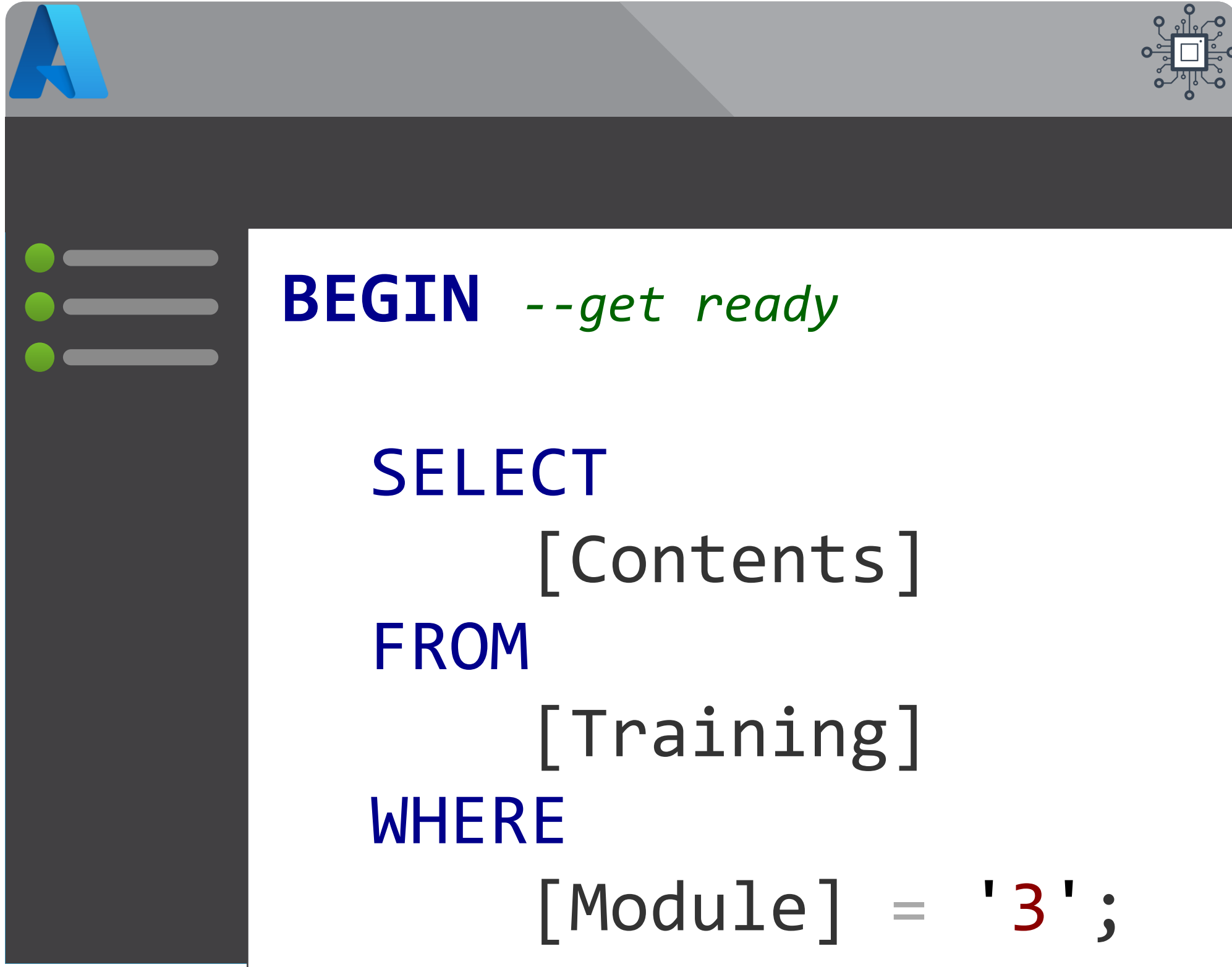


Module 3

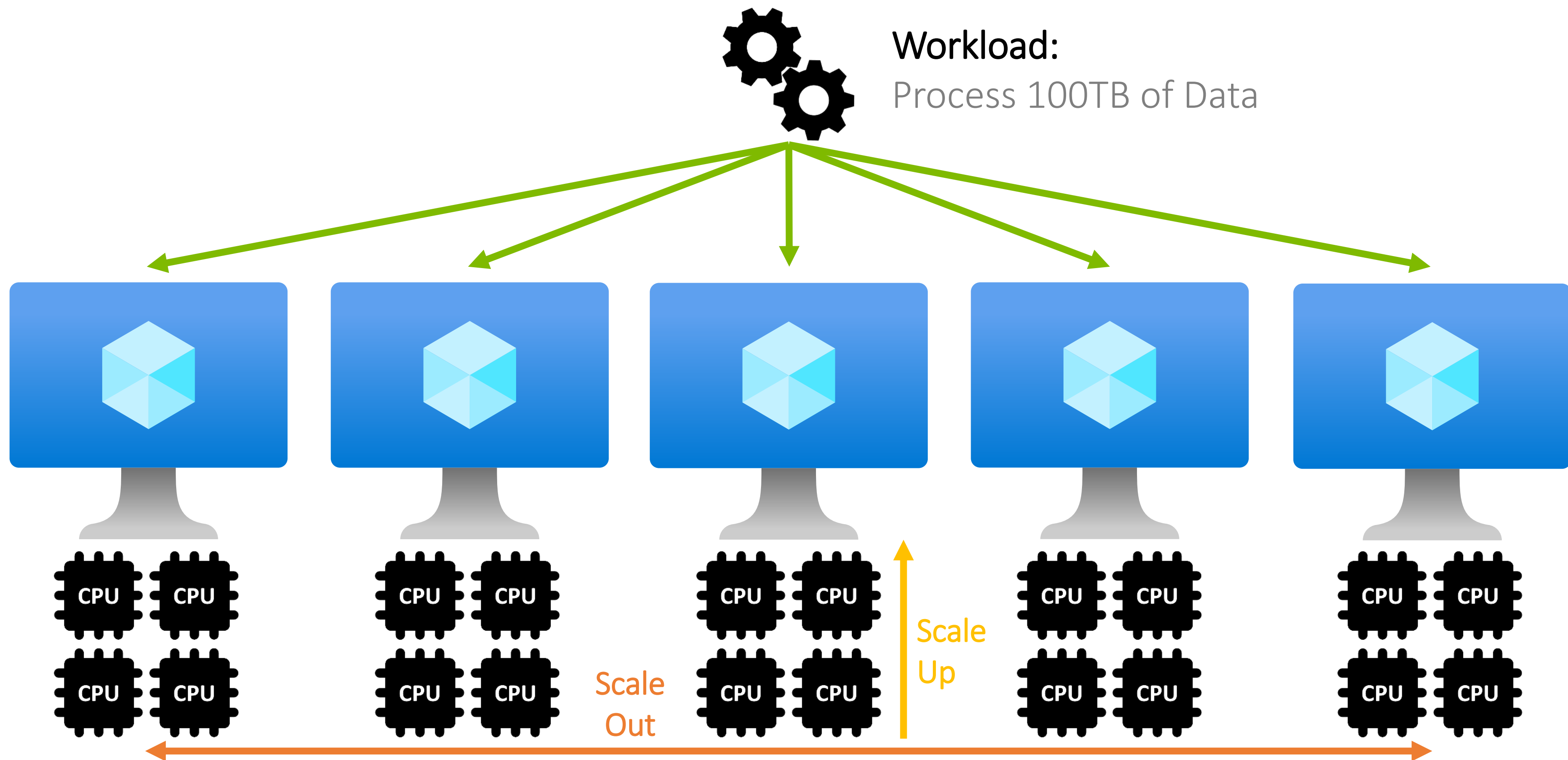
Data Transformation



- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

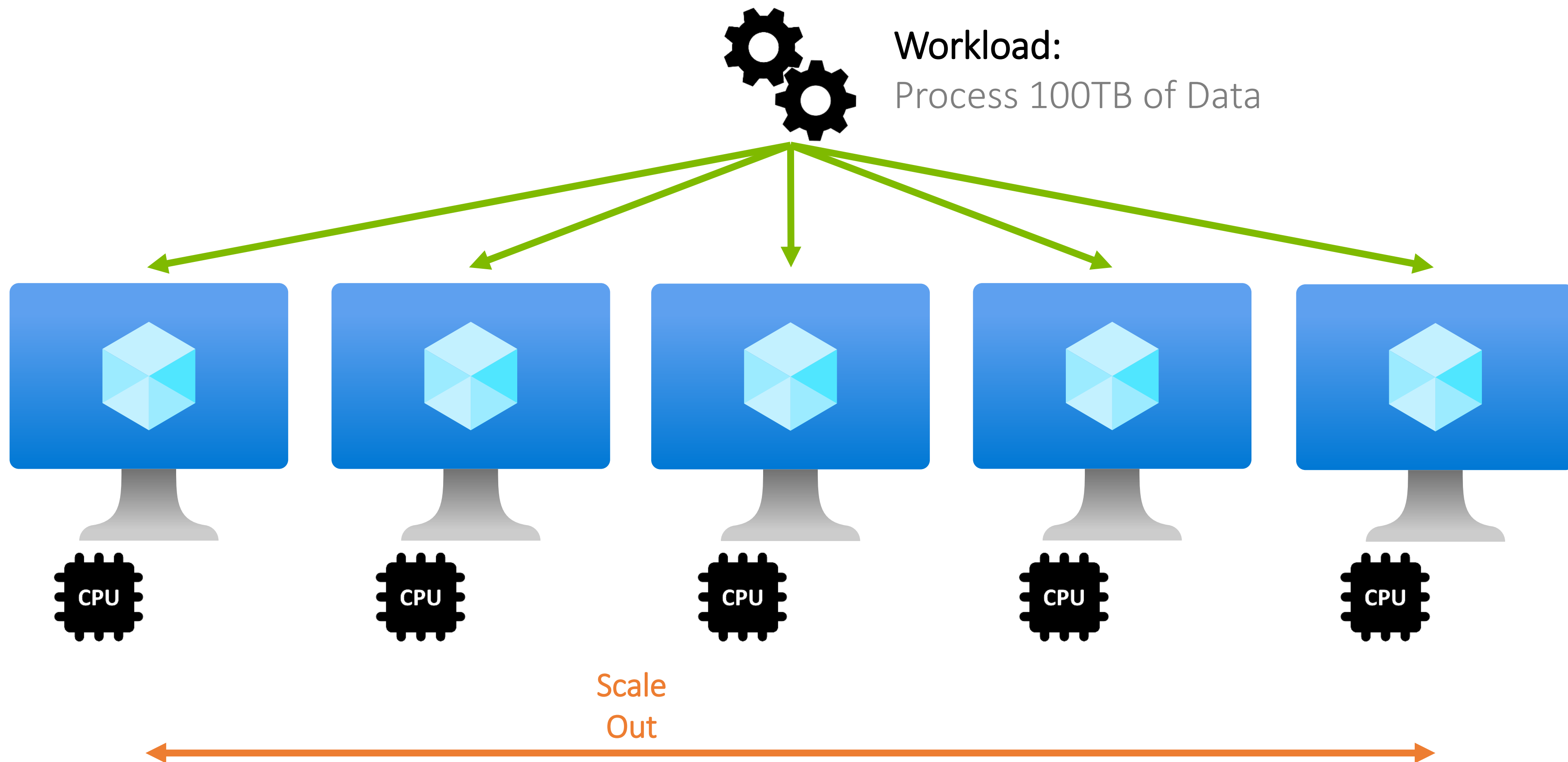


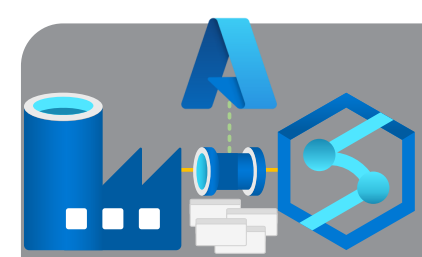
Scaling Up and/or Scaling Out



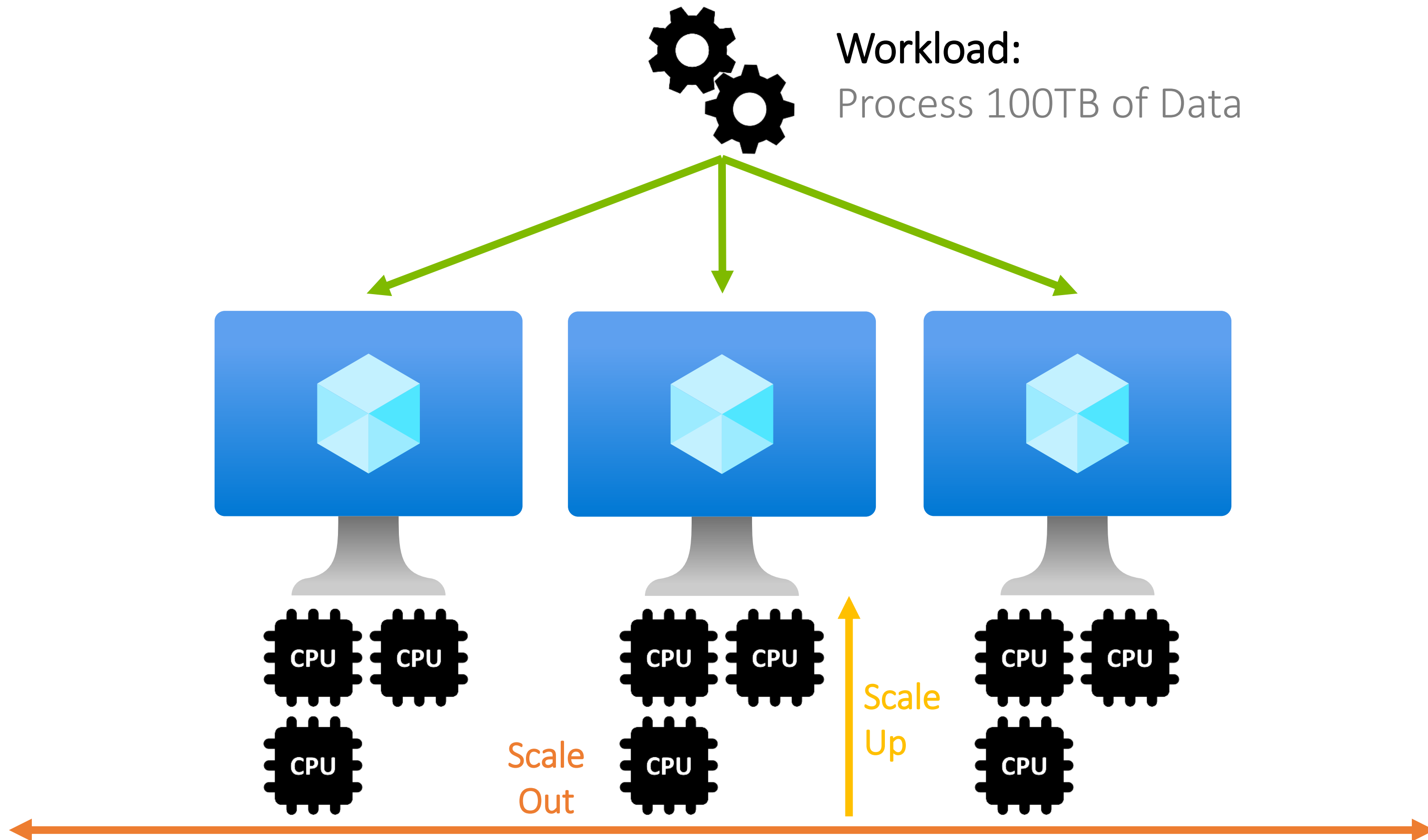
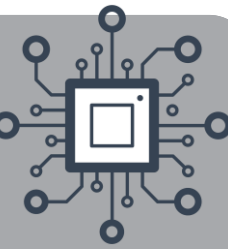


