



# Microsoft Ignite The Tour

Comes to Dublin!

Learn. Explore. Connect.

~~London, England~~



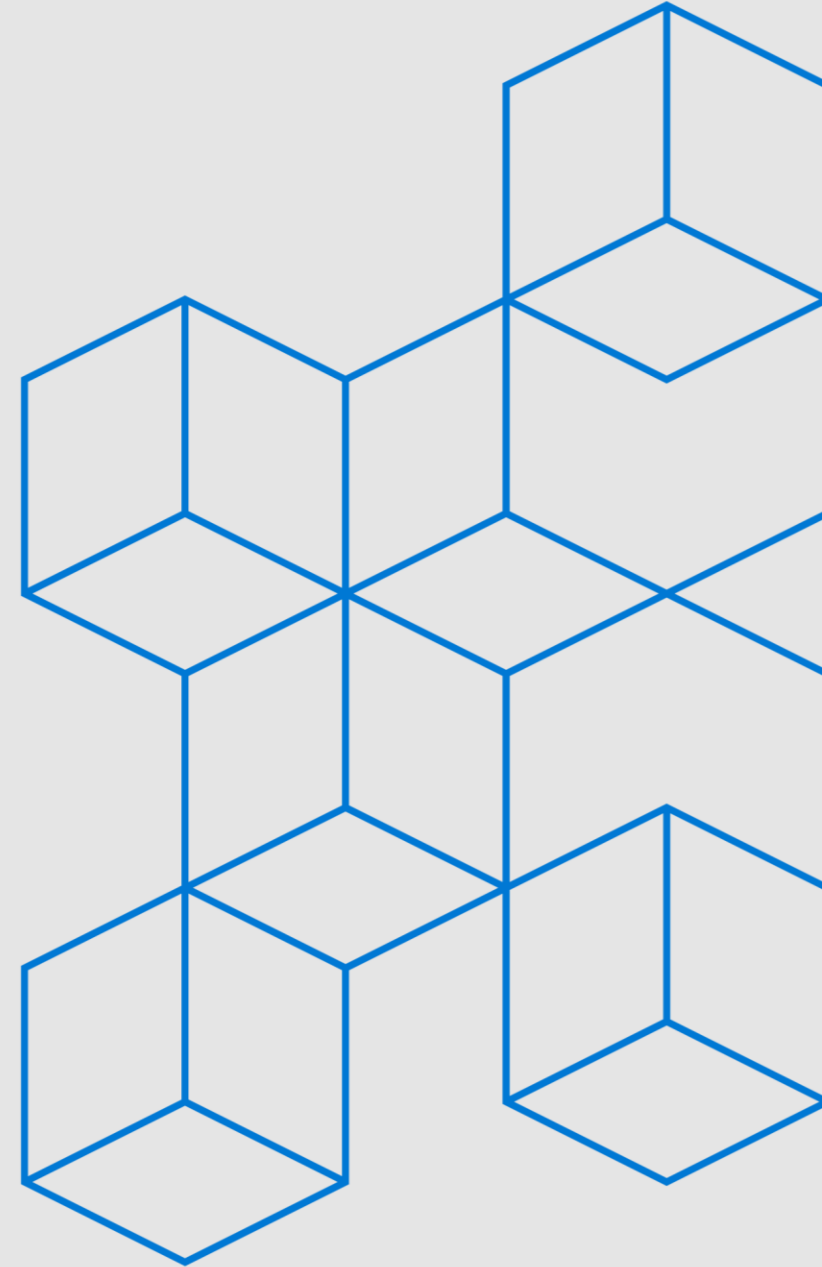


# ETL in Azure Made Easy with Data Factory Data Flows



**Paul Andrew**

Data Platform MVP & Solution Architect



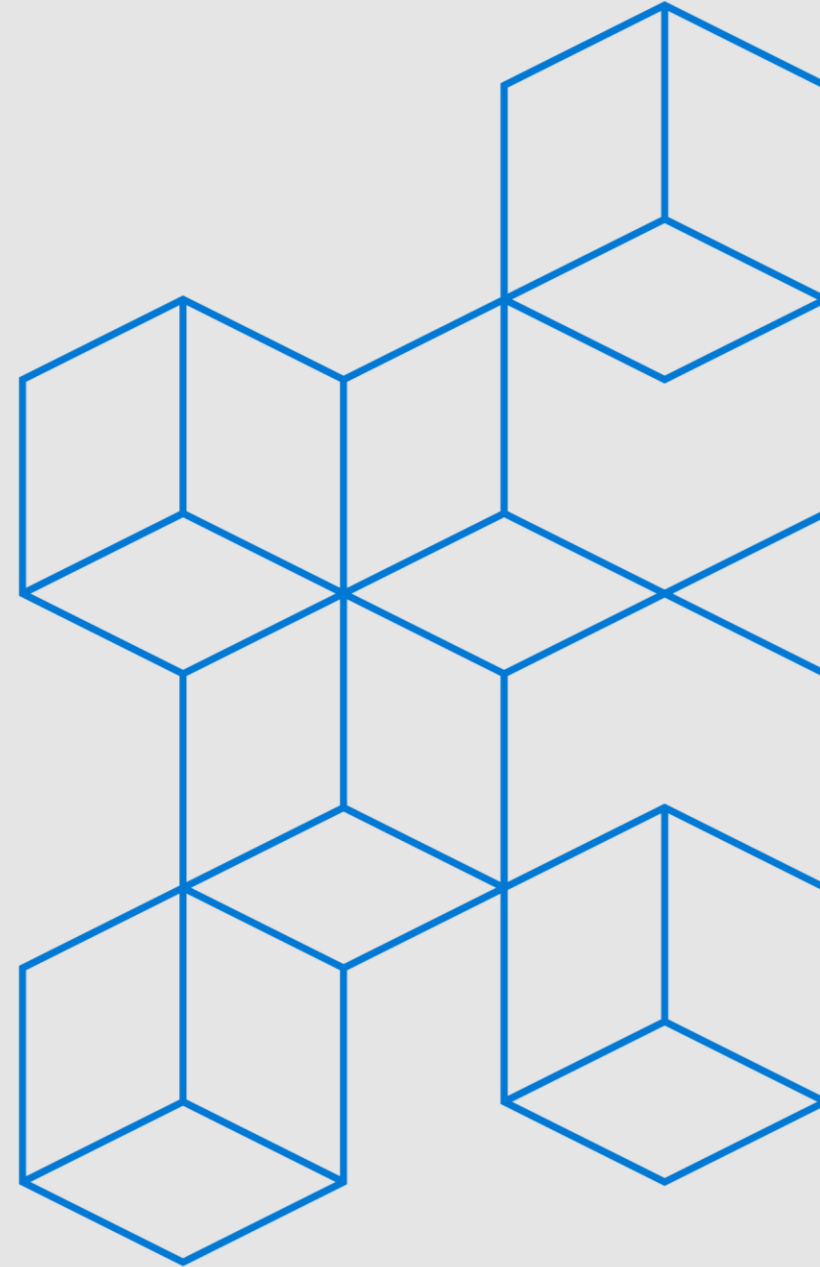


Extract Transform Load

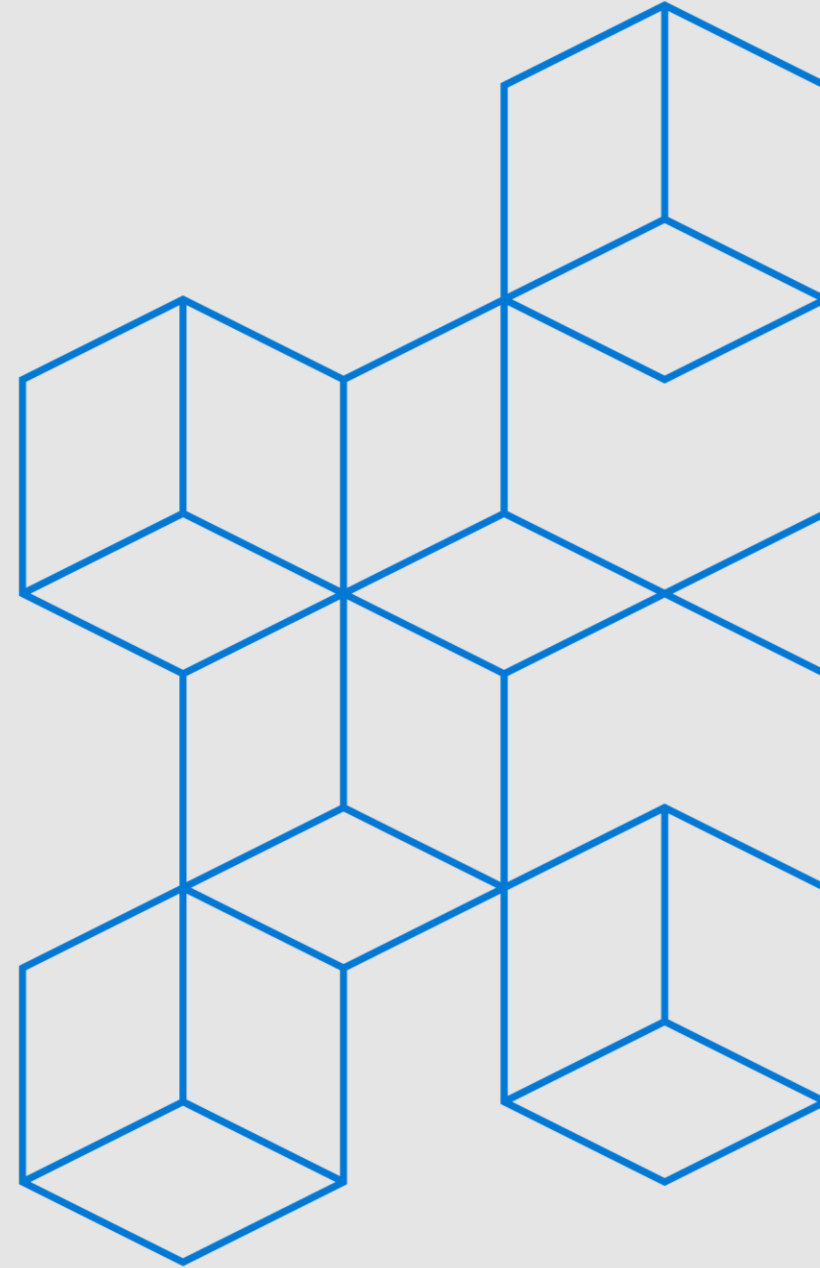
# ETL in Azure Made Easy with Data Factory Mapping Data Flows

**Paul Andrew**

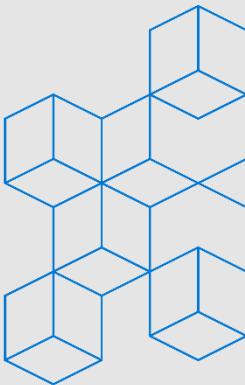
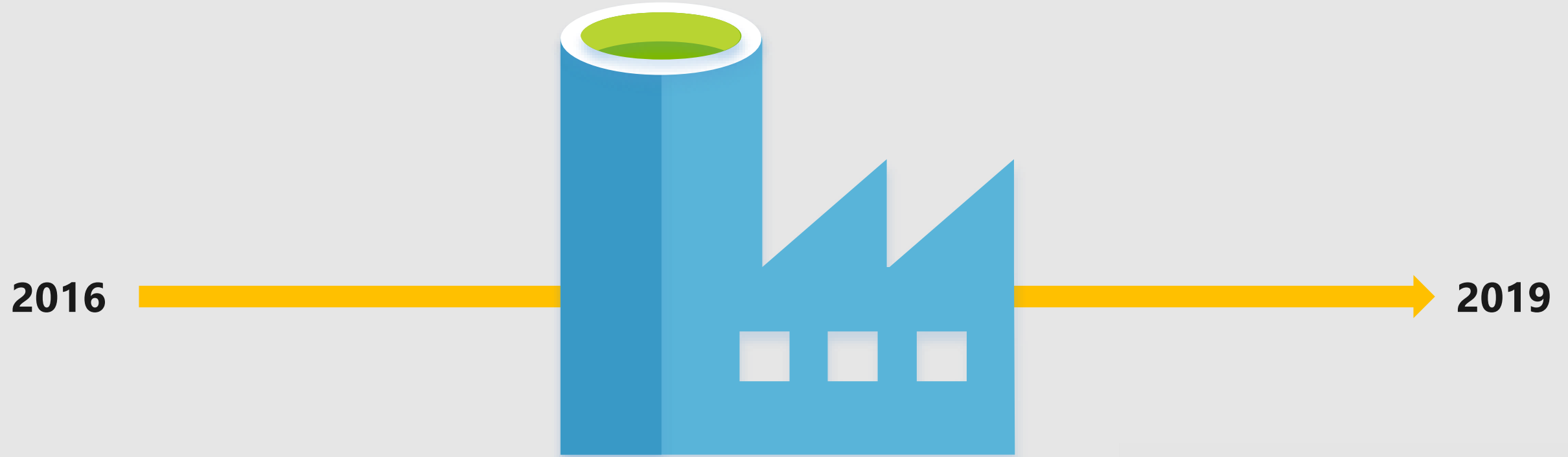
Data Platform MVP & Solution Architect



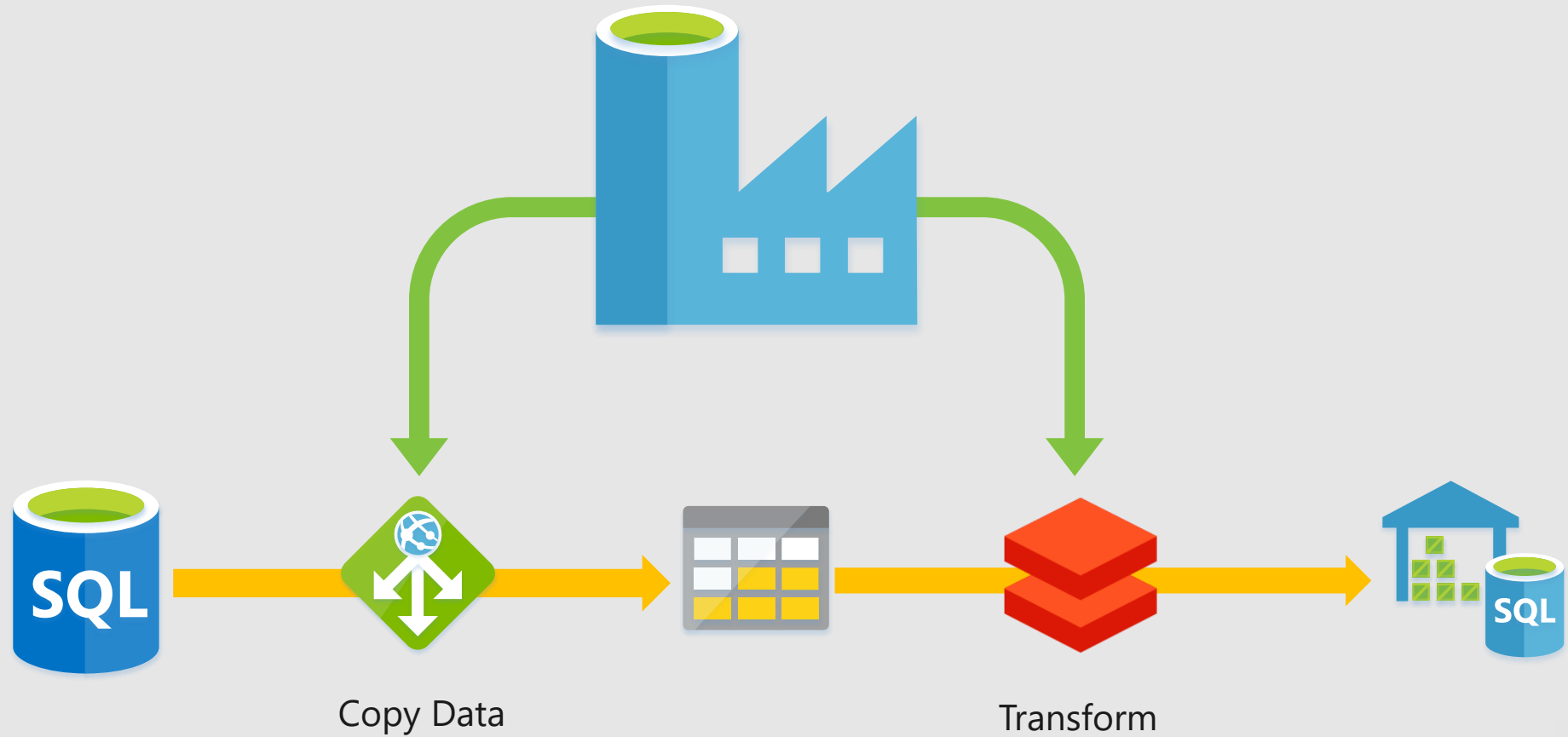
# Azure Data Factory



# What is Azure Data Factory?

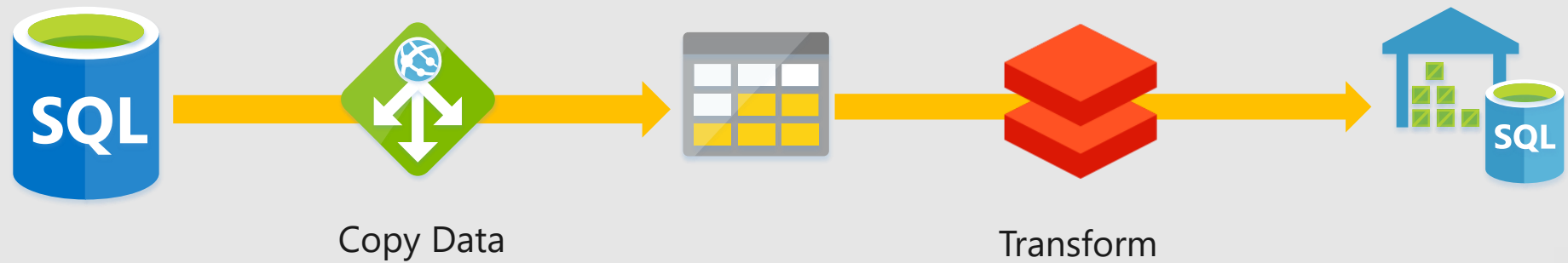


# What is Azure Data Factory?

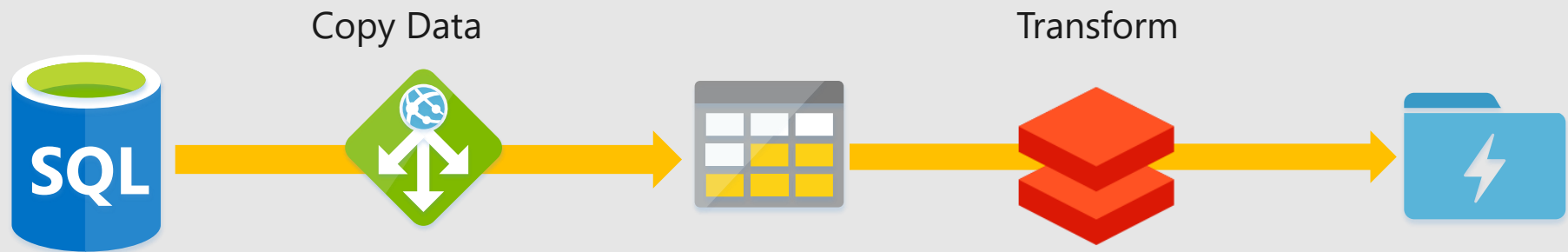




# What is Azure Data Factory?

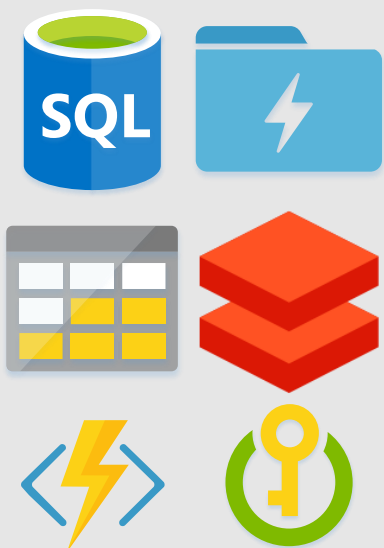


# Data Factory Components



1

**Linked Services** – How do I connect? Like the SSIS connection manager.





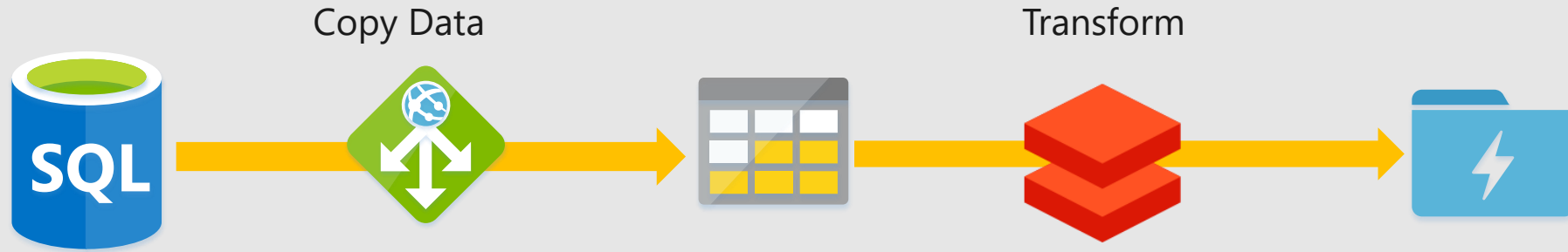
SQLDBLinkedService

ConnectionString: *Server=MyServer;Database=myDataBase*  
UserName: *"MrPaulAndrew"*  
Password: *\*\*\*\*\**

