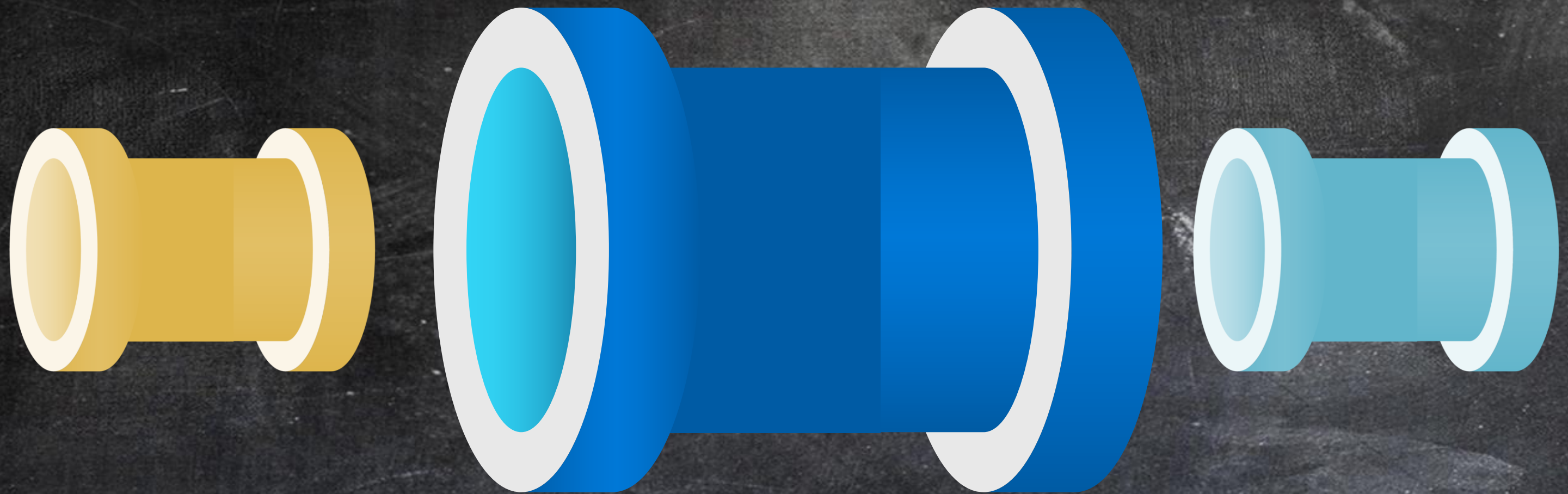


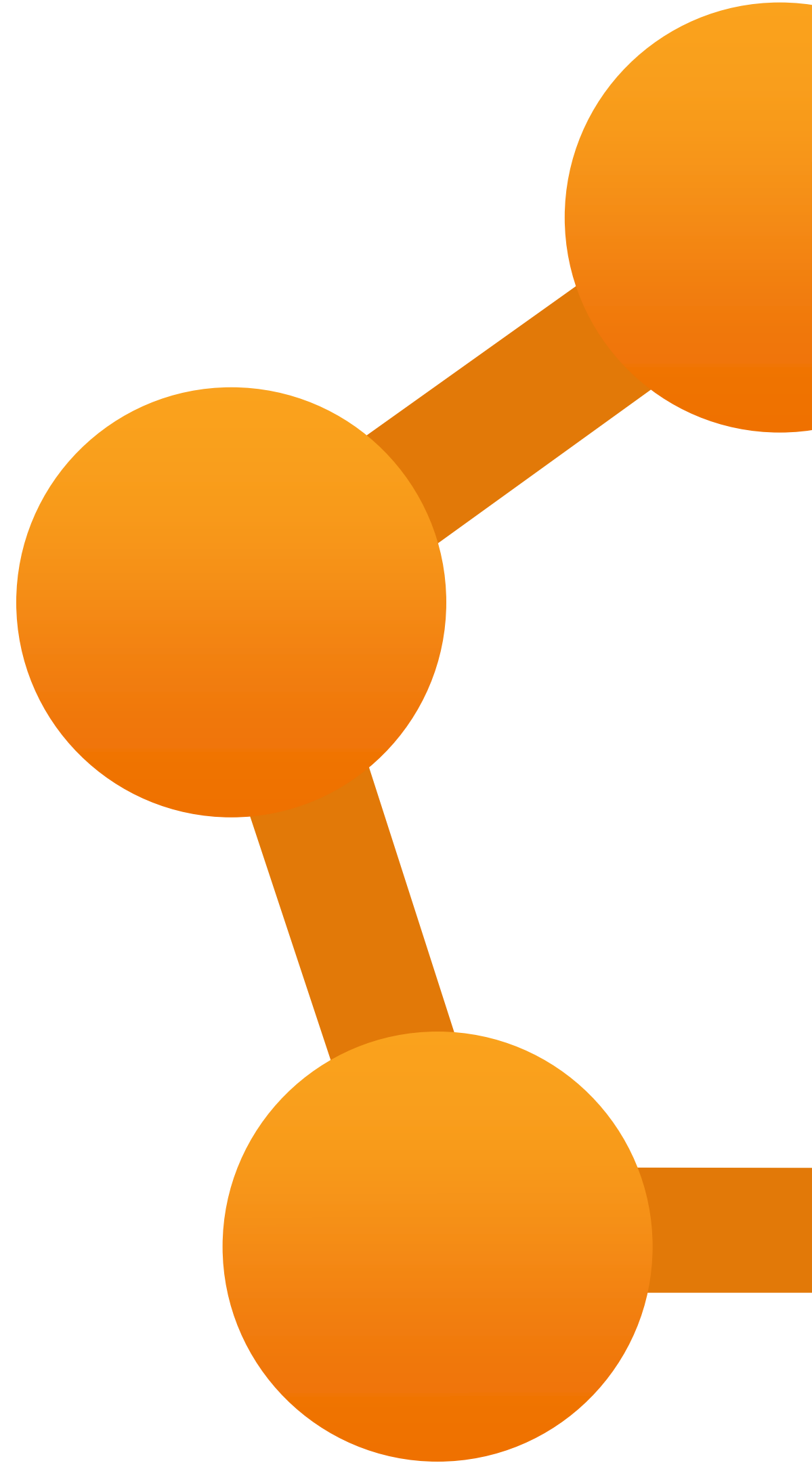
Integration Pipelines



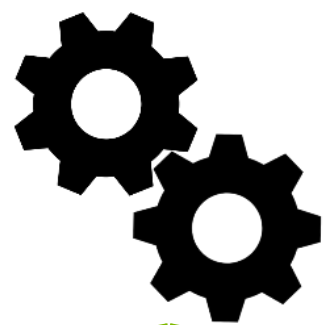
Module 3 – Data Transformation

Scaling Up vs Scaling Out

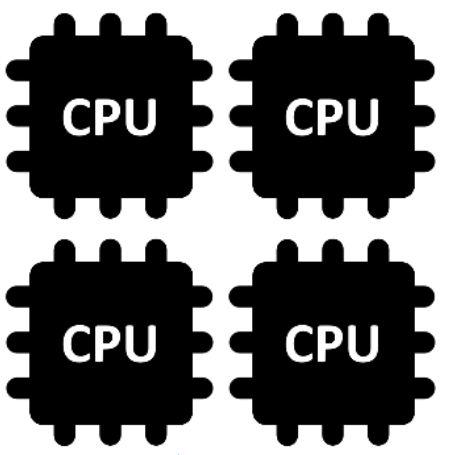
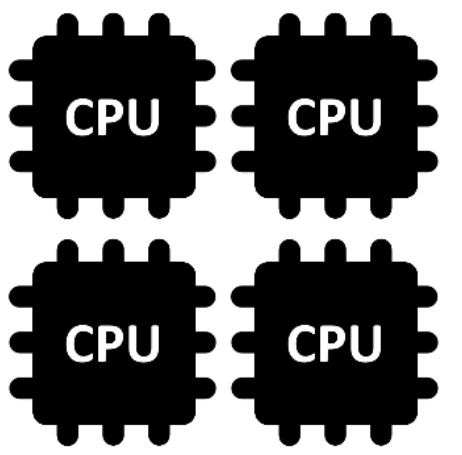
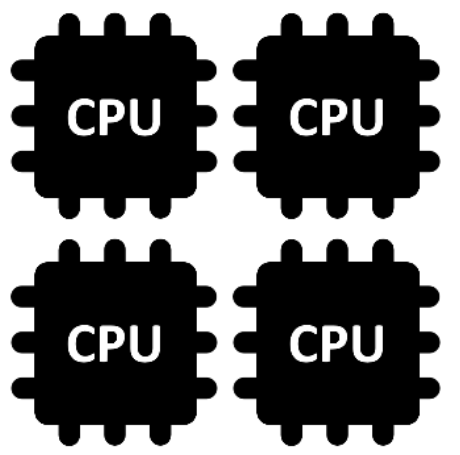
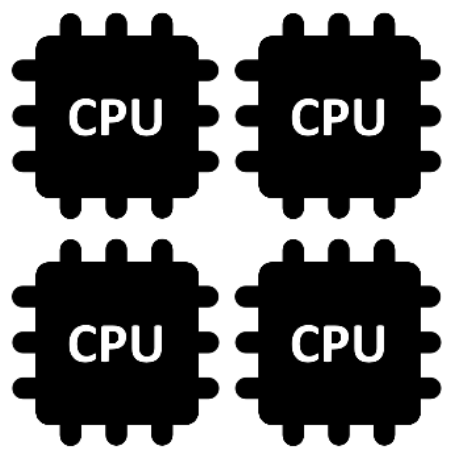
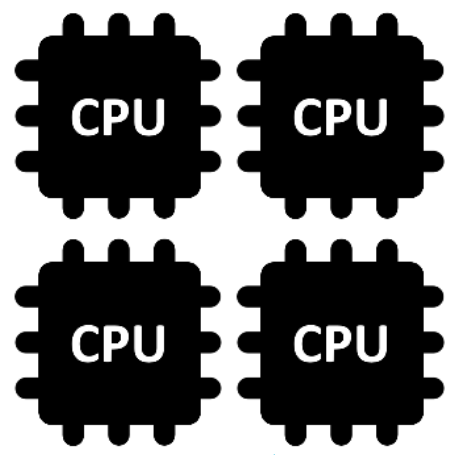
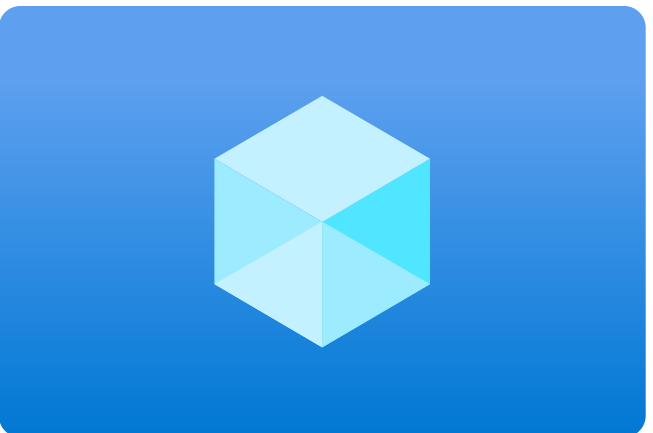
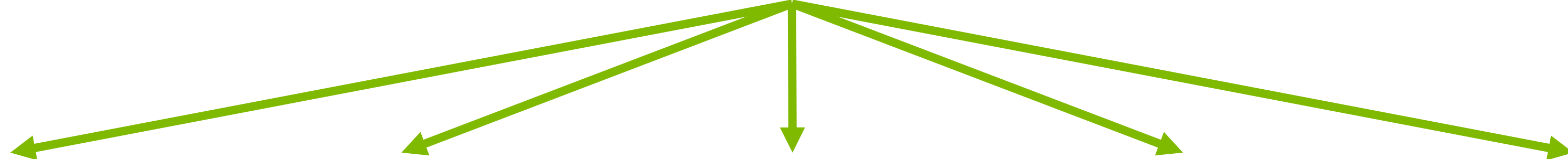
Cloud Formations