Scaling Up and/or Scaling Out



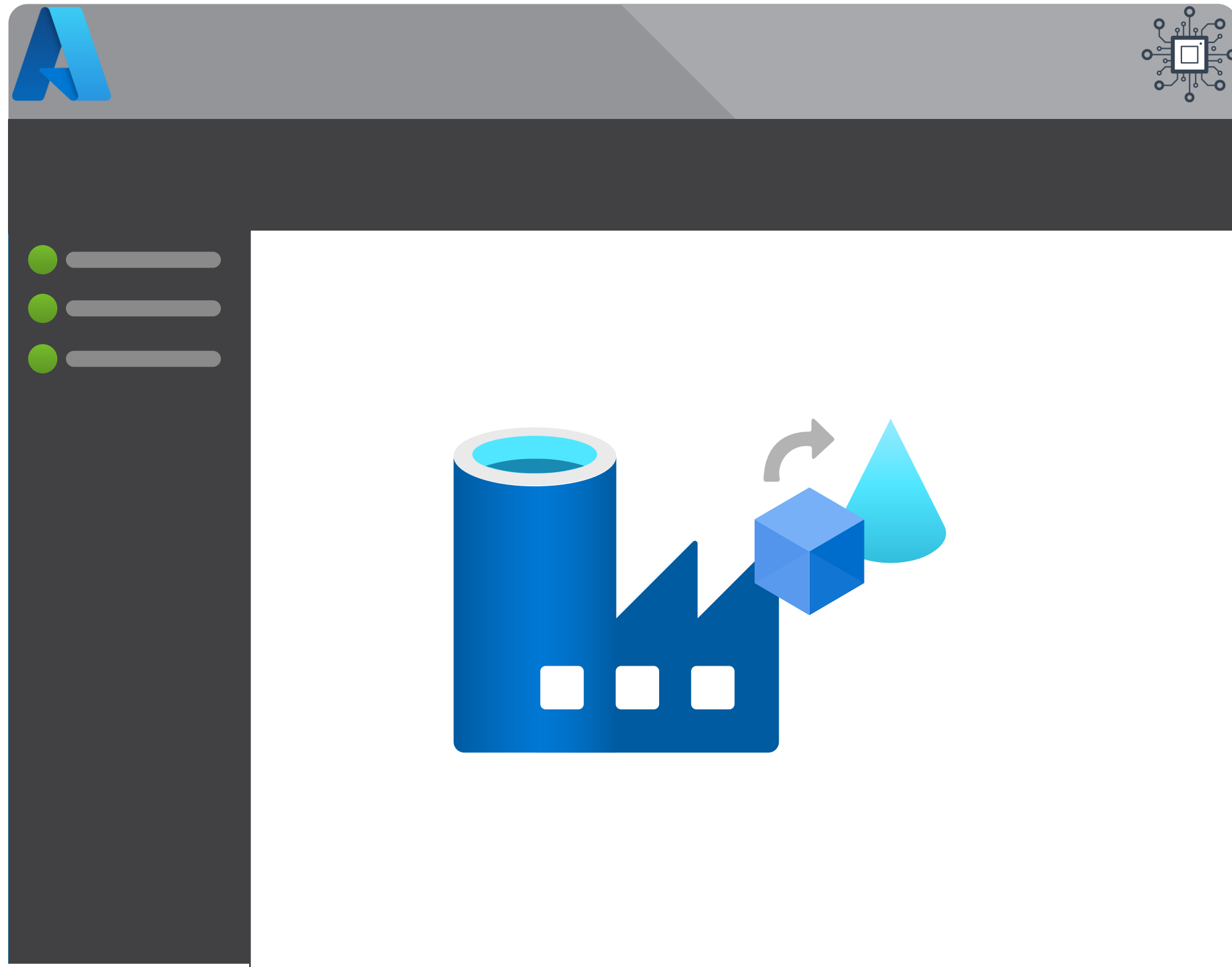


Scaling Up and/or Scaling Out



Module 3

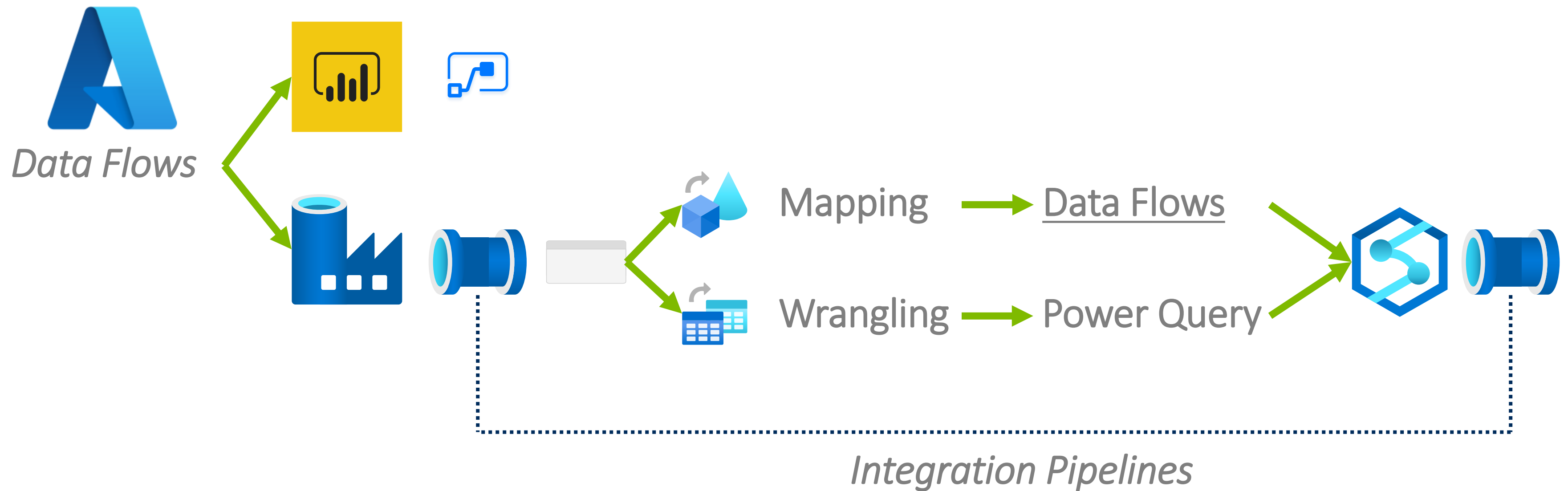
Data Transformation

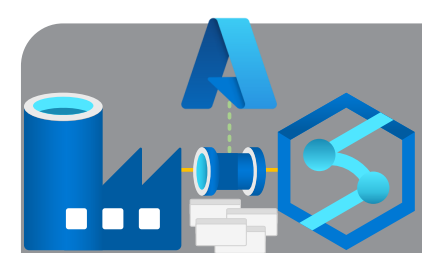


- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

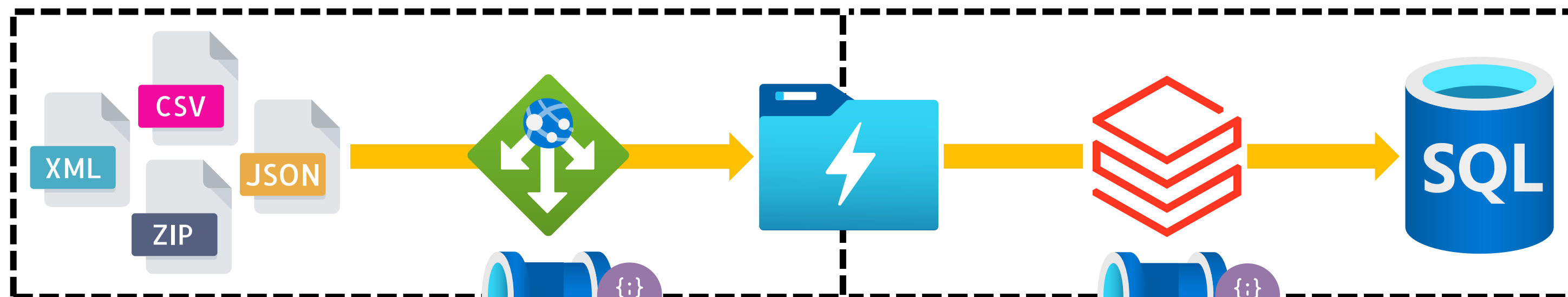
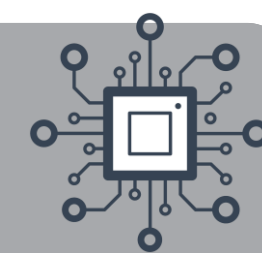


Terminology Clarification





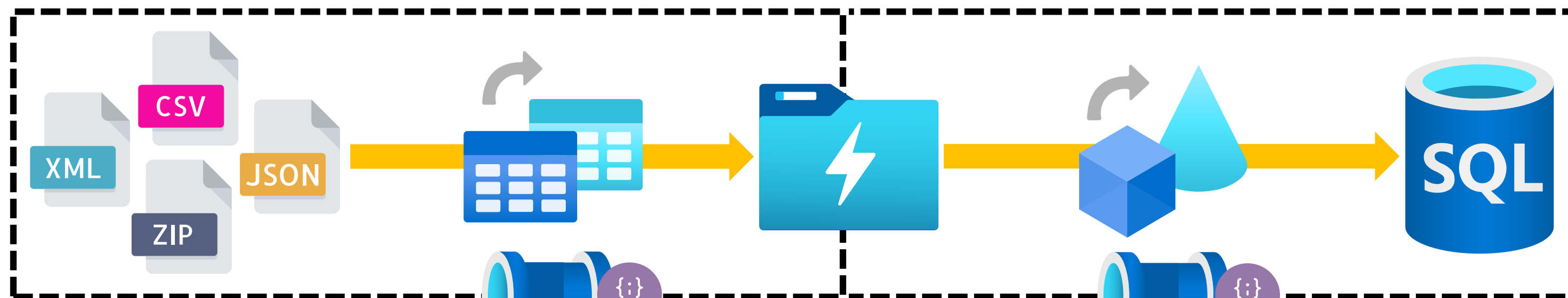
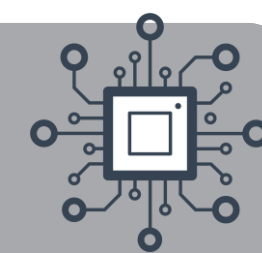
Control Flow Components



- 1 Linked Services
- 2 Datasets
- 3 Activities
- 4 Pipelines
- 5 Triggers



Control Data Flow Components

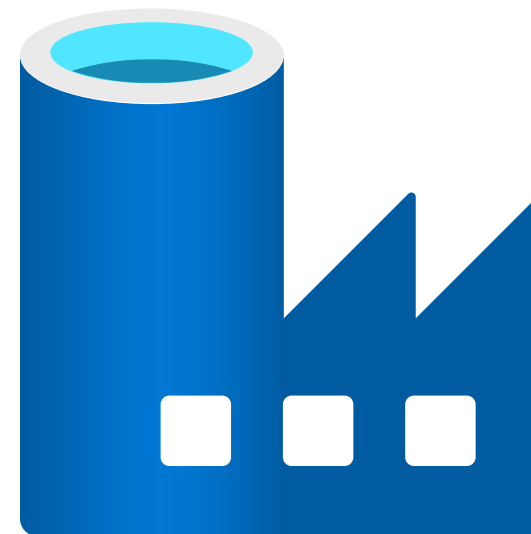
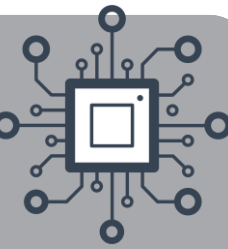


- 1 Linked Services
- 2 Datasets
- 3 Activities
- 4 Pipelines
- 5 Triggers



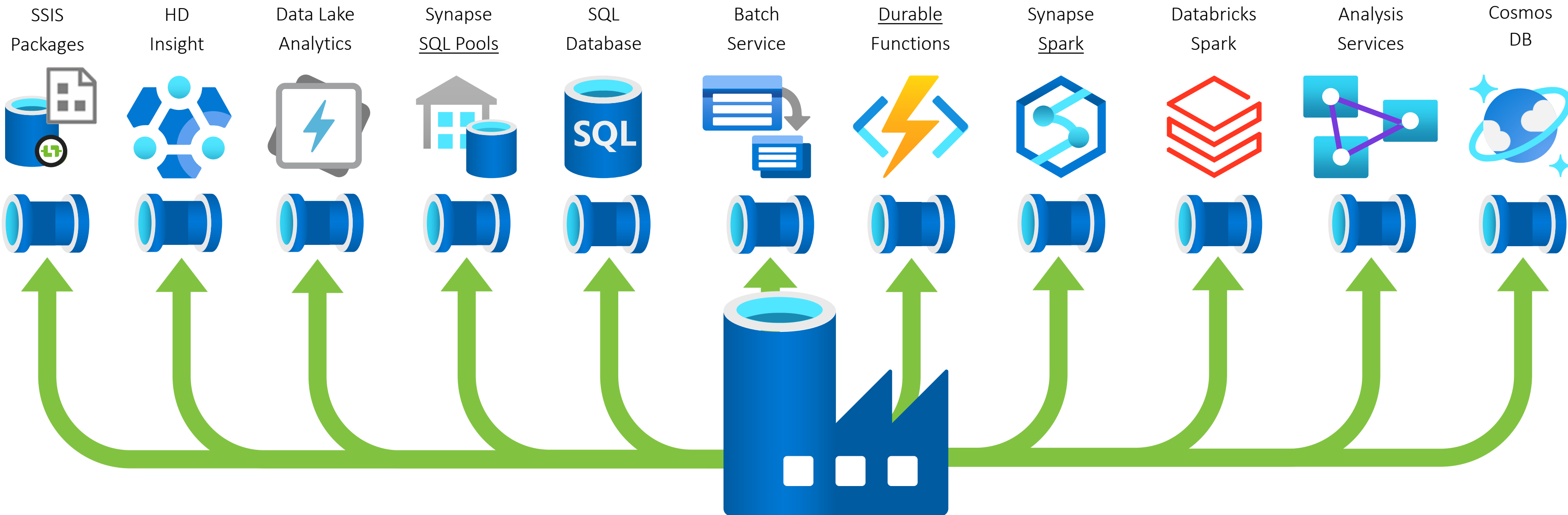
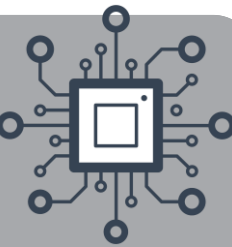


