



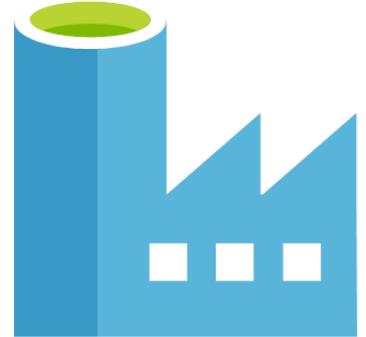
Thank you to
our Sponsors

claranet





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



Kamil Nowiński

Principal Microsoft Consultant

altius



Azure Data Factory with Mapping Data Flow (first blood)

altius @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



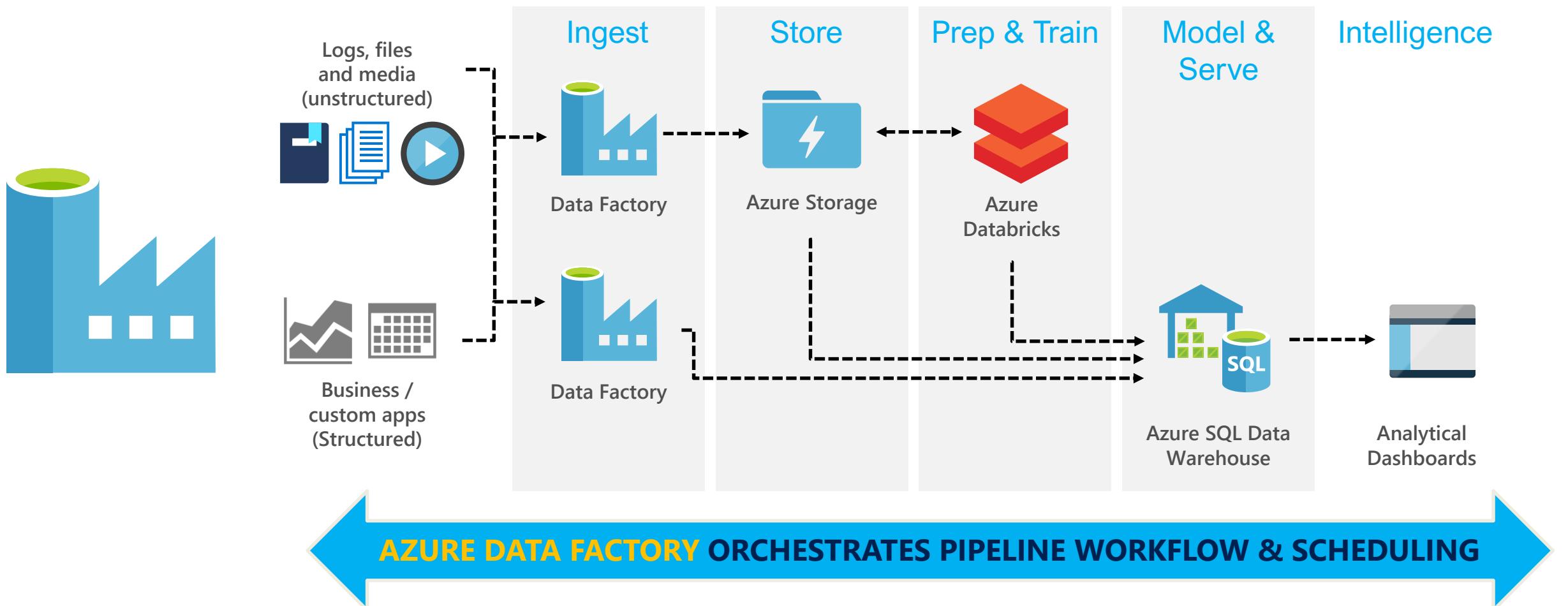
www.SQLPlayer.net

PODCAST – interviews with...

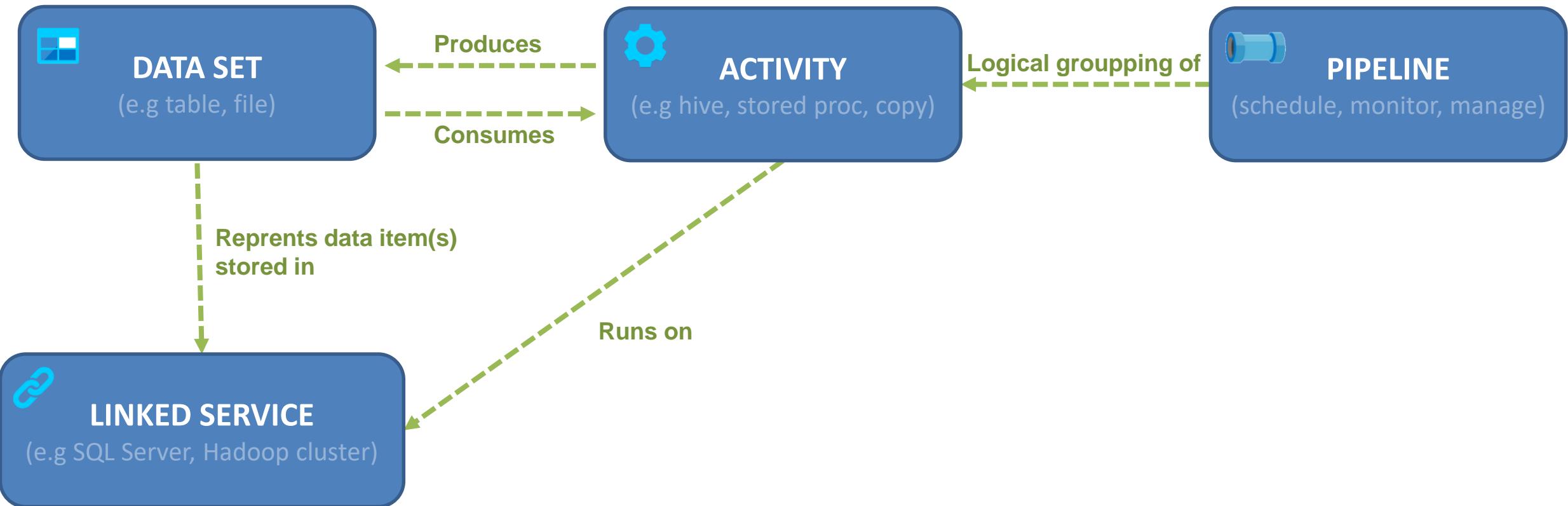


Scan me

What the Azure Data Factory is?



