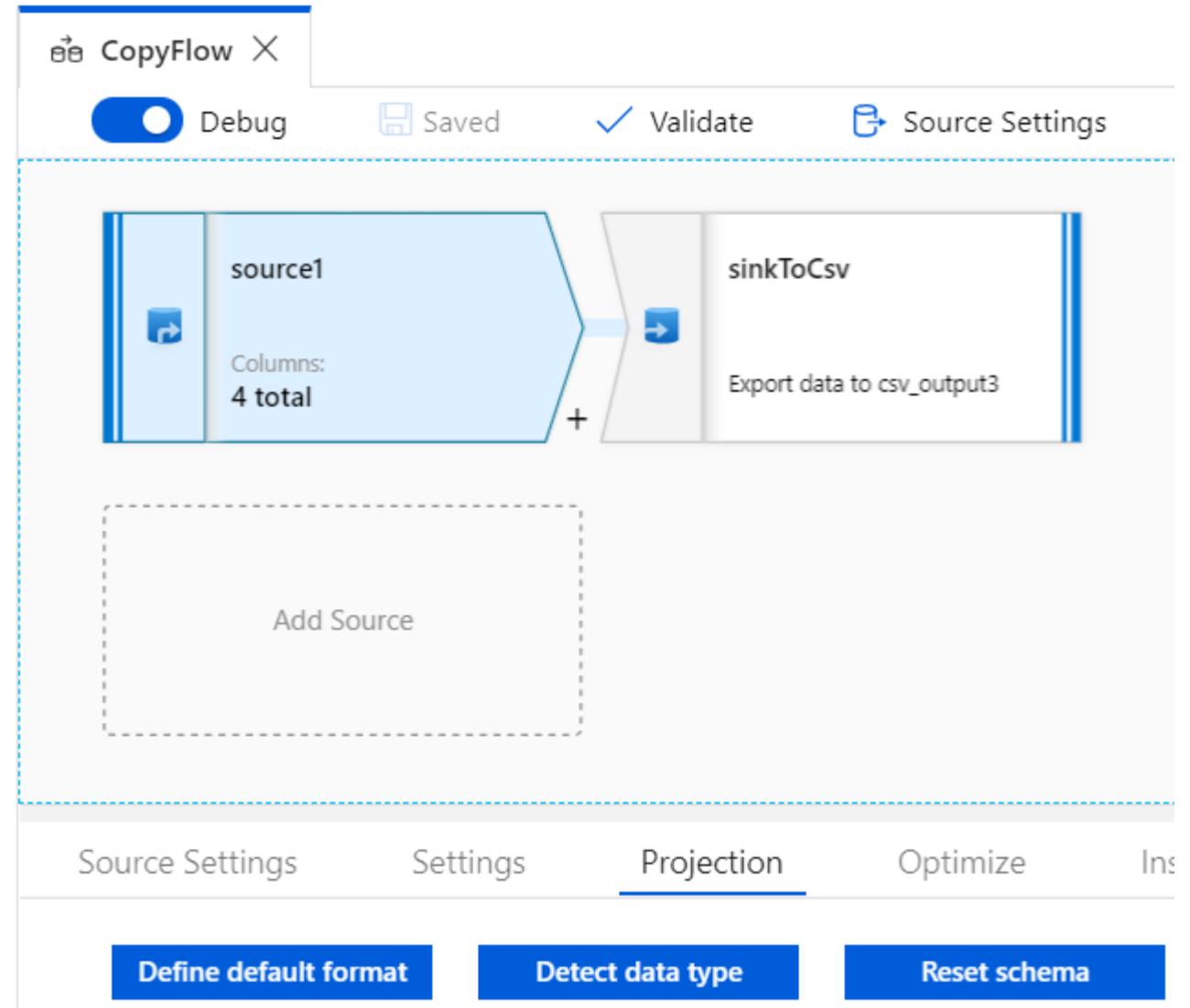
The screenshot shows the Microsoft Azure Data Factory v2 interface. At the top, the navigation bar displays "Microsoft Azure | Data Factory > SQLPlayerDemo2". To the right of the navigation is a search bar labeled "Search resources". Below the navigation, there are several action buttons: "Data Factory" (dropdown), "Publish All", "Validate All" (with a checkmark), "Refresh", and "Discard All".

The main area is titled "Factory Resources" with a dropdown arrow and a magnifying glass icon. A search bar below it says "Filter resources by name". To the right of the search bar is a "+" button. The "Factory Resources" section lists four categories:

- Pipelines: 2 items
- Datasets: 12 items
- Data Flows (Preview): 5 items (highlighted with an orange rounded rectangle and an orange arrow pointing to it)

On the far left, there is a vertical sidebar with three icons: a blue square with a chart, a pencil, and a red circle with a play button.