Data Transformation in Azure



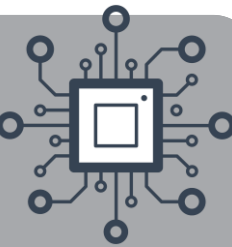


Other Data Transformation Services in Azure

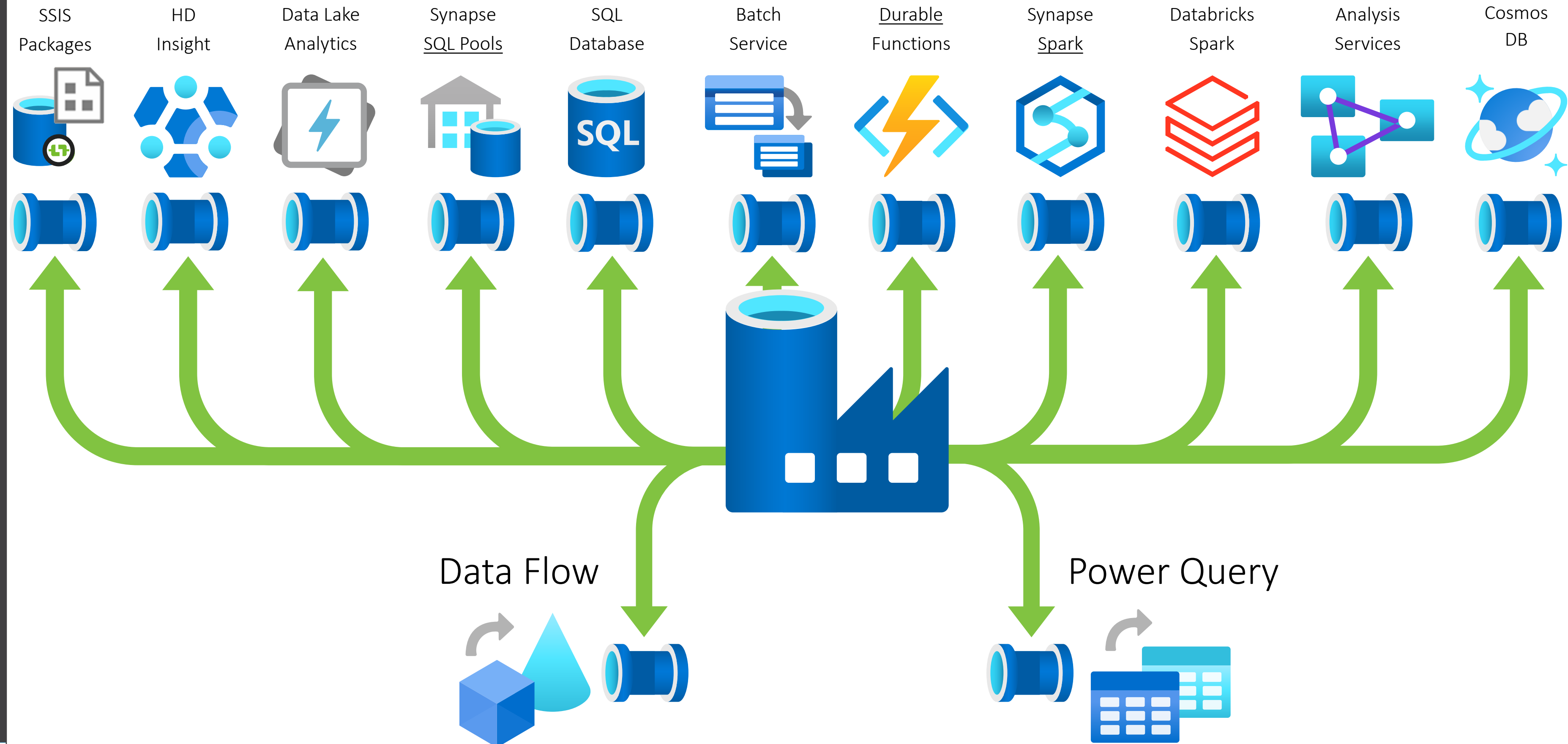


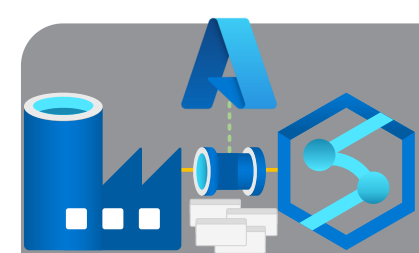


Other Data Transformation Services in Azure



When Should We Use These Integration Pipeline Transformation Activities?





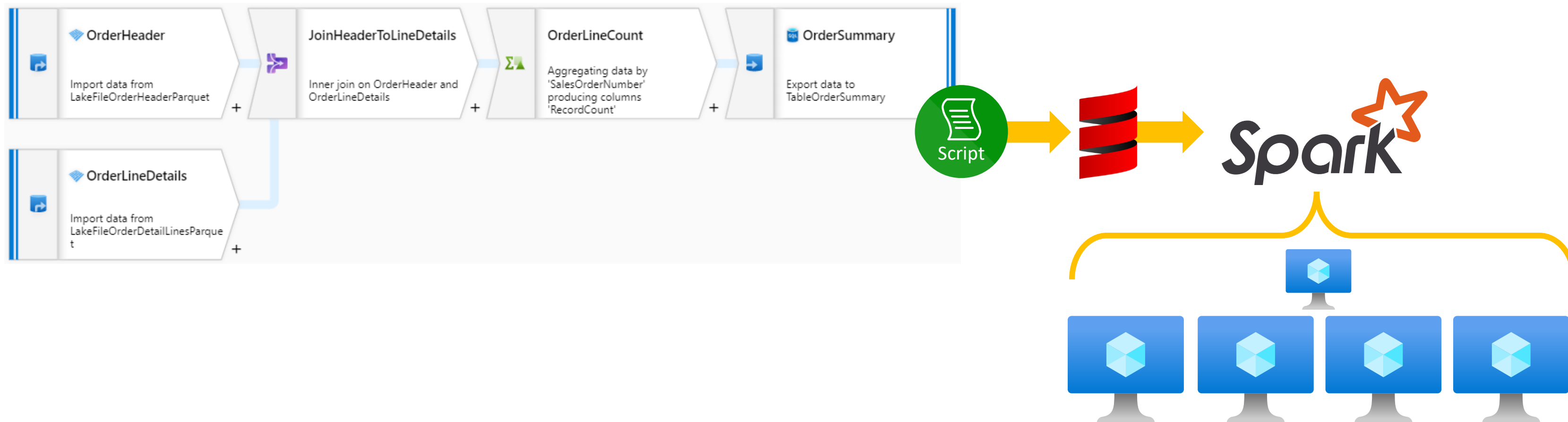
What is a Mapping Data Flow?

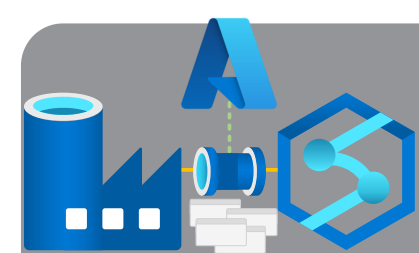


Control Flow

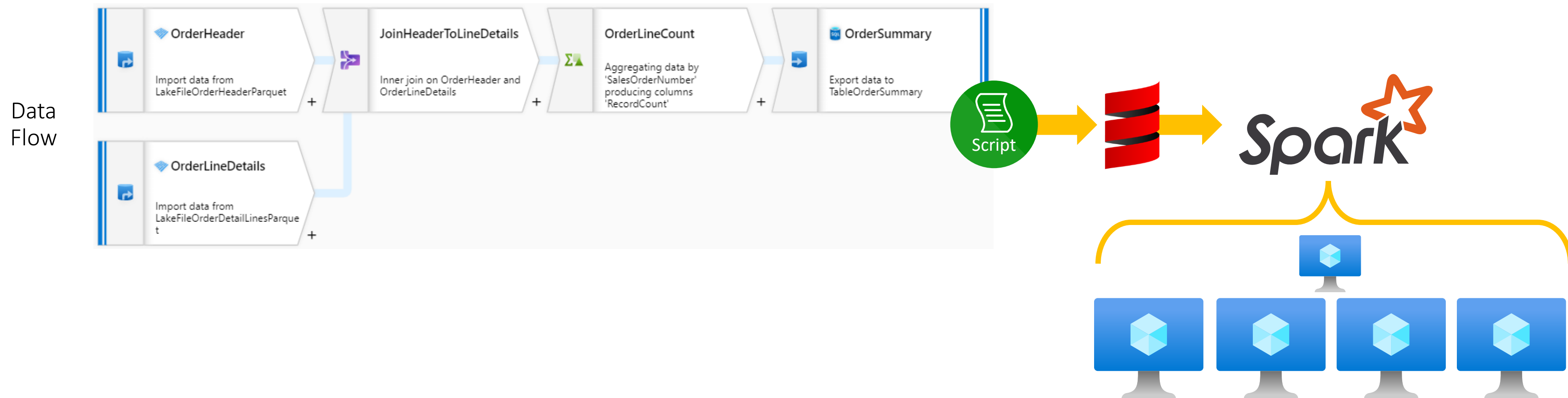


Data Flow

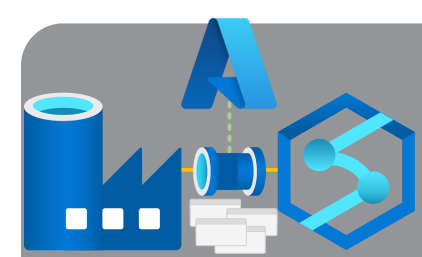




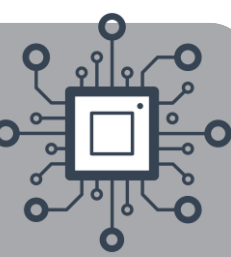
Q: What is a Mapping Data Flow?



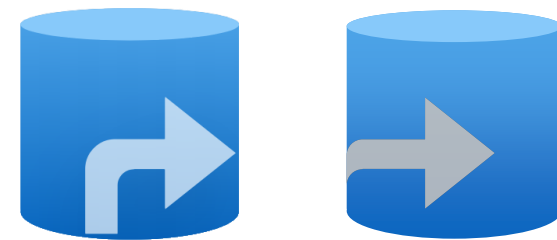
A: Graphic no low/low code data transformation tool that sits on top of Apache Spark.



Data Flows – Inputs & Outputs



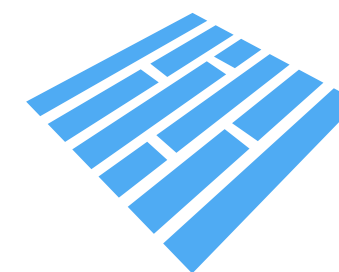
Source & Sink



Linked Services

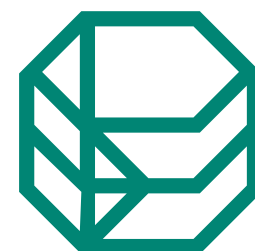


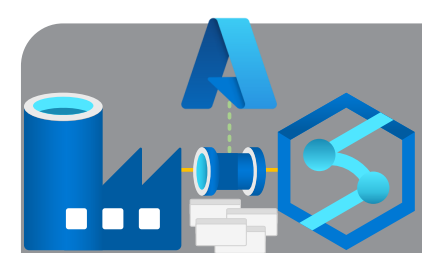
Dataset



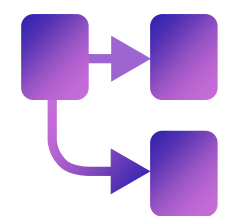
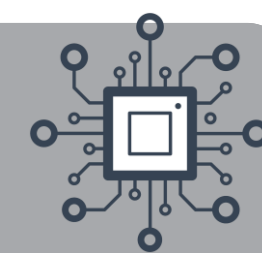
Source
Types

Inline

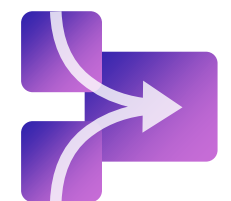




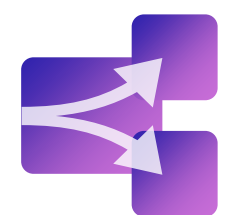
Data Flows – Transformations



New Branch



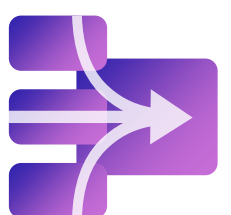
Join



Conditional Split



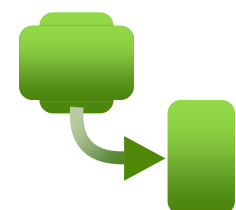
Exists



Union



Lookup



Derived Column



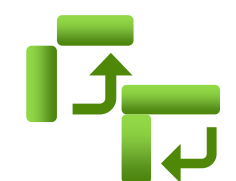
Select



Aggregate



Surrogate Key



Pivot/Unpivot



Window



Rank



External Call



Cast



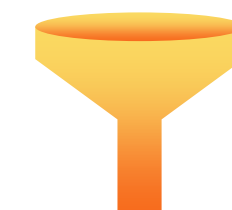
Flatten



Parse



Stringify



Filter



Sort



Alter Row



Assert



Flowlet

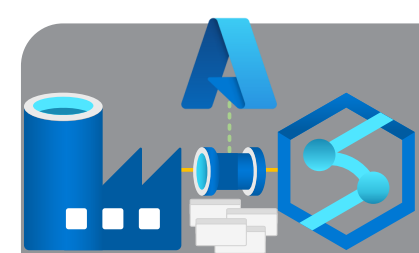
Key

Input & Output Modifiers

Schema Modifiers

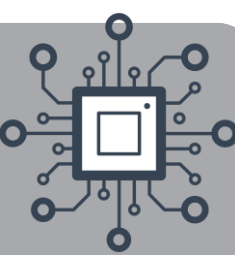
Formatters

Row Modifiers



















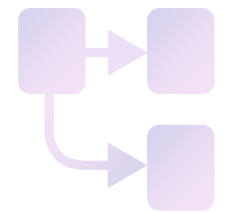
Data Flows – Transformations

<https://sqlplayer.net/2018/12/azure-data-factory-v2-and-its-available-components-in-data-flows/>