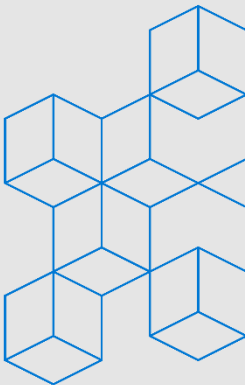




























































































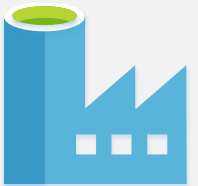
# Data Factory Components



## 1 Linked Services

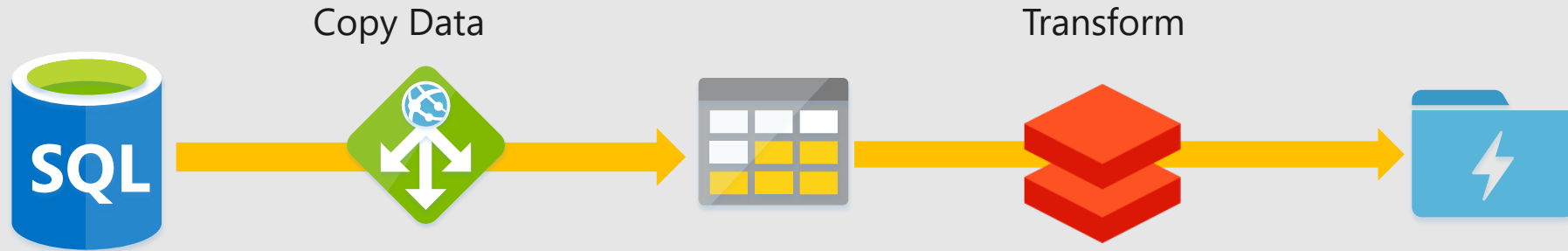


 Amazon Marketplace Web Service (Preview)	 Amazon Redshift	 Amazon S3	 HDFS	 HTTP	 Hive	 Nettezza	 ODBC	 OData	 Azure Batch	 Azure Data Lake Analytics	 Azure Databricks
 Apache Impala (Preview)	 Azure Blob Storage	 Azure Cosmos DB (MongoDB API)	 HubSpot (Preview)	 Informix	 Jira (Preview)	 Office 365 (Preview)	 Oracle	 Oracle Eloqua (Preview)	 Azure Function	 Azure HDInsight	 Azure ML
 Azure Cosmos DB (SQL API)	 Azure Data Explorer (Kusto)	 Azure Data Lake Storage Gen1	 Magento (Preview)	 MariaDB	 Marketo (Preview)	 Oracle Responsys (Preview)	 Oracle Service Cloud (Preview)	 Paypal (Preview)	 ServiceNow	 Shopify (Preview)	 Spark
 Azure Data Lake Storage Gen2 (Preview)	 Azure Database for MariaDB	 Azure Database for MySQL	 Microsoft Access	 MongoDB	 MySQL	 Phoenix	 PostgreSQL	 Presto (Preview)	 Square (Preview)	 Sybase	 Teradata
 Azure Database for PostgreSQL	 Azure File Storage	 Azure Key Vault	 DB2	 Drill (Preview)	 Dynamics 365	 QuickBooks (Preview)	 REST	 SAP BW Open Hub	 Vertica	 Web Table	 Xero (Preview)
 Azure SQL Data Warehouse	 Azure SQL Database	 Azure SQL Database Managed Instance	 Dynamics AX (Preview)	 Dynamics CRM	 FTP	 SAP BW via MDX	 SAP Cloud For Customer	 SAP ECC	 Zoho (Preview)		
 Azure Search	 Azure Table Storage	 Cassandra	 File System	 Google AdWords (Preview)	 Google BigQuery	 SAP HANA	 SFTP	 SQL Server			
 Common Data Service for Apps	 Concur (Preview)	 Couchbase (Preview)	 Google Cloud Storage (S3 API)	 Greenplum	 HBase	 Salesforce	 Salesforce Marketing Cloud (Preview)	 Salesforce Service Cloud			



**88x** linked services supported as 1<sup>st</sup> class citizens within Azure Data Factory. As of 5<sup>th</sup> Feb 2019.

# Data Factory Components



1

## Linked Services

2

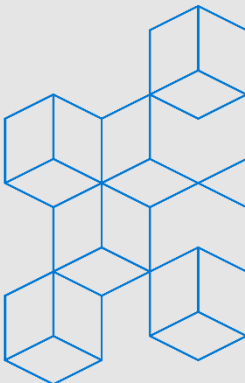
**Data Sets** – Where is my data? What format? What file path/table do I need?



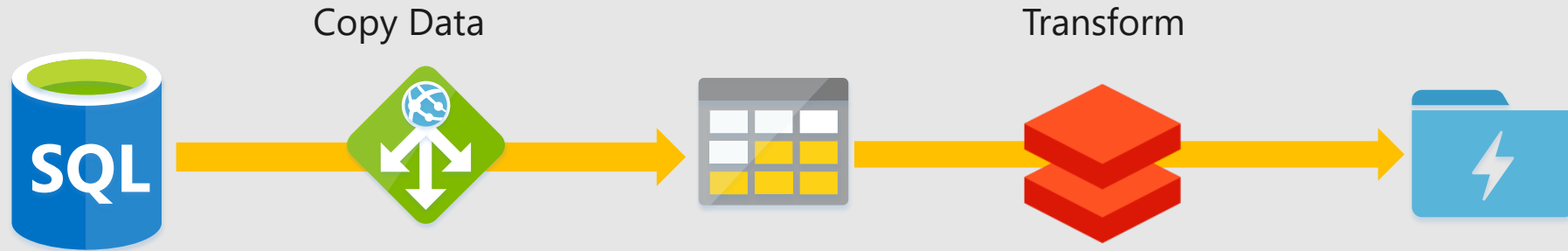
dbo.DimCustomer



/RAW/Orders/2018/01/01/Orders.csv



# Data Factory Components



1

**Linked Services**

2

**Data Sets**

3

**Activities** – What do we want to happen?  
With what conditions?

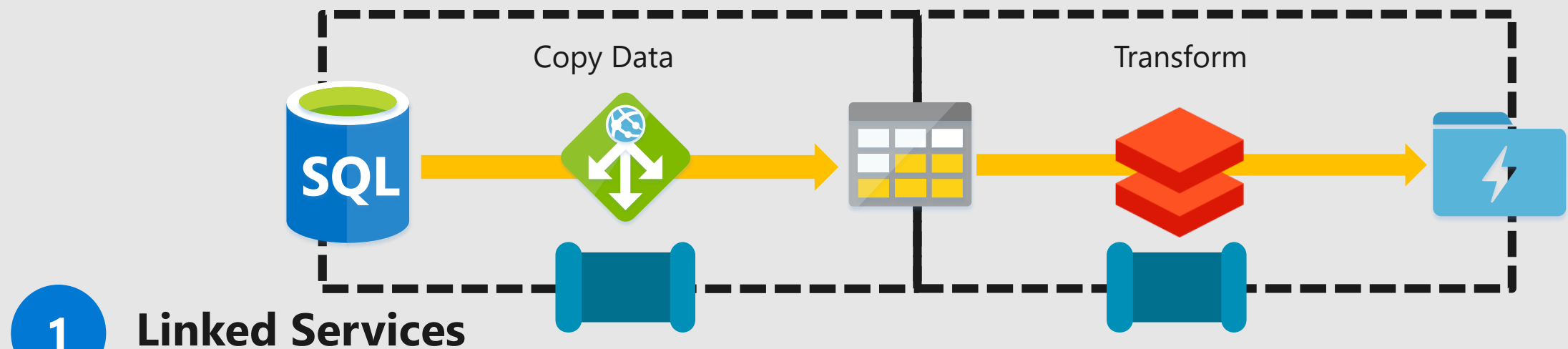


## Databricks Notebook Activity

```
notebookPath: /Playground/Playing
baseParameters: Testing
libraries[jar]: dbfs:/lib1.jar
linkedServiceName: BricksOfData01
```



# Data Factory Components



1

**Linked Services**

2

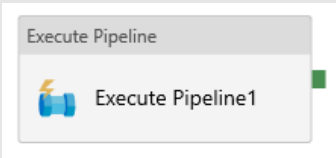
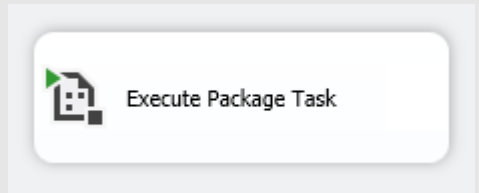
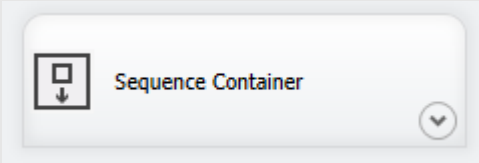
**Data Sets**

3

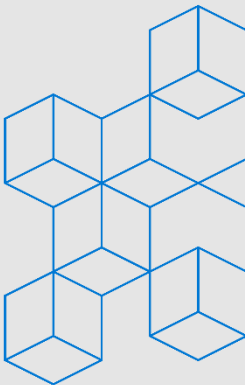
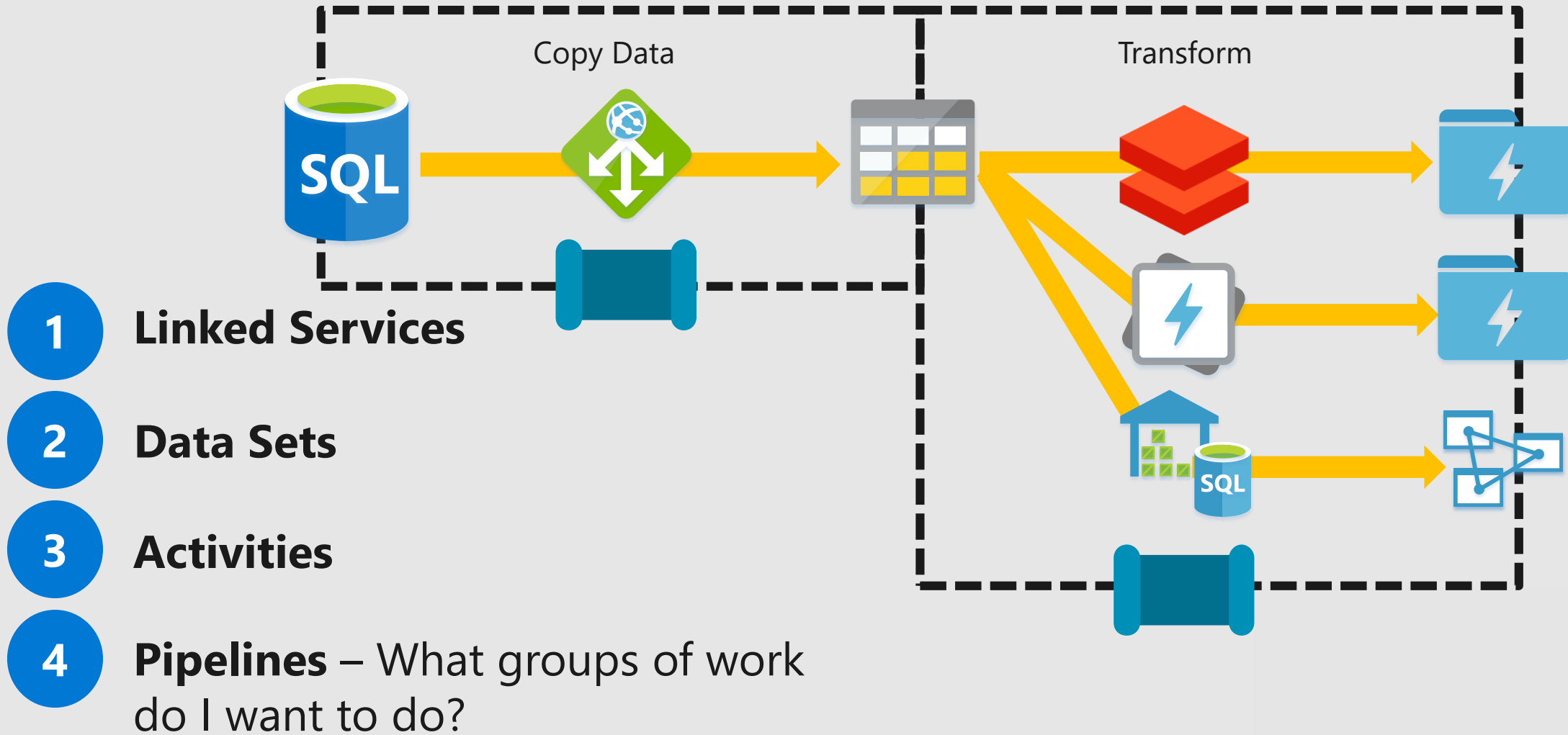
**Activities**

4

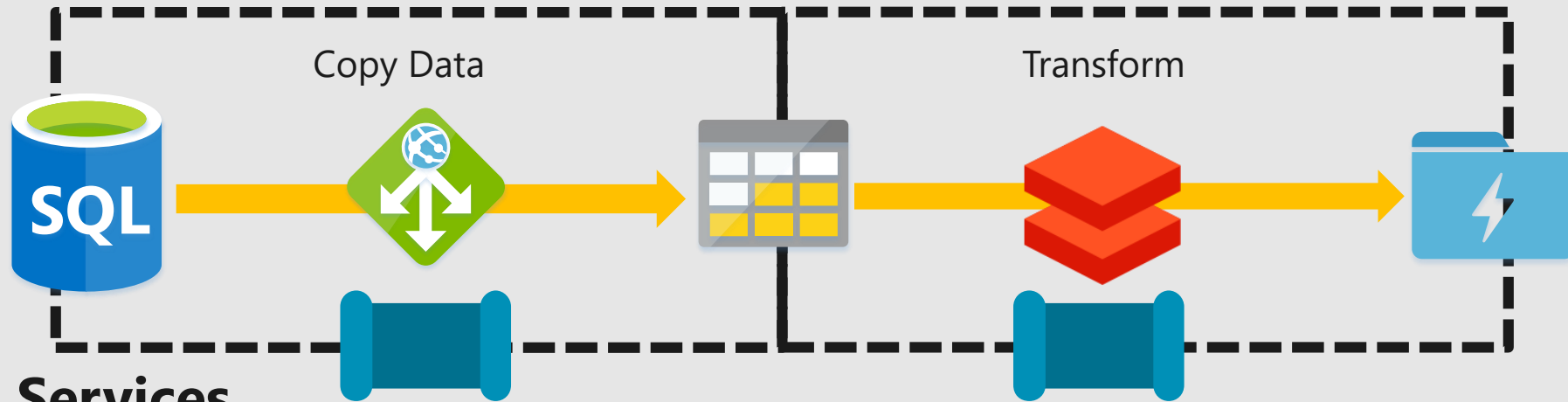
**Pipelines** – What groups of work do I want to do?



# Data Factory Components



# Data Factory Components



1

**Linked Services**

2

**Data Sets**

3

**Activities**

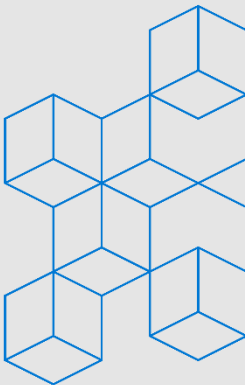
4

**Pipelines**

5

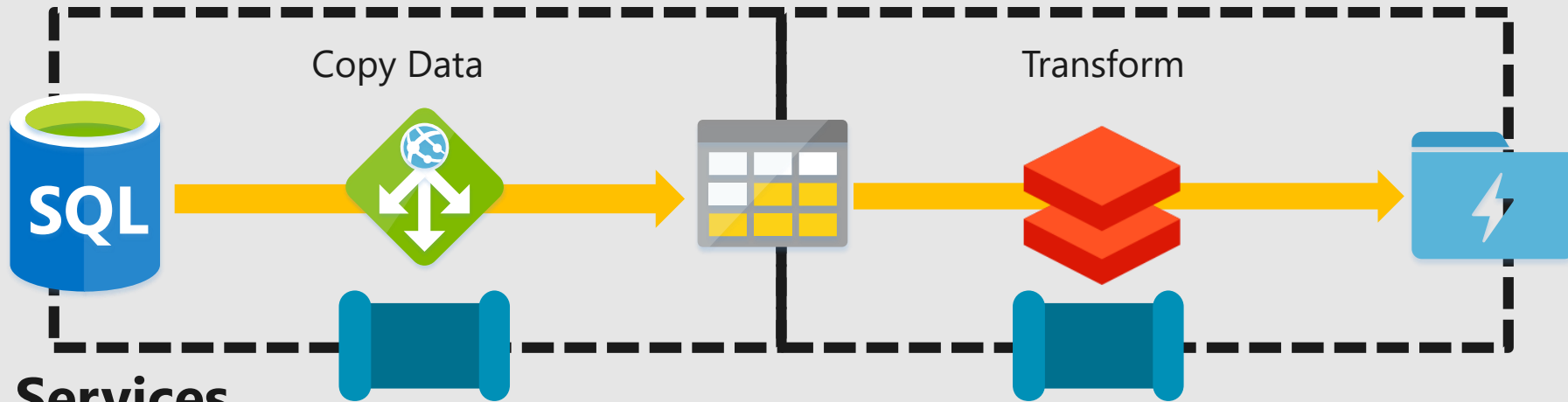
**Triggers** – How are we going to tell our pipeline(s) to execute?

