




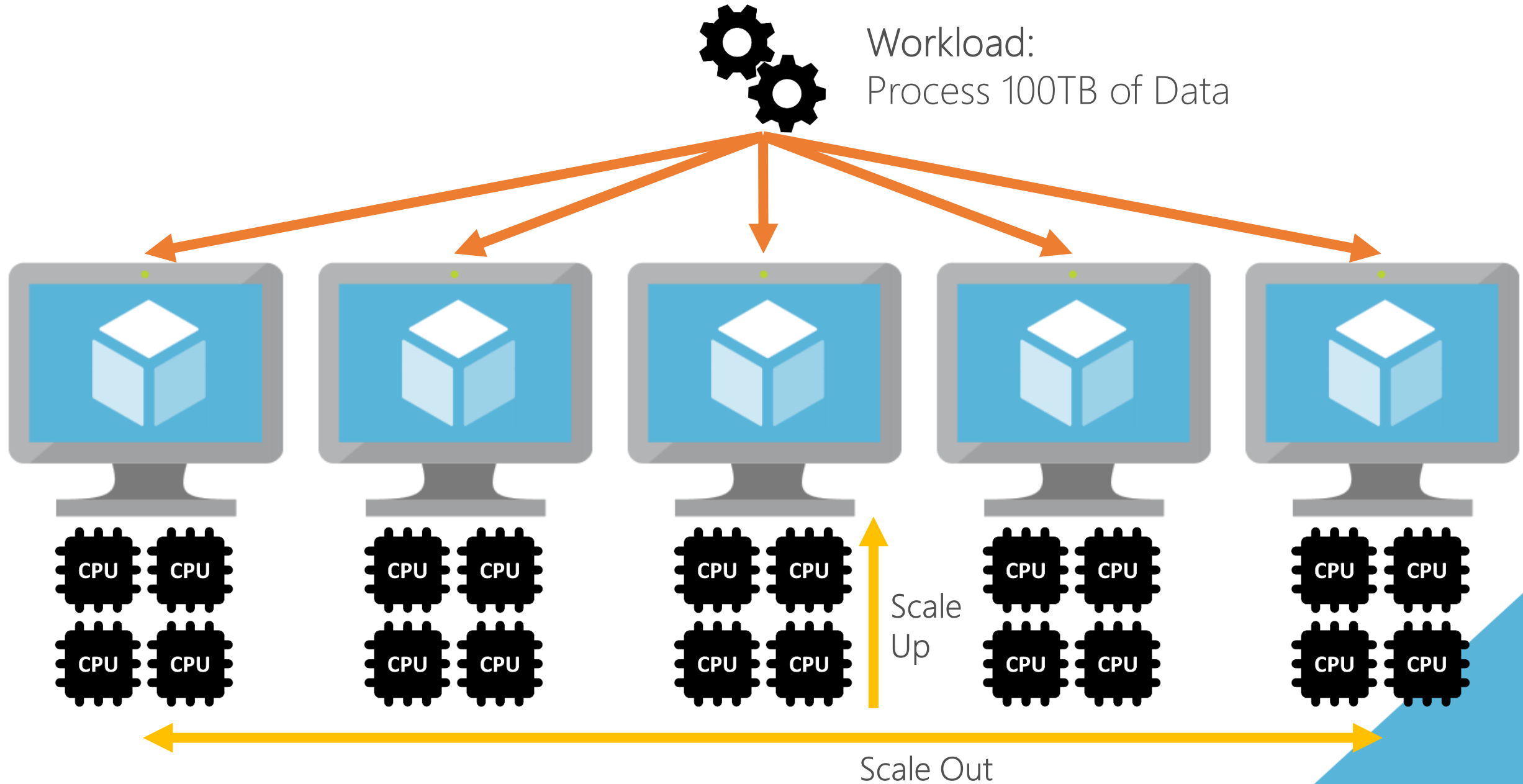
# Module 4: Data Flows

- Mapping Data Flows
  - Databricks Cluster Configuration
  - Wrangling Data Flows
- 