Components

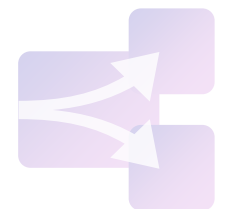
Operation / Activity	Description	SSIS equivalent	SQL Server equivalent
 New branch	Create a new flow branch with the same data	 Multicast (+icon)	<pre>1 SELECT INTO 2 SELECT OUTPUT</pre>
 Join	Join data from two streams based on a condition	 Merge join	<pre>1 INNER/LEFT/RIGHT JOIN, 2 CROSS/FULL OUTER JOIN</pre>
 Conditional Split	Route data into different streams based on conditions	 Conditional Split	<pre>SELECT INTO WHERE condition1 SELECT INTO WHERE condition2 CASE ... WHEN</pre>
 Union	Collect data from multiple streams	 Union All	<pre>SELECT colla UNION (ALL) SELECT collb</pre>
 Lookup	Lookup additional data from another stream	 Lookup	<i>Subselect, function,</i> <pre>LEFT/RIGHT JOIN</pre>
 Derived Column	Compute new columns based on the existing once	 Derived Column	<pre>SELECT Column1 * 1.09 as NewColumn</pre>
 Aggregate	Calculate aggregation on the stream	 Aggregate	<pre>SELECT Year(DateOfBirth) as YearOnly, MIN(), MAX(), AVG() GROUP BY Year(DateOfBirth)</pre>
 Surrogate Key	Add a surrogate key column to output stream from a specific value	 Script Component	<pre>SELECT ROW_NUMBER() OVER(ORDER BY name ASC) AS Row#, name FROM sys.databases</pre>



New Branch



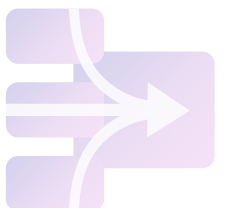
Join



Conditional Split



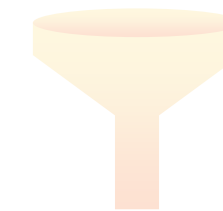
Exists



Union



Lookup



Filter



Sort



Alter Row

Key

Input & Output Modifiers

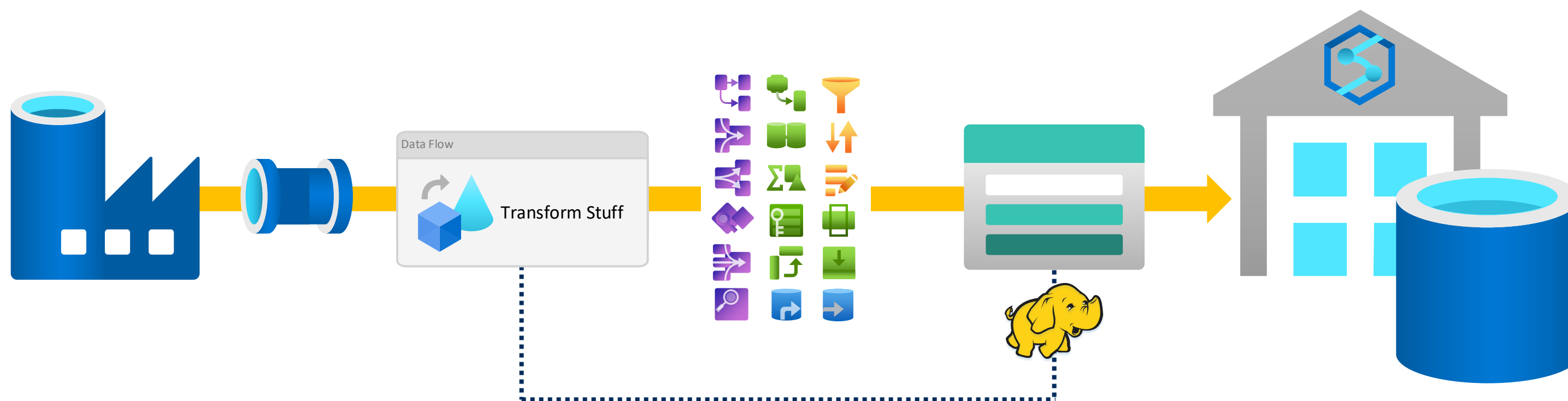
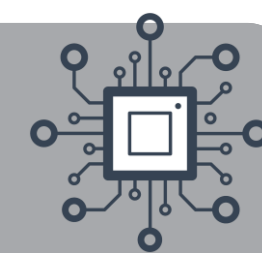
Schema Modifiers

Formatters

Row Modifiers



Data Flows – Data Warehouse Loading (PolyBase)



Staging

▲ PolyBase ⓘ

Staging linked service

Select...



+ New

Staging storage folder

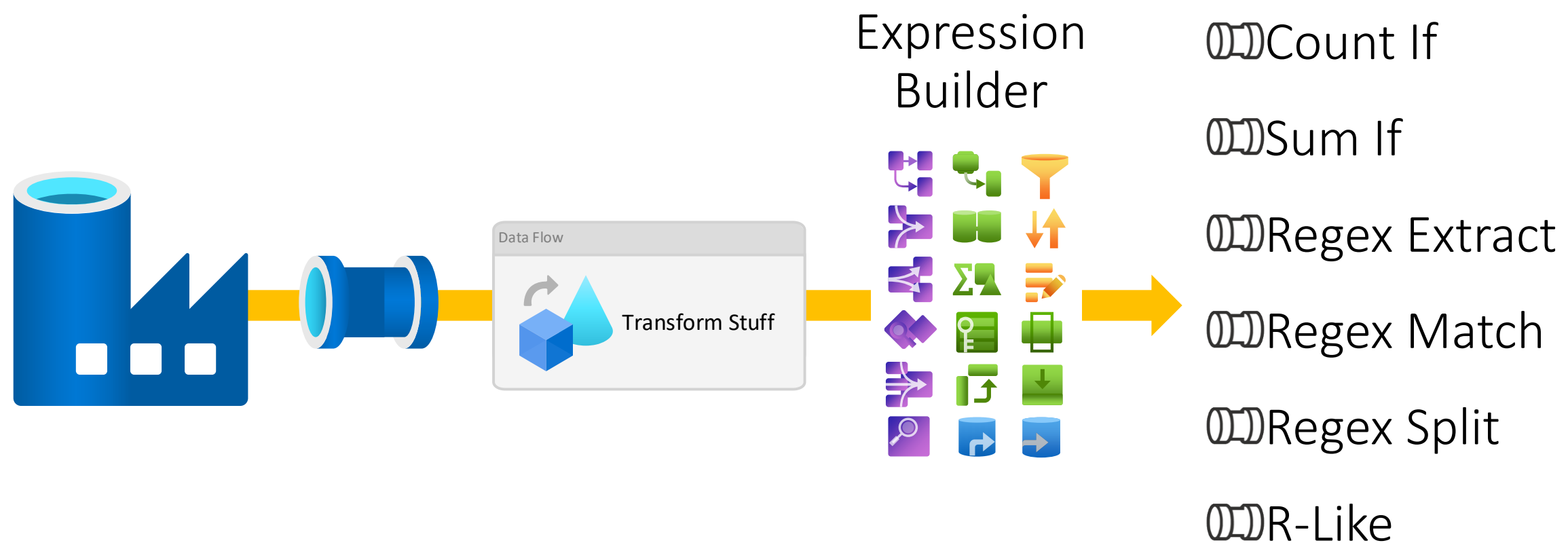
Container

/ Directory

📁 Browse | ▾

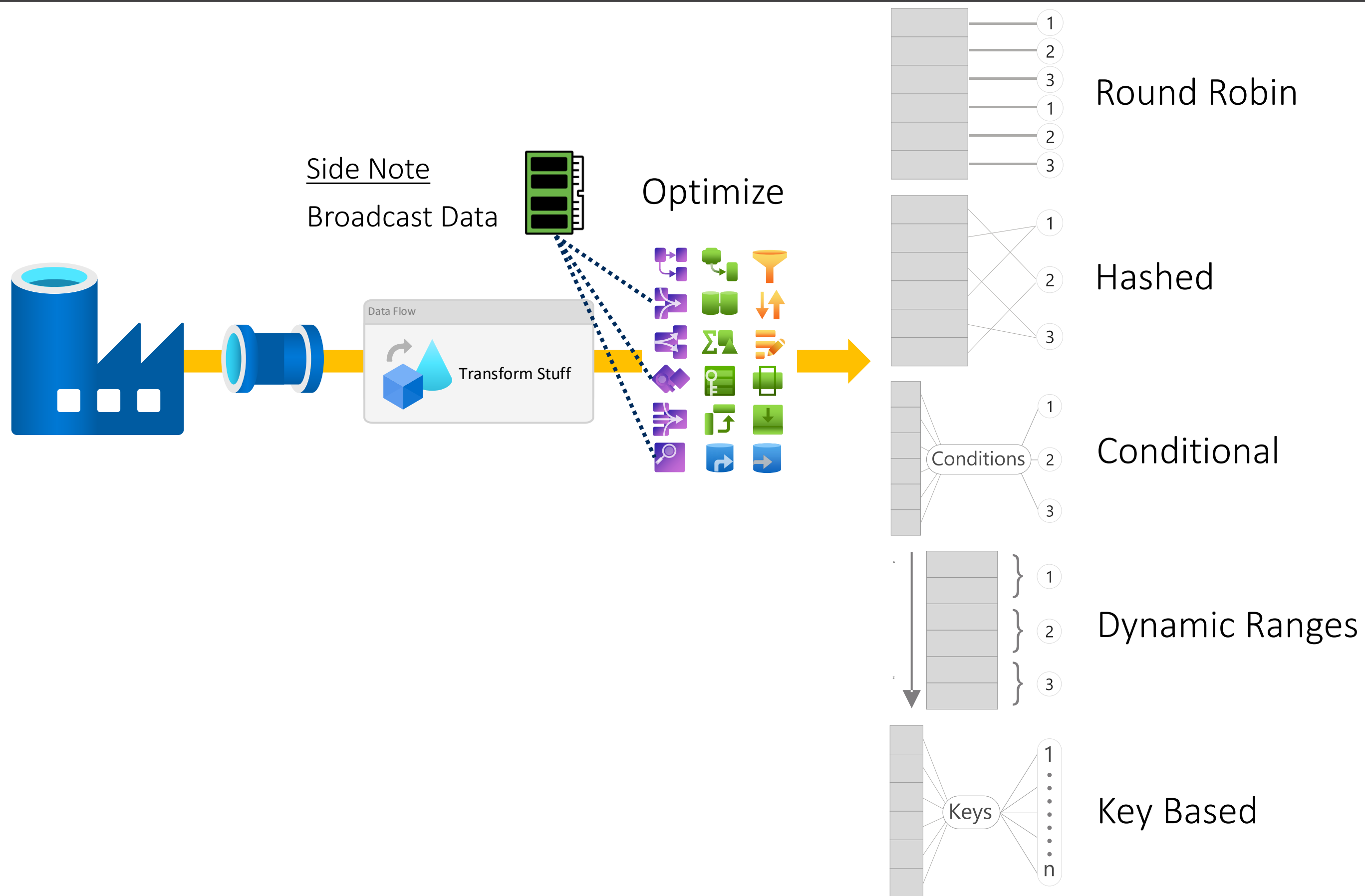


Data Flows – Expression Builder



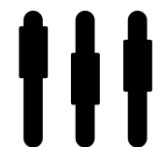
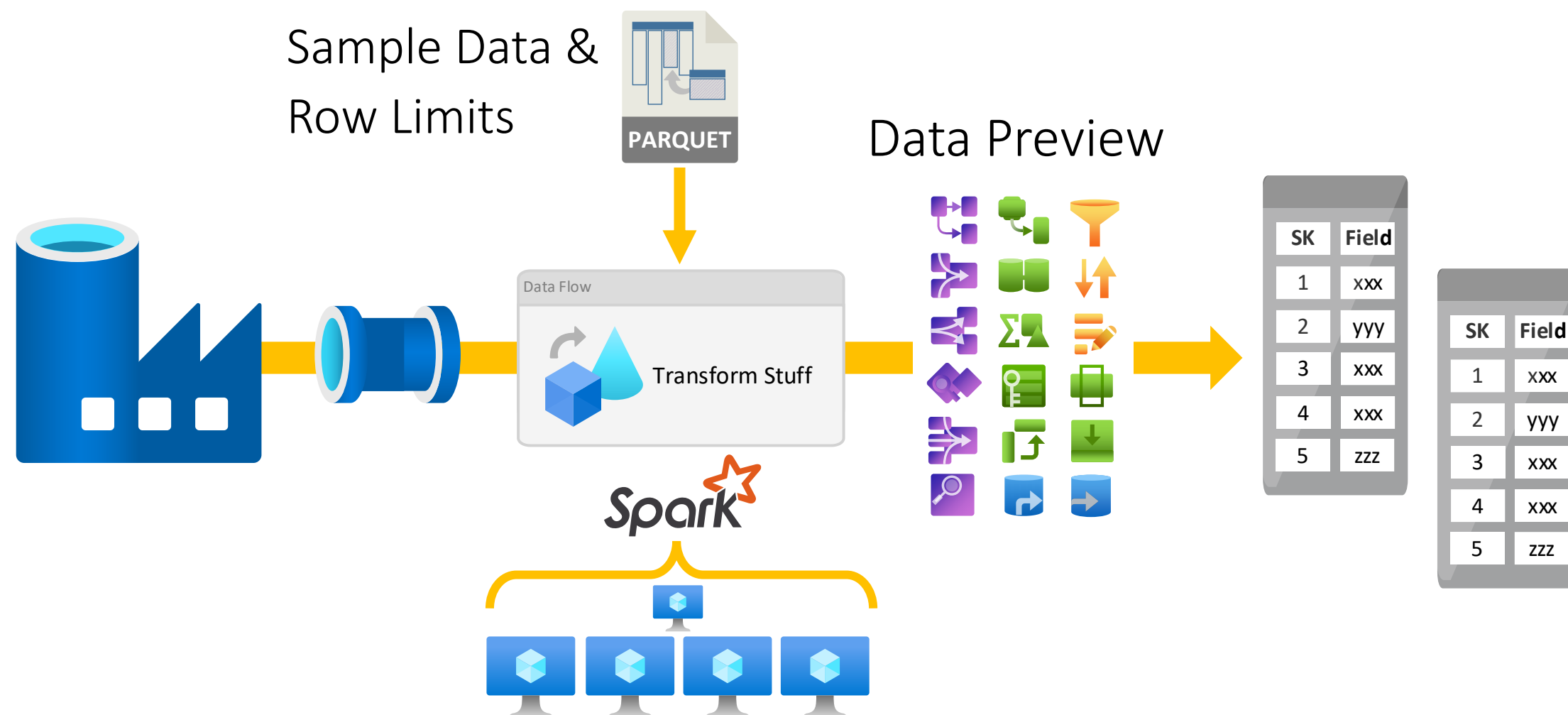


