

ETL in Azure Made Easy

with Data Factory Data Flows



Paul Andrew

Principal Consultant & Solution Architect



altius



@MrPaulAndrew



<https://github.com/mrpaulandrew>


CommunityEvents

Demo code, content and slides from various community events.

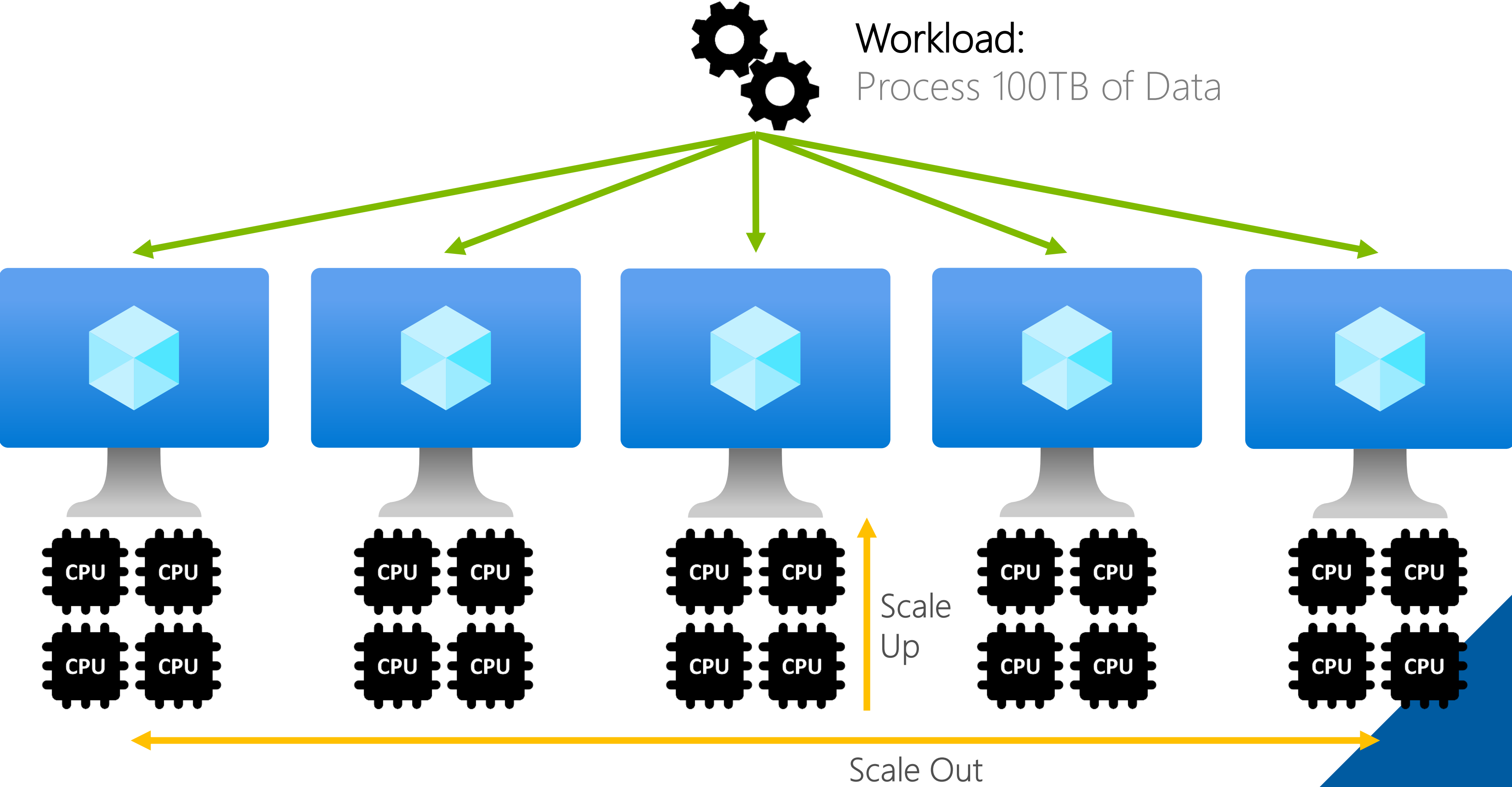
● C++

[{Event/Location}-{Month}-{Year}](#)

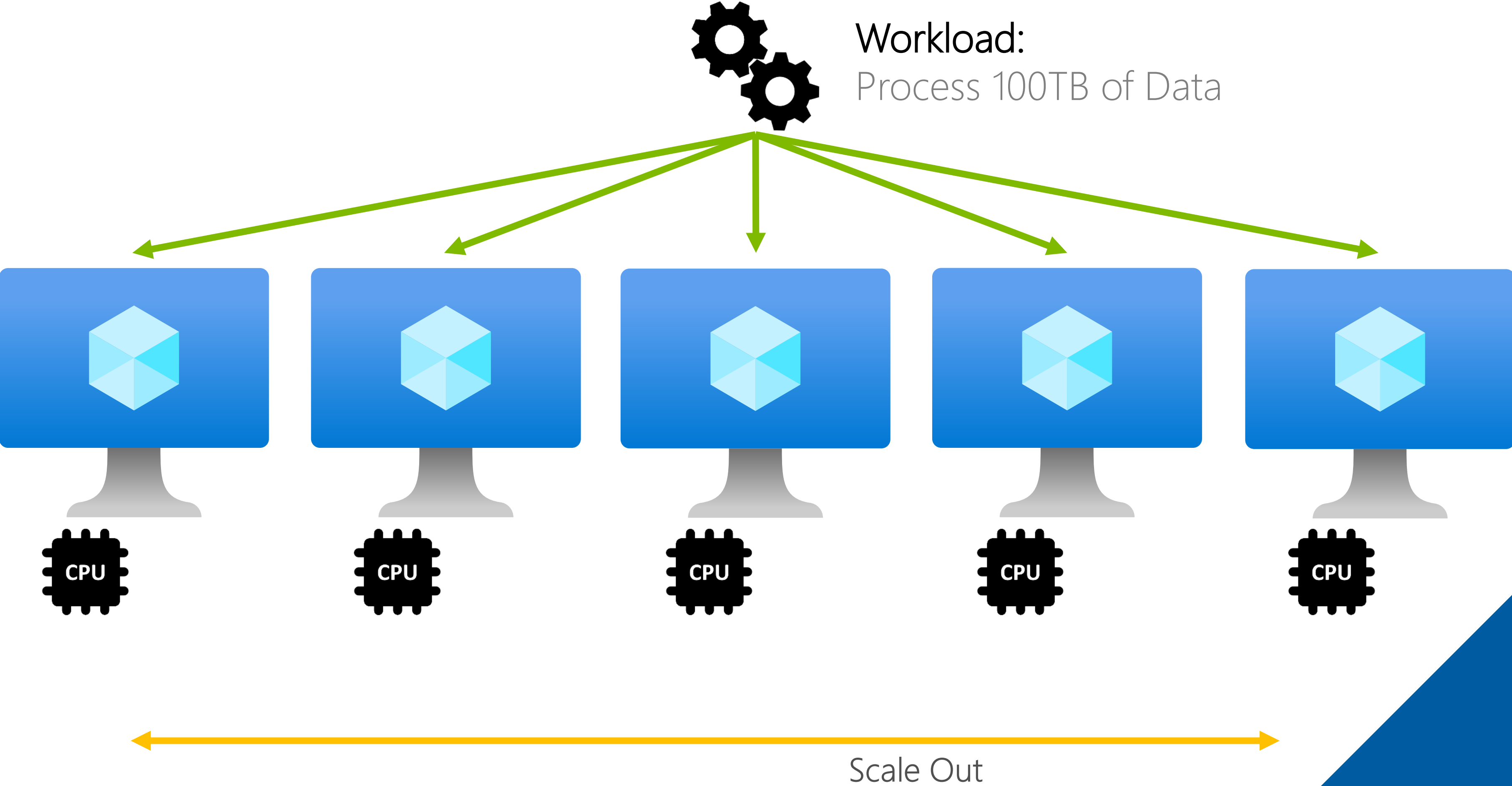
Scaling Up vs Scaling Out



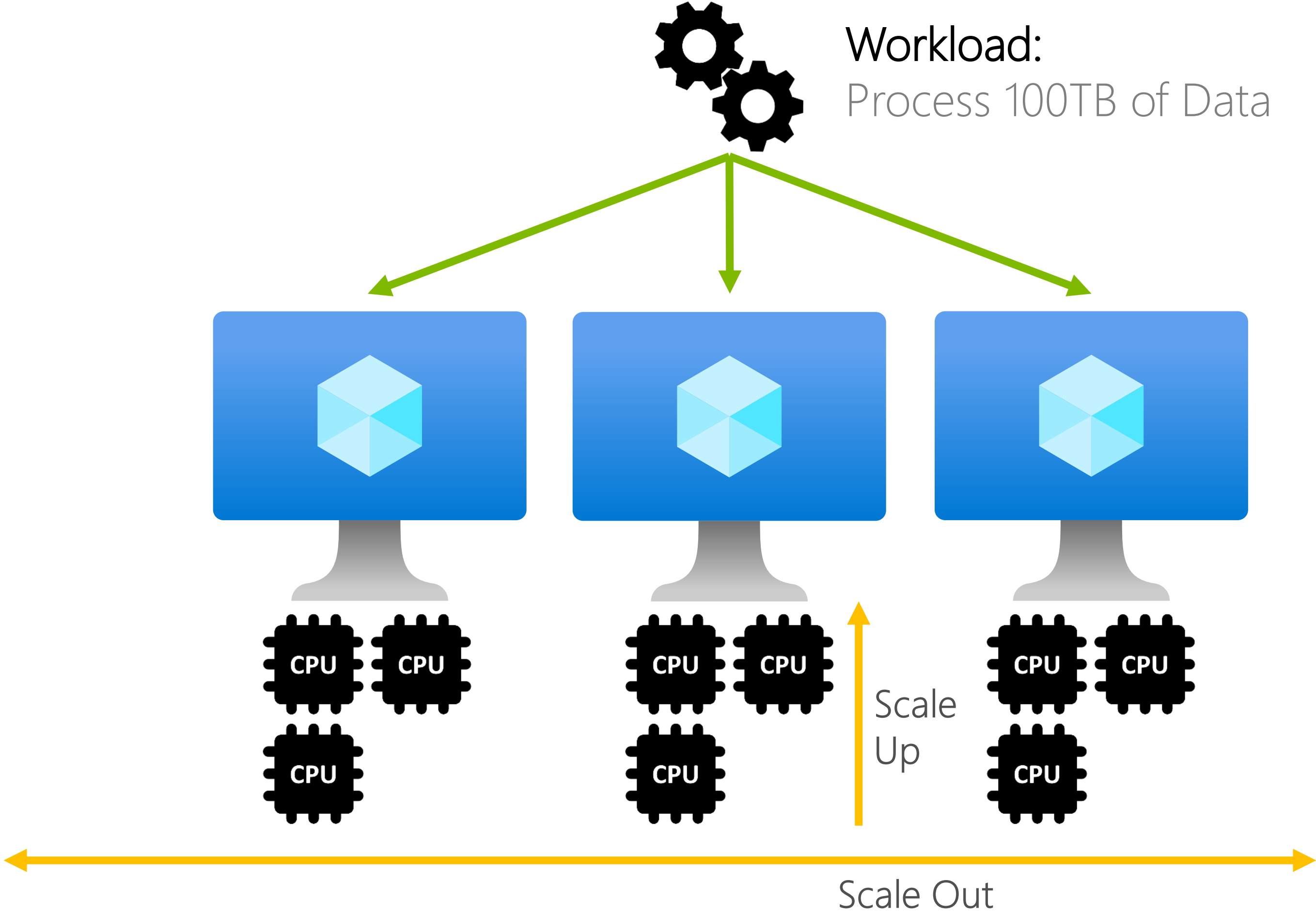
Scaling Up and/or Scaling Out



Scaling Up and/or Scaling Out



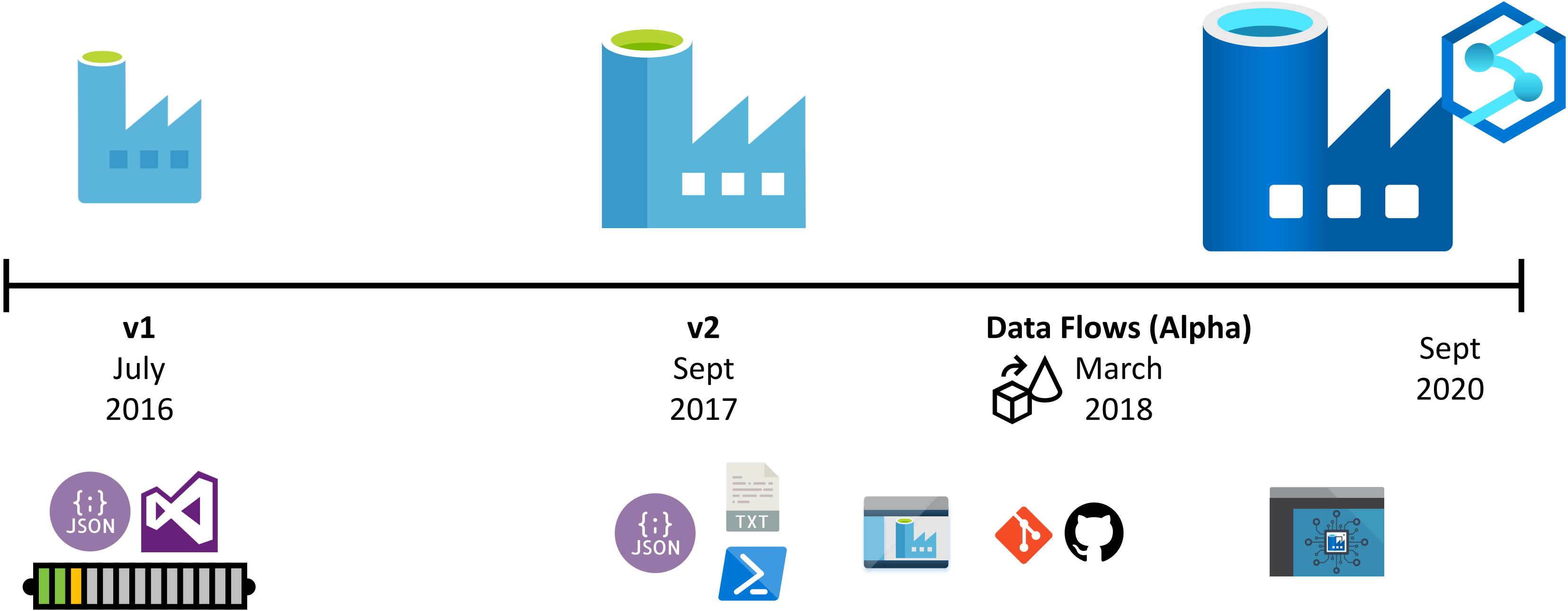
Scaling Up and/or Scaling Out



Azure Data Factory –

What is it?
Why use it?