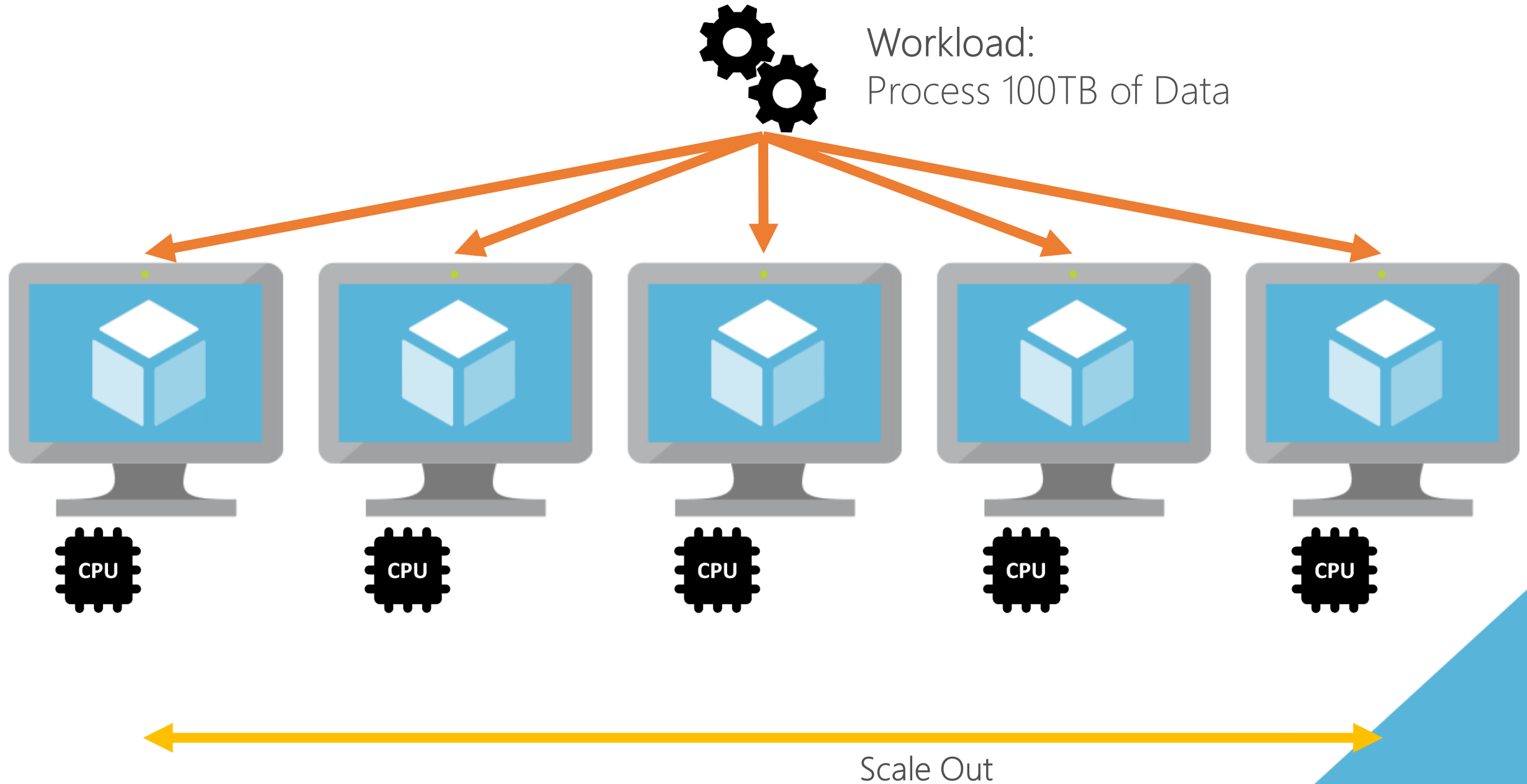
# Scale Up vs Scale Out

A large, solid blue triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

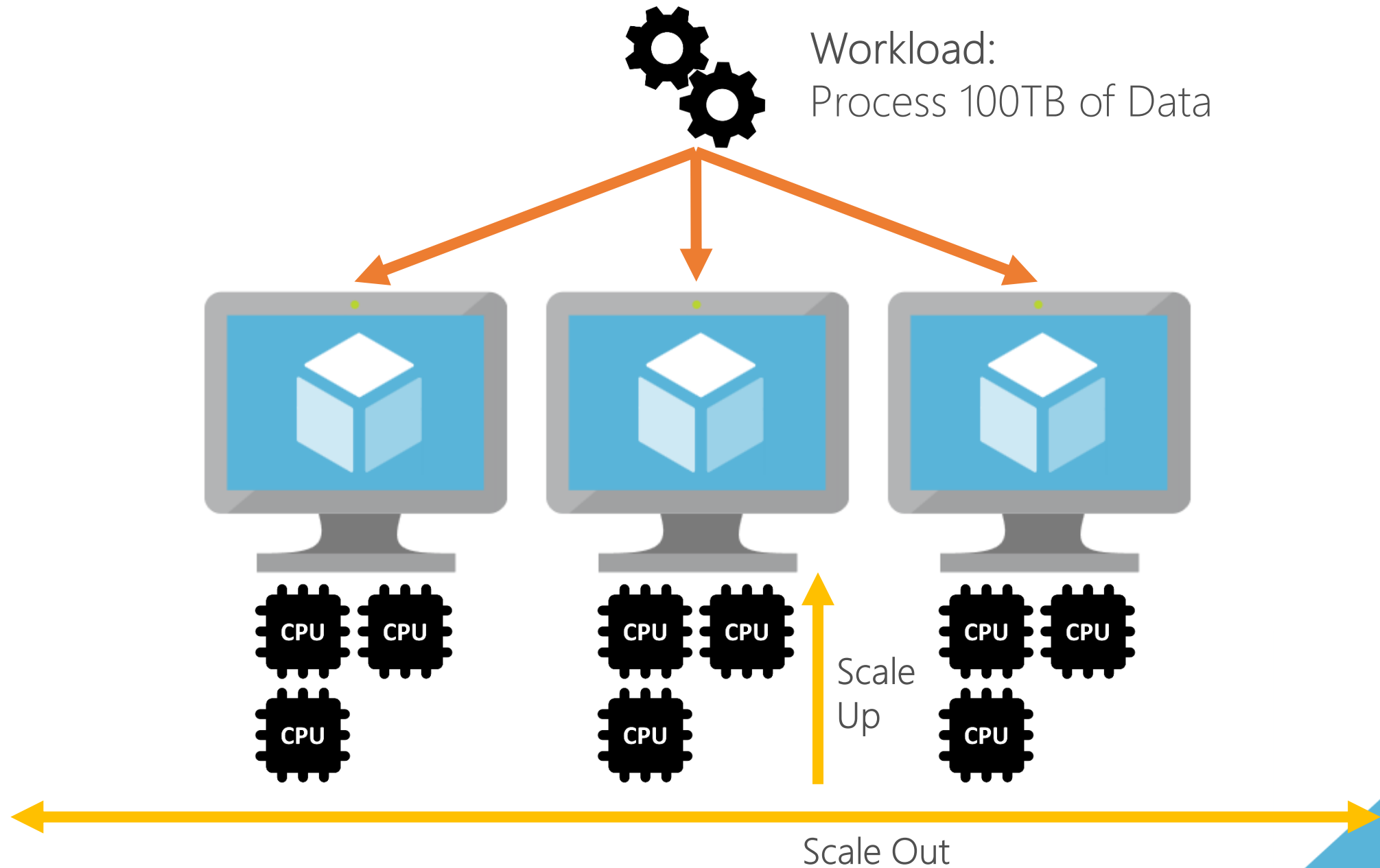
# Scale Up and Scale Out



# Scale Up and Scale Out



# Scale Up and Scale Out



# Azure Data Factory

A large, solid blue triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

# Data Transformation in zure

# Data Factory What & Why - Recap

**1 Linked Services**

**2 Data Sets**

**3 Activities**

**4 Pipelines**

**5 Triggers**

**1**

**Azure**

Integration Runtime

