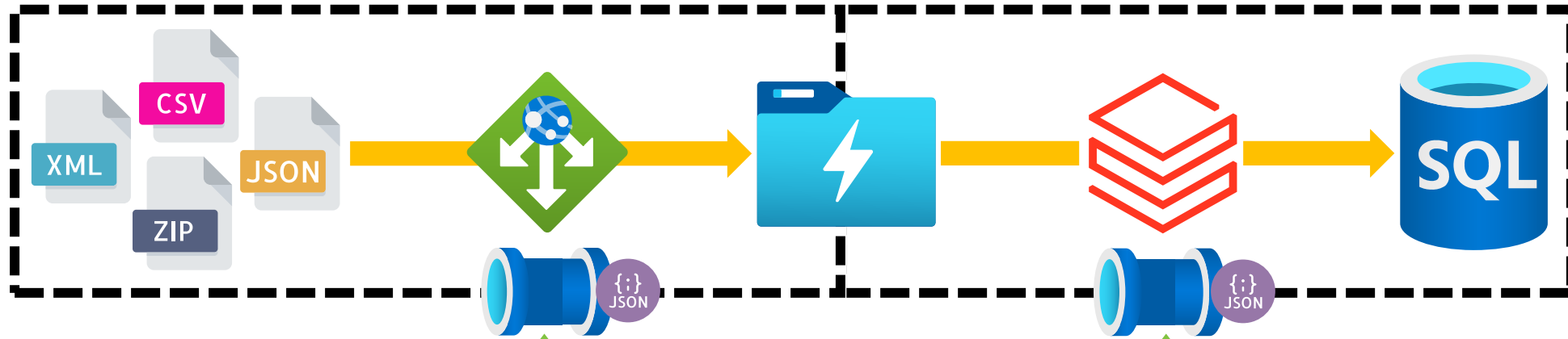


# Module 4: Data Flows

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



# Data Factory Control Flow Components



1 Linked Services

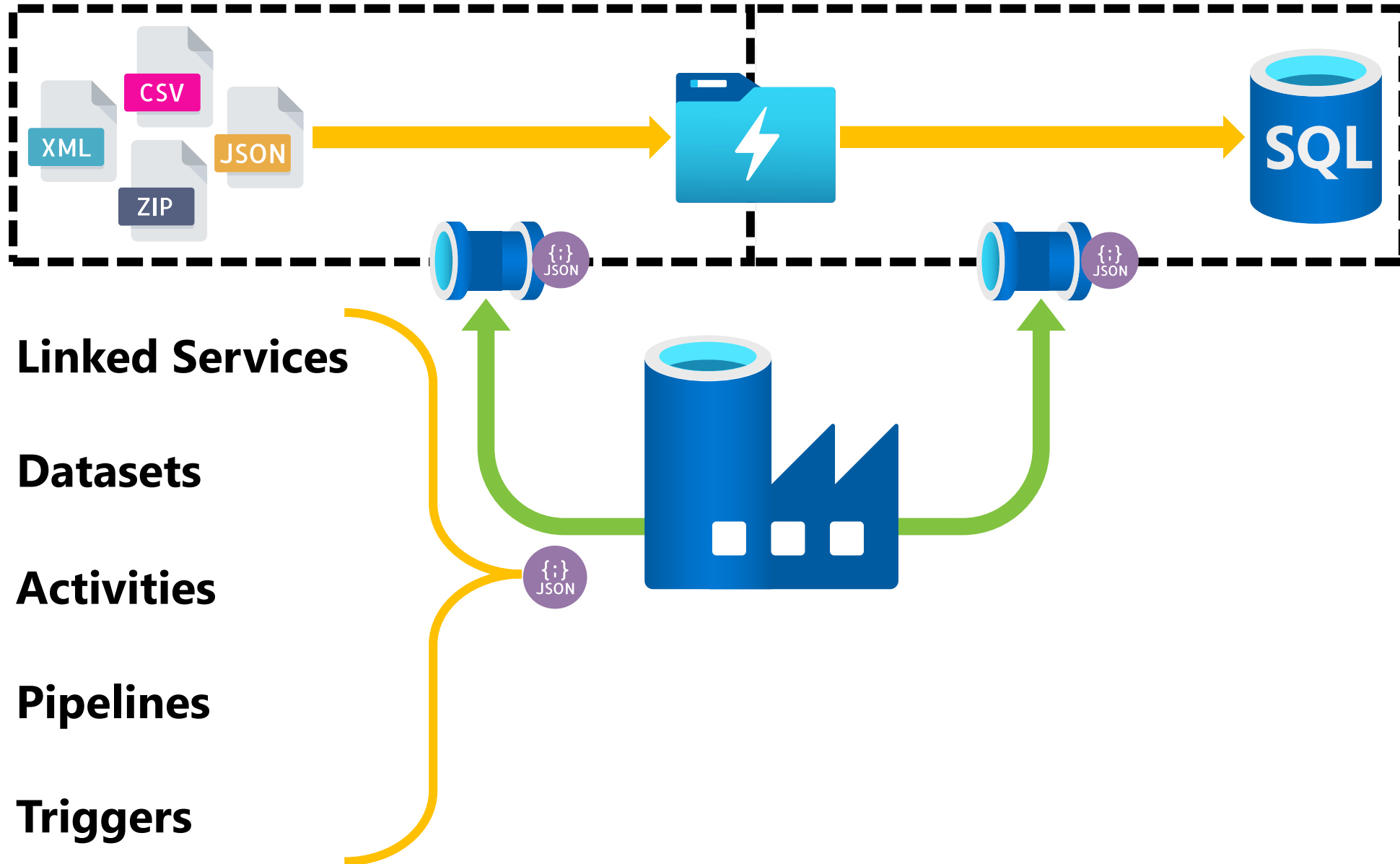
2 Datasets

3 Activities

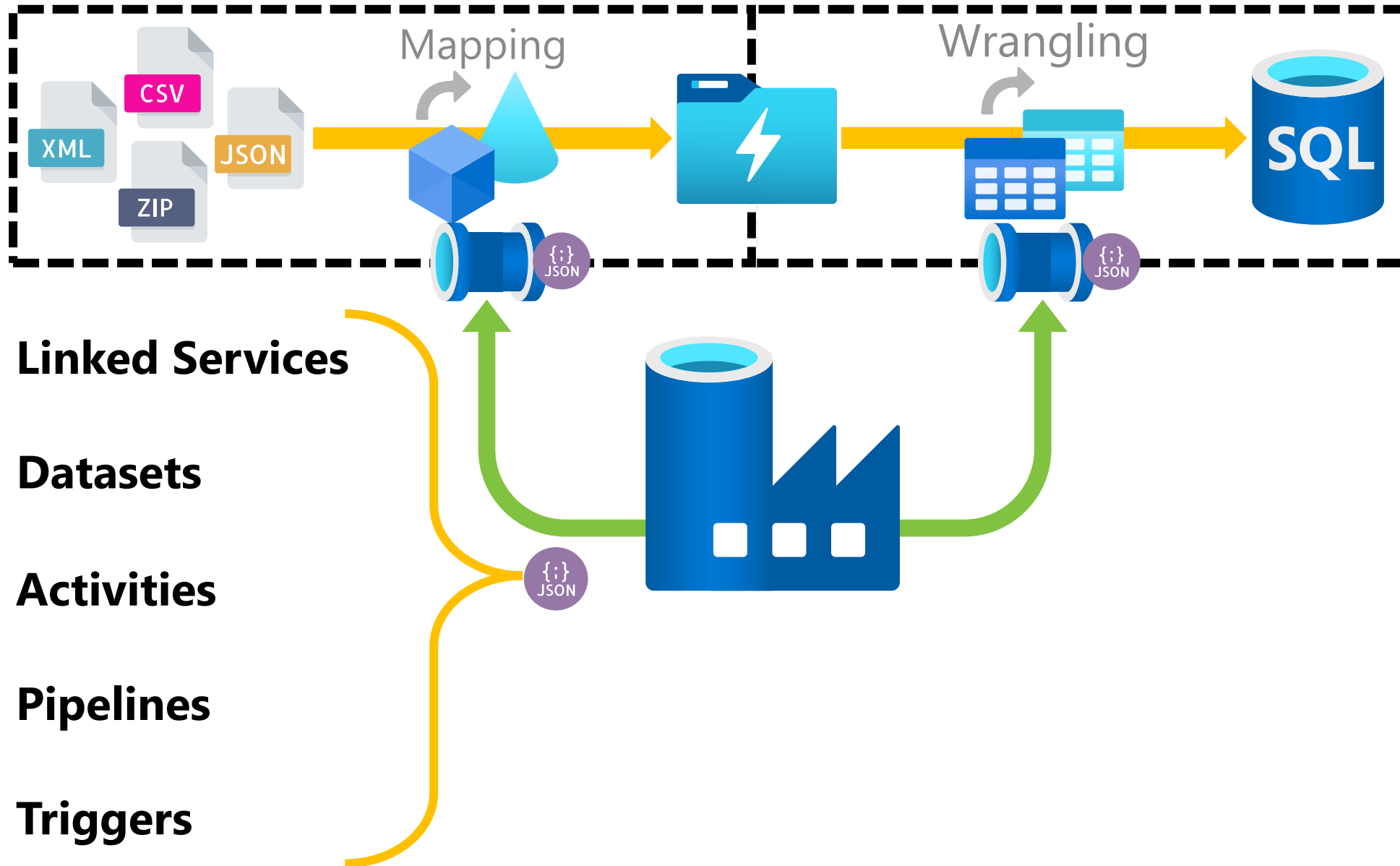
4 Pipelines

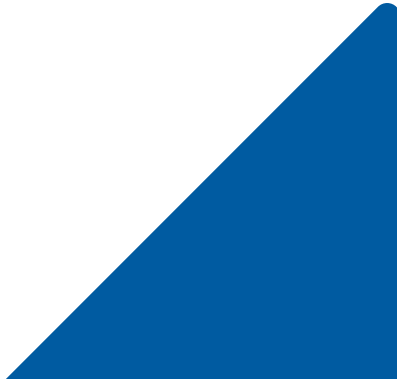
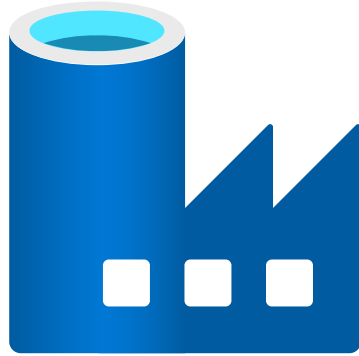
5 Triggers