A Quick History Lesson



What is Azure Data Factory (ADF)?

[Home](#) / [Products](#) / Data Factory

Data Factory

Hybrid data integration service that simplifies ETL at scale

Start for free >

Already an Azure customer? [Getting started](#) >

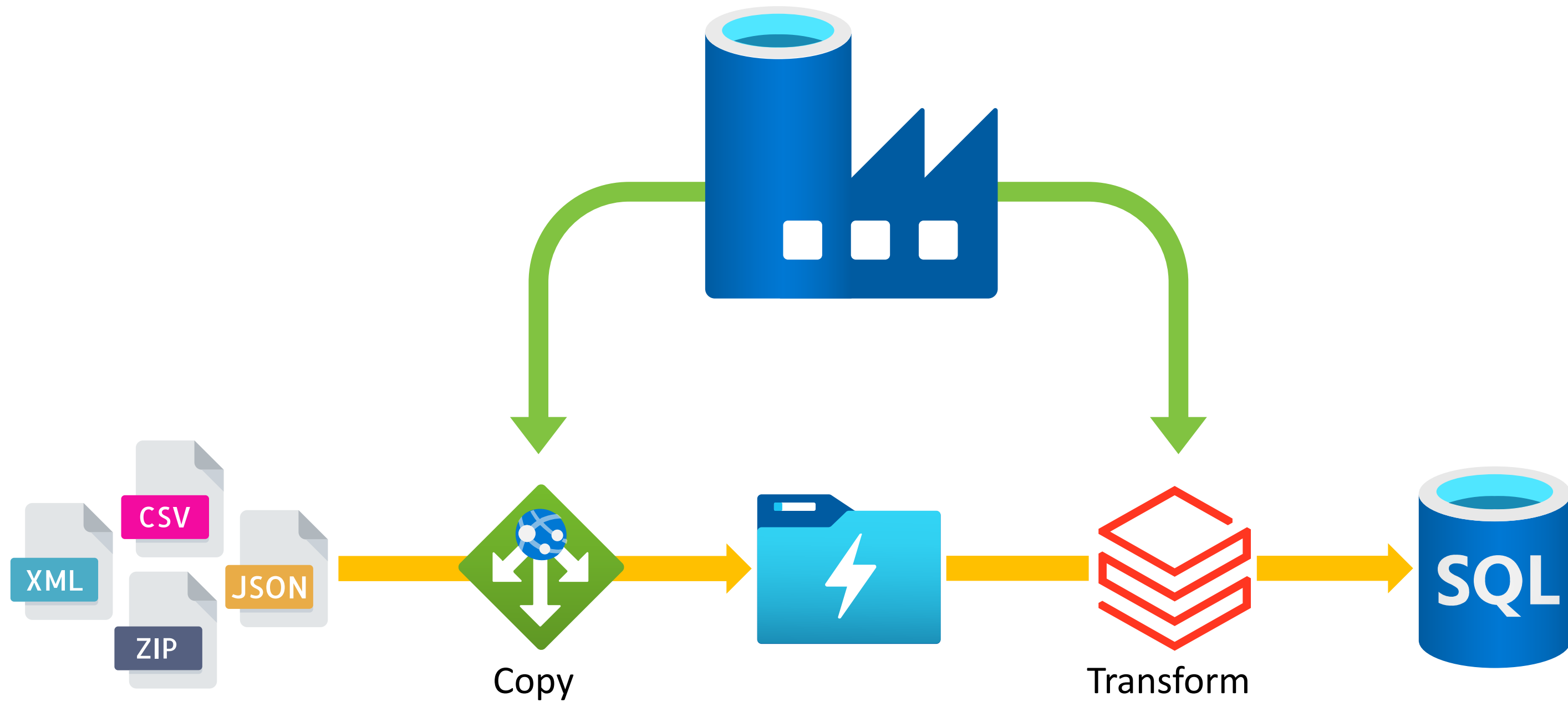
[Product overview](#) [Features](#) [Security](#) [Pricing](#) [Customer stories](#) [Getting started](#) [Documentation](#) [FAQs](#)

Accelerate data integration

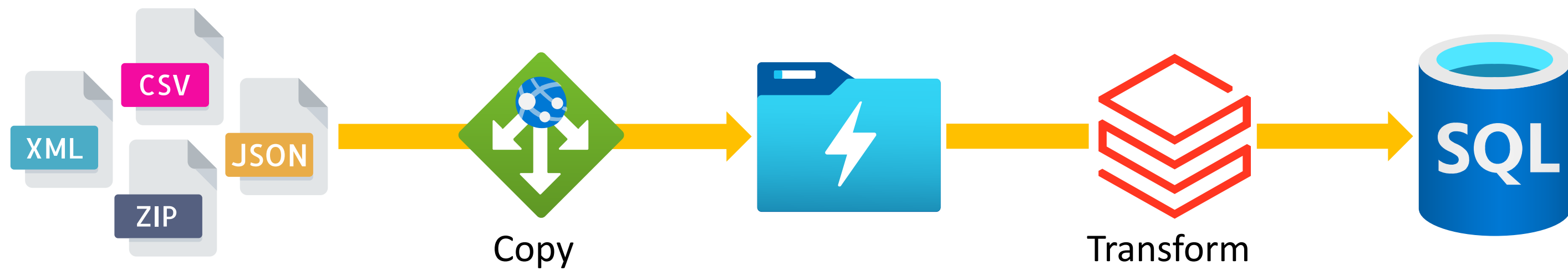
Integrate data silos with Azure Data Factory, a service built for all data integration needs and skill levels. Easily construct ETL and ELT processes code-free within the intuitive visual environment, or write your own code. Visually integrate data sources using more than 90+ natively built and maintenance-free connectors at no added cost. Focus on your data – the serverless integration service does the rest.

<https://azure.microsoft.com/en-gb/services/data-factory/>

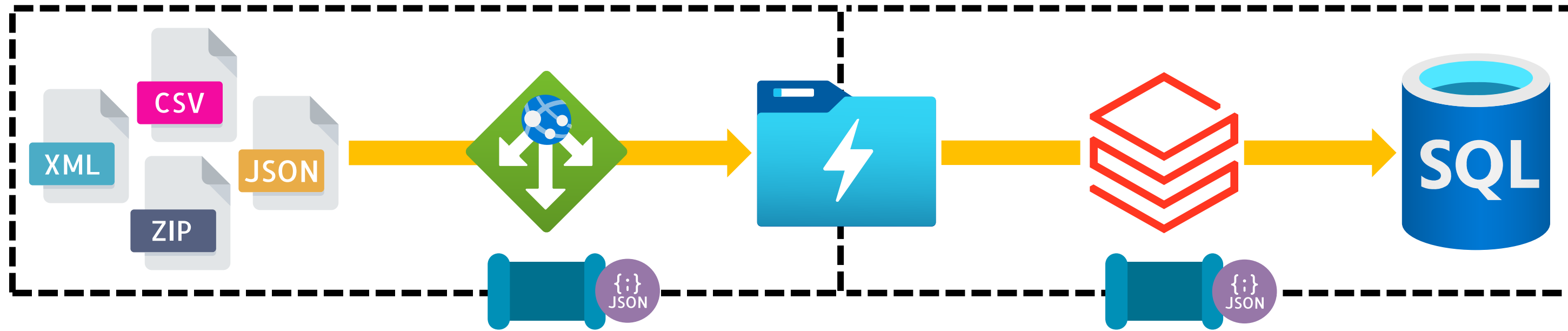
What is Azure Data Factory (ADF)?



What is Azure Data Factory (ADF)?