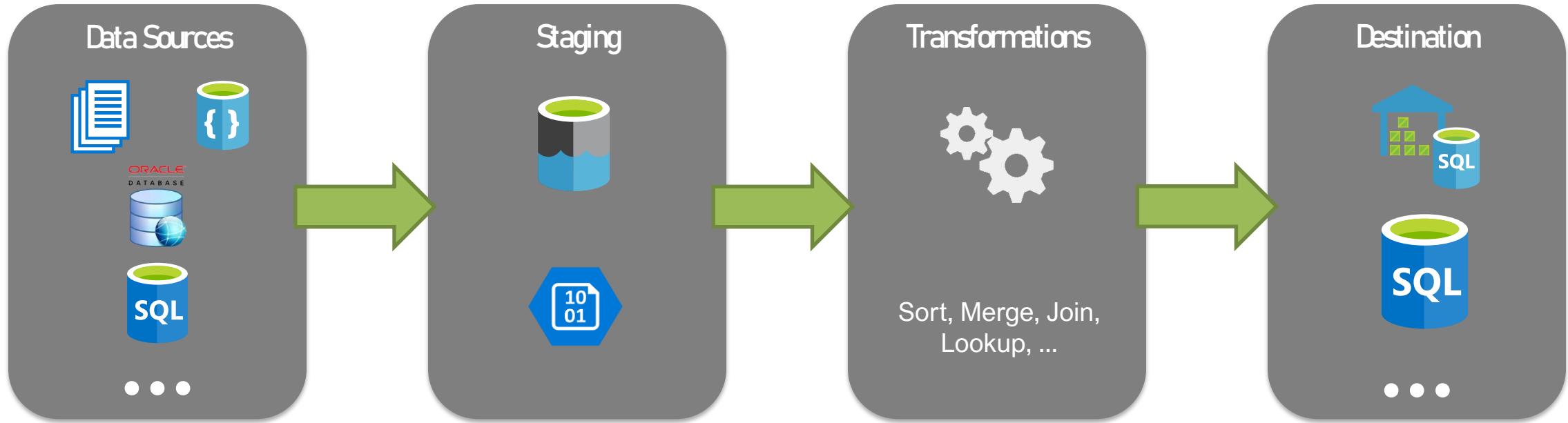
ADF Key Concepts



Visual Data Transformations with

MAPPING DATA FLOW

What the hell (Mapping) Data Flows are?