# Data Factory Components

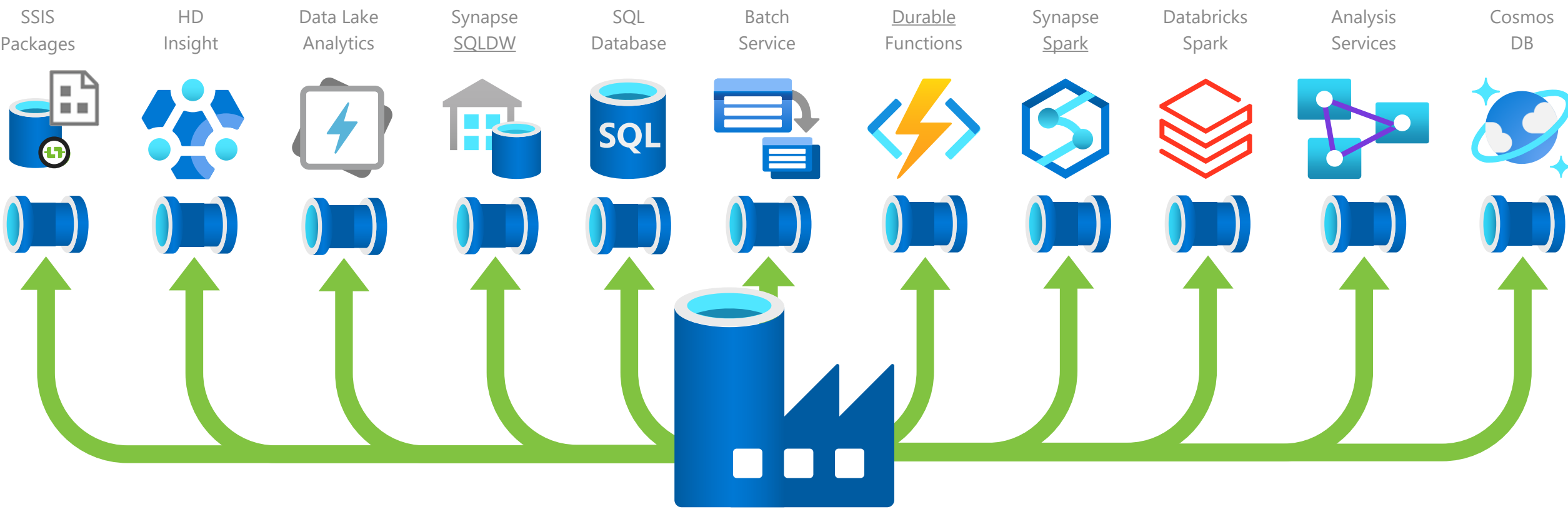


# Data Factory Data Flows

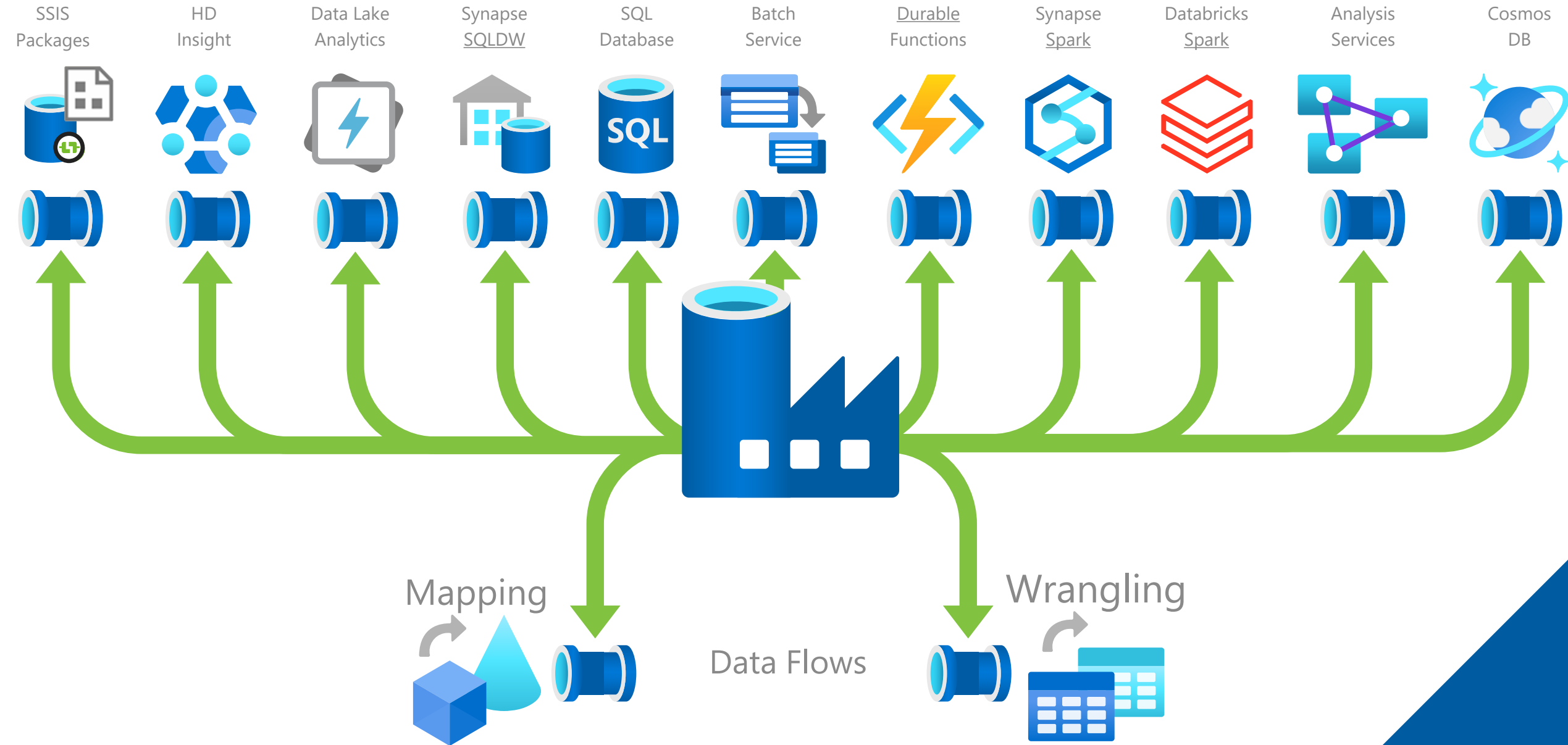




# Other Data Transformation Services in Azure



# When Should We Use Data Flows?

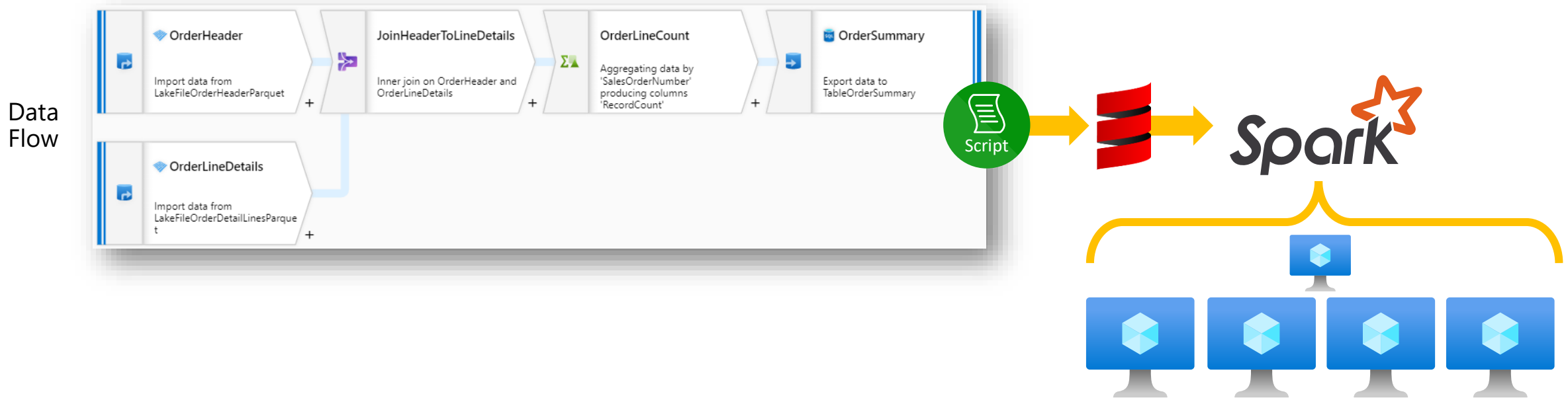


# Mapping Data Flows

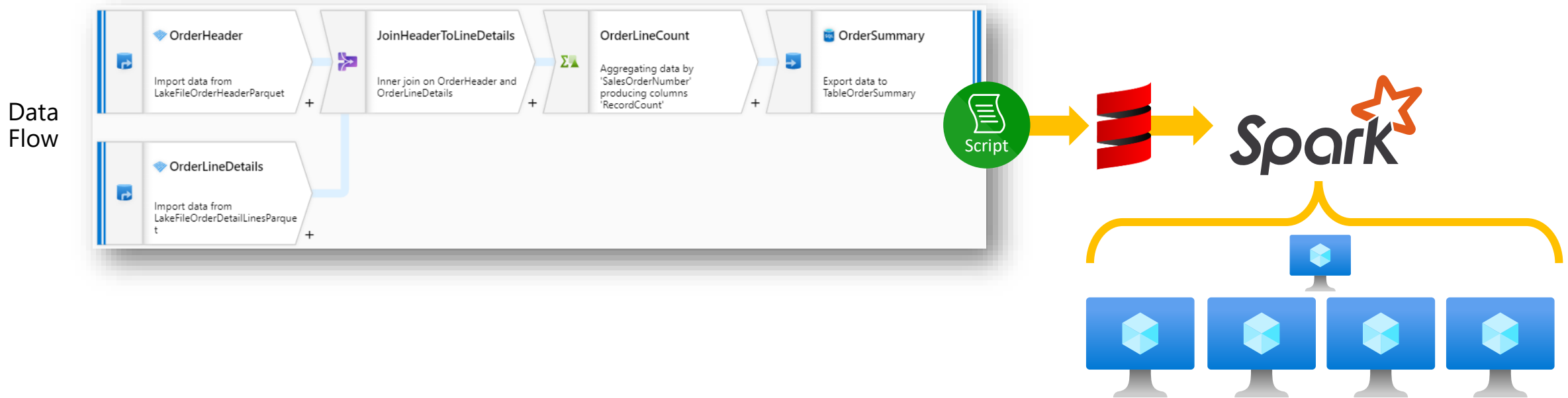




# What is a Mapping Data Flow?



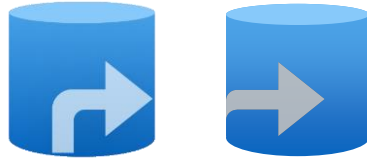
# Q: What is a Mapping Data Flow?



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

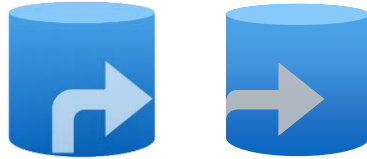


Limited File Type Support



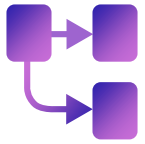
# What can a Mapping Data Flow do? - Inputs and Outputs