Data Factory Components



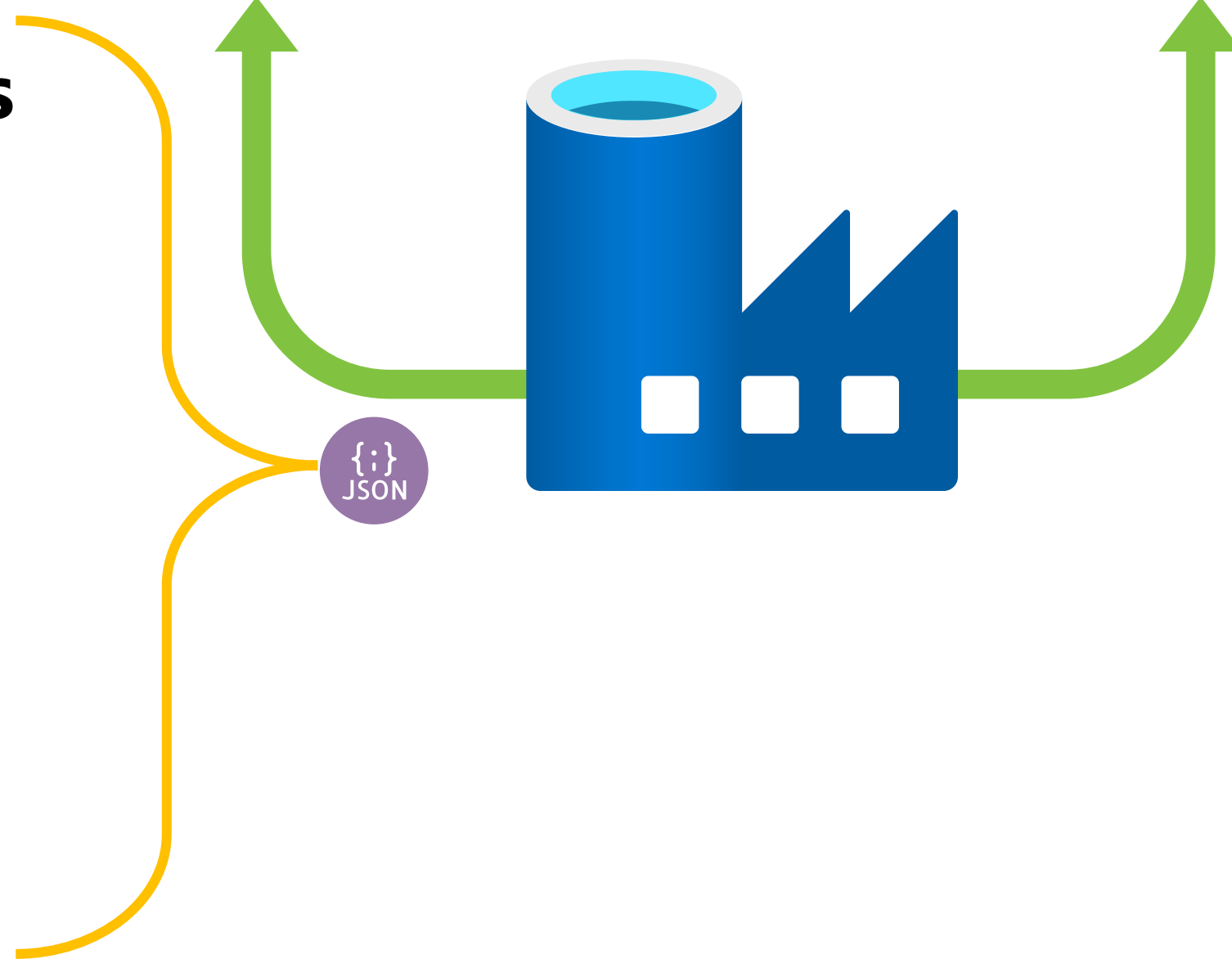
1 **Linked Services**

2 **Datasets**

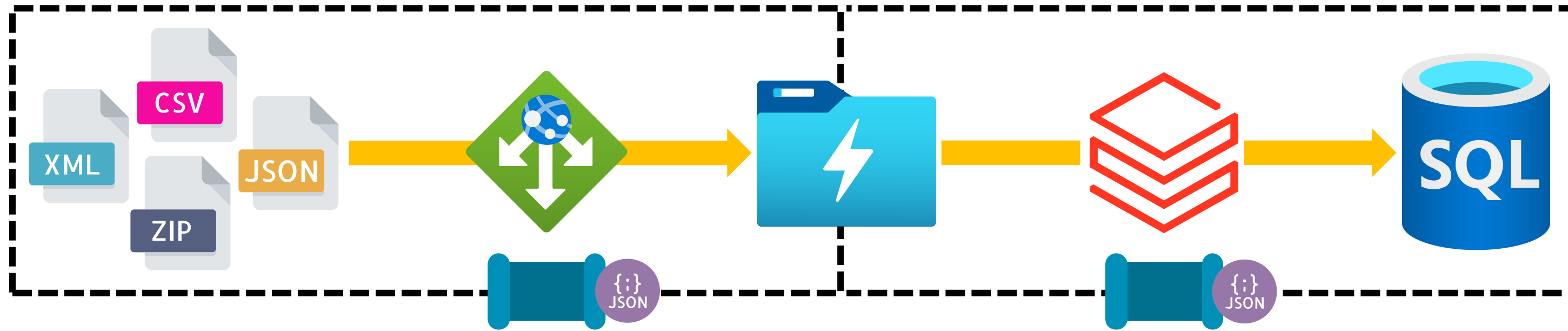
3 **Activities**

4 **Pipelines**

5 **Triggers**



Data Factory Control Flow Components



1 **Linked Services**

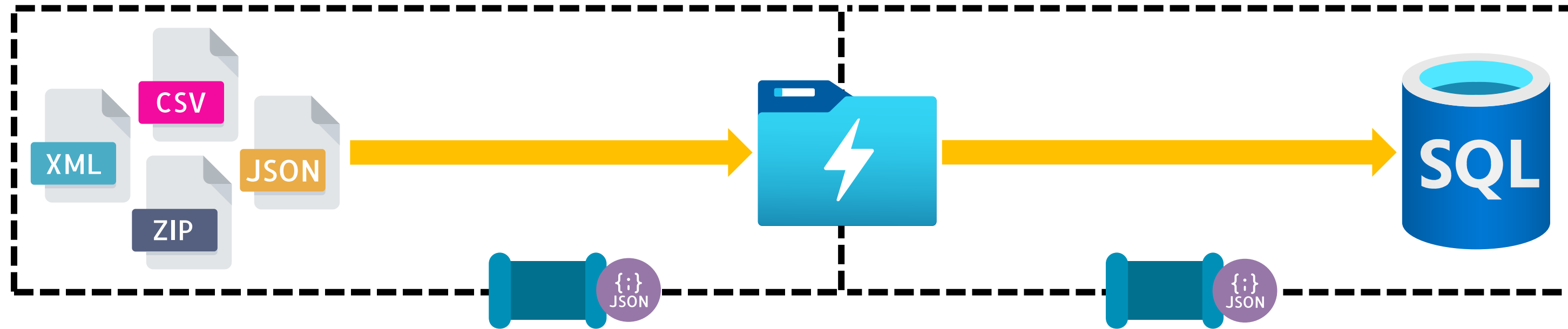
2 **Datasets**

3 **Activities**

4 **Pipelines**

5 **Triggers**

Data Factory Components



1 **Linked Services**

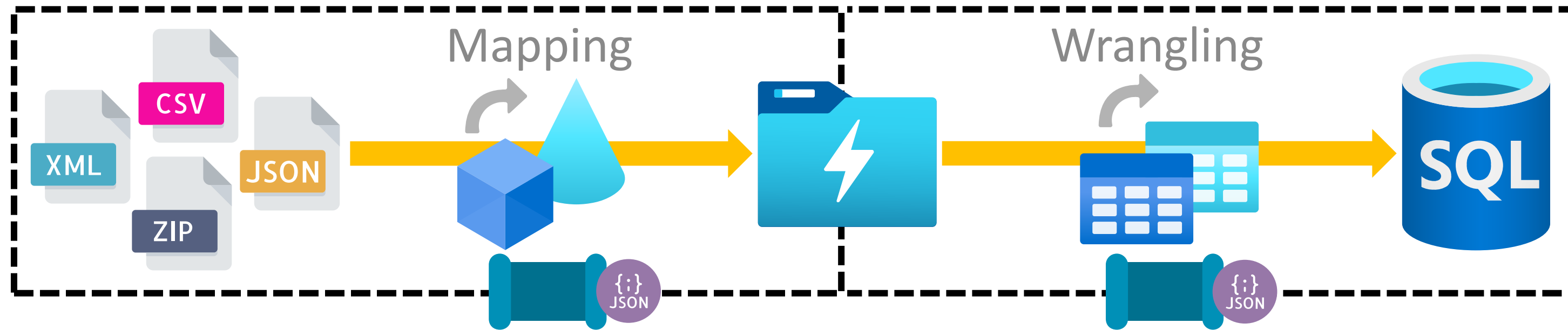
2 **Datasets**

3 **Activities**

4 **Pipelines**

5 **Triggers**

Data Factory Data Flows



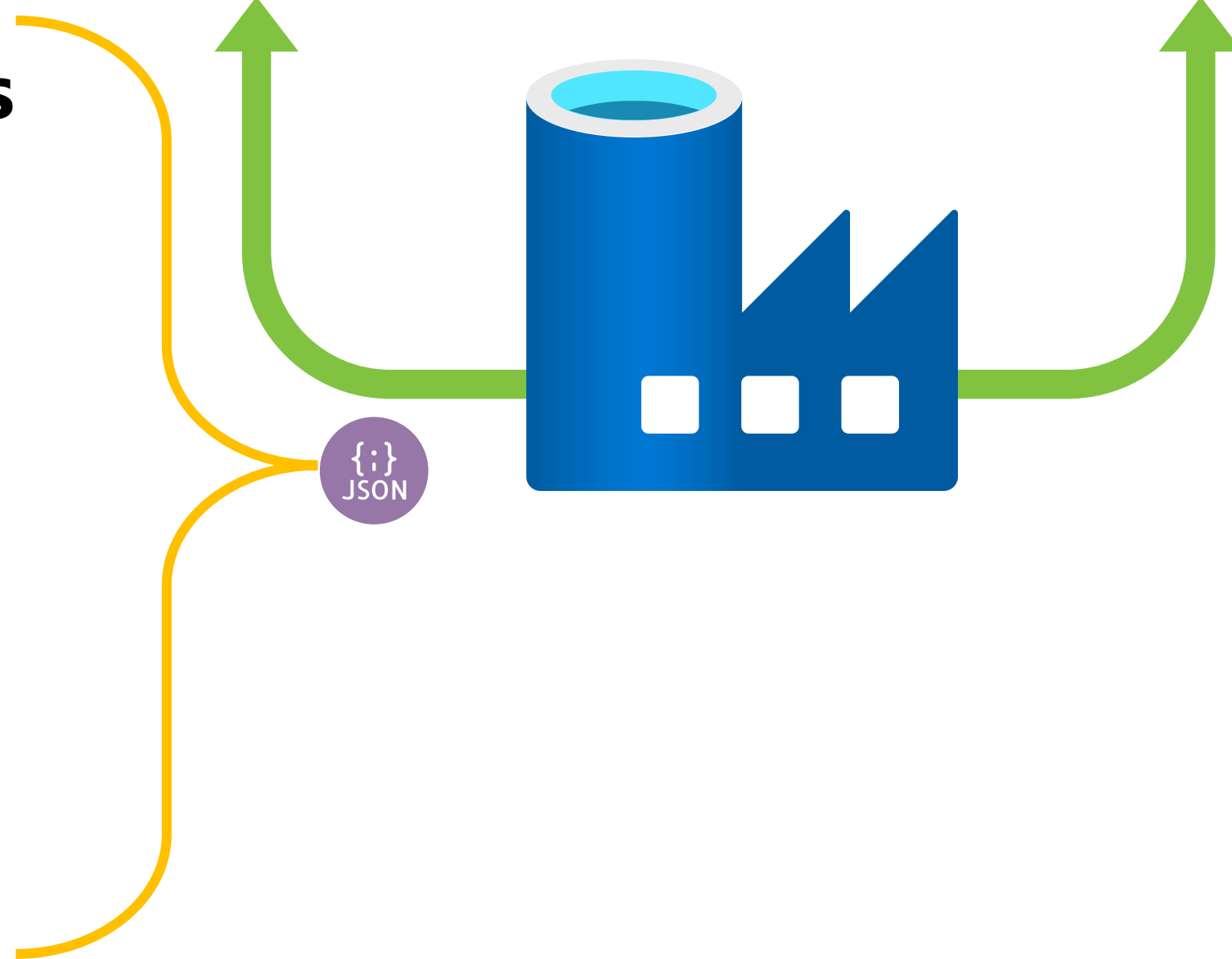
1 **Linked Services**

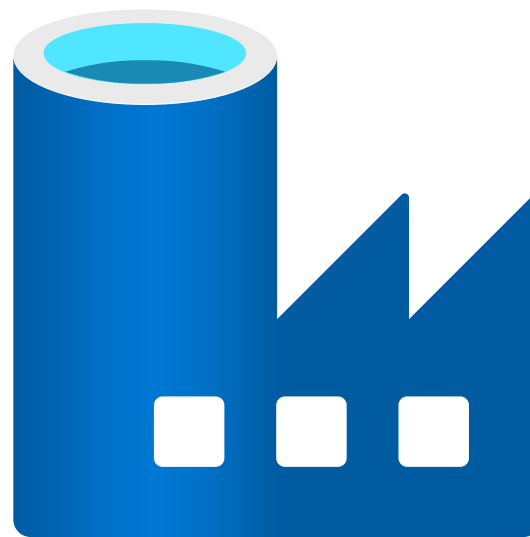
2 **Datasets**

3 **Activities**

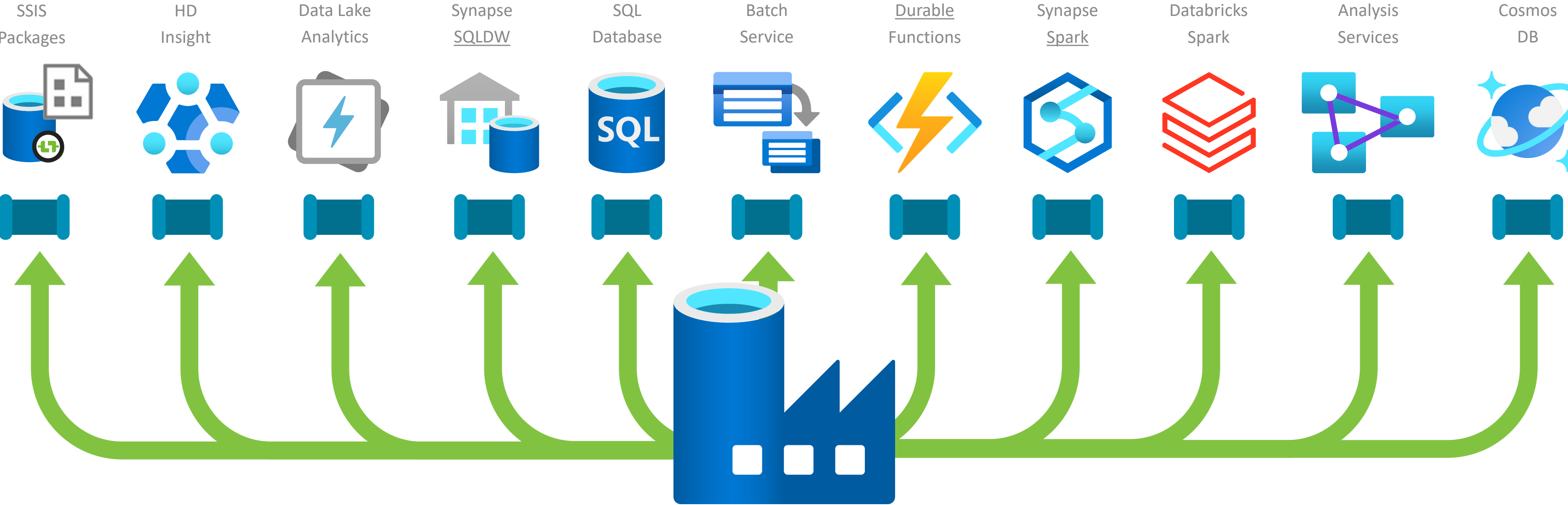
4 **Pipelines**

5 **Triggers**

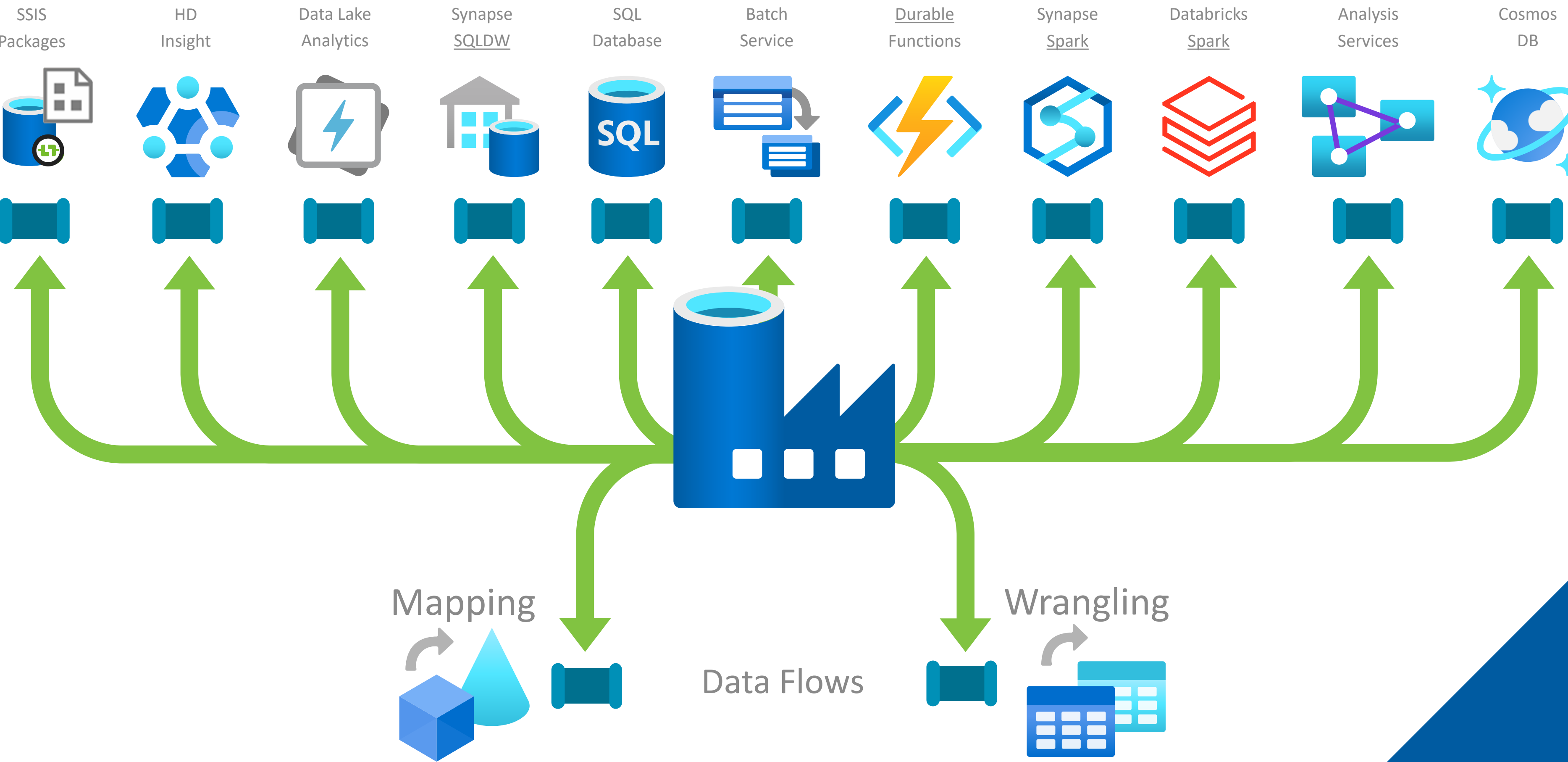




Other Data Transformation Services in Azure



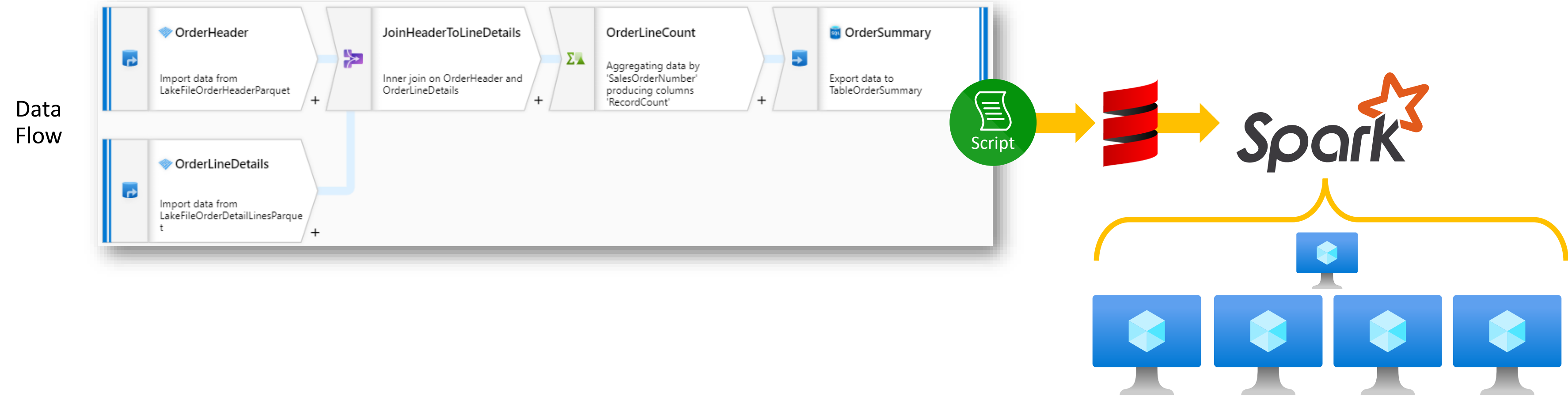
When Should We Use Data Flows?



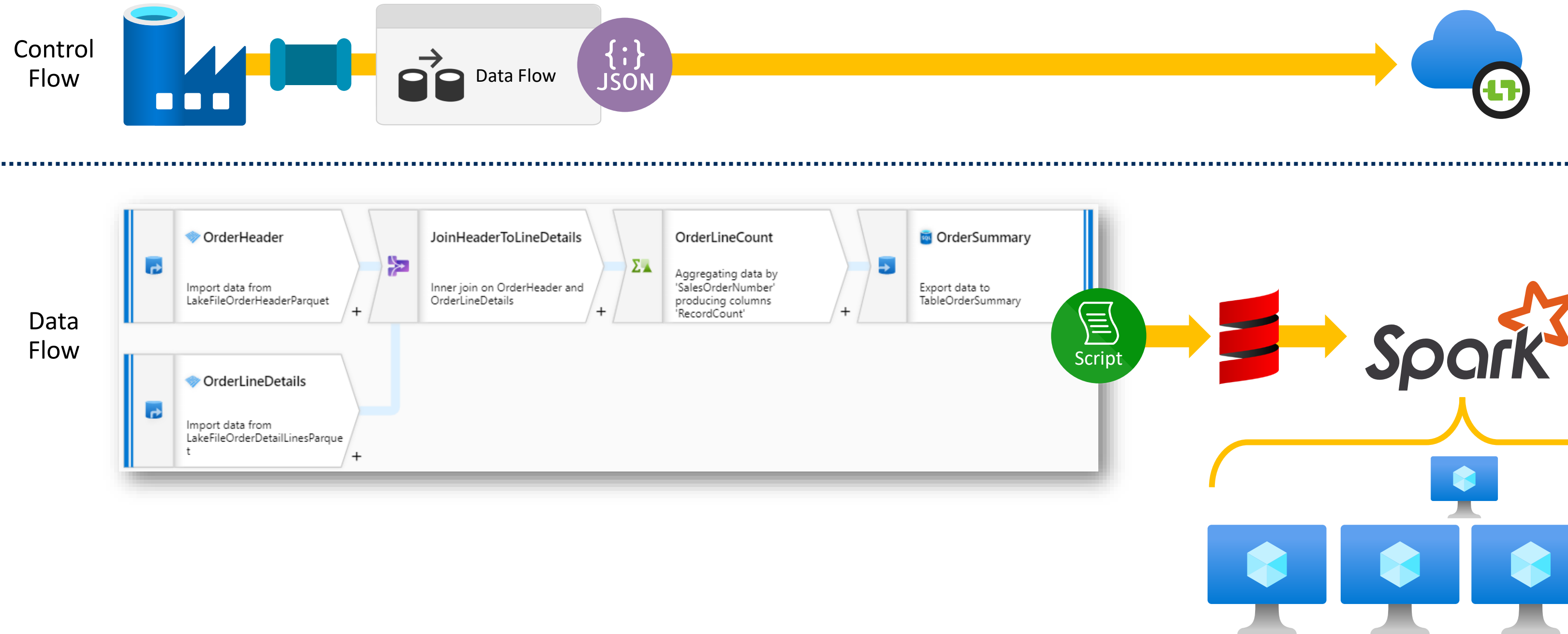
Mapping Data Flows



What is a Mapping Data Flow?



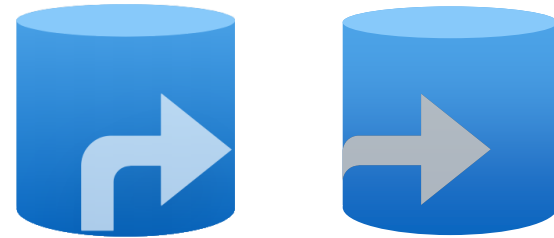
Q: What is a Mapping Data Flow?



A: Graphic data transformation tool that sits on top of Apache Spark.

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

Source & Sink



Limited Connectors

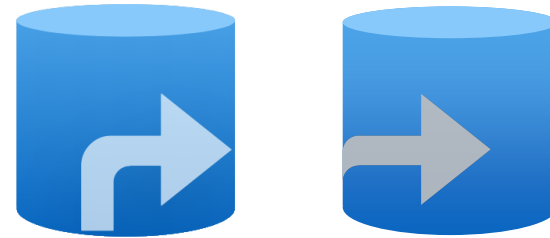


Limited File Type Support



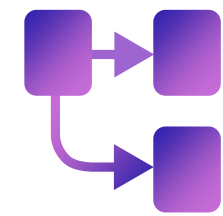
What can a Mapping Data Flow do? - Inputs and Outputs

Source & Sink

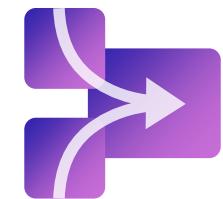


- Schema Drift & Validation
- Inferred Drifted Column Types
- File Lists
- Delete/Move Operations
- File Modified Date Filtering
- Pre-Execute Scripts & Operations (Truncate)

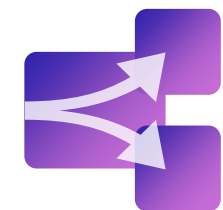
What can a Mapping Data Flow do? - Transformations