Scaling Up and/or Scaling Out



Workload:
Process 100TB of Data

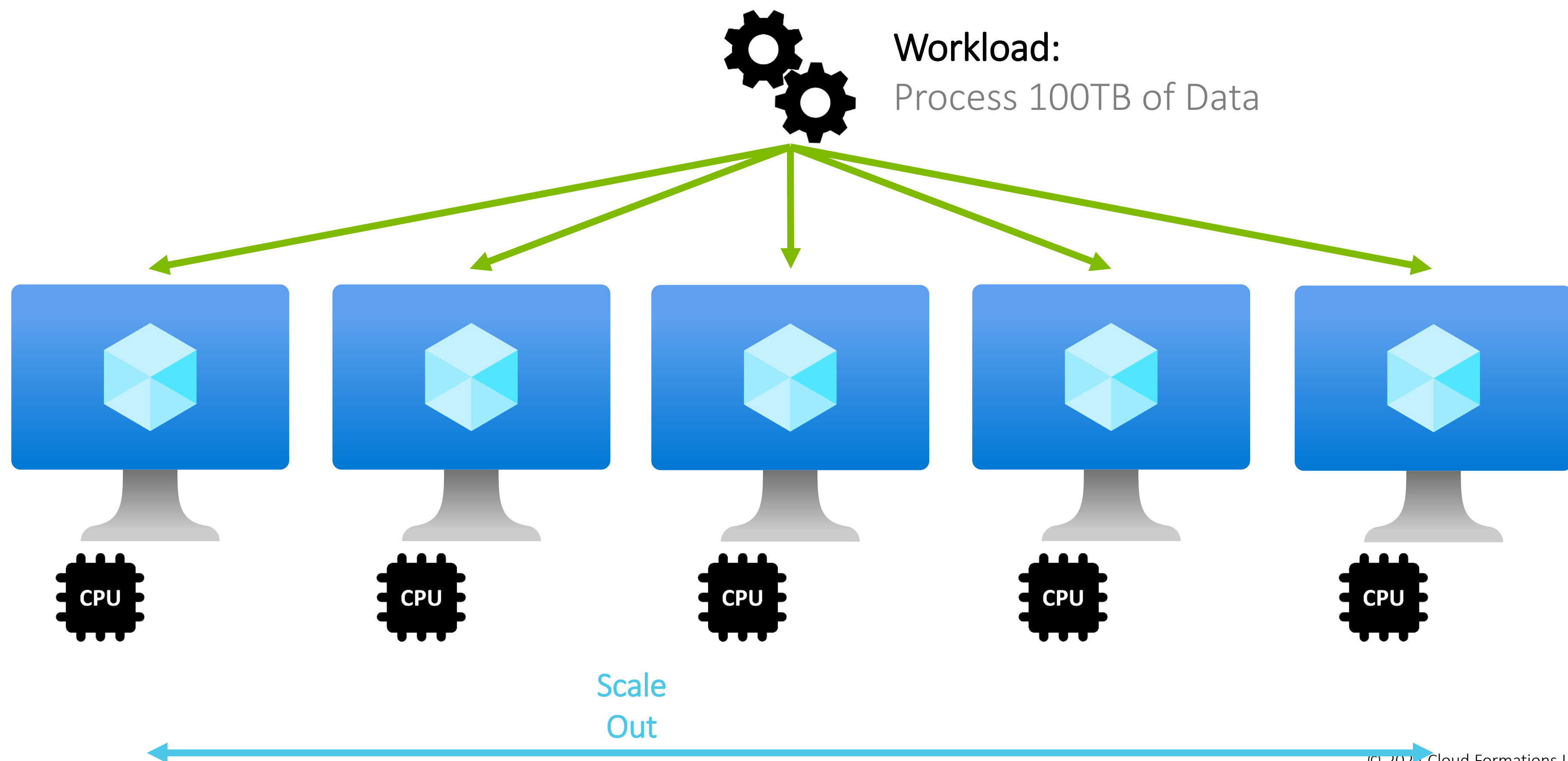


Scale
Out

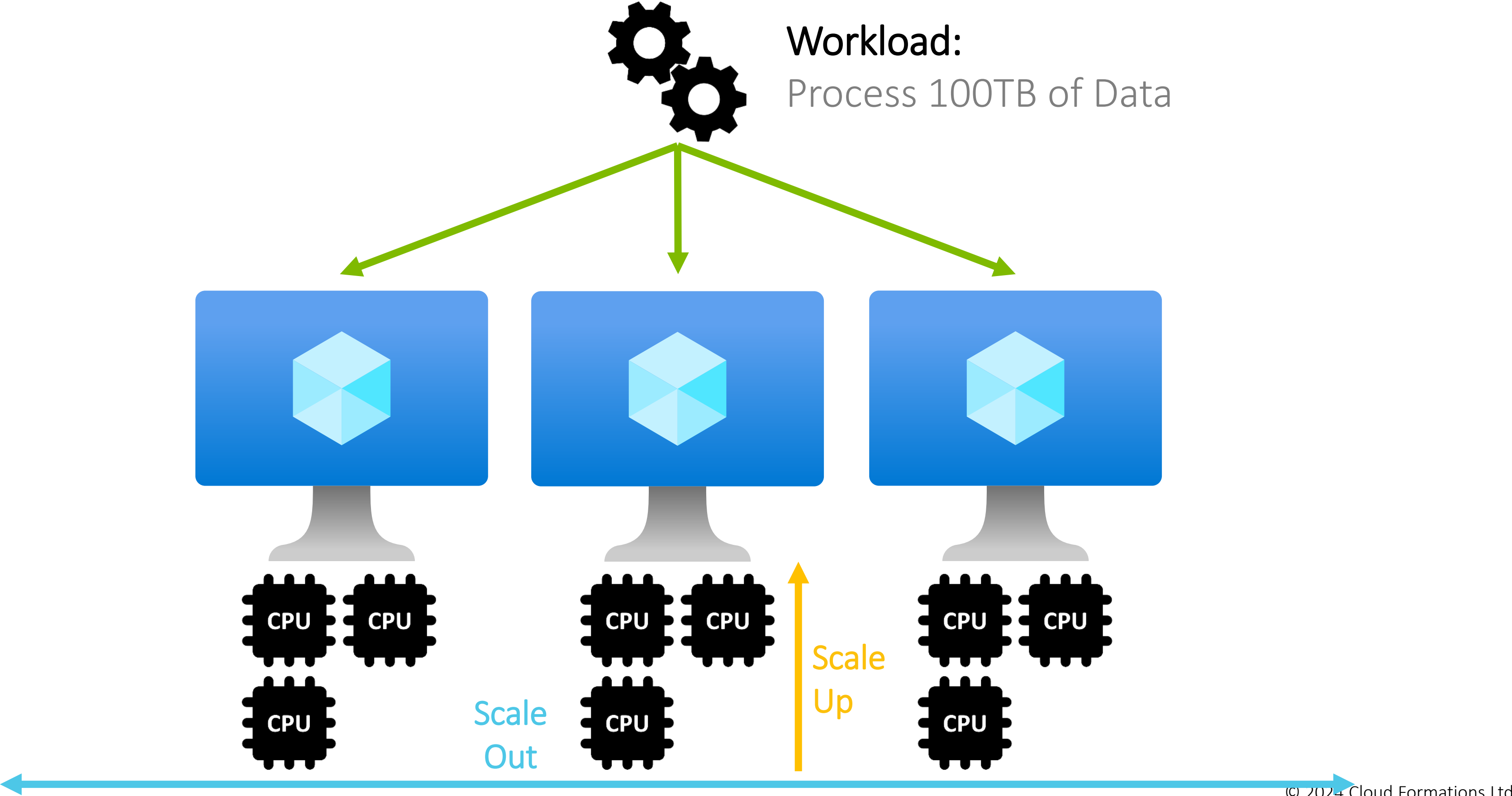
Scale
Up



Scaling Up and/or Scaling Out



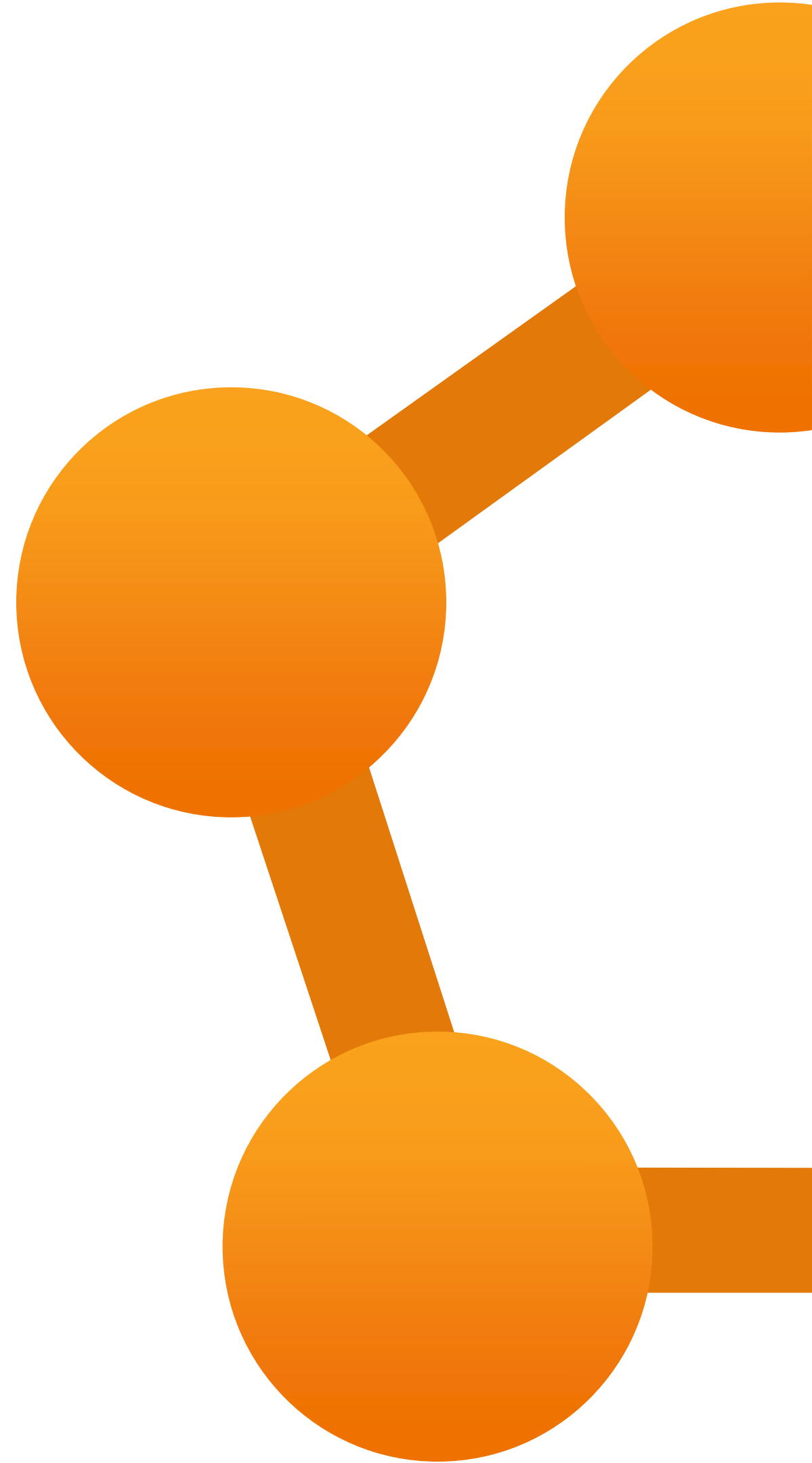
Scaling Up and/or Scaling Out



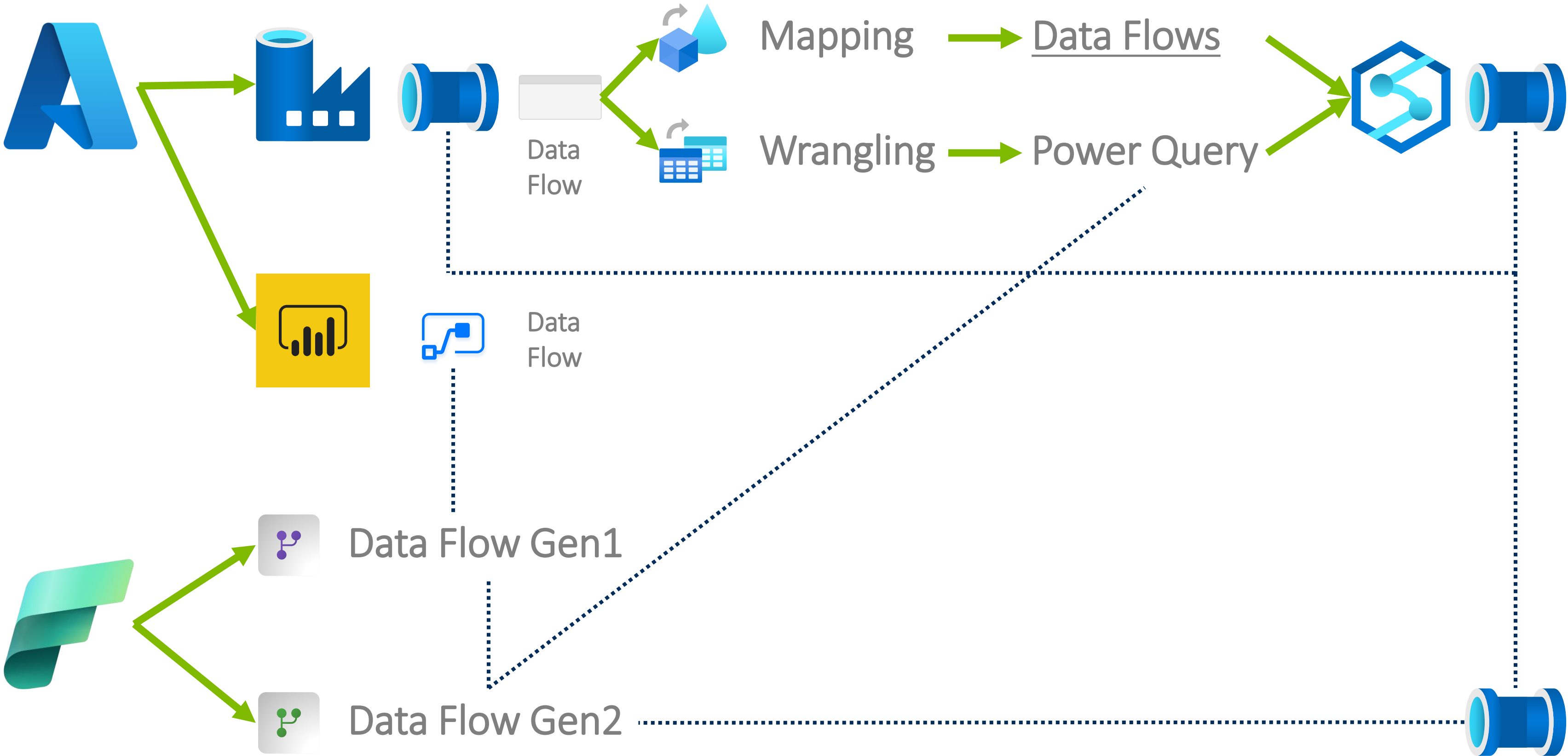
Module 3 – Data Transformation

Data Flows

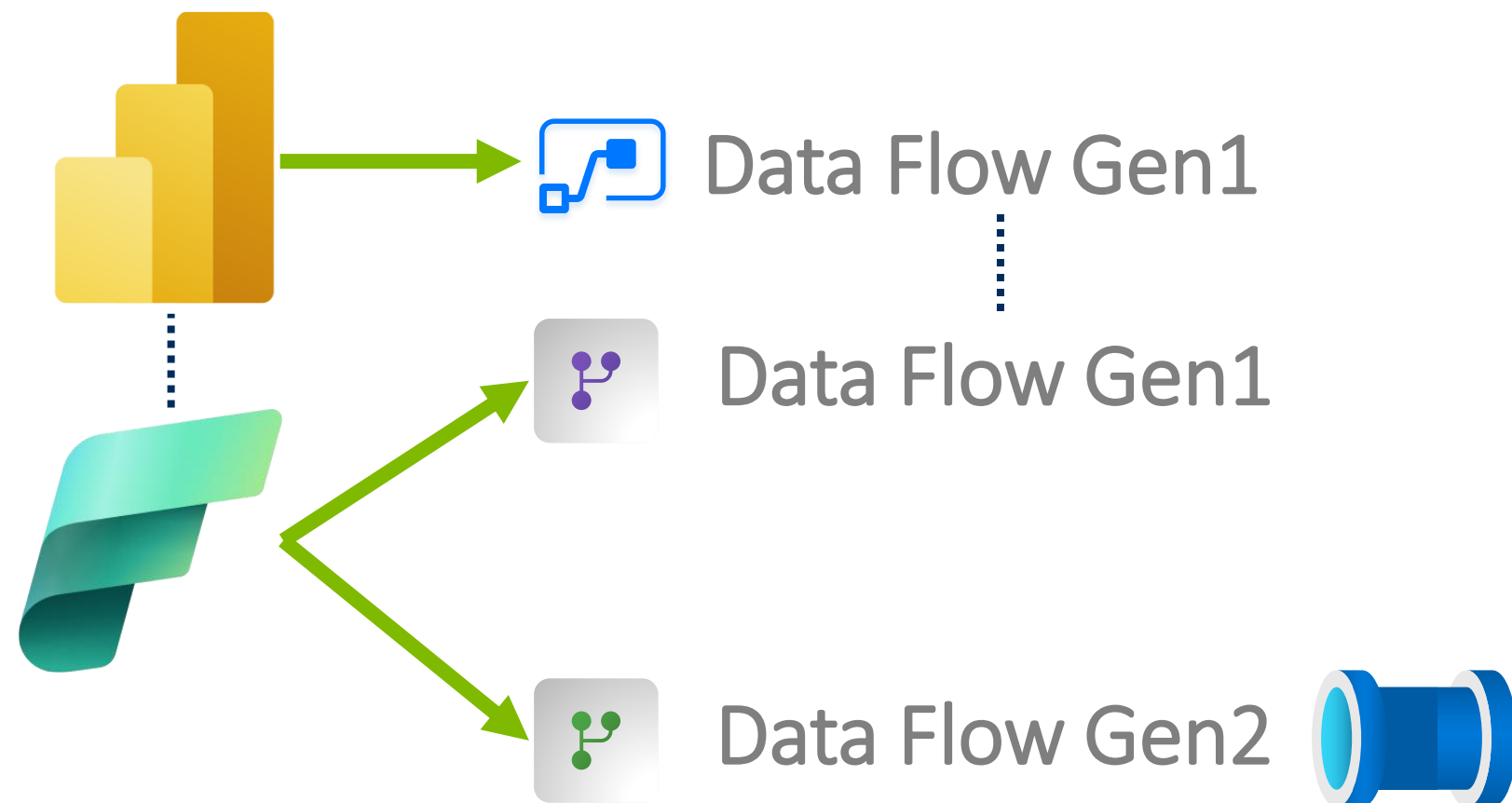
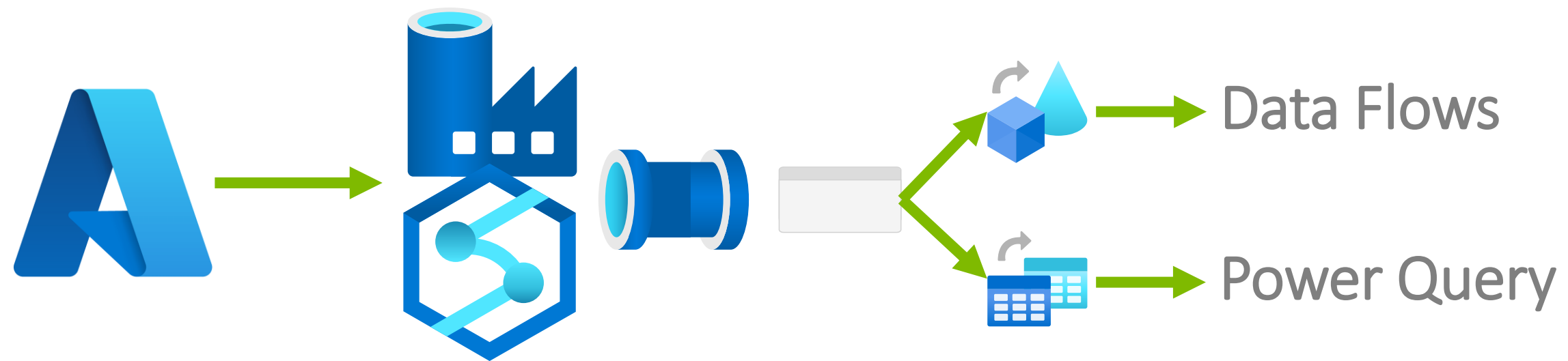
Cloud Formations