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



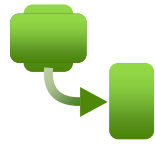
Exists



Union



Lookup



Derived Column



Select



Aggregate



Surrogate Key



Pivot



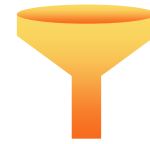
Unpivot



Window



Flatten



Filter



Sort



Alter Row

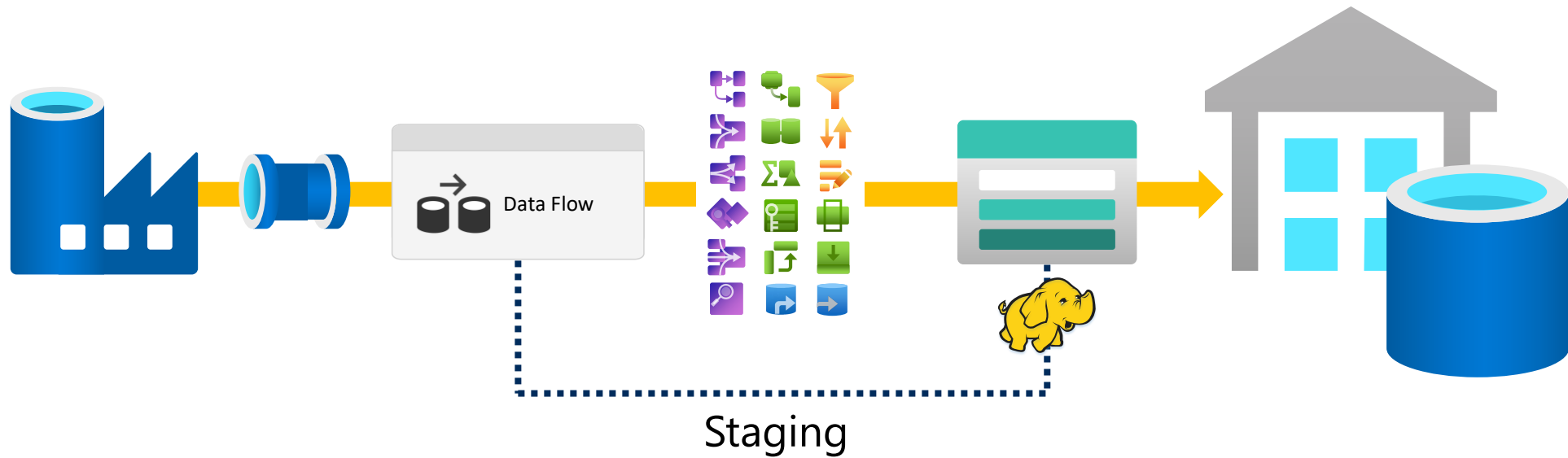
Key

**Input & Output Modifiers**

**Schema Modifiers**

**Row Modifiers**

# What can a Mapping Data Flow do? - PolyBase

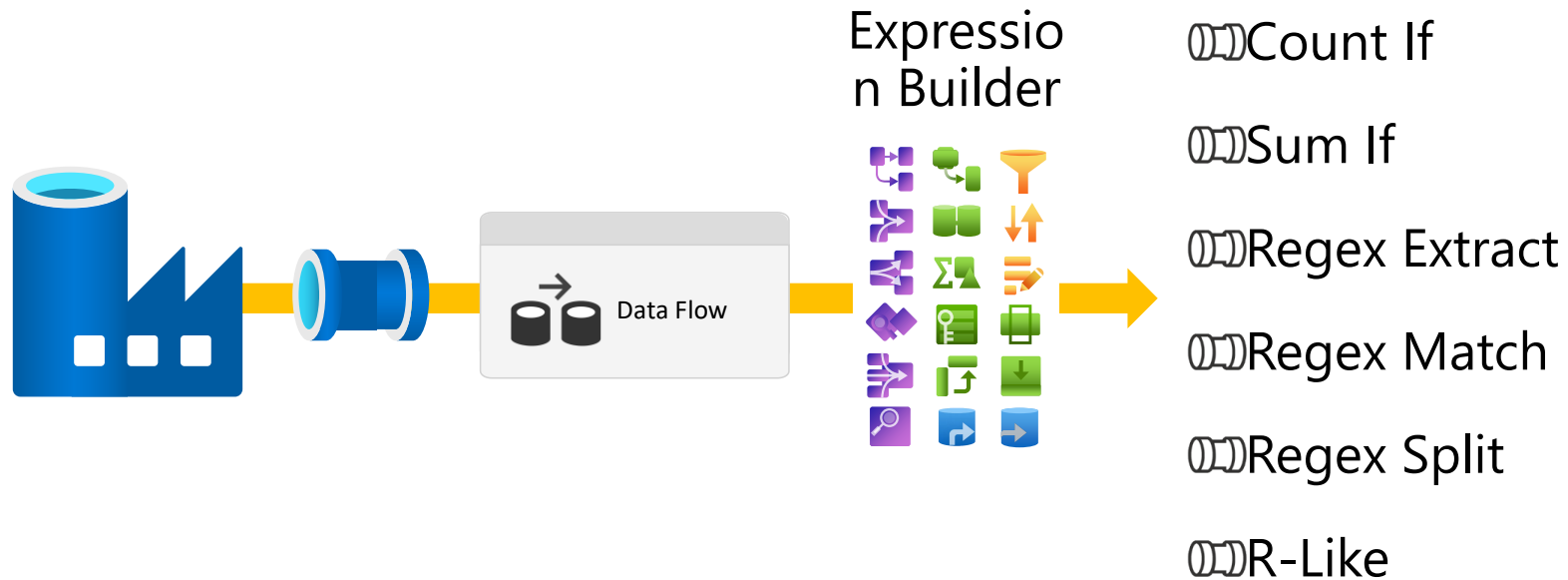


PolyBase ⓘ

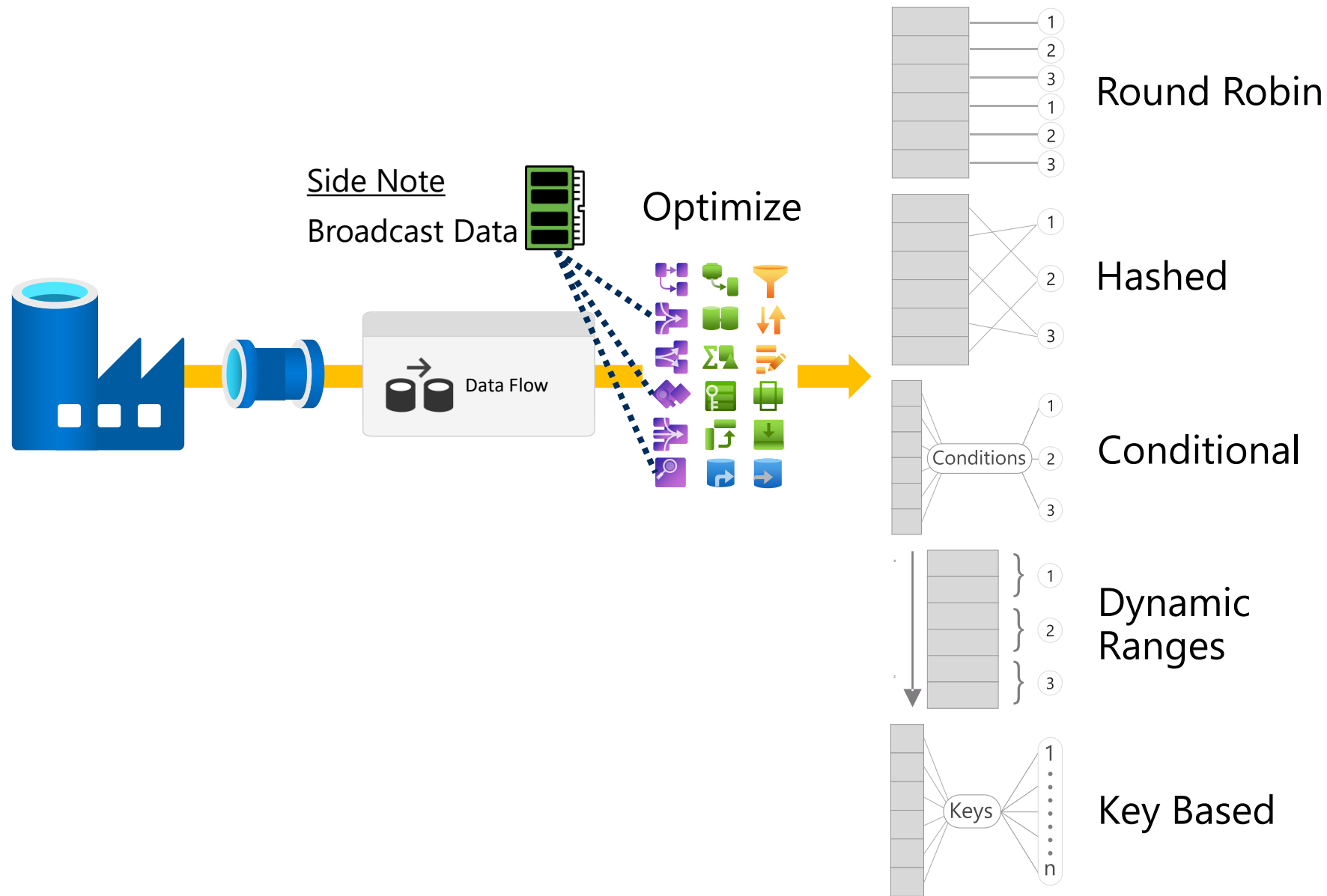
Staging linked service  ⓘ + New

Staging storage folder  /   ⓘ

# What can a Mapping Data Flow do? - Expression Builder



# What can a Mapping Data Flow do? - Partition Handling



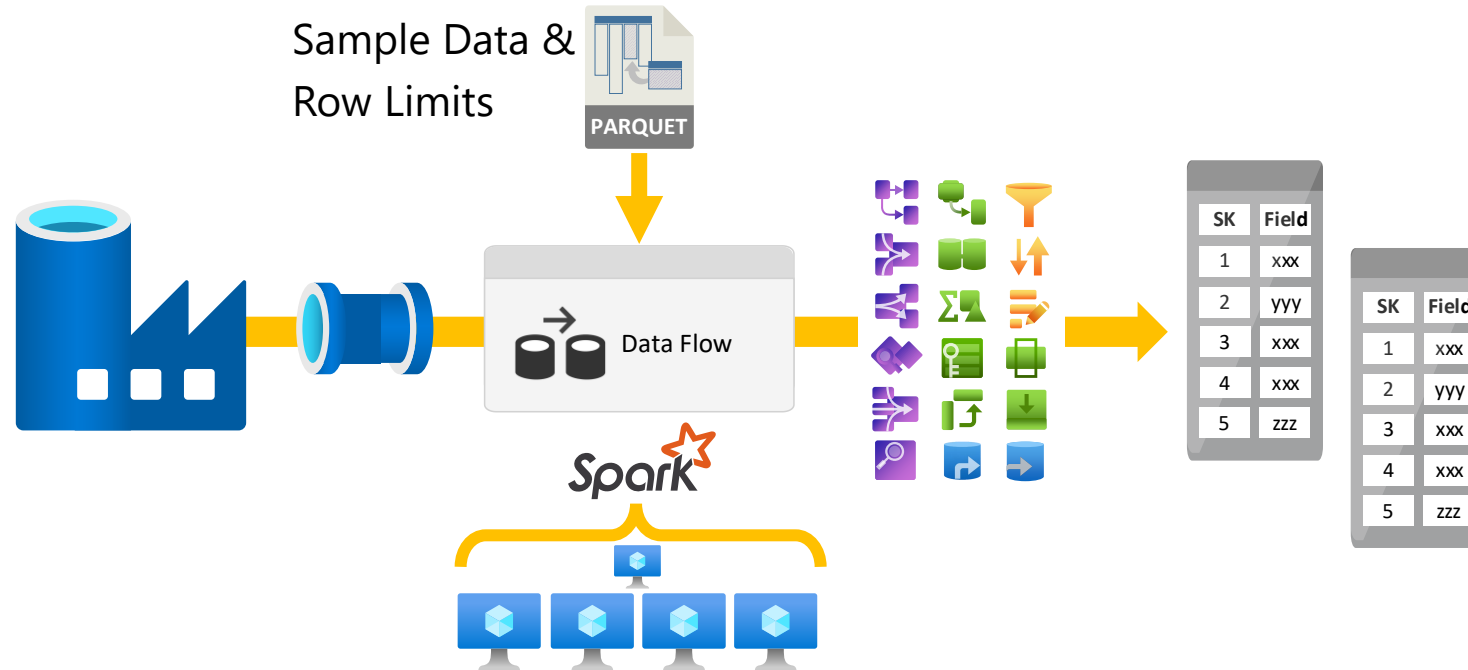


# What can a Mapping Data Flow do? - Debugging

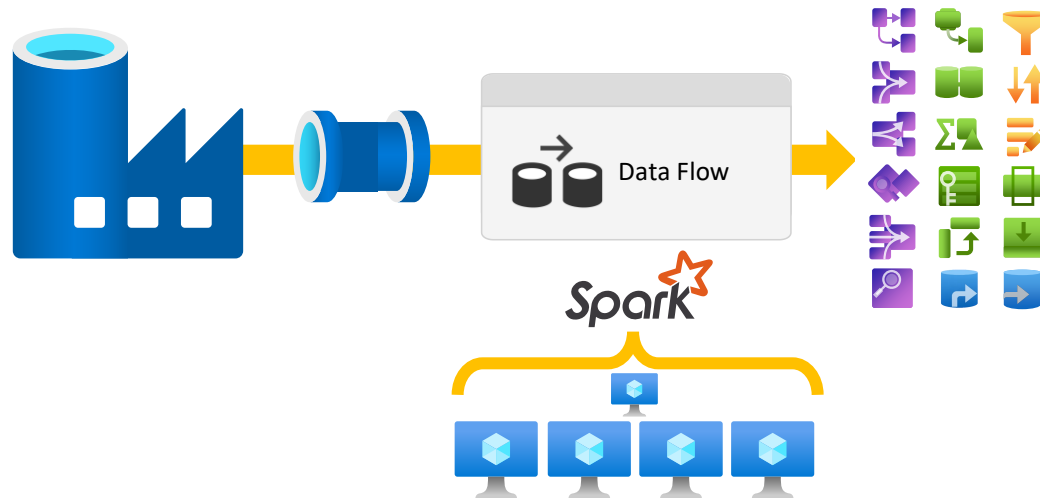


Enable Data Flow Debug Mode

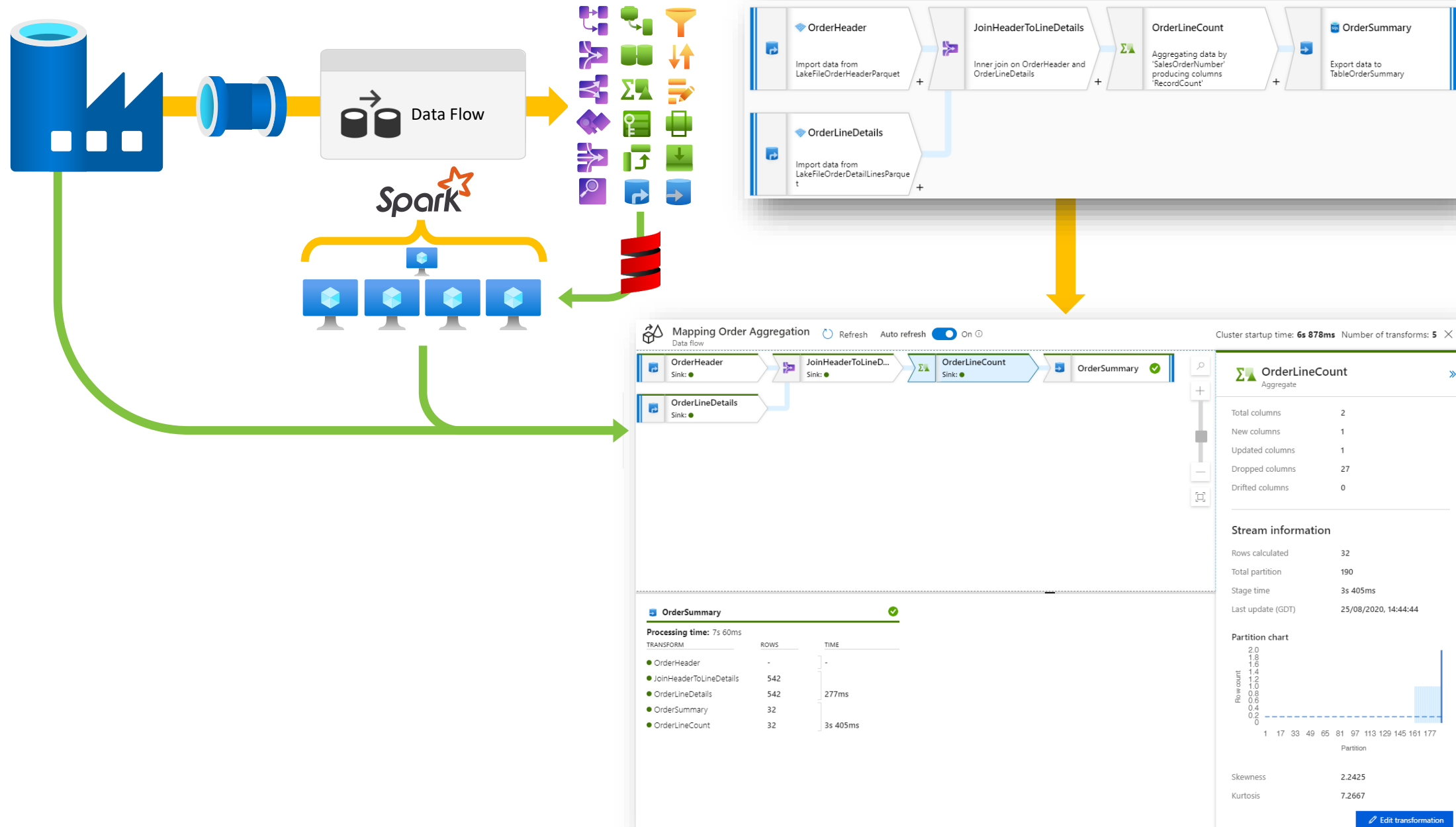
Data  
Preview

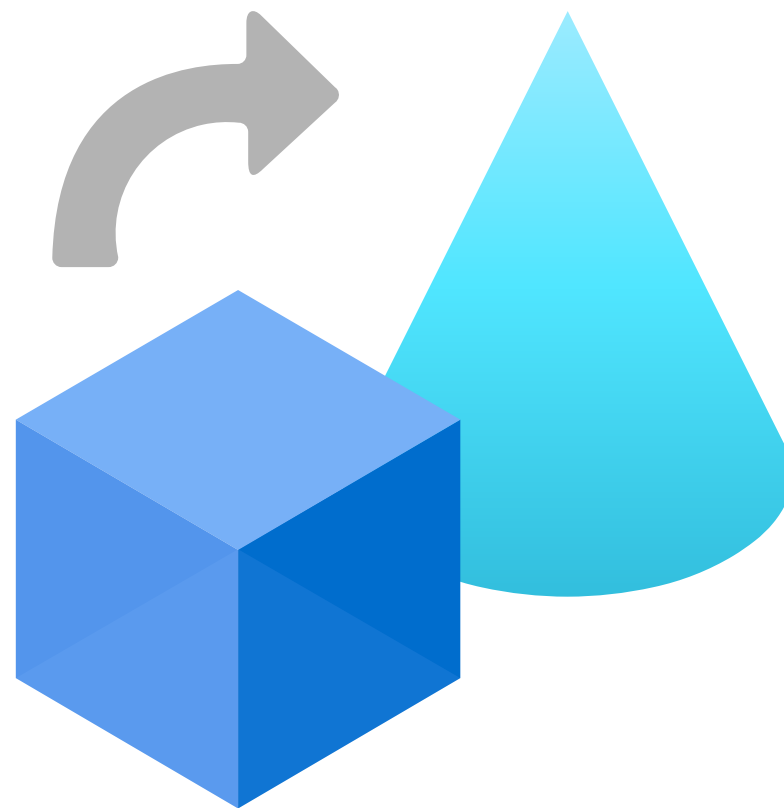


# What can a Mapping Data Flow do? - Monitoring



# What can a Mapping Data Flow do? - Monitoring





Mapping Data Flow

# Wrangling Data Flows

*(Preview)*



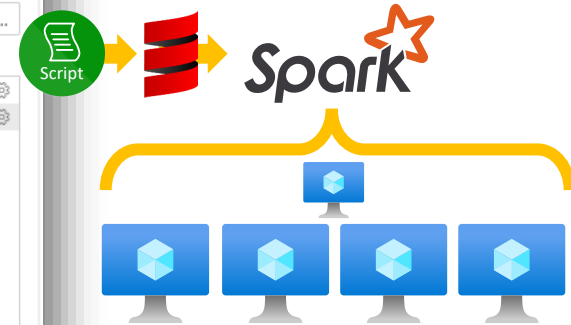
# What is a Wrangling Data Flow?



Data Flow