New Branch



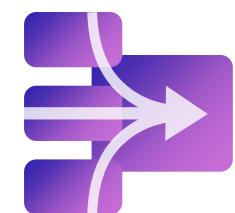
Join



Conditional Split



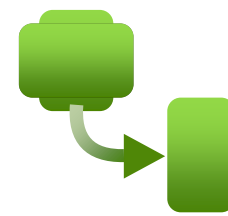
Exists



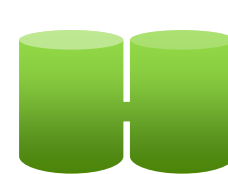
Union



Lookup



Derived Column



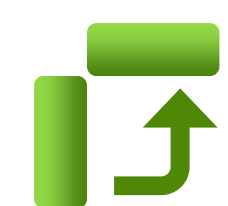
Select



Aggregate



Surrogate Key



Pivot



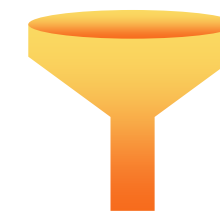
Unpivot



Window



Flatten



Filter



Sort



Alter Row

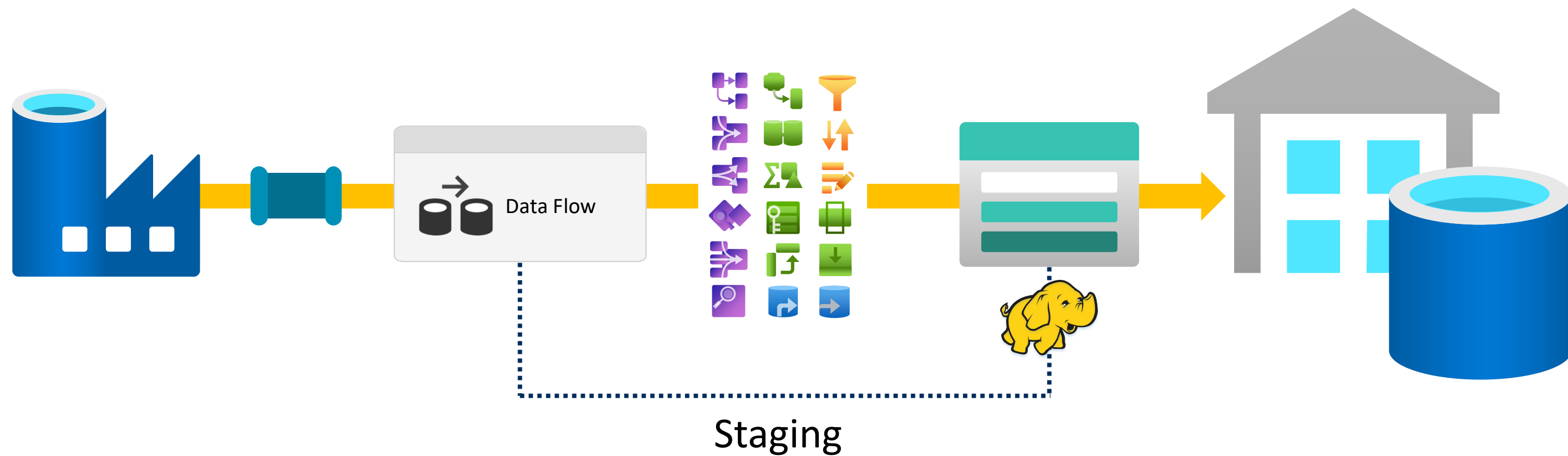
Key

Input & Output Modifiers

Schema Modifiers

Row Modifiers

What can a Mapping Data Flow do? - PolyBase

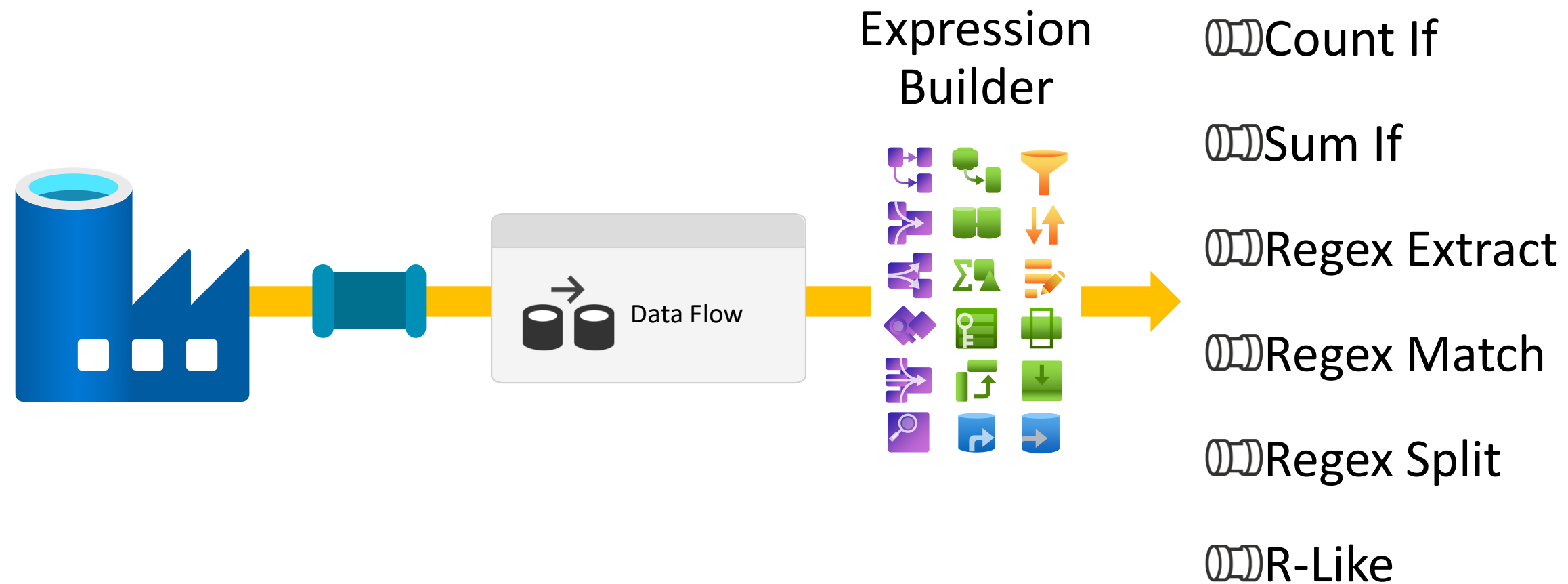


▲ PolyBase ⓘ

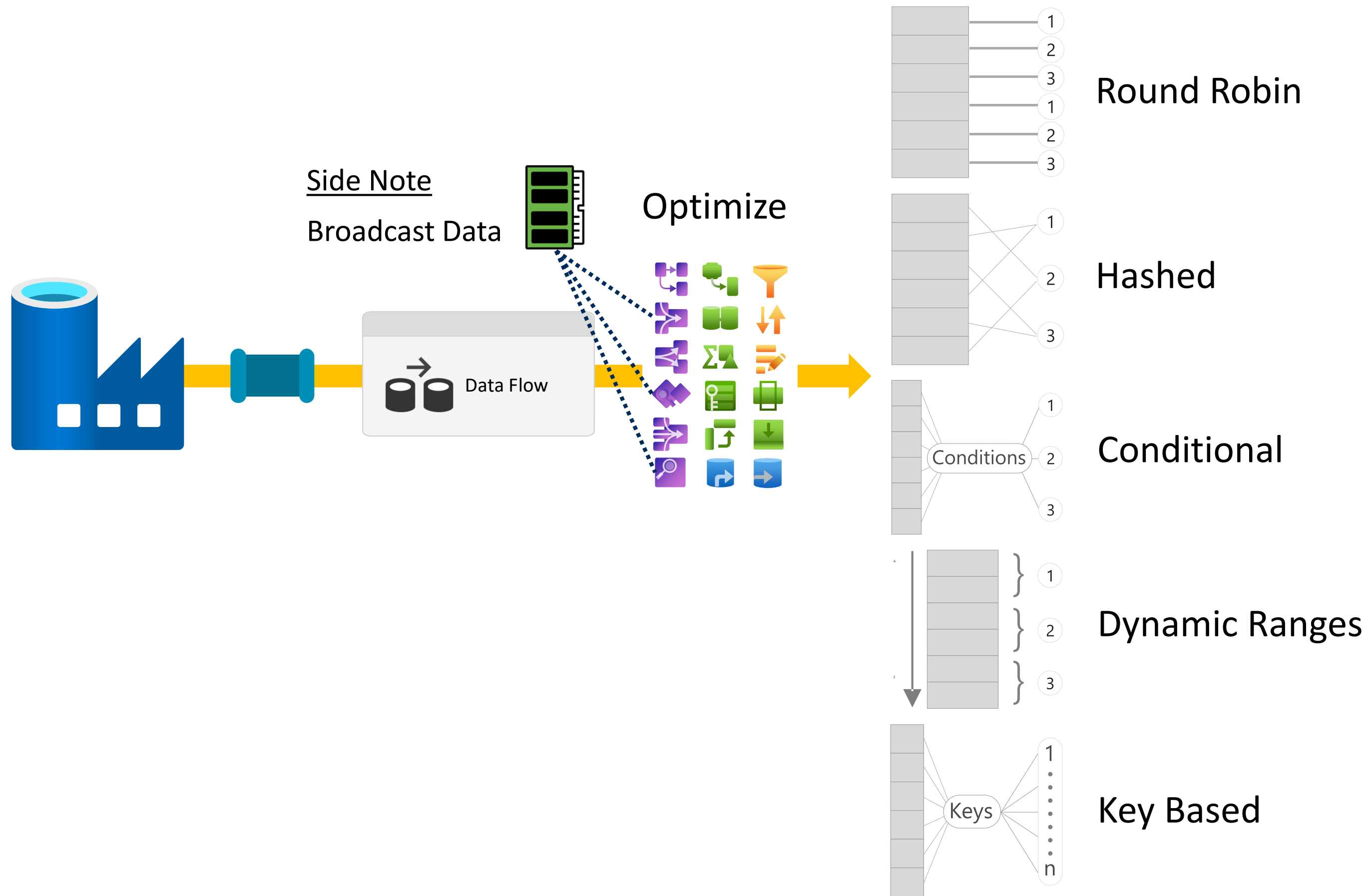
Staging linked service ⓘ + New

Staging storage folder / ▼

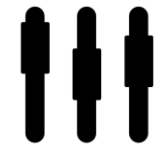
What can a Mapping Data Flow do? - Expression Builder



