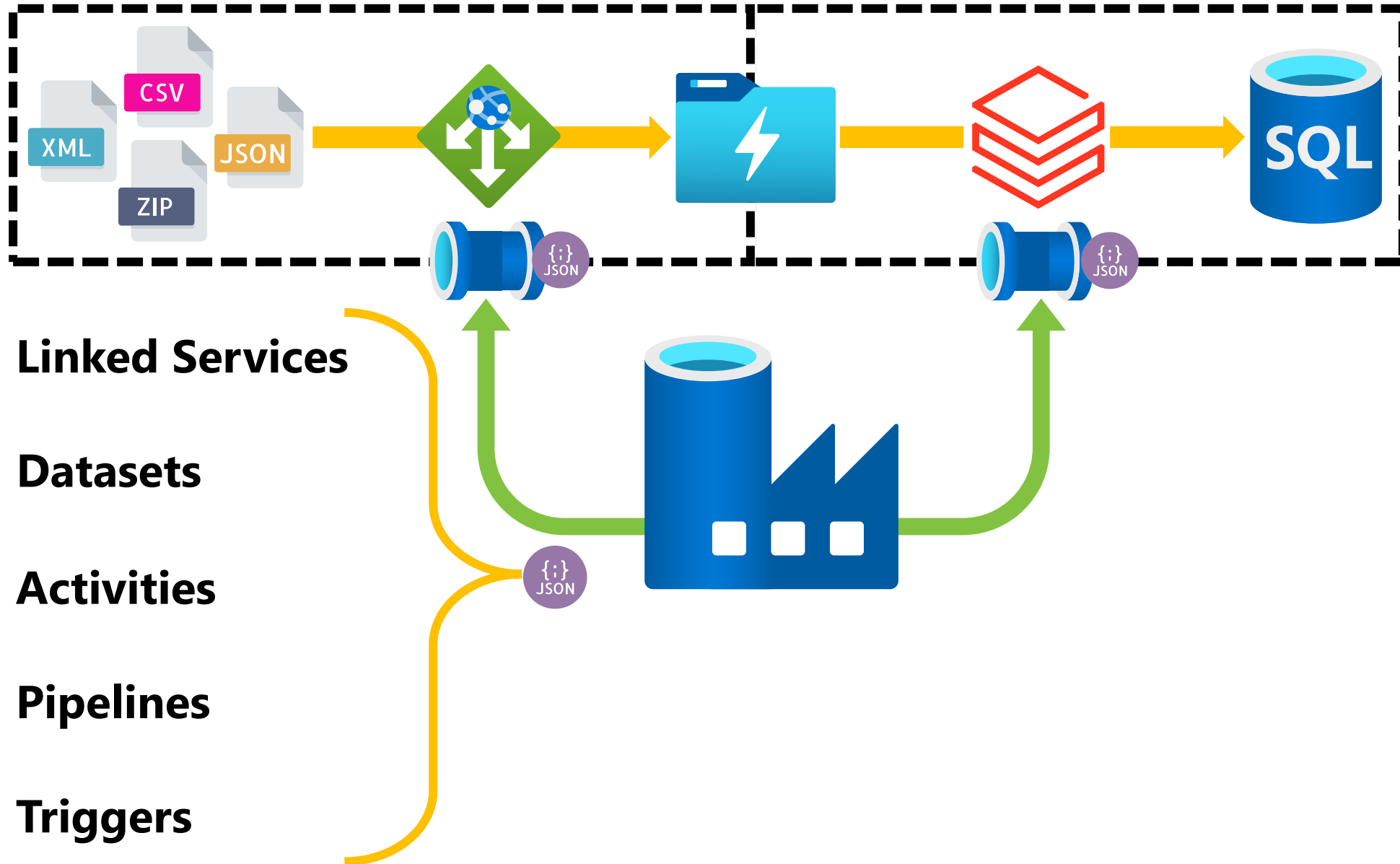


Module 4: Data Flows

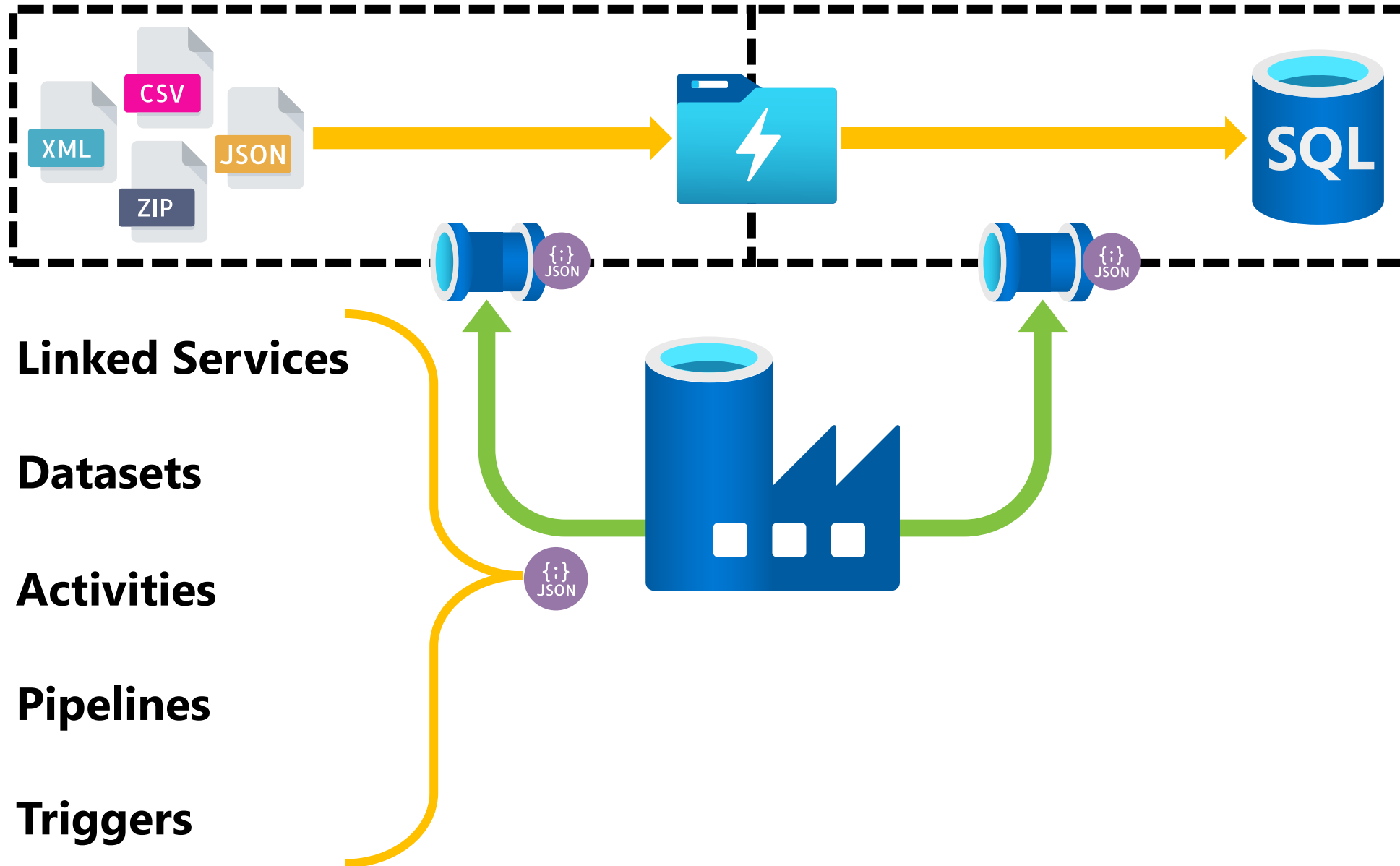
- ▢ Mapping Data Flows
- ▢ Wrangling Data Flows
- ▢ Configuration
- ▢ Use Cases