- Manual via UI
- Tumbling Windows
- Scheduled
- Blob File Events
- Logic App Calls





# Data Factory Components



1

Linked Services

2

Data Sets

3

Activities

4

Pipelines

5

Triggers

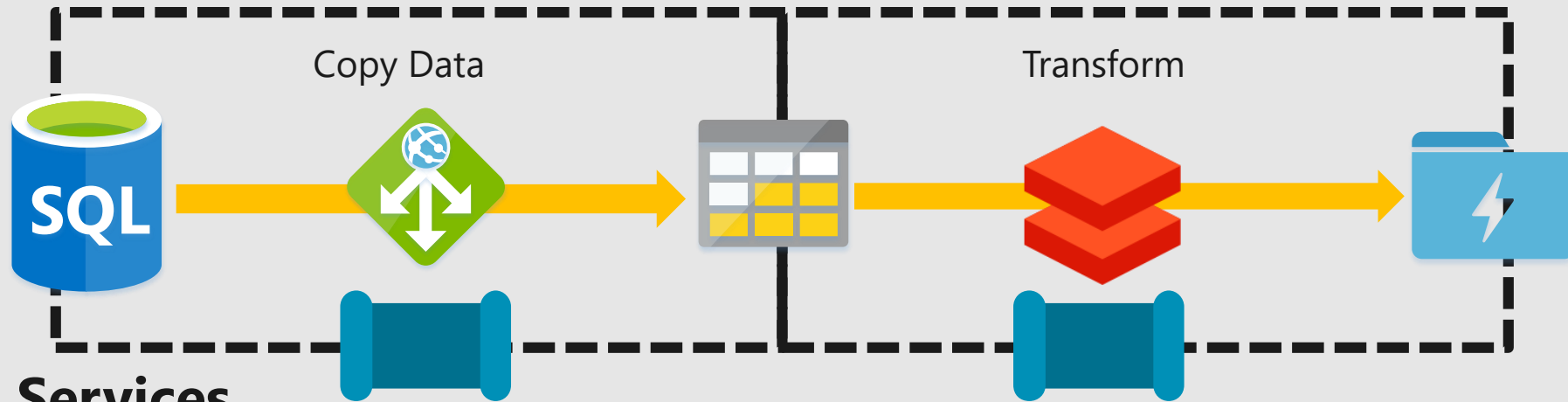
- **Manual**
- Tumbling Windows
- Scheduled
- Blob File Events
- Logic App Calls



```
Invoke-AzureRmDataFactoryV2Pipeline  
-DataFactoryName $dataFactoryName  
-ResourceGroupName $resourceGroupName  
-PipelineName $pipelineName
```



# Data Factory Components



1

Linked Services

2

Data Sets

3

Activities

4

Pipelines

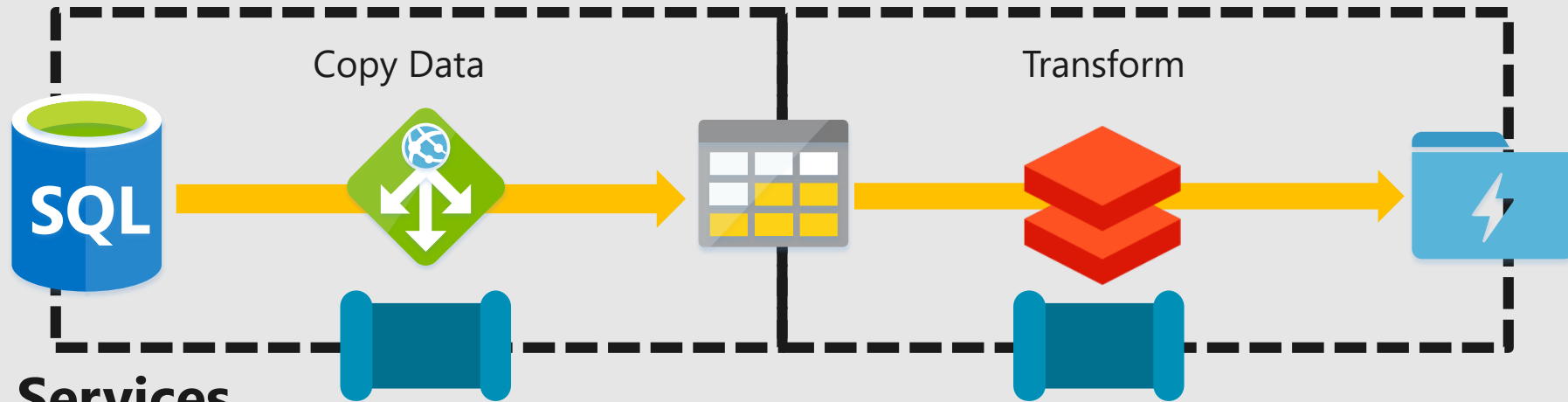
5

Triggers

- Manual via UI
- **Tumbling Windows** - AKA Time Slices
- Scheduled
- Blob File Events
- Logic App Calls



# Data Factory Components



1

Linked Services

2

Data Sets

3

Activities

4

Pipelines

5

Triggers

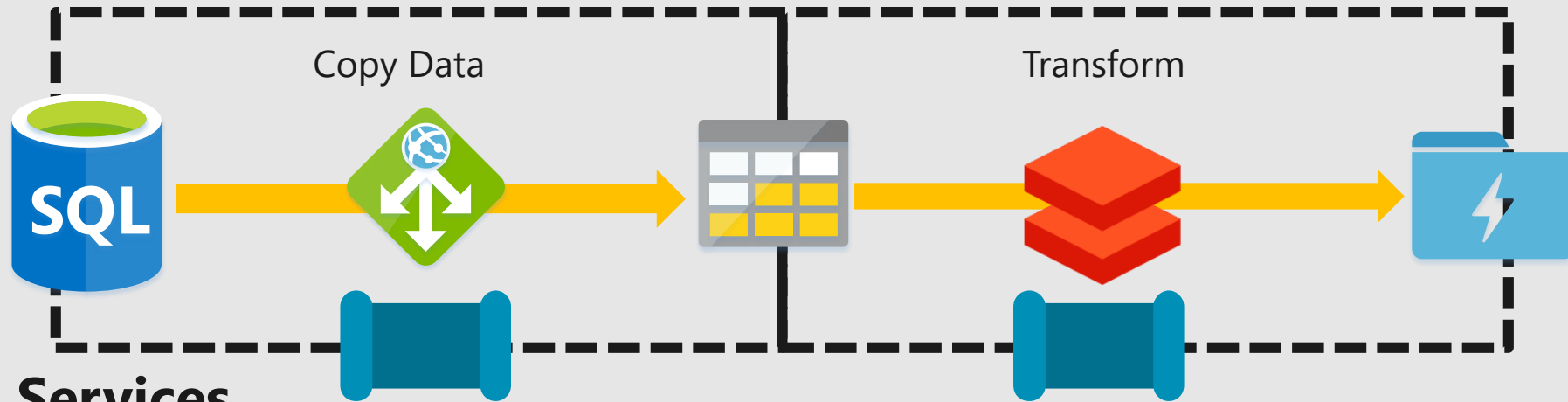
- Manual via UI
- Tumbling Windows
- **Scheduled**
- Blob File Events
- Logic App Calls



- Every 1 minute.
- UTC



# Data Factory Components



1

Linked Services

2

Data Sets

3

Activities

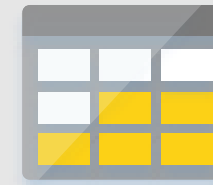
4

Pipelines

5

Triggers

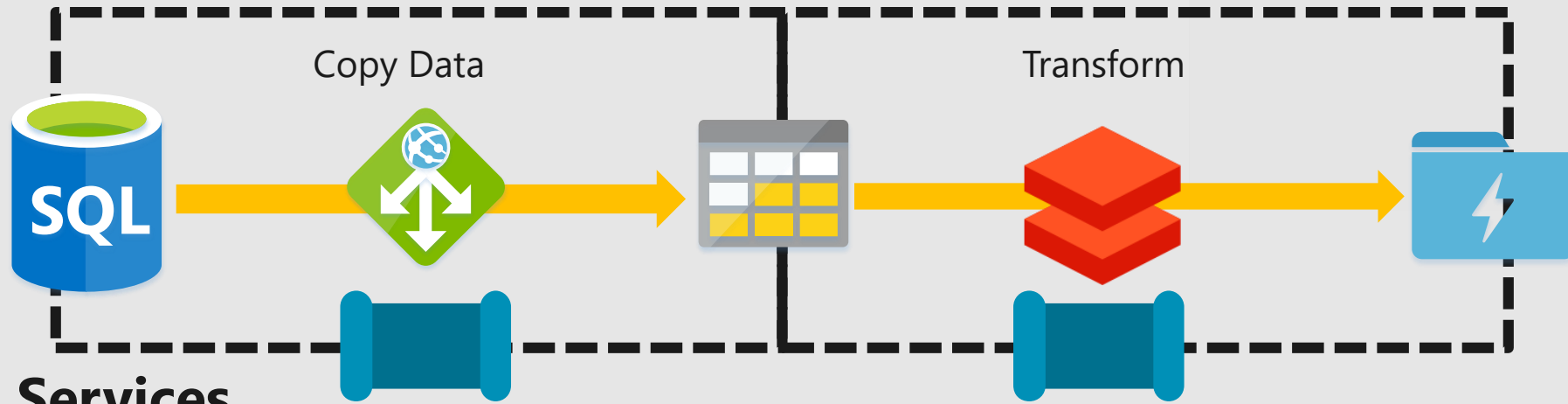
- Manual via UI
- Tumbling Windows
- Scheduled
- **Blob File Events**
- Logic App Calls



- {Path} Created
- {Path} Deleted



# Data Factory Components



1

Linked Services

2

Data Sets

3

Activities

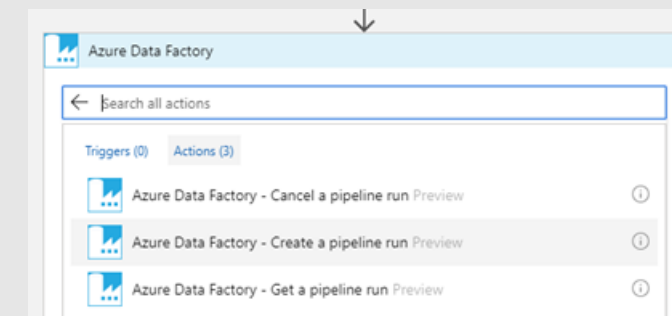
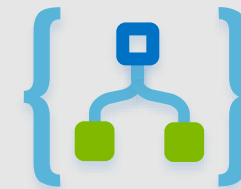
4

Pipelines

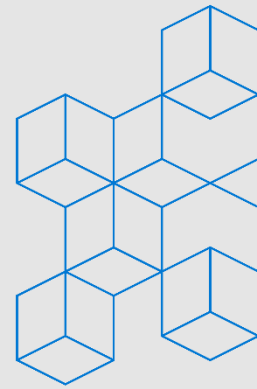
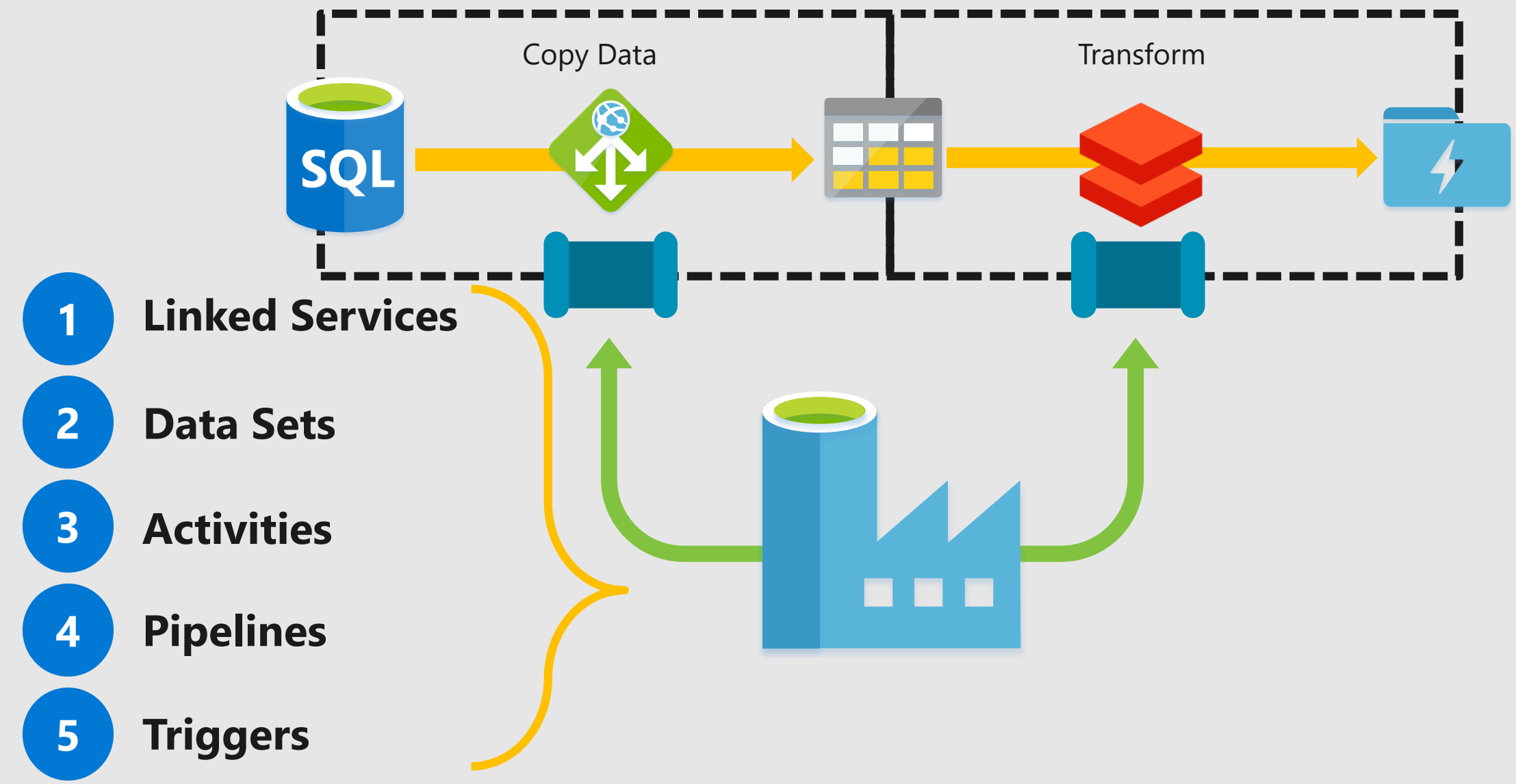
5

Triggers

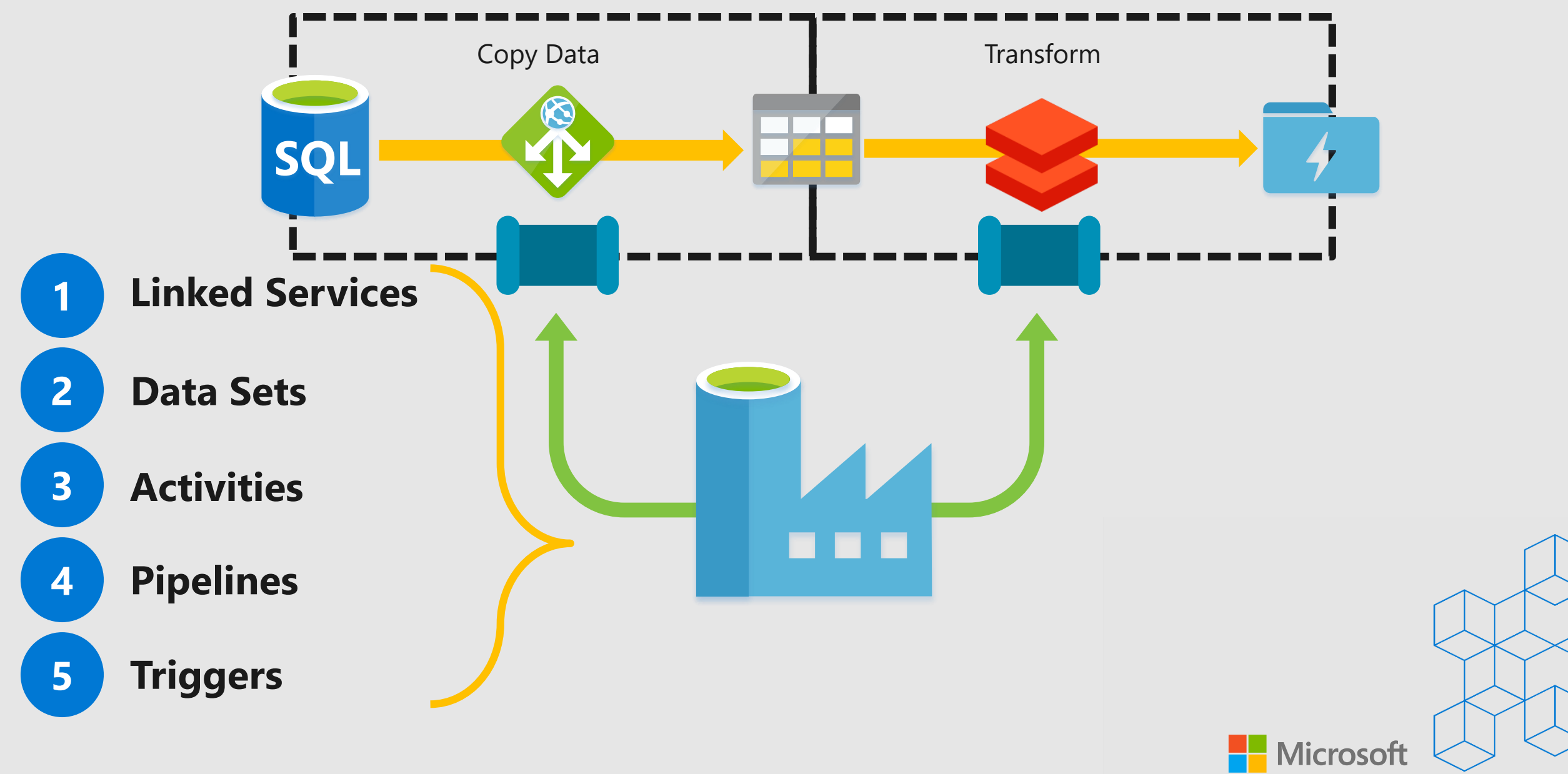
- Manual via UI
- Tumbling Windows
- Scheduled
- Blob File Events
- **Logic App Calls**