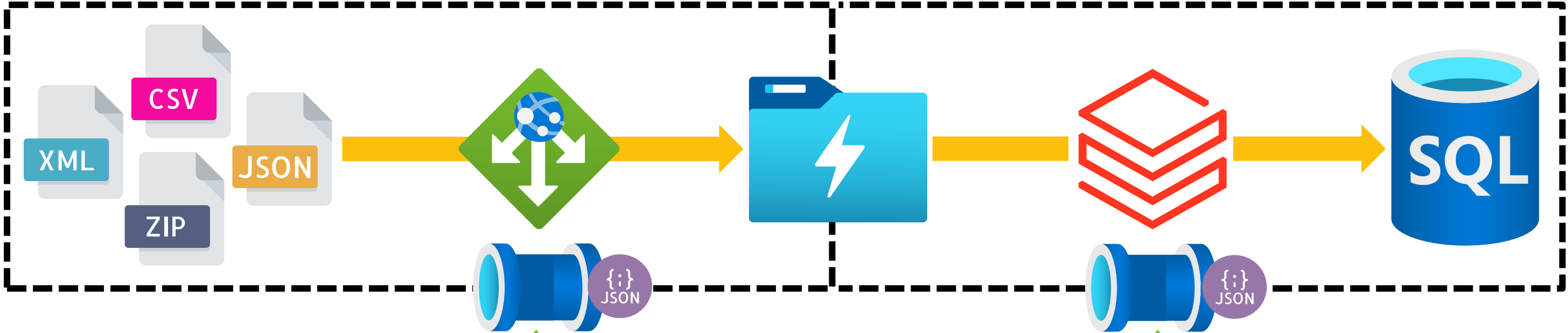
Terminology Clarification



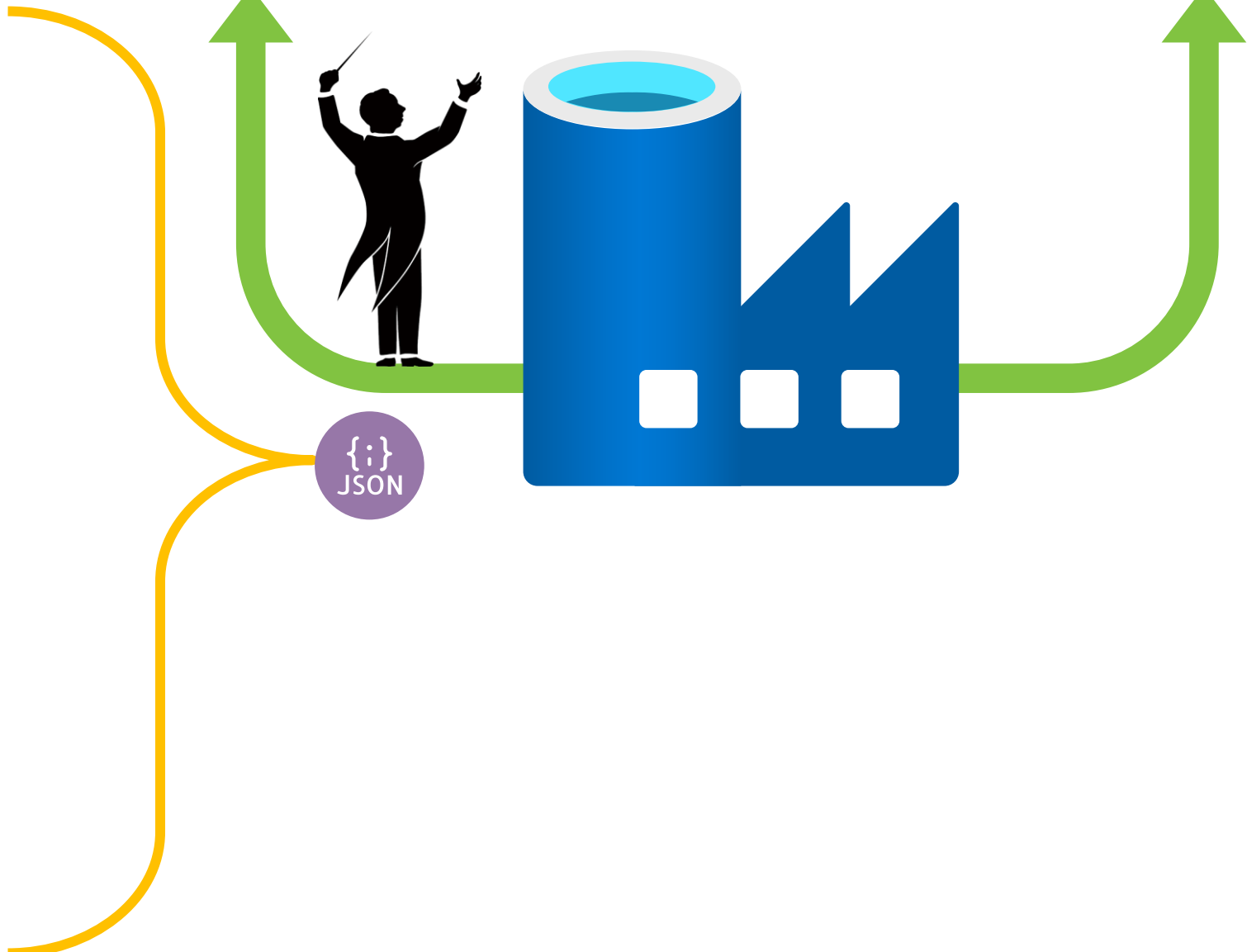
Terminology Clarification



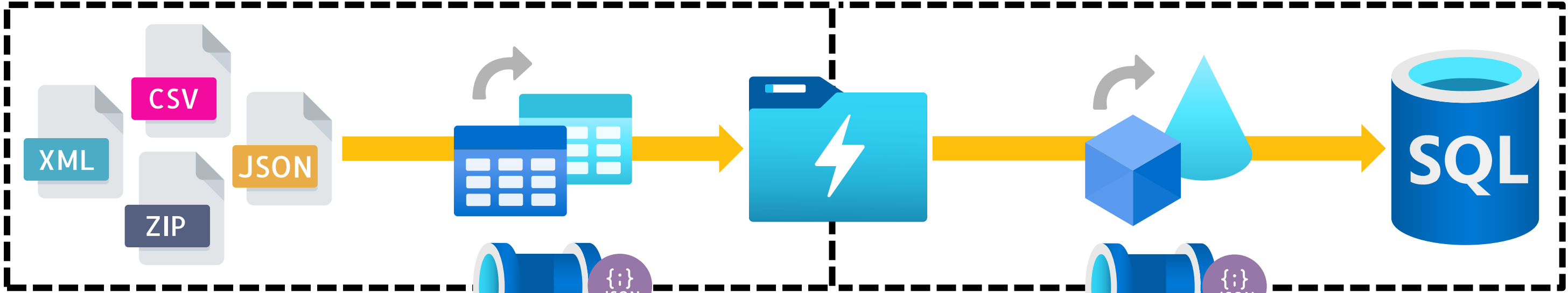
Control Flow Components



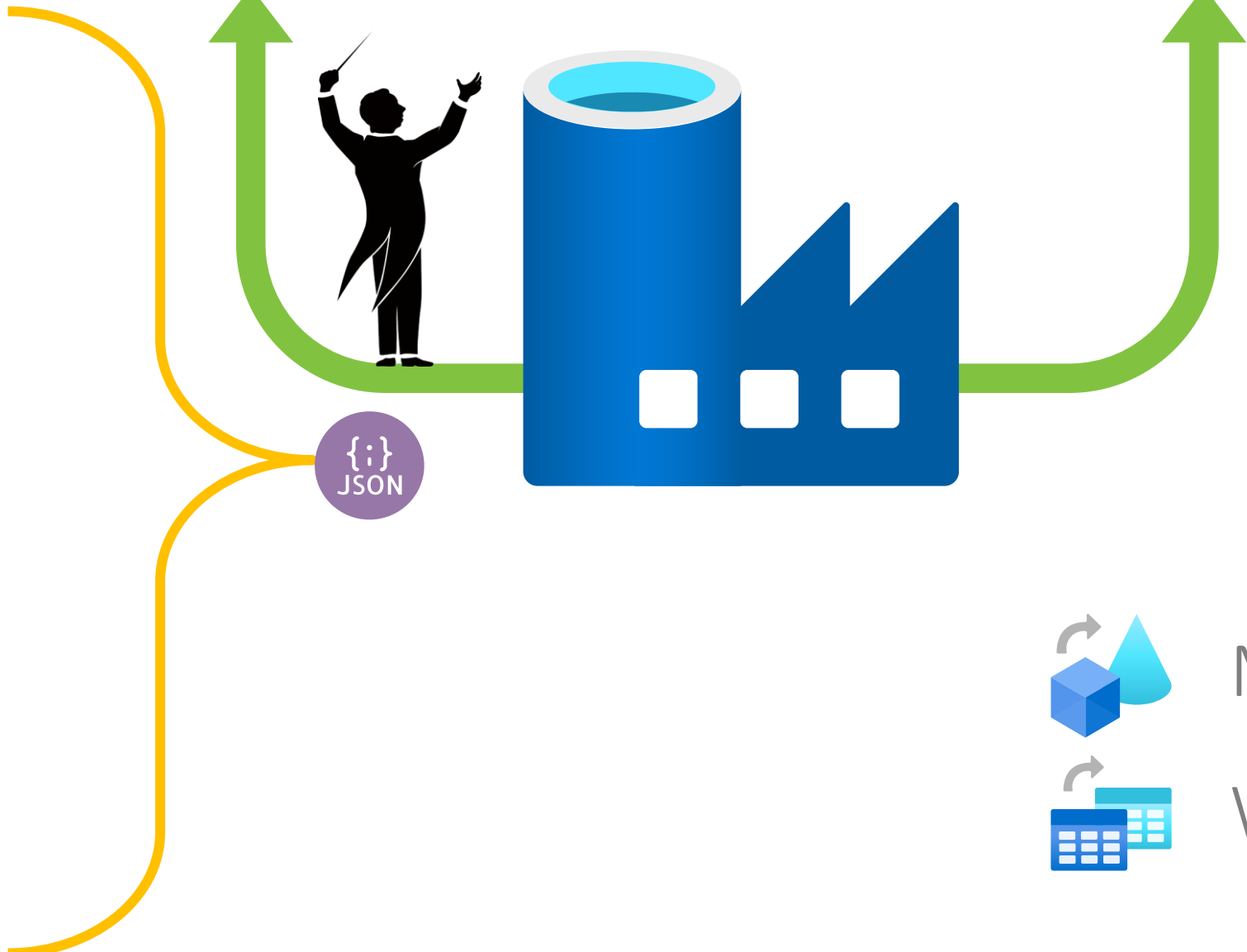
- 1 Linked Services
- 2 Datasets
- 3 Activities
- 4 Pipelines
- 5 Triggers




Control Data Flow Components



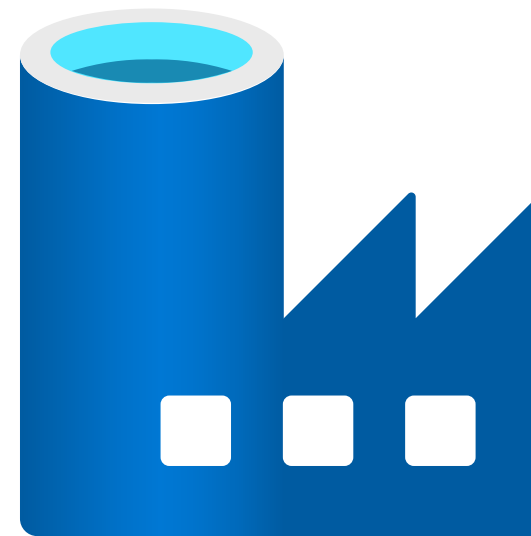
- 1 Linked Services
- 2 Datasets
- 3 Activities
- 4 Pipelines
- 5 Triggers



 Mapping  Data Flows

 Wrangling  Power Query

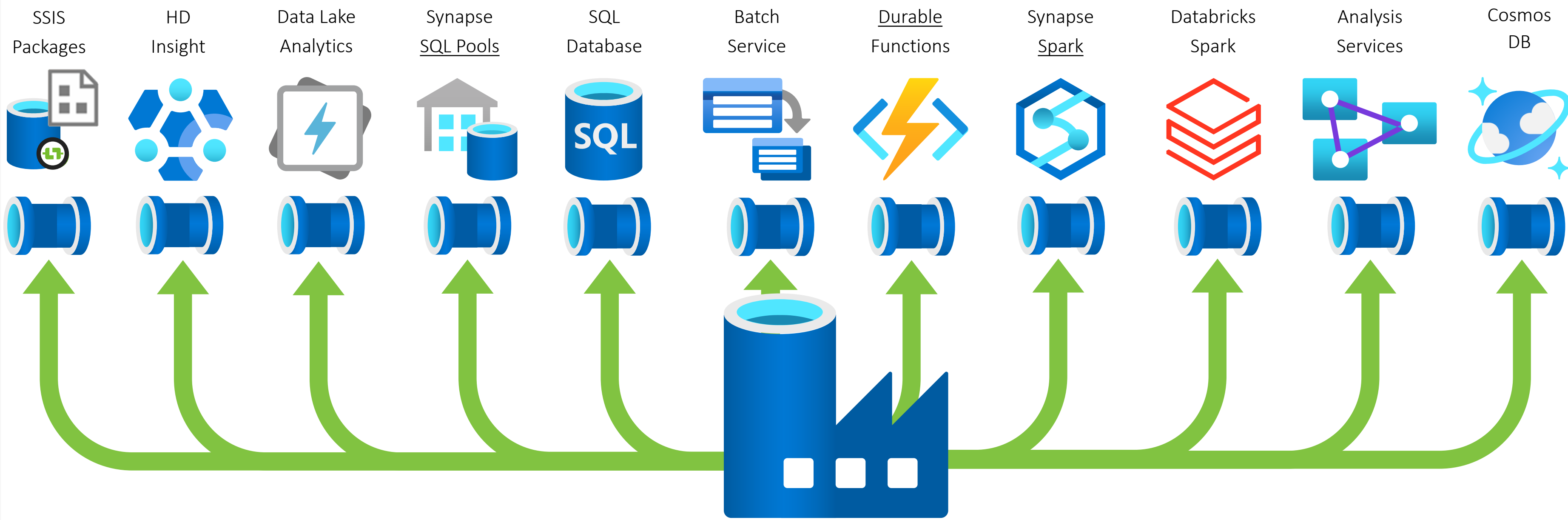
Data Transformation in Azure



Other Data Transformation Services in Azure



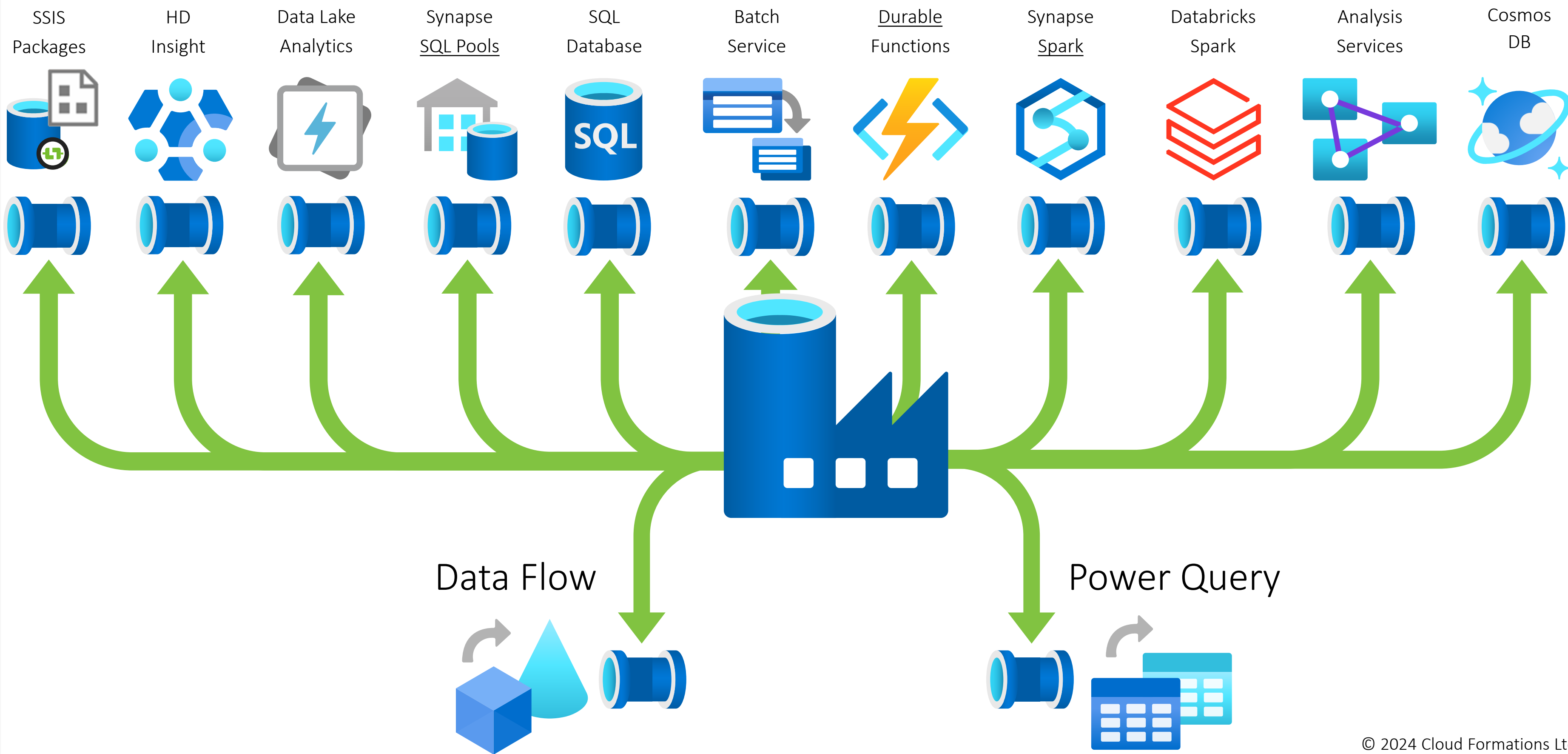
Cloud Formations - Knowledge Transfer & Training