Data Flows – Data Distribution





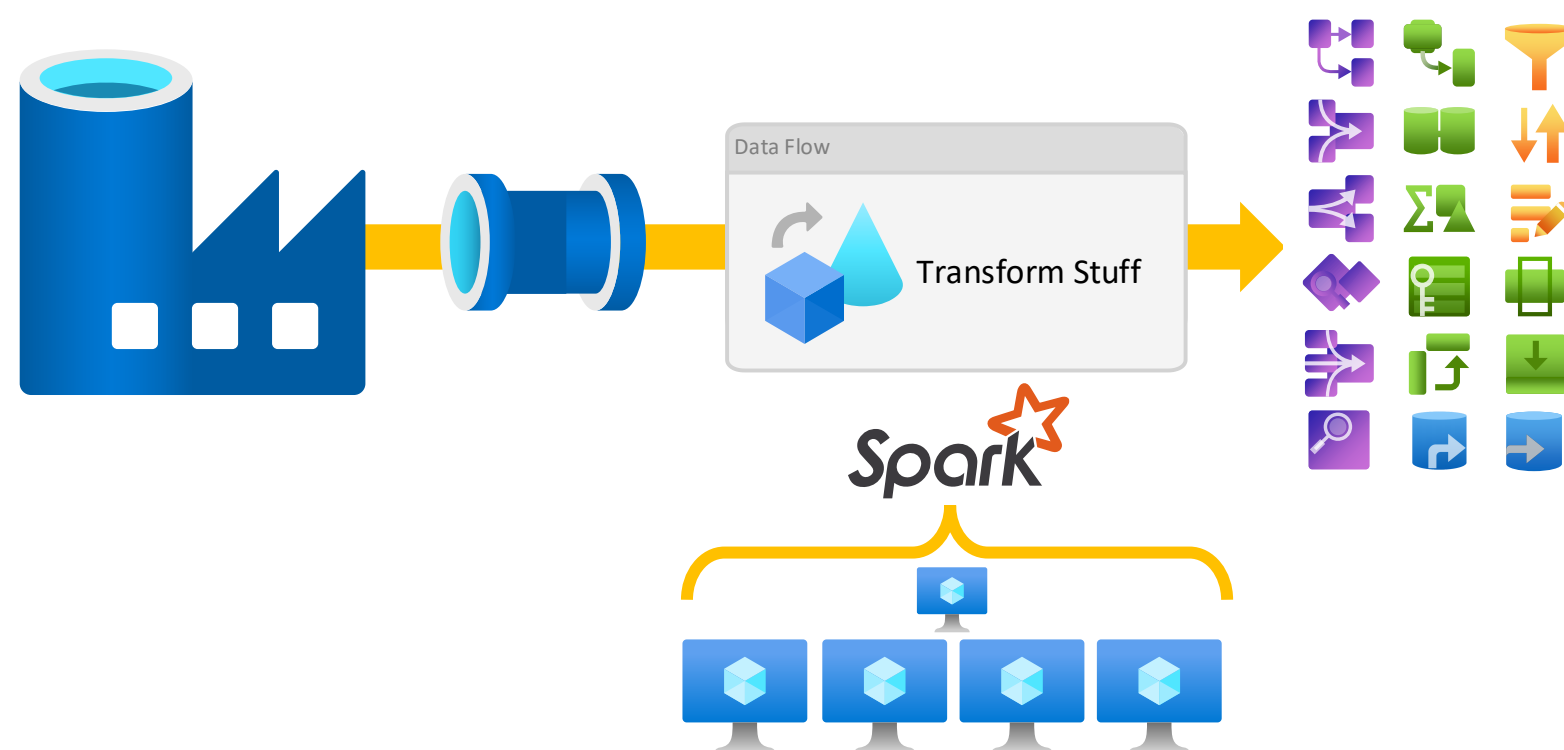
Data Flows – Debugging

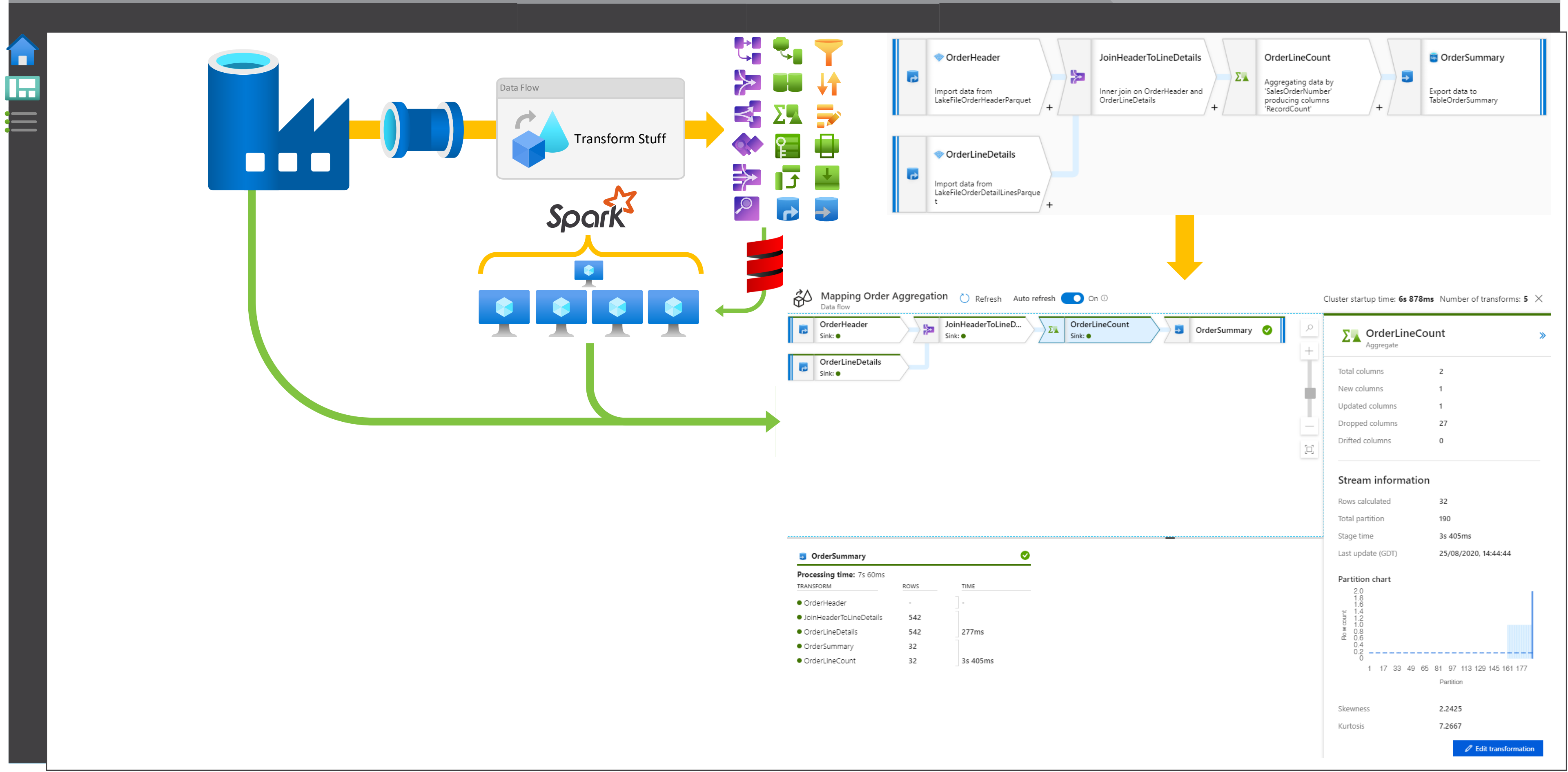


Enable Data Flow Debug Mode



Data Flows – Monitoring





The dashboard illustrates a data flow pipeline for monitoring. It starts with a data source (factory icon) feeding into a 'Transform Stuff' box, which then connects to a 'Spark' cluster (represented by four server icons). The data flow continues through a series of transformations and sinks, including 'OrderHeader', 'JoinHeaderToLineDetails', 'OrderLineCount', and 'OrderSummary'. A detailed view of the 'OrderLineCount' transformation is shown on the right, including its configuration, stream information, and a partition chart.

OrderLineCount Configuration

Step	Operation
1	Import data from LakeFileOrderHeaderParquet
2	Inner join on OrderHeader and OrderLineDetails
3	Aggregating data by 'SalesOrderNumber' producing columns 'RecordCount'
4	Export data to TableOrderSummary

Mapping Order Aggregation

Refresh Auto refresh On


Transform	Rows	Time
OrderHeader Sink	-	-
OrderLineDetails Sink	-	-
JoinHeaderToLineD...	542	277ms
OrderLineCount Sink	32	3s 405ms
OrderSummary	32	-

OrderLineCount Stream Information

Metric	Value
Rows calculated	32
Total partition	190
Stage time	3s 405ms
Last update (GDT)	25/08/2020, 14:44:44

OrderLineCount Partition Chart

Row count vs Partition



Partition	Row count
1	0.0
177	1.0

OrderSummary Processing Time

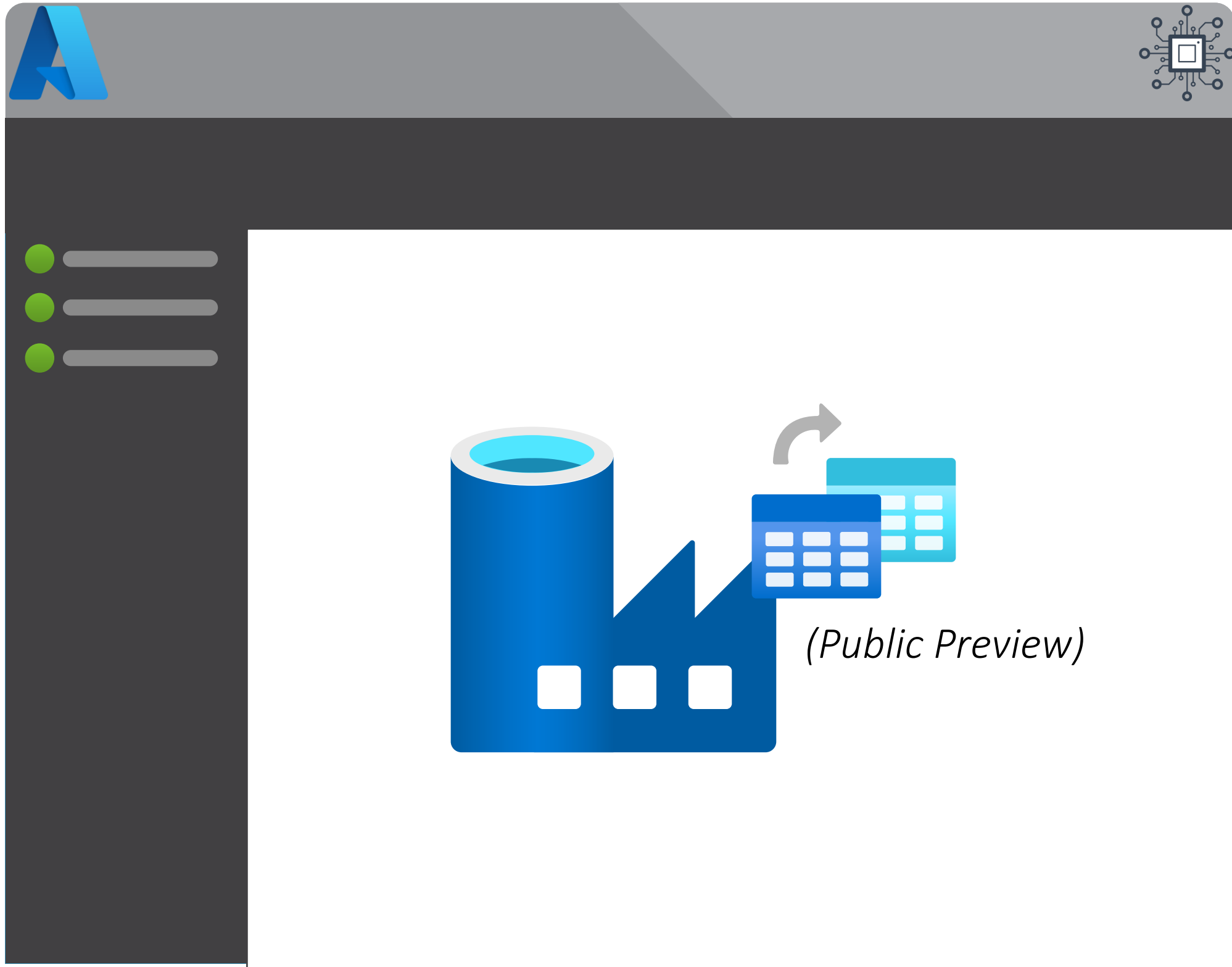
Transform	Rows	Time
OrderHeader	-	-
JoinHeaderToLineDetails	542	277ms
OrderLineDetails	542	277ms
OrderSummary	32	-
OrderLineCount	32	3s 405ms

Cluster startup time: 6s 878ms Number of transforms: 5

Edit transformation

Module 3

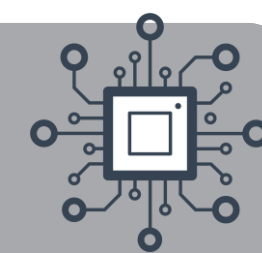
Data Transformation



- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases



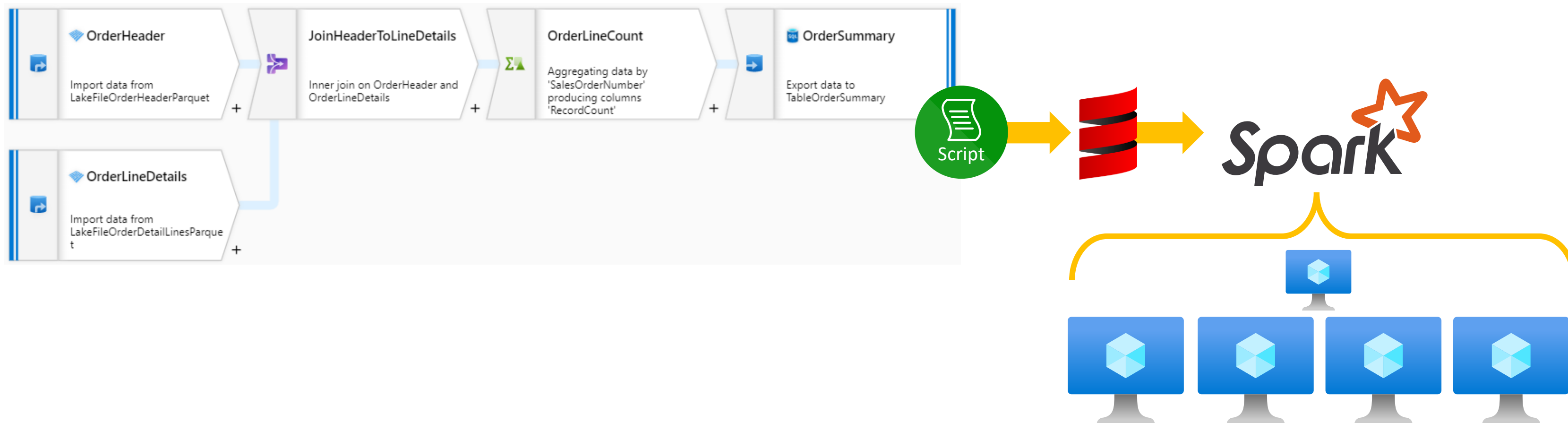
What is a Data Flow?