# Data Factory Components

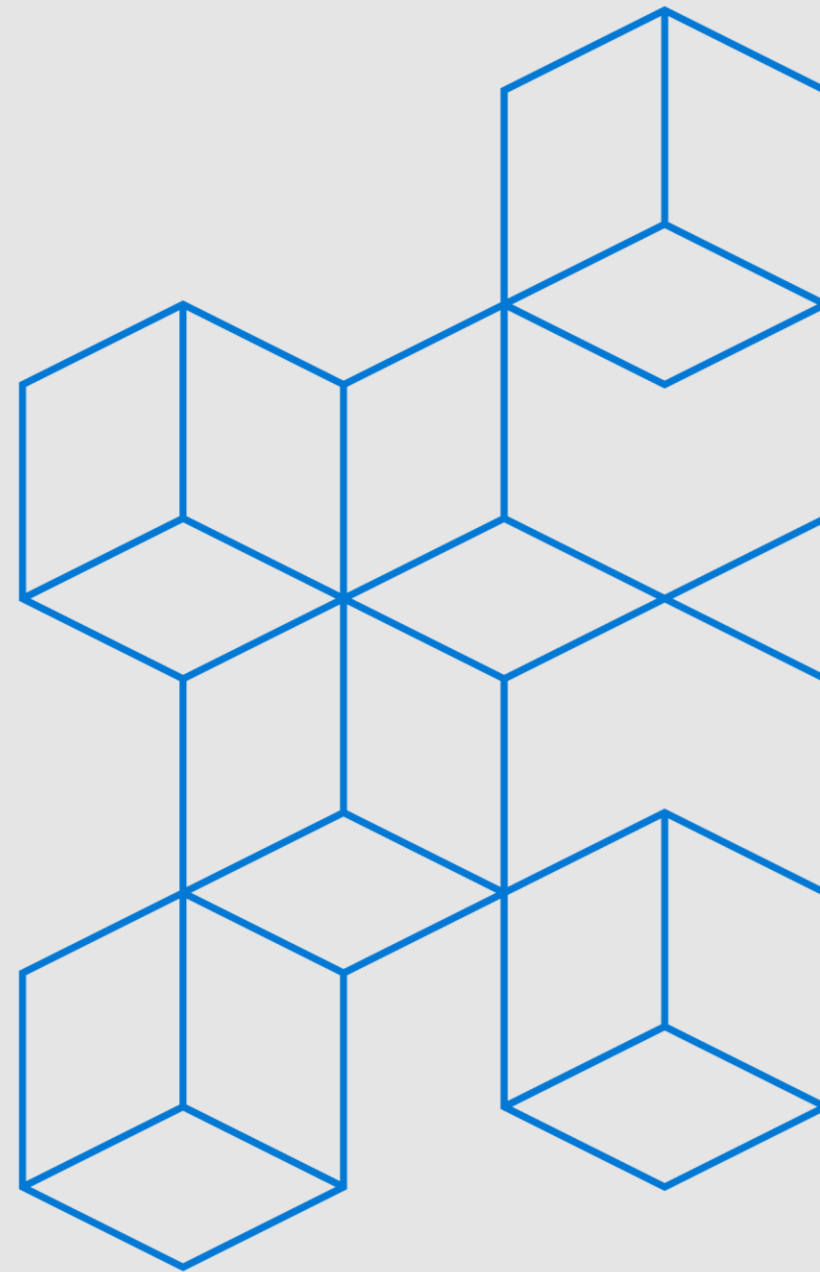


# Data Factory Control Flow Components

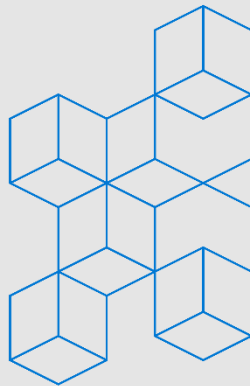
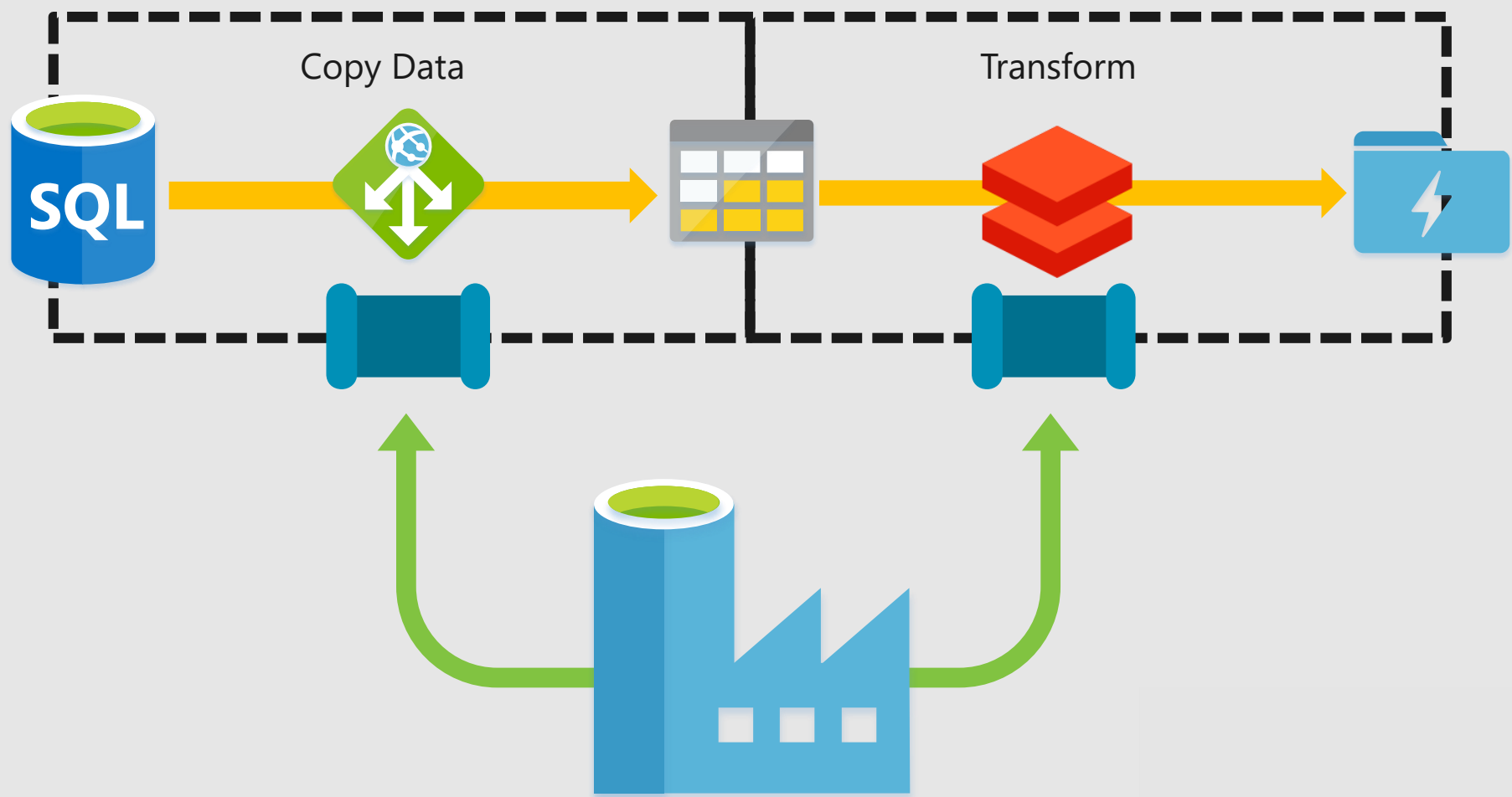




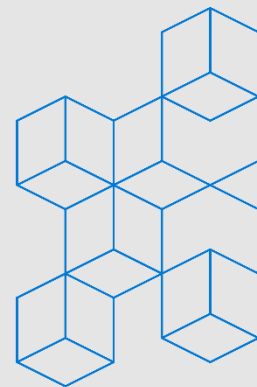
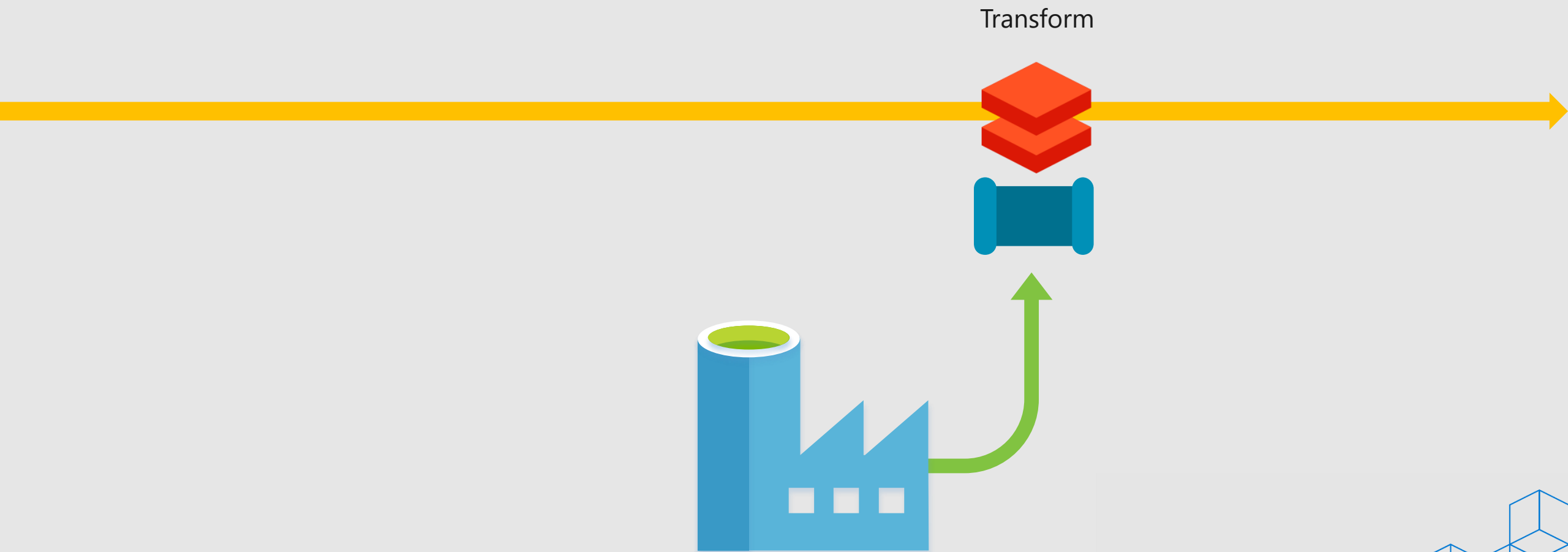
# Data Transformation in zure



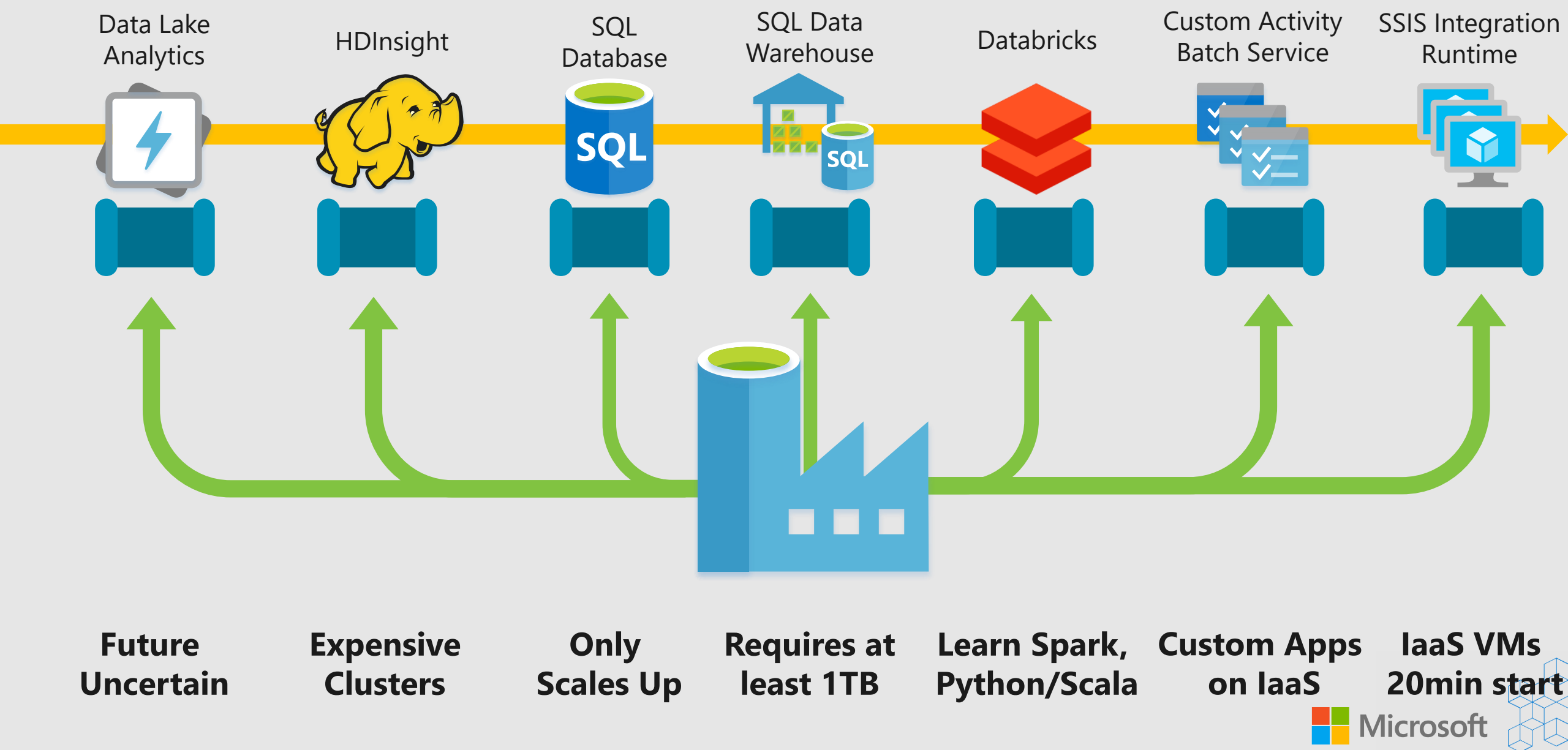
# Data Factory Control Flow Components



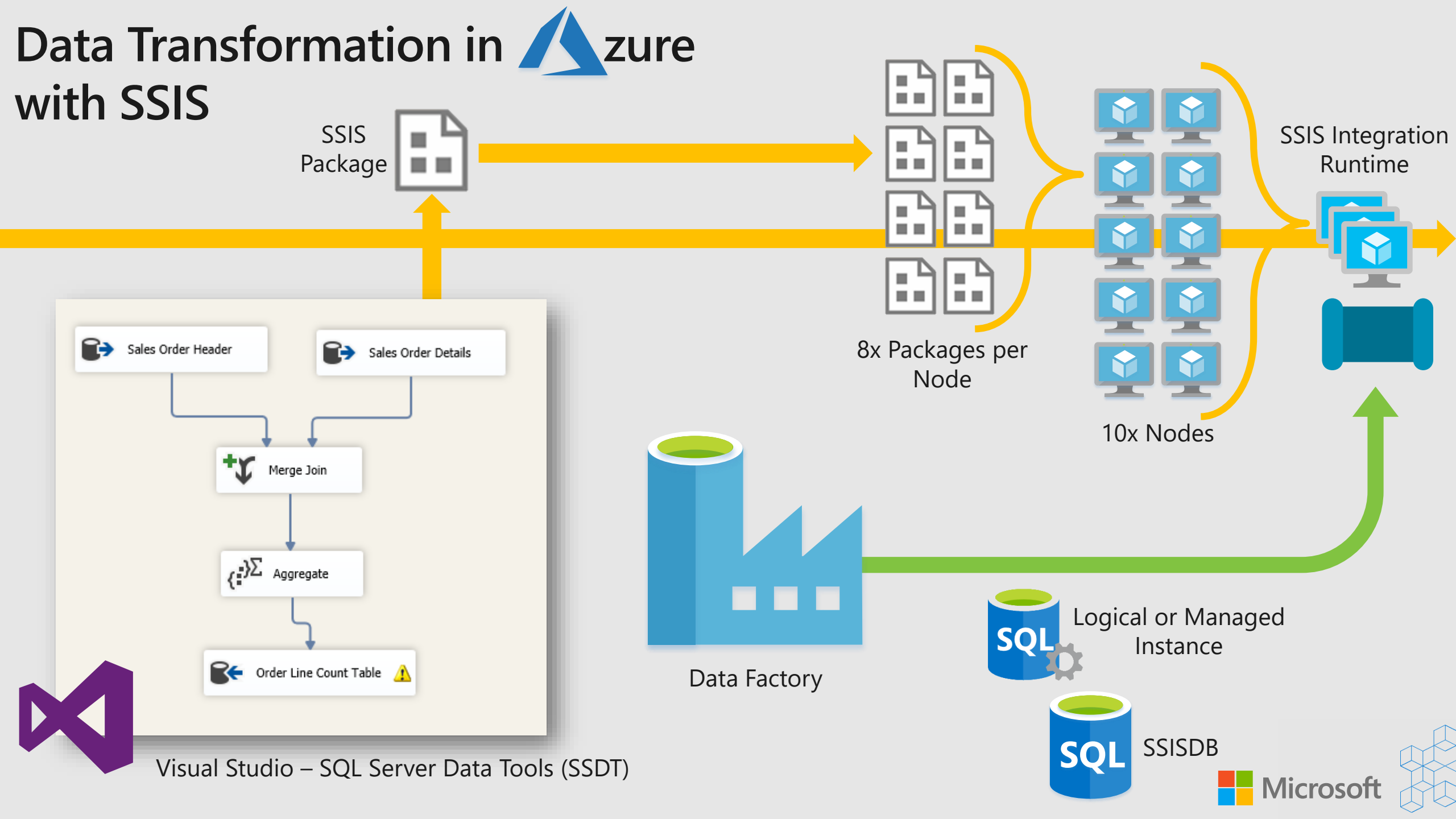
# Data Transformation in zure



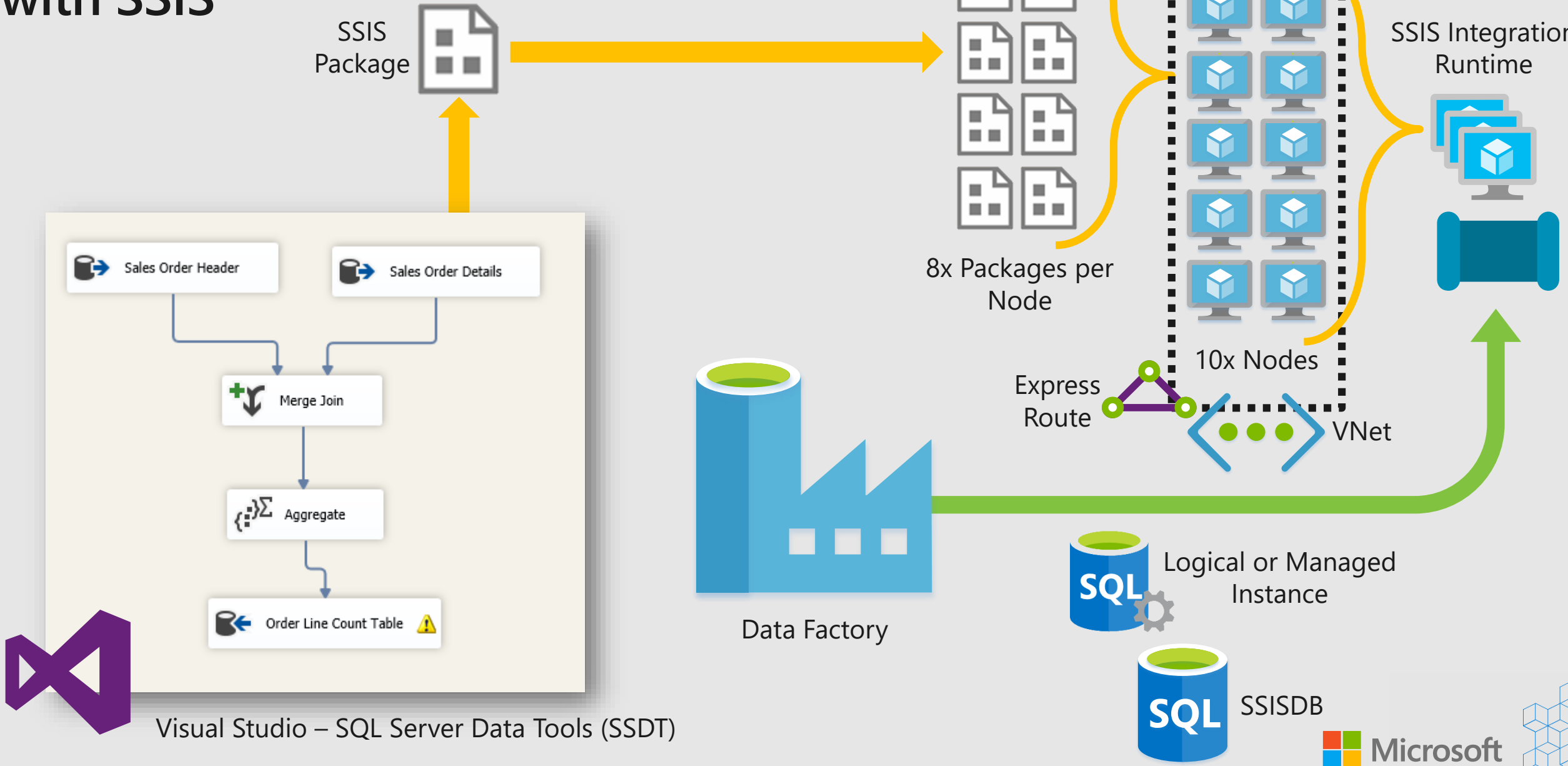
# Data Transformation in zure



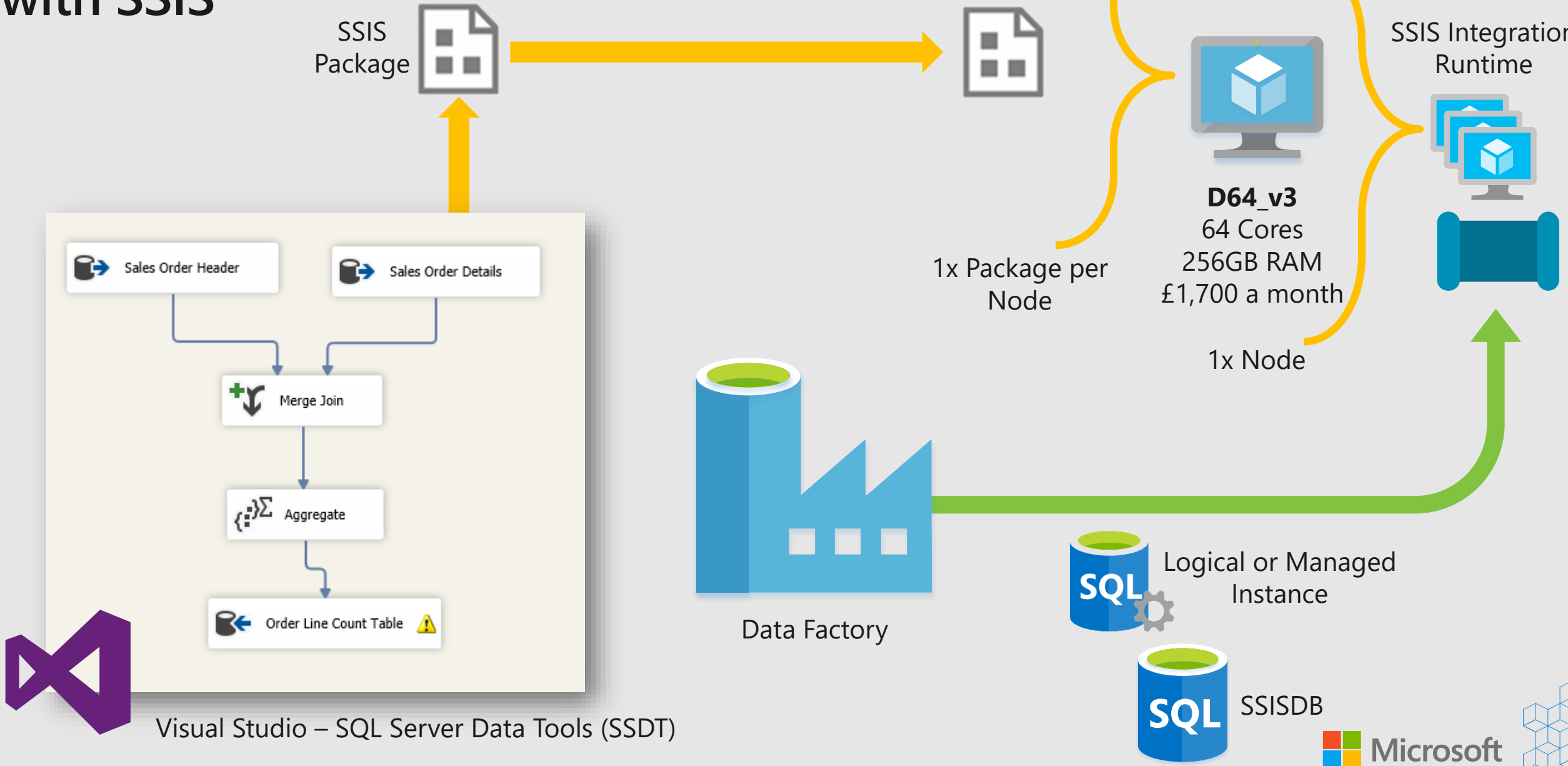
# Data Transformation in Azure with SSIS



# Data Transformation in Azure with SSIS

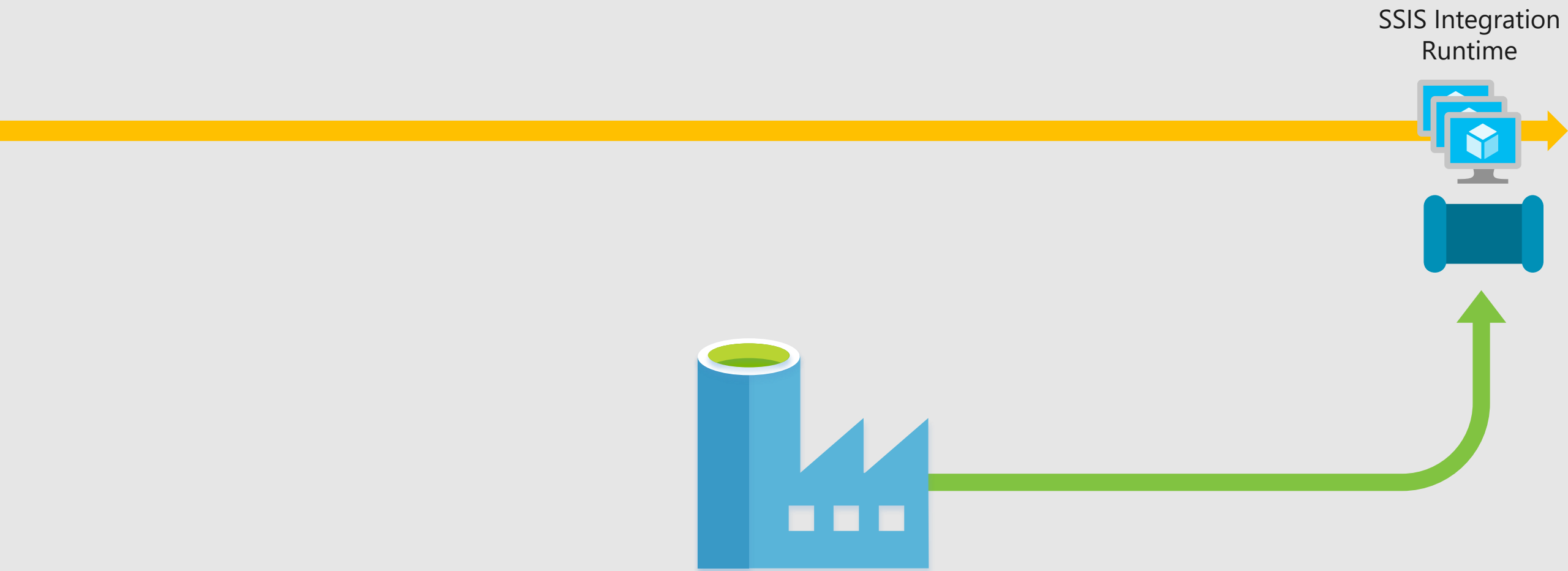


# Data Transformation in Azure with SSIS

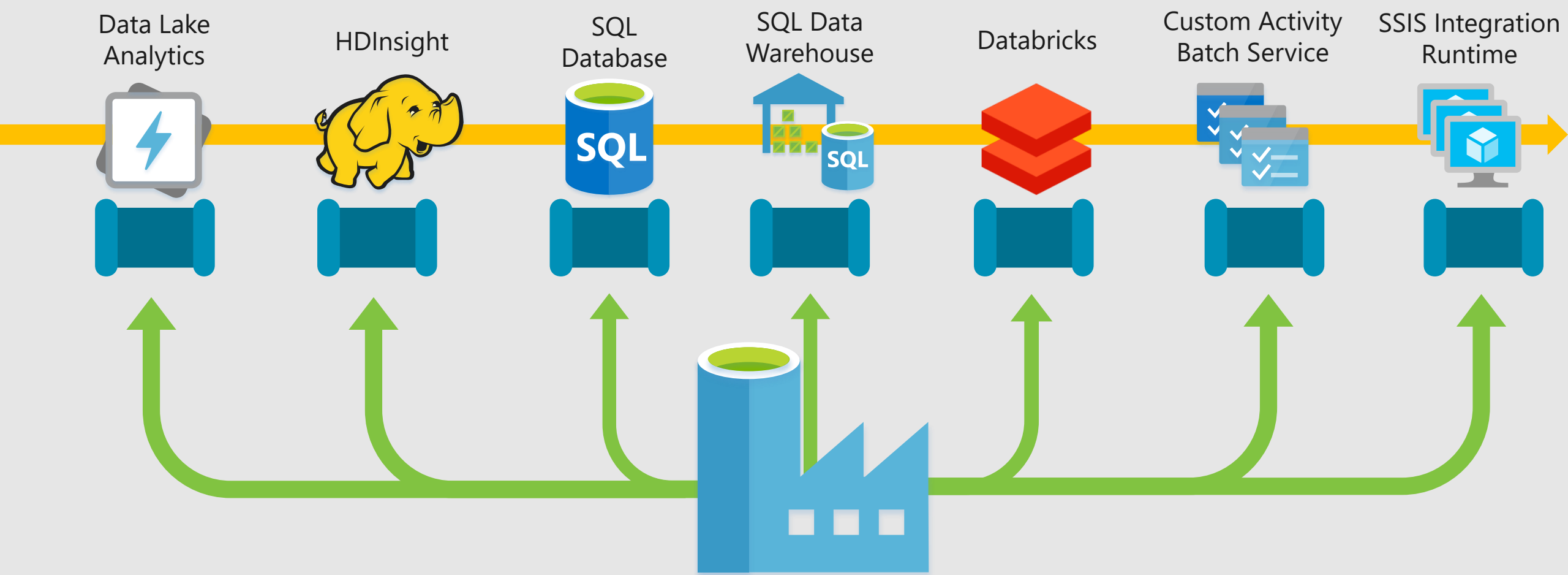




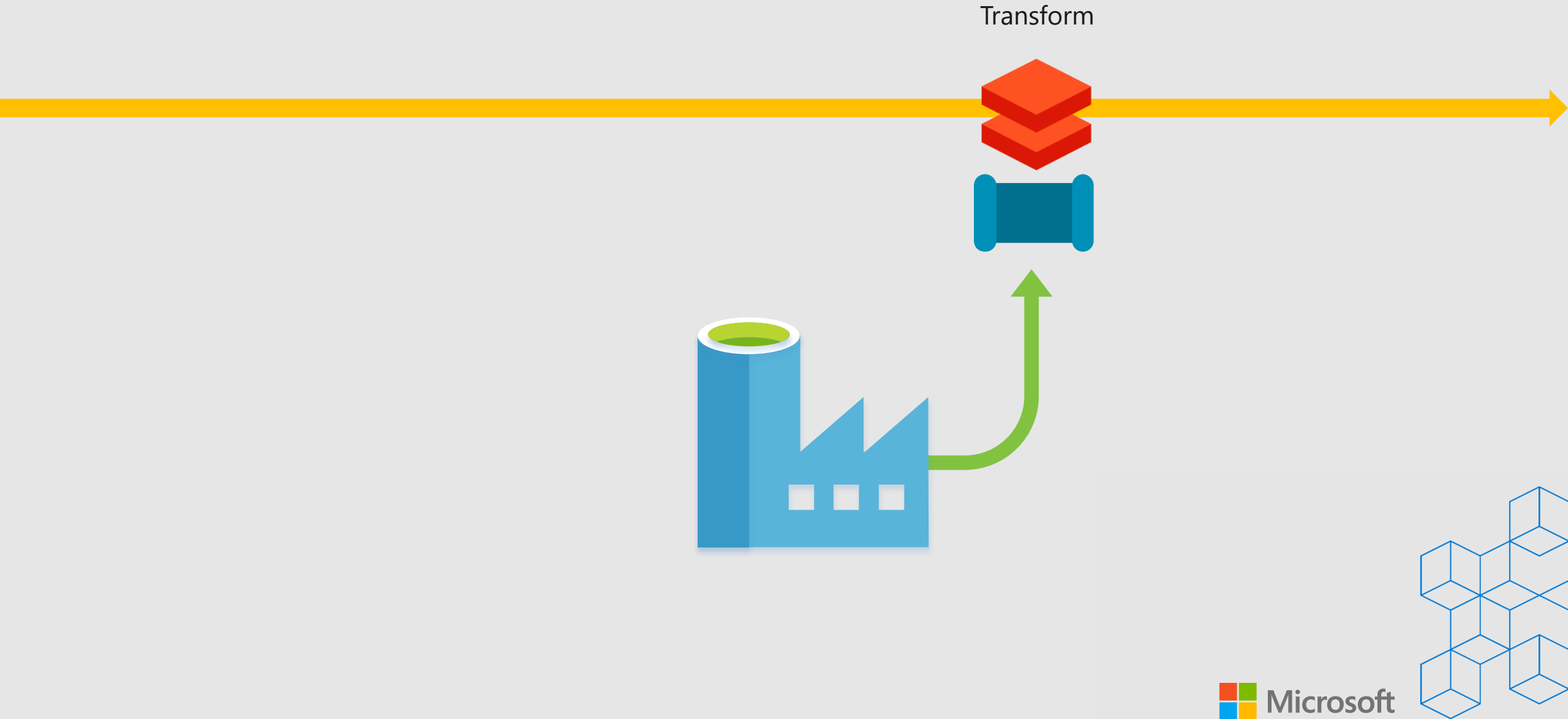
# Data Transformation in zure



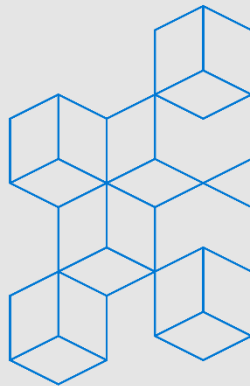
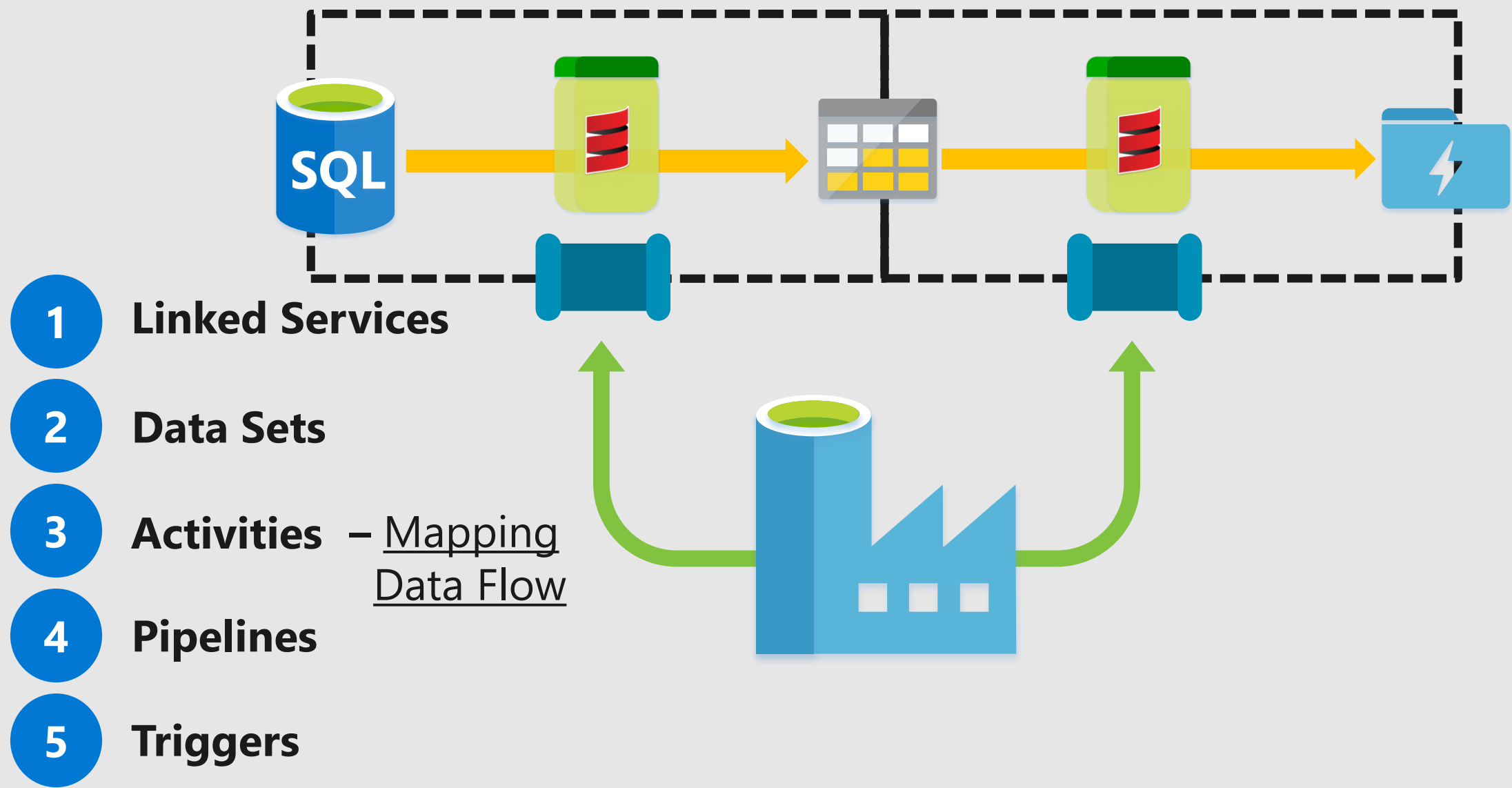
# Data Transformation in zure



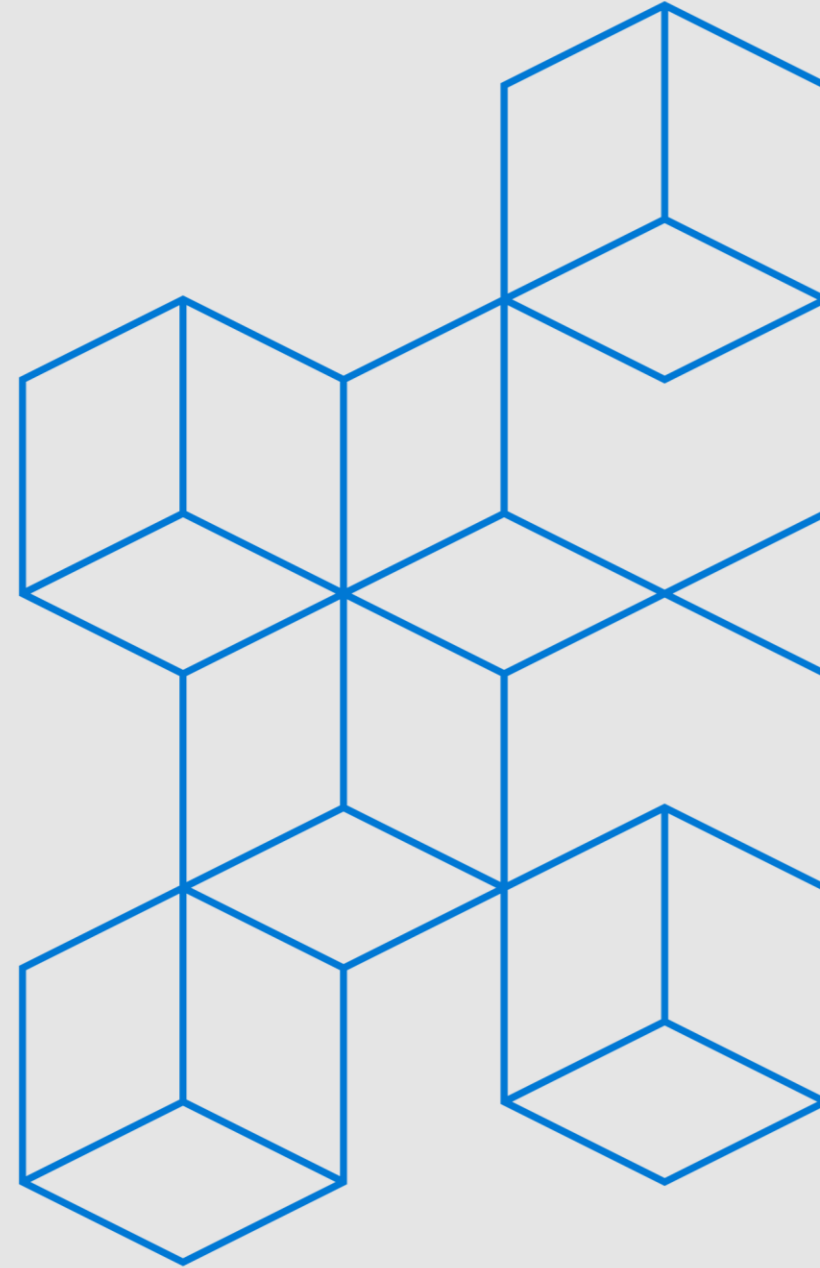
# Data Transformation in zure



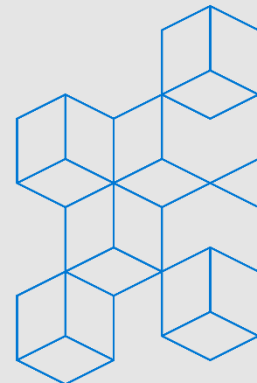
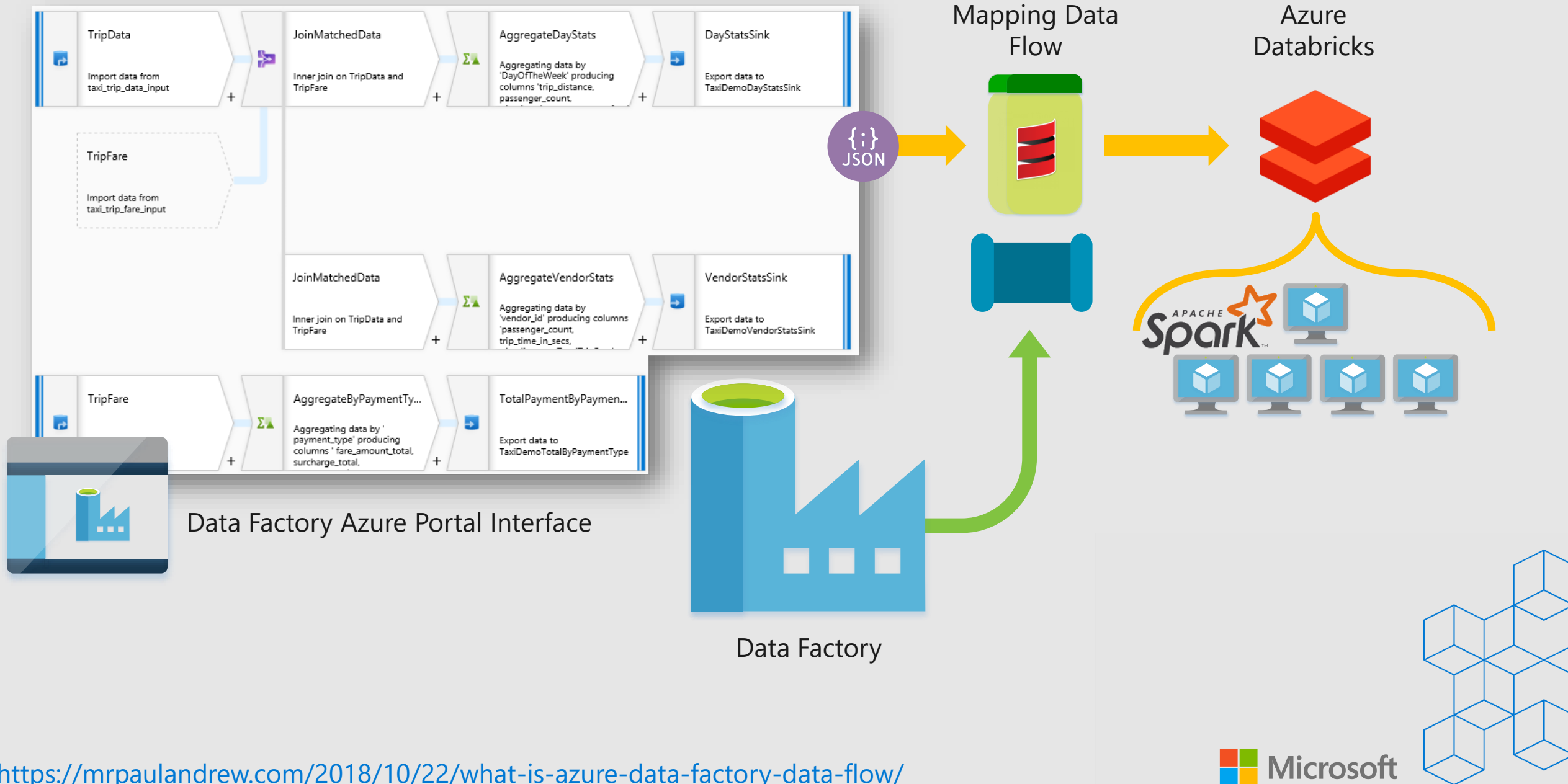
# Data Factory Control Flow Components



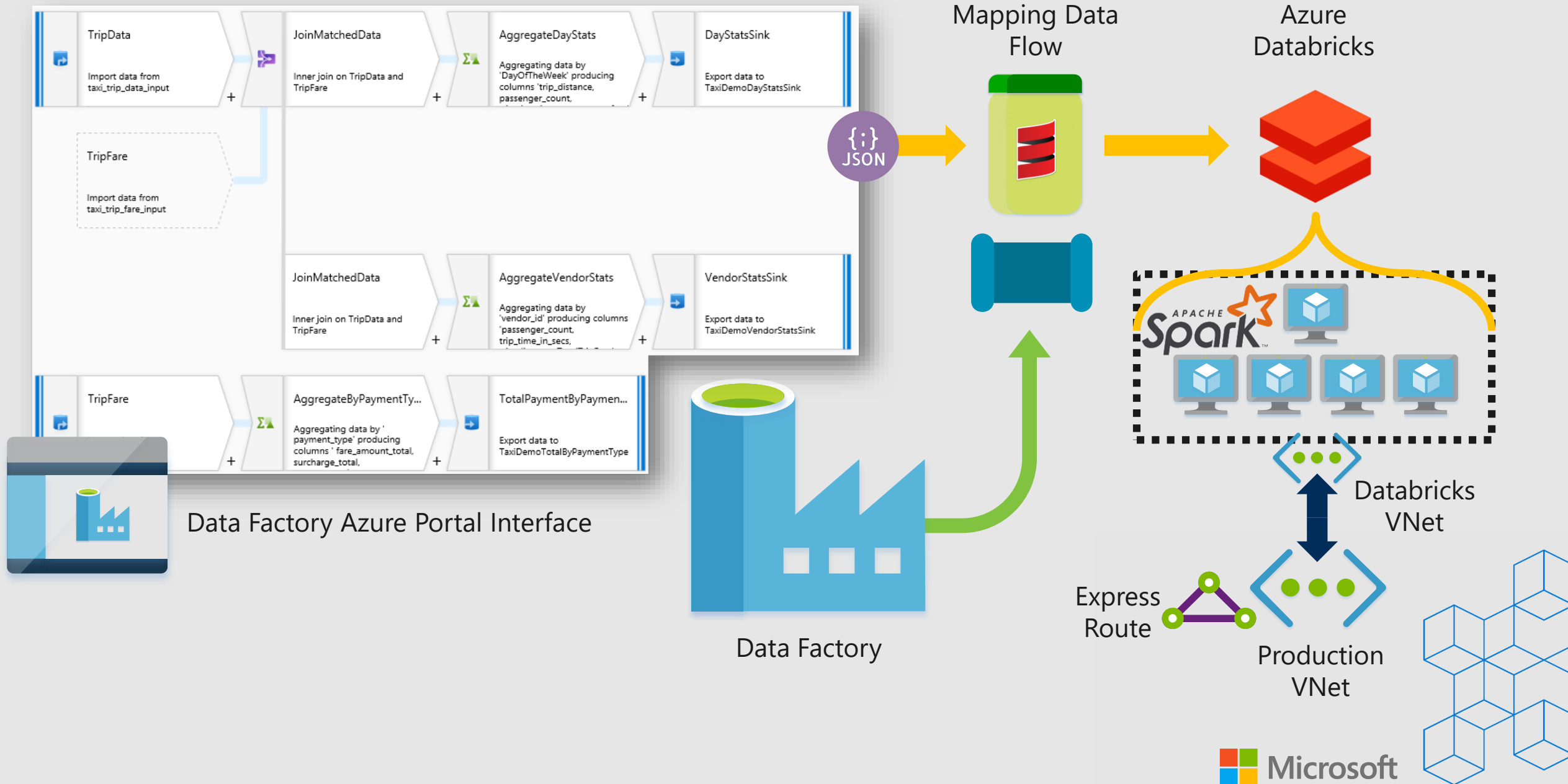
# Mapping Data Flows



# What is a Mapping Data Flow?

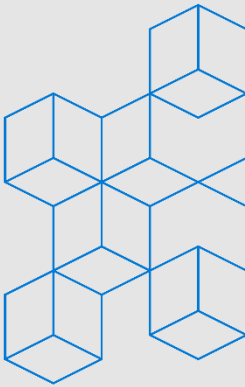
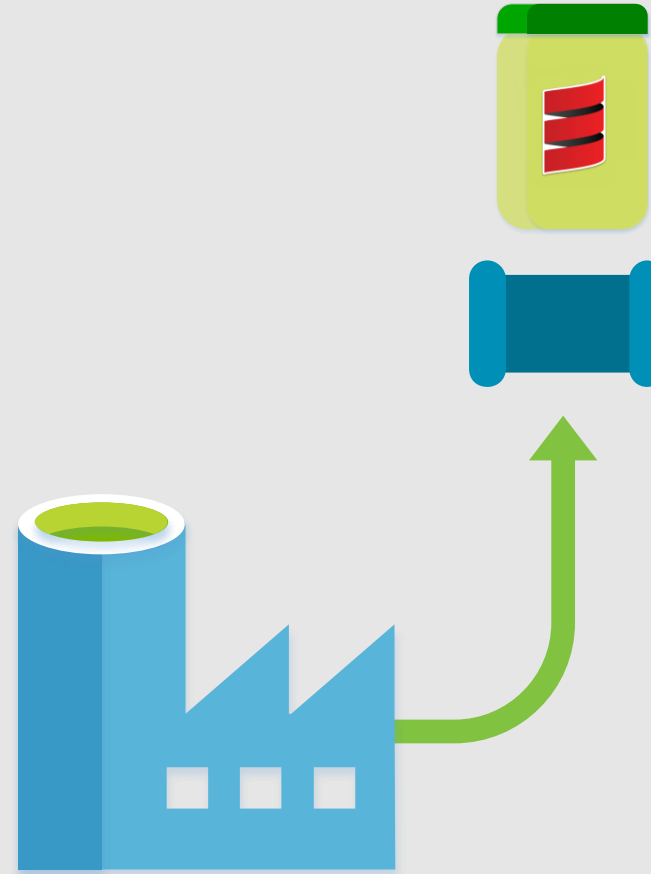


# What is a Mapping Data Flow?

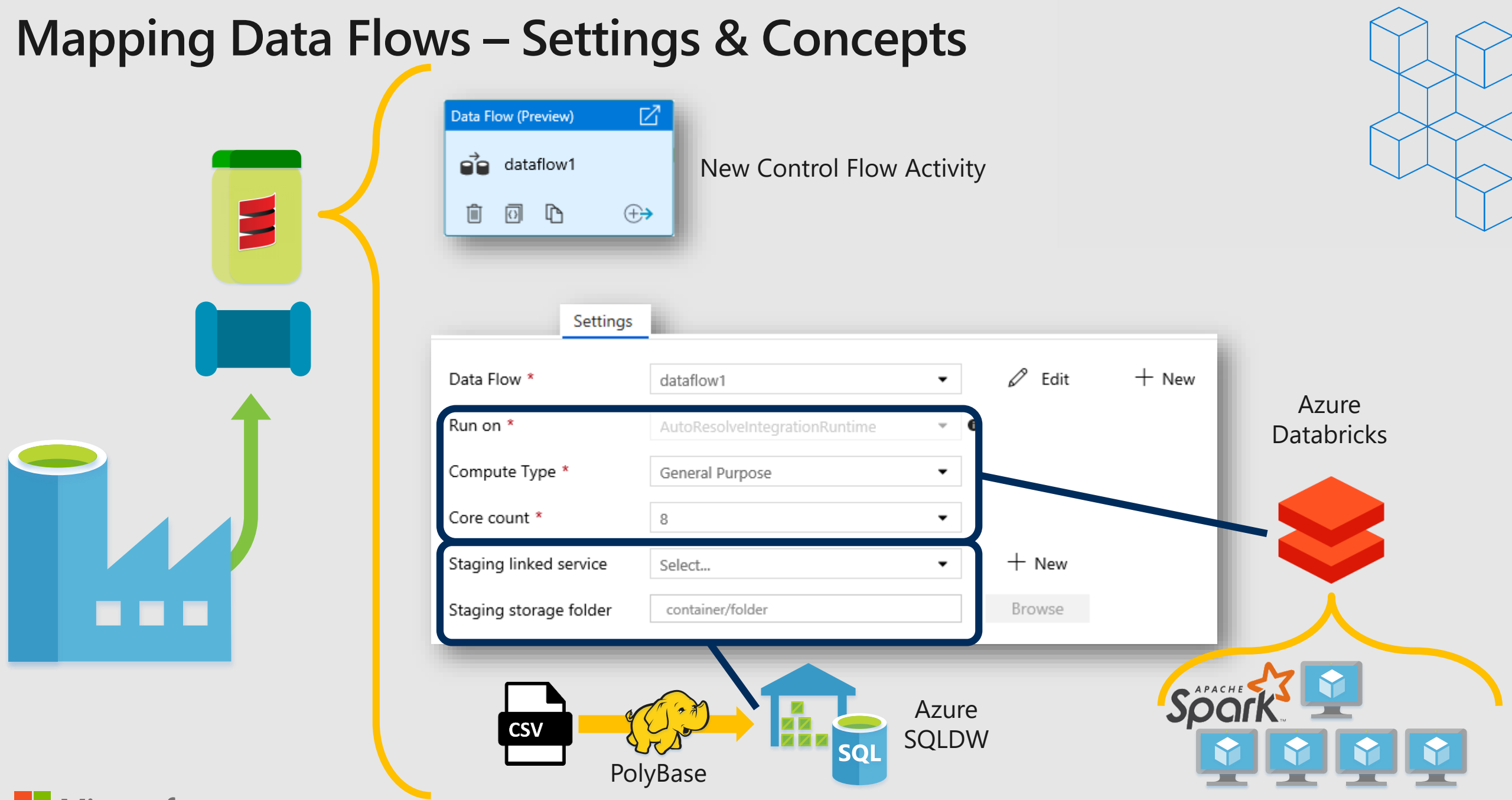




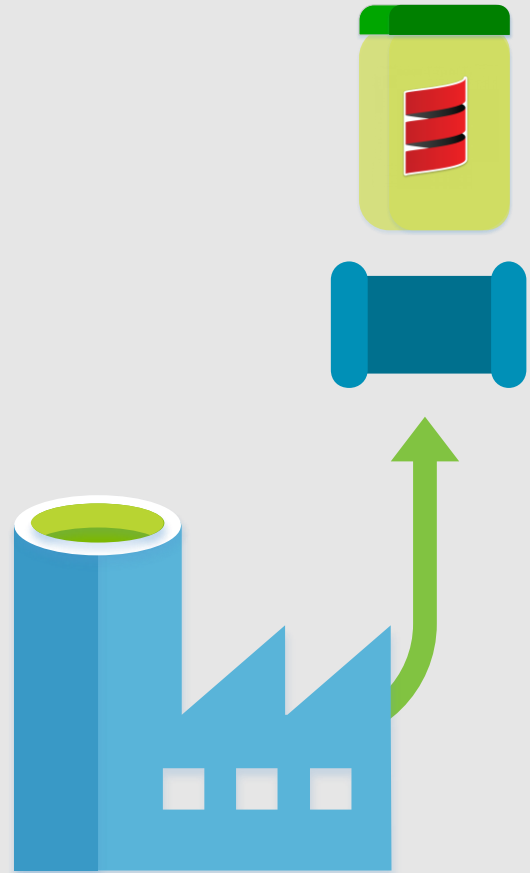
# Mapping Data Flows



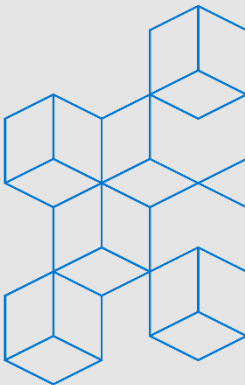
# Mapping Data Flows – Settings & Concepts



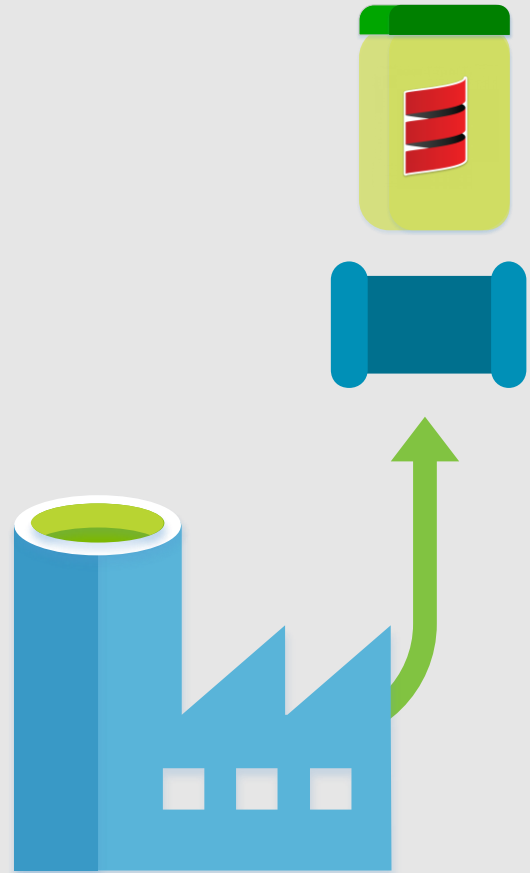
# Mapping Data Flows – Settings & Concepts



Currently Available:



# Mapping Data Flows – Settings & Concepts



source1

Add source dataset

+

Source Settings

Output stream name \* Table1

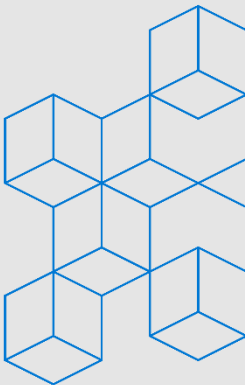
Source dataset \* GenericSQLTable Edit + New