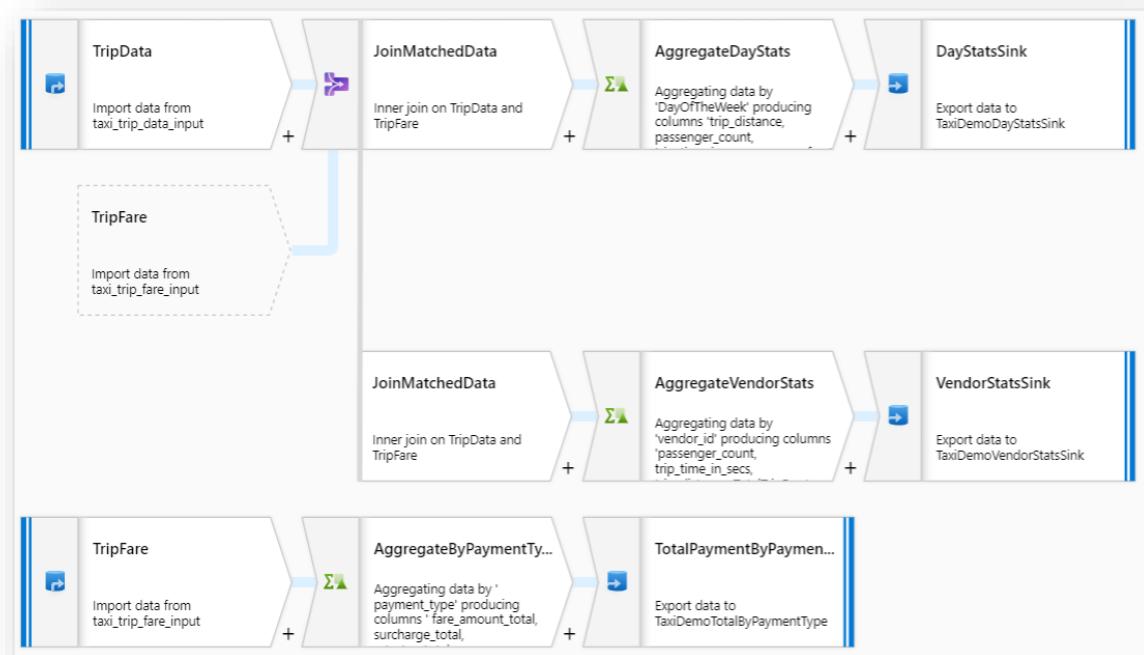


- Explicit user action
- User places data source(s) on design surface, from toolbox
- Select explicit sources
- Implicit/Explicit
- Data Lake staging area as default
- User does not need to configure this manually
- Advanced feature to set staging area options
- File formats/types:
(Parquet, JSON, txt, 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 '\n' STORED AS TEXTFILE LOCATION '${hiveconf:MovieRatings}';
24
25 DROP TABLE IF EXISTS MovieTitles;
26 CREATE EXTERNAL TABLE MovieTitles
27 (
28   MovieID int,
29   MovieName string
30 ) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE LOCATION '${hiveconf:MovieTitles}';
```

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

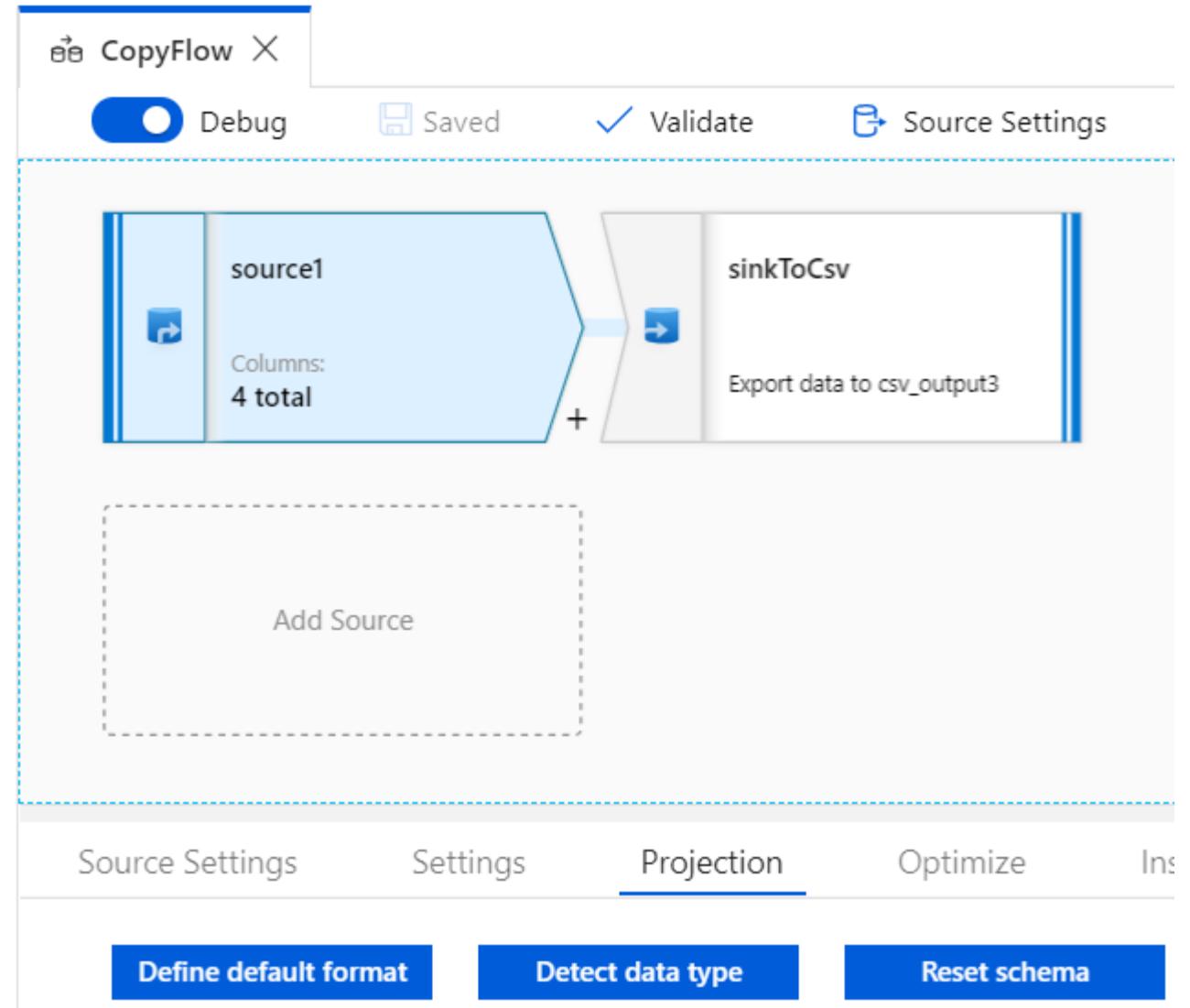
The screenshot shows the Microsoft Azure Data Factory v2 interface. At the top, the navigation bar displays "Microsoft Azure | Data Factory > SQLPlayerDemo2". To the right of the navigation is a search bar labeled "Search resources". Below the navigation, there are several action buttons: "Data Factory" (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, a vertical sidebar contains icons for Data Factory, Pipelines, Datasets, and Data Flows (Preview). Above the sidebar, there are three tabs: "CopyFlow X", "users X", and "dstUsersBlob X". Below these tabs are buttons for "Debug" (blue toggle switch), "Validate" (checkmark), and "Source Settings".

Simple Copy Flow

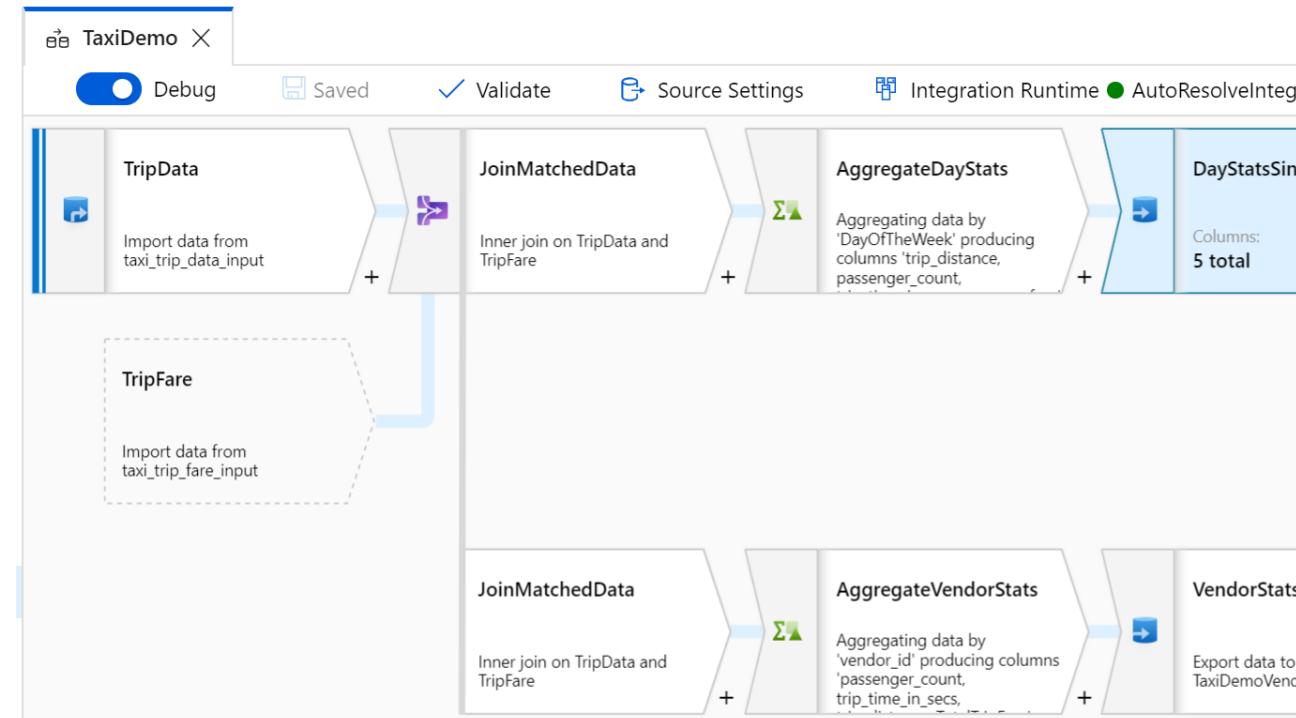
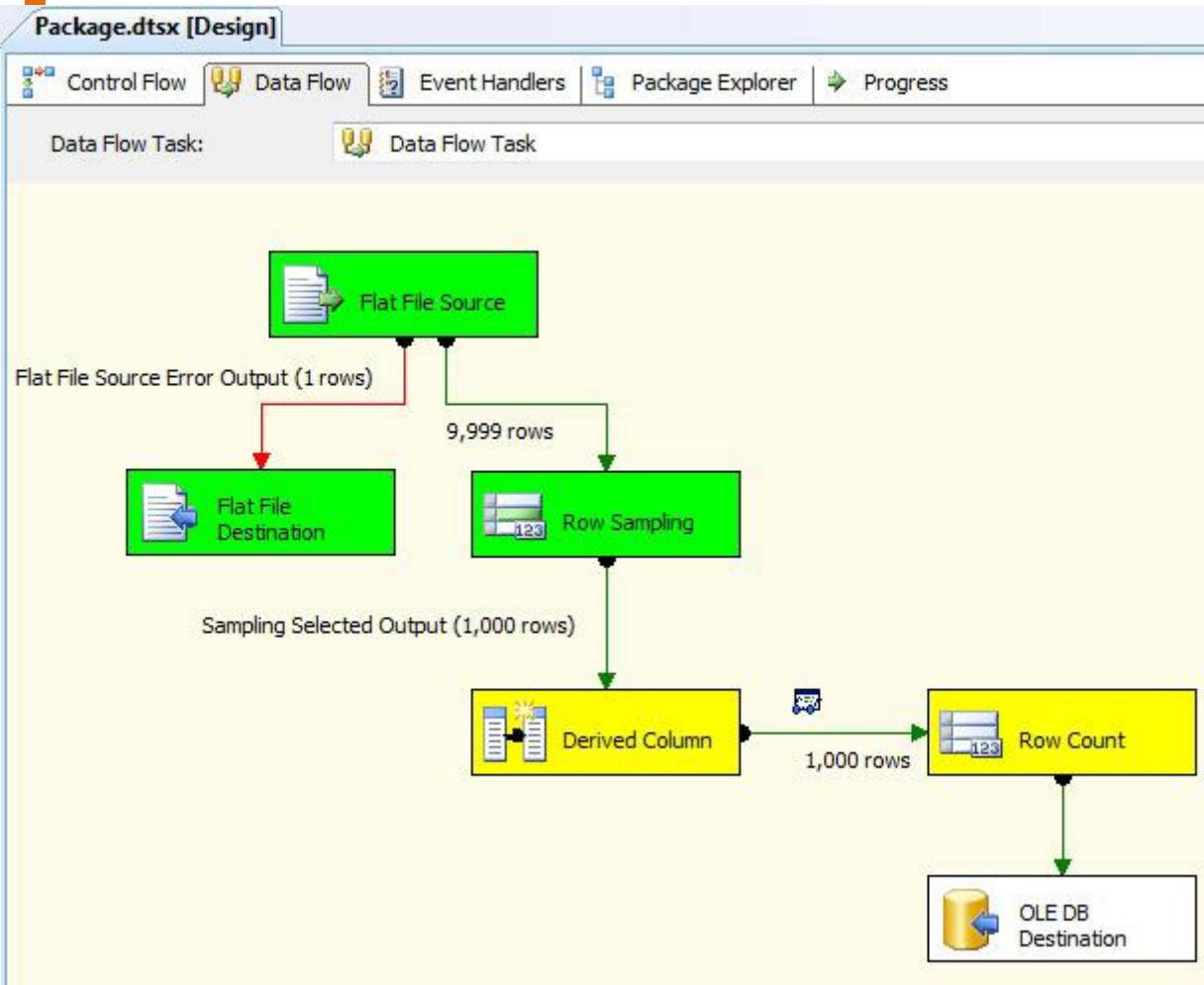


Mapping Data Flow: Components = Actions *

Multiple inputs/outputs	Schema modifier	Row modifier	Destination
 New branch	 Derived Column	 Filter	 Sink
 Join	 Select	 Sort	
 Conditional Split	 Aggregate	 Alter Row	
 Exists	 Surrogate Key		
 Union	 Pivot		
 Lookup	 Unpivot		
	 Window		

* 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 [users]; direction LR; sourceUsers --> Select1; Select1 --> FilterByReputation; FilterByReputation --> GroupByLocation; GroupByLocation --> SortByLocation; SortByLocation --> Wrong; end; AllRight[AllRight] --- Wrong;
```