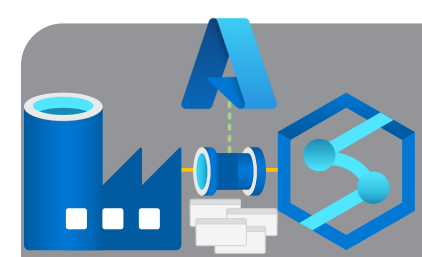


Control Flow

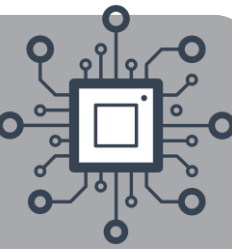


Data Flow

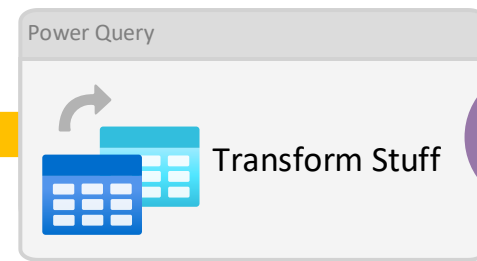
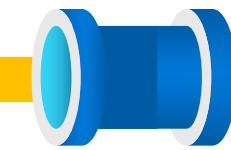
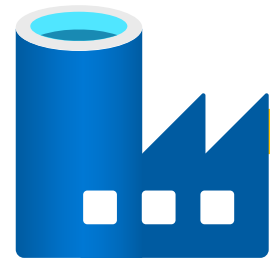




What is a Power Query Activity?



Control
Flow





Power Query

Home

Transform

Add column

View

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: Whole number

Use first row as headers

Replace values

Merge queries

Append queries

Combine files

New query

Options

Parameters

Query

Manage columns

Reduce rows

Transform

Combine

Queries

ADfResource [1]

LakeFileOrderDetailL...

UserQuery

fx

= 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-42e6-4d00-9a14-83i...
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...

Query settings

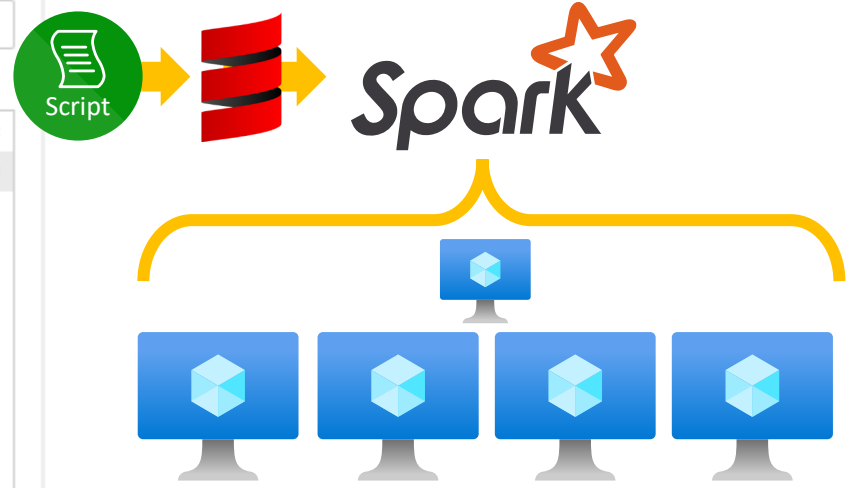
Name

LakeFileOrderDetailLinesP...

Applied steps

AdfDoc

Parquet





Power Query

Home Transform Add column View

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: Whole number Use first row as headers Replace values Merge queries Append queries Combine files

Queries

ADFResource [1]

LakeFileOrderDetailL...

UserQuery

	1 ² 3 SalesOrderID	1 ² 3 SalesOrderDetailID	1 ² 3 OrderQty	1 ² 3 ProductID	1.2 UnitPrice	1.2 UnitPriceDiscount	1.2 LineTotal	A ^B C 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

Query settings

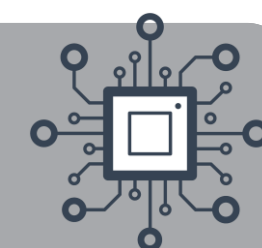
Name

LakeFileOrderDetailLinesP...

Applied steps

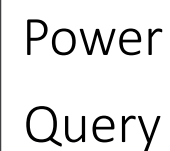
AdfDoc

Parquet



Home

Control Flow



What can a Power Query Activity do?

Transform

Control Flow



Power Query

The screenshot displays the Power Query Editor interface. The top ribbon shows the 'Transform' tab with various data manipulation options. The left pane lists queries: 'ADFSResource', 'LakeFileOrderDetail...', and 'UserQuery'. The main area shows a table with columns: 'SalesOrderID', 'SalesOrderDetailID', 'OrderQty', 'ProductID', 'UnitPrice', and 'UnitPrice'. The bottom right pane shows the 'Query Settings' for 'OrderDetailLines', including 'Properties' and 'Applied Steps'.

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 Power Query Activity do?

Add Column



Power Query

Power Query Editor interface showing the 'Add Column' tab and a data table.

Queries

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

OrderDetailLines

	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	

Query Settings

PROPERTIES

Name: OrderDetailLines

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type



What can a Power Query Activity do?



View

Control Flow



Power Query

Power Query Editor interface showing the 'View' tab and a data preview table.

Queries

- ADFSResource [1]
- LakeFileOrderDetailL...
- UserQuery

OrderDetailLines

	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	

Query Settings

PROPERTIES

Name: OrderDetailLines

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type



What can a Power Query Activity do?