Options

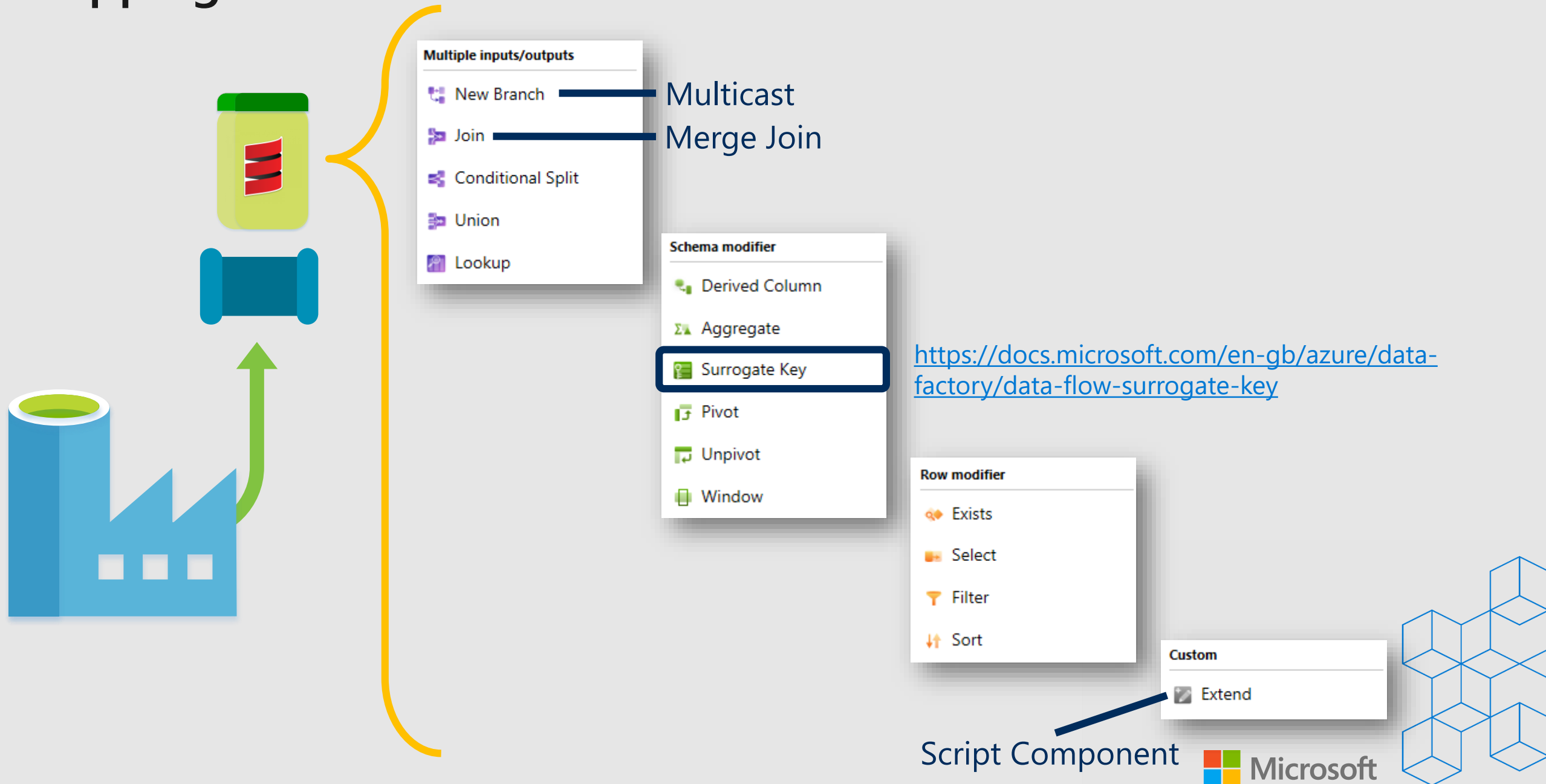
- ☒ Allow schema drift ⓘ
- ☒ Validate schema ⓘ

Sampling \* ☒ Enable ☐ Disable ⓘ

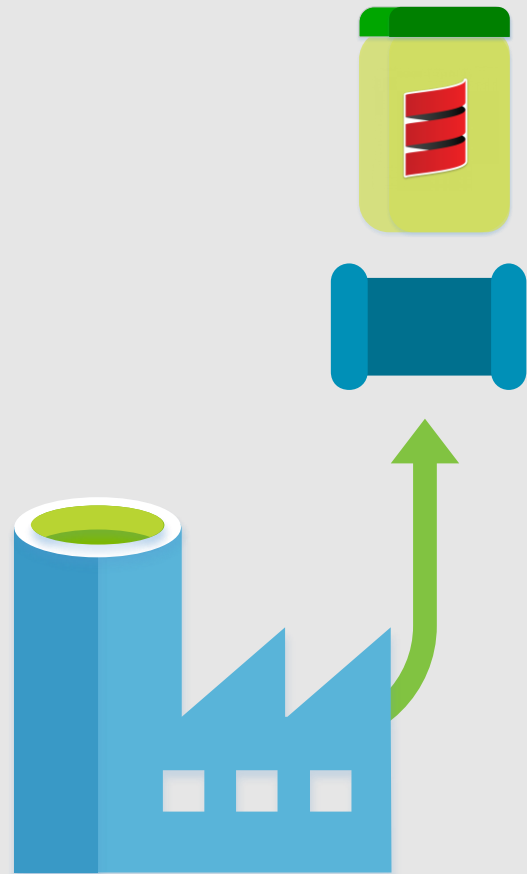
Rows limit 100



# Mapping Data Flows – Transformations



# 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)

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

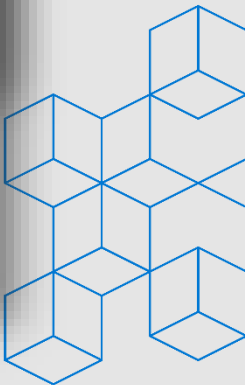
Examples  
1. regexExtract('Cost is between 600 and 800 dollars', '(\\d+) and (\\d+)', 2) -> '800'  
2. regexExtract('Cost is between 600 and 800 dollars', '(\\d+) and (\\d+)', 2) -> '800'

+ - \* / ||

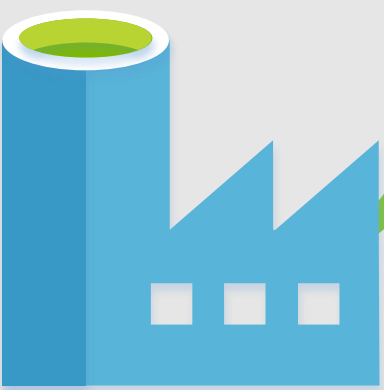
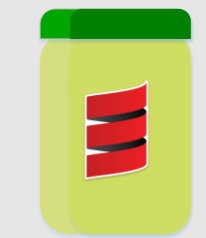
Data preview

⚠ 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

+

Join1

Inner join on OrderHeader and OrderDetails

+

Aggregate1

Aggregating data by 'SalesOrderNumber' producing columns 'OrderLineCount'

+

sink1

Export data to ADWOrderLineCountTable

OrderDetails

Import data from ADWSalesOrderDetail

+

Add Source

Source Settings

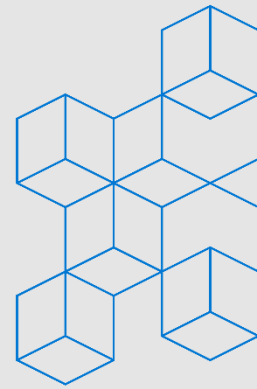
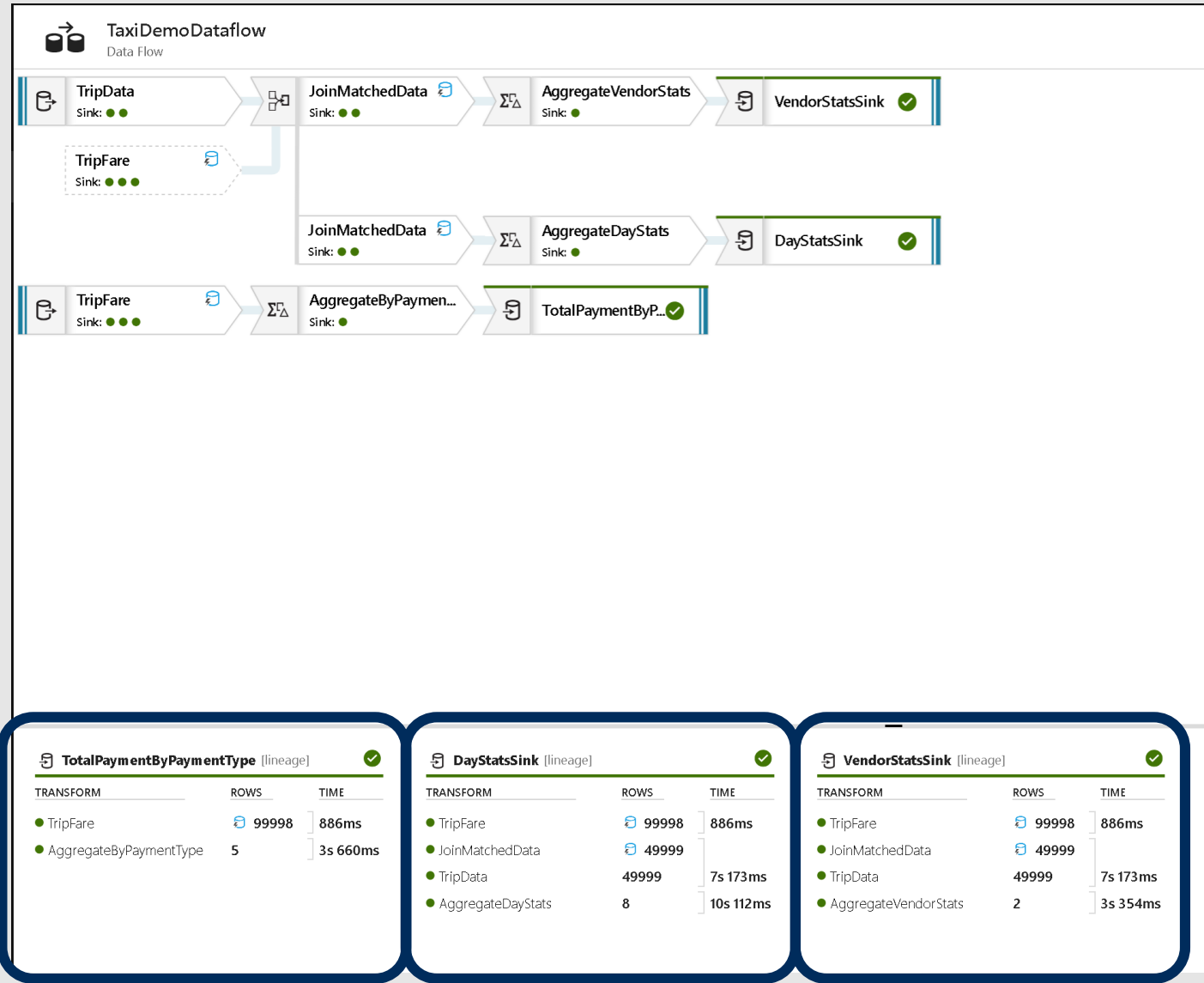
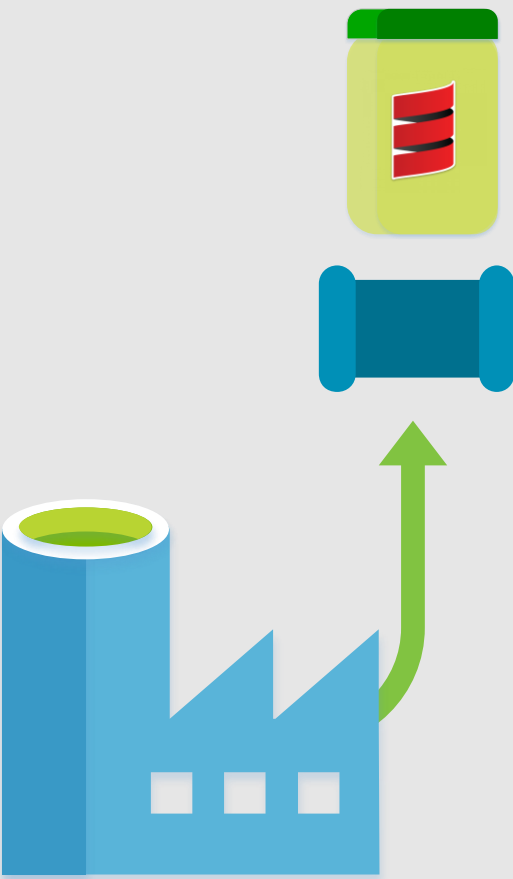
Define schema

Optimize

Inspect

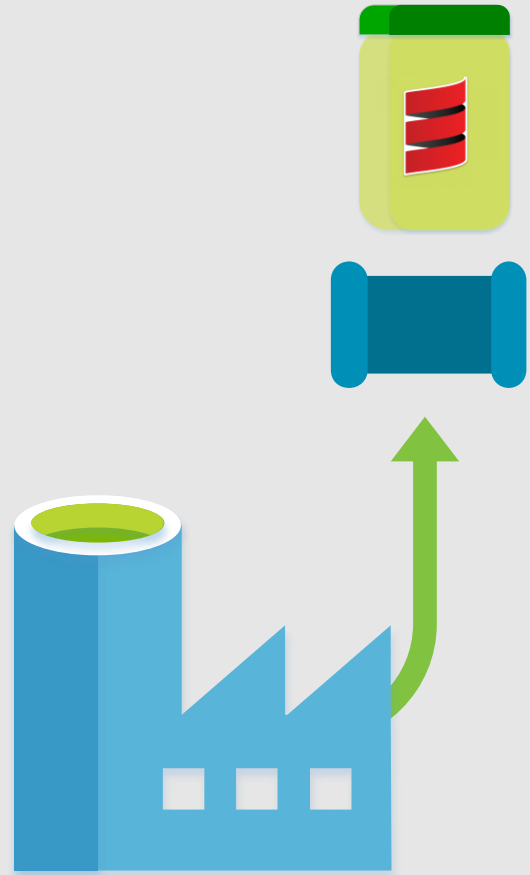
Data Preview

# 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

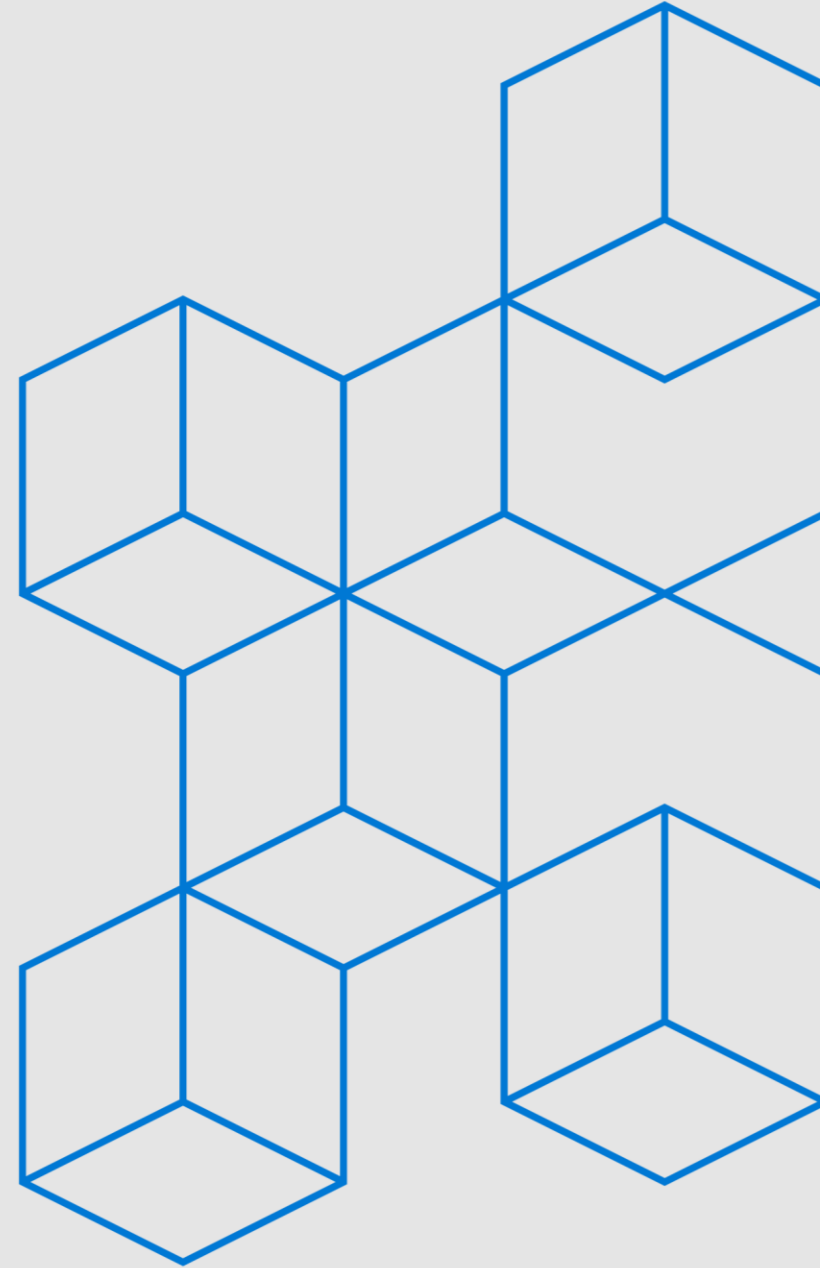
<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-debug-mode>

6





## Monitoring

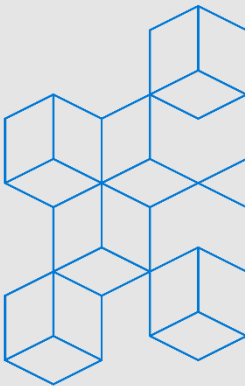
<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-monitoring>