**2**

**SSIS**

Integration Runtime

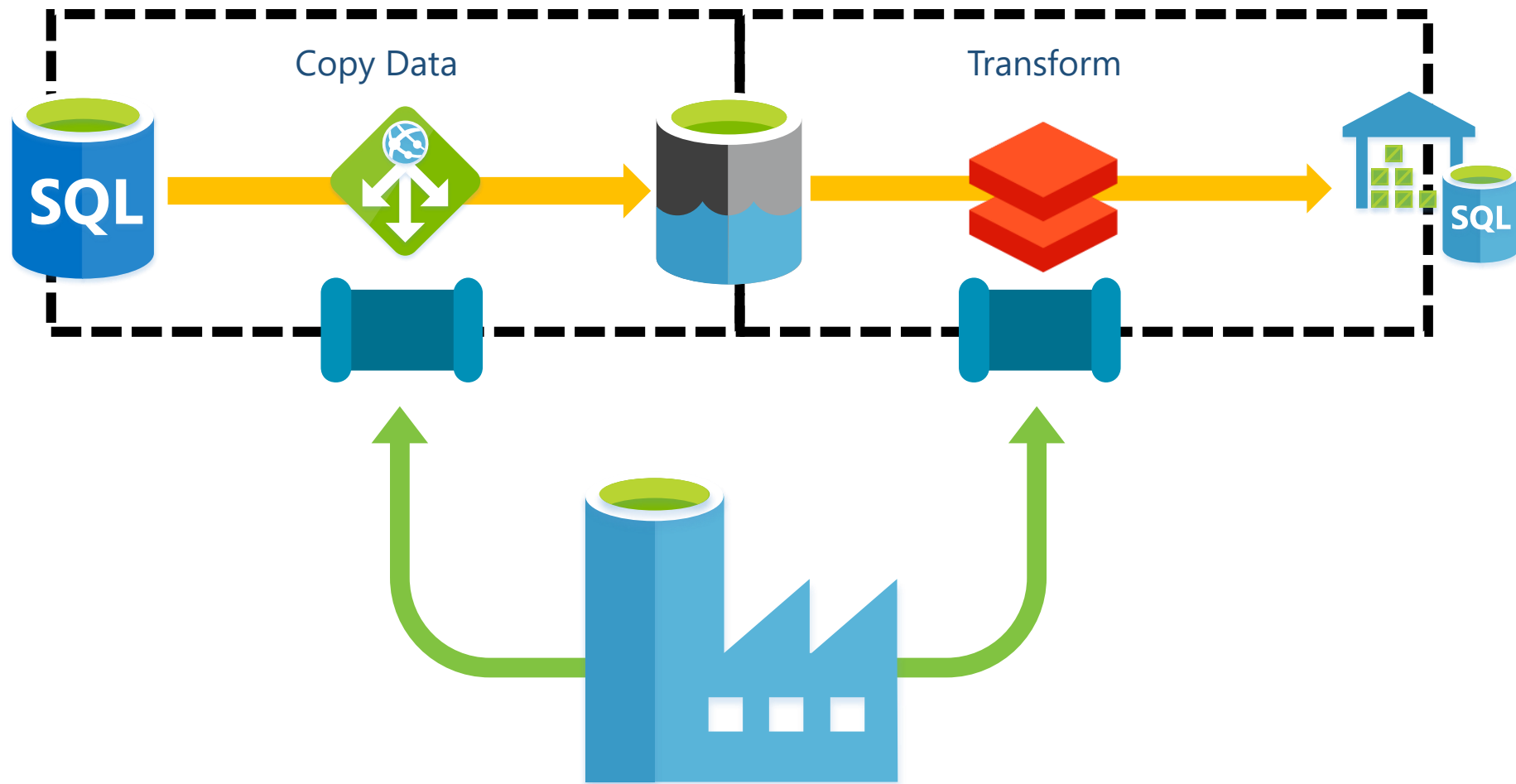
**3**

**Self Hosted**

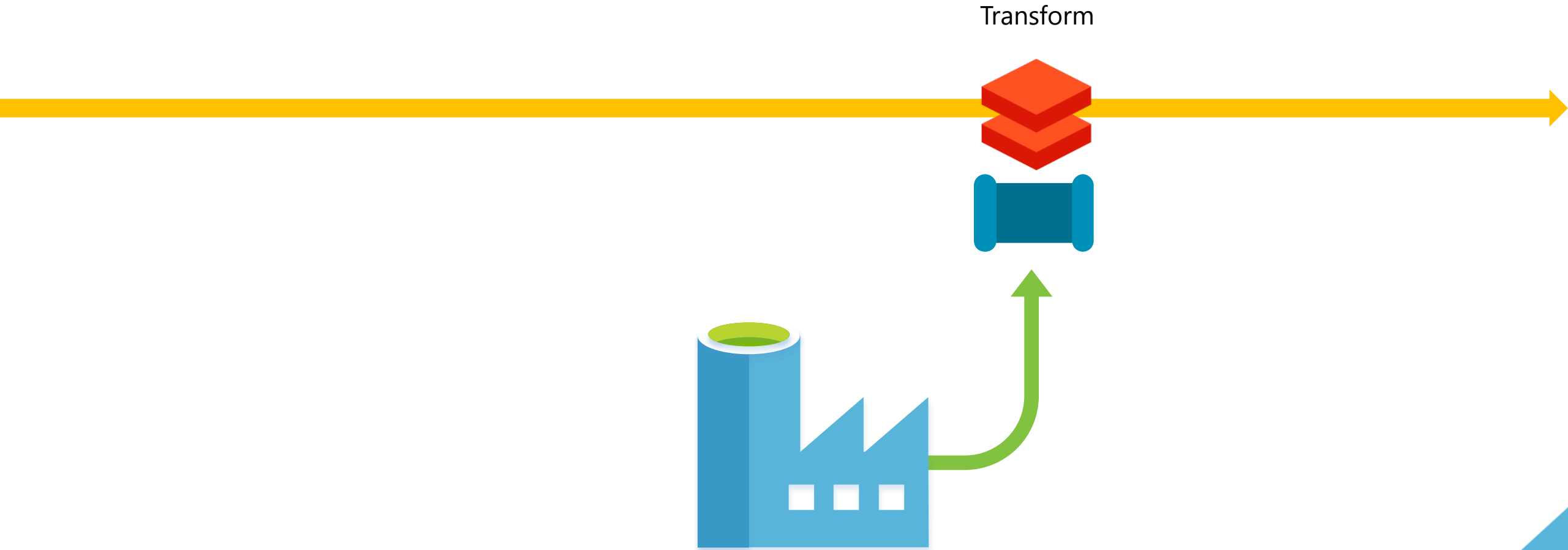
Integration Runtime



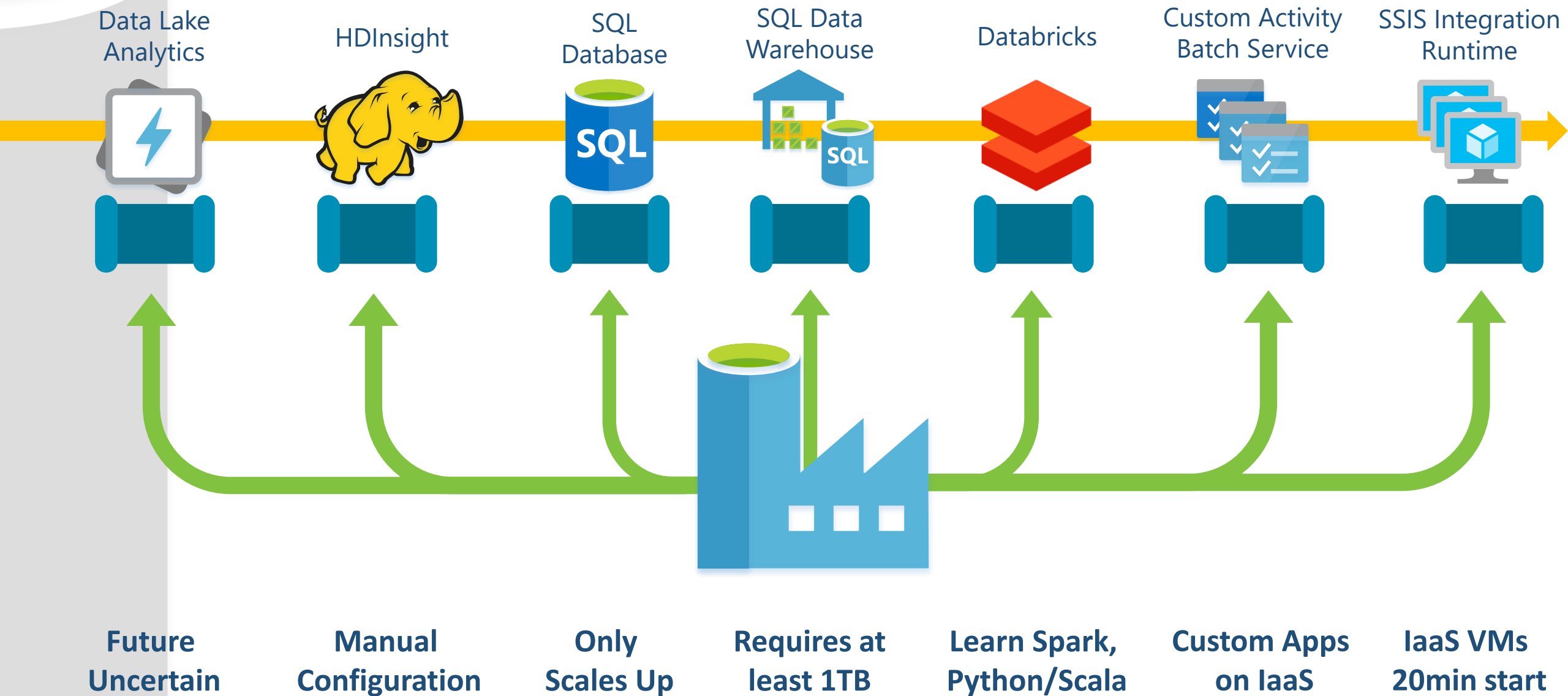
# Data Factory Control Flow Components



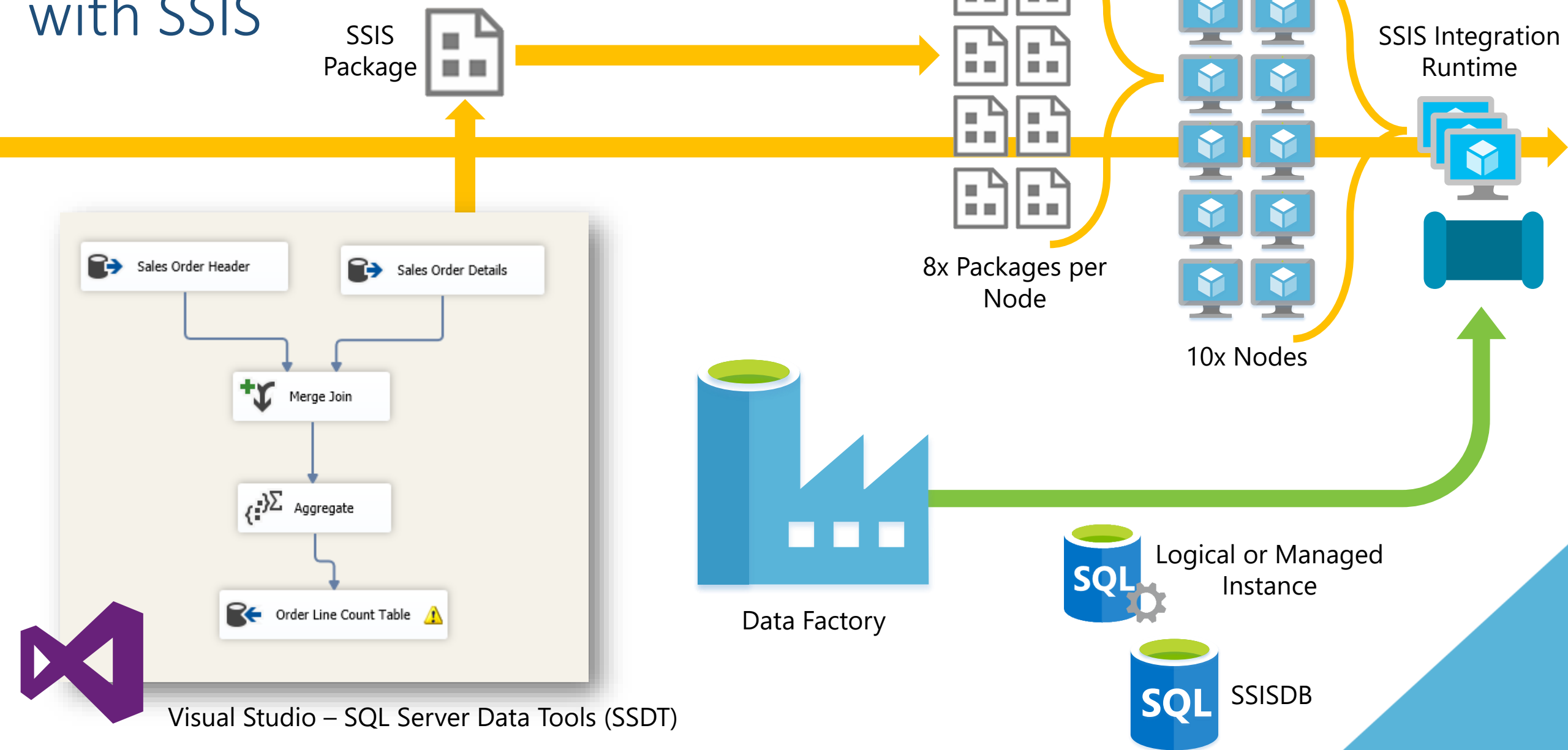
# Data Transformation in zure



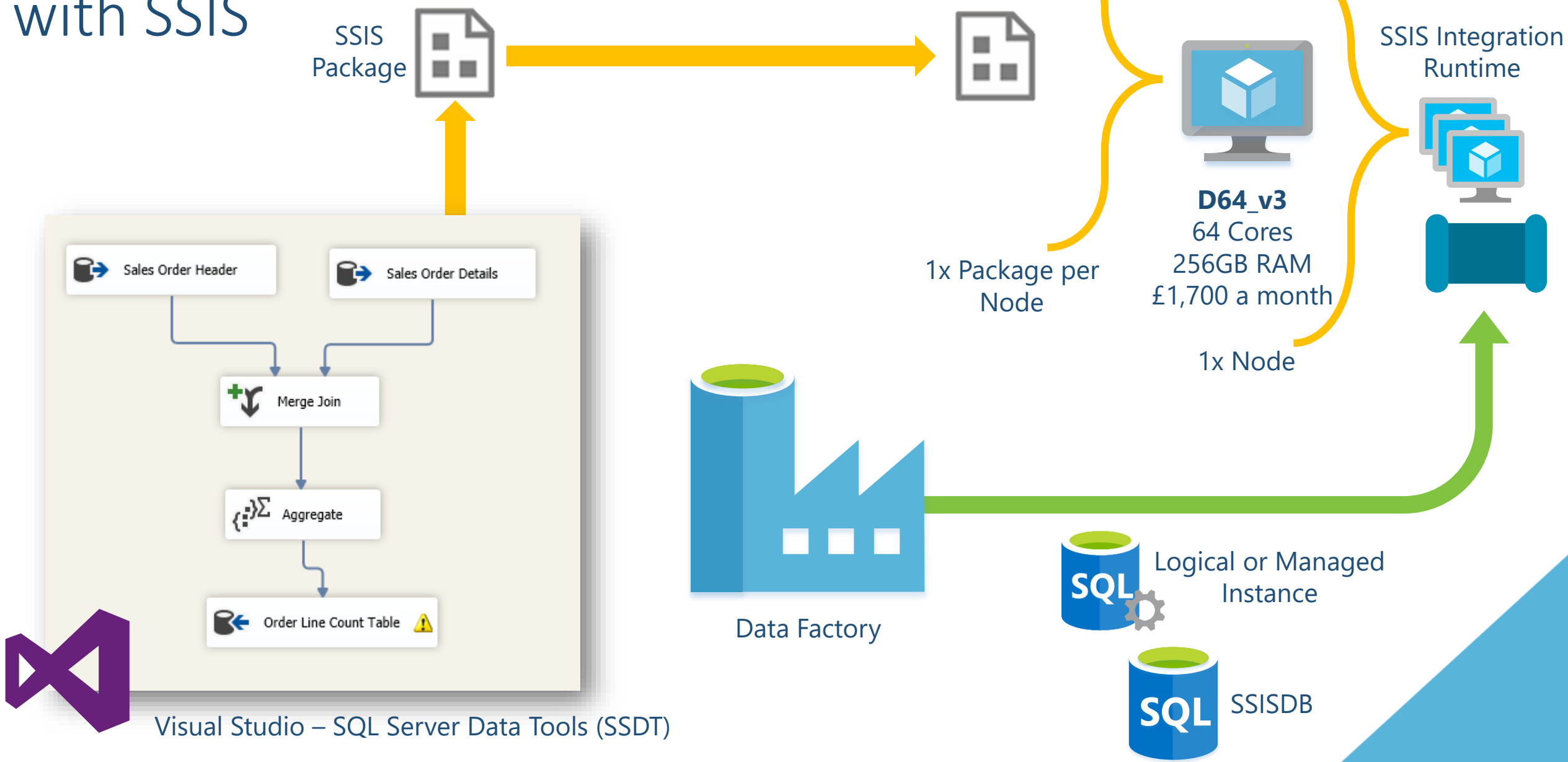
# Data Transformation in zure



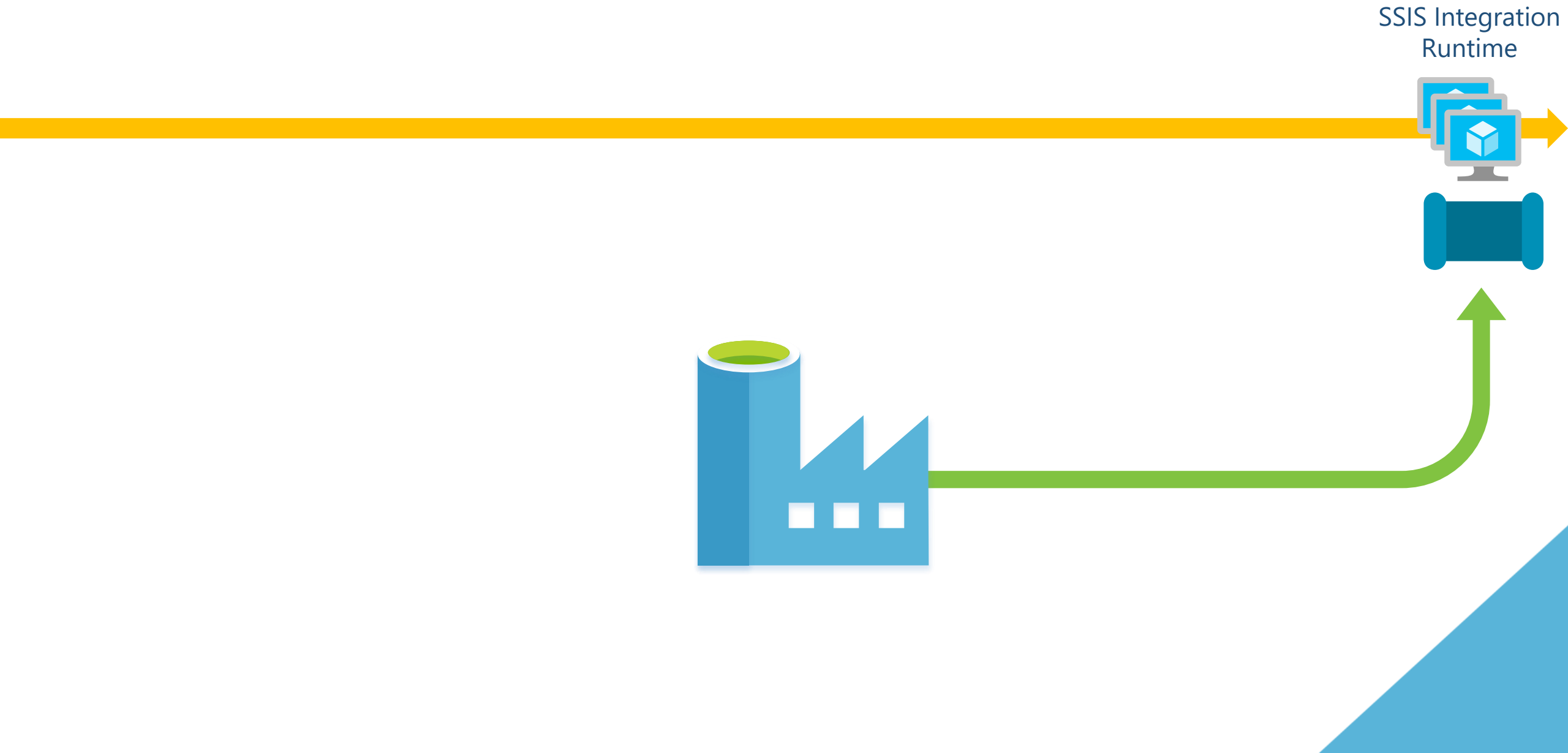
# Data Transformation in Azure with SSIS