The screenshot shows the Databricks Data Wrangling interface. The top menu bar includes 'Home', 'Transform', 'Add column', and 'View'. The 'Home' tab is active, showing a toolbar with various data manipulation tools like 'Enter data', 'Options', 'Manage parameters', 'Refresh', 'Advanced editor', 'Manage', 'Choose columns', 'Remove columns', 'Keep rows', 'Remove rows', 'Sort', 'Split column', 'Group by', 'Data type', 'Merge queries', 'Append queries', 'Combine files', and 'Combine'. The main area displays a table with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, UnitPrice, UnitPriceDiscount, LineTotal, and rowguid. The table contains 17 rows of data. The right sidebar shows 'Query settings' with the name 'LakeFileOrderDetailLinesP...' and 'Applied steps' including 'AdfDoc' and 'Parquet'.

1 <sup>2</sup> SalesOrderID	1 <sup>2</sup> SalesOrderDetailID	1 <sup>2</sup> OrderQty	1 <sup>2</sup> ProductID	1.2 UnitPrice	1.2 UnitPriceDiscount	1.2 LineTotal	A <sup>B</sup> 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

The screenshot shows the Databricks Data Flow interface. The top menu bar includes Home, Transform, Add column, and View. Below the menu is a toolbar with various actions like Enter data, Options, Manage parameters, Refresh, Properties, Advanced editor, Manage, Choose columns, Remove columns, Keep rows, Remove rows, Sort, Split column, Group by, Data type, Use first row as headers, Replace values, Merge queries, Append queries, and Combine files.

The main area displays a table titled "Parquet.Document (AdfDoc)". The table has columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, UnitPrice, UnitPriceDiscount, LineTotal, and rowguid. The data is organized into 17 rows.