Simple Copy Flow



Mapping Data Flow: Components = Actions *

Multiple inputs/outputs

 New branch

 Join

 Conditional Split

 Exists

 Union

 Lookup

Schema modifier

 Derived Column

 Select

 Aggregate

 Surrogate Key

 Pivot

 Unpivot

 Window

Row modifier

 Filter

 Sort

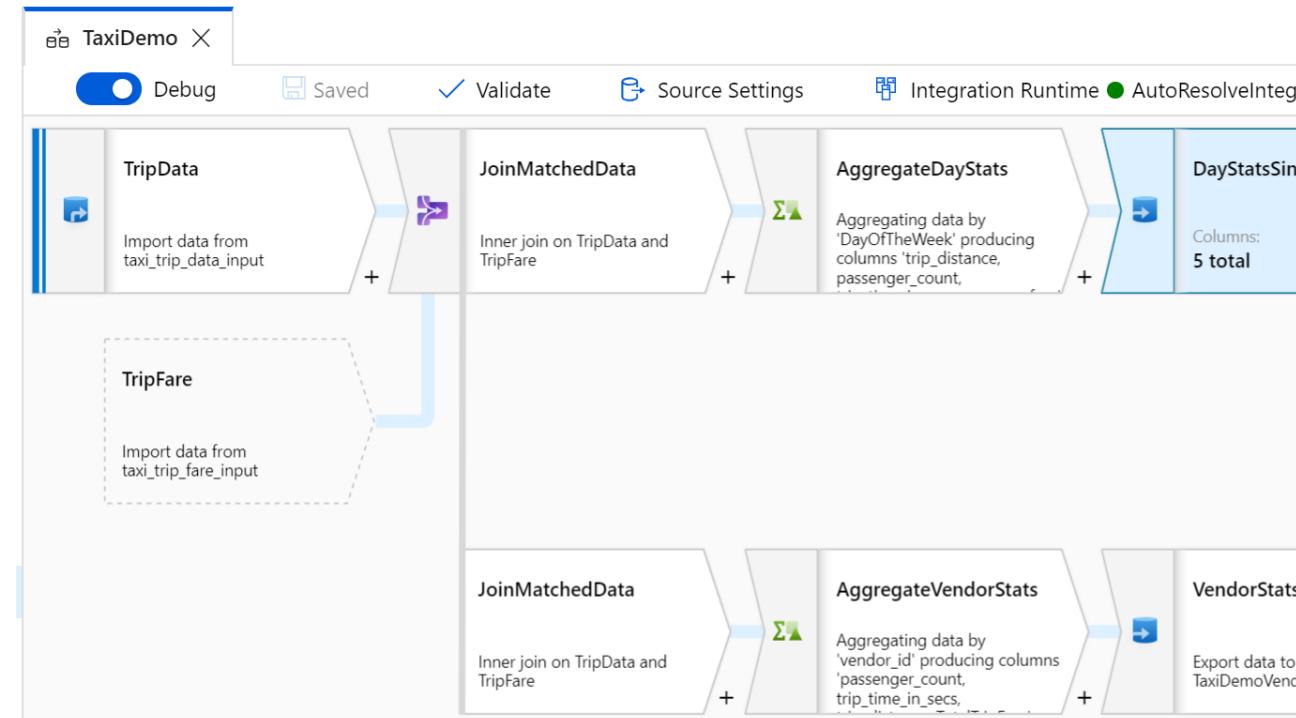
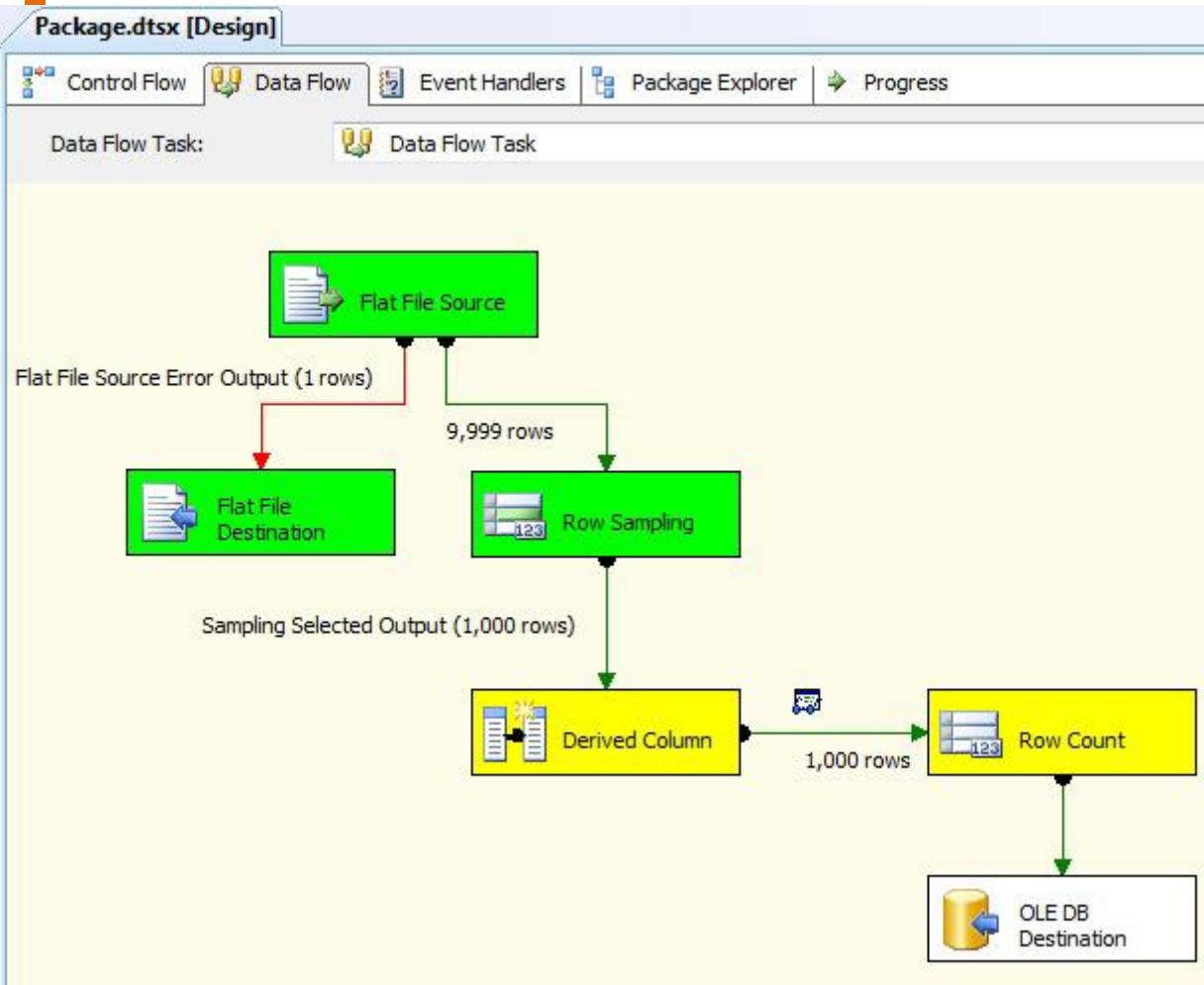
 Alter Row

