

Azure Data Factory <u>Mapping Data Flow - capabilities</u>





Kamil Nowinski

Kamil Nowinski

FRIEND OF redgate 2019 Solutions Associate SQL Server 2012

Microsoft Data Platform MVP Speaker, blogger, data enthusiast Senior Data Engineer at ASOS (<u>www.asos.com</u>) 15+ yrs experience as DBA/DEV/BI Member of the Data Community PL Project member of "SCD Merge Wizard" Founder of blog SQLPlayer (www.SQLplayer.net)

Data

Platform

Microsof

Professiona

Most Valuable

SQL Server Certificates: MCITP, MCP, MCTS, MCSA, MCSE Data Platform, MCSE Data Management & Analytics Moreover: Bicycle, Running, Digital photography @NowinskiK, @SQLPlayer





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)	Databa	ise (24)	File Storage (5)	NoSQL (3)	Services ar	nd Apps (28)	Generic (4)
Blob Storage	Amazon Redshift	Netezza	Amazon S3	Cassandra	Amazon MWS	Office 365 *	НТТР
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	l Sink					
	Supported as Source onl	У					
	Supported as Sink only						

ADF Key Concepts



Visual Data Transformations with



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 doe 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

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



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

Micro	osoft Azure Data Factory	SQLPla	yerDemo2		$^{ ho}$ Search resources			
»	🔛 Data Factory 🗸	l∱∣ Publ	ish All 🛛 🗸 Validat	te All 🛛 📿 Ref	resh 🔋 Discard All			
	Factory Resources	» «	$\vec{_{\mathfrak{B}}}$ CopyFlow $ imes$	III users $ imes$	\boxplus dstUsersBlob $ imes$			
		+	Debug	🗸 Validate	C Source Settings			
(3)	III Pipelines	2						
	Datasets	12						
	මේ Data Flows (Preview)	5]					

Simple Copy Flow



Mapping Data Flow: Components = Actions *

Multiple inputs/outputs

≰ Conditional Split

눩 Union

省 Lookup

Schema modifier Derived Column Maggregate Surrogate Key Pivot

🗖 Unpivot





材 Sort

Extend Destination

* 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)

crosoft Azure					× (ī.	Q © 0 X
BigPlayer f냂 Data Factory >	🗅 Publish All 🗸 Validate All	📿 Refresh 💿 Discard All 🛛 💱 ARN	M Template 🗸			
Factory Resources ≥ 🛠	_{BB} users ×					
$\mathbf{\nabla}$ Filter Resources \times \bullet +	Debug 🗸 Validate					Code
D Pipelines ··· 2 Datasets ··· 9	sourceUsers	Select1 Renaming sourceUsers to Select1 with columns	FilterByReputation	Standard Strategy Content of State S	SortByLocation	Conditionally distributing the data in 2 groups, based on
Badges	Users_BlobCsv	+ LastAccessDate, Location,	+ Reputation'	+ SumOfReputation, SumOfViews, Count' +	'Location'	columns 'Location, Location, Location, Location, Location'
BadgesBlob						AllBight
BadgesBlobWithHeader						Conditionally distributing the
BadgesStatsByName						data in 2 groups, based on columns 'Location, Location, Location, Location' +
BadgesStatsByNameBlob						
Crimes_BlobCsv				_		
Src_Users	General External depe	endencies				
Users_BlobCsv						
UsersTest	Name *	users				
B Data Flows ··· 3	Description					
▲ 🗁 StackOverflow 3						
badgesGroupByName						
badgesGroupByName2						
users						