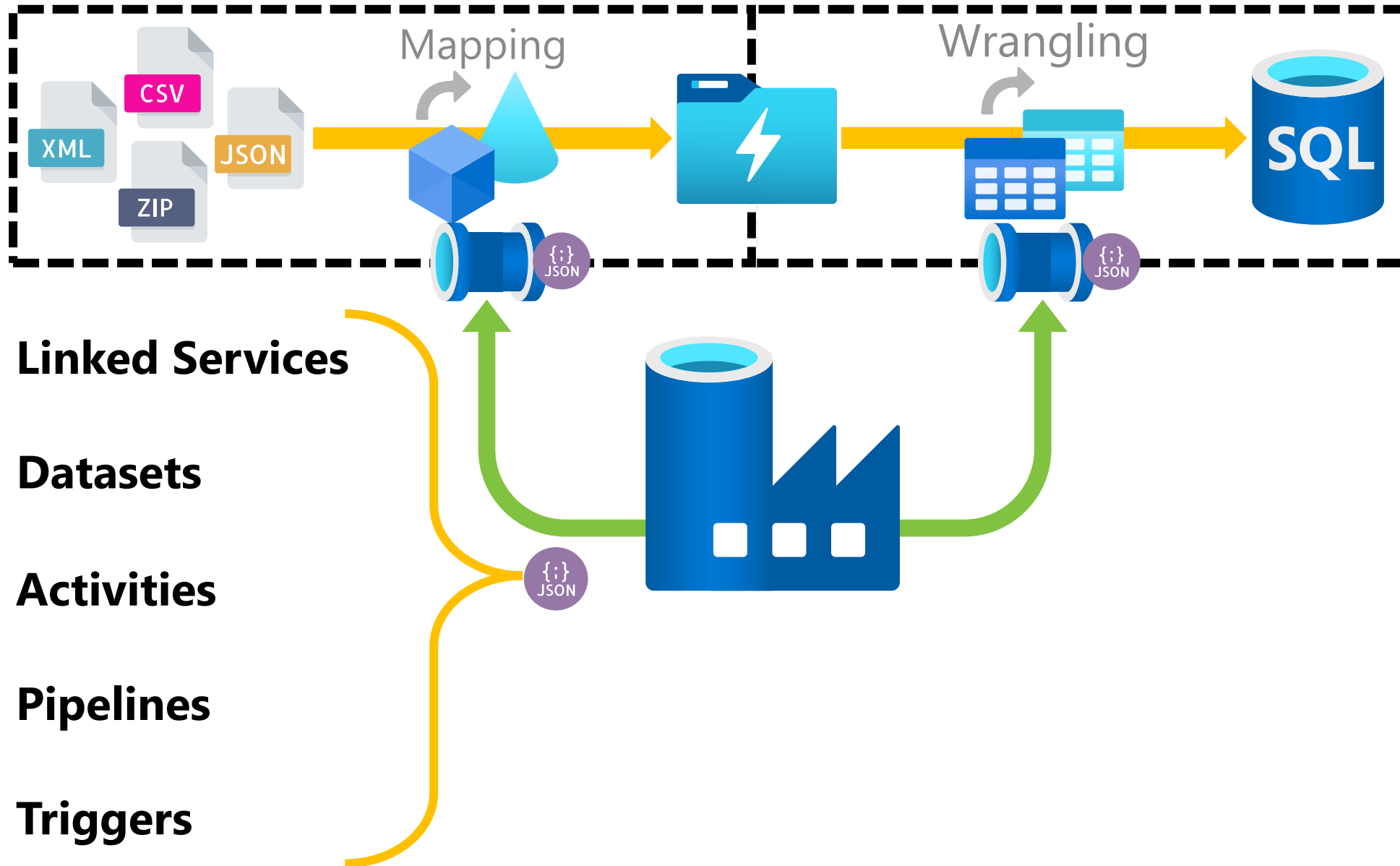
Data Factory Control Flow Components

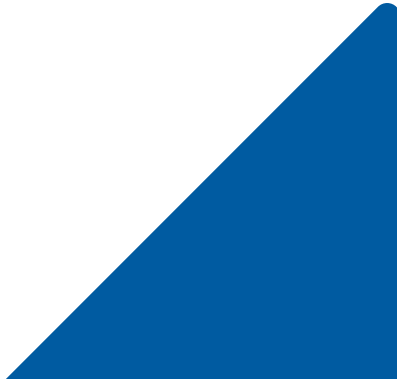
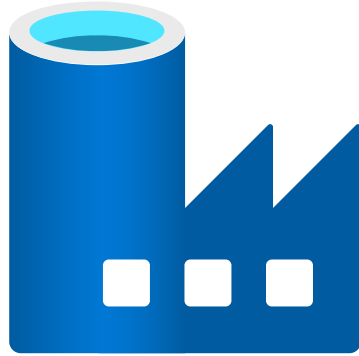