On the right side, there is a "Query settings" panel with a "Name" field containing "LakeFileOrderDetailLinesP..." and an "Applied steps" list showing "AdfDoc" and "Parquet".

1 <sup>2</sup> SalesOrderID	1 <sup>2</sup> SalesOrderDetailID	1 <sup>2</sup> OrderQty	1 <sup>2</sup> ProductID	1.2 UnitPrice	1.2 UnitPriceDiscount	1.2 LineTotal	A <sup>B</sup> 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

# What can a Wrangling Data Flow do? - Home

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 Home tab is active, showing various data manipulation options like 'Enter data', 'Options', 'Manage parameters', 'Refresh', 'Advanced editor', 'Manage', 'Choose columns', 'Remove columns', 'Keep rows', 'Remove rows', 'Sort', 'Split column', 'Group by', 'Data type', 'Merge queries', 'Append queries', and 'Combine files'. The 'Queries' pane on the left shows a list of queries: 'ADFRsource [1]', 'LakeFileOrderDetail...', and 'UserQuery'. The main workspace shows a table with columns: 'SalesOrderID', 'SalesOrderDetailID', 'OrderQty', 'ProductID', 'UnitPrice', and 'UnitPrice'. The table contains 17 rows of data. The 'Query Settings' pane on the right shows the 'Properties' tab with the name 'OrderDetailLines' and the 'Applied Steps' tab with a list of steps: 'Source', 'Promoted Headers', and 'Changed Type'.

SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice	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? - Transform

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 Transform tab is active, showing various data manipulation options like Transpose, Reverse rows, Replace values, Detect data type, Rename, Pivot column, Unpivot columns, Split column, Format, Merge columns, Statistics, Standard Scientific, Trigonometry, Rounding, and Information. The main area shows a data table with columns SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, UnitPrice, and UnitPrice. 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	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? - 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', 'Extract', 'Parse', 'Statistics', 'Standard Scientific', 'Trigonometry', 'Rounding', 'Information', 'Date', 'Time', and 'Duration'. The main area shows a table with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, and UnitPrice. The 'Query Settings' pane on the right shows the 'APPLIED STEPS' list with 'Changed Type' selected.

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

Home Transform Add column View

Data view Schema view Go to column Advanced editor

Preview Columns Advanced

Queries

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

File Home Transform Add Column View Tools Help

Query Settings

Layout Data Preview

Columns Parameters Advanced Dependencies

Query Settings

PROPERTIES

Name

OrderDetailLines

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- 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



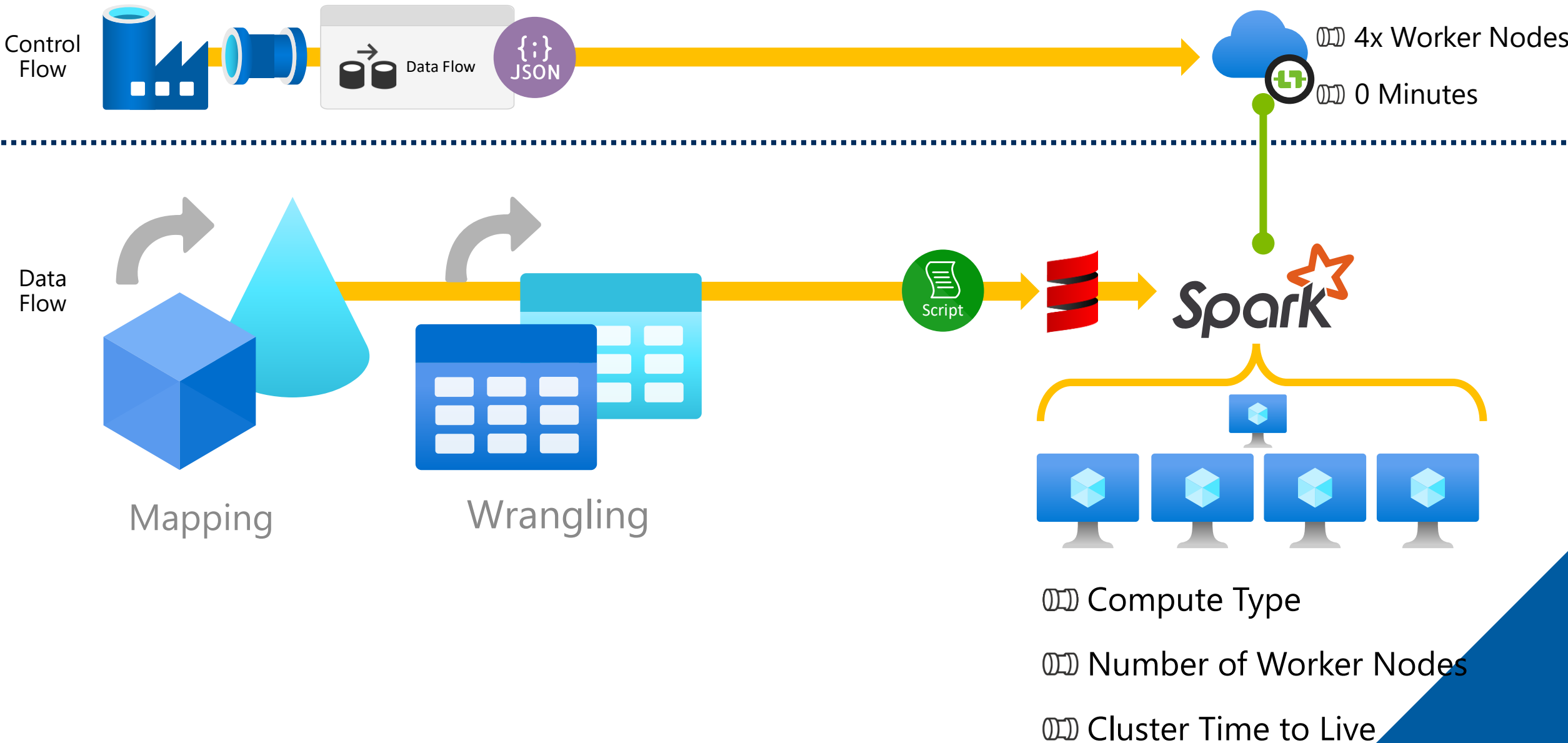


Wrangling Data Flow

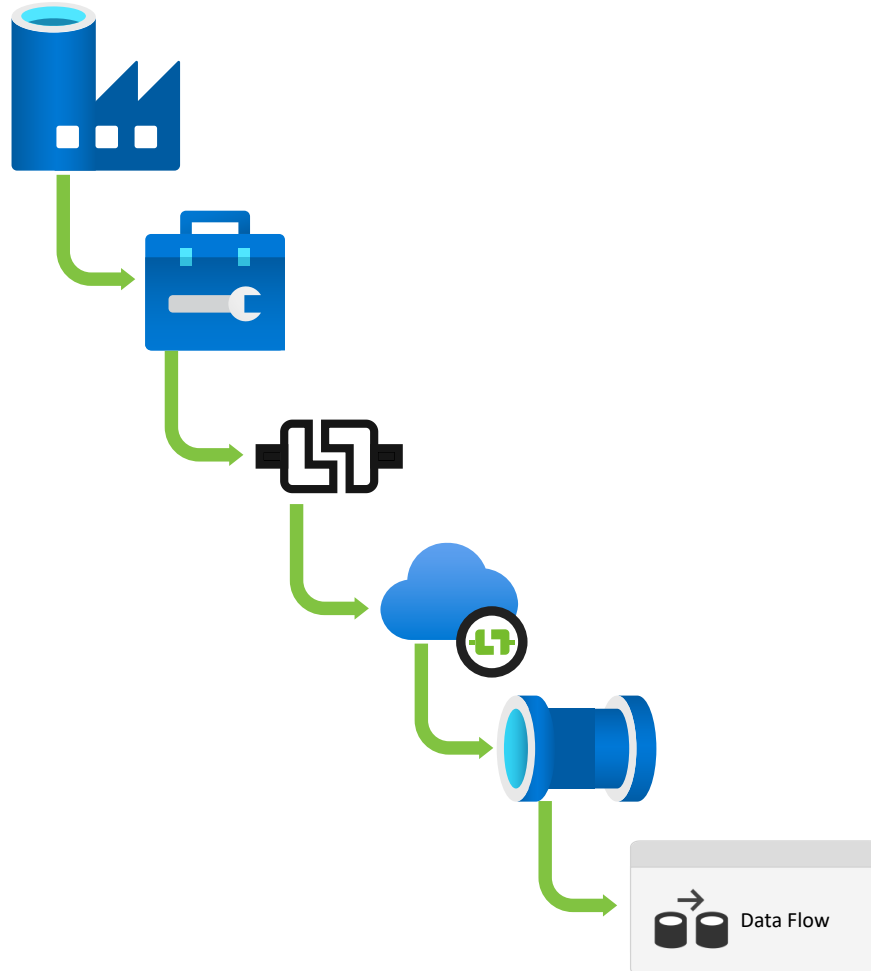
# Configuration



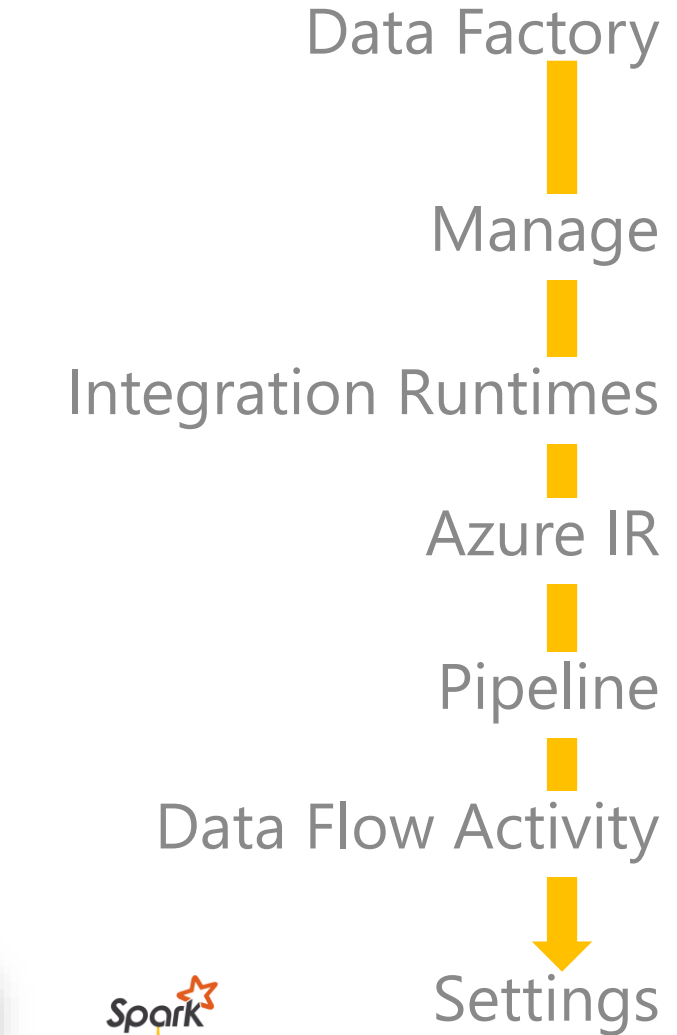
# Data Flow Cluster Configuration



# Setting the Data Flow Cluster (IR Configuration)



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




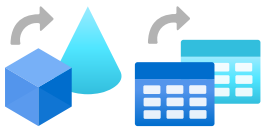




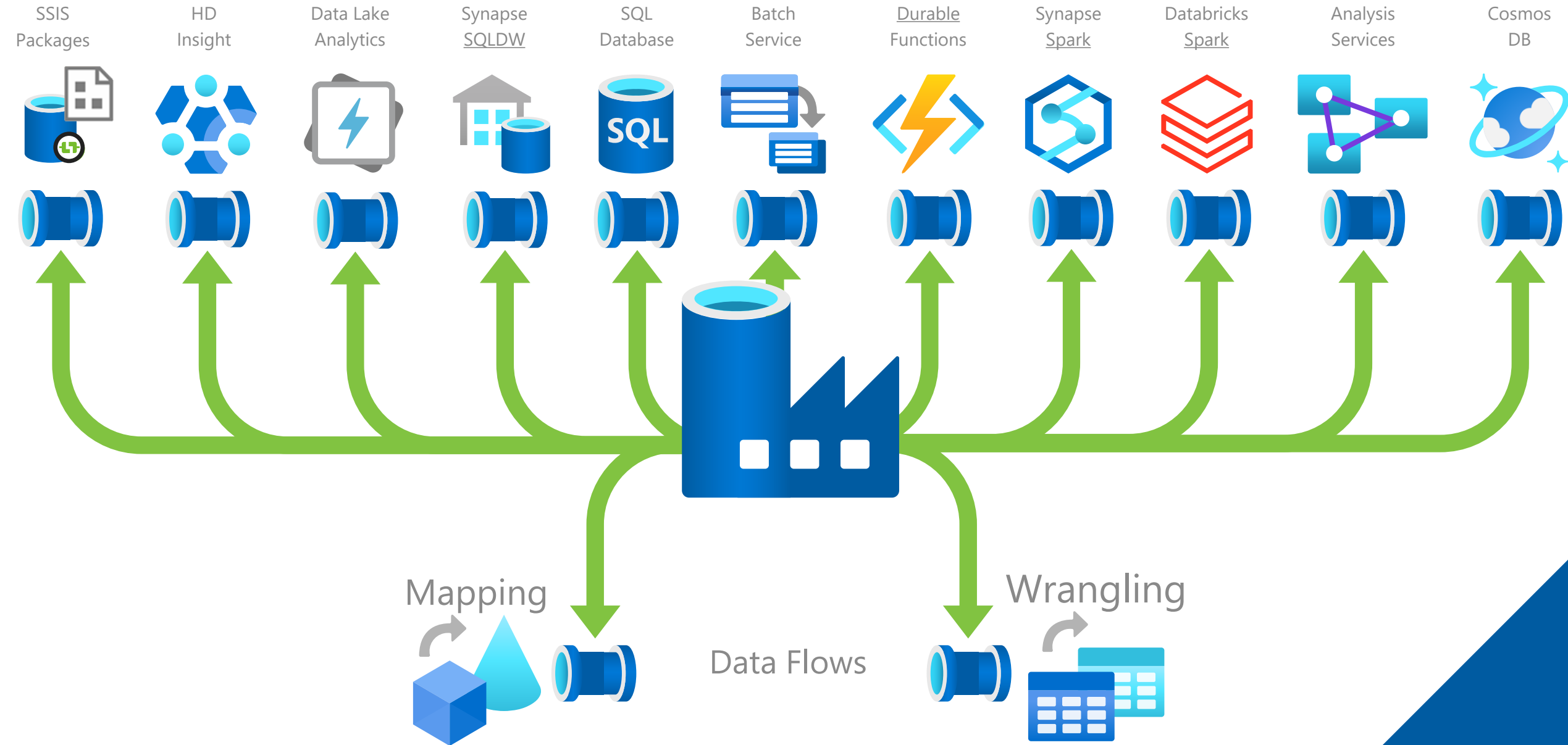
# Use Cases & Conclusions

A large, solid blue triangular shape that points towards the top right corner of the slide, occupying the right half of the image.

# 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?



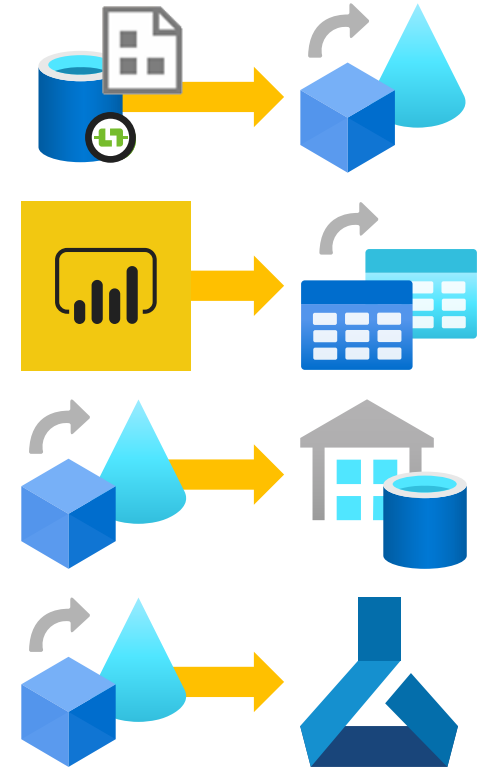
# Use Cases

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