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)

Microsoft Azure

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 3 badgesGroupName badgesGroupName2 users

users X

Debug ✓ Validate

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'

General External dependencies

Name * users

Description

```
graph LR; subgraph users [ ]; direction LR; sourceUsers --> Select1; Select1 --> FilterByReputation; FilterByReputation --> GroupByLocation; GroupByLocation --> SortByLocation; SortByLocation --> Wrong; SortByLocation --> AllRight; end; sourceUsers["sourceUsers<br/>Import data from Users_BlobCsv"]; Select1["Select1<br/>Renaming sourceUsers to Select1 with columns 'DisplayName, DownVotes, LastAccessDate, Location'"]; FilterByReputation["FilterByReputation<br/>Filtering rows using expressions on columns 'Reputation'"]; GroupByLocation["GroupByLocation<br/>Aggregating data by 'Location' producing columns 'SumOfReputation, SumOfViews, Count'"]; SortByLocation["SortByLocation<br/>Sorting rows on columns 'Location'"]; Wrong["Wrong<br/>Conditionally distributing the data in 2 groups, based on columns 'Location, Location, Location, Location, Location'"]; AllRight["AllRight<br/>Conditionally distributing the data in 2 groups, based on columns 'Location, Location, Location, Location, Location'"];
```

Guided experience to build data flows

The screenshot shows the Microsoft Azure Data Factory Data Flow designer interface. A pipeline named "usersql" is displayed, consisting of several components: source1, Select1, FilterByReputation, GroupByLocation, Sort1, Filter1, and sink1. The "source1" component is currently selected, and a context menu is open over it, highlighted with a red box. The menu includes options like "Multiple inputs/outputs", "New Branch", "Join", "Conditional Split", "Union", "Lookup", "Schema modifier" (with sub-options "Derived Column", "Aggregate", "Surrogate Key", "Pivot", "Unpivot", and "Window"), and "Row modifier". Below the menu, the "Source Settings" tab is active, showing the configuration for the "source1" stream. The "Output stream name" is set to "source1", and the "Source Dataset" is "stack_users". Other settings include "Allow schema drift" checked, "Sampling" set to "Enable", and a "Rows limit" of 1000.

Microsoft Azure | Data Factory > SQLPlayerDemo

Factory Resources <>

usersql X

source1

Select1

FilterByReputation

GroupByLocation

Sort1

Filter1

sink1

Search resources

Pipelines: 5

Datasets: 16

Data Flows (Preview): 6

usersql

Beta: 2

StackOverflow: 3

badgesGroupName

badgesGroupName2

users

Add Source

Multiple inputs/outputs

- New Branch
- Join
- Conditional Split
- Union
- Lookup
- Schema modifier
- Derived Column
- Aggregate
- Surrogate Key
- Pivot
- Unpivot
- Window