View

Control
Flow



Power
Query



Module 3

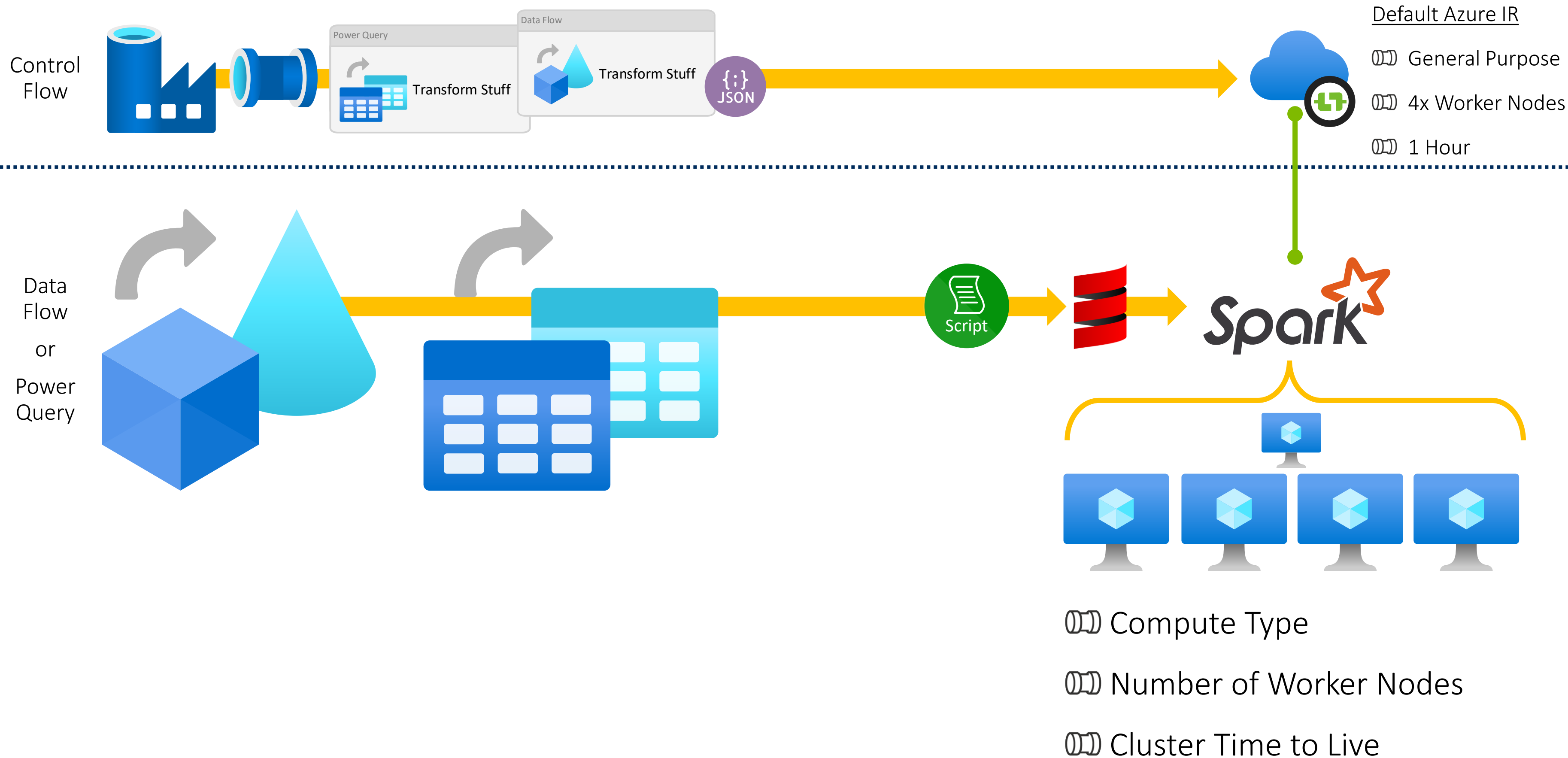
Data Transformation



- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

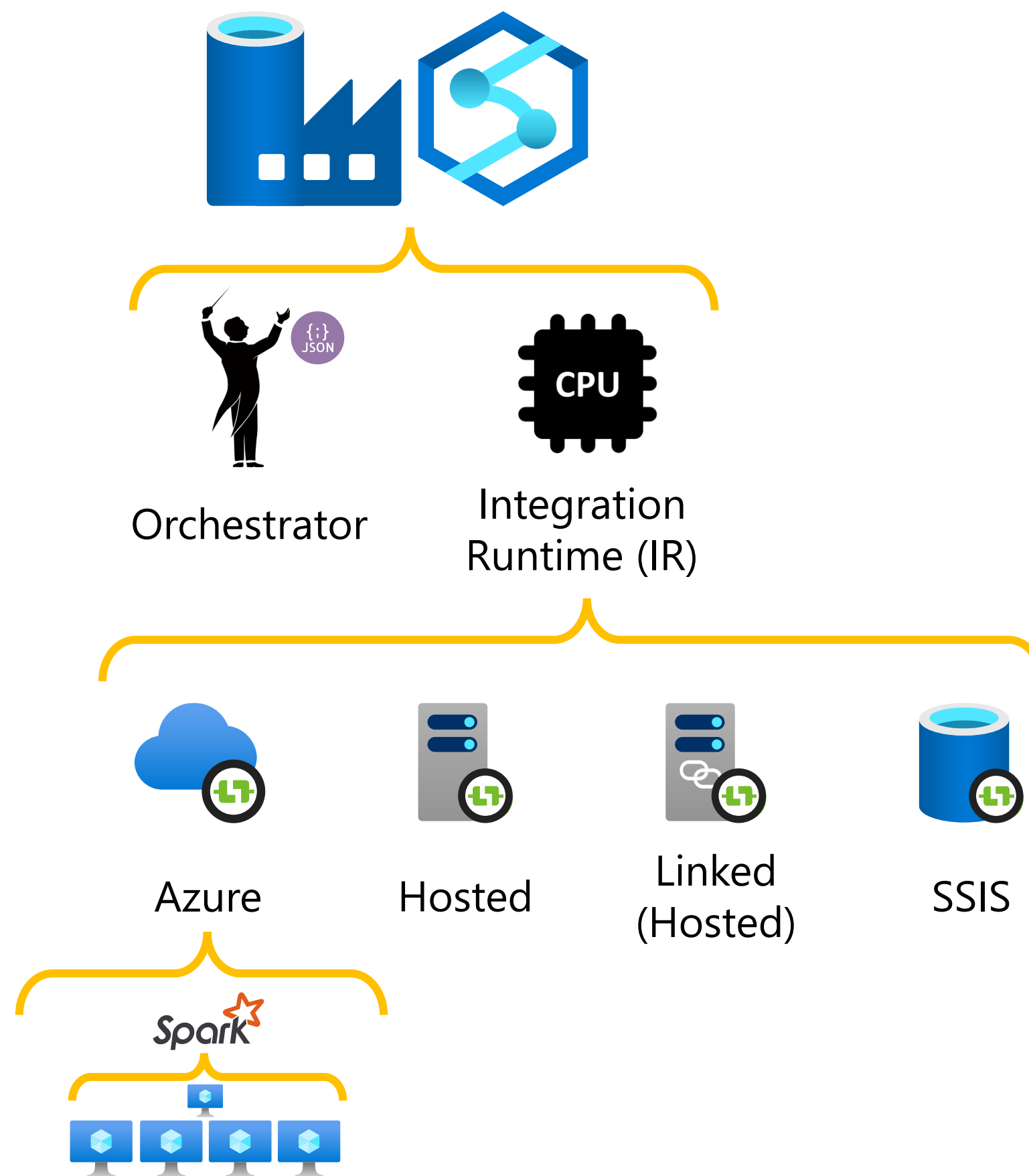


Spark Configuration



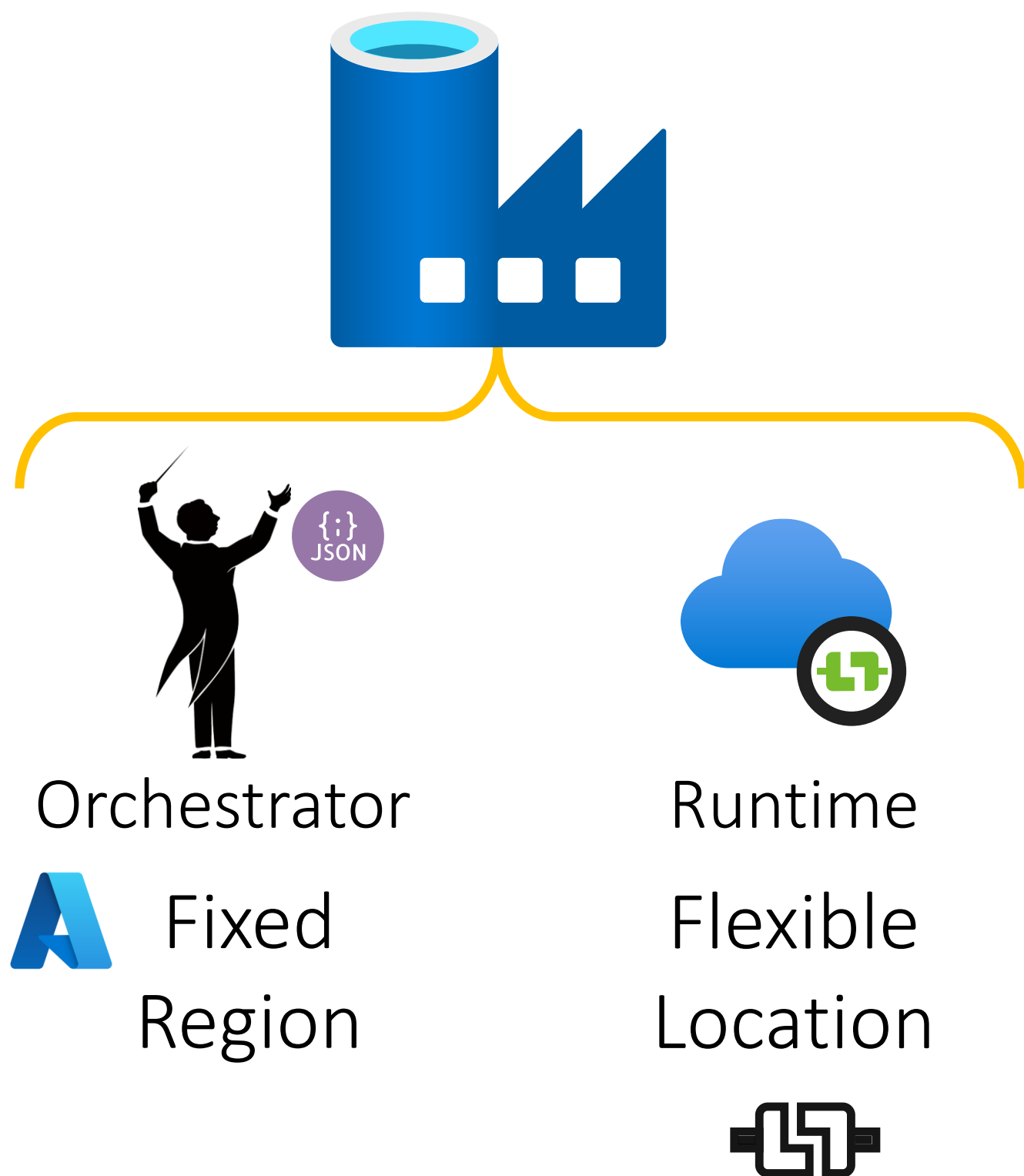


Data Flow Compute – IR's vs Spark





What is an Integration Runtime?



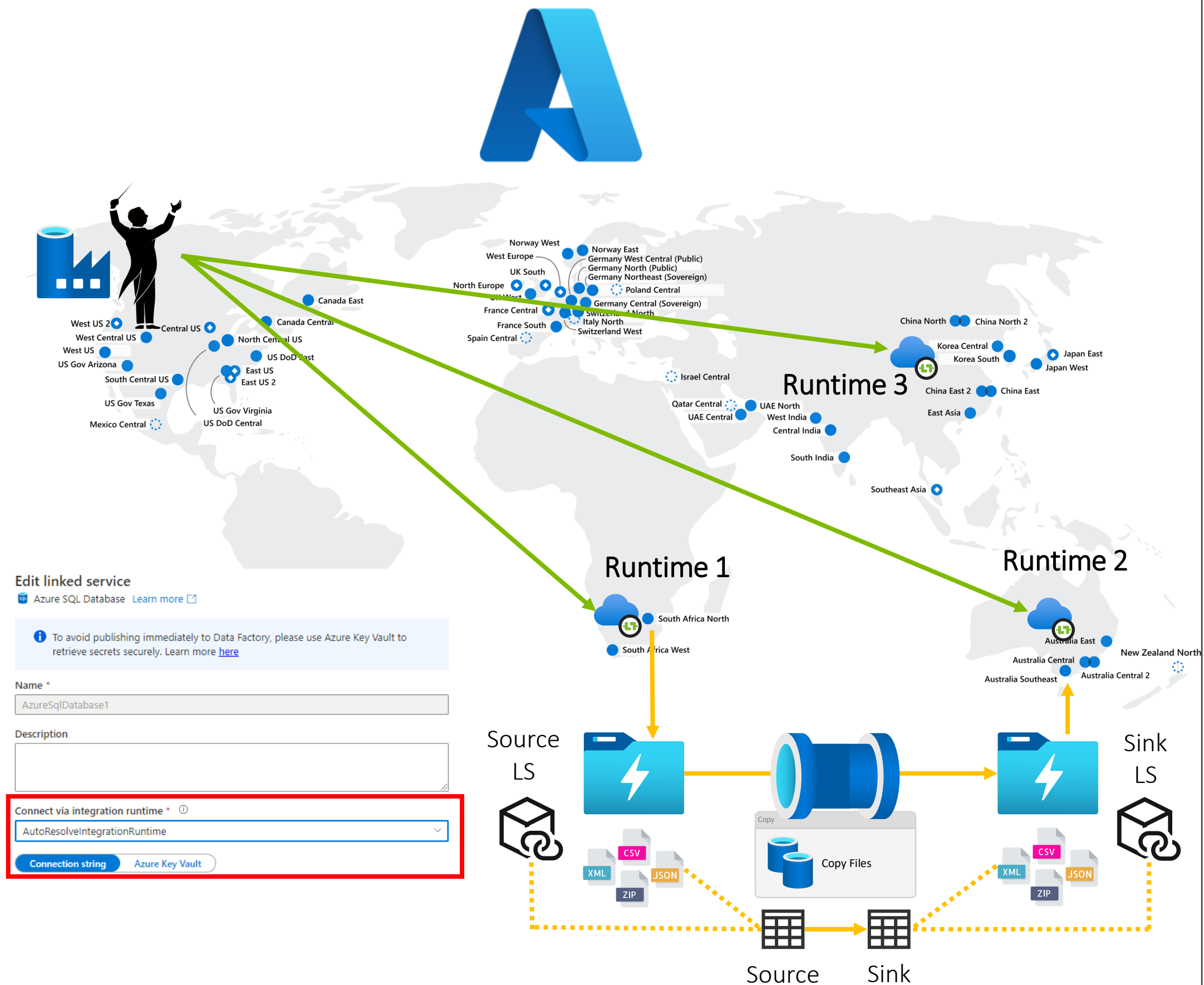
Orchestrator

A Fixed
Region

Runtime
Flexible
Location



AutoResolveIntegrationRuntime



Edit linked service

Azure SQL Database [Learn more](#)

To avoid publishing immediately to Data Factory, please use Azure Key Vault to retrieve secrets securely. [Learn more](#)

Name *

AzureSqlDatabase1

Description

Connect via integration runtime *

AutoResolveIntegrationRuntime

Connection string

Azure Key Vault

Source
LS



Sink
LS

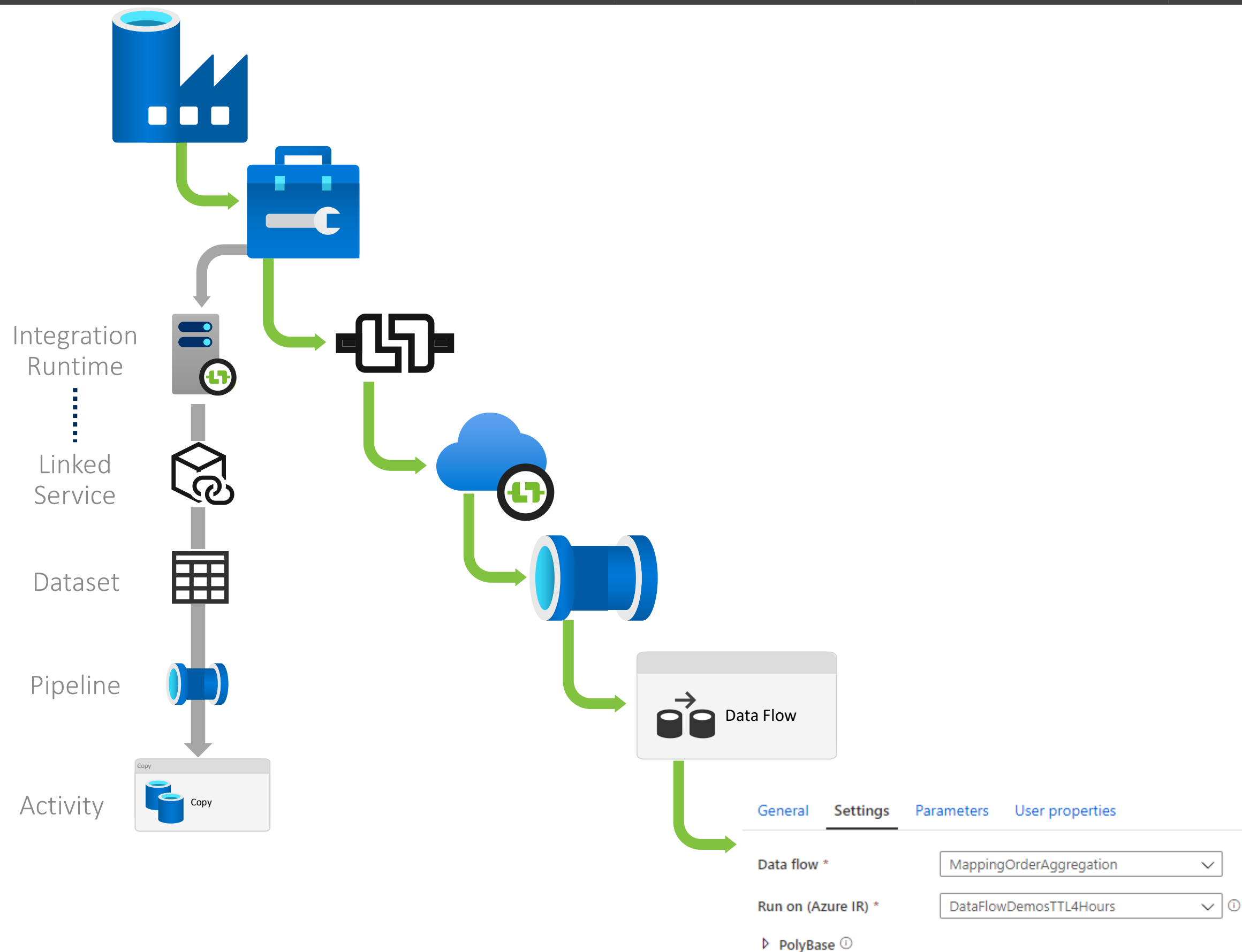


Source

Sink



Setting the Data Flow Cluster (IR Configuration)