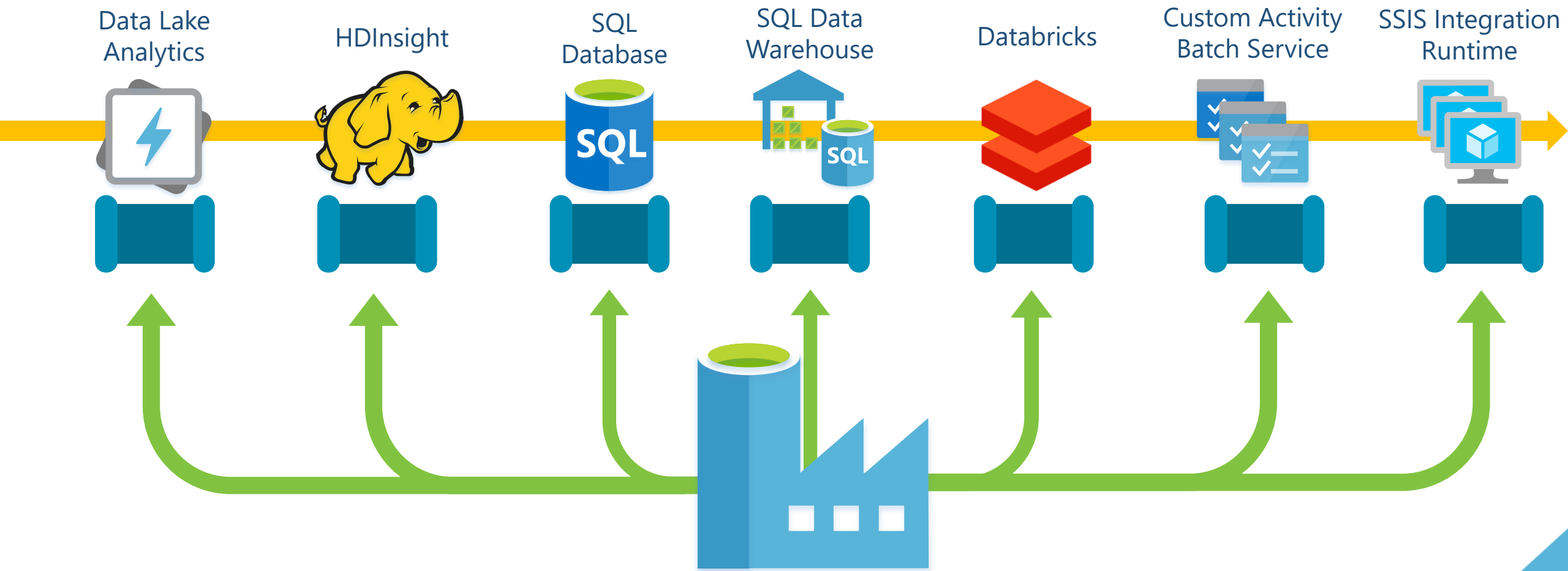
# Data Transformation in zure with SSIS



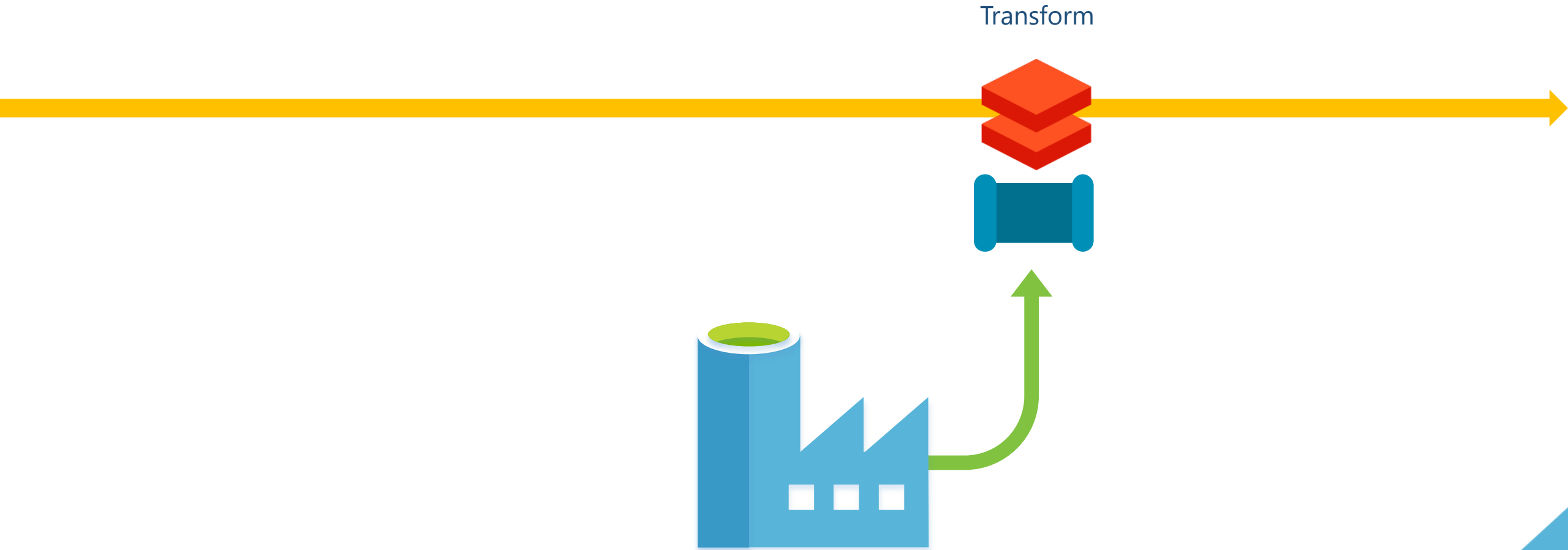
# Data Transformation in zure



# Data Transformation in zure

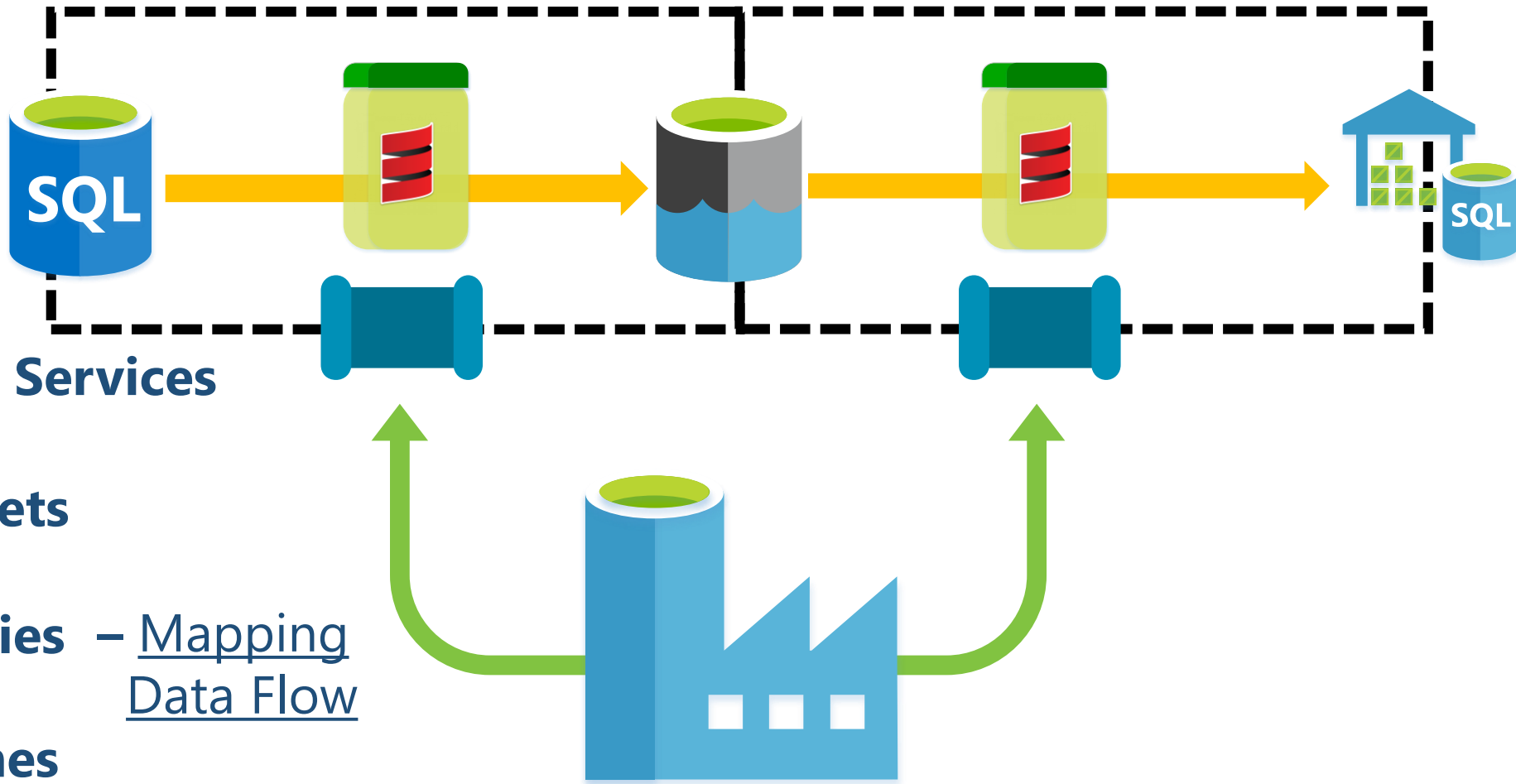


# Data Transformation in zure





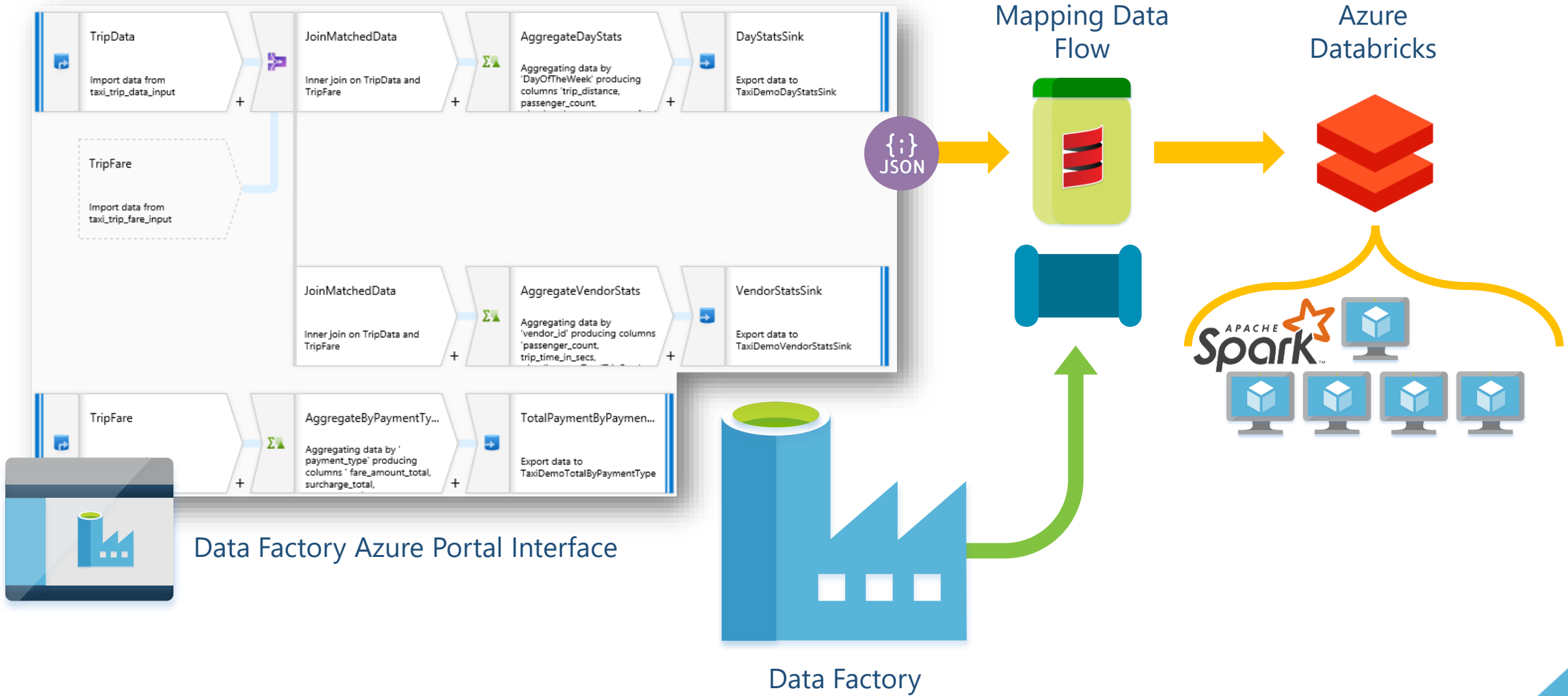
# Data Factory Control Flow Components



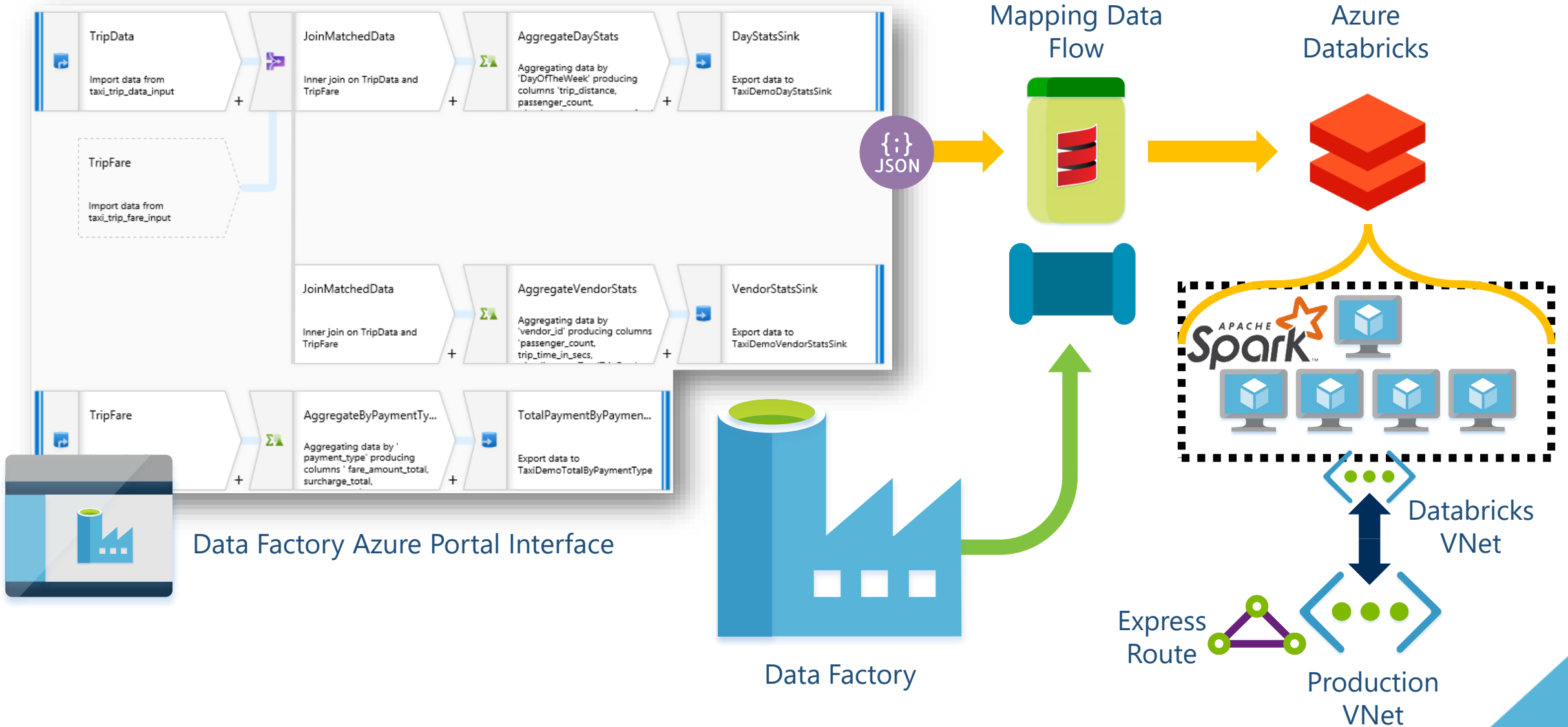
# Mapping Data Flows

A large, solid blue triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

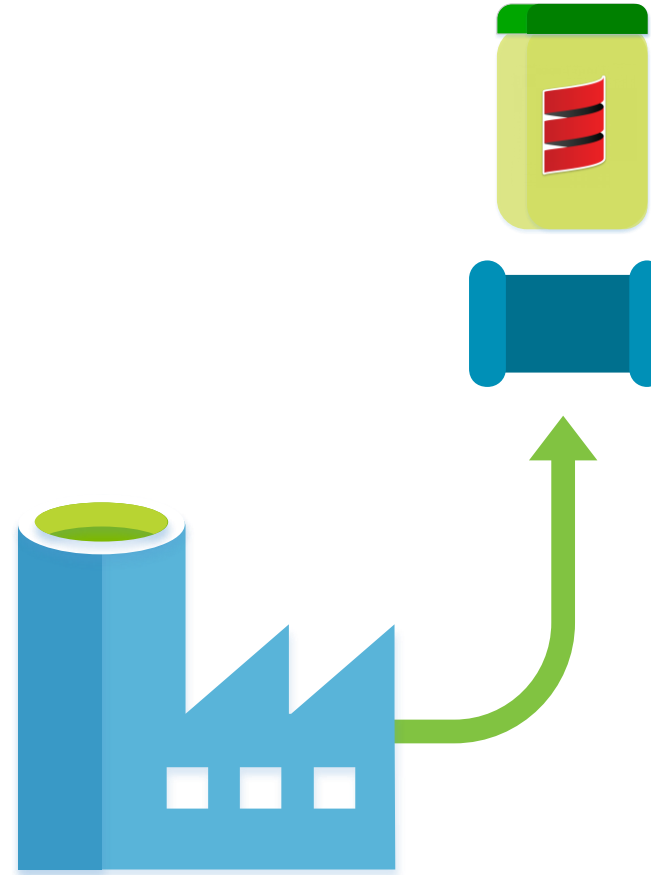
# What is a Mapping Data Flow?



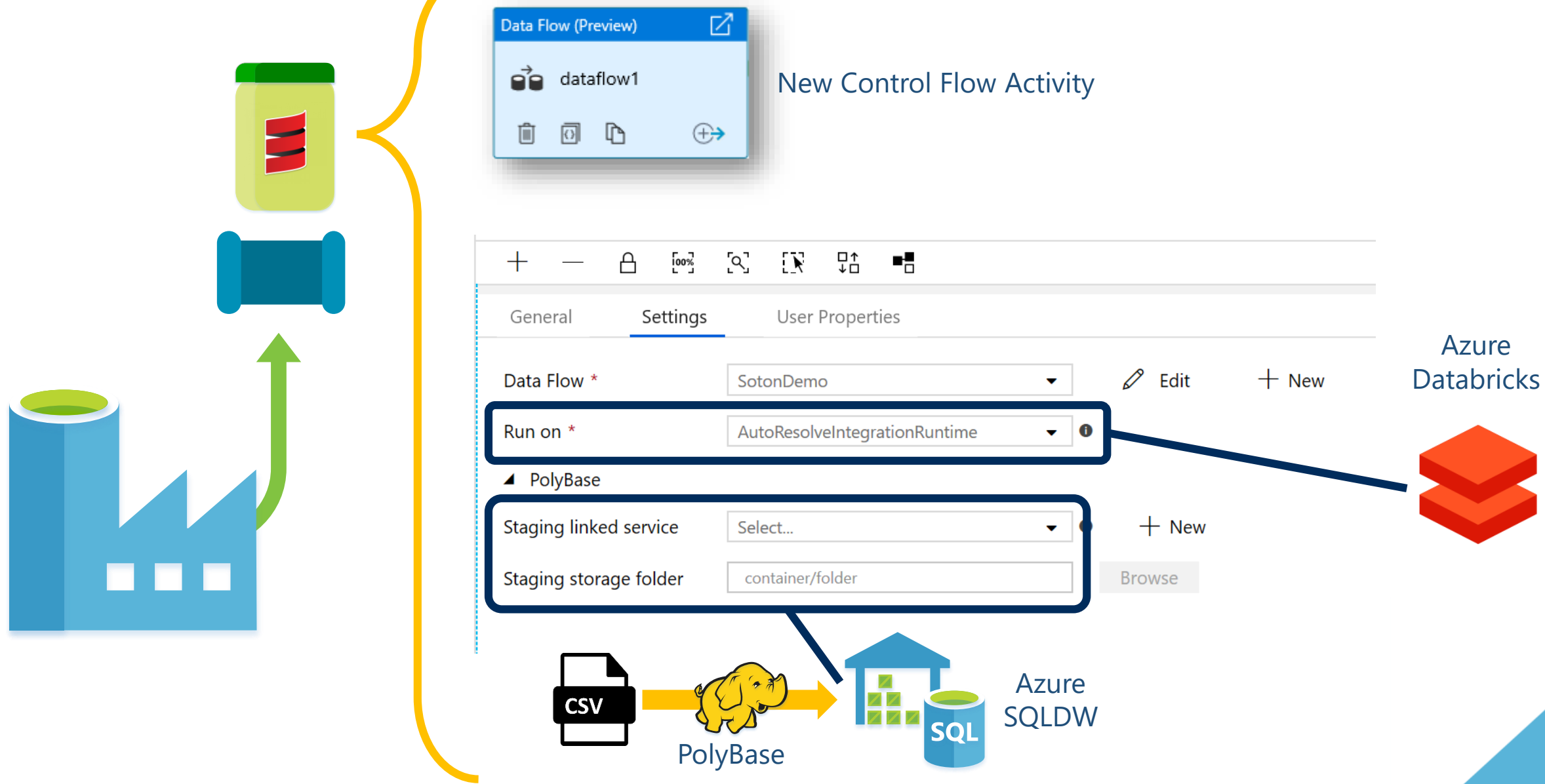
# What is a Mapping Data Flow?



# Mapping Data Flows



# Mapping Data Flows – Settings & Concepts



# Integration Runtimes



1

**Azure**  
Integration Runtime

Movement Hours



Activity  
Orchestration







2

**SSIS**  
Integration Runtime

SSIS Package  
Execution







3

**Self Hosted**  
Integration Runtime

Gateway Access



Activity  
Orchestration





# Integration Runtimes – Mapping Data Flow Cluster



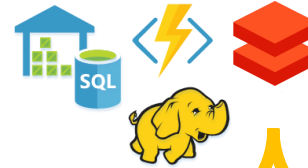
1

## Azure Integration Runtime

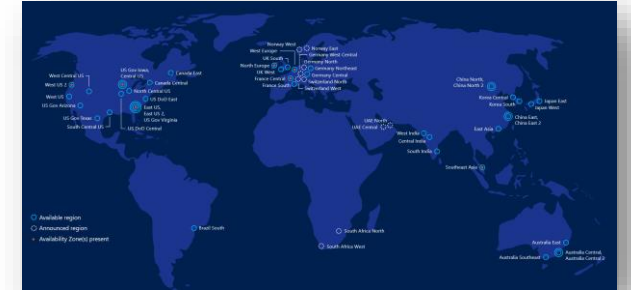
Movement Hours



Activity  
Orchestration



Flexible Region



### ▲ Data Flow run time

Compute Type \*

General Purpose

Core count \*

4 (4 Driver Cores)

Time to live (in minutes)

Time to live feature is coming soon

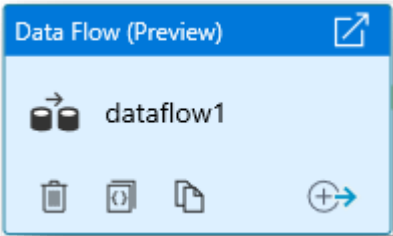
- General Purpose
- Memory Optimised
- Compute Optimised

=





# Mapping Data Flows – Settings & Concepts



New Control Flow Activity

General

Settings

User Properties

Data Flow \*

SotonDemo

Edit

New

Run on \*

AutoResolveIntegrationRuntime

PolyBase

Staging linked service

Select...

New

Staging storage folder

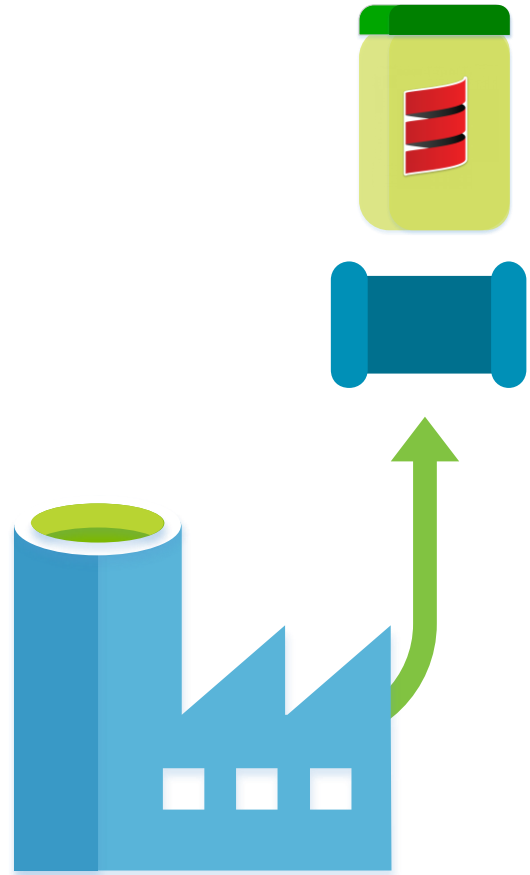
container/folder

Browse

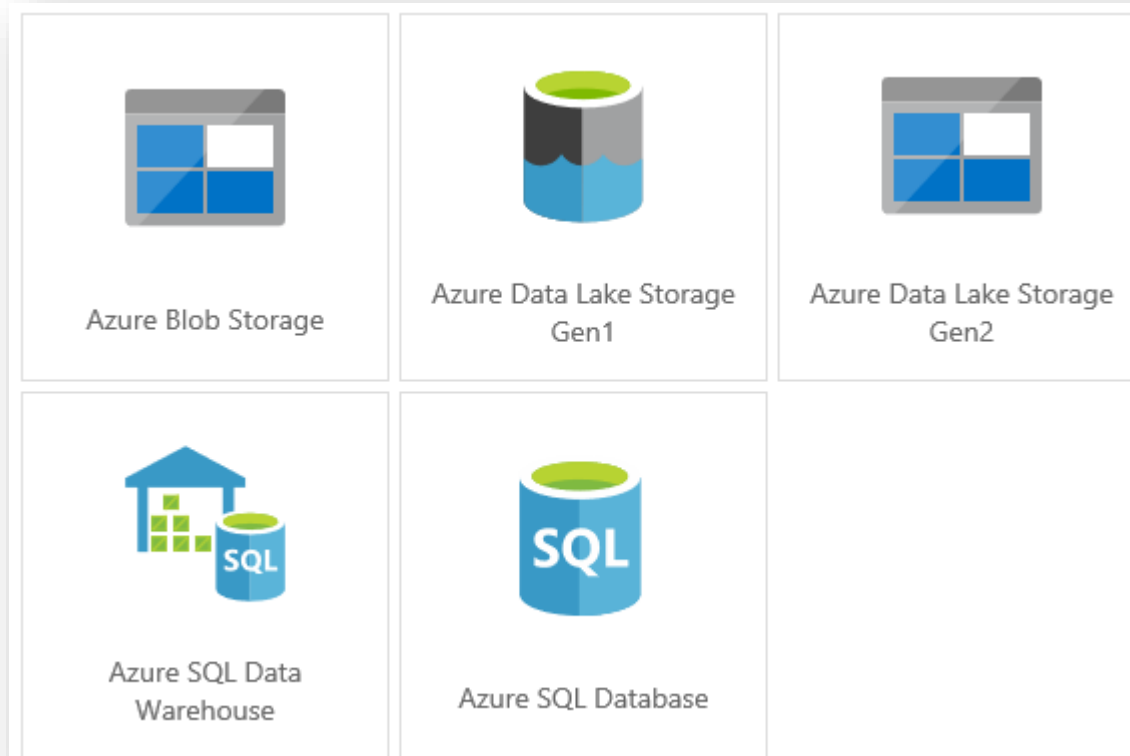
Azure Databricks



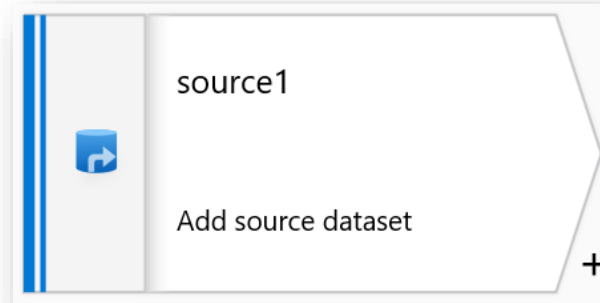
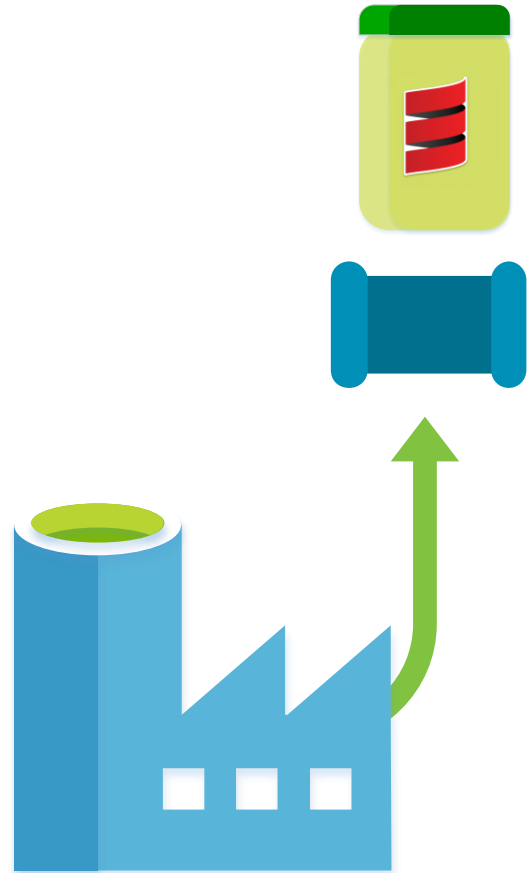
# Mapping Data Flows – Settings & Concepts



Currently Available:



# Mapping Data Flows – Settings & Concepts



Source Settings

Output stream name \*

Source dataset \*  GenericSQLTable  Edit  New

Options

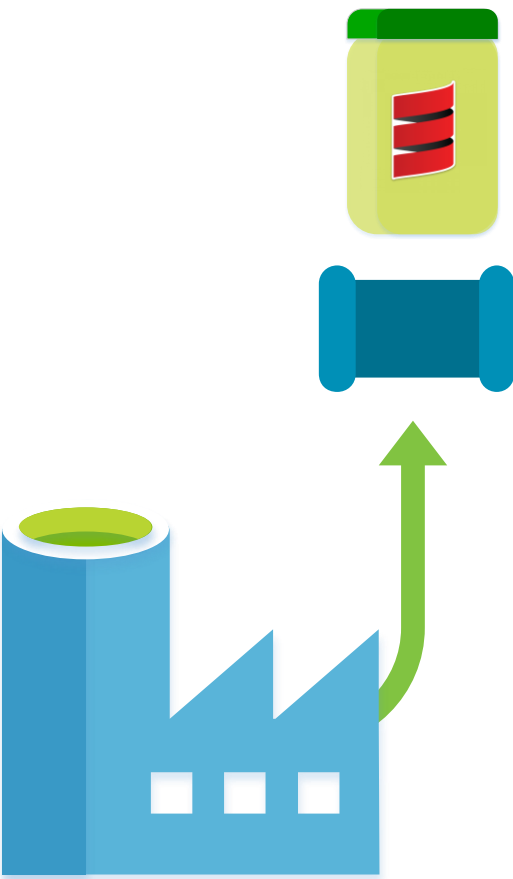
☒ Allow schema drift


☒ Validate schema

Sampling \* ☒ Enable ☐ Disable

Rows limit

# Mapping Data Flows – Settings & Concepts



sink1

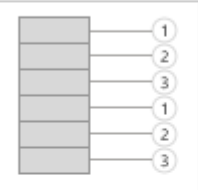
Add sink dataset

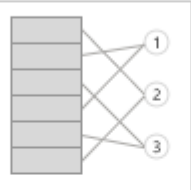
SinkSettingsMappingOptimizeInspectData Preview

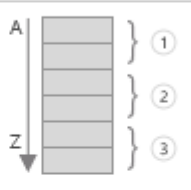
Partition option \*

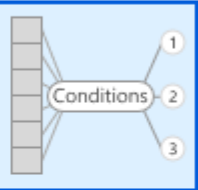
☐ Use current partitioning☐ Single partition☒ Set Partitioning

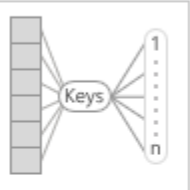
Partition type \*

Round Robin

Hash

Dynamic Range

Fixed Range

Key

Number of partitions \*

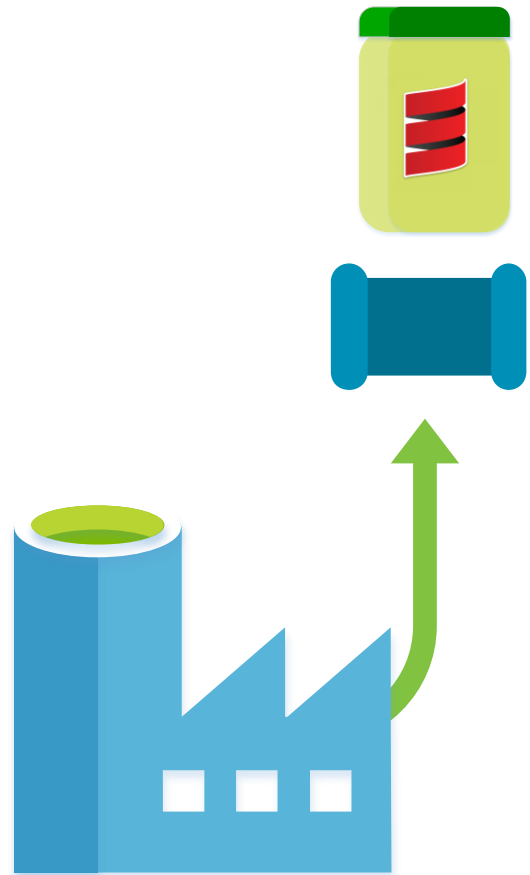
2






Condition to partition \*

Condition







Enter expression...ANY+🗑️

# Mapping Data Flows – Transformations








Multiple inputs/outputs	
	New Branch
	Join
	Conditional Split
	Union
	Lookup

Multicast  
Merge Join

Schema modifier	
	Derived Column
	Aggregate
	Surrogate Key
	Pivot
	Unpivot
	Window

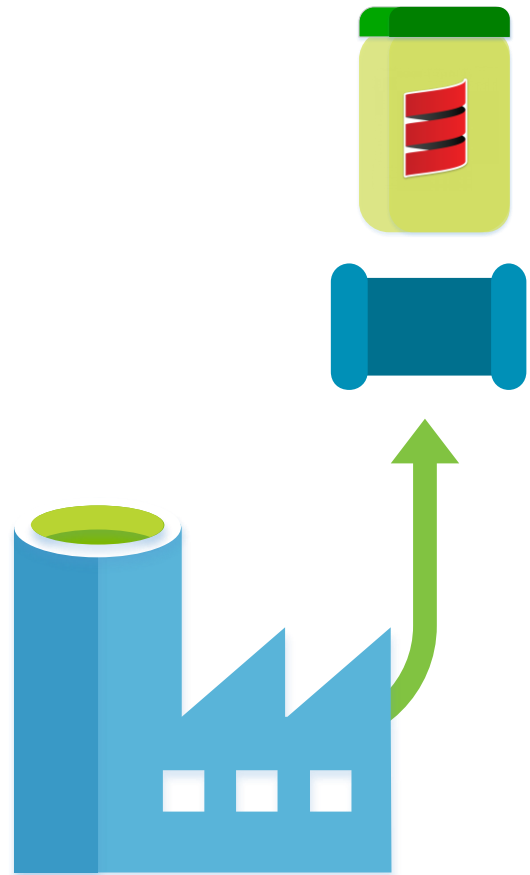
<https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-surrogate-key>

Row modifier	
	Exists
	Select
	Filter
	Sort

Custom	
	Extend

Script Component

# Mapping Data Flows – Expression Builder



### Visual Expression Builder

Currently working on year

«

AllStringMathDateLogicalInput

abc md5(ANY expression)

123 nextSequence()

abc regexExtract(abc string, abc regex to find, ANY match group 1-based index)

✕ regexMatch(abc string, abc regex to match)

abc right(abc string to subset, ANY number of characters)

toInteger(trim(right(title, 6), '()'))

Extract a matching substring for a given regex pattern. The last parameter identifies the match group and is defaulted to 1 if omitted. Use '<regex>' (back quote) to match a string without escaping

Examples  
1. regexExtract('Cost is between 600 and 800 dollars', '(\\d+) and (\\d+)', 2) -> '800'  
2. regexExtract('Cost is between 600 and 800 dollars', '(\\d+) and (\\d+)', 2) -> '800'

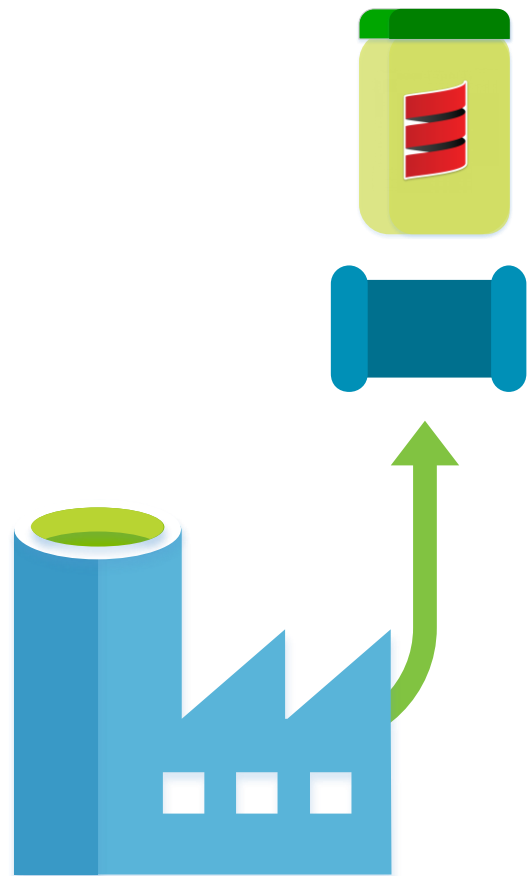
+ - \* / ||

Data preview

⚠ Please turn on the debug mode and wait until cluster is ready to preview data...