Source Settings

Output stream name * source1

Source Dataset * stack_users

Options Allow schema drift

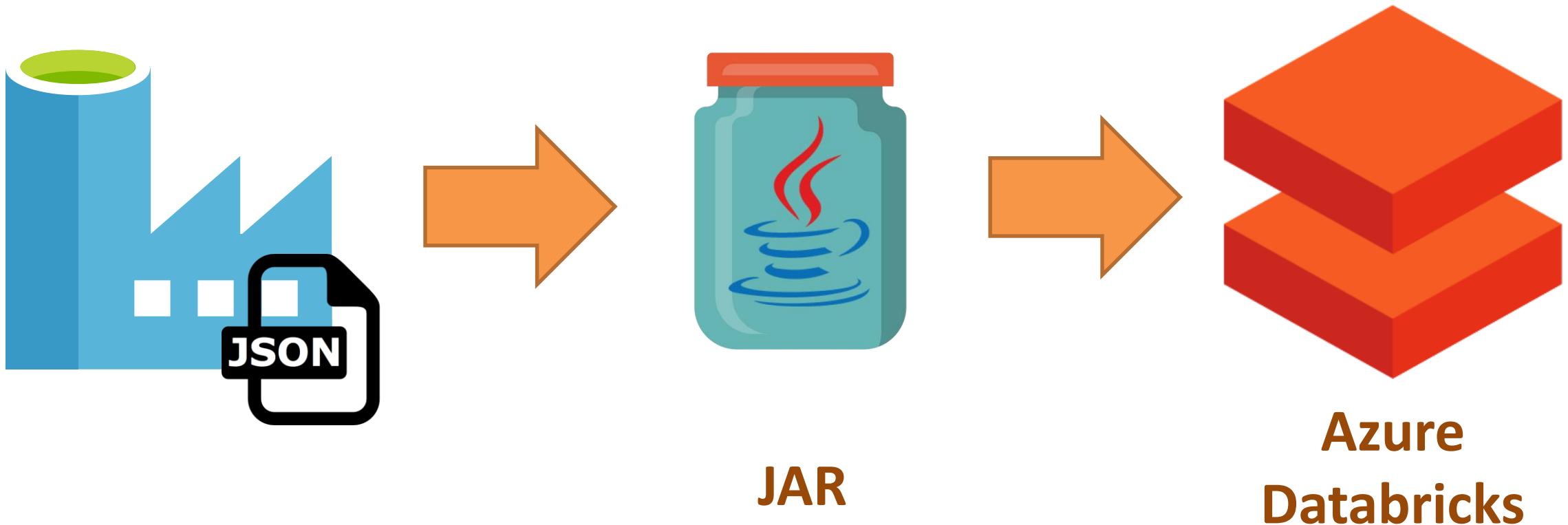
Sampling * Enable Disable

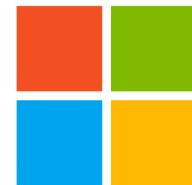
Rows limit 1000

Connections

Triggers

What is going on behind the scenes?