Guided experience to build data flows

Microsoft Azure Data Factory + SQLPIa	ayerDemo	٩,	Search resources			<i>(</i> ., Q 🙂 🔊 R
» 🔛 Data Factory 🗸 ሰ Publish All	Validate All 📿 Refre	esh 🔋 Discard All 🛛 🕅 ARM	Template 🗸			⊳ «
Factory Resources × «	_{eie} usersql ×					
✓ Pilter resources by name +	💽 Debug 🗸 Val	lidate				O Code
© Pipelines 5	source1	Select1	FilterByReputation	GroupByLocation Sort1	Filter1	sink1 +
Datasets 16	Columos	Renaming source1 to Select1	Filtering rows using	Aggregating data by 'Location'	Filtering rows using	
de Data Flows (Preview) 6	13 total	DisplayName, DownVotes,	+ expressions on columns 'Reputation' +	Producing columns' Soluting to	+ expressions on columns + "Location, Location"	+ Export data to AzureBlob2
usersql						<u>[</u> 4]
▶ 🗖 Beta 2		Multiple inputs/outputs				
▲ 🗁 StackOverflow 3	Add Source	😋 New Branch				
badgesGroupByName		🎾 Join				
badgesGroupByName2		🛸 Conditional Split				
users		🎦 Union				
		쮬 Lookup				
		Schema modifier				
		🔩 Derived Column				
		∑ Aggregate				
		🚝 Surrogate Key				
		Pivot				
		🗔 Unpivot				
		Window				
	Source Settings Defi	Row modifier	nspect Data Preview			\Box
	Output stream name *	source1				
	Course Dataset *	-	- 2 E-14 Nov			
	Source Dataset *	stack_users	- V Ealt + New	v		
	Options	Allow schema drift 0				
	Sampling *	Enable Disable				
	Rows limit	1000				
🔀 Connections						
Triggers						

What is going on behind the scenes?



JAR

Azure Databricks

FEEDBACK FORMS

PLEASE FILL OUT AND PASS TO YOUR ROOM HELPER BEFORE YOU LEAVE THE SESSION



Data Preview in Debug mode



Micro	soft Azı	ure 🛛 Data Fact	ory 🕨 BigF	actorySample2			I	ſł.	Q (<u>;</u>)	?
»		Data Flow	Dataflow					Num	per of tra	nsforn	ns: 9
P	6	TripFare Sink: •••	8	AggregateByPayme	TotalPaymentBy 🤡						
6	•	TripData Sink: ● ●		JoinMatchedData	AggregateDayStats Sink: •	DayStatsSink 🔗					
		TripFare Sink: • • •	Ð								
				JoinMatchedData Sink: • • 🕄 🕄	AggregateVendorSt Sink: •	VendorStatsSink					

TotalPaymentByPaymentType							
TRANSFORM	ROWS	TIME					
• TripFare	ඩ් 49999	3s 708ms					
 AggregateByPaymentType 	5	7s 801ms					

VendorStatsSink		0
TRANSFORM	ROWS	TIME
 TripFare 	ට් 49999	3s 708ms
 JoinMatchedData 	ට් 49999	
● TripData	49999	3s 290ms
 AggregateVendorStats 	2	3s 449ms

DayStatsSink		C
TRANSFORM	ROWS	TIME
• TripFare	ඩ් 49999	3s 708ms
 JoinMatchedData 	ට් 49999	
• TripData	49999	3s 290ms
 AggregateDayStats 	8	6s 561ms

Data Flow Execution Plan

												(?.	0) A
Data Flow	Dataflow											Number of	transforms:	9 ×
TripData Sink: ••	st at	oinMatchedData	2	AggregateDayStats	DayStatsSink	0	1			→	VendorSt	atsSink		
TripFare Sink:										Total colu	mns	5		
										New colu	minis	0		
	J	oinMatchedData) Za	AggregateVendorSta	VendorStatsSink	0				Updated of	columns	0		
	2	nic ••• 0		Sinc •						Dropped	columns	0		
TripFare Sink: •••	8	ggregateByPaymen nk: •	>	TotalPaymentByP						Drifted co	lumns	0		
										Stream	informati	on		
										Rows calc	ulated	2		
										Total parti	tion	1		
										Stage tim	0	24 210 mm		
										stage titte		35 3 19 115		
										Partition	chart	33 319005		
										Partition 2.0	chart	35 319005		
										Partition 2.0 1.5	chart	33 319005	•	
										Partition 2.0 1.5 1.0	chart	22 31300	•	
_						_				Partition 2.0 1.5 1.0 0.5	chart	32,31900		
UendorStatsSink						_			0	Partition 2.0 1.5 1.0 0.5 0	chart	35 31900		
VendorStatsSink COLUMN	METHOD	ORIGINAL SI	OURCE			_			0	Partition 2.0 1.5 1.0 0.5 0	chart	1 Partition		
VendorStatsSink COLUMN A passenger_count	METHOD Calculated	ORIGINAL SI	OURCE Data			-			0	Partition 2.0 1.5 0.5 0	chart	1 Partition		
VendorStatsSink COLUMN passenger_count	METHOD Calculated	ORIGINAL SI Tript paste	OURCE Data eng <mark>er_</mark> cour	nt					0	Partition 2.0 1.5 0.5 0 Skewness	chart	1 Partition		
VendorStatsSink COLUMN passenger_count trip_time_in_secs	METHOD Calculated Calculated	ORIGINAL SI Tript paste Z Tript	OURCE Data eng <mark>er_</mark> cour Data(trip_ti	nt ime_in_secs)					0	Partition 2.0 1.5 0.5 0 Skewness Kurtosis	chart	1 Partition -		
VendorStatsSink COLUMN passenger_count trip_time_in_secs trip_distance	METHOD Calculated Calculated Calculated	ORIGINAL SI Tript Paste Tript	OURCE Data eng <mark>er_</mark> cour Data(trip_ti Data(trip_d	nt ime_in_secs) listance)					0	Partition 2.0 1.5 0 0.5 0 Skewness Kurtosis	chart	1 Partition - -		
VendorStatsSink COLUMN passenger_count trip_time_in_secs trip_distance TotalTripFare	METHOD Calculated Calculated Calculated Calculated	ORIGINAL SI Tript past Tript Tript	iOURCE Data enger_cour Data(trip_ti Data(trip_d Fare(total_	nt ime_in_secs) listance) amount)					0	Partition 2.0 1.5 0 0.5 0 Skewness Kurtosis	chart	1 Partition -		