Output: year 123	title abc
-	-

# Mapping Data Flows – Debug Mode



ADWAnalysis X ADWAnalysisB... X

☒ Debug

Saved

✓ Validate

Source Settings

Cluster ● ForDataFlow

OrderHeader

Columns:  
22 total

OrderDetails

Import data from  
ADWSalesOrderDetail

Join1

Inner join on OrderHeader and  
OrderDetails

Aggregate1

Aggregating data by  
'SalesOrderNumber'  
producing columns  
'OrderLineCount'

sink1

Export data to  
ADWOrderLineCountTable

Add Source

Source Settings


Define schema

Optimize

Inspect

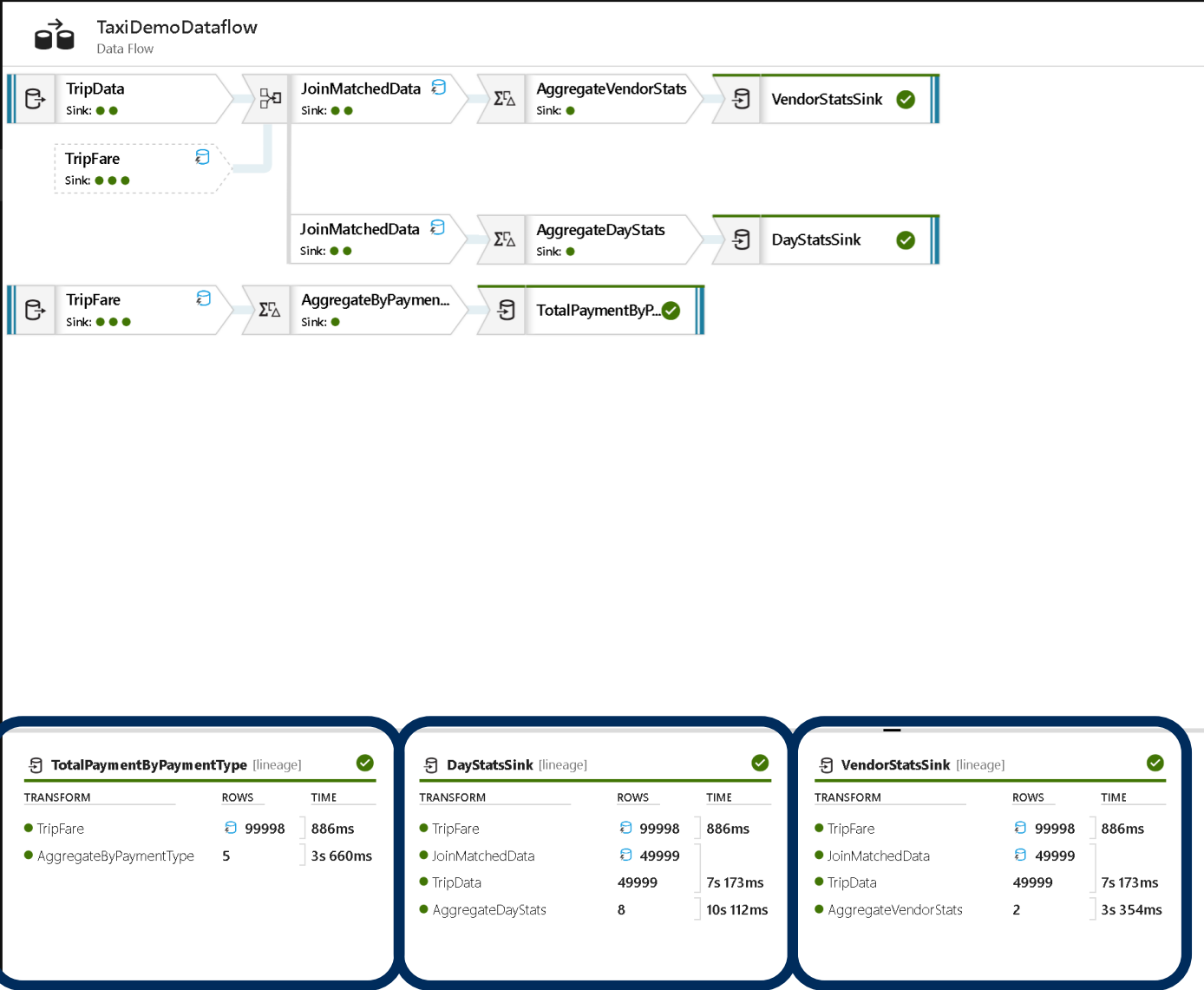
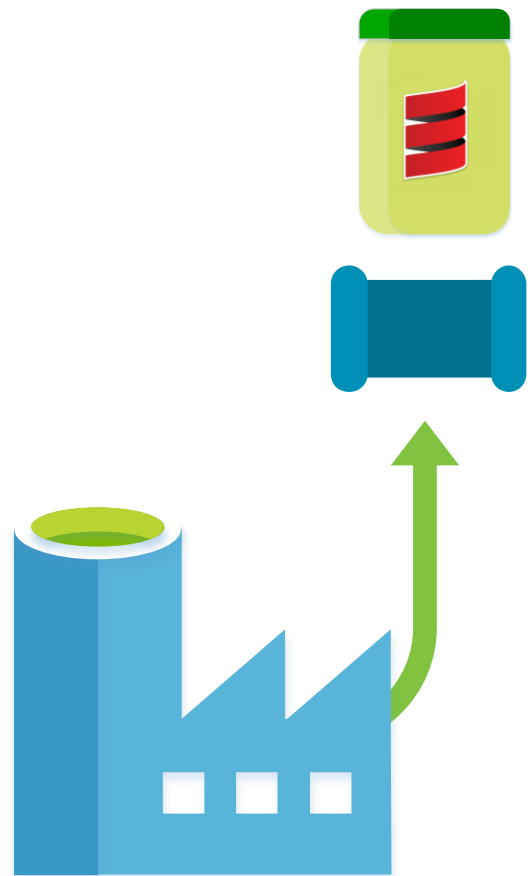
Data Preview

	Updated*	New*	Unchanged	Total
Number of rows	N/A	N/A	N/A	32
SalesOrderID 123	RevisionNumber abc	OrderDate 🕒	DueDate 🕒	ShipDate
71774	2	06/01/2008 00:06:00	06/13/2008 00:06:00	06/08/2008
71776	2	06/01/2008 00:06:00	06/13/2008 00:06:00	06/08/2008



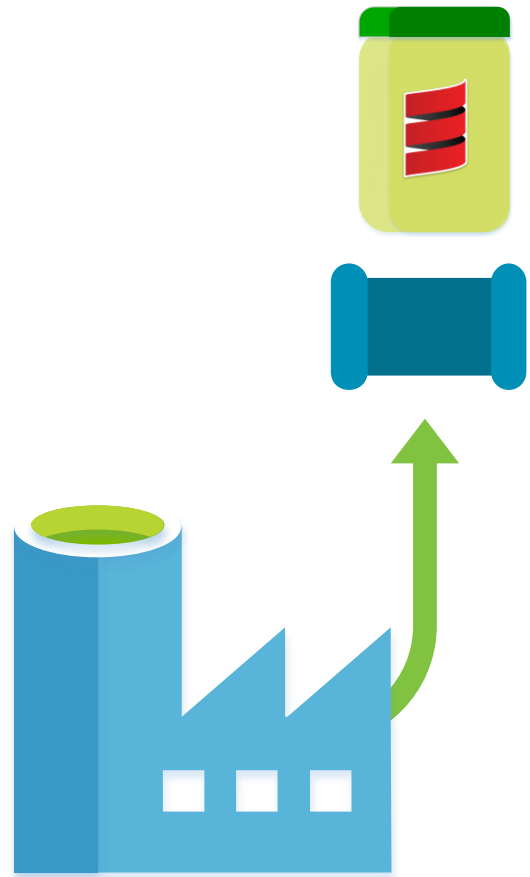
Only gives you a General Purpose cluster

# Mapping Data Flows – Monitoring





# Mapping Data Flows



1

## Activity

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-overview>

2

## Source & Sink

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-schema-drift>

3

## Transformations

<https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-aggregate>

4

## Expression Builder

<https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-expression-functions>

5

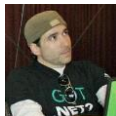
## Debug Mode

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-debug-mode>

6

## Monitoring

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-monitoring>







## Mark Kromer

<https://github.com/kromerm/adfdataflowdocs>

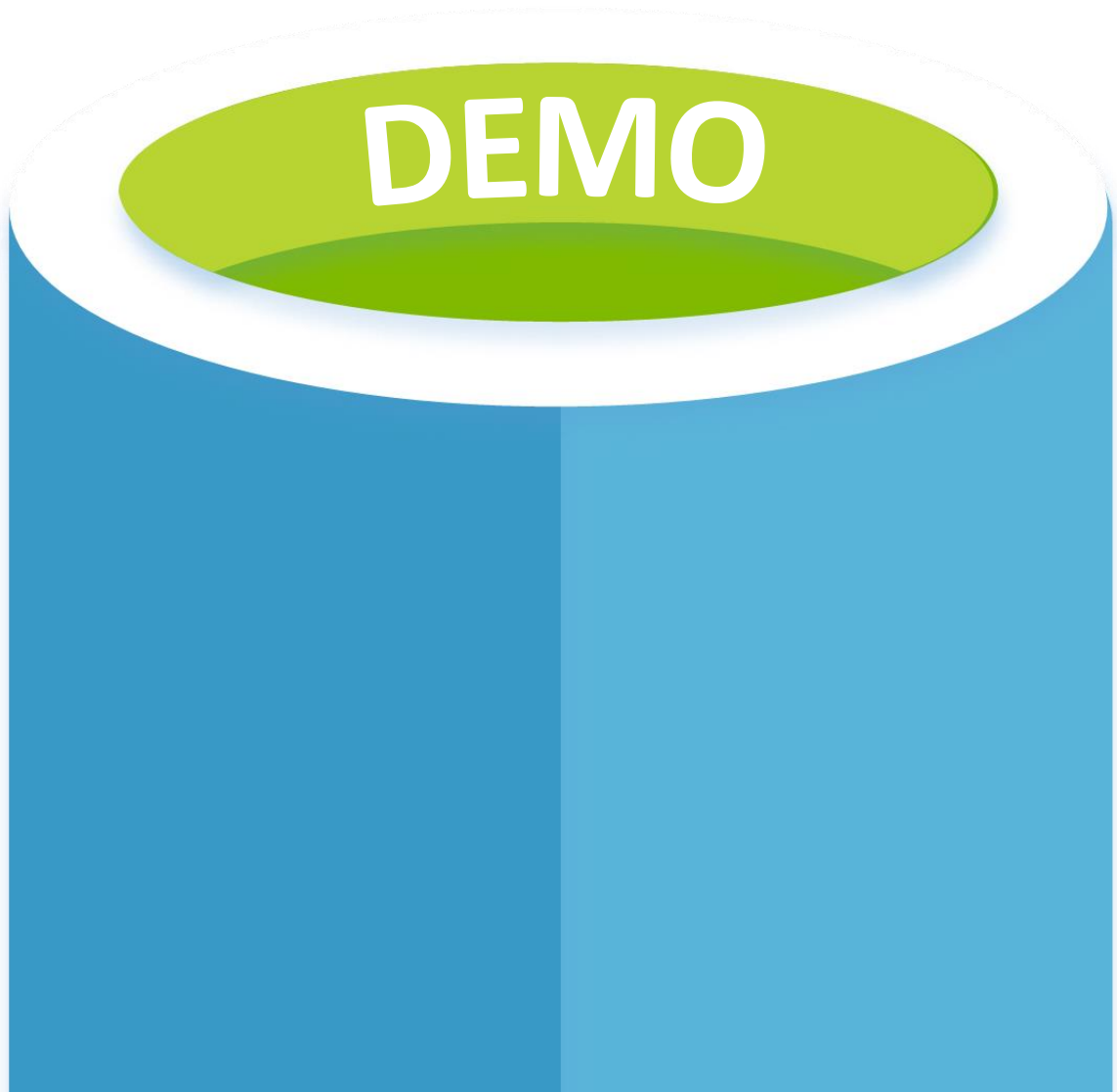
# Demo

**Start clusters!**

# Demo Summary

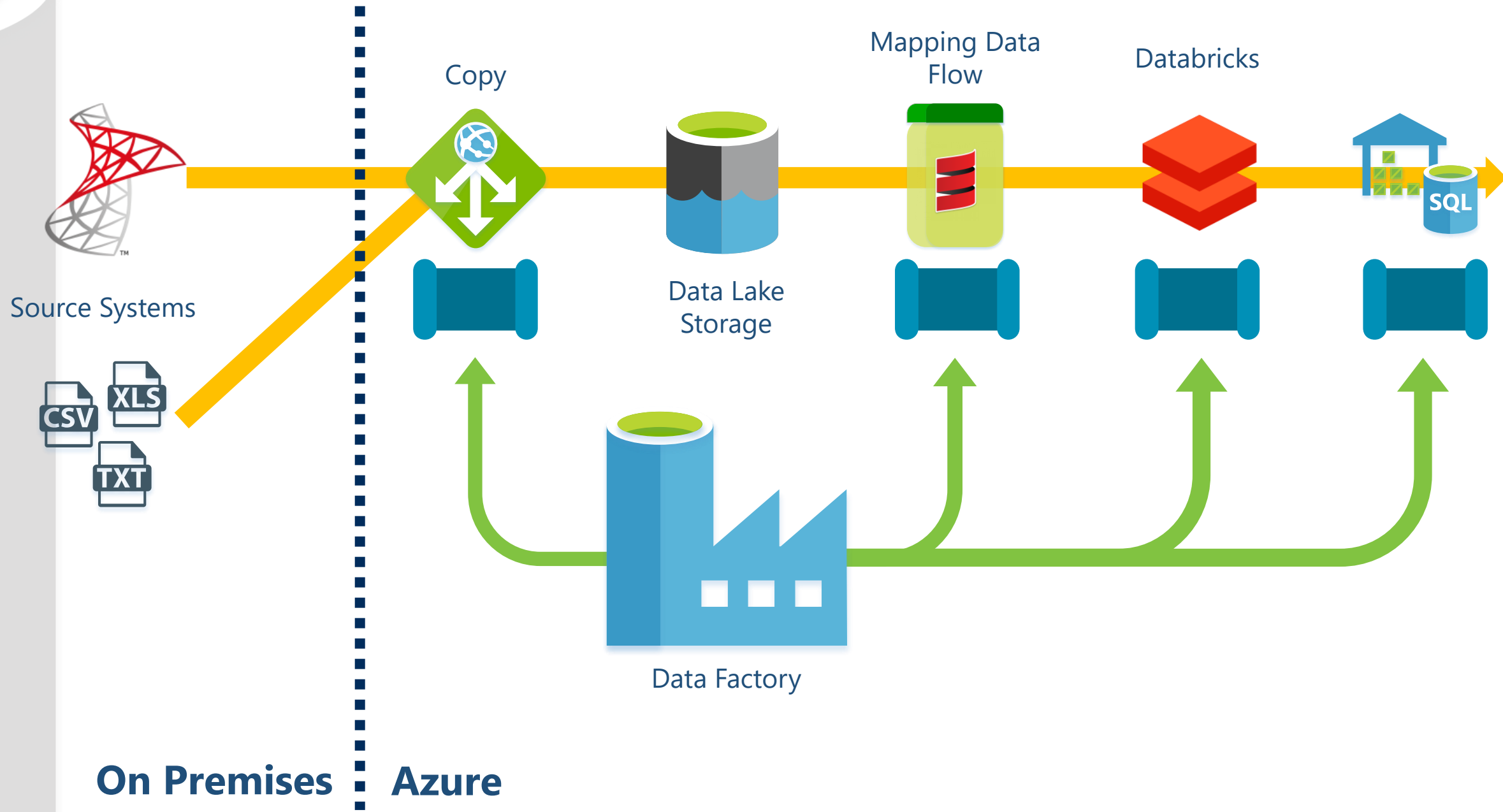
Transformation Method		Graphical UI	Scales Out	Scales Up	Cloud Native Tech
	T-SQL (SQLDB)	✗	✗	✓	✗
	SSIS	✓	✗	✓	✗
	Scala (Databricks)	✗	✓	✓	✓
	Mapping Data Flow	✓	✓	✓	✓