Microsoft

GOLD

 adatis

 dbWatch
DATABASE CONTROL

Quest®

 PYRAMID
ANALYTICS

 ZAP

KNOW YOUR BUSINESS

BRONZE


coeo
Making SQL sense


DLM
Consultants

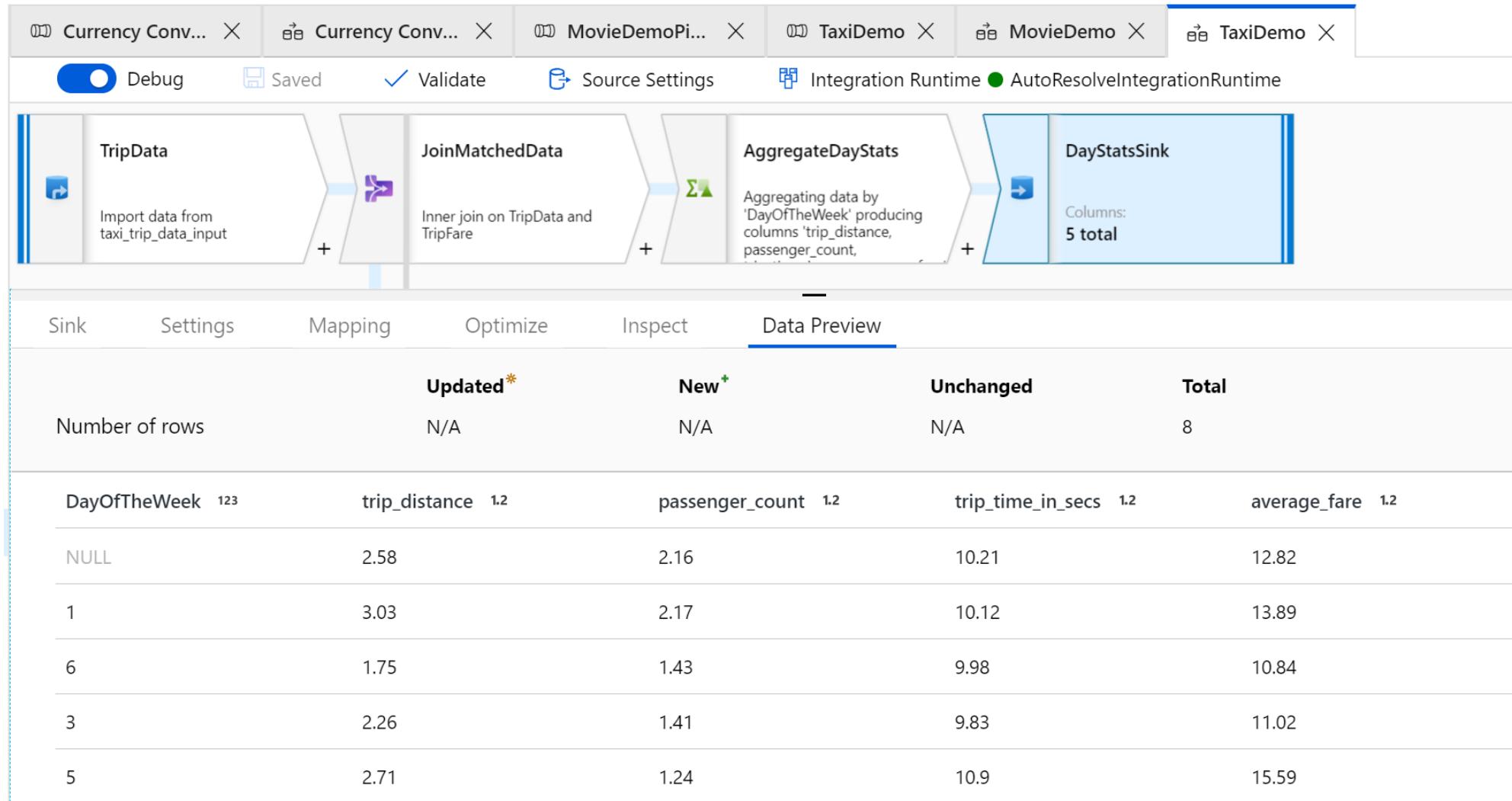

Beacon
INTELLIGENCE

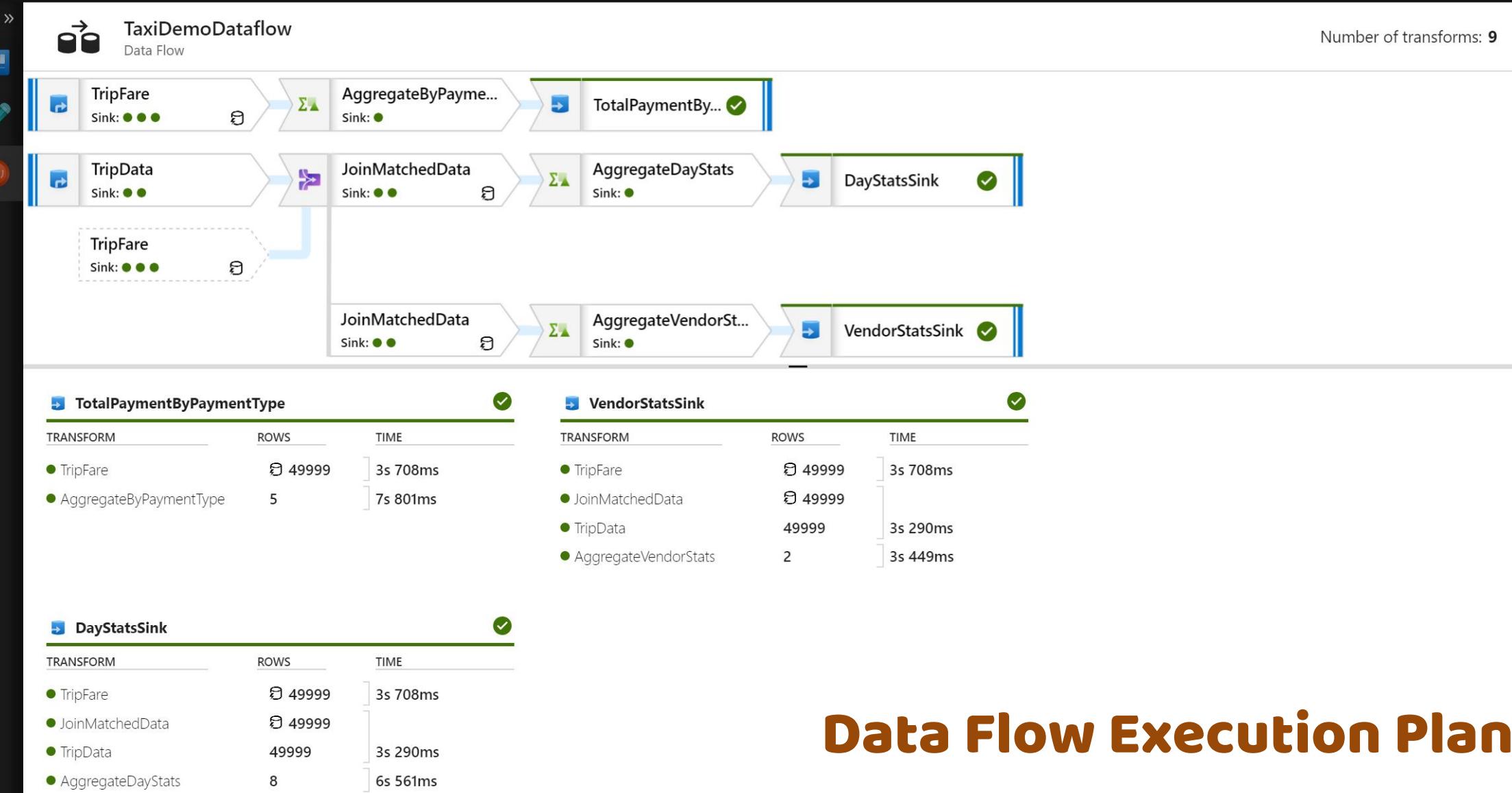

redgate

<https://datarelay.co.uk/Feedback>

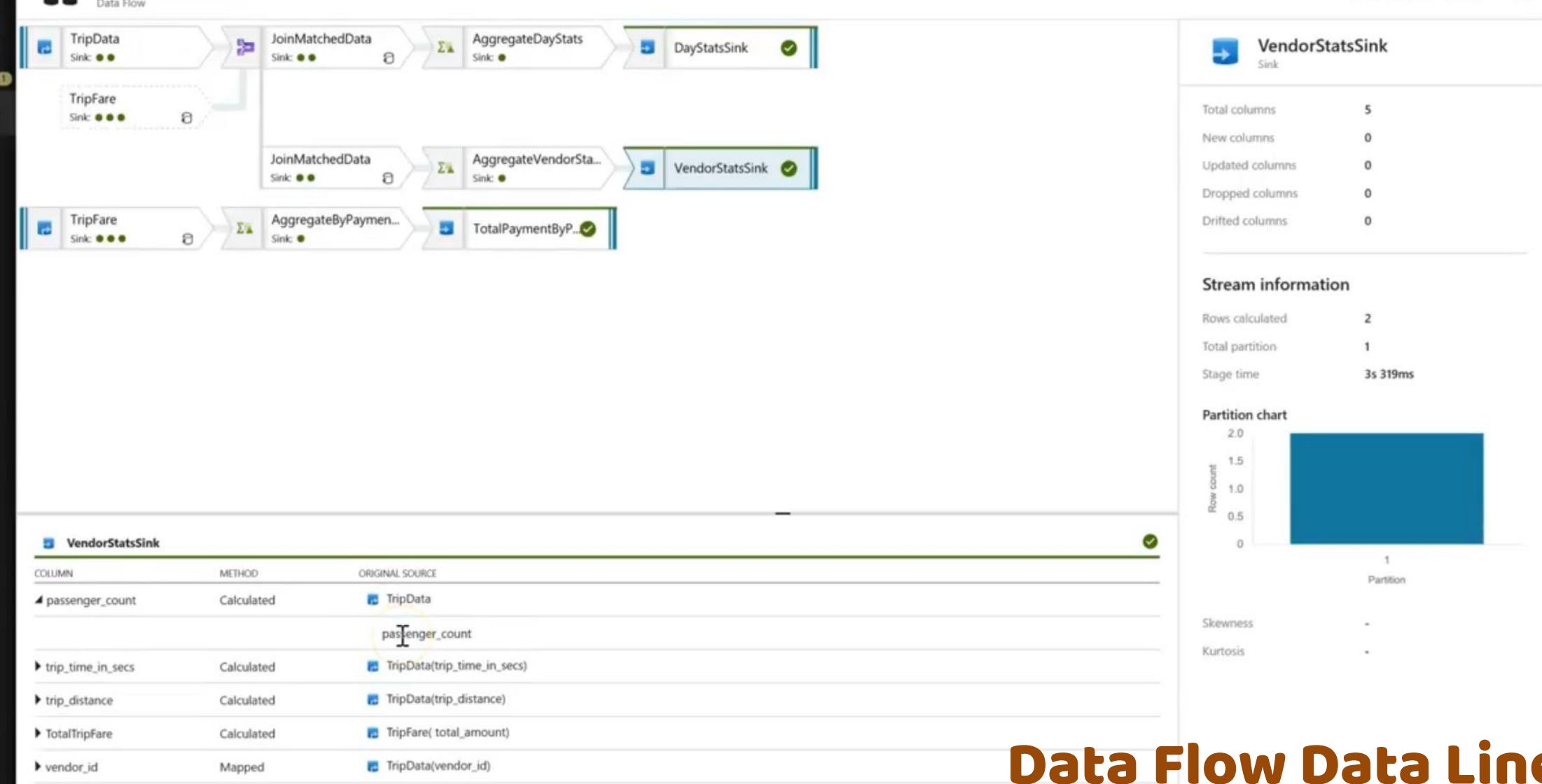
DEMO TIME

Data Preview in Debug mode





Data Flow Execution Plan



Data Flow Data Lineage

ADF Template Gallery

Template gallery X

Filter Reset all filter

Search templates

Categories

- Copy
- Data Flow
- SSIS
- Transform

Create by

- Microsoft
- My templates

Tag

All

Services used

All

Import template

Template Name	Description	Icon
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...	
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.	
Copy data from Google BigQuery to Azure Data Lake Store	Use this template to copy data from Google BigQuery to Azure Data Lake Storage.	
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.	
Copy data from Netezza to Azure Data Lake Store	Use this template to copy data from Netezza server to Azure Data Lake Storage.	
Copy data from on premise SQL Server to SQL Azure	Use this template to copy data from on premise SQL Server to SQL Azure.	
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.	
Copy data from Oracle to SQL Data Warehouse	Use this template to copy data from Oracle server to SQL Data Warehouse.	
Copy delta data from AWS S3 to Azure Data Lake Storage Gen2	Use this template to copy delta data from netabutes.	
Copy multiple files containers between File Stores	Use this template to copy all files from AWS S3 to	
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.	