# Demo

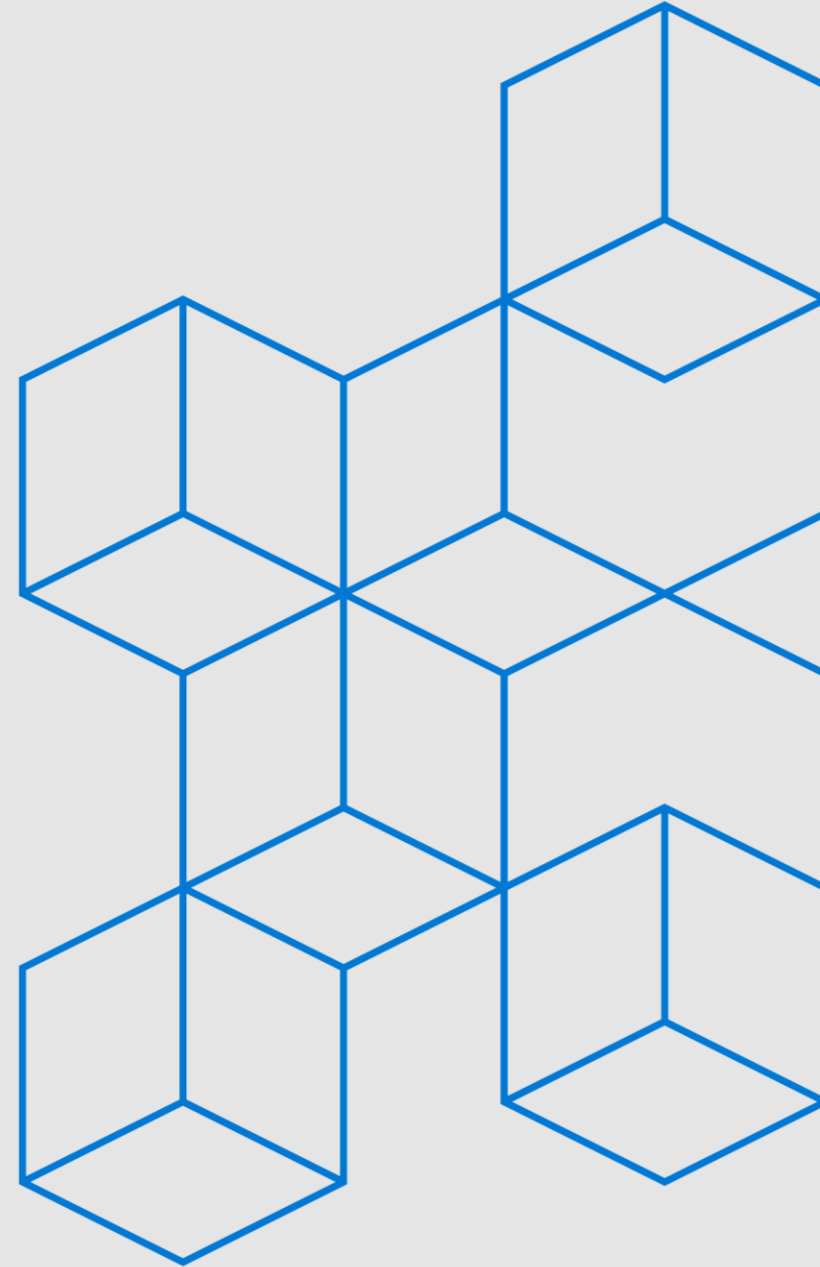


# Demo Summary

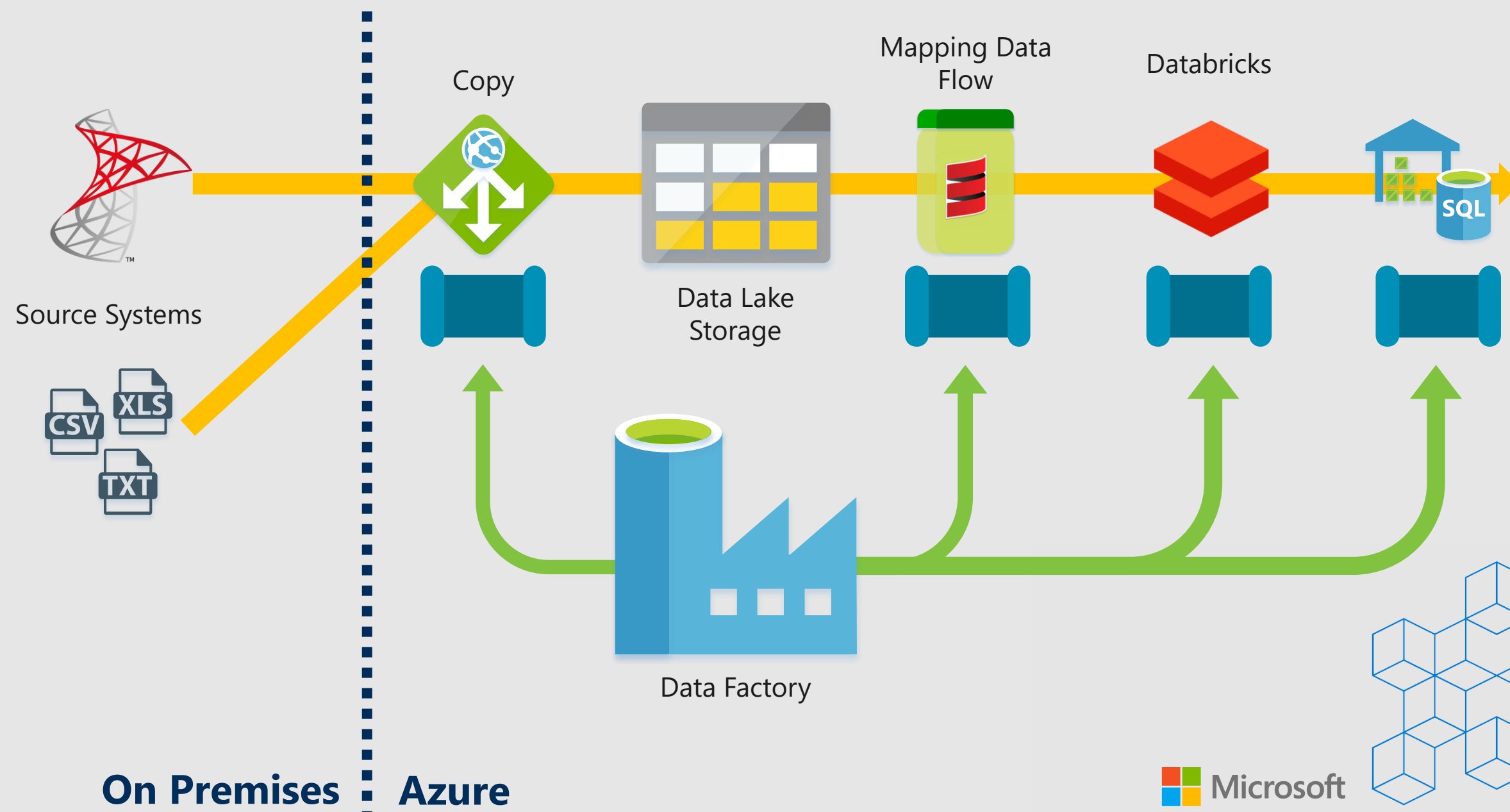
Transformation Method		Graphical Development	Scales Out	Scales Up	Cloud Native Tech
	T-SQL (SQLDB)	✗	✗	✓	✗
	SSIS	✓	✗	✓	✗
	Scala (Databricks)	✗	✓	✓	✓
	Mapping Data Flow	✓	✓	✓	✓



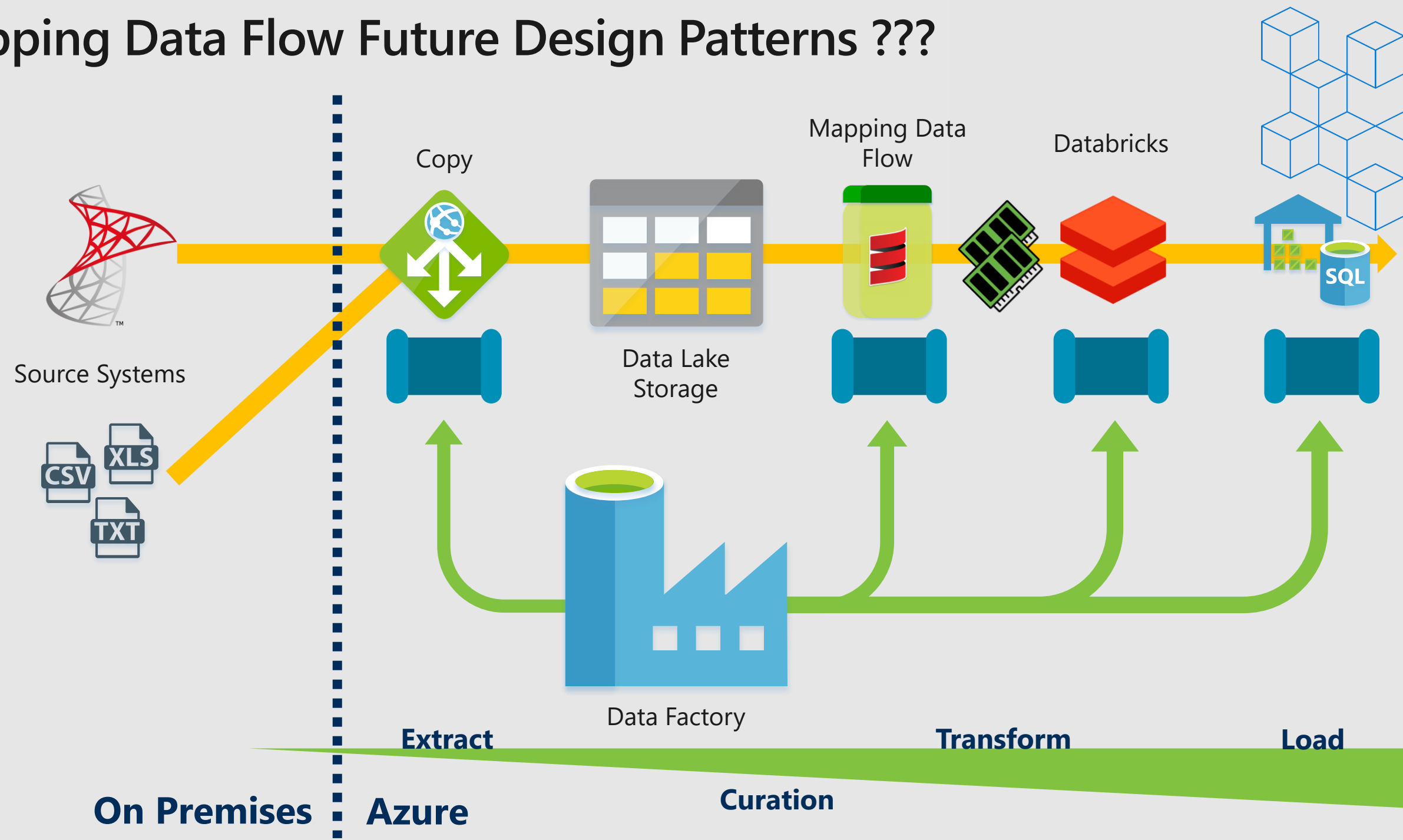
# Design Patterns



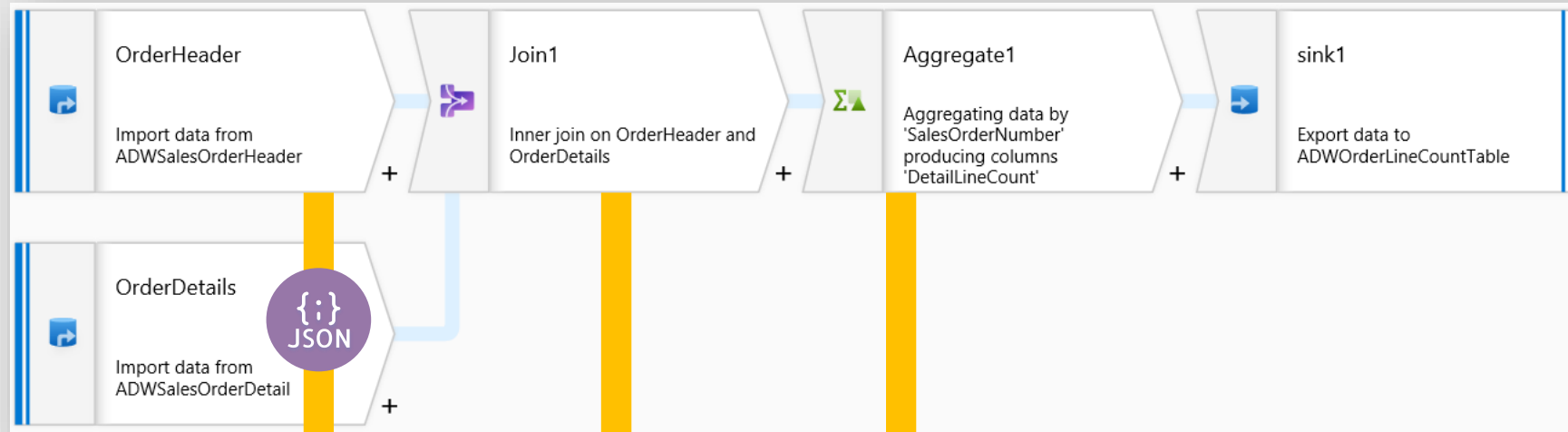
# Mapping Data Flow Future Design Patterns ???



# Mapping Data Flow Future Design Patterns ???



# Mapping Data Flow Future Design Patterns ???

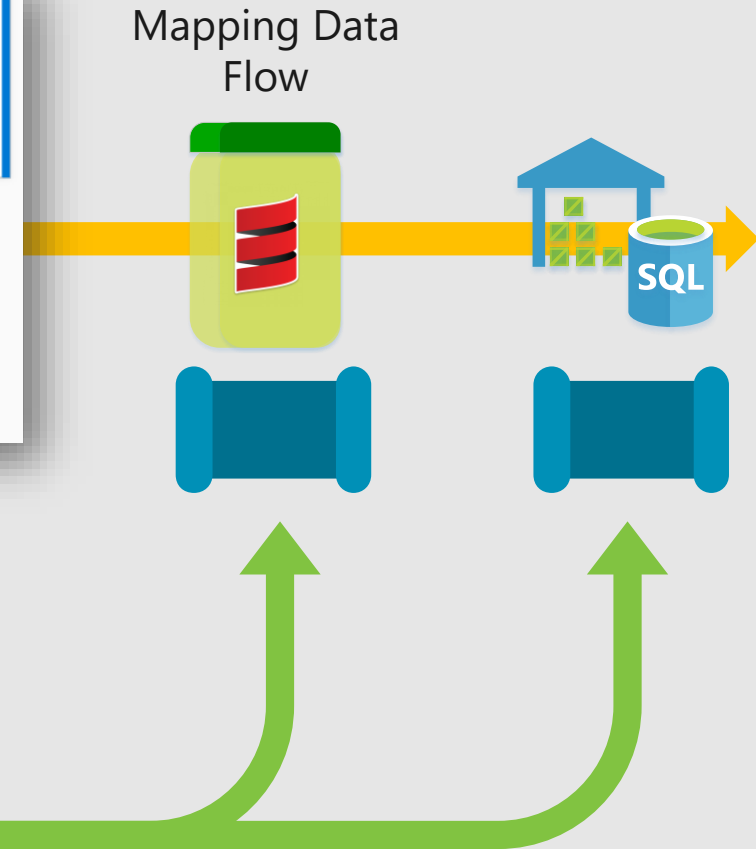


```
"fileName": {
  "value": "@dataset().FileName",
  "type": "Expression"
},
"folderPath": {
  "value": "@dataset().SourceDIR",
  "type": "Expression"
}
```

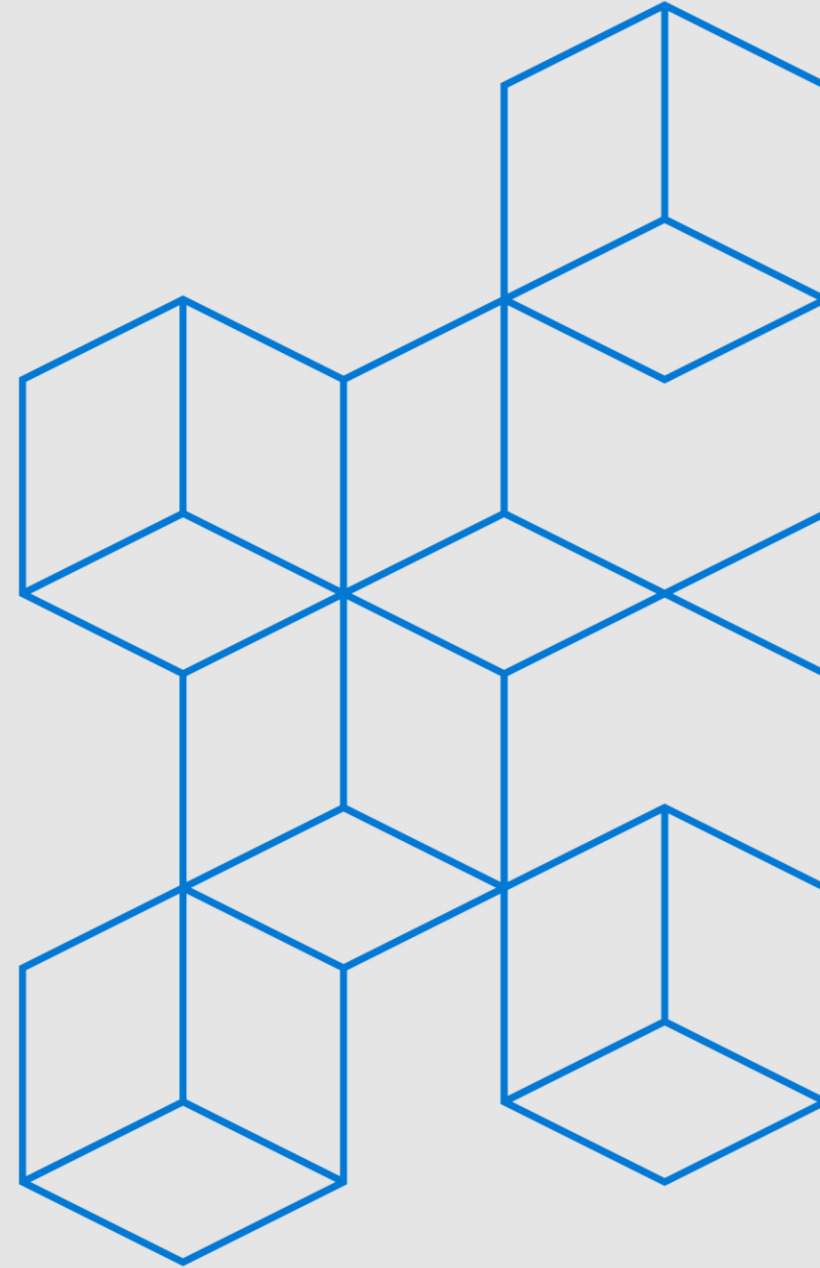
```

"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"
  }
]

```

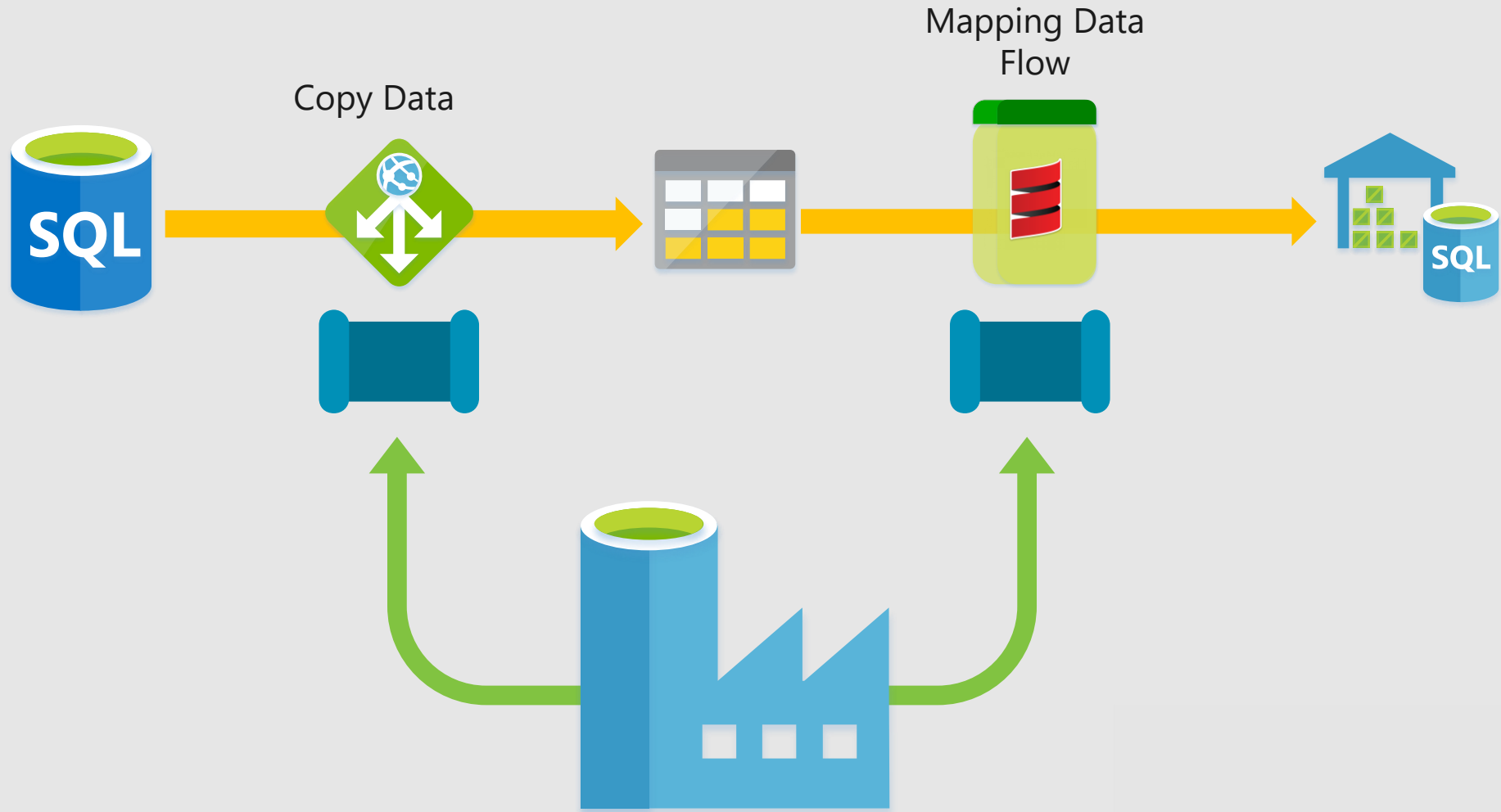


# Conclusion




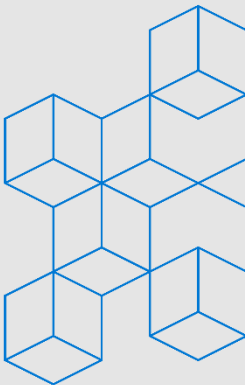


# What is Azure Data Factory?



Orchestrator of our solution Control Flow operations.  
Orchestrator of our solution Data Flow transformations.

... using cloud native technology in  zure and now with an easy developer interface for both.



# Thanks for Listening

## Paul Andrew



@MrPaulAndrew



**Email:** paul@mrpaulandrew.com

**Blog:** mrpaulandrew.com

**GitHub:** github.com/mrpaulandrew