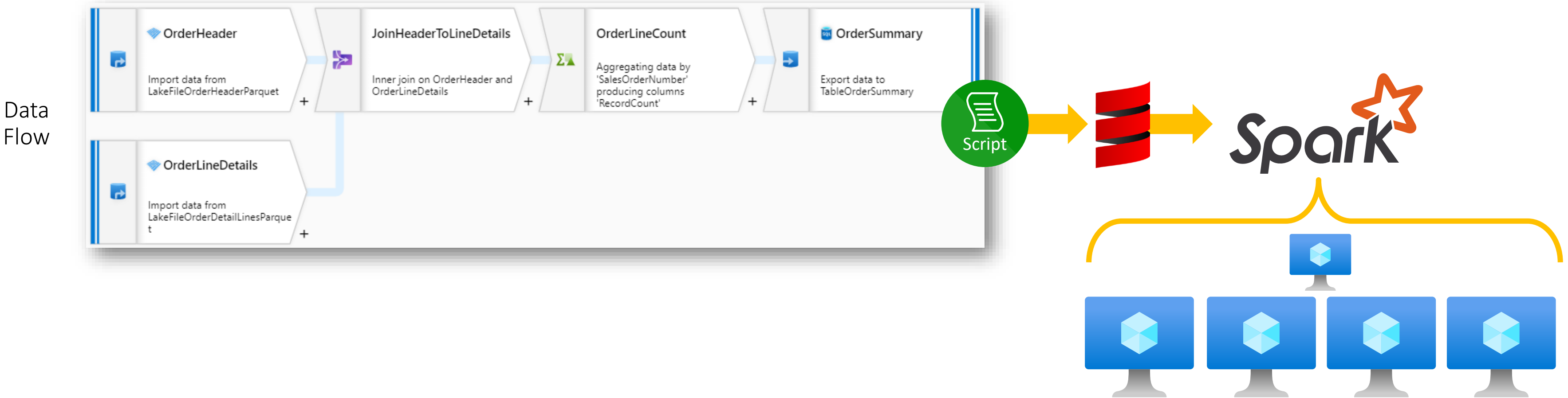
Other Data Transformation Services in Azure



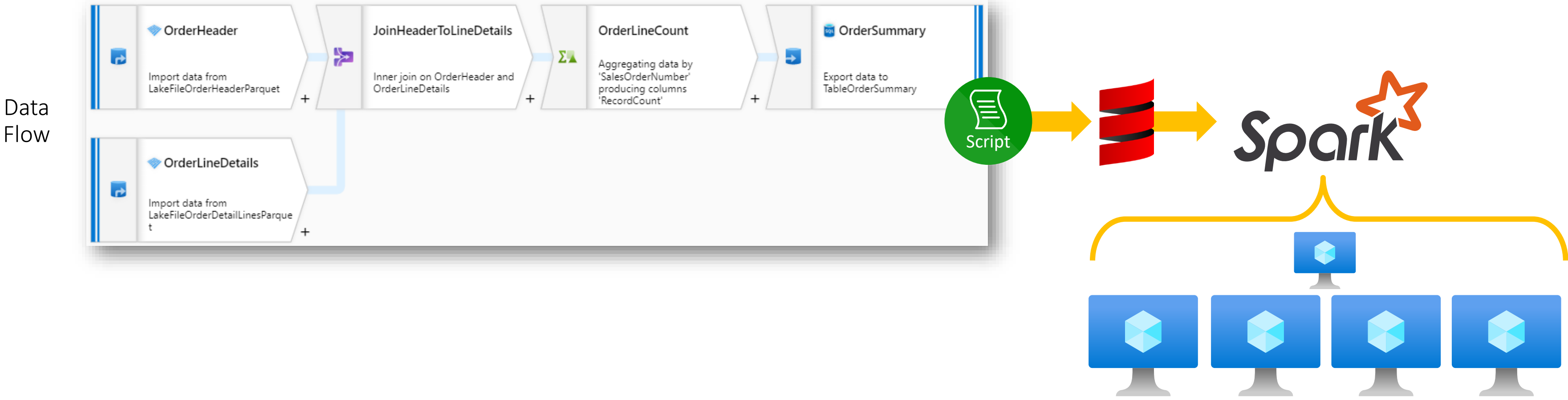
When Should We Use These Integration Pipeline Transformation Activities?



What is a Mapping Data Flow?



Q: What is a Mapping Data Flow?



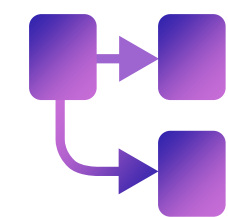
A: Graphic no low/low code data transformation tool that sits on top of Apache Spark.

Data Flows – Inputs & Outputs

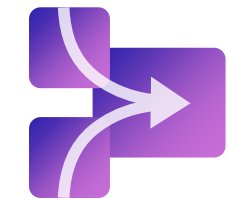


Source & Sink		Two blue cylinder icons representing data sources and sinks. The first cylinder has a white arrow pointing right, and the second has a grey arrow pointing right.
Linked Services		A row of ten icons representing various linked services: a blue folder with a lightning bolt, a green and grey database icon, a blue cylinder with 'SQL', a blue cylinder with an elephant head, a blue cylinder with 'My', a grey server rack with 'SQL', a blue globe with a ring, a grey house with a blue window, a blue cylinder, and a blue snowflake.
Source Types	Dataset A small icon of a grid with a black border.	A row of seven icons representing different dataset formats: a blue 'AVRO' logo with wings, a black circle with a green gear-like shape, a green 'X' logo, a grey document with a green 'X' and 'CSV', a purple document with a white '{ }' and 'JSON', a blue grid pattern, and a blue document with 'XML'.
	Inline A small icon of a square with diagonal lines.	A row of three icons representing inline data sources: a green 3D cube, a blue stylized 'A' shape, and a yellow bee with 'HIVE' text.

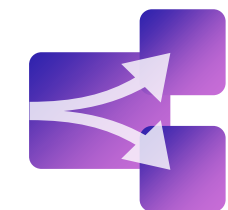
Data Flows – Transformations



New Branch



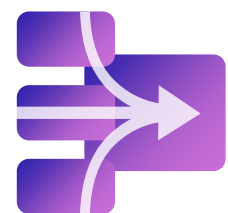
Join



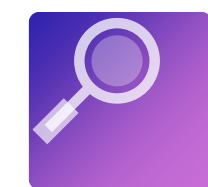
Conditional Split



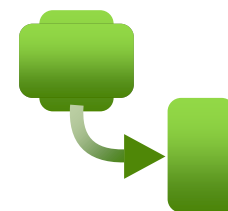
Exists



Union



Lookup



Derived Column



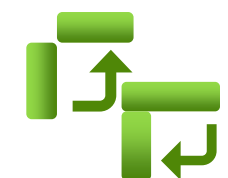
Select



Aggregate



Surrogate Key



Pivot/Unpivot



Window



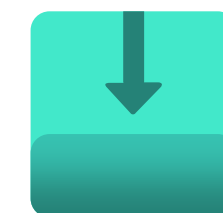
Rank



External Call



Cast



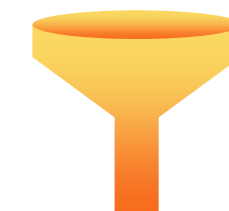
Flatten



Parse



Stringify



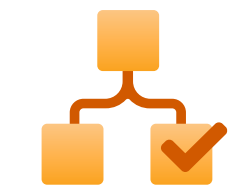
Filter



Sort



Alter Row



Assert



Flowlet

Key

Input & Output Modifiers

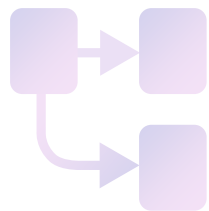
Schema Modifiers

Formatters

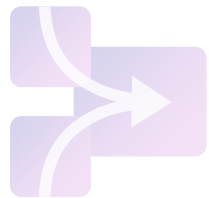
Row Modifiers

Data Flows – Transformations

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



New Branch



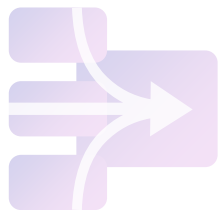
Join



Conditional Split



Exists



















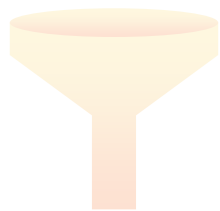
Union



Lookup

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>



Filter



Sort



Alter Row

Key

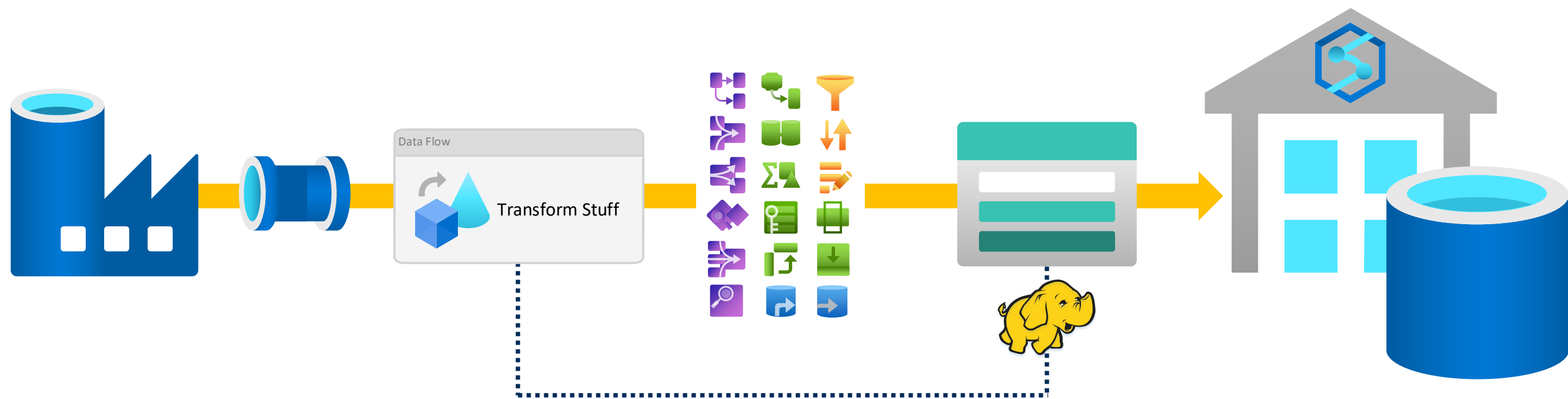
Input & Output Modifiers

Schema Modifiers

Formatters

Row Modifiers

Data Flows – Data Warehouse Loading (PolyBase)