Azure DevOps GIT

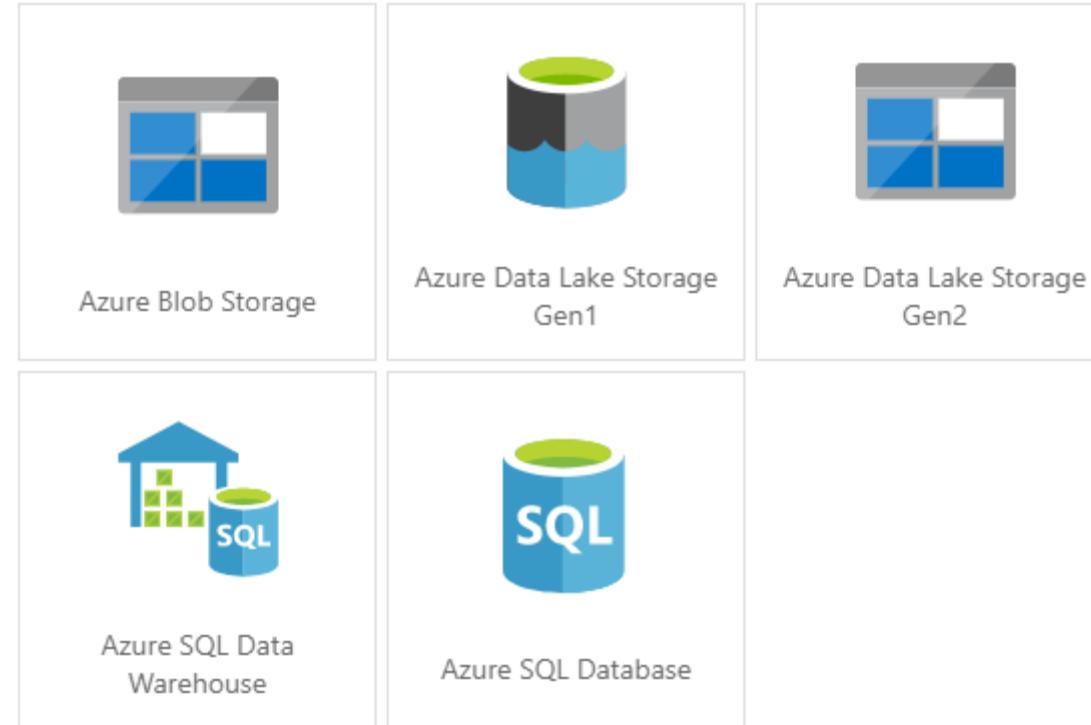
Factory Resources

Filter resources by name

Pipeline

Pipeline from template

Mapping Data Flow – Source & Sink



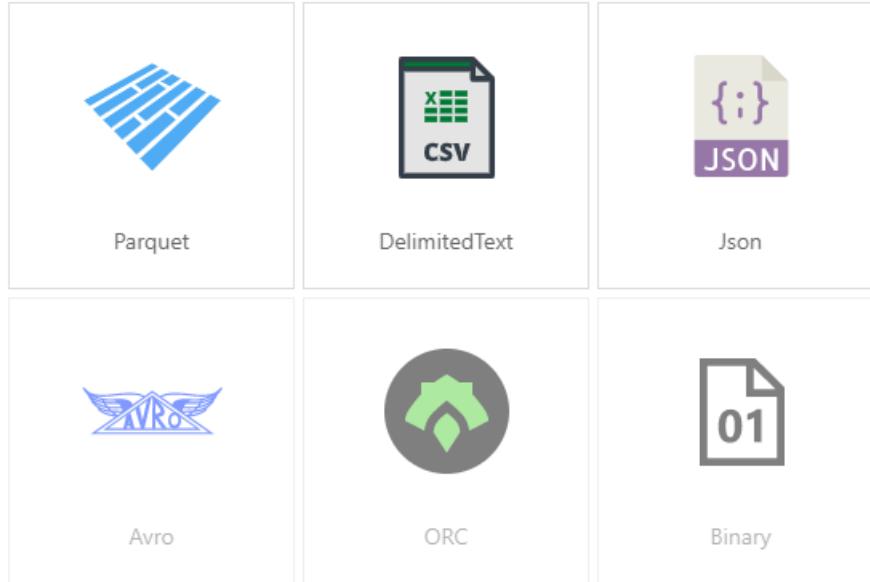
Mapping Data Flow – Source & Sink capabilities