Data Flow Data Lineage



Mapping Data Flow – the latest update

- The Preview version of ADF with Data Flows has been deprecated (26 February)
- You will no longer need to stand-up Azure Databricks clusters. ADF will handle cluster management for you on-demand.
- Data Flow samples have been into the new ADF Template Gallery



Mapping Data Flow – the latest update for v2



- 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 – New Datasets





- New datasets for Data Flow (only):
 - Parquet
 - Delimited Text

General	Settings	User Propertie	es				
Data Flow *		dataflow2		•	🖉 Edit	+ New	
		▲ source1 0					
		NAME	VALUE				
		mypath	@pipeli	ne().paramet	ers.mypath		
lun on *		AutoResolveIntegrati	ionRuntime	- 0			
Compute Type	*	General Purpose		•			
Core count *		8		•			
taging linked s	ervice	Select		•	+ New		
taging storage	folder	container/folder			Browse		

- The Execute Data Flow transformation:
 - Now support **parameterized datasets**
 - Control size of cluster for specific data flow execution

SSIS vs ADF activities vs T-SQL

Activity	Description	SSIS e	quivalent	SQL Server equivalent
New branch	Create a new flow branch with the same data	Υ	Multicast (+icon)	SELECT INTO SELECT OUTPUT
>>> Join	Join data from two streams based on a condition	т	Merge join	INNER LEFT RIGHT JOIN, CROSS FULL OUTER JOIN
Conditional Split	Route data into different streams based on conditions	乙	Conditional Split	SELECT INTO WHERE condition1 SELECT INTO WHERE condition2 CASE WHEN
Dnion	Collect data from multiple streams	¥	Union All	SELECT colla UNION (ALL) SELECT collb
Lookup	Lookup additional data from another stream	1	Lookup	LEFT RIGHT JOIN
Column	Compute new columns based on the existing once	f.x	Derived Column	SELECT Column1 * 1.09 as NewColumn
E Aggregate	Calculate aggregation on the stream	{: ^{}∑}	Aggregate	<pre>SELECT Year(DateOfBirth) as Year, MIN(), MAX(), AVG() GROUP BY Year(DateOfBirth)</pre>

http://bit.ly/ADFDFvsSSIS

http://bit.ly/ADFDF-CheatSheet



- Microsoft Azure Data Factory <u>Tutorials & API Reference</u>
- Azure Data Factory <u>Overview</u>
- Azure Data Factory <u>Data integration service</u>
- ADF Mapping Data Flow's <u>documentation</u>
- ADF Mapping Data Flow's <u>videos</u>
- SQLPlayer blog:
 - <u>Azure Data Factory v2 and its available components in Data Flows</u>
 - Follow this tag on SQLPlayer blog: <u>#ADFDF</u>



Thank you!



kamil@nowinski.net

@NowinskiK **@SQLPlayer**

Ø **SQLPlayer.net**

https://github.com/NowinskiK/CommunityEvents

Kamil Nowinski

Microsoft Data Platform MVP MCSE Data Platform & MCSE Data Management and Analytics