Data Factory Components

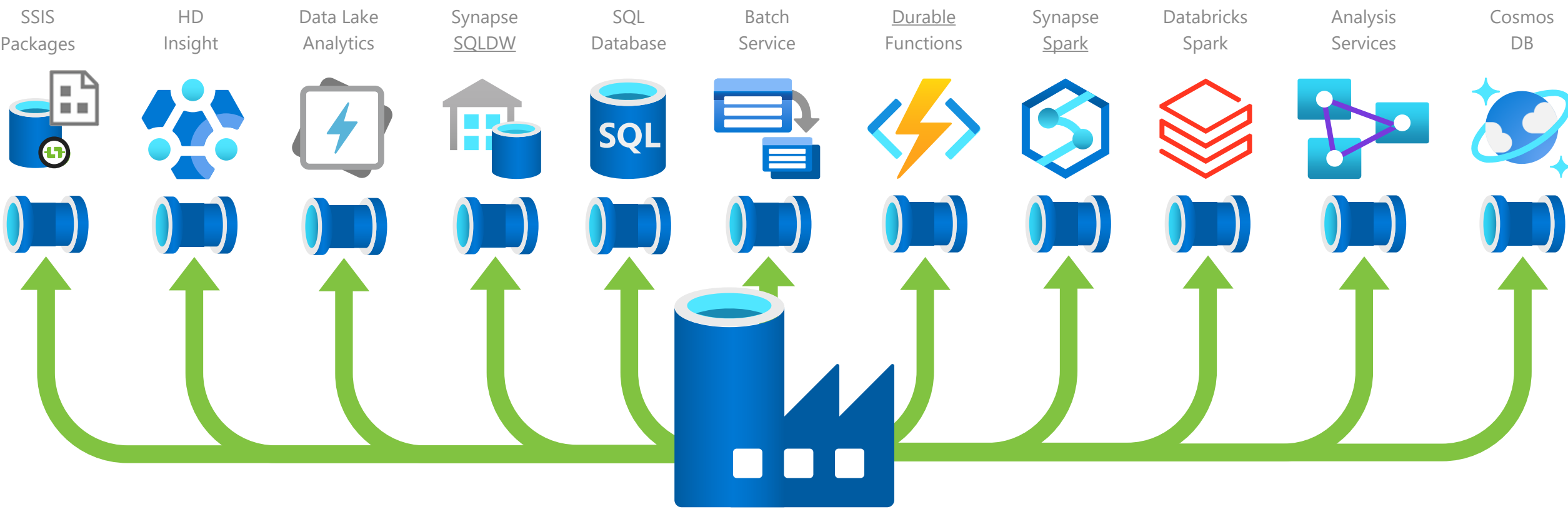


Data Factory Data Flows

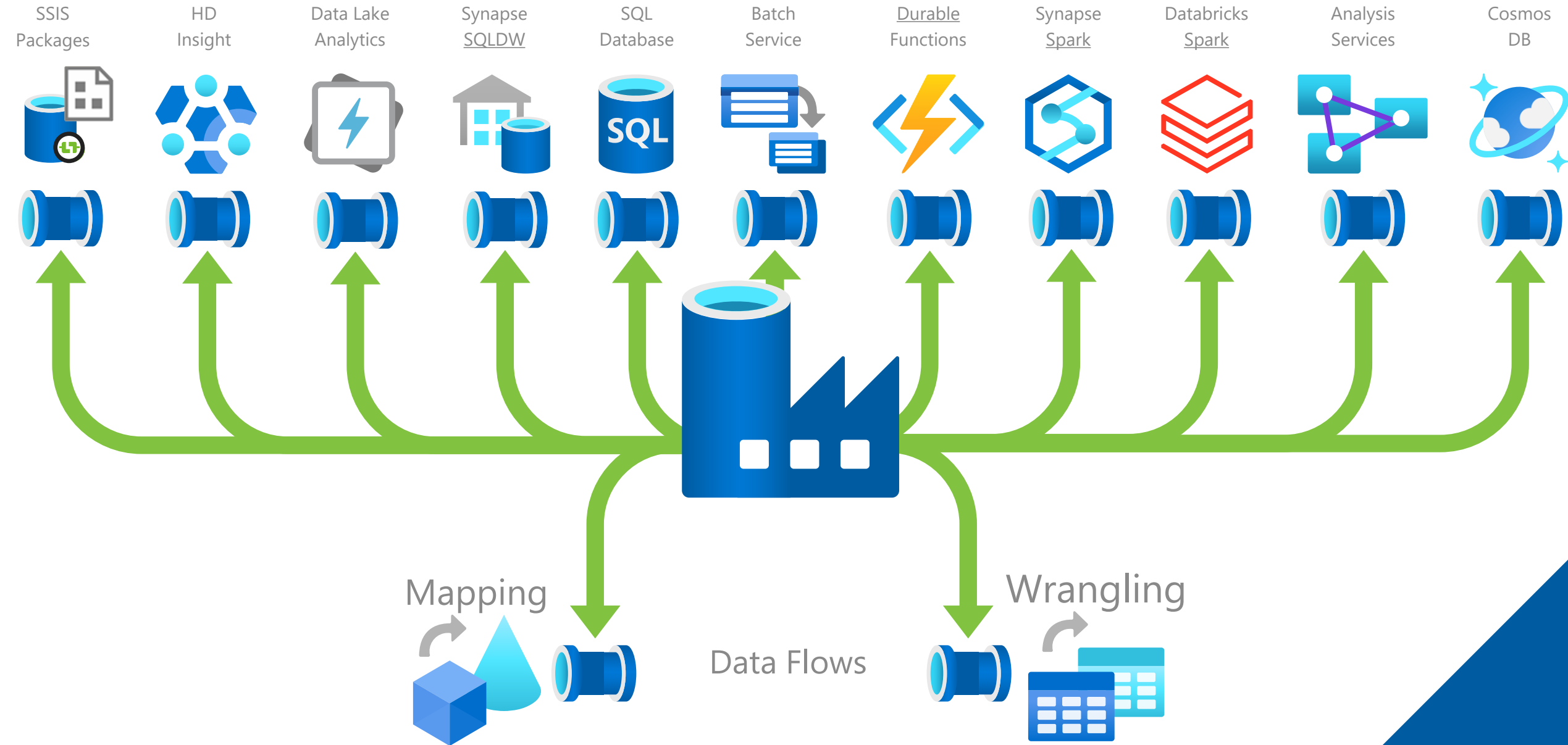




Other Data Transformation Services in Azure



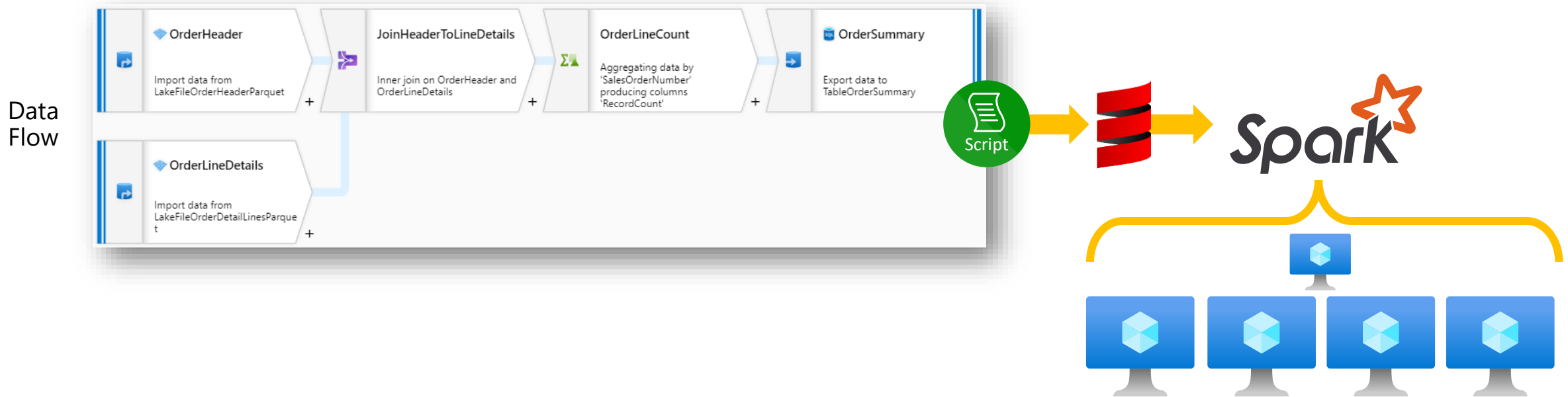
When Should We Use Data Flows?



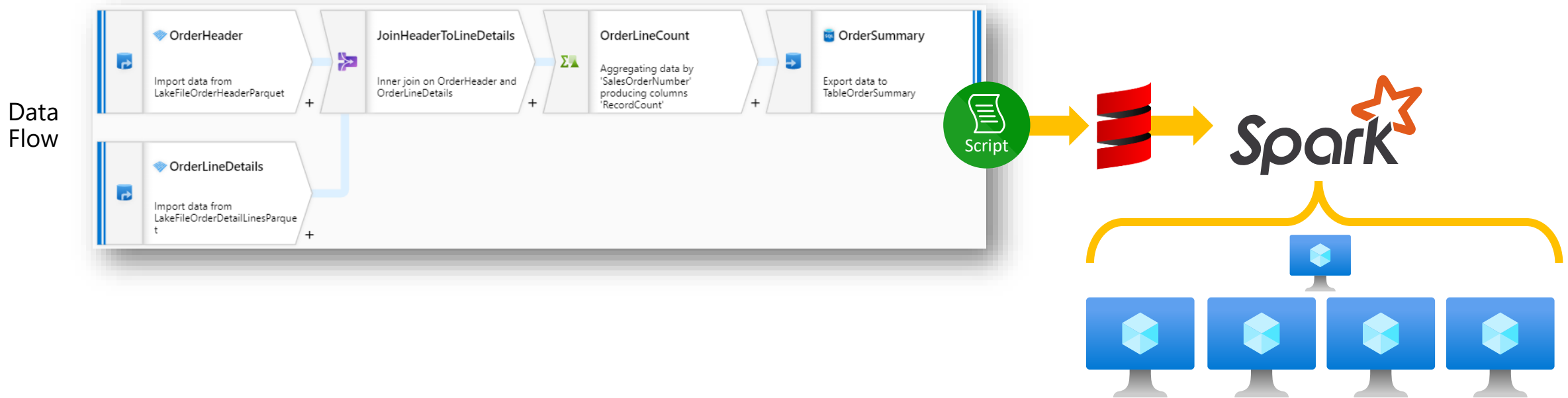
Mapping Data Flows



What is a Mapping Data Flow?



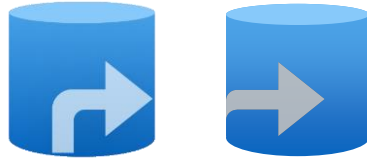
Q: What is a Mapping Data Flow?



A: Graphic data transformation tool that sits on top of Apache Spark.

What can a Mapping Data Flow do? - Inputs and Outputs

Source &
Sink



Limited
Connectors

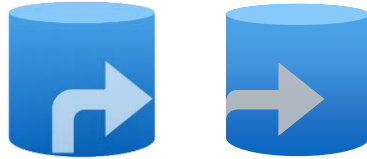


Limited File
Type Support



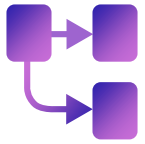
What can a Mapping Data Flow do? - Inputs and Outputs

Source &
Sink



- Schema Drift & Validation
- Inferred Drifted Column Types
- File Lists
- Delete/Move Operations
- File Modified Date Filtering
- Pre-Execute Scripts & Operations (Truncate)