What can a Mapping Data Flow do? - Partition Handling

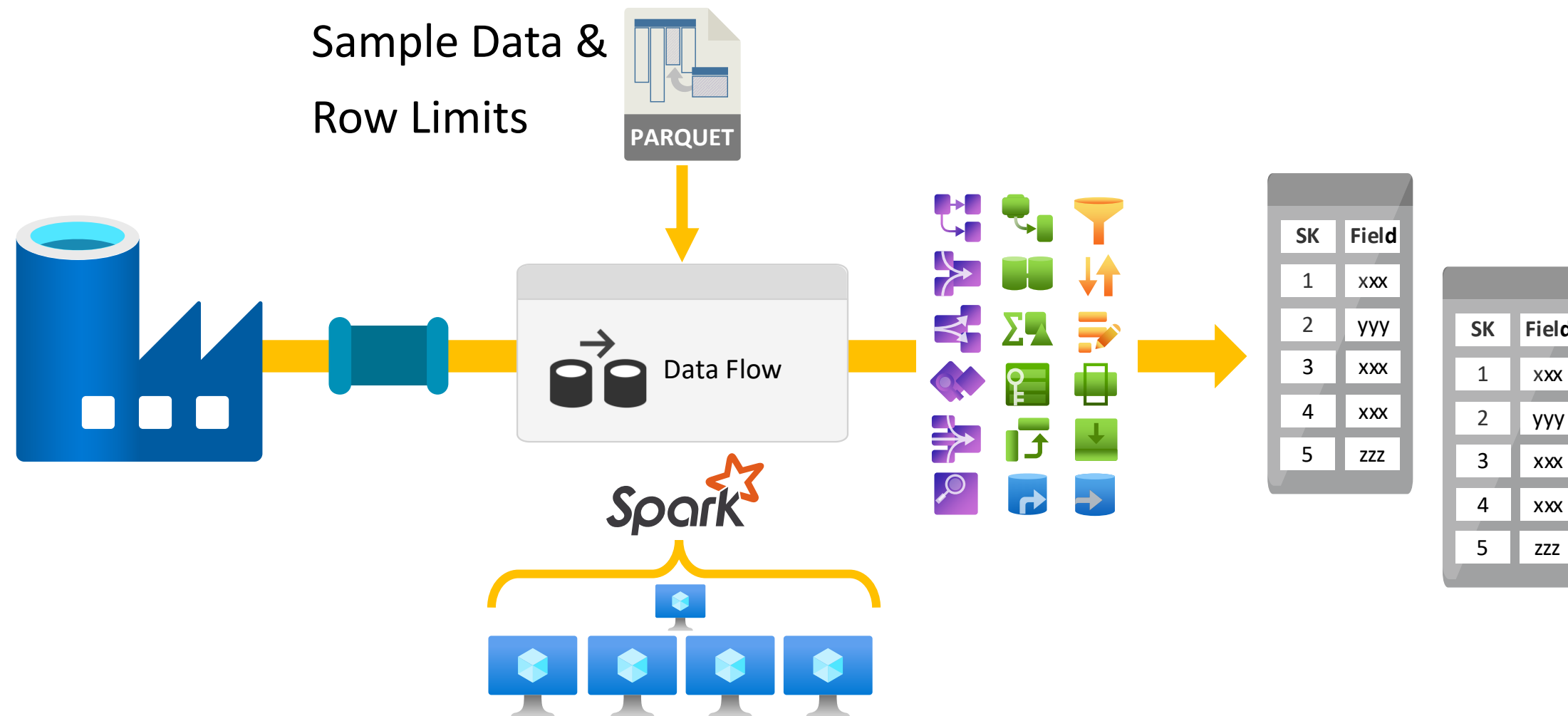


What can a Mapping Data Flow do? - Debugging

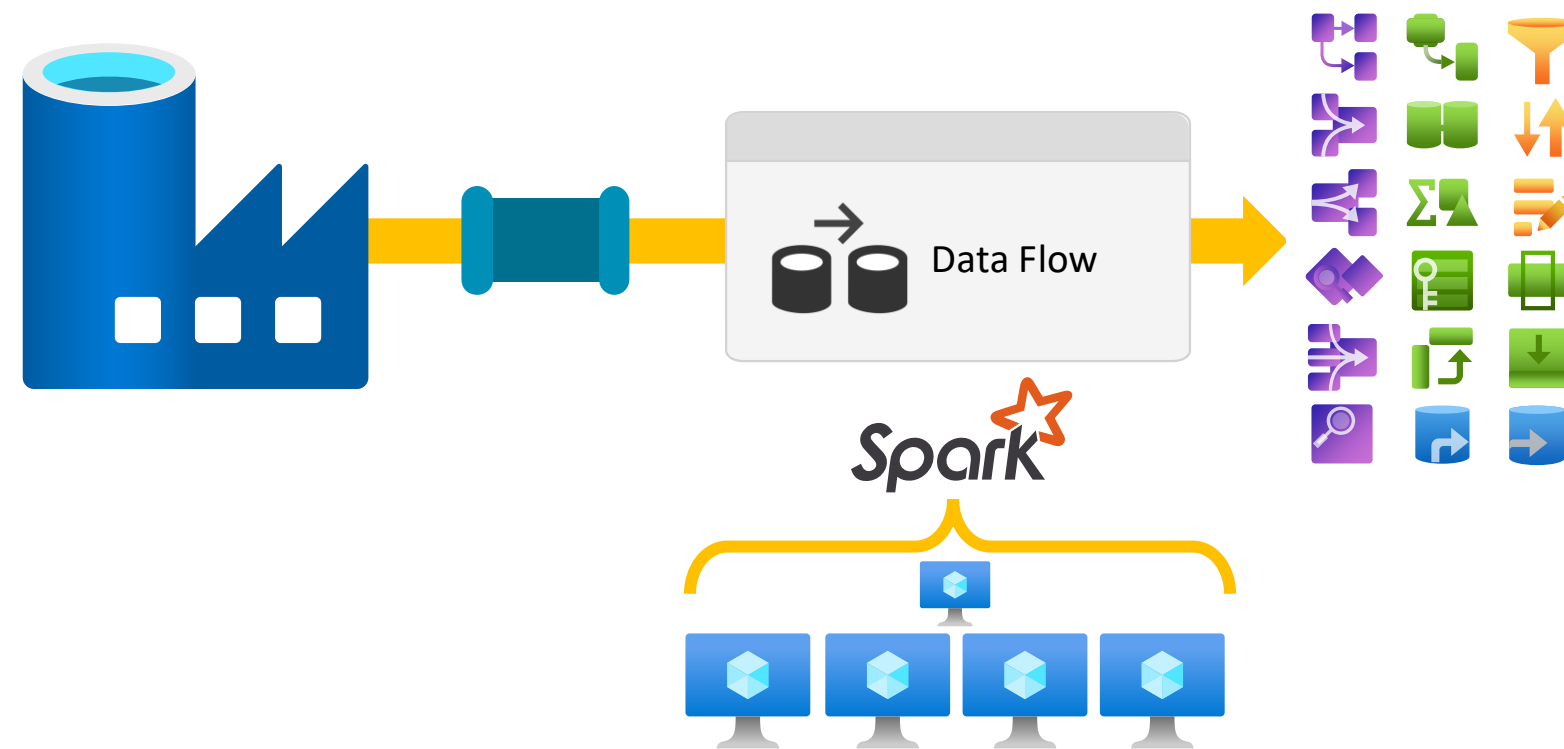


Enable Data Flow Debug Mode

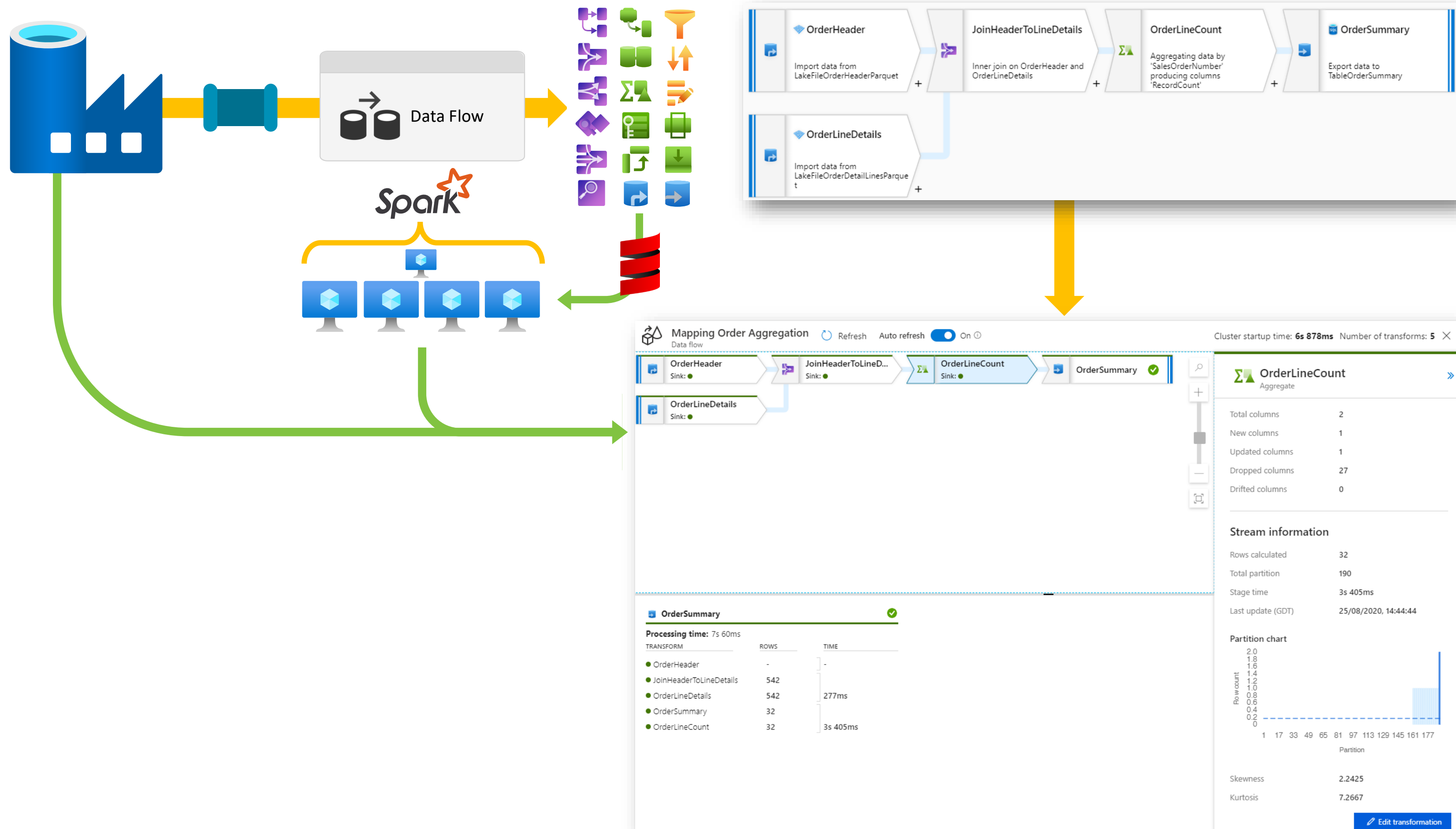
Data Preview

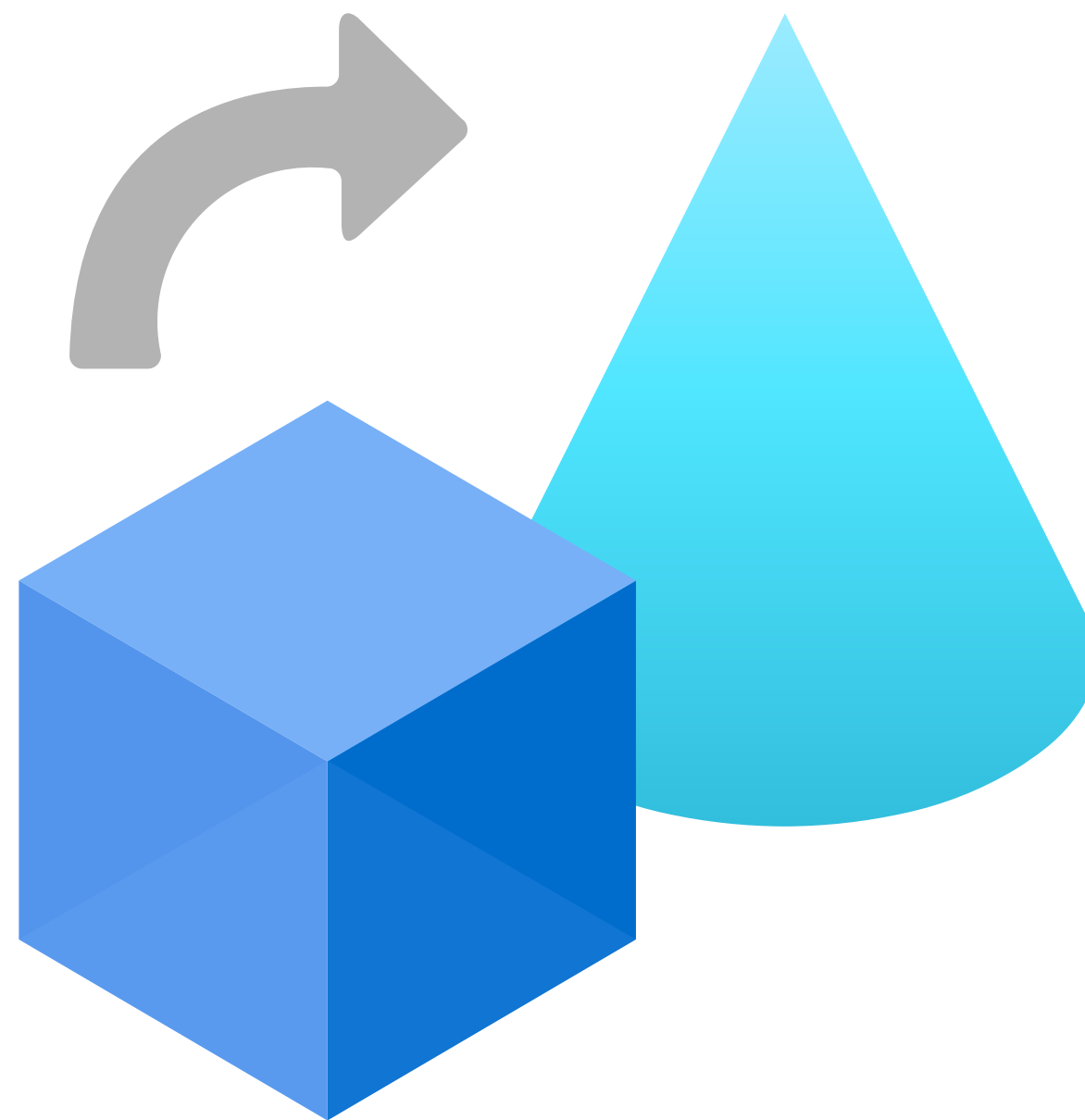


What can a Mapping Data Flow do? - Monitoring



What can a Mapping Data Flow do? - Monitoring






Mapping Data Flow

Wrangling Data Flows

(Preview)



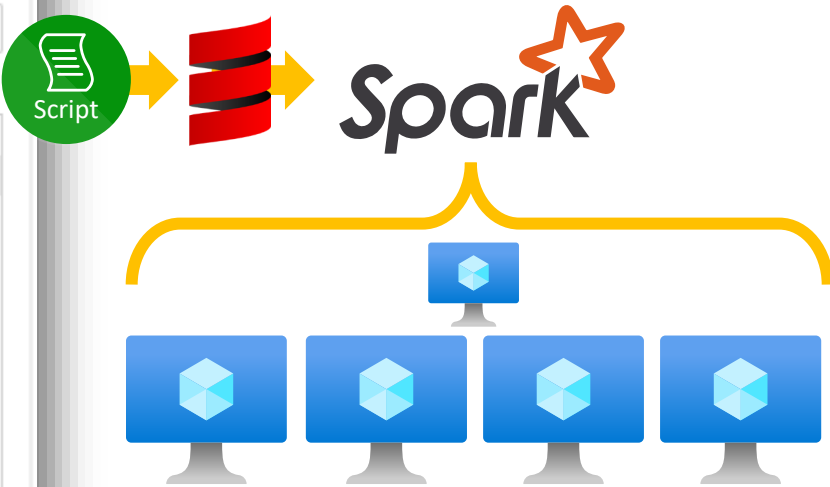
What is a Wrangling Data Flow?



Data Flow

The screenshot shows the Databricks Data Wrangler interface. The top menu bar includes 'Home', 'Transform', 'Add column', and 'View'. Below this is a toolbar with various icons for data manipulation. The main area displays a table with columns: SalesOrderID, SalesOrderDetailID, OrderQty, ProductID, UnitPrice, UnitPriceDiscount, LineTotal, and rowguid. The table contains 17 rows of data. On the right side, the 'Query settings' panel is visible, showing the name 'LakeFileOrderDetailLinesP...' and a list of applied steps: 'AdfDoc' and 'Parquet'.

| | SalesOrderID | SalesOrderDetailID | OrderQty | ProductID | UnitPrice | UnitPriceDiscount | 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 Wrangling Data Flow do?



Data Flow

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

- ADFRResource [1]
- LakeFileOrderDetail...
- 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

What can a Wrangling Data Flow do? - Home



Data Flow

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

Merge queries

Append queries

Combine files

New query

Options

Parameters

Query

Manage columns

Reduce rows

Sort

Transform

Combine

Queries

ADFRResource [1]

LakeFileOrderDetail...

UserQuery

fx

= Parquet.Document(AdfDoc)

Query settings

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

Merge Queries

Append Queries

Combine Files

Text Analytics

Vision

Azure Machine Learning

AI Insights

Queries [1]

OrderDetailLines

fx

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

Query Settings

| 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 | |

Properties

Name

OrderDetailLines

All Properties

Applied Steps

Source

Promoted Headers

Changed Type



What can a Wrangling Data Flow do? - Transform



Data Flow

Home Transform Add column View

Group by Use first row as headers Count rows Transpose Reverse rows Replace values

Data type: Whole number Detect data type Mark as key Rename Pivot column Unpivot columns Move Convert to list Fill

Split column Format Merge columns Extract Parse Statistics Standard Scientific Trigonometry Rounding Information Date Time Duration

Query settings

Queries

- ADFRResource [1]
- LakeFileOrderDetail...
- UserQuery

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

File Home Transform Add Column View Tools Help

Data Type: Whole Number Replace Values Unpivot Columns Split Column Format Merge Columns Extract Parse Statistics Standard Scientific Trigonometry Rounding Information Date Time Duration

Queries [1]

- OrderDetailLines

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

| SalesOrderID | SalesOrderDetailID | OrderQty | ProductID | UnitPrice |
|--------------|--------------------|----------|-----------|-----------|
| 71774 | 110562 | 1 | 836 | 356.898 |
| 71774 | 110563 | 1 | 822 | 356.898 |
| 71776 | 110567 | 1 | 907 | 63.9 |
| 71780 | 110616 | 4 | 905 | 218.454 |
| 71780 | 110617 | 2 | 983 | 461.694 |
| 71780 | 110618 | 6 | 988 | 112.998 |
| 71780 | 110619 | 2 | 748 | 818.7 |
| 71780 | 110620 | 1 | 990 | 323.994 |
| 71780 | 110621 | 1 | 926 | 149.874 |
| 71780 | 110622 | 1 | 743 | 809.76 |
| 71780 | 110623 | 4 | 782 | 1376.994 |
| 71780 | 110624 | 2 | 918 | 158.43 |
| 71780 | 110625 | 4 | 780 | 1391.994 |
| 71780 | 110626 | 1 | 937 | 48.594 |
| 71780 | 110627 | 6 | 867 | 41.994 |
| 71780 | 110628 | 1 | 985 | 112.998 |
| 71780 | 110629 | 2 | 989 | 323.994 |