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

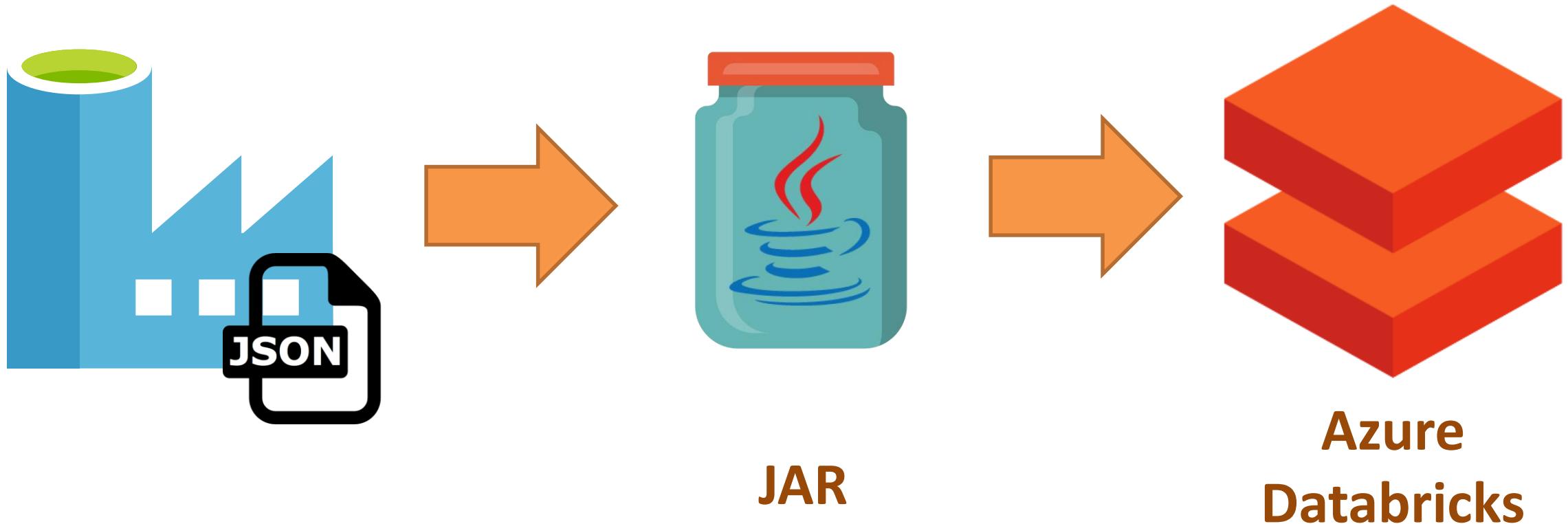
Sampling * Enable Disable

Rows limit 1000