What can a Mapping Data Flow do? - Transformations



New Branch



Join



Conditional Split



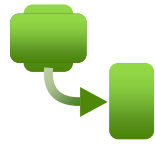
Exists



Union



Lookup



Derived Column



Select



Aggregate



Surrogate Key



Pivot



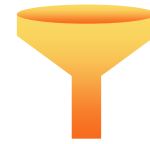
Unpivot



Window



Flatten



Filter



Sort



Alter Row

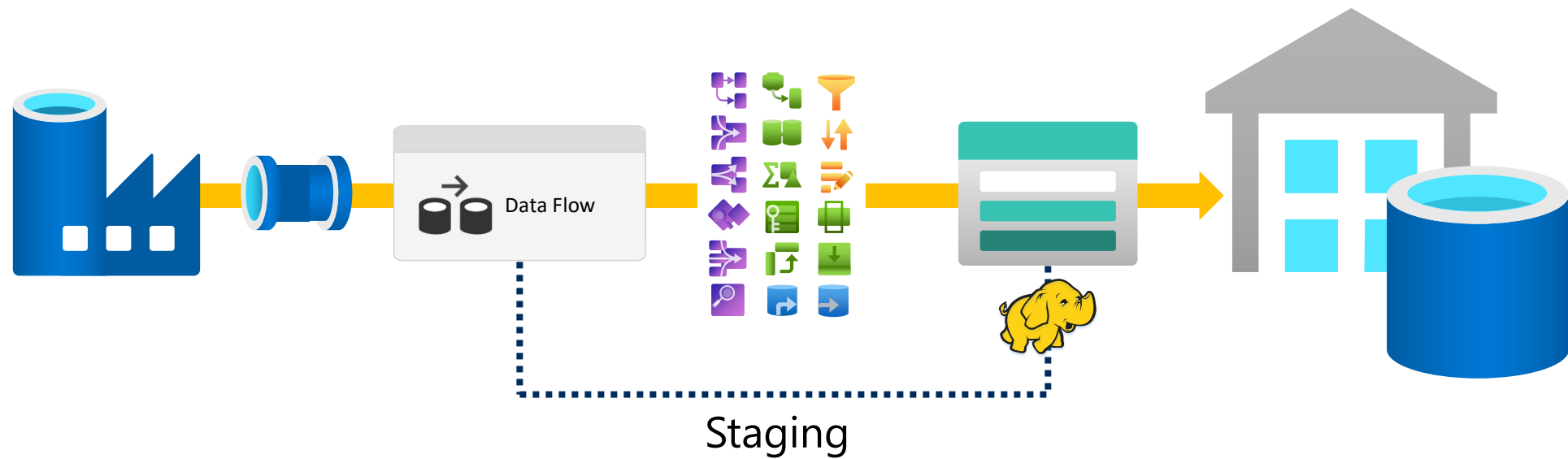
Key

Input & Output Modifiers

Schema Modifiers

Row Modifiers

What can a Mapping Data Flow do? - PolyBase

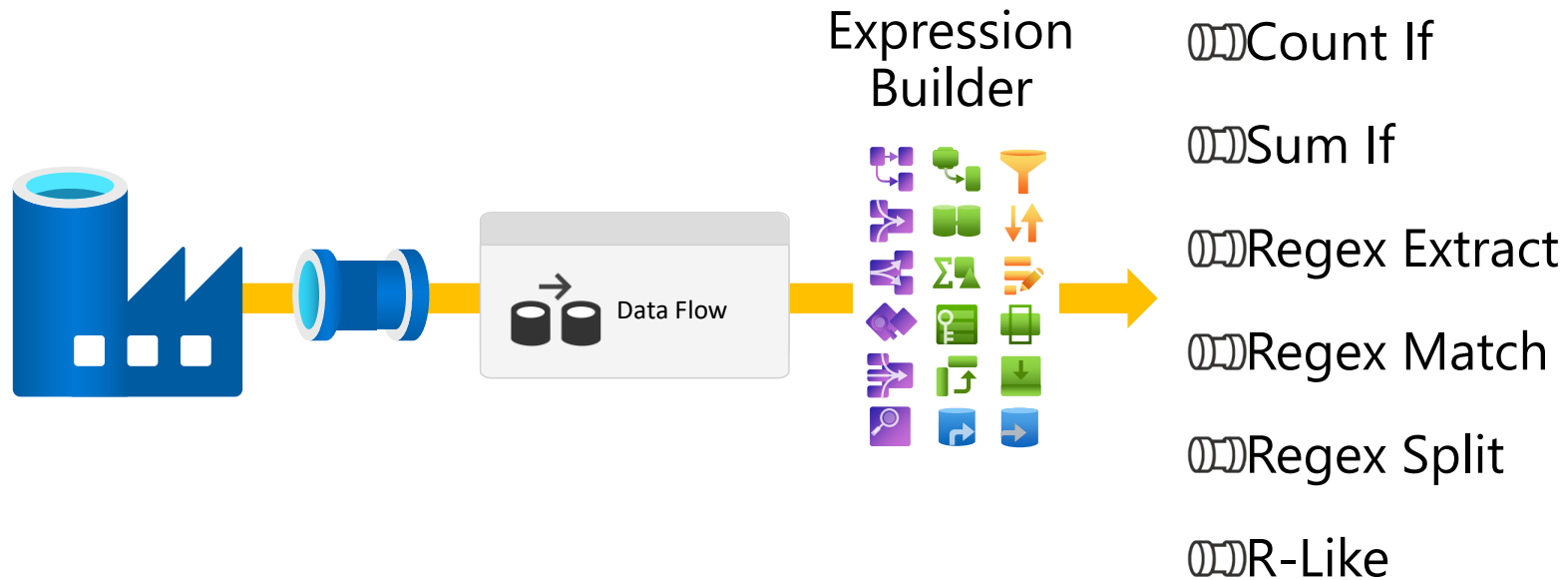


PolyBase ⓘ

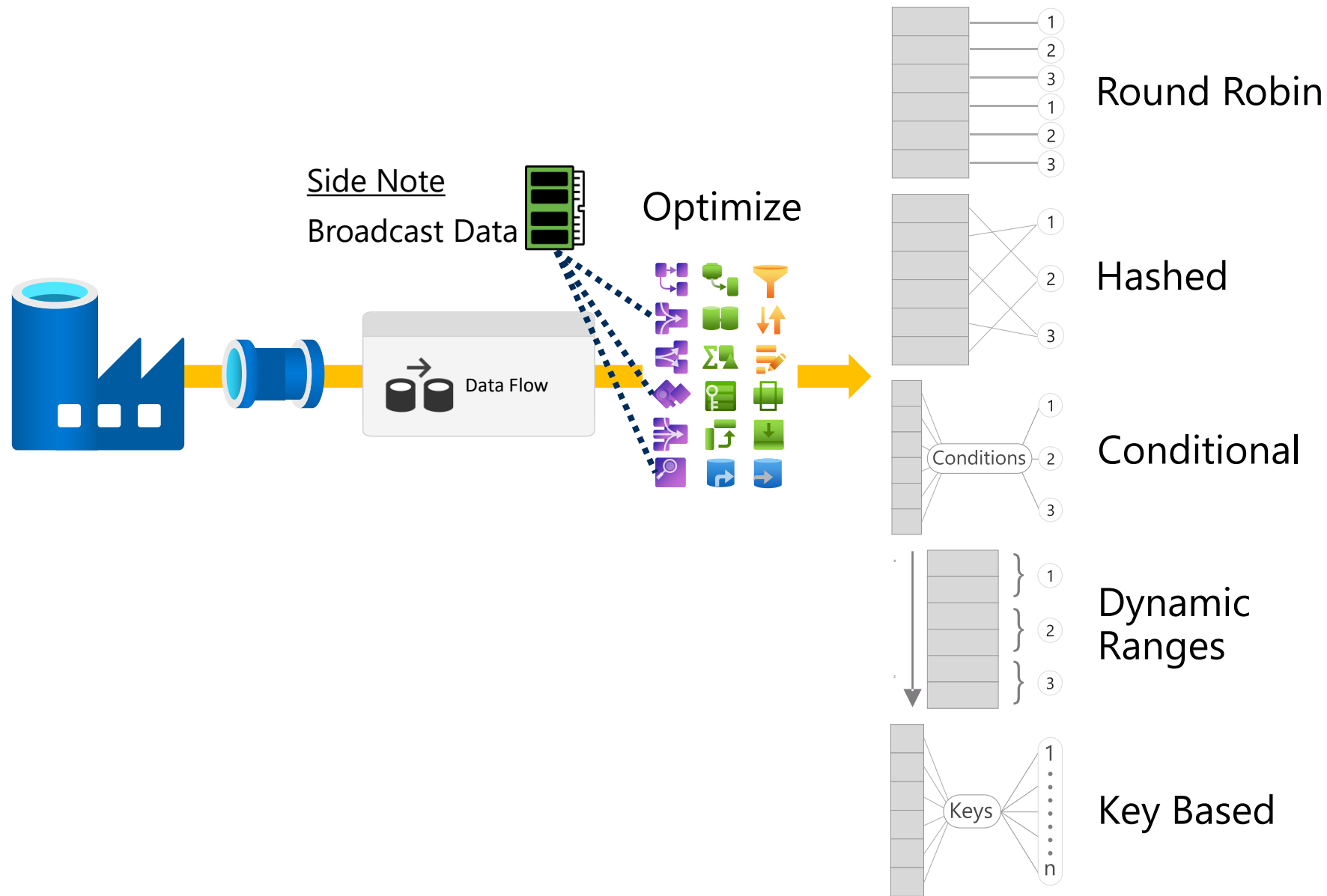
Staging linked service ⓘ + New

Staging storage folder / ⓘ

What can a Mapping Data Flow do? - Expression Builder



What can a Mapping Data Flow do? - Partition Handling

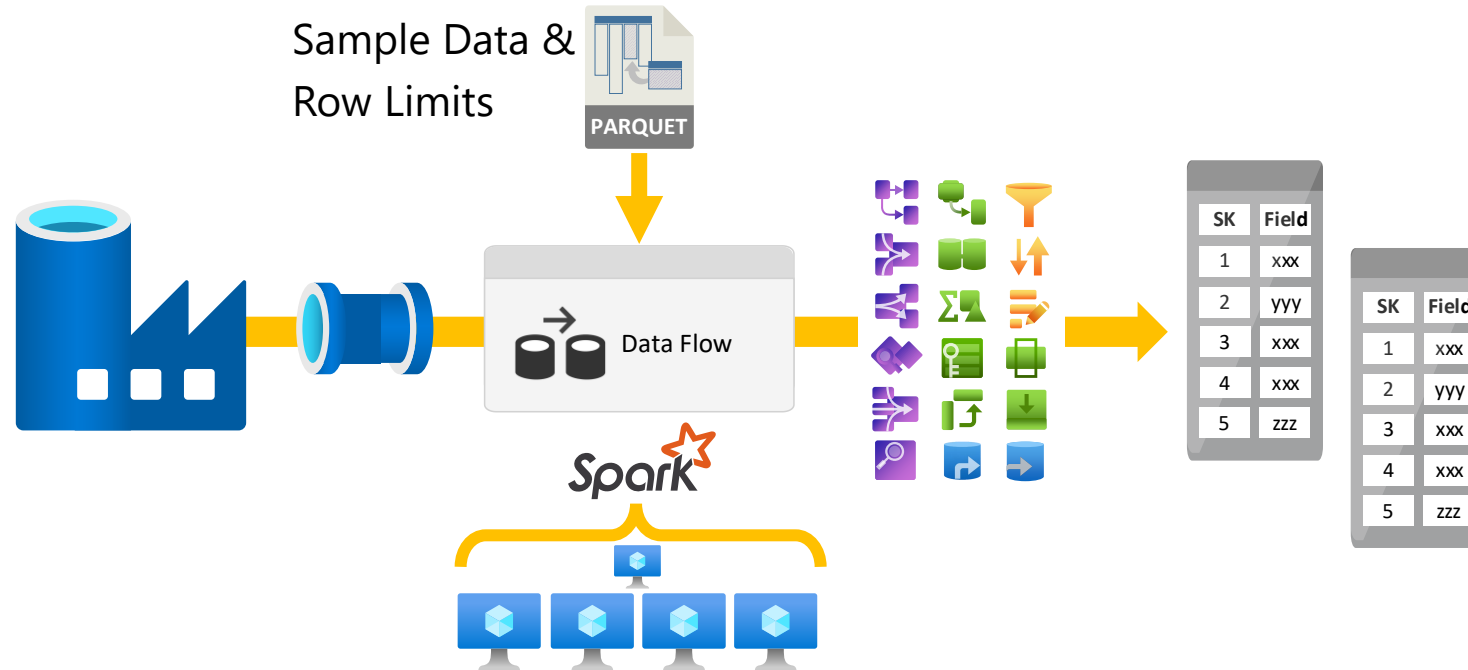


What can a Mapping Data Flow do? - Debugging

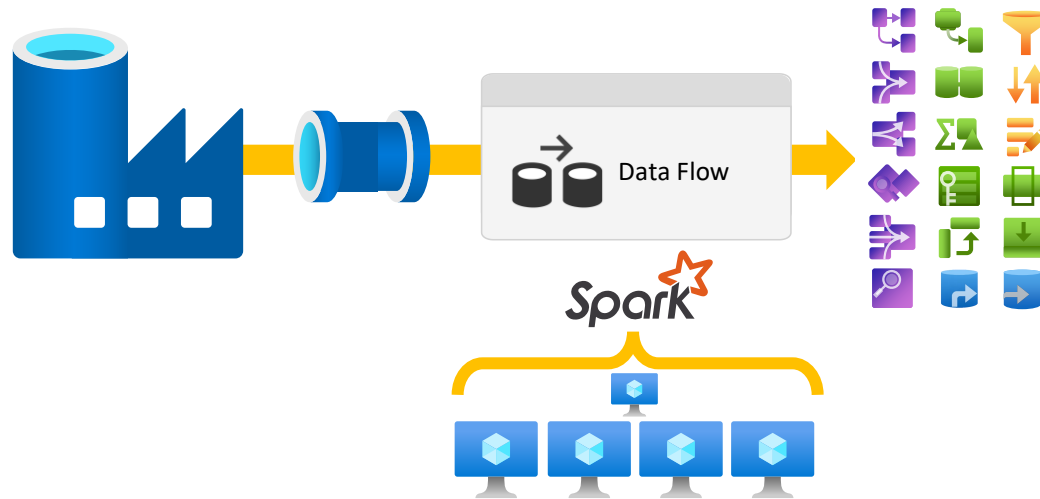


Enable Data Flow Debug Mode

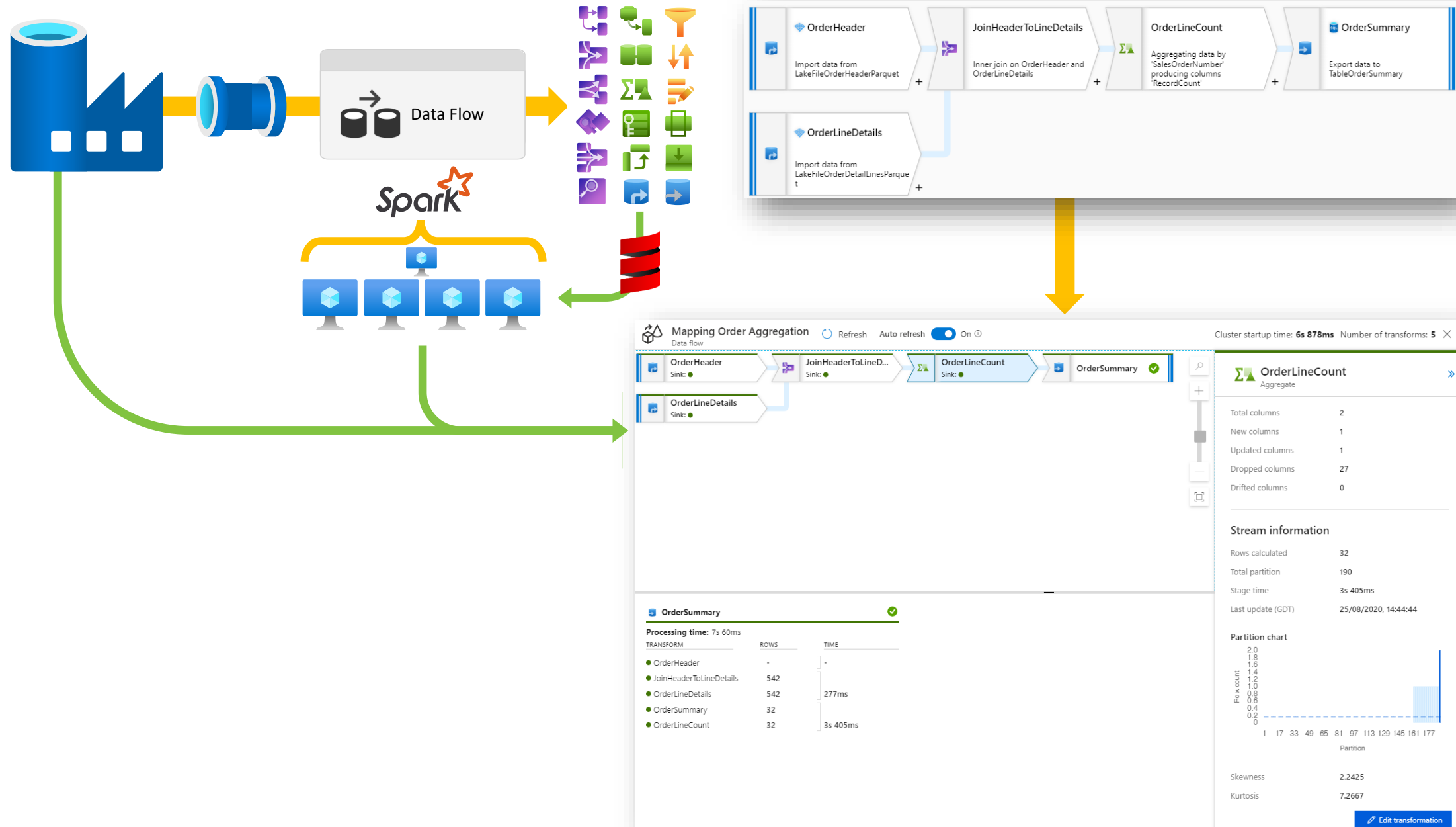
Data
Preview

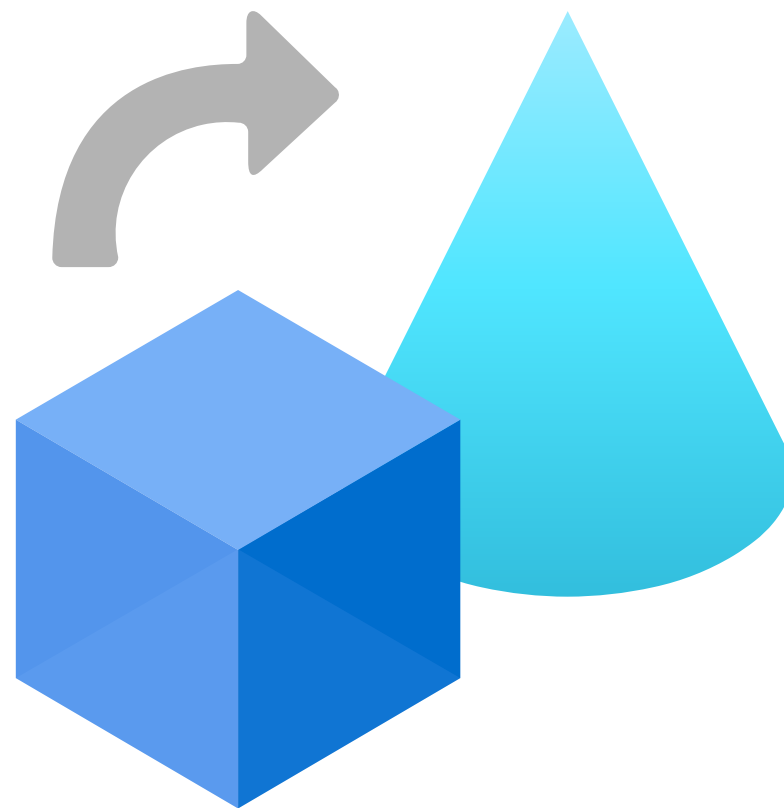


What can a Mapping Data Flow do? - Monitoring



What can a Mapping Data Flow do? - Monitoring





Mapping Data Flow

Wrangling Data Flows

(Preview)