- 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

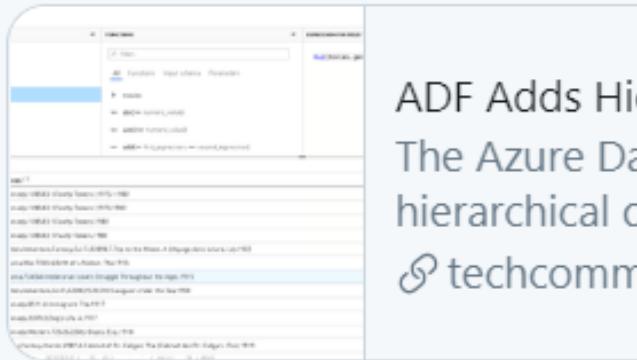


The screenshot shows the "Settings" tab of the Execute Data Flow transformation. It includes fields for "Data Flow" (set to "DF_DistinctRows"), "Run on (Azure IR)" (set to "AutoResolveIntegrationRuntime"), "Compute type" (set to "General Purpose"), and "Core count" (set to "4 (+ 4 Driver cores)"). The "Time to live" field is set to "0 minutes". A dropdown menu for "Time to live" shows options: "Filter...", "0 minutes", "10 minutes", "30 minutes", "1 hour", and "4 hours". An orange arrow points from the "Run on (Azure IR)" dropdown to the "Time to live" dropdown, highlighting the relationship between the two settings.

Latest updates? Go Twitter!



Mark Kromer @KromerBigData ·
#Azure #datafactory has released
capabilities via #mappingdataflow
in Data Flow, build and manage co
hierarchical data.



8



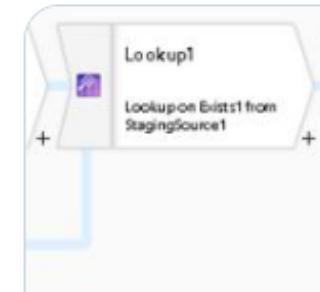
12



Azure Data Factory
@DataAzure

Azure Data Factory Mapping Data Flows are now generally available

#Azure #DataFactory #mappingdataflows



Azure Data Factory Mapping Data Flows are now generally av...
In today's data-driven world, big data processing is a critical task for every organization. To unlock transformational insight...
[azur...
e.microsoft.com](https://azure.microsoft.com)

4:34 pm · 7 Oct 2019 · Twitter for Android

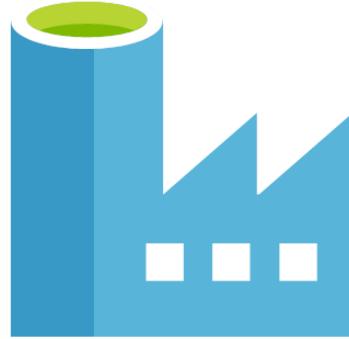
SSIS vs ADF activities vs T-SQL

Activity	Description	SSIS equivalent	SQL Server equivalent
 New branch	Create a new flow branch with the same data	 Multicast (+icon)	<code>SELECT INTO SELECT OUTPUT</code>
 Join	Join data from two streams based on a condition	 Merge join	<code>INNER LEFT RIGHT JOIN, CROSS FULL OUTER JOIN</code>
 Conditional Split	Route data into different streams based on conditions	 Conditional Split	<code>SELECT INTO WHERE condition1 SELECT INTO WHERE condition2 CASE ... WHEN</code>
 Union	Collect data from multiple streams	 Union All	<code>SELECT colla UNION (ALL) SELECT collb</code>
 Lookup	Lookup additional data from another stream	 Lookup	<code>LEFT RIGHT JOIN</code>
 Derived Column	Compute new columns based on the existing once	 Derived Column	<code>SELECT Column1 * 1.09 as NewColumn</code>
 Aggregate	Calculate aggregation on the stream	 Aggregate	<code>SELECT Year(DateOfBirth) as Year, MIN(), MAX(), AVG() GROUP BY Year(DateOfBirth)</code>

<http://bit.ly/ADFDFvsSSIS>

<http://bit.ly/ADFDF-CheatSheet>

Resources



<http://sqlplayer.net/ADF>

Q&A



Thank you!



kamil@nowinski.net



@NowinskiK



SQLPlayer.net



<https://github.com/NowinskiK/CommunityEvents>



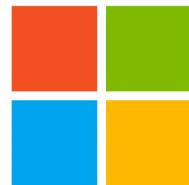
Kamil Nowinski

Microsoft Data Platform MVP

MCSE Data Platform & MCSE Data Management and Analytics



PLATINUM



Microsoft

GOLD



adatis



dbWatch
DATABASE CONTROL

Quest®



PYRAMID
ANALYTICS



ZAP

KNOW YOUR BUSINESS

BRONZE



coeo
Making SQL sense



DLM
Consultants



Beacon
INTELLIGENCE



redgate

@DataRelay_UK

#DataRelay

DataRelay.co.uk