Query Settings

PROPERTIES

Name

OrderDetailLines

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type

What can a Wrangling Data Flow do? - Add Column



Data Flow

Home Transform Add column View

Conditional column
Index column
Duplicate column

Custom column

General

Format
Merge columns
Extract
Parse

From text

Statistics
Standard
Scientific

From number

Trigonometry
Rounding
Information

Date
Time
Duration

Date and time column

Query settings

Queries

ADFRResource [1]
LakeFileOrderDetail...
UserQuery

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

Column From Examples
Custom Column
Invoke Custom Function

General

Conditional Column
Index Column
Duplicate Column

Format
Merge Columns
Extract
Parse

From Text

Statistics
Standard
Scientific

From Number

Trigonometry
Rounding
Information

Date
Time
Duration

From Date & Time

Text Analytics
Vision
Azure Machine Learning
AI Insights

Queries [1]
OrderDetailLines

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
All Properties

APPLIED STEPS
Source
Promoted Headers
X Changed Type

Table.TransformColumnTypes(#"Promoted Headers",{{"SalesOrderID", Int64.Type}, {"SalesOrderDetailID", Int64.Type})

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

Source
Promoted Headers
X Changed Type

What can a Wrangling Data Flow do? - View



Data Flow

Home Transform Add column View

Data view Schema view Go to column Advanced editor

Preview Columns Advanced

Queries

- ADFResource [1]
- LakeFileOrderDetailL...
- UserQuery

Query settings

File Home Transform Add Column View Tools Help

☒ Formula Bar ☐ Monospaced ☐ Column distribution ☐ Always allow ☐ Advanced Editor ☐ Query Dependencies

☒ Show whitespace ☐ Column profile

☐ Column quality

Layout Data Preview Columns Parameters Advanced Dependencies

Queries [1]

- OrderDetailLines

Query Settings

PROPERTIES

Name

OrderDetailLines

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type

| 123 SalesOrderID | 123 SalesOrderDetailID | 123 OrderQty | 123 ProductID | 1.2 UnitPrice | 1.2 UnitPrice |
|------------------|------------------------|--------------|---------------|---------------|---------------|
| 71774 | 110562 | 1 | 836 | 356.898 | |
| 71774 | 110563 | 1 | 822 | 356.898 | |
| 71776 | 110567 | 1 | 907 | 63.9 | |
| 71780 | 110616 | 4 | 905 | 218.454 | |
| 71780 | 110617 | 2 | 983 | 461.694 | |
| 71780 | 110618 | 6 | 988 | 112.998 | |
| 71780 | 110619 | 2 | 748 | 818.7 | |
| 71780 | 110620 | 1 | 990 | 323.994 | |
| 71780 | 110621 | 1 | 926 | 149.874 | |
| 71780 | 110622 | 1 | 743 | 809.76 | |
| 71780 | 110623 | 4 | 782 | 1376.994 | |
| 71780 | 110624 | 2 | 918 | 158.43 | |
| 71780 | 110625 | 4 | 780 | 1391.994 | |
| 71780 | 110626 | 1 | 937 | 48.594 | |
| 71780 | 110627 | 6 | 867 | 41.994 | |
| 71780 | 110628 | 1 | 985 | 112.998 | |
| 71780 | 110629 | 2 | 989 | 323.994 | |

What can a Wrangling Data Flow do? - View



Data Flow

Home Transform Add column View

Data view Schema view Go to column Preview Columns Advanced

Queries

- ADFResource [1]
- LakeFileOrderDetailL...
- UserQuery

Advanced editor

```
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"
```

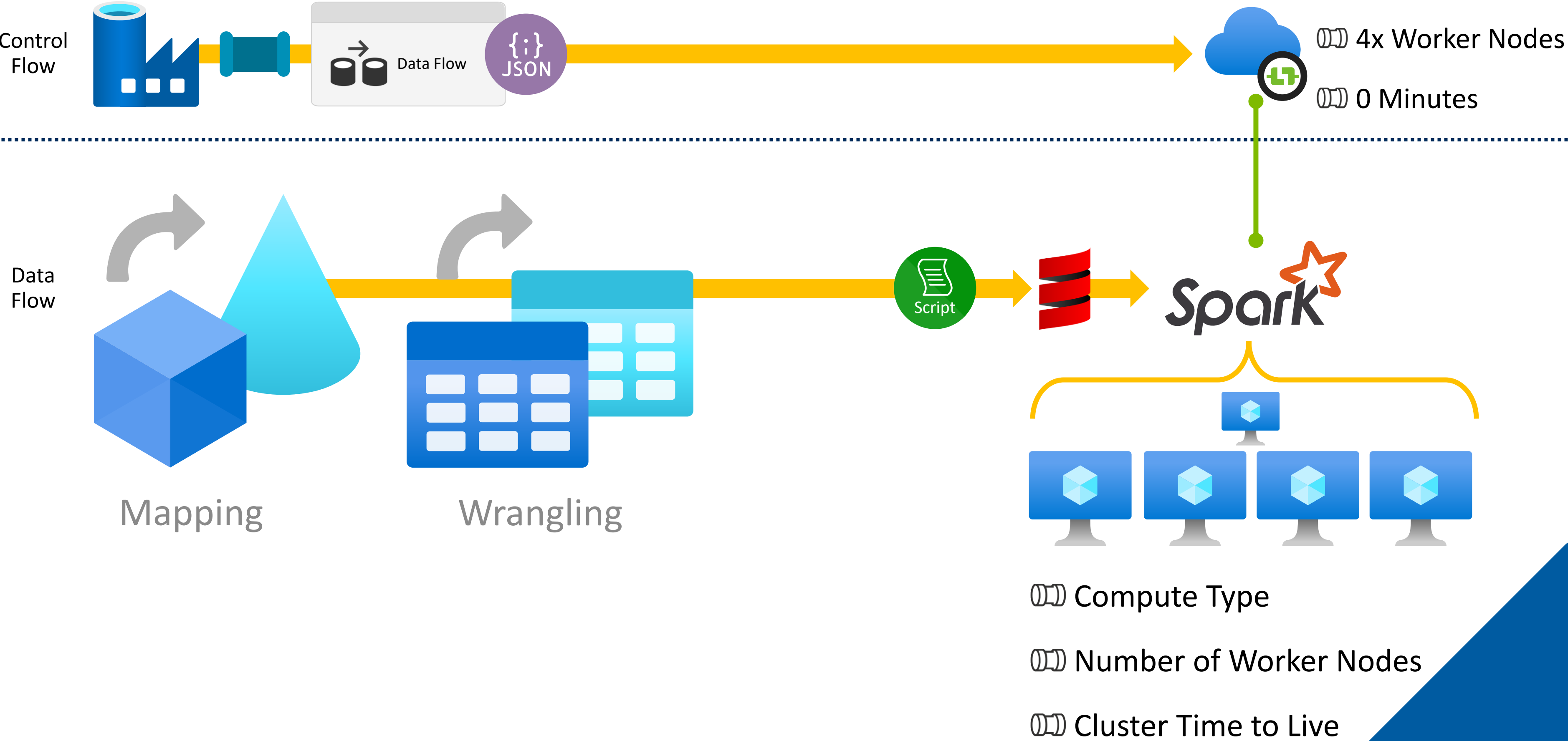
OK Cancel



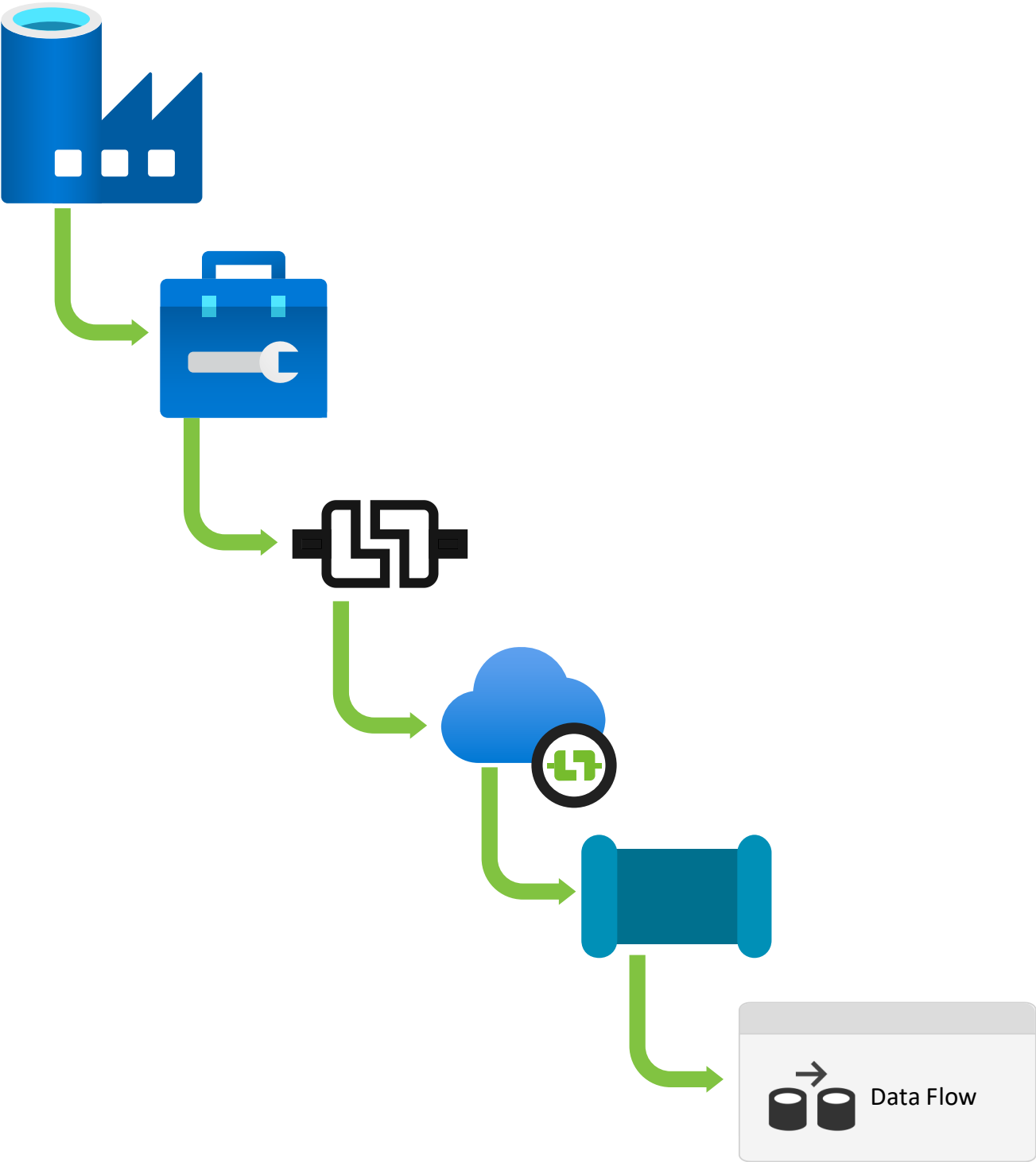
Wrangling Data Flow

Configuration

Data Flow Cluster Configuration



Setting the Data Flow Cluster (IR Configuration)