Connections

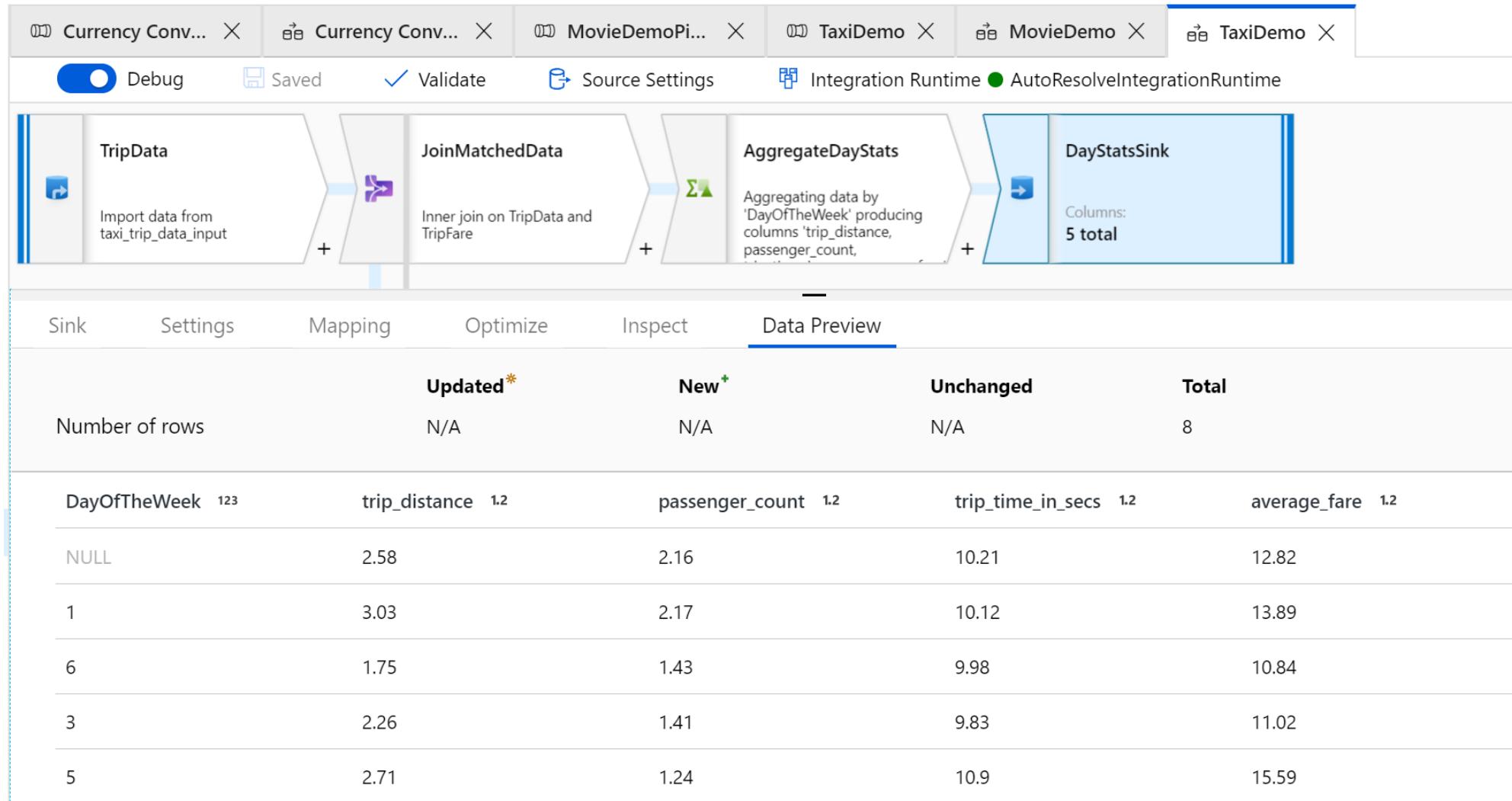
Triggers

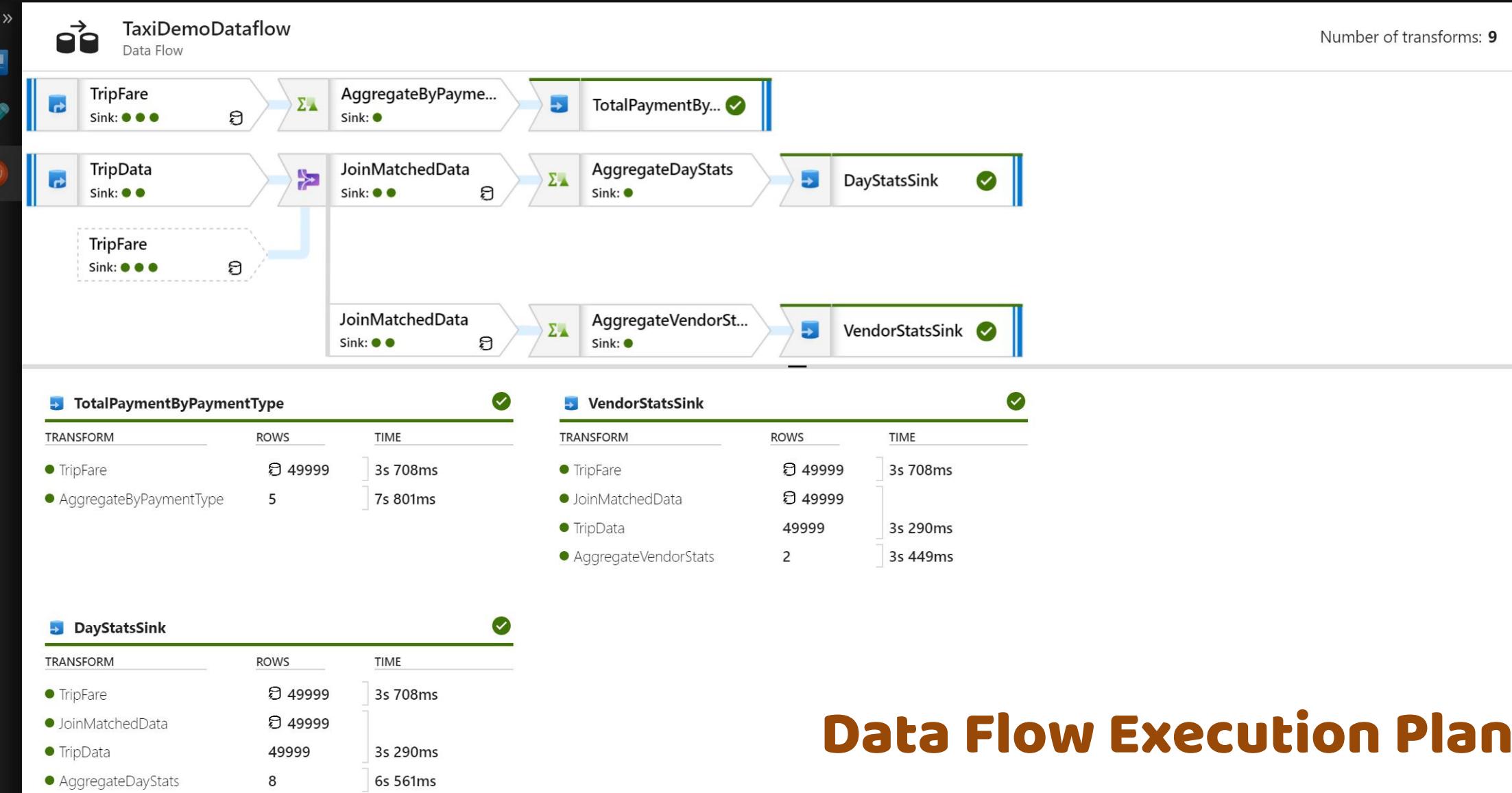
What is going on behind the scenes?



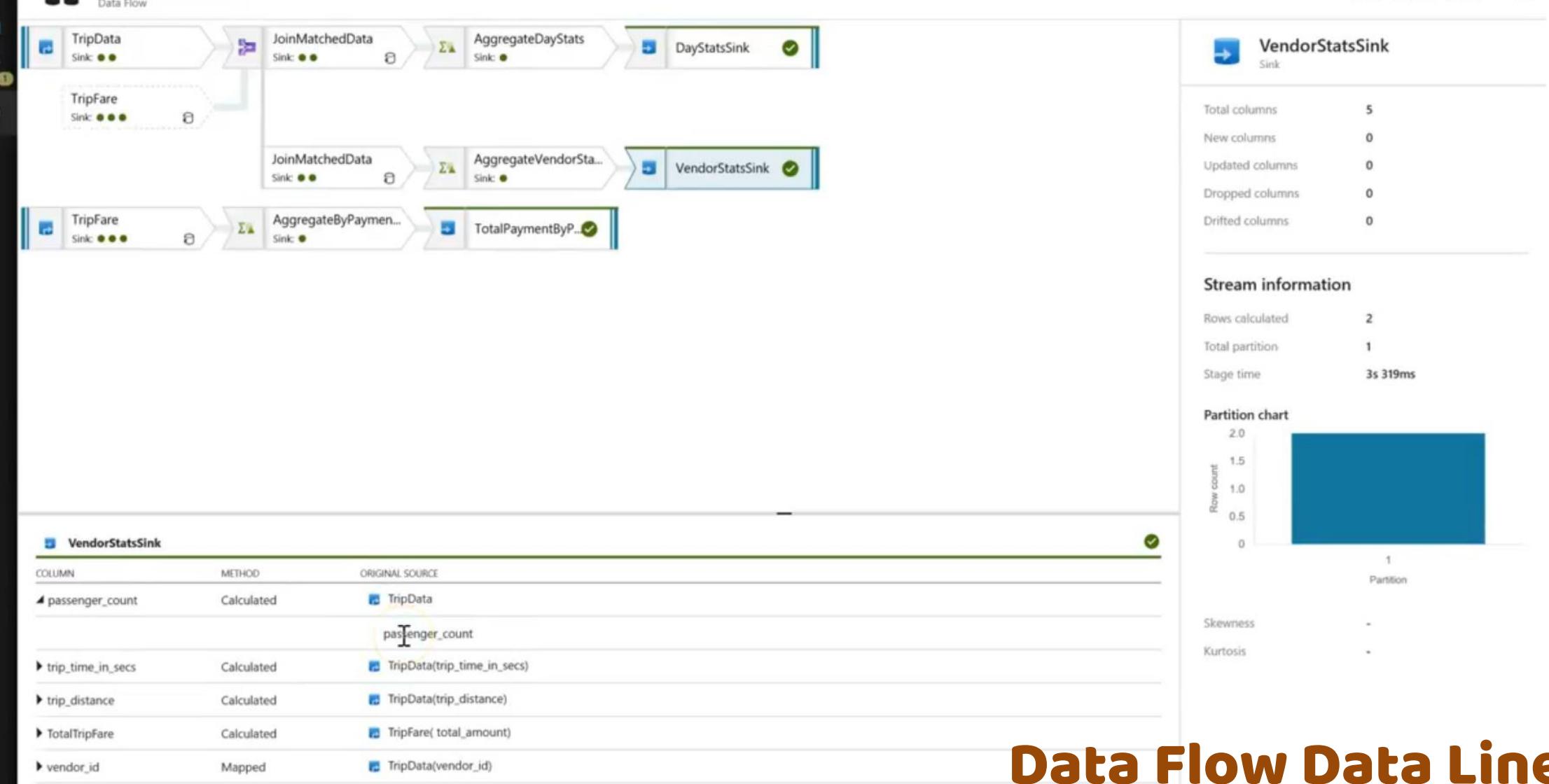
DEMO TIME

Data Preview in Debug mode





Data Flow Execution Plan



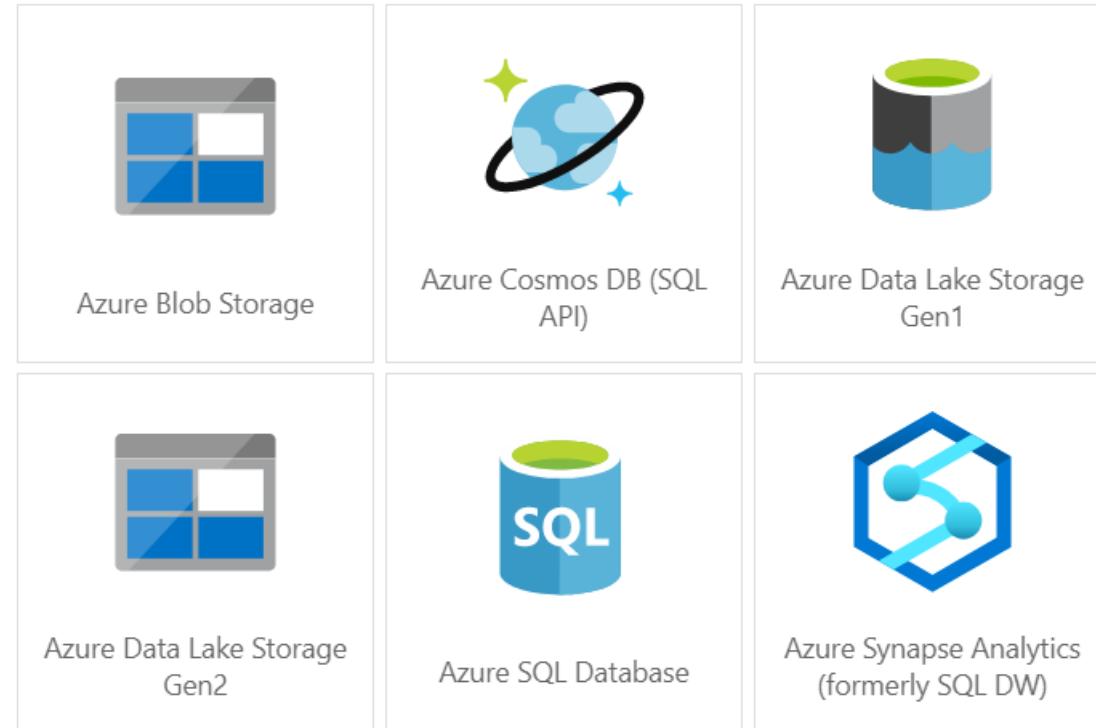
Data Flow Data Lineage

ADF Template Gallery

The image shows two side-by-side screenshots. On the left is the 'Template gallery' interface, which includes a sidebar for filtering by categories like Copy, Data Flow, SSIS, and Transform, and services used like Microsoft and My templates. It lists 12 data transfer templates with icons and brief descriptions. On the right is the 'Factory Resources' section of the Azure DevOps GIT interface, showing a navigation bar with 'Pipeline' and 'Pipeline from template' selected, and a search bar with a plus sign icon highlighted with a red circle.

Template Description	Source	Destination
Bulk Copy from Database to Azure Data Explorer	by Microsoft	Azure Data Explorer (ADX)
Bulk Copy from Database	by Microsoft	Azure Data Lake Store
Copy data from Google BigQuery to Azure Data Lake Store	by Microsoft	Azure Data Lake Storage
Copy data from HDFS to Azure Data Lake Store	by Microsoft	Azure Data Lake Storage
Copy data from Netezza to Azure Data Lake Store	by Microsoft	Azure Data Lake Storage
Copy data from on premise SQL Server to SQL Azure	by Microsoft	SQL Azure
Copy data from on premise SQL Server to SQL Data Warehouse	by Microsoft	SQL Data Warehouse
Copy data from Oracle to SQL Data Warehouse	by Microsoft	SQL Data Warehouse
Copy delta data from AWS S3 to Azure Data Lake Storage Gen2	by Microsoft	Azure Data Lake Storage Gen2
Copy multiple files containers between File Stores		Azure Data Lake Storage Gen2
Copy new files only by LastModifiedDate		Azure Data Lake Storage Gen2
Data Flow Search Log Analytics		Log Analytics