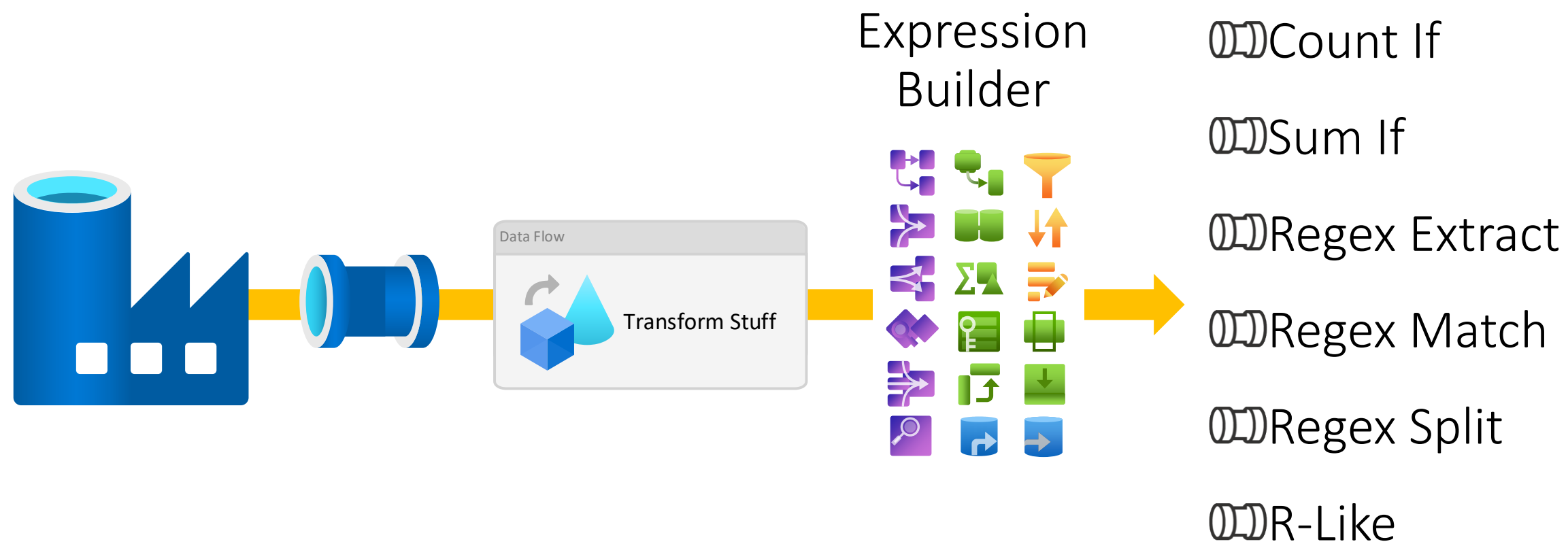


▲ PolyBase ⓘ

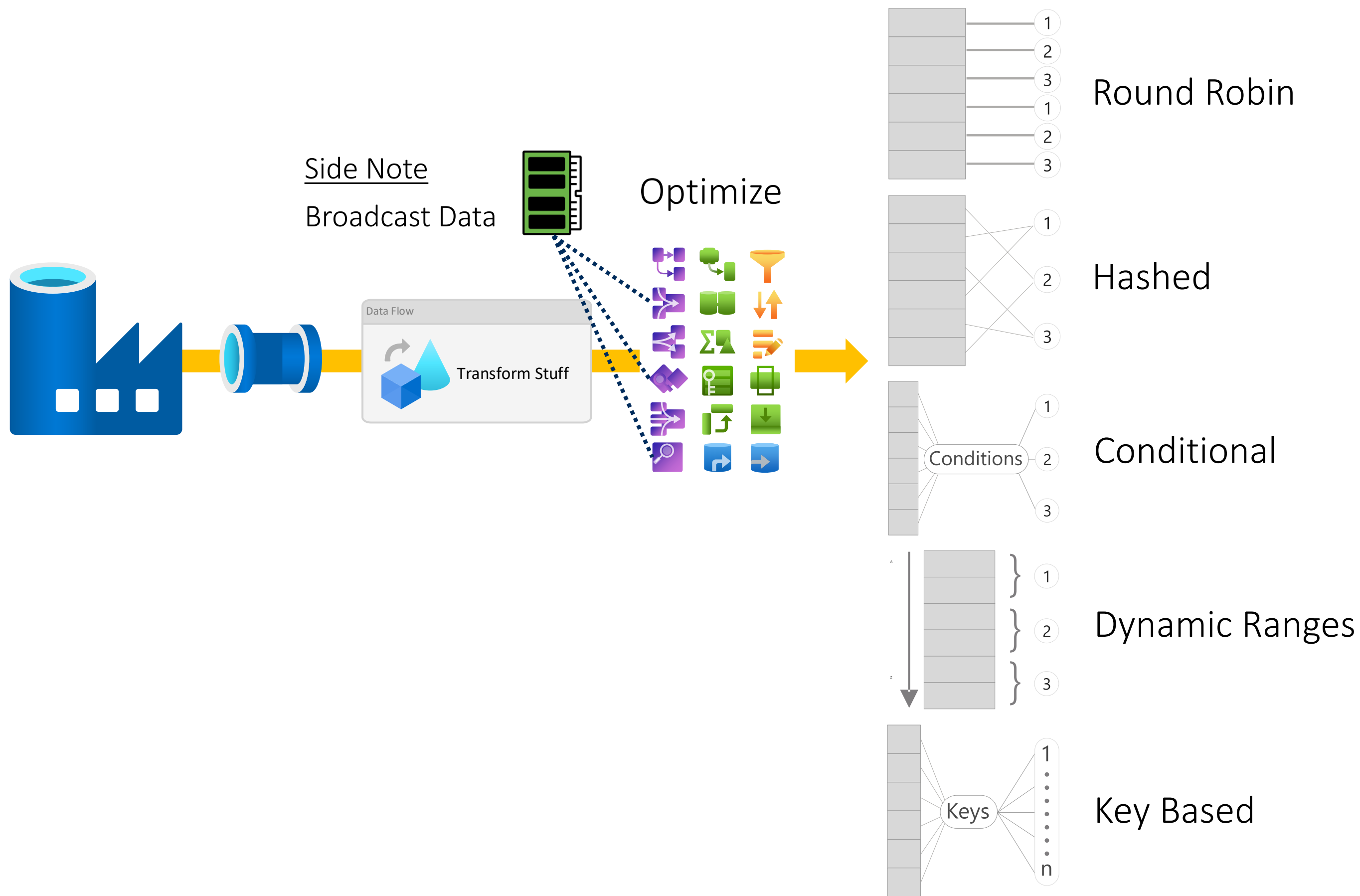
Staging linked service ⓘ + New

Staging storage folder / | ▼

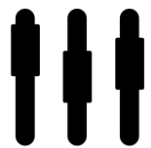
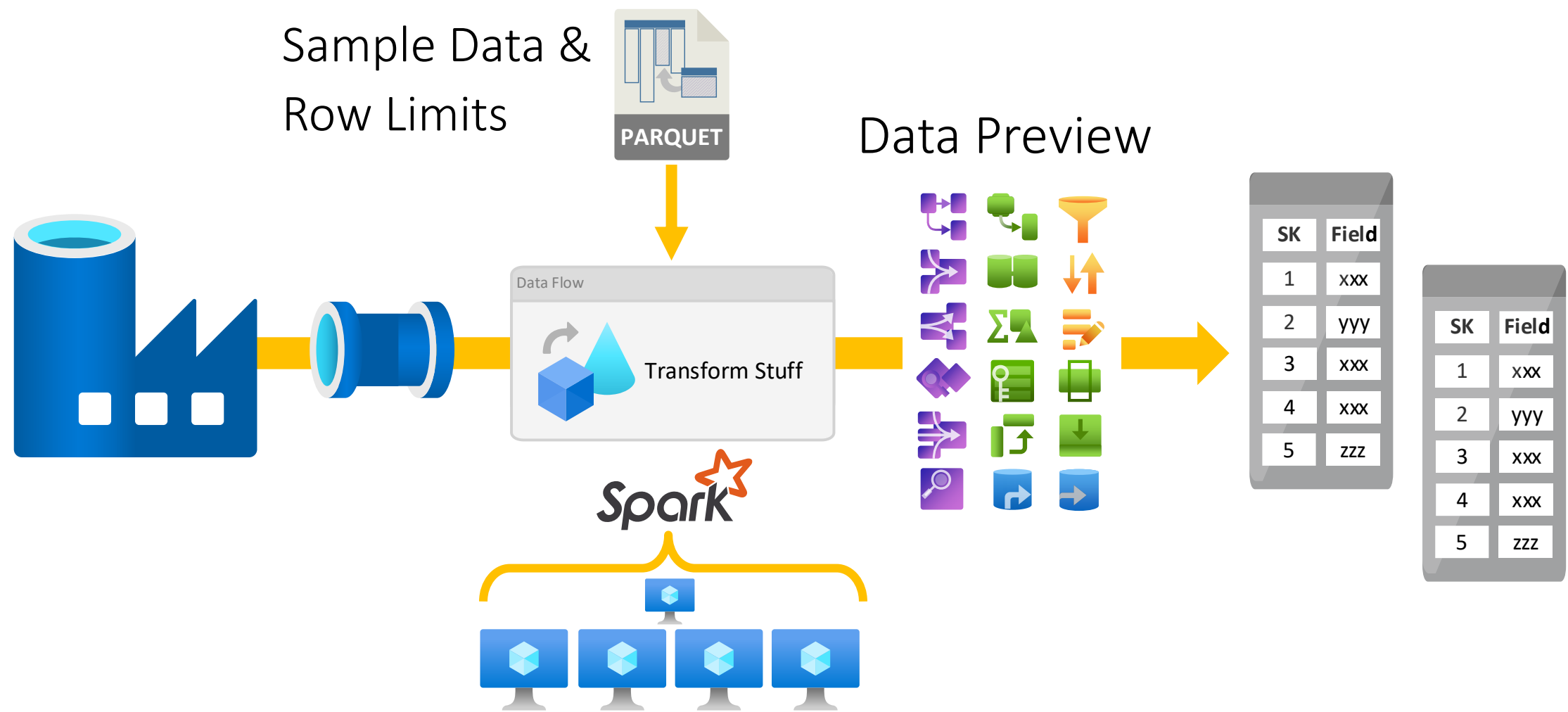
Data Flows – Expression Builder