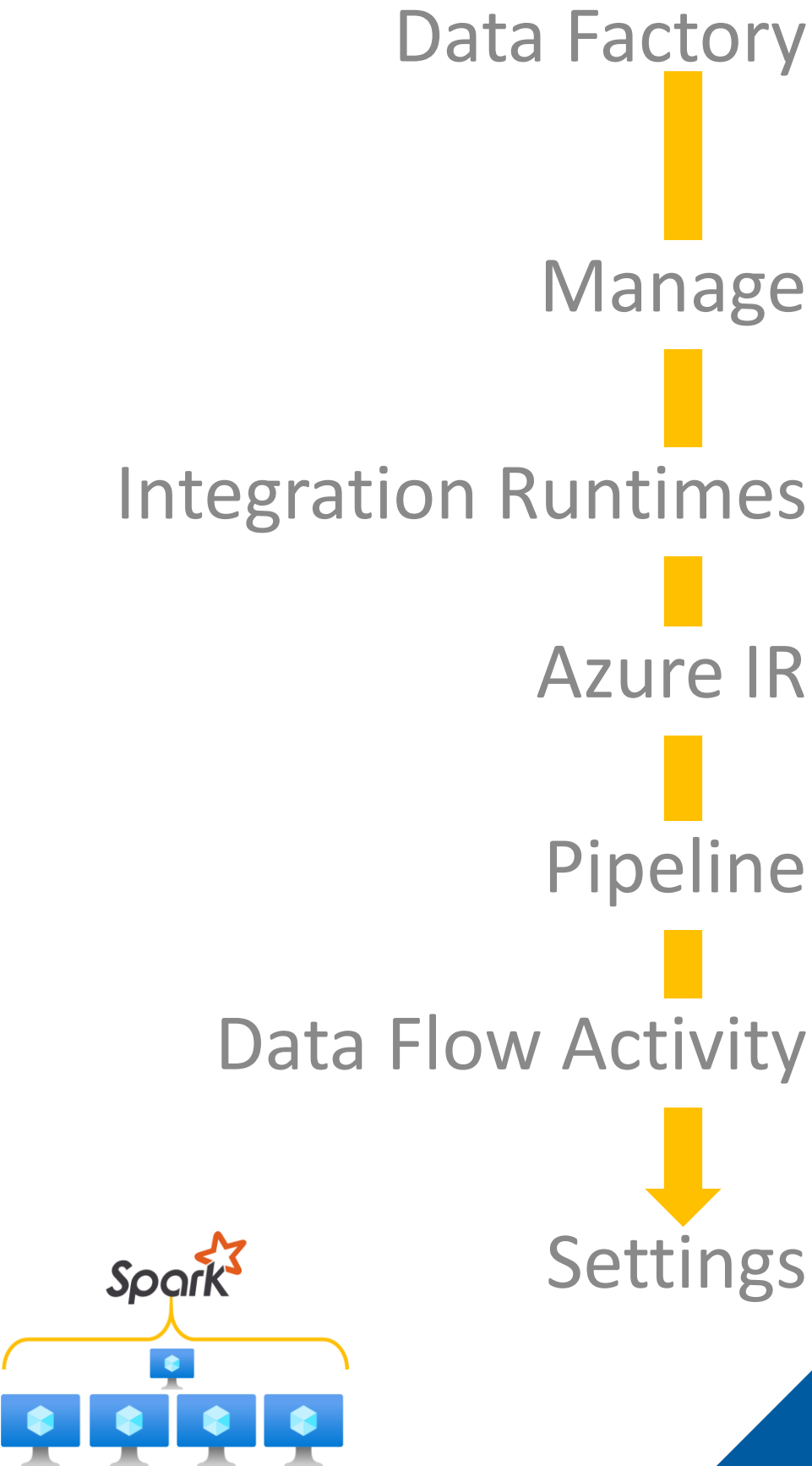


General Settings Parameters User properties


Data flow * MappingOrderAggregation

Run on (Azure IR) * DataFlowDemosTTL4Hours




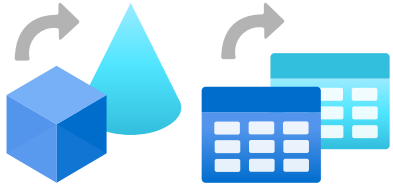
PolyBase



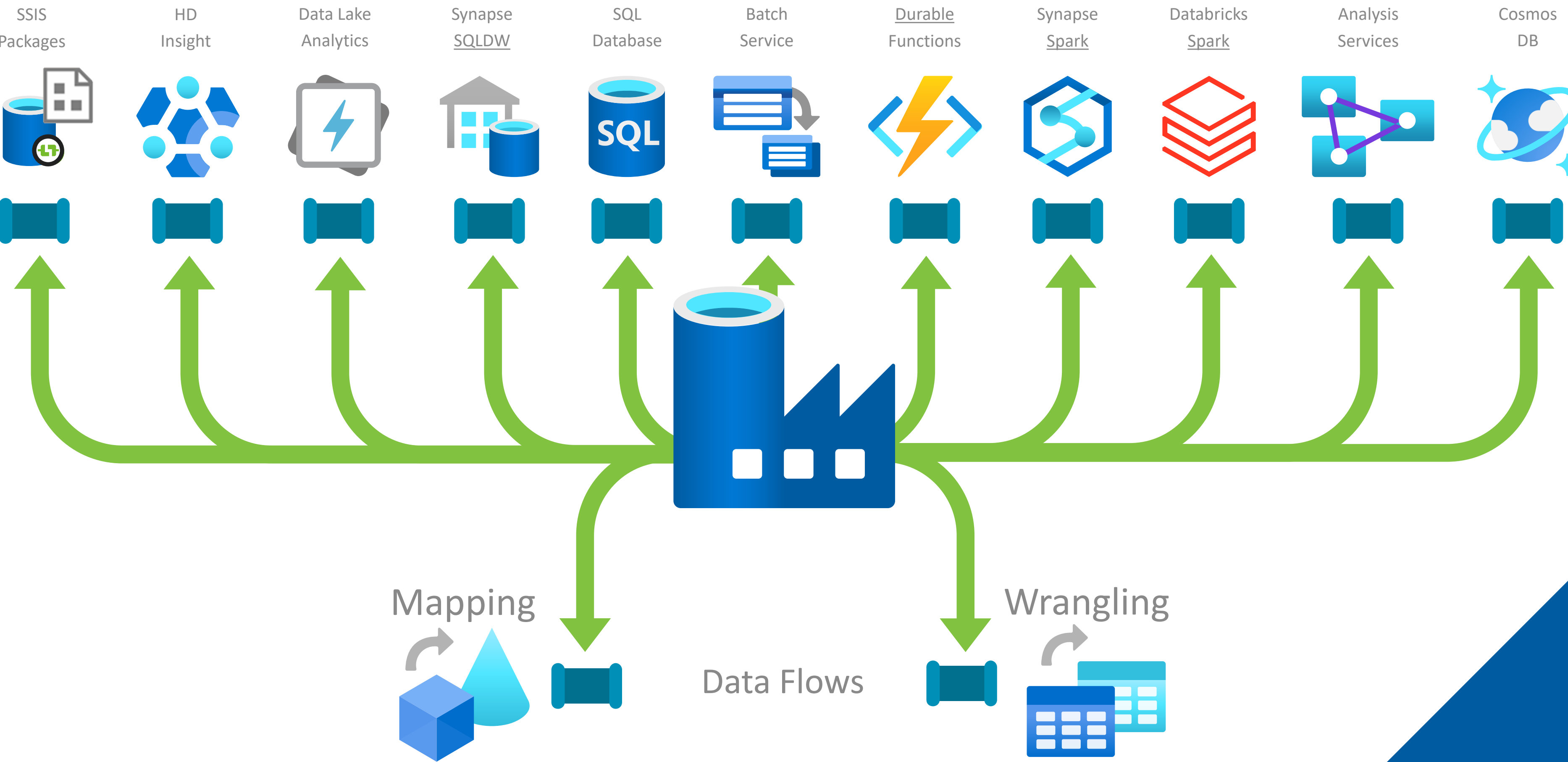
Use Cases & Conclusions



Data Transformations in Azure Comparisons

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

When Should We Use Data Flows?



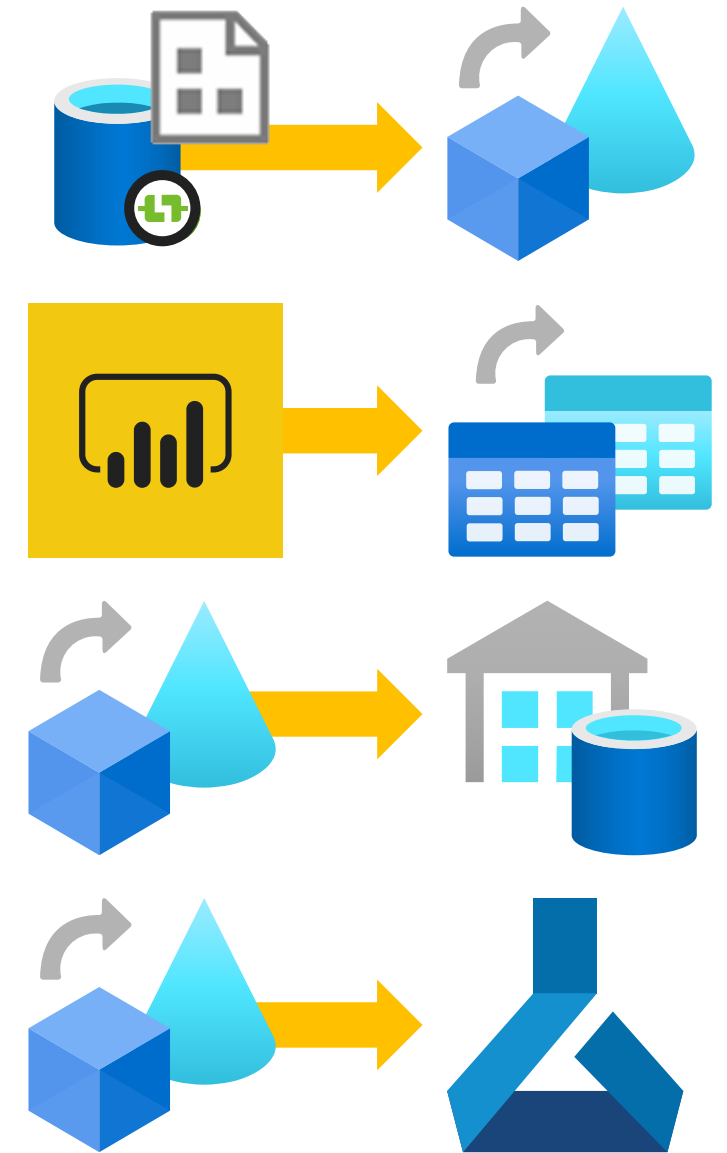
Use Cases

SSIS developers who are transferring existing skills to cloud native technologies have a very low barrier to entry and don't need to worry about distributed compute to get started.

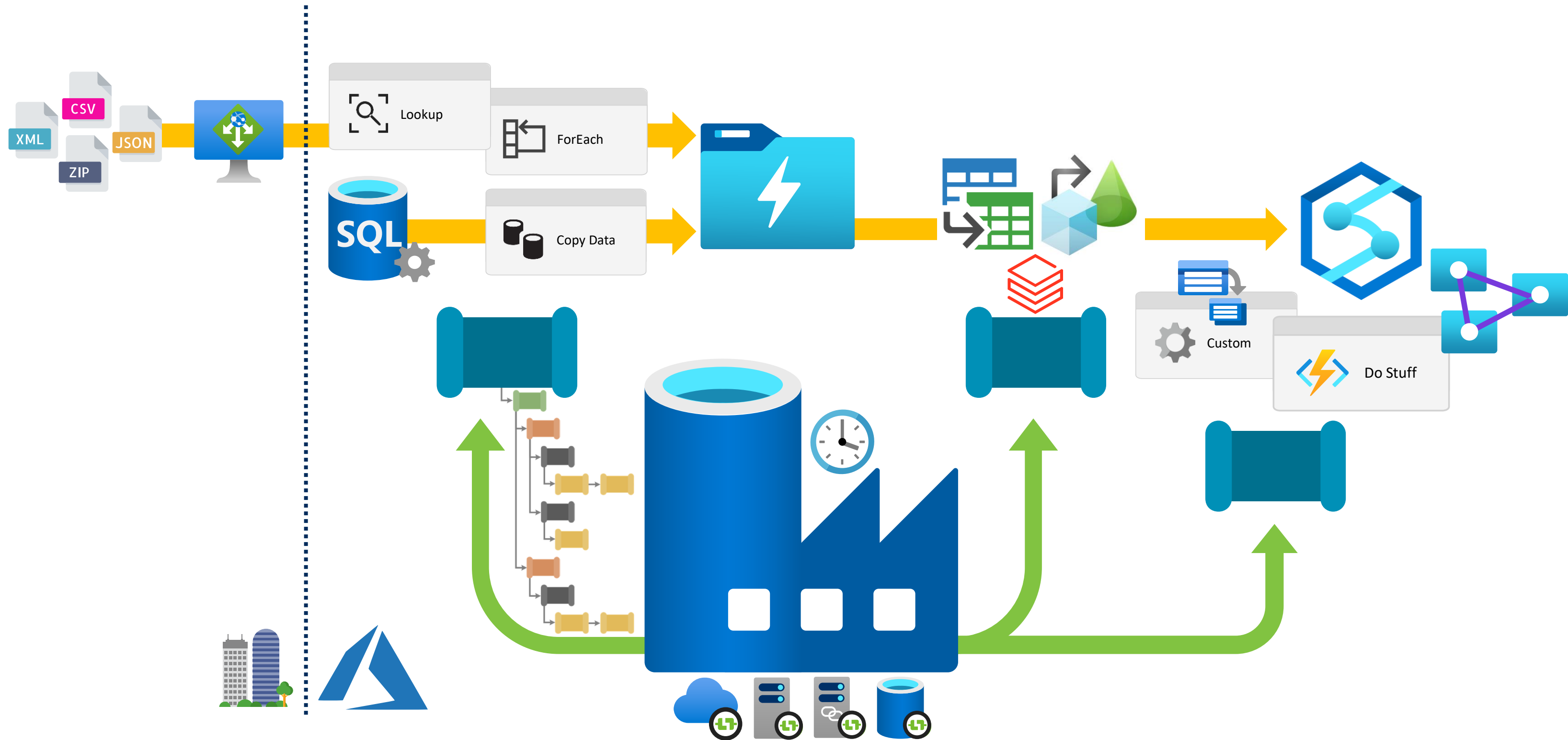
Data engineering made easy for the power users who has grown out of Power BI following a series of Data Lake exploration sessions.

Data insight teams needing to do rapid prototyping and data warehouse loading within a single Azure Resource making deployments simple and release cycles short.

Simpler and quicker data engineering for data scientists that want to quickly prepare raw data for model training and testing, also with the ability to use large amounts of compute.



What is Azure Data Factory?



1. A complete Microsoft Azure integration tool.
2. Orchestrator of our Control Flow operations – with scale out Activities.
3. Orchestrator of our Data Flow transformations – using cloud native services.
4. The scheduler of solutions – using a variety of Pipeline Triggers.

Thank you for listening...

Paul Andrew



altius

Blog: mrpaulandrew.com

Email: paul@mrpaulandrew.com

Twitter: [@mrpaulandrew](https://twitter.com/mrpaulandrew)

LinkedIn: [In/mrpaulandrew](https://in.linkedin.com/in/mrpaulandrew)

GitHub: github.com/mrpaulandrew