Mapping Data Flow – Source & Sink



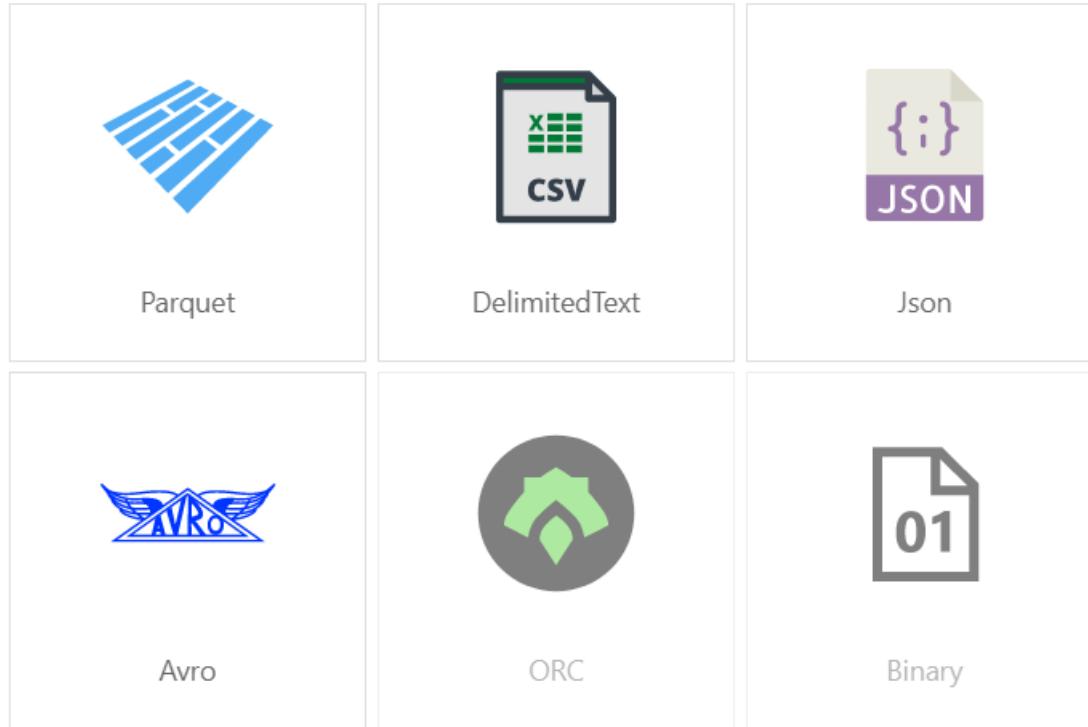
Mapping Data Flow – Source & Sink capabilities



- 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

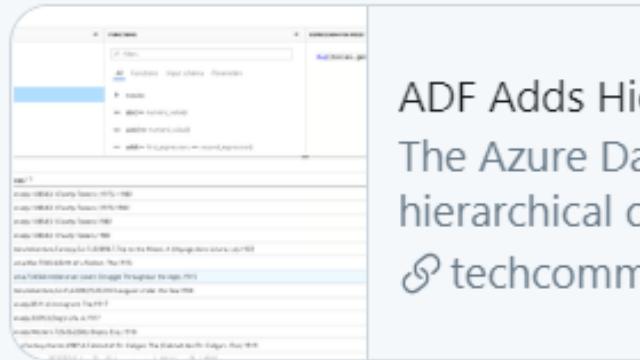


The screenshot shows the "Settings" tab of the Execute Data Flow transformation. The "Run on (Azure IR)" dropdown is set to "AutoResolveIntegrationRuntime". An orange arrow points from this dropdown to the "Time to live" section of the execution settings panel. The execution settings panel includes fields for Region, Compute type (General Purpose), Core count (4 + 4 Driver cores), and Time to live (0 minutes). A "Filter..." button is also present. The "Time to live" dropdown menu lists options: 0 minutes, 10 minutes, 30 minutes, 1 hour, and 4 hours.

Latest updates? Go Twitter!



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



8



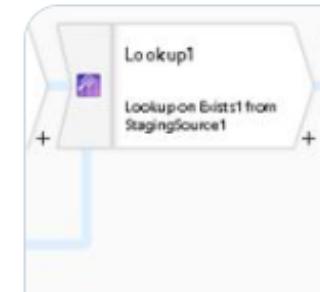
12



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 available...
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

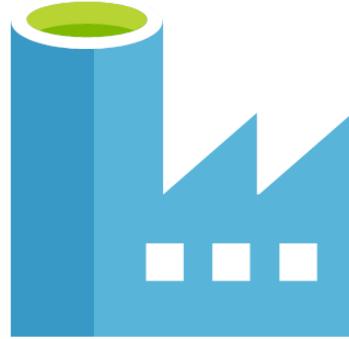
SSIS vs ADF activities vs T-SQL

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 ones	 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!

Obrigado!



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



09:30 AM - 10:30 AM	<u>SQL Server & Containers</u> Andrew Pruski Level: Beginner	<u>Azure Data Factory v2: Mapping Data Flows - first blood</u> Kamil Nowinski Level: Intermediate	<u>Index Tuning for the developer</u> Stijn Wynants Level: Beginner
Coffee Break			
10:45 AM - 11:45 AM	<u>From Docker to Big Data Clusters: a new era for SQL server</u> Christophe Laporte Level: Intermediate	<u>My experience implementing DistinctCount measures in SSAS</u> Tiago Rente Level: Beginner	<u>Will my workload run faster with In-Memory OLTP?</u> Ned Otter Level: Intermediate
Lunch			
01:00 PM - 02:00 PM	<u>Running SQL Server as Stateful(set) application on Kubernetes</u> David Barbarin Level: Intermediate	<u>Disposable Development environments with Azure DevOps and AzureRM</u> Gavin Campbell Level: Intermediate	<u>My Company is going to cloud, what can I do?</u> Paresh Motiwala Level: Beginner
02:00 PM - 02:20 PM	<u>Claranet: Manage monoliths as containers using DevOps solutions</u> Eduardo Namba Level: Advanced	<u>Microsoft: What's up with SQL Server & Azure SQLDB (including MI)</u> Vitor Faria Tomaz Level: Beginner	
02:30 PM - 03:30 PM	<u>SQL Server 2019 what's new for the DBA's</u> Ivan Campos Level: Intermediate	<u>DevOps and automation in a modern Microsoft DataWarehouse and BI/AI world</u> Gergely Csom Level: Intermediate	<u>Adjusting management rules in a business environment</u> Ana Raquel Teixeira Level: Beginner
Coffee Break			
03:50 PM - 04:50 PM	<u>dbatools' recipes for Data Professionals</u> Cláudio Silva Level: Intermediate	<u>Power up Power BI... with AI</u> Rita Dias Level: Intermediate	<u>Batch Execution Mode on Rowstore Indexes</u> Niko Neugebauer Level: Intermediate