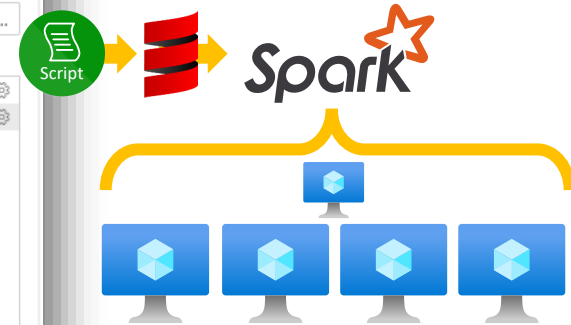
What is a Wrangling Data Flow?



Data Flow

The screenshot shows the Databricks Data Wrangling interface. The top menu bar includes 'Home', 'Transform', 'Add column', and 'View'. Below the menu is a toolbar with various icons for data manipulation. The main area displays a table with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, UnitPrice, UnitPriceDiscount, LineTotal, and rowguid. The table contains 17 rows of data. On the right side, the 'Query settings' panel is visible, showing the 'Name' as 'LakeFileOrderDetailLinesP...' and the 'Applied steps' as 'AdfDoc' and 'Parquet'.

1 ² SalesOrderID	1 ² SalesOrderDetailID	1 ² OrderQty	1 ² ProductID	1.2 UnitPrice	1.2 UnitPriceDiscount	1.2 LineTotal	A ^B rowguid
1	71774	110562	1	836	356.898	0	356.898 e3a1994c-7a68-4ce8-96a3-77f
2	71774	110563	1	822	356.898	0	356.898 5c77f557-fdb6-43ba-90b9-9a7
3	71776	110567	1	907	63.9	0	63.9 6dbfe398-d15d-425e-aa58-88
4	71780	110616	4	905	218.454	0	873.816 377246c9-4483-48ed-a5b9-e5
5	71780	110617	2	983	461.694	0	923.388 43a54bcd-536d-4a1b-8e69-24
6	71780	110618	6	988	112.998	0.4	406.793 12706fab-f3a2-48c6-b7c7-1cc
7	71780	110619	2	748	818.7	0	1637.4 b12f0d3b-5b4e-4f1f-b2f0-f7cc
8	71780	110620	1	990	323.994	0	323.994 f117a449-039d-44b8-a4b2-b1
9	71780	110621	1	926	149.874	0	149.874 92e5052b-72d0-4c91-9a8c-42
10	71780	110622	1	743	809.76	0	809.76 8bd33bed-c4f6-4d44-84fb-a7c
11	71780	110623	4	782	1376.994	0	5507.976 686999fb-42e6-4d00-9a14-83i
12	71780	110624	2	918	158.43	0	316.86 82940b03-c70b-4183-8660-6b
13	71780	110625	4	780	1391.994	0	5567.976 644b0cd6-b2c3-4e4d-ab43-09
14	71780	110626	1	937	48.594	0	48.594 7f5feb17-8ef4-4236-9f1c-1504
15	71780	110627	6	867	41.994	0	251.964 ac78838d-b503-41a5-9791-48
16	71780	110628	1	985	112.998	0.4	67.799 2c10a282-a13d-442a-8f45-f4d
17	71780	110629	2	989	323.994	0	647.988 654fb79e-70df-4b92-9832-9fa



What can a Wrangling Data Flow do?



Data Flow

Home Transform Add column View

Enter data Options Manage parameters Refresh Advanced editor Properties Choose columns Remove columns Keep rows Remove rows Sort Split column Group by Data type: Whole number Use first row as headers Replace values Merge queries Append queries Combine files

Queries

- ADFRsource [1]
- LakeFileOrderDetailL...
- UserQuery

Parquet.Document (AdfDoc)

	1 ² SalesOrderID	1 ² SalesOrderDetailID	1 ² OrderQty	1 ² ProductID	1.2 UnitPrice	1.2 UnitPriceDiscount	1.2 LineTotal	A ^B rowguid
1	71774	110562	1	836	356.898	0	356.898	e3a1994c-7a68-4ce8-96a3-77f
2	71774	110563	1	822	356.898	0	356.898	5c77f557-fdb6-43ba-90b9-9a7
3	71776	110567	1	907	63.9	0	63.9	6dbfe398-d15d-425e-aa58-88
4	71780	110616	4	905	218.454	0	873.816	377246c9-4483-48ed-a5b9-e5
5	71780	110617	2	983	461.694	0	923.388	43a54bcd-536d-4a1b-8e69-24
6	71780	110618	6	988	112.998	0.4	406.793	12706fab-f3a2-48c6-b7c7-1cc
7	71780	110619	2	748	818.7	0	1637.4	b12f0d3b-5b4e-4f1f-b2f0-f7cc
8	71780	110620	1	990	323.994	0	323.994	f117a449-039d-44b8-a4b2-b1
9	71780	110621	1	926	149.874	0	149.874	92e5052b-72d0-4c91-9a8c-42
10	71780	110622	1	743	809.76	0	809.76	8bd33bed-c4f6-4d44-84fb-a7c
11	71780	110623	4	782	1376.994	0	5507.976	686999fb-42e6-4d00-9a14-83
12	71780	110624	2	918	158.43	0	316.86	82940b03-c70b-4183-8660-6b
13	71780	110625	4	780	1391.994	0	5567.976	644b0cd6-b2c3-4e4d-ab43-09
14	71780	110626	1	937	48.594	0	48.594	7f5feb17-8ef4-4236-9f1c-1504
15	71780	110627	6	867	41.994	0	251.964	ac78838d-b503-41a5-9791-48
16	71780	110628	1	985	112.998	0.4	67.799	2c10a262-a13d-442a-8f45-f4d
17	71780	110629	2	989	323.994	0	647.988	654fb79e-70df-4b92-9832-9fa

Query settings

Name

LakeFileOrderDetailLinesP...

Applied steps

- AdfDoc
- Parquet

What can a Wrangling Data Flow do? - Home

Control Flow



Data Flow

The screenshot displays the Power Query Editor interface. The main area shows a table with the following data:

SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice
71774	110562	1	836	356.898
71774	110563	1	822	356.898
71776	110567	1	907	63.9
71780	110616	4	905	218.454
71780	110617	2	983	461.694
71780	110618	6	988	112.998
71780	110619	2	748	818.7
71780	110620	1	990	323.994
71780	110621	1	926	149.874
71780	110622	1	743	809.76
71780	110623	4	782	1376.994
71780	110624	2	918	158.43
71780	110625	4	780	1391.994
71780	110626	1	937	48.594
71780	110627	6	867	41.994
71780	110628	1	985	112.998
71780	110629	2	989	323.994

The interface includes a ribbon with tabs: Home, Transform, Add Column, View, Tools, and Help. The right-hand pane shows the 'Query Settings' and 'Applied Steps' sections. The 'Applied Steps' list includes 'Source', 'Promoted Headers', and 'Changed Type'.

What can a Wrangling Data Flow do? - Transform

Control Flow



Data Flow

The screenshot displays the Power Query Editor with the 'Transform' tab selected. The main data view shows a table with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, UnitPrice, and UnitPrice. The 'Query Settings' pane on the right shows the 'APPLIED STEPS' list, which includes 'Source', 'Promoted Headers', and 'Changed Type'.

Below is a table representing the data shown in the editor:

	SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice	UnitPrice
1	71774	110562	1	836	356.898	
2	71774	110563	1	822	356.898	
3	71776	110567	1	907	63.9	
4	71780	110616	4	905	218.454	
5	71780	110617	2	983	461.694	
6	71780	110618	6	988	112.998	
7	71780	110619	2	748	818.7	
8	71780	110620	1	990	323.994	
9	71780	110621	1	926	149.874	
10	71780	110622	1	743	809.76	
11	71780	110623	4	782	1376.994	
12	71780	110624	2	918	158.43	
13	71780	110625	4	780	1391.994	
14	71780	110626	1	937	48.594	
15	71780	110627	6	867	41.994	
16	71780	110628	1	985	112.998	
17	71780	110629	2	989	323.994	

What can a Wrangling Data Flow do? - Add Column

Control Flow



Data Flow

The screenshot shows the Power Query Editor interface. The 'Add Column' tab is active, displaying various transformation options like 'Conditional Column', 'Index Column', 'Duplicate Column', 'Format', 'Merge Columns', 'Statistics', 'Standard Scientific', 'Trigonometry', 'Rounding', 'Date', 'Time', 'Duration', 'Text Analytics', 'Vision', and 'Azure Machine Learning'. The main area shows a table with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, and UnitPrice. The 'Query Settings' pane on the right shows the 'Properties' and 'Applied Steps' sections.

SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice
71774	110562	1	836	356.898
71774	110563	1	822	356.898
71776	110567	1	907	63.9
71780	110616	4	905	218.454
71780	110617	2	983	461.694
71780	110618	6	988	112.998
71780	110619	2	748	818.7
71780	110620	1	990	323.994
71780	110621	1	926	149.874
71780	110622	1	743	809.76
71780	110623	4	782	1376.994
71780	110624	2	918	158.43
71780	110625	4	780	1391.994
71780	110626	1	937	48.594
71780	110627	6	867	41.994
71780	110628	1	985	112.998
71780	110629	2	989	323.994

What can a Wrangling Data Flow do? - View

Control Flow



Data Flow

The screenshot displays the Power Query Editor interface. The top ribbon includes tabs for Home, Transform, Add column, and View. The left sidebar shows a list of queries, with 'OrderDetailLines' selected. The main area displays a data table with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, and UnitPrice. The table contains 17 rows of data. The bottom right pane shows the 'Query Settings' for 'OrderDetailLines', including the 'APPLIED STEPS' list which includes 'Source', 'Promoted Headers', and 'Changed Type'.

SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice
71774	110562	1	836	356.898
71774	110563	1	822	356.898
71776	110567	1	907	63.9
71780	110616	4	905	218.454
71780	110617	2	983	461.694
71780	110618	6	988	112.998
71780	110619	2	748	818.7
71780	110620	1	990	323.994
71780	110621	1	926	149.874
71780	110622	1	743	809.76
71780	110623	4	782	1376.994
71780	110624	2	918	158.43
71780	110625	4	780	1391.994
71780	110626	1	937	48.594
71780	110627	6	867	41.994
71780	110628	1	985	112.998
71780	110629	2	989	323.994