# Design Patterns

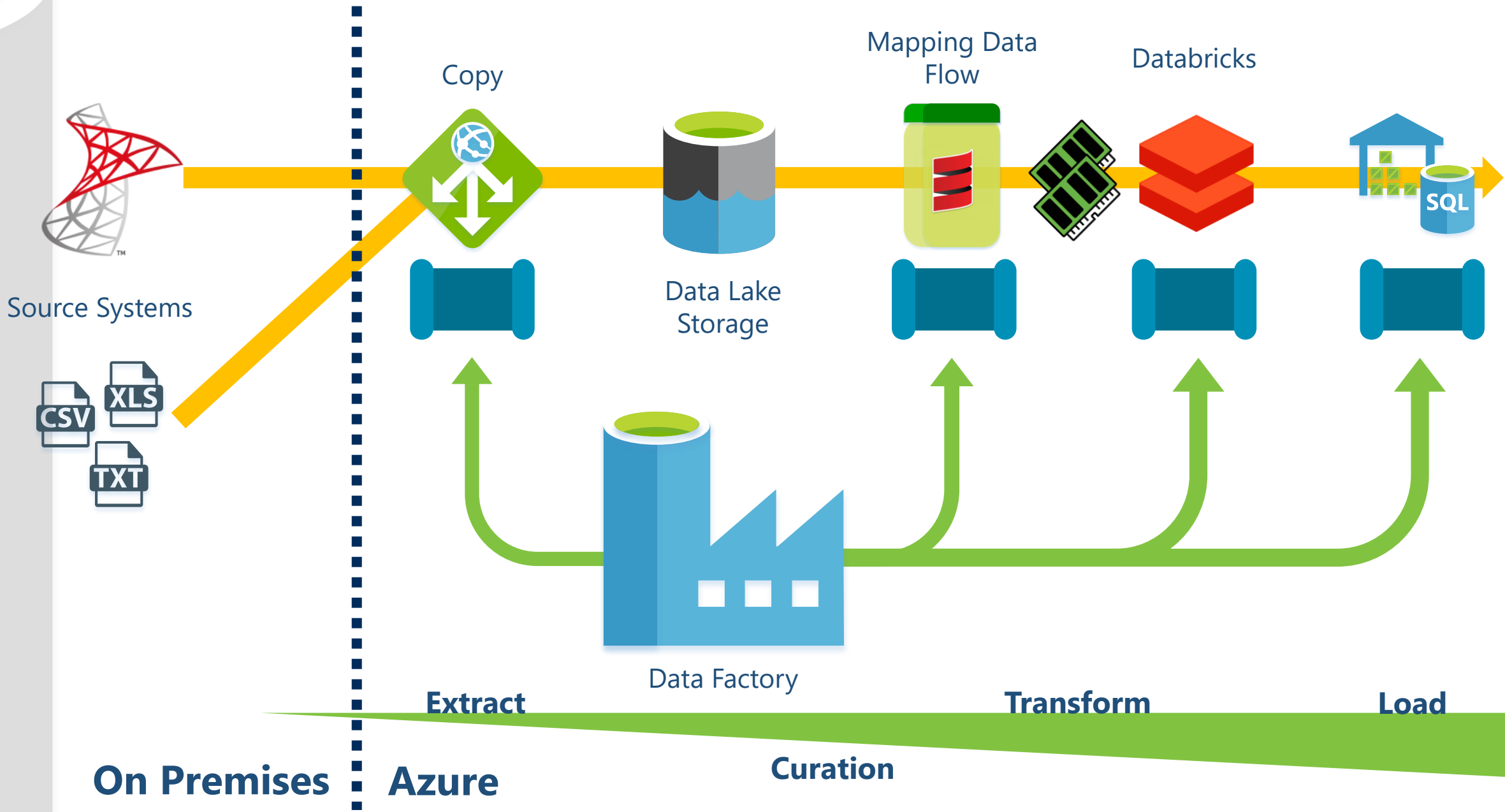


**Start Clusters!**

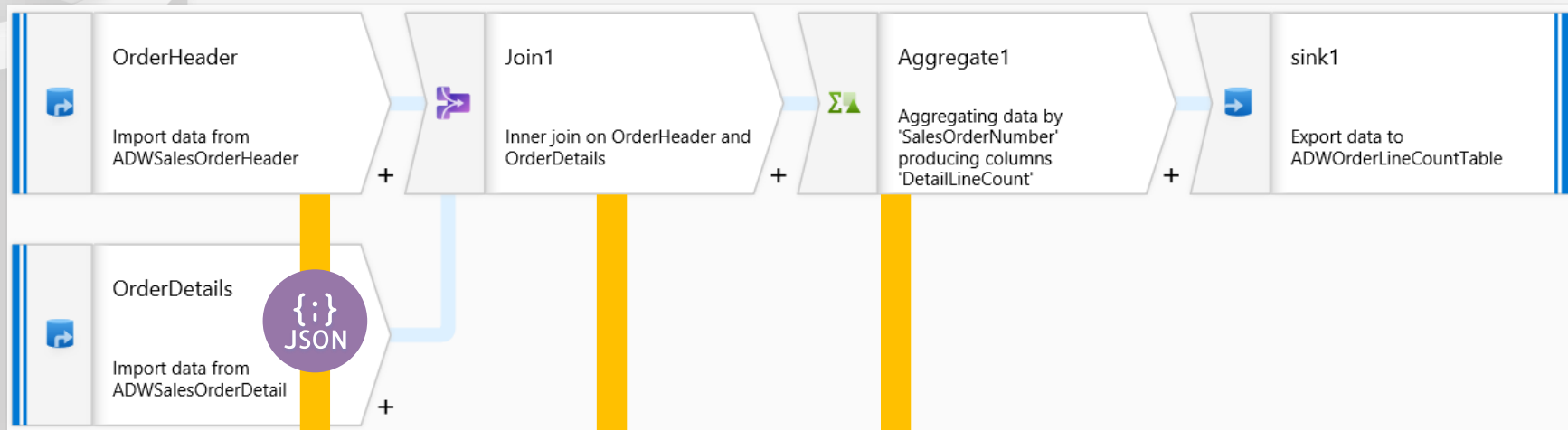
# Mapping Data Flow Future Design Patterns ???



# Mapping Data Flow Future Design Patterns ???

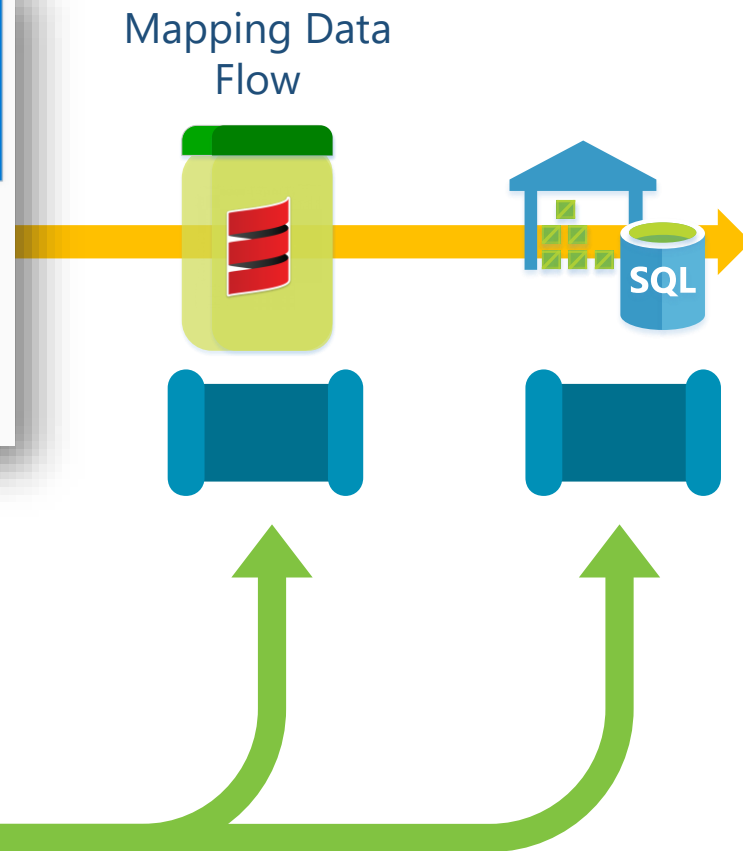


# Mapping Data Flow Future Design Patterns ???



```
"fileName": {
  "value": "@dataset().FileName",
  "type": "Expression"
},
"folderPath": {
  "value": "@dataset().SourceDIR",
  "type": "Expression"
}
```

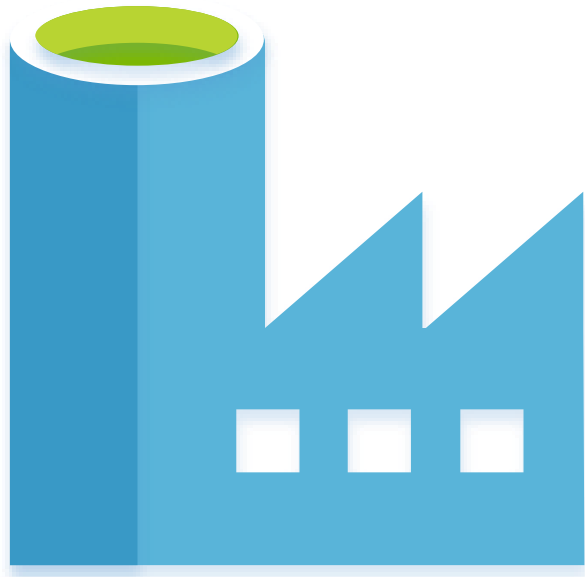
```
"transformations": [
{
  "name": "Join1",
  "script": "OrderHeader, OrderDetail join(OrderHeader@SalesOrderID == OrderDetail@SalesOrderID,\n\tjoinType:'inner',\n\tbroadcast: 'none') ~> Join1"
},
{
  "name": "Aggregate1",
  "script": "Join1 aggregate(groupBy(SalesOrderNumber),\n\tDetailLineCount = count(SalesOrderDetailID)) ~> Aggregate1"
}
]
```



# Wrangling Data Flows

A large, solid blue triangle is positioned in the bottom right corner of the slide, pointing towards the top right.