Data Flows – Data Distribution

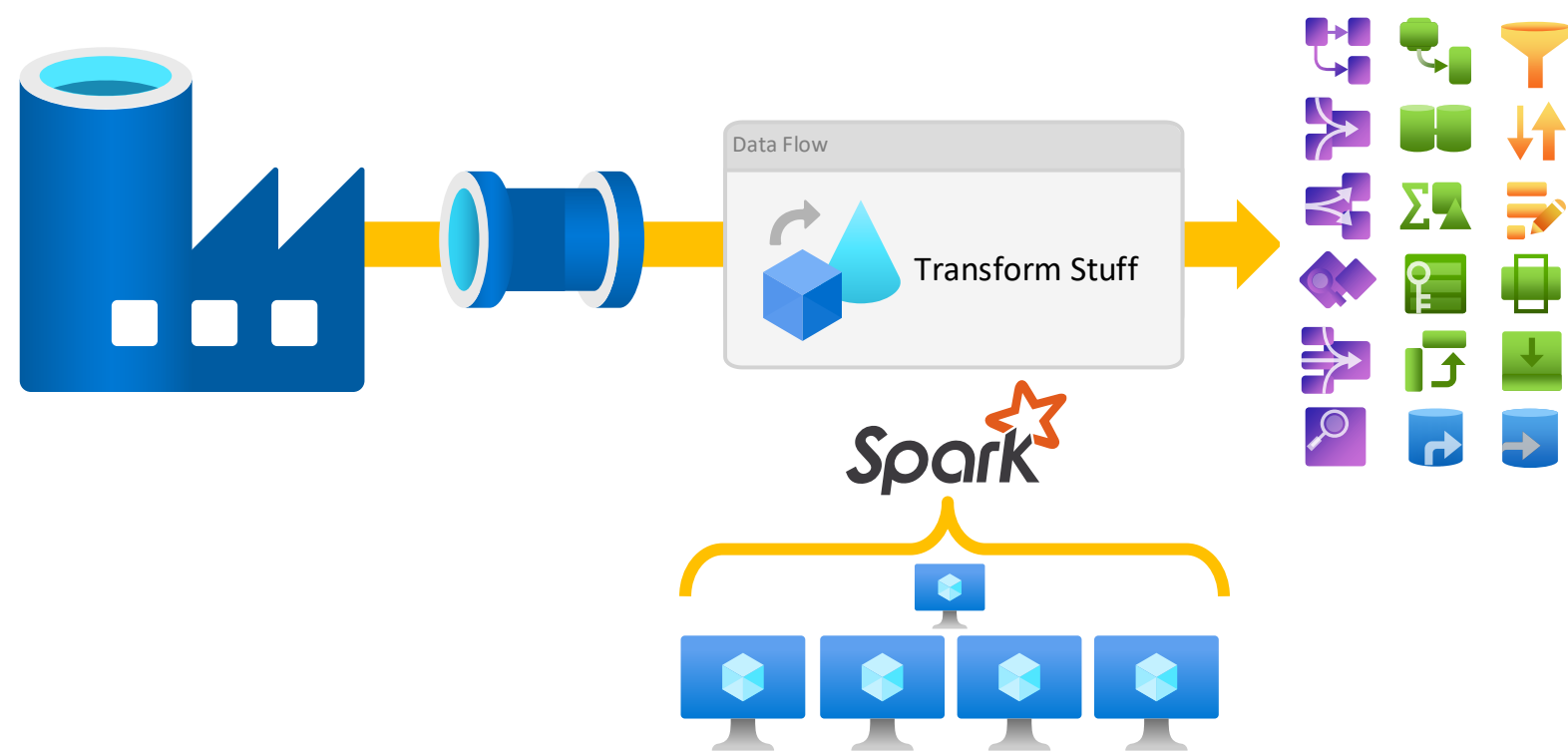


Data Flows – Debugging

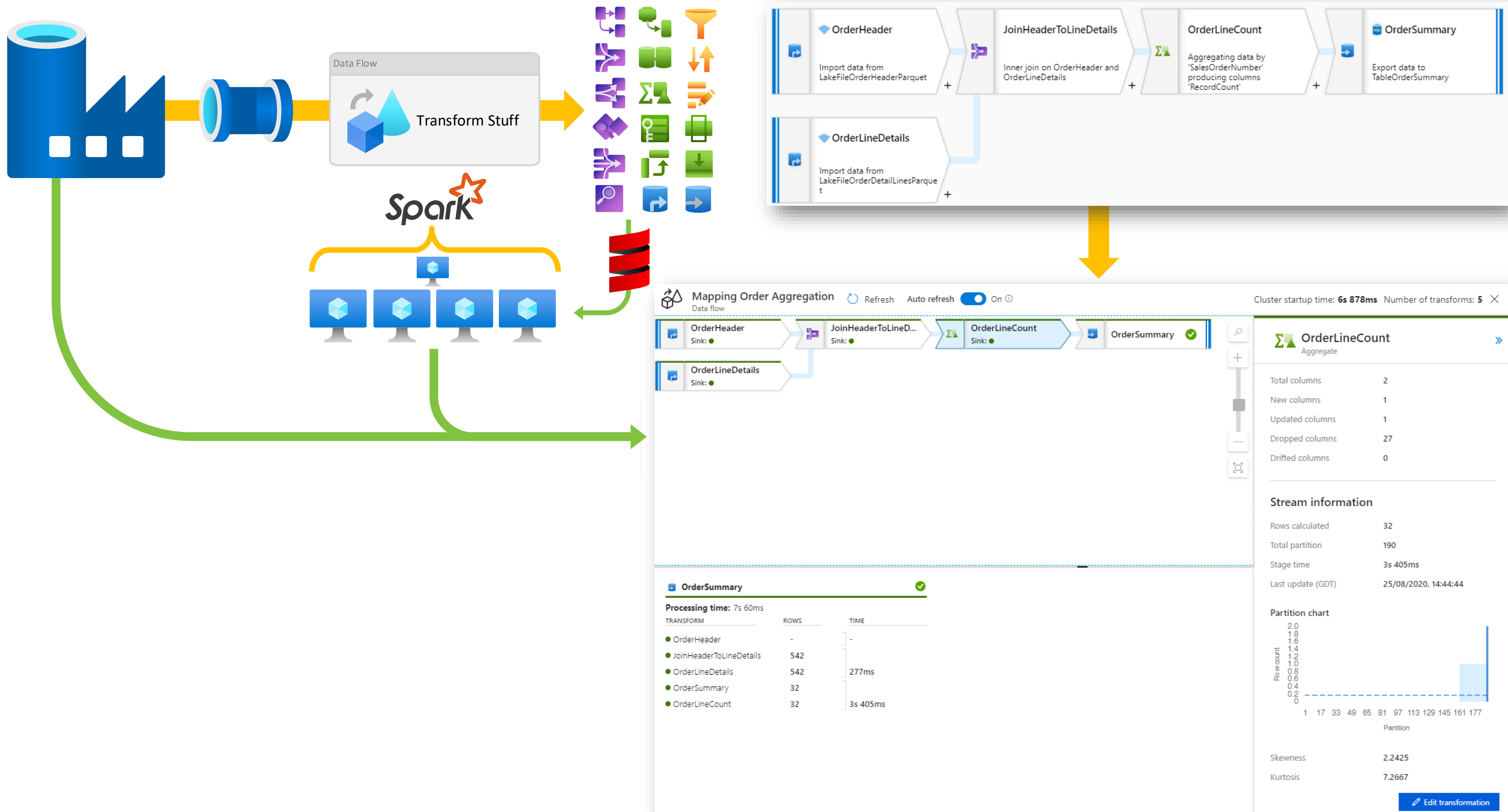


Enable Data Flow Debug Mode

Data Flows – Monitoring



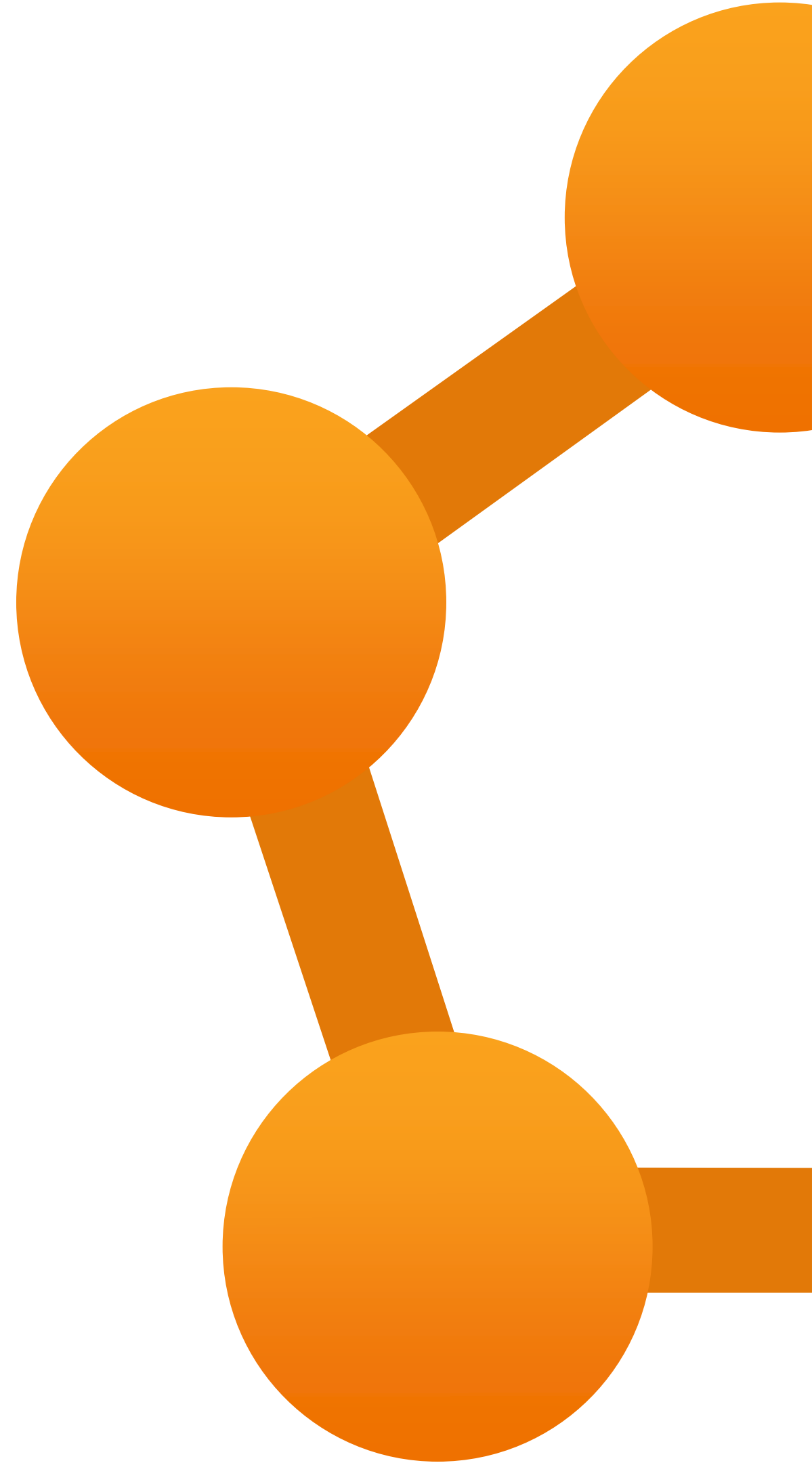
Data Flows – Monitoring



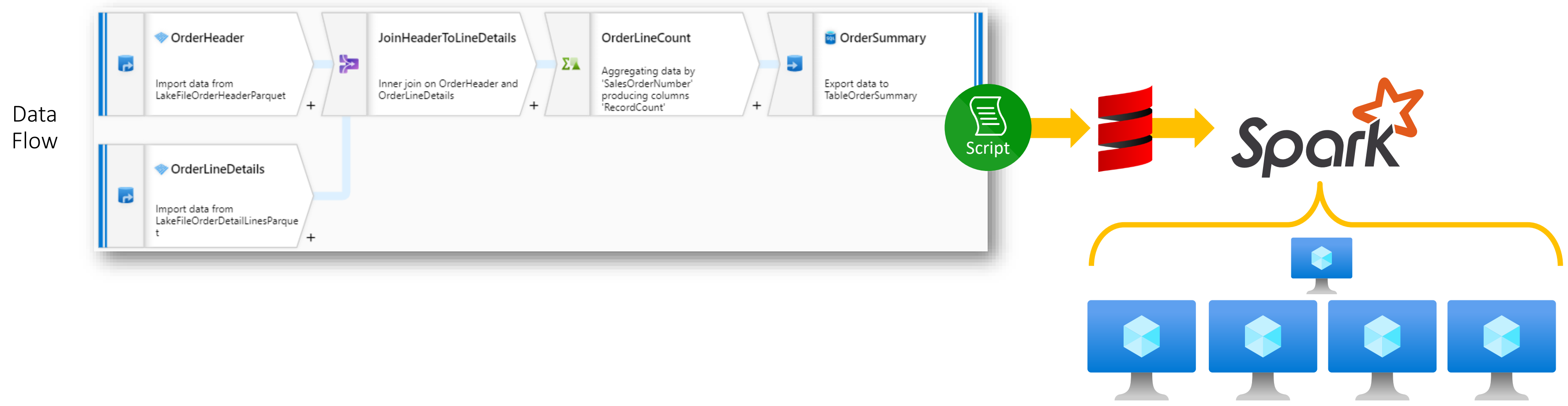
Module 3 – Data Transformation

Power Query

Cloud Formations



What is a Data Flow?



What is a Power Query Activity?



What is a Power Query Activity?



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

Parquet.Document (AdfDoc)

	1.2 SalesOrderID	1.2 SalesOrderDetailID	1.2 OrderQty	1.2 ProductID	1.2 UnitPrice	1.2 UnitPriceDiscount	1.2 LineTotal	ABC 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

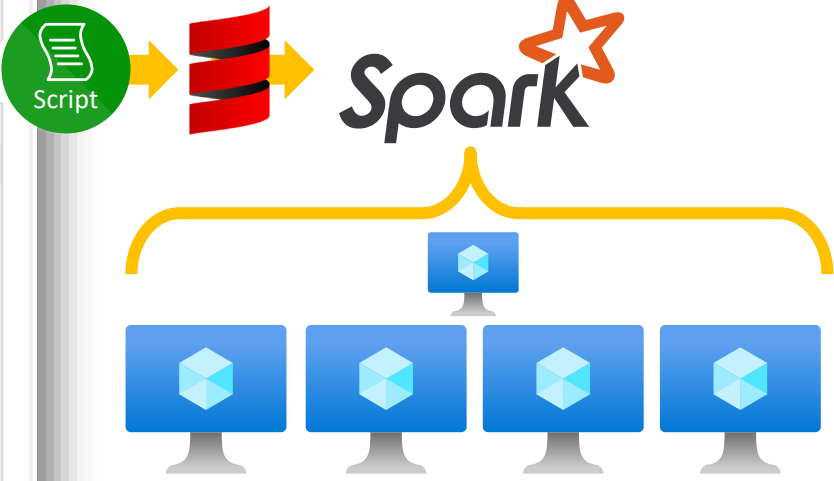
Name

LakeFileOrderDetailLinesP...

Applied steps

AdfDoc

Parquet



What can a Power Query Activity do?



Power Query

Power Query interface showing a table of data with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, UnitPrice, UnitPriceDiscount, LineTotal, and rowguid. The interface includes a ribbon with tabs: Home, Transform, Add column, and View. The ribbon contains various icons for data manipulation, such as 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.

	1.2 SalesOrderID	1.2 SalesOrderDetailID	1.2 OrderQty	1.2 ProductID	1.2 UnitPrice	1.2 UnitPriceDiscount	1.2 LineTotal	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 Power Query Activity do?



Home

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

Replace values

Use first row as headers

Replace values

Merge queries

Append queries

Combine files

Queries

ADFSResource [1]

LakeFileOrderDetailL...

UserQuery

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

123 SalesOrderID

123

71774

71774

71776

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

File

Home

Transform

Add Column

View

Tools

Help

Close & Apply

New Source

Recent Sources

Enter Data

Data source settings

Manage Parameters

Refresh Preview

Manage

Choose Columns

Remove Columns

Keep Rows

Remove Rows

Sort

Split Column

Group By

Replace Values

Use First Row as Headers

Replace Values

Merge Queries

Append Queries

Combine Files

Queries [1]

OrderDetailLines

123 SalesOrderID

123 SalesOrderDetailID

123 OrderQty

123 ProductID

1.2 UnitPrice

1.2 UnitPrice

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

71774

71774

71776

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

71780

110562

110563

110567

110616

110617

110618

110619

110620

110621

110622

110623

110624

110625

110626

110627

110628

110629

1

1

1

4

2

6

2

1

1

1

4

2

4

1

6

1

2

836

822

907

905

983

988

748

990

926

743

782

918

780

937

867

985

989

356.898

356.898

63.9

218.454

461.694

112.998

818.7

323.994

149.874

809.76

1376.994

158.43

1391.994

48.594

41.994

112.998

323.994

Query Settings

PROPERTIES

Name

OrderDetailLines

APPLIED STEPS

Source

Promoted Headers

Changed Type



What can a Power Query Activity do?



Transform



Power Query

Power Query Editor interface showing the Transform tab and a data table.

Transform Tab Options:

- Table: Group by, Use first row as headers, Transpose, Reverse rows, Count rows, Replace values, Detect data type, Mark as key, Rename, Pivot column, Unpivot columns, Move, Convert to list, Split column, Format, Merge columns, Extract, Parse, Statistics, Standard, Scientific, Rounding, Information, Date, Time, Duration.
- Any Column: Split column, Format, Merge columns, Extract, Parse, Statistics, Standard, Scientific, Rounding, Information, Date, Time, Duration.
- Text Column: Split column, Format, Merge columns, Extract, Parse, Statistics, Standard, Scientific, Rounding, Information, Date, Time, Duration.
- Number Column: Split column, Format, Merge columns, Extract, Parse, Statistics, Standard, Scientific, Rounding, Information, Date, Time, Duration.
- Date & Time Column: Split column, Format, Merge columns, Extract, Parse, Statistics, Standard, Scientific, Rounding, Information, Date, Time, Duration.

Queries [1]:

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

Formula Bar: = Table.TransformColumnTypes(#"Promoted Headers",{{"SalesOrderID", Int64.Type}, {"SalesOrderDetailID", Int64.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	

Query Settings:

- NAME: OrderDetailLines
- APPLIED STEPS: Source, Promoted Headers, Changed Type

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

- 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



Power Query

Home Transform Add column View

Data view Schema view Go to column Advanced editor

Preview Columns Advanced

Queries

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

	SalesOrderID
1	71774
2	71774
3	71776
4	71780
5	71780
6	71780
7	71780
8	71780
9	71780
10	71780
11	71780
12	71780
13	71780
14	71780
15	71780
16	71780
17	71780

File Home Transform Add Column View Tools Help

Formula Bar

Monospaced Column distribution

Show whitespace Column profile

Column quality

Layout

Data Preview

Columns Parameters Advanced Dependencies

Queries [1]

OrderDetailLines

SalesOrderID

SalesOrderDetailID

OrderQty

ProductID

UnitPrice

UnitPrice

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	

Query Settings

PROPERTIES

Name

- OrderDetailLines

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type

What can a Power Query Activity do?



View



Power Query

The screenshot shows the 'Advanced editor' window in Power Query. The window has a ribbon with 'Home', 'Transform', 'Add column', and 'View' tabs. The 'Advanced editor' tab is selected. The main area displays a M query script:

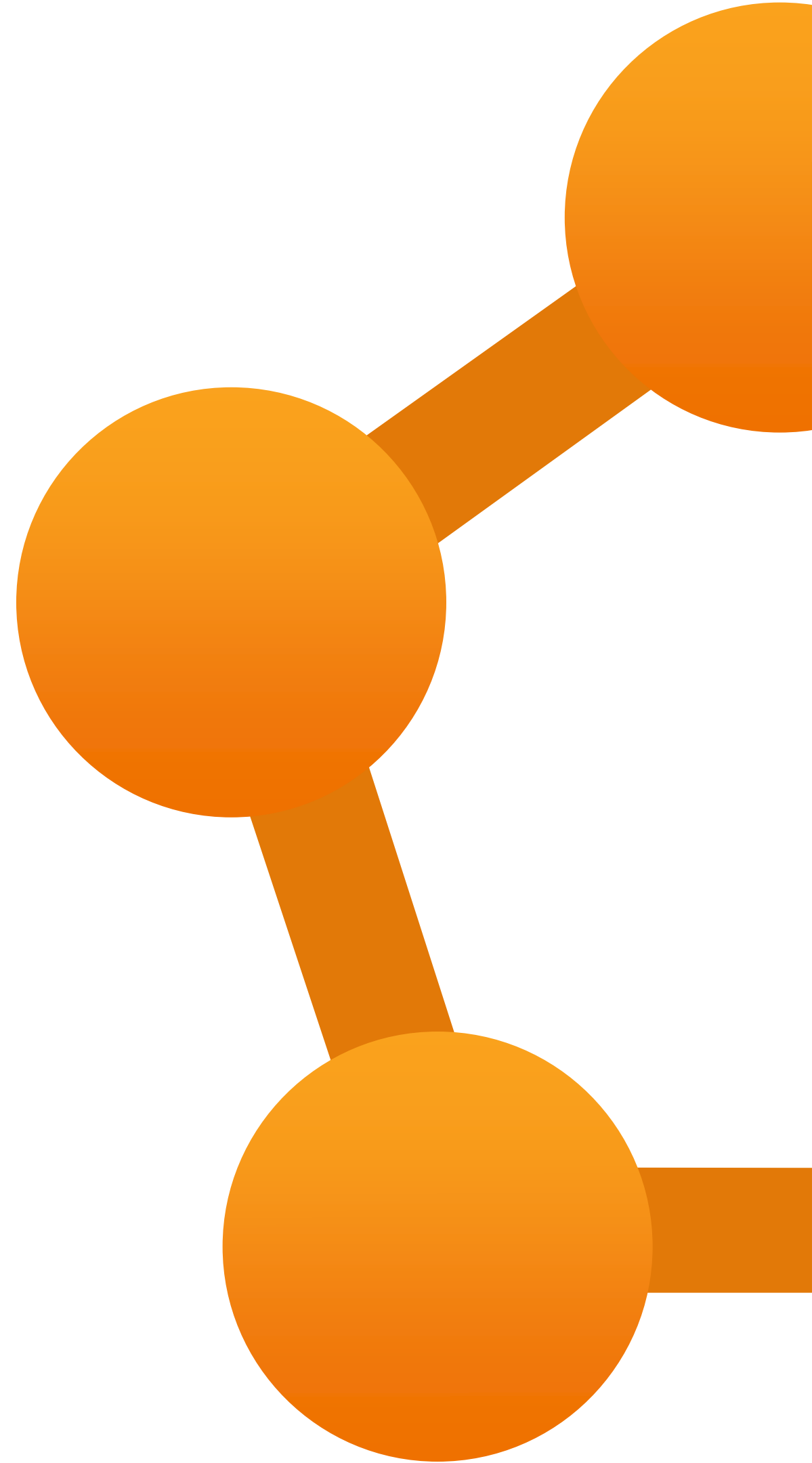
```
1 let
2   AdfDoc = Web.Contents("https://traininglake01.dfs.core.windows.net/datawarehouse/Raw/OrderDetailLines.parquet"),
3   Parquet = Parquet.Document(AdfDoc),
4   #"Grouped rows" = Table.Group(Parquet, {"SalesOrderID"}, {"Count", each Table.RowCount(_), Int64.Type})
5 in
6   #"Grouped rows"
```

The script is written in a light blue font on a white background. The 'Advanced editor' window is overlaid on the main Power Query interface, which shows a list of queries on the left and a data preview on the right.

Module 3 – Data Transformation

Spark Cluster Configuration

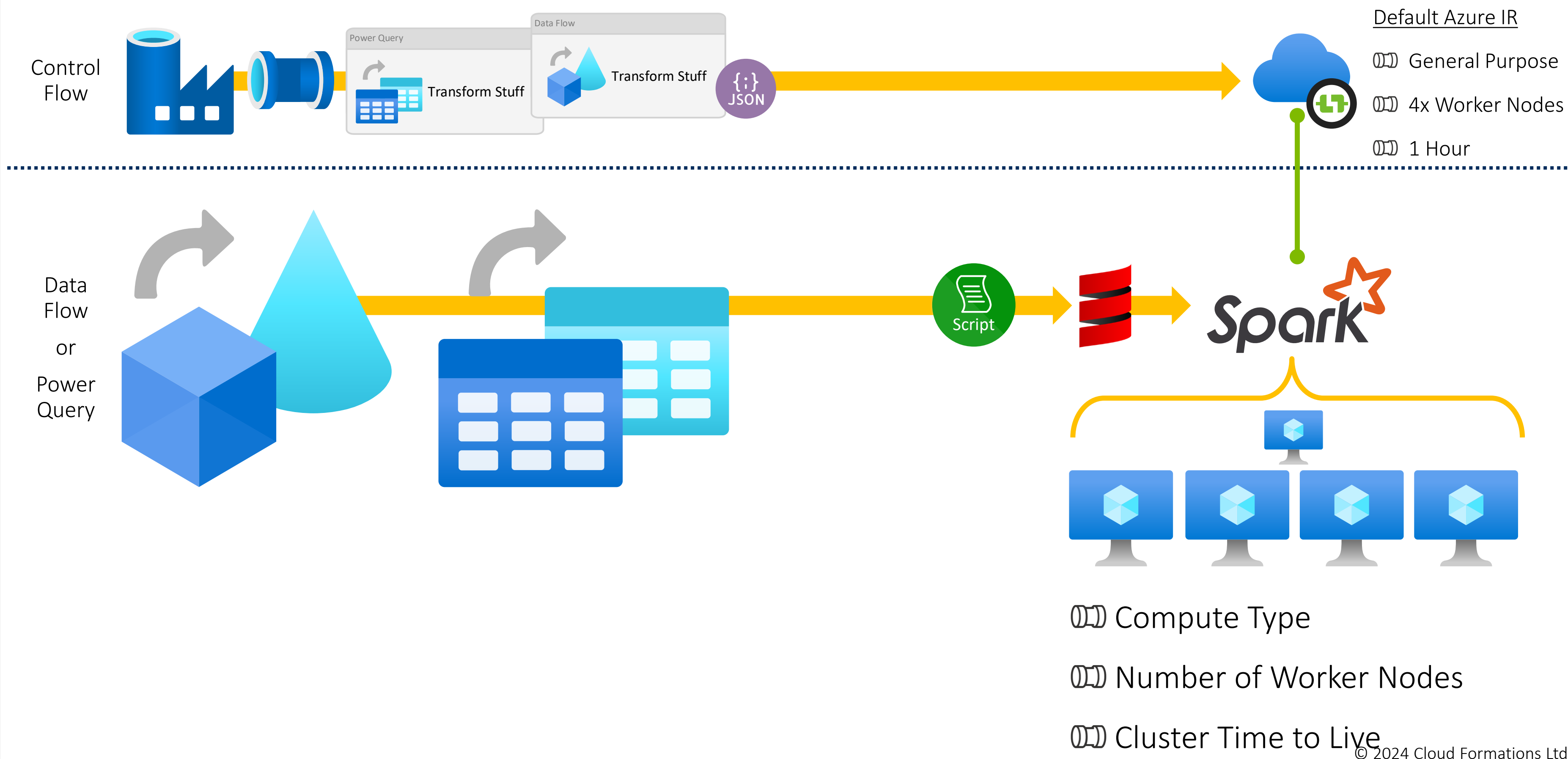
Cloud Formations



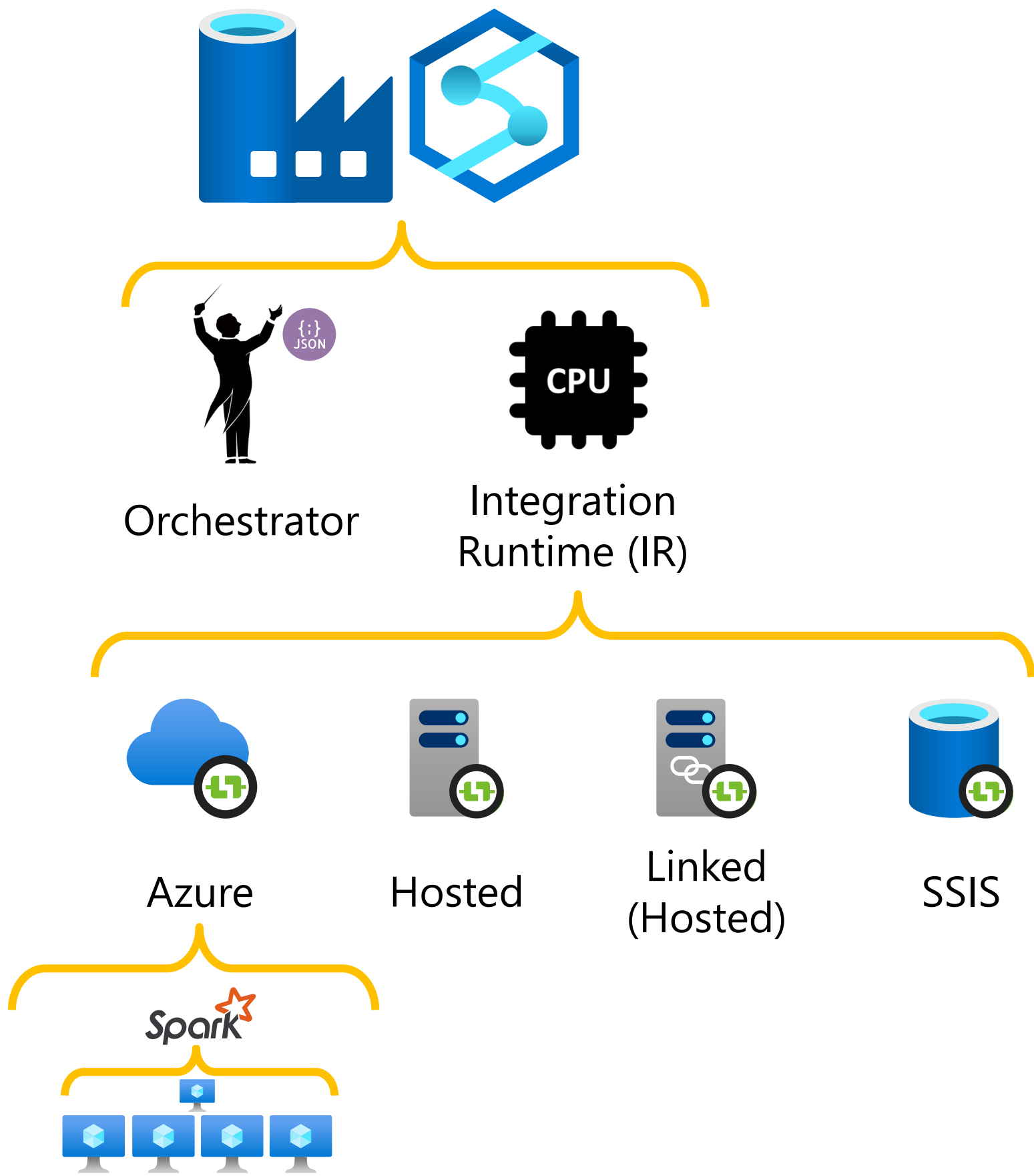
Spark Configuration



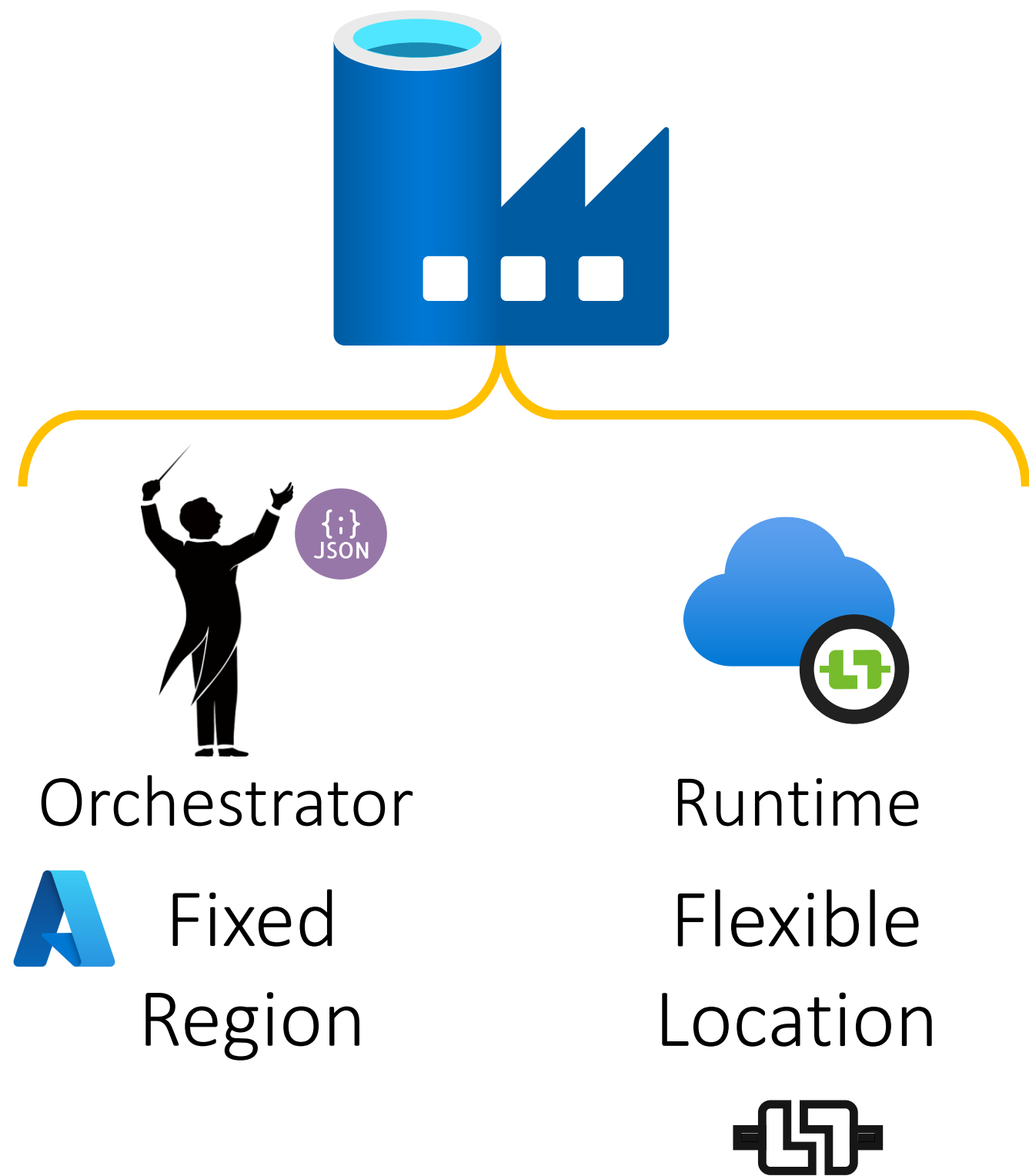
Cloud Formations - Knowledge Transfer & Training



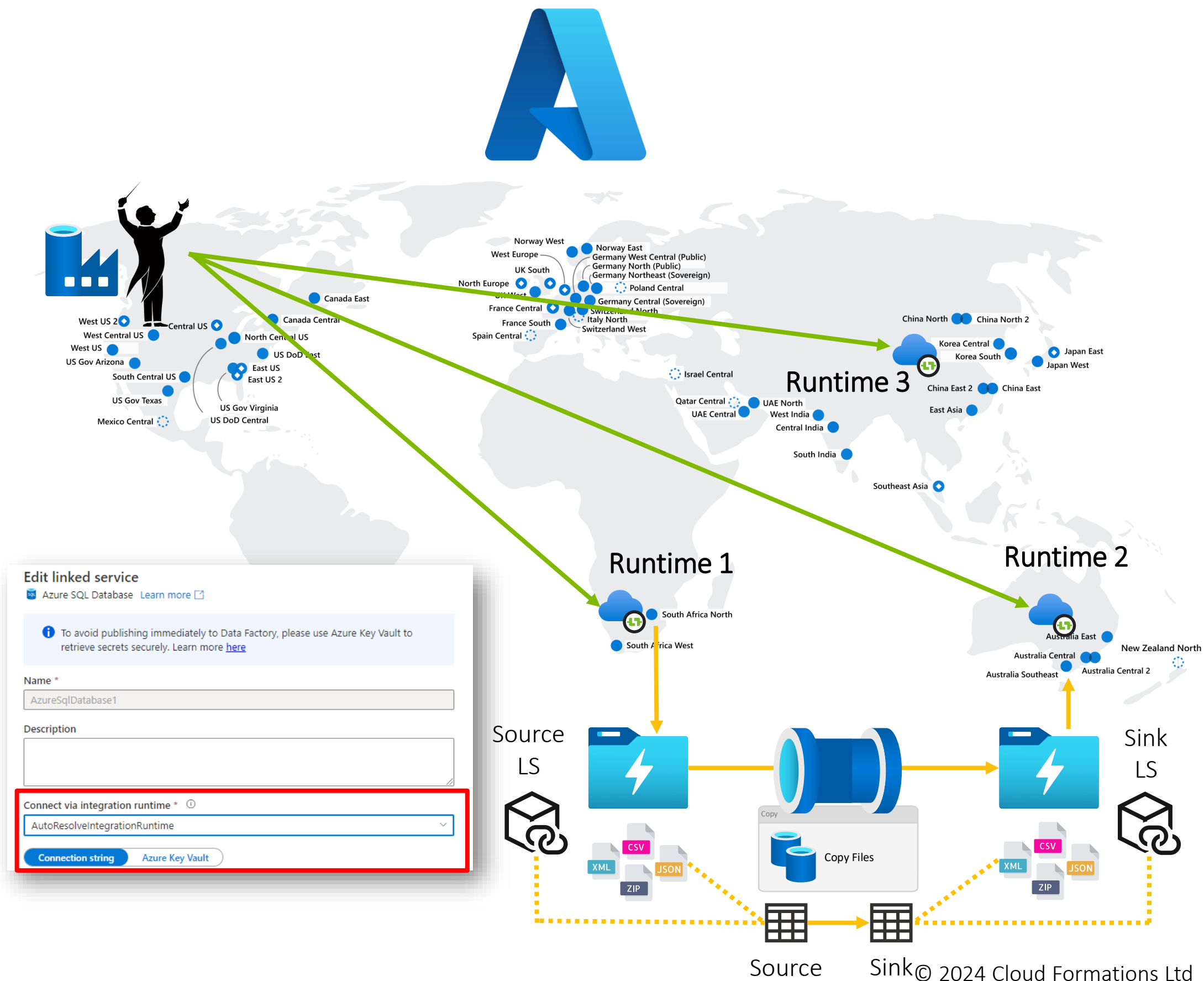
Data Flow Compute – IR's vs Spark



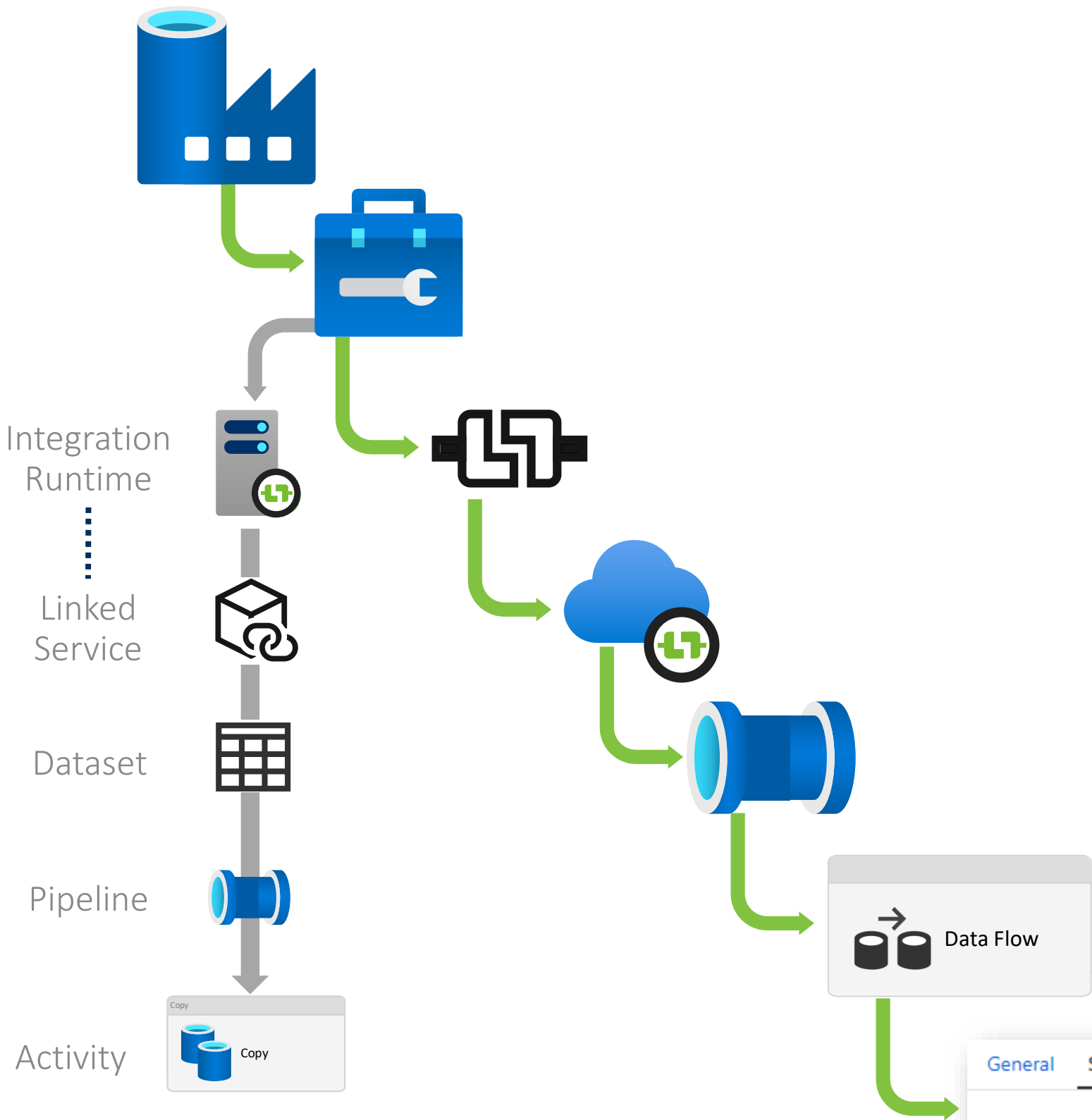
What is an Integration Runtime?