What can a Wrangling Data Flow do? - View



Data Flow



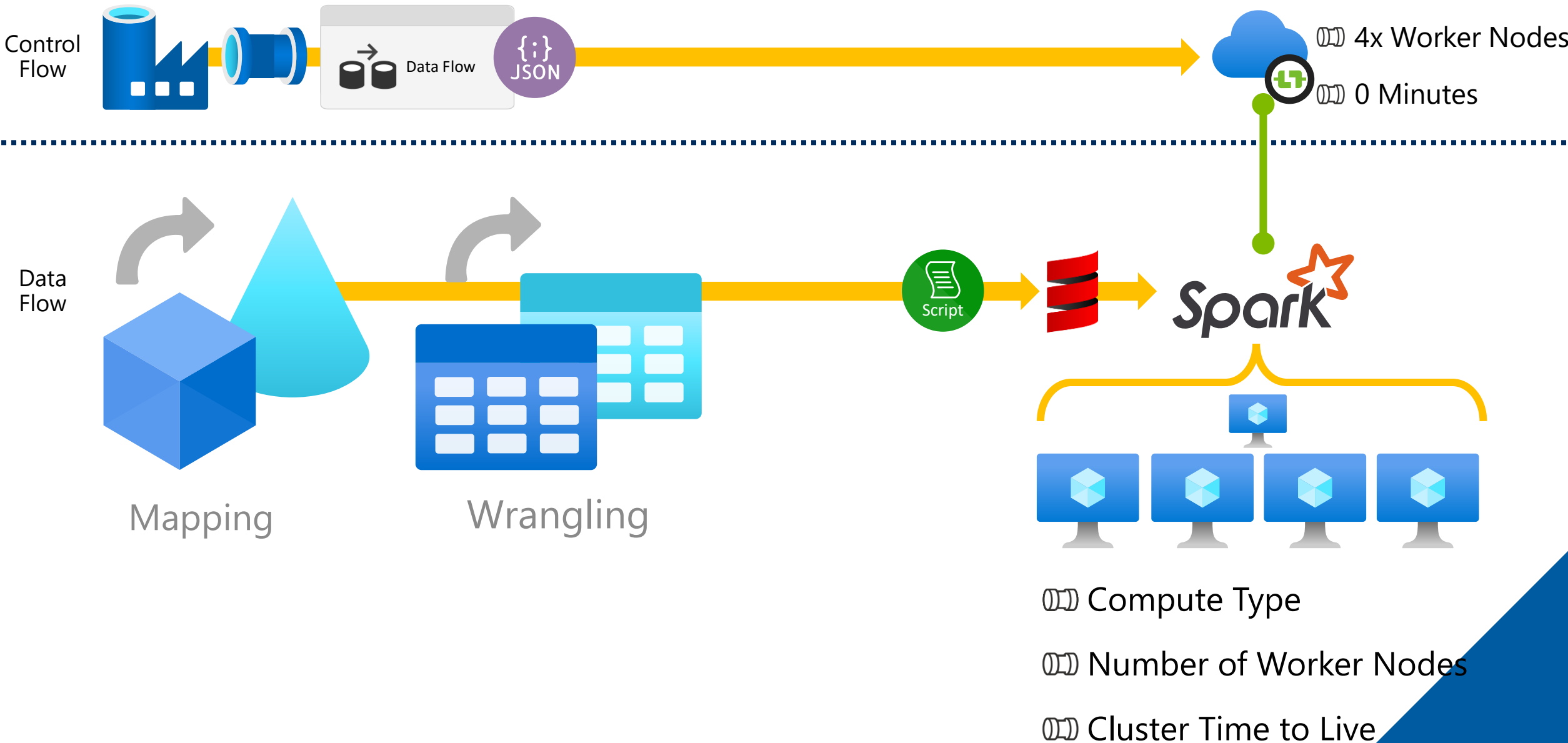


Wrangling Data Flow

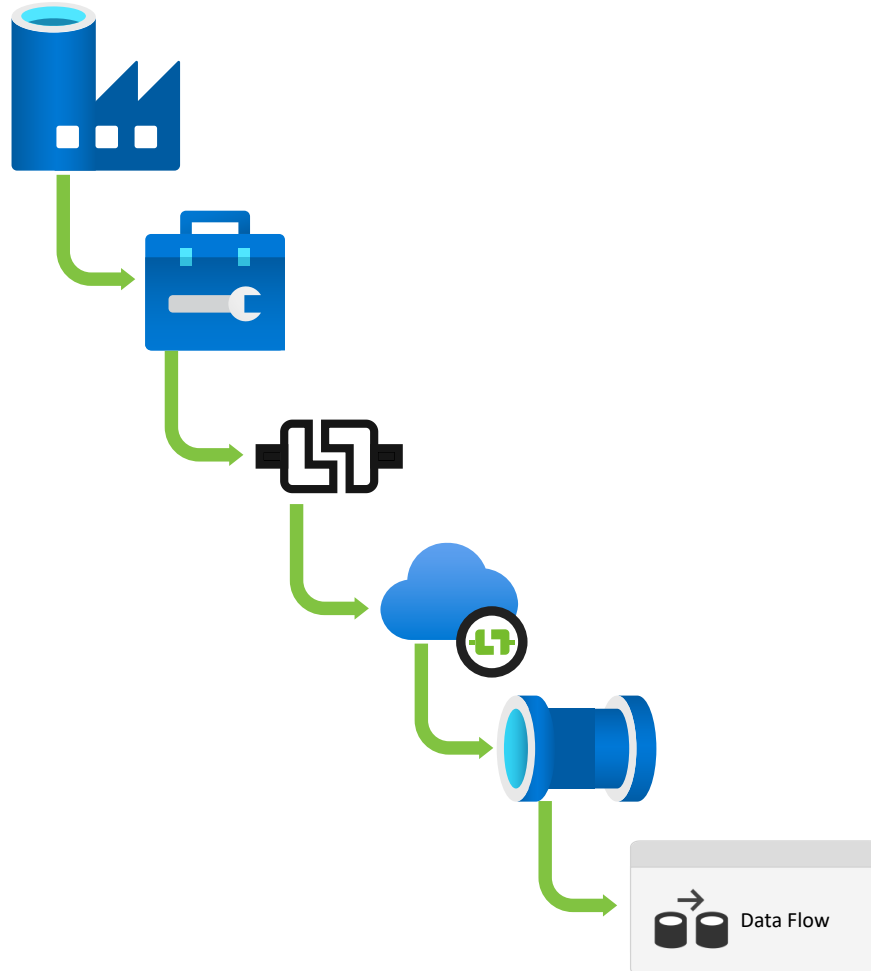
Configuration



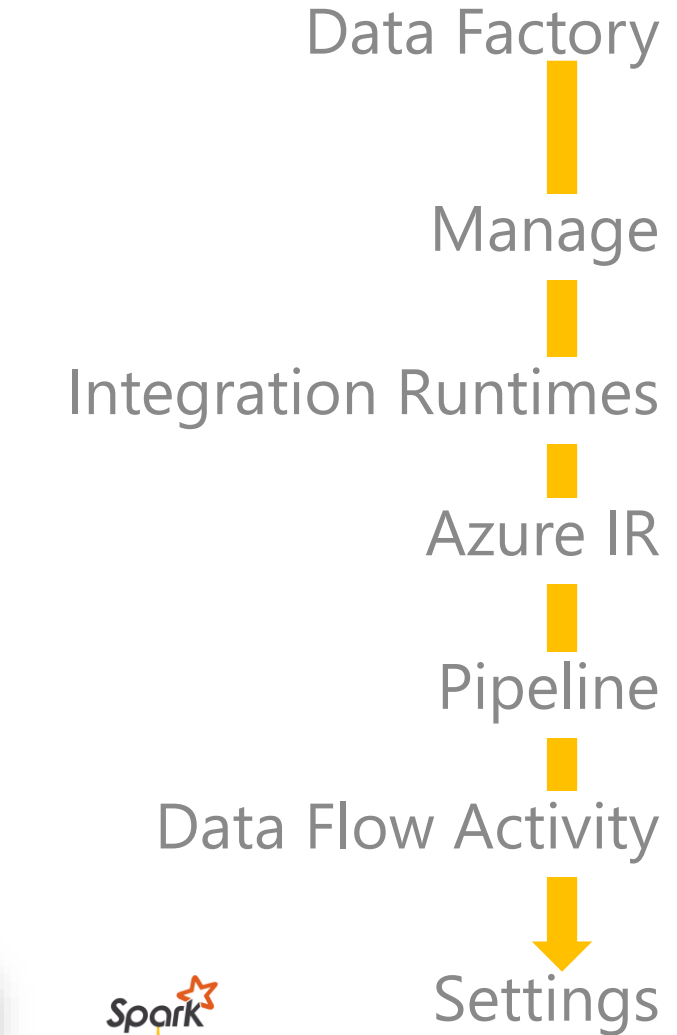
Data Flow Cluster Configuration



Setting the Data Flow Cluster (IR Configuration)






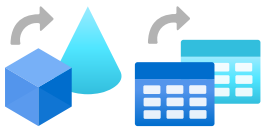
General	Settings	Parameters	User properties
Data flow *			
		MappingOrderAggregation	▼
Run on (Azure IR) *			
		DataFlowDemosTTL4Hours	▼ ⓘ
▶ PolyBase ⓘ			



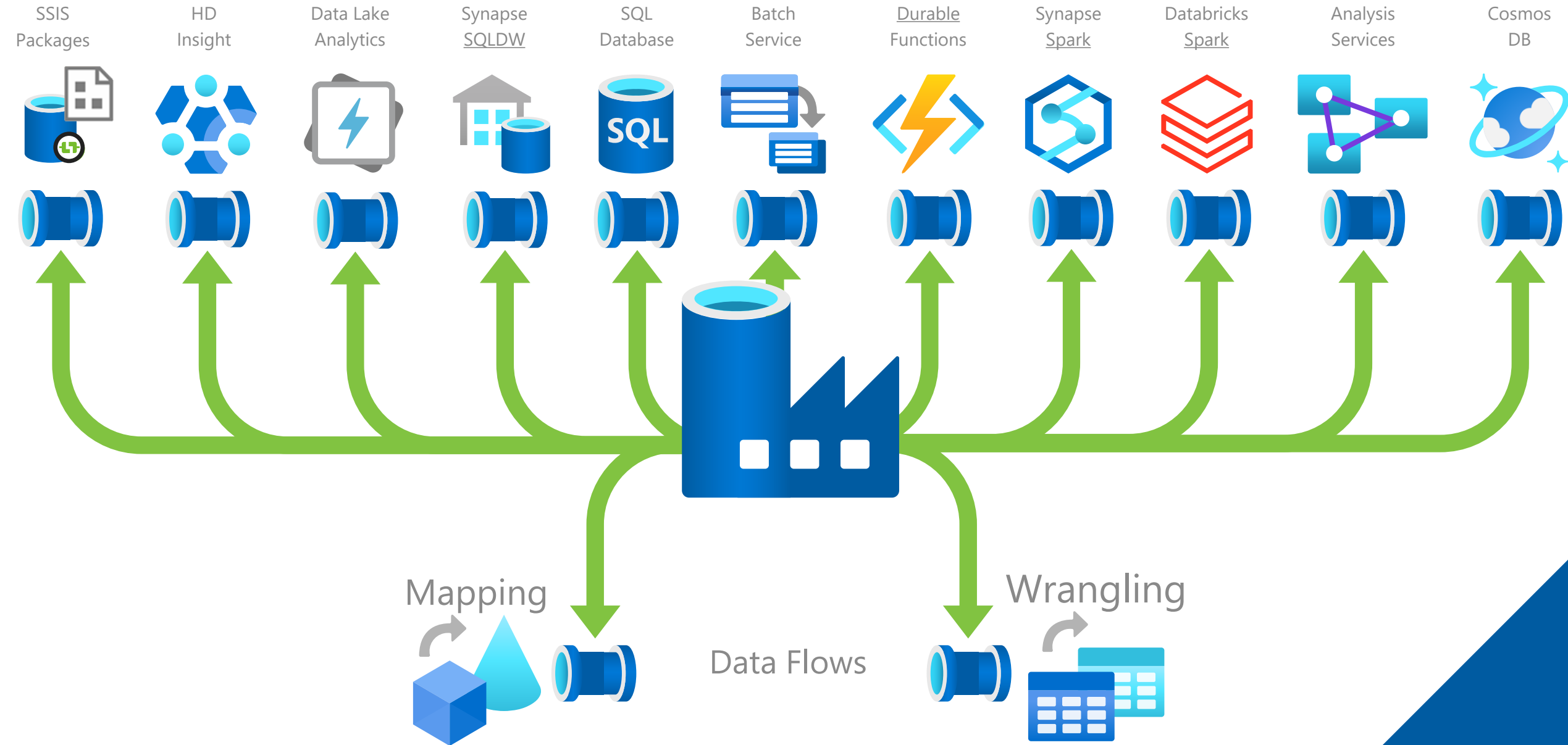
Use Cases & Conclusions



Data Transformations in Azure Comparisons

Transformation Method		Graphical UI	Scales Out	Scales Up	Cloud Native Tech
	T-SQL (SQLDB)	✗	✗	✓	✗
	SSIS	✓	✗	✓	✗
	Scala (Databricks)	✗	✓	✓	✓
	Data Factory Data Flows	✓	✓	✓	✓

When Should We Use Data Flows?



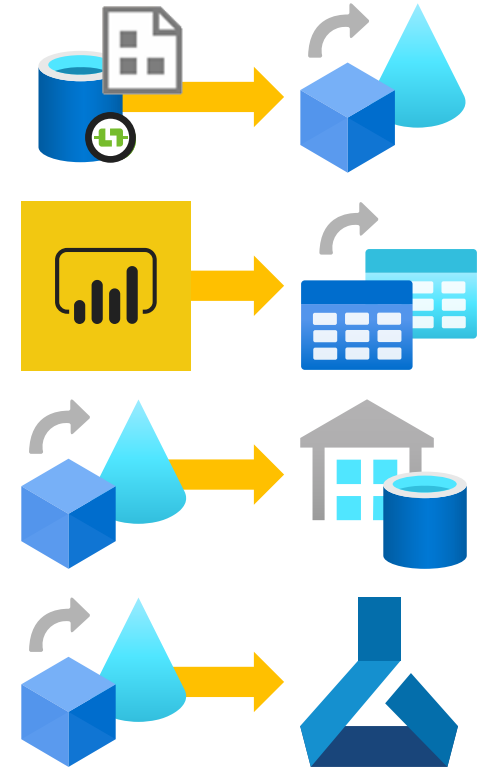
Use Cases

SSIS developers who are transferring existing skills to cloud native technologies have a very low barrier to entry and don't need to worry about distributed compute to get started.

Data engineering made easy for the power users who has grown out of Power BI following a series of Data Lake exploration sessions.

Data insight teams needing to do rapid prototyping and data warehouse loading within a single Azure Resource making deployments simple and release cycles short.

Simpler and quicker data engineering for data scientists that want to quickly prepare raw data for model training and testing, also with the ability to use large amounts of compute.



Module 4: Data Flows

🔌 Mapping Data Flows



🔌 Wrangling Data Flows



🔌 Configuration



🔌 Use Cases