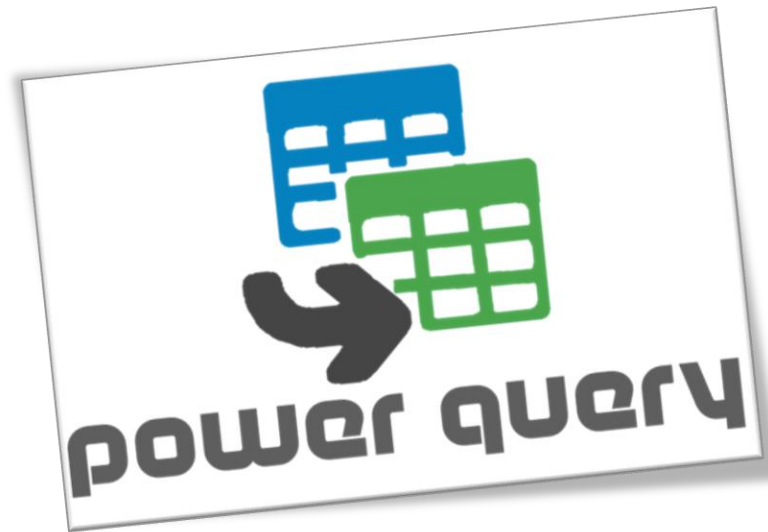
## Mapping Data Flow

Code free data transformation at scale

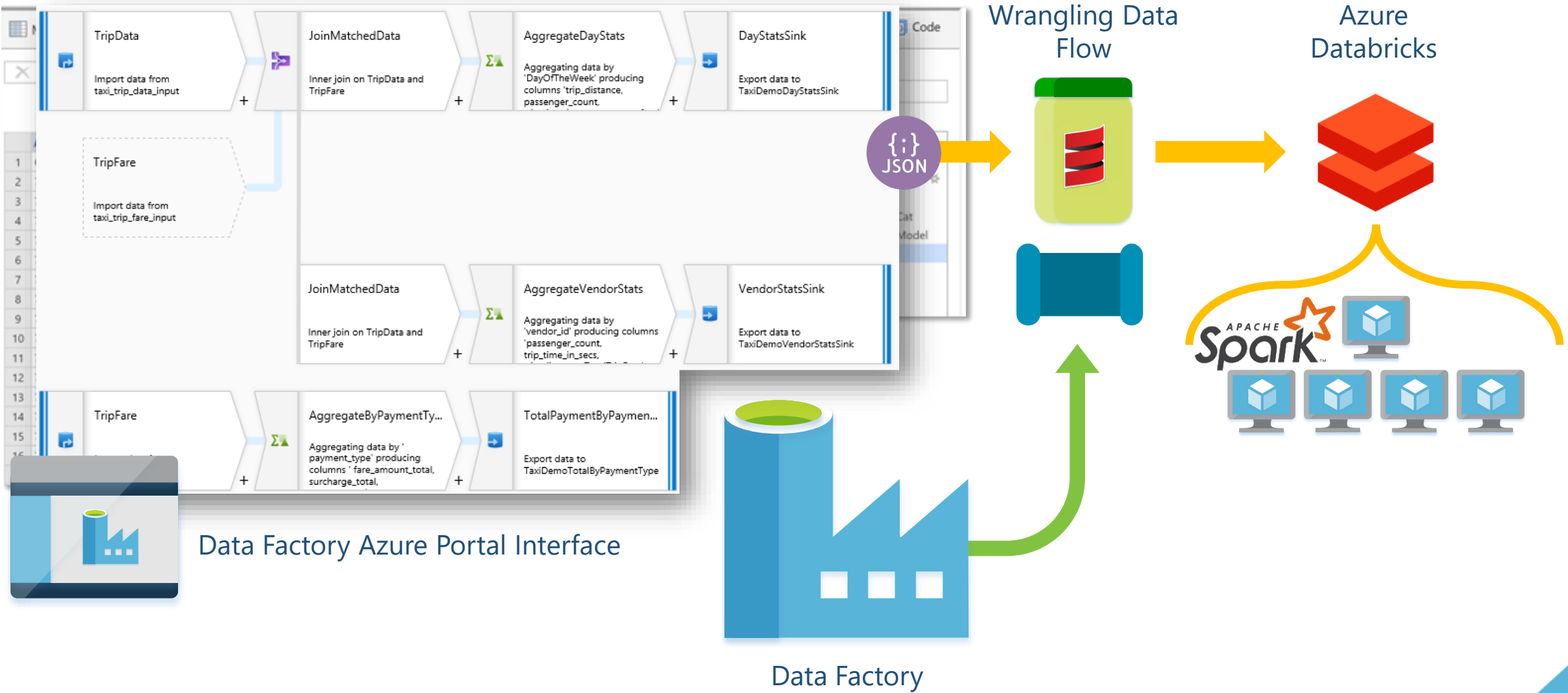


## Wrangling Data Flow (Preview)

Code free data preparation at scale



# What is a Wrangling Data Flow?



Are you...



- Productionising
- Engineering
- Warehousing

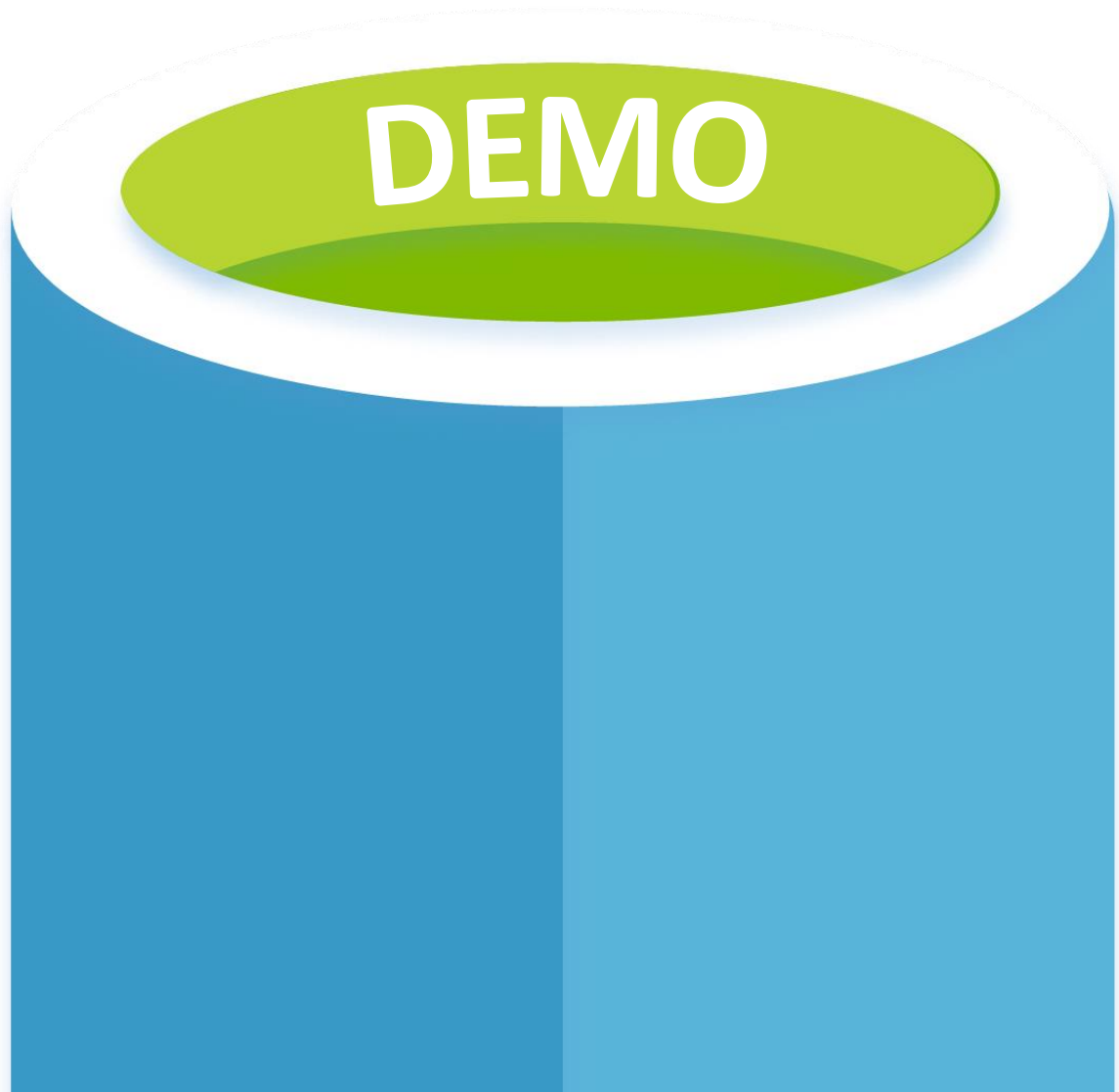
MAPPING



- Exploring
- Experimenting
- Analysing

WRANGLING

# Wrangling Data Flows

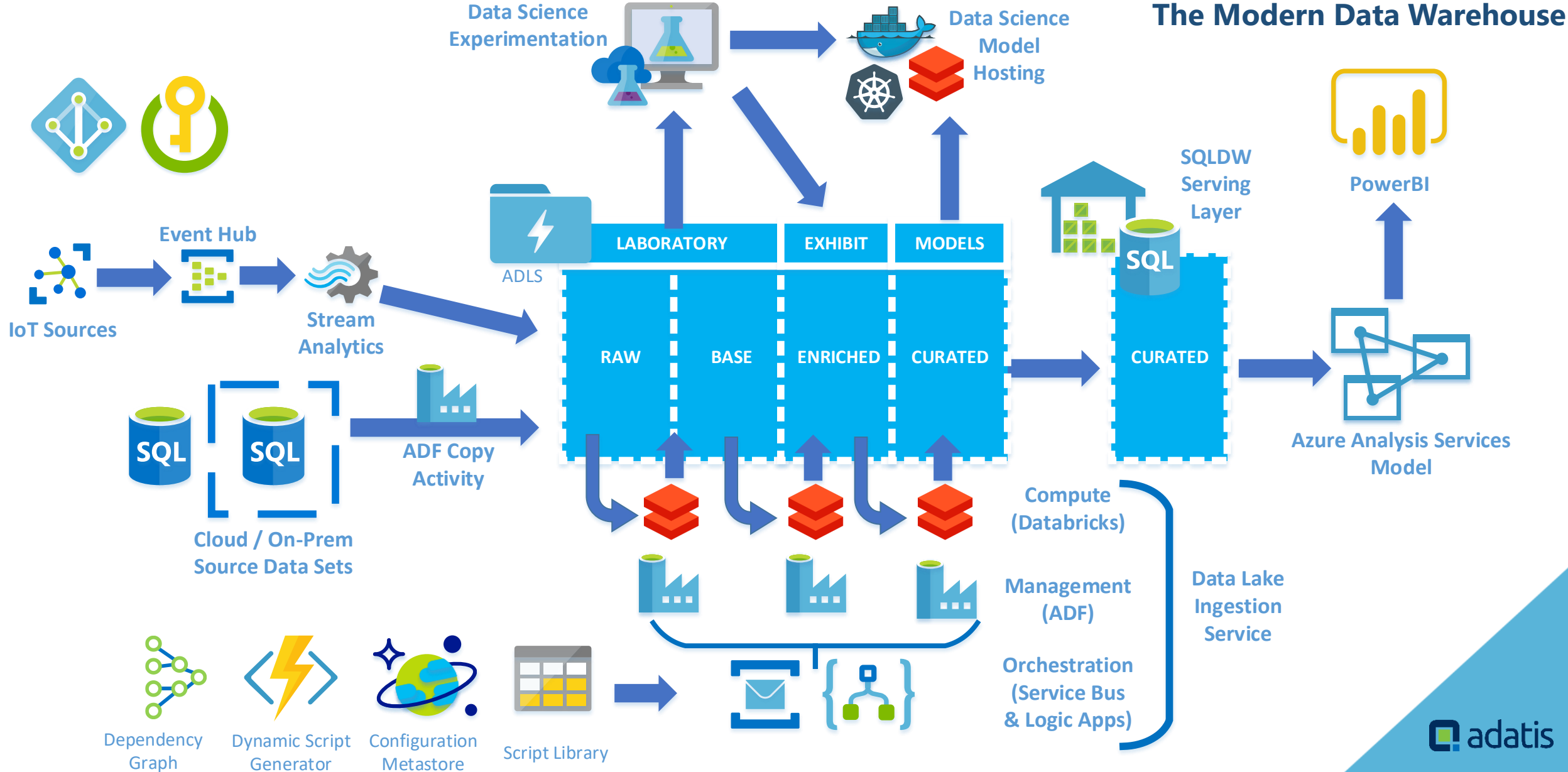


- Building a Wrangling Data Flow

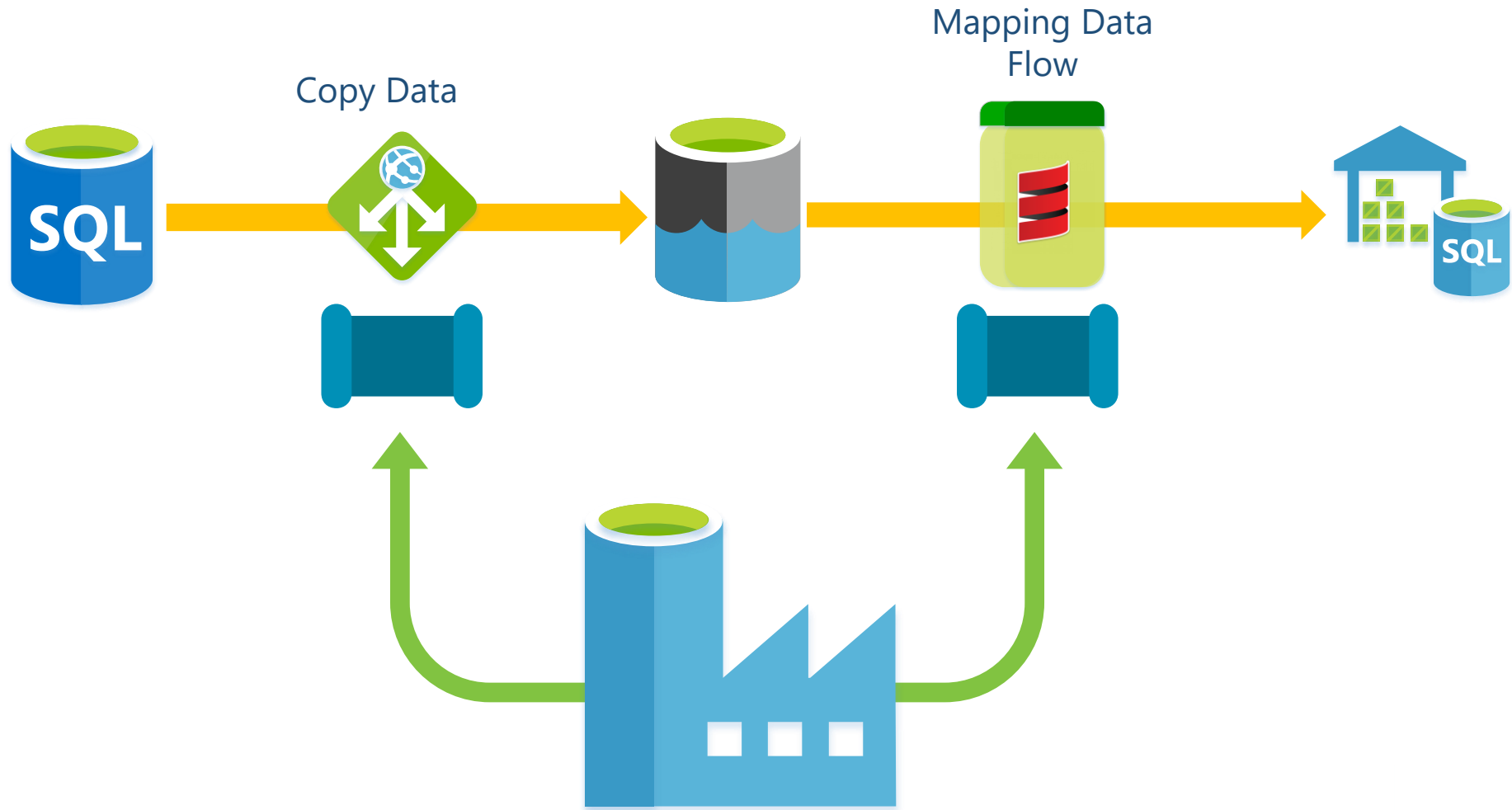
Conclusions

A large, solid blue triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

# Why use Azure Data Factory?



# What is Azure Data Factory?



Orchestrator of our solution Control Flow operations.

Orchestrator of our solution Data Flow transformations.

... using cloud native technology in  zure and now with a user interface for both.

# Module 4: Data Flows

- Mapping Data Flows
- Databricks Cluster Configuration
- Wrangling Data Flows
- What is Data Factory? UPDATE