AutoResolveIntegrationRuntime



Setting the Data Flow Cluster (IR Configuration)

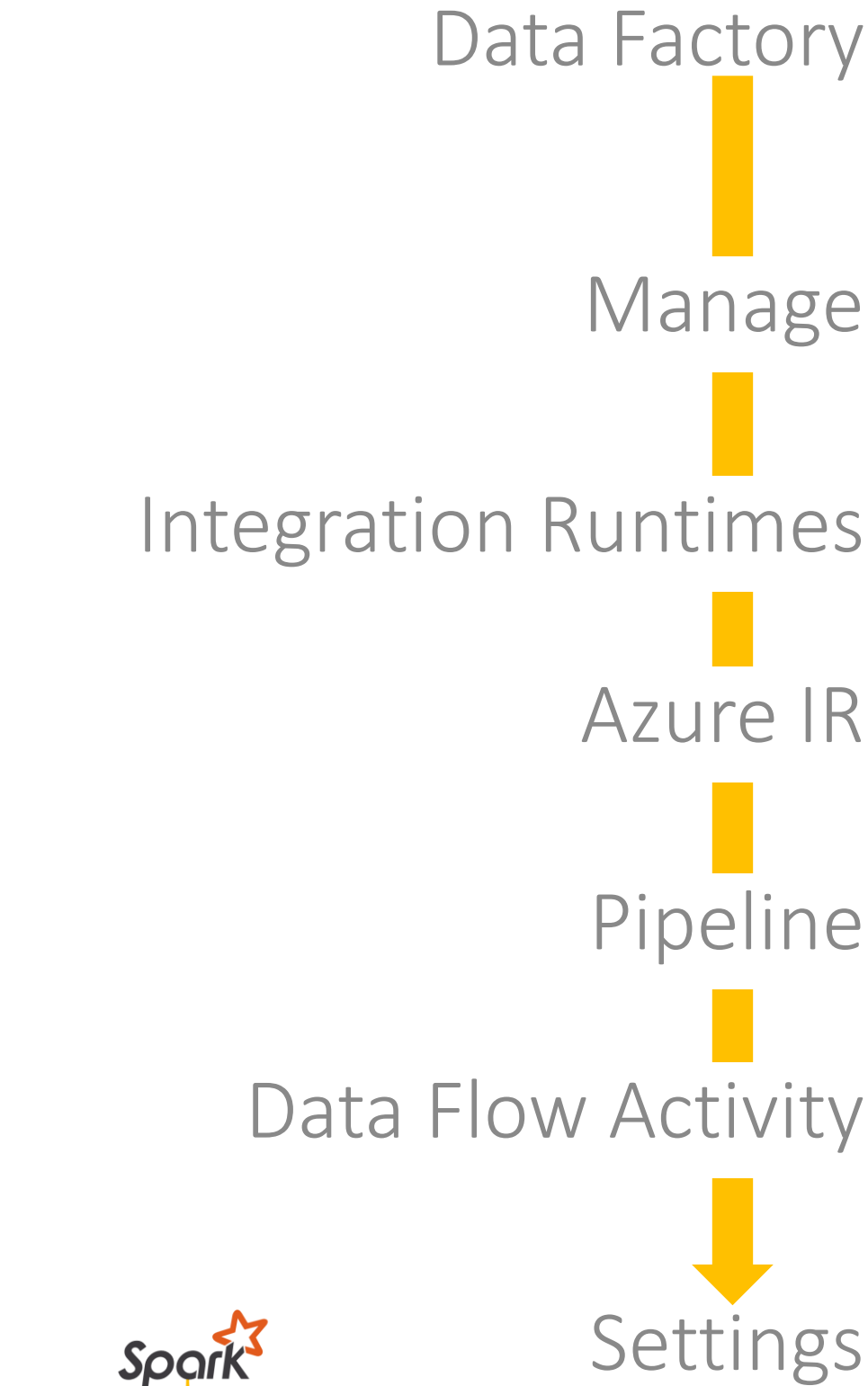


General Settings Parameters User properties

Data flow * MappingOrderAggregation

Run on (Azure IR) * DataFlowDemosTTL4Hours ⓘ

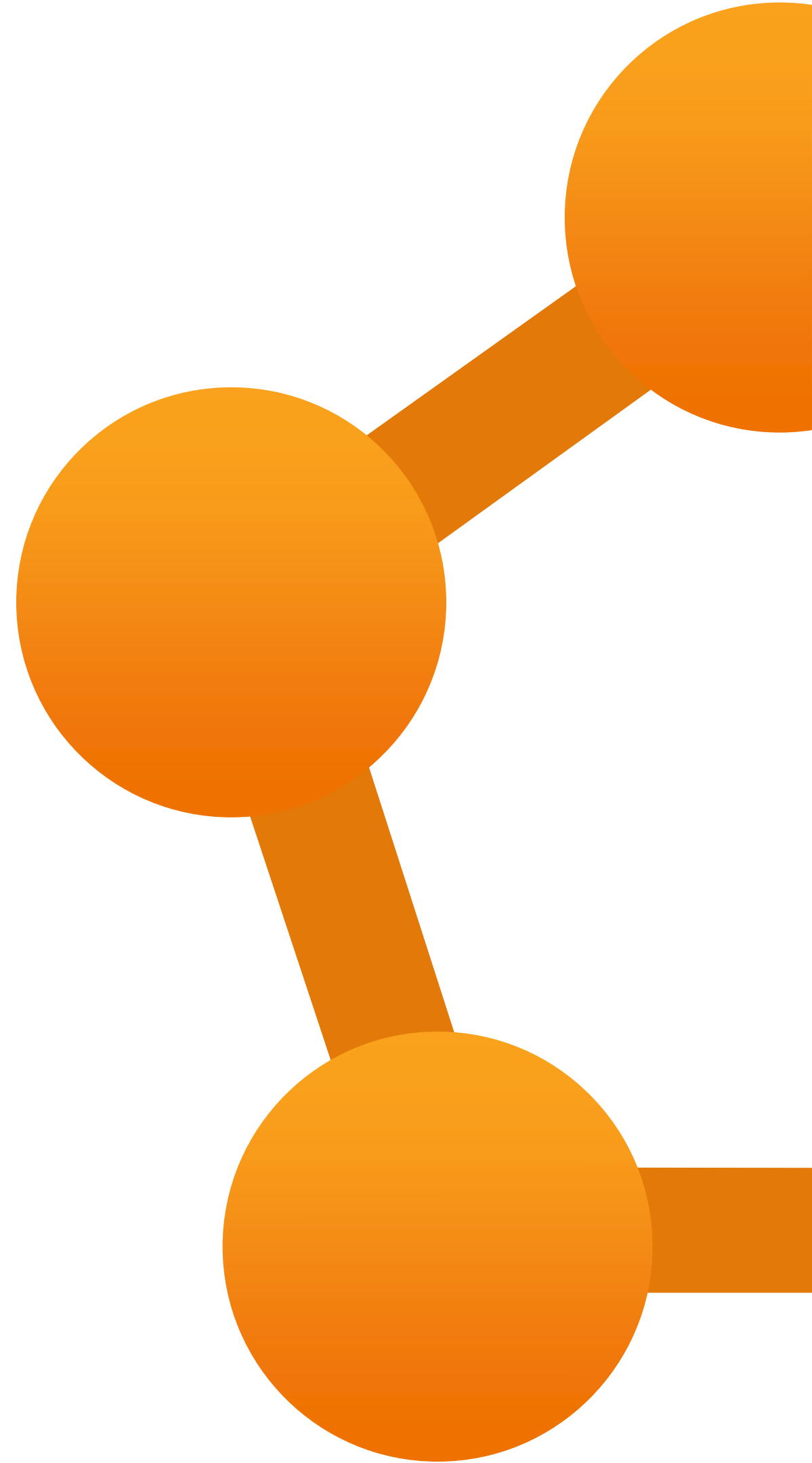
PolyBase ⓘ



Module 3 – Data Transformation




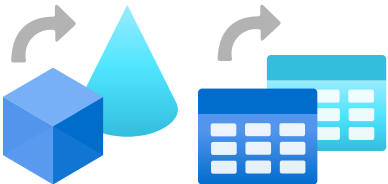
Use Cases

Cloud Formations



Data Transformation Resources in Azure Comparison

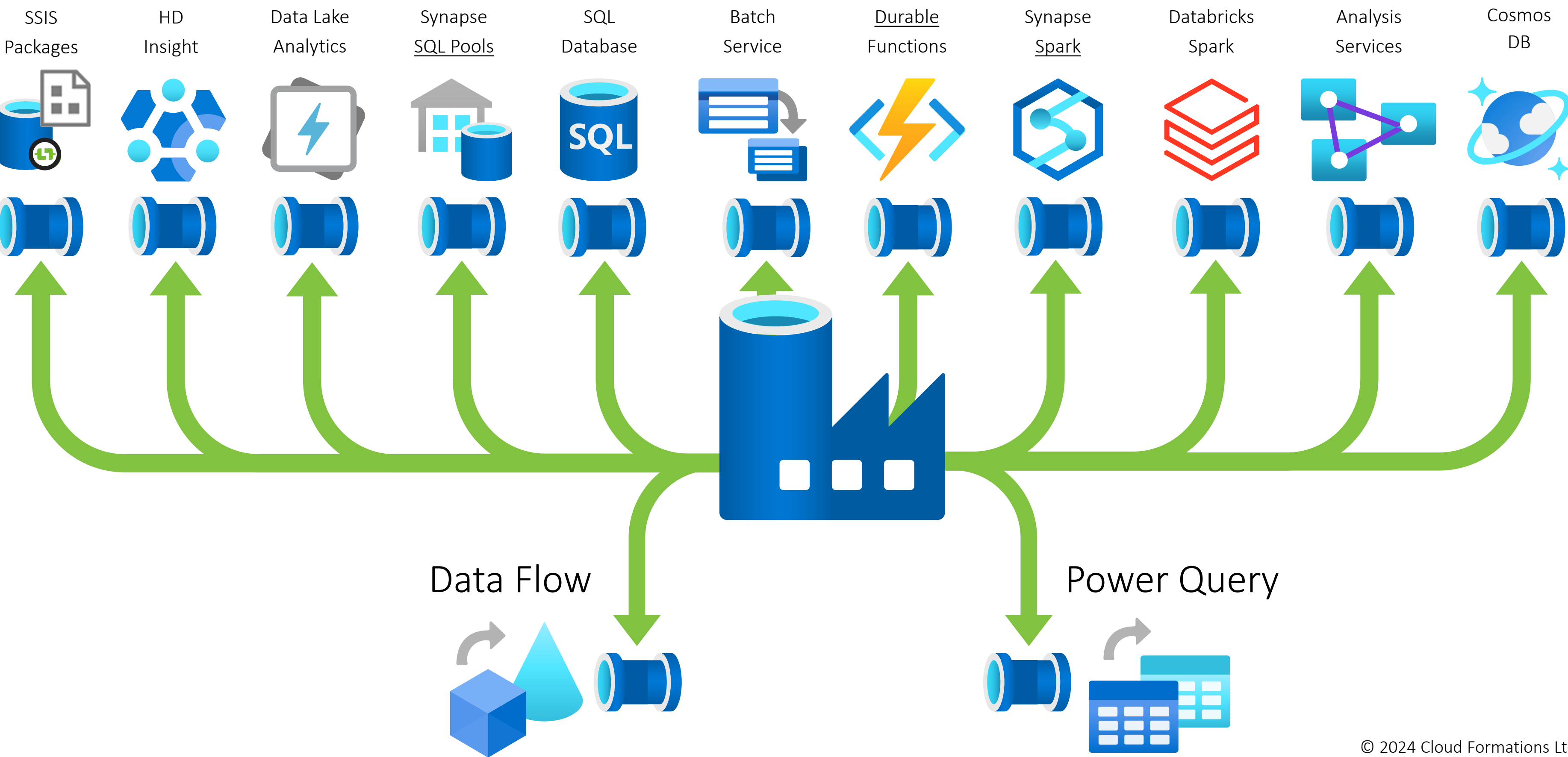


Transformation Tools		Graphical UI (Low/No Code)	Scales Out	Scales Up	Cloud Native Tech
	T-SQL with SQLDB	✗	✗	✓	✗
	SSIS Packages	✓	✗	✓	✗
	Scala/Python/SQL with Databricks	✗	✓	✓	✓
	Data Flows & Power Query	✓	✓	✓	✓

Other Data Transformation Services in Azure



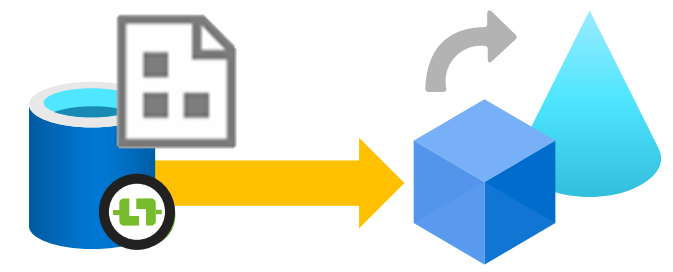
When Should We Use These Integration Pipeline Transformation Activities?



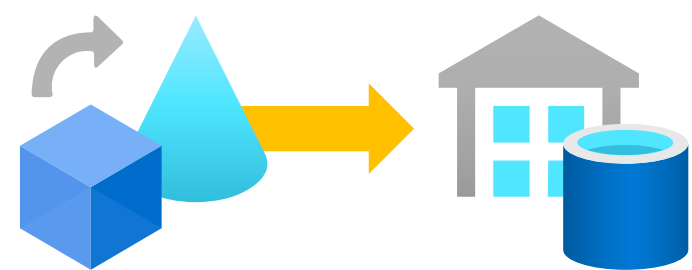
Use Cases



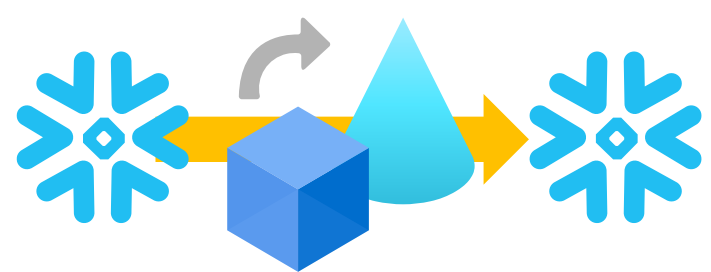
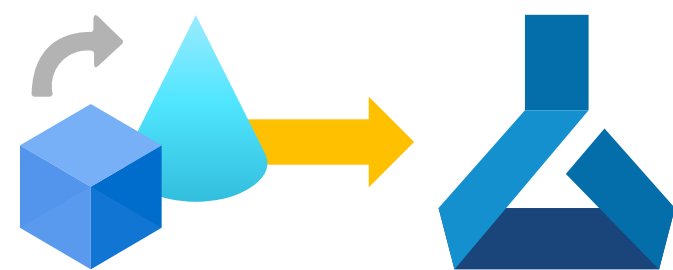
SSIS Package rebuild
and skills migration.



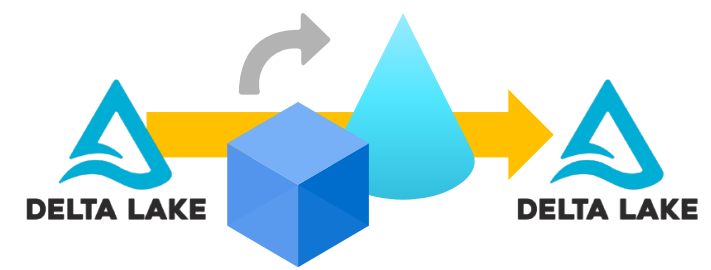
Warehouse data
distribution & loading.



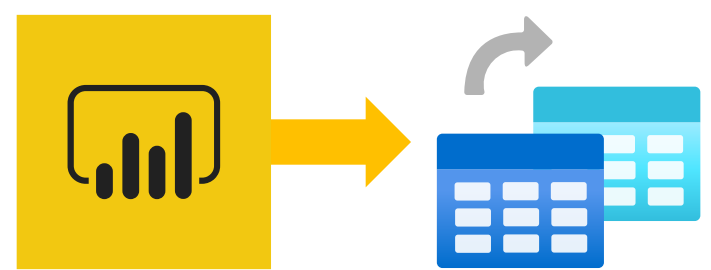
Data model dataset
preparation.



Inline dataset
transformations.



Power Query
industrialisation.



Module 3

Data Transformation ✓

Any questions?

Cloud Formations