Data Factory

Manage

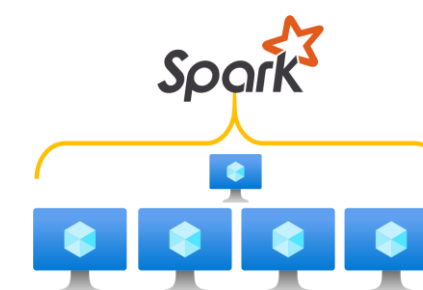
Integration Runtimes

Azure IR

Pipeline

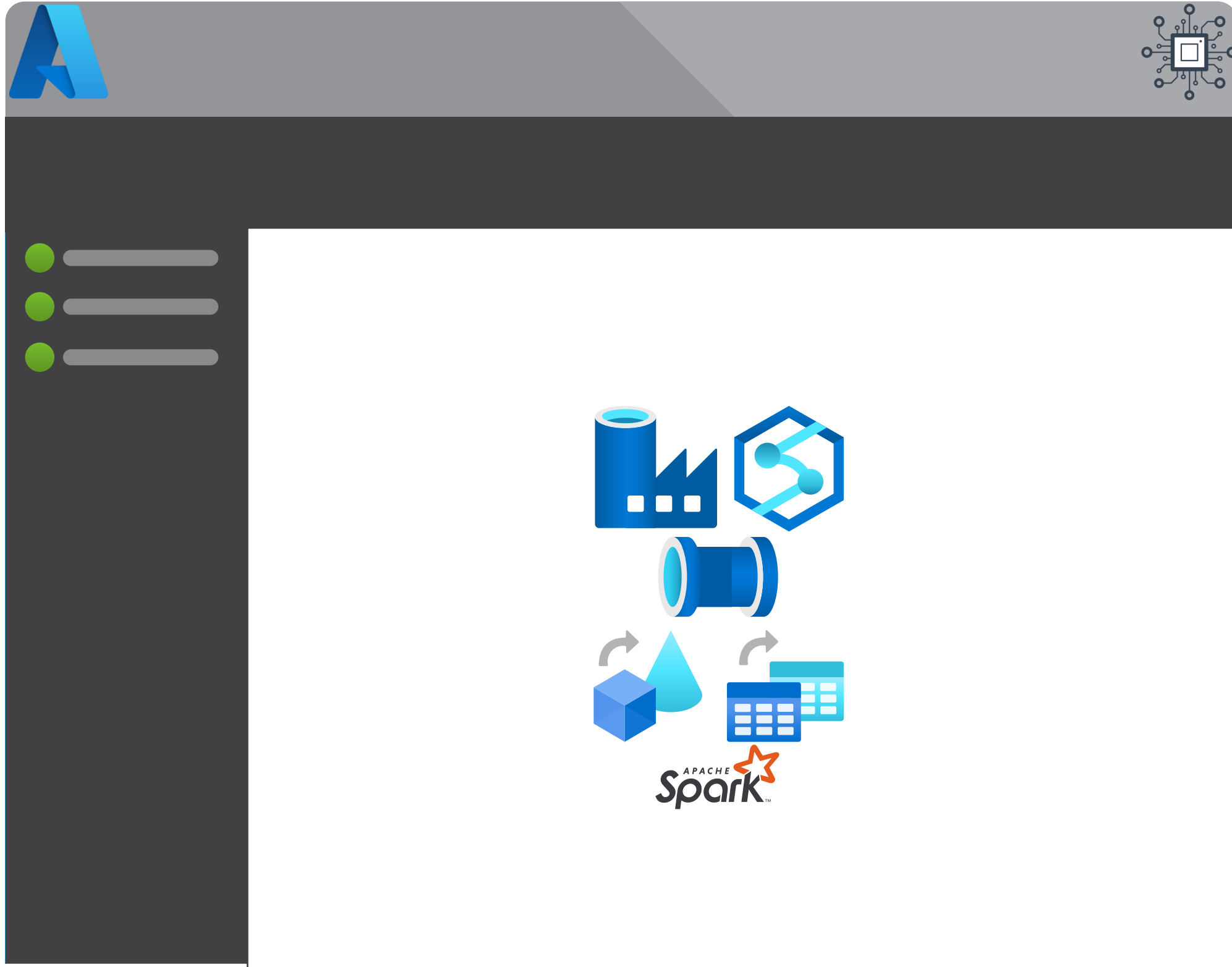
Data Flow Activity

Settings



Module 3

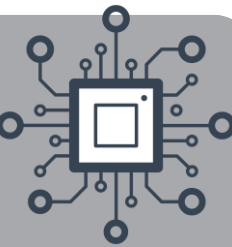
Data Transformation


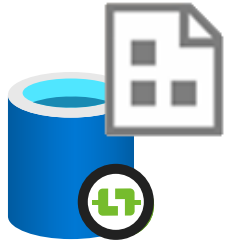

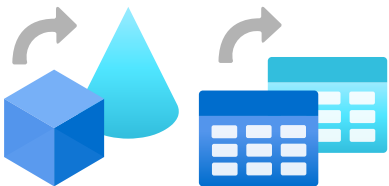


- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases



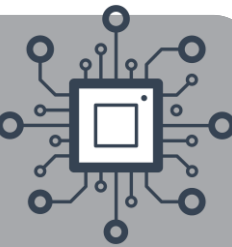
Data Transformation Resources in Azure Comparison



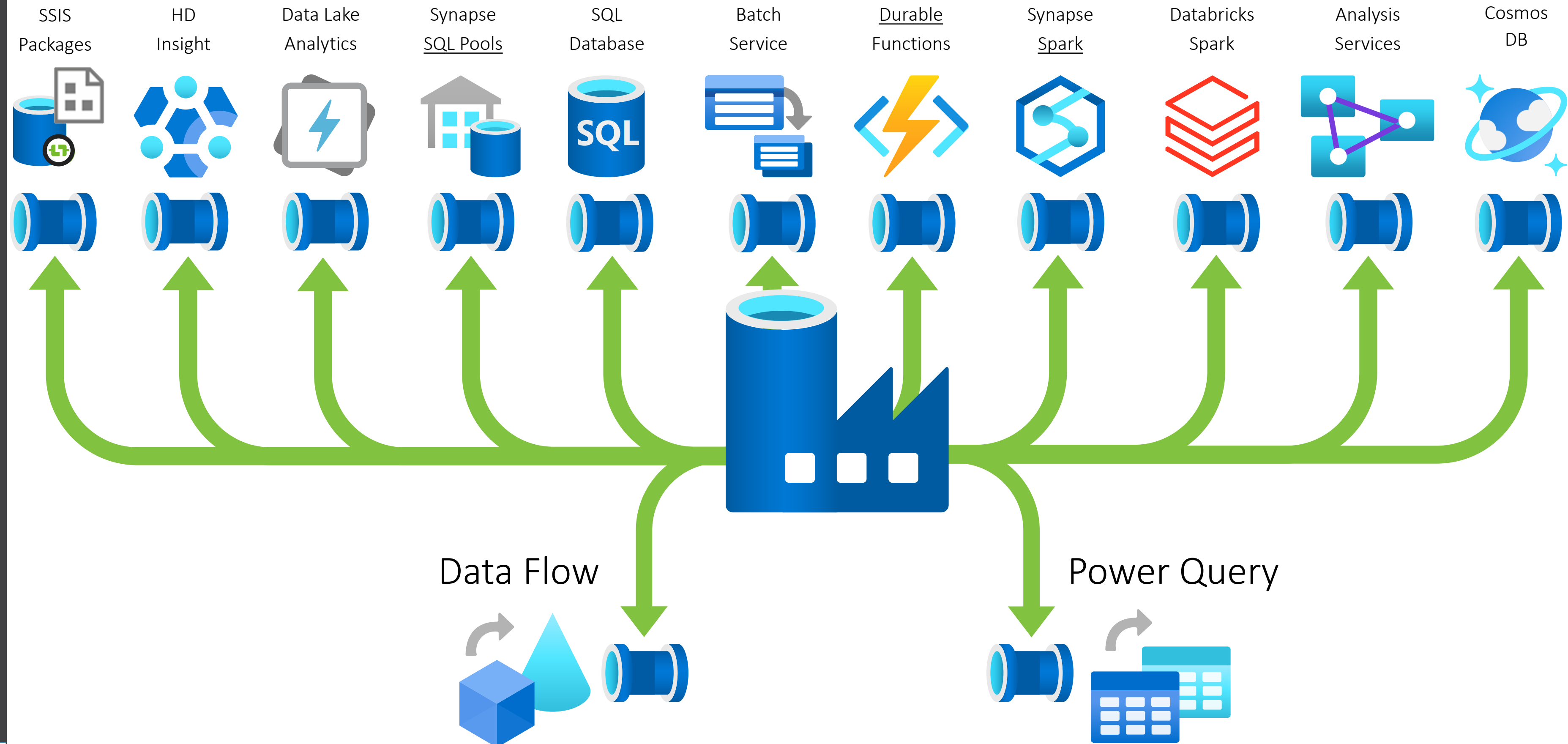
Transformation Tools		Graphical UI (Low/No Code)	Scales Out	Scales Up	Cloud Native Tech
	T-SQL with SQLDB	✗	✗	✓	✗
	SSIS Packages	✓	✗	✓	✗
	Scala/Python/SQL with Databricks	✗	✓	✓	✓
	Data Flows & Power Query	✓	✓	✓	✓



Other Data Transformation Services in Azure

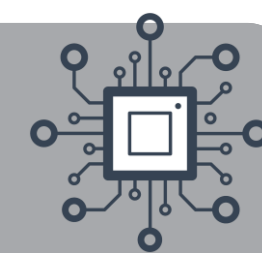


When Should We Use These Integration Pipeline Transformation Activities?

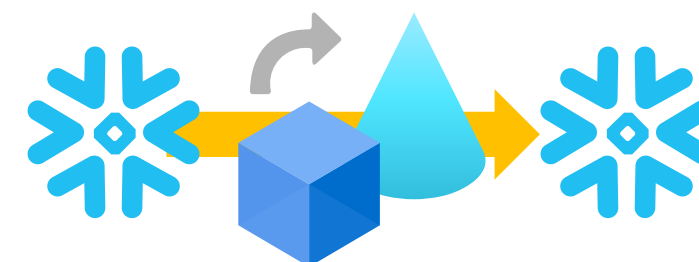
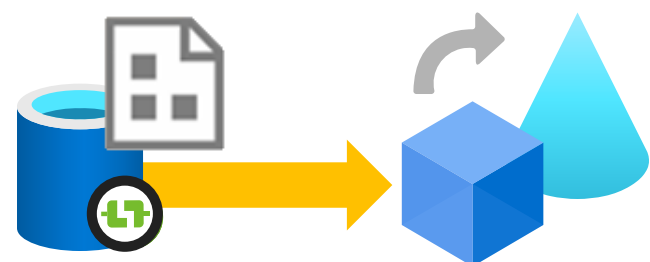




Use Cases

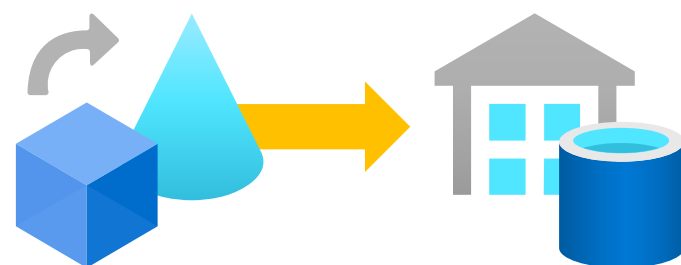


SSIS Package rebuild
and skills migration.

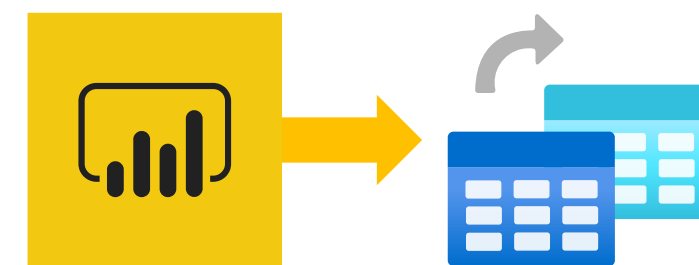
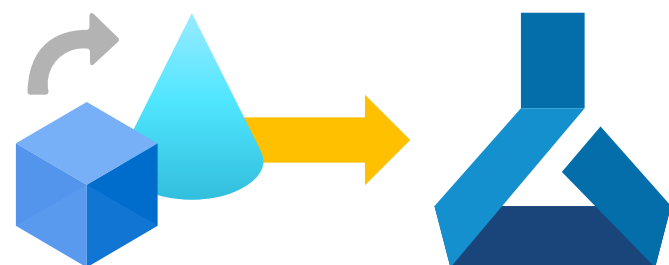


Inline dataset
transformations.

Warehouse data
distribution & loading.



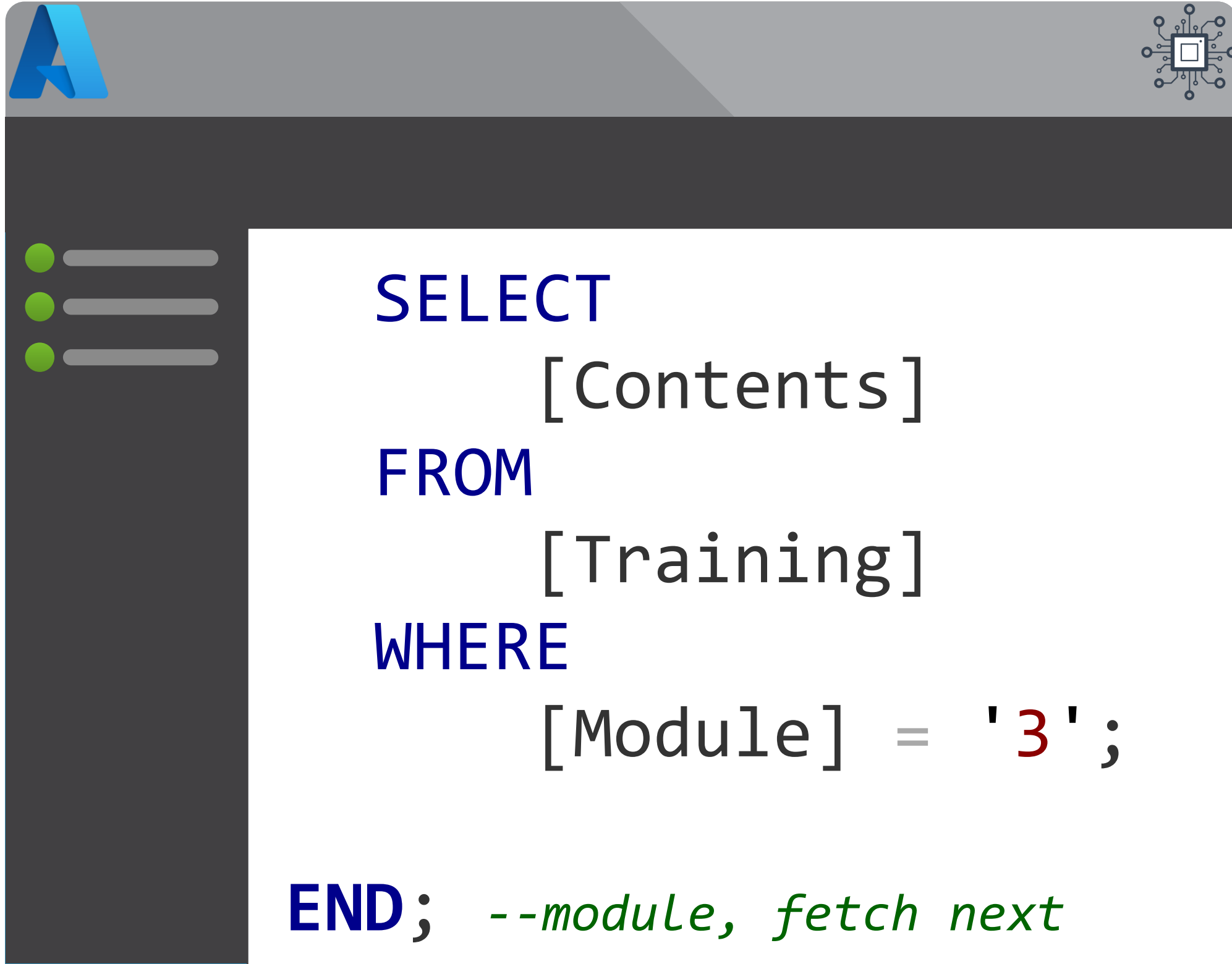
Data model dataset
preparation.



Power Query
industrialisation.

Module 3

Data Transformation



- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases