ETL in Azure Made Easy

with Data Factory Data Flows

Paul Andrew

Principal Consultant & Solution Architect

altius PASS

MVP Microsoft® Most Valuable Professional

@MrPaulAndrew

GitHub

https://github.com/mrpaulandrew

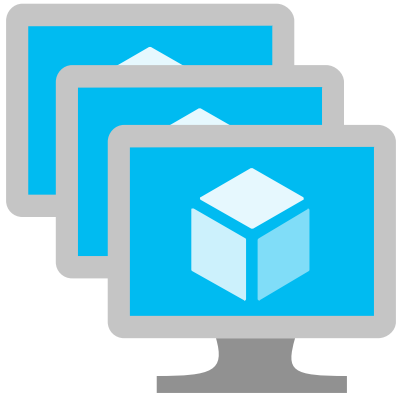
CommunityEvents

Demo code, content and slides from various community events.

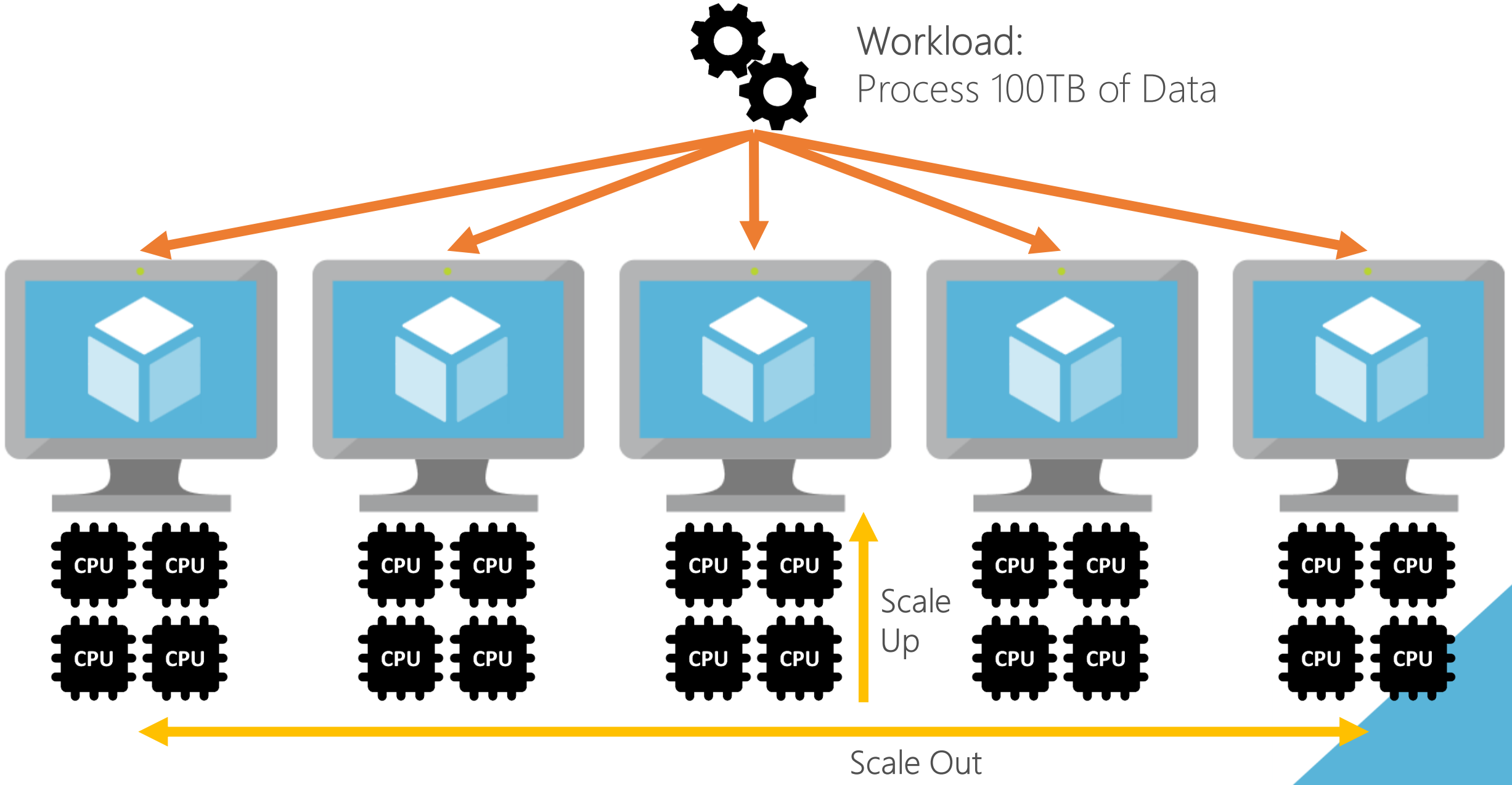
C++

{Event/Location}-{Month}-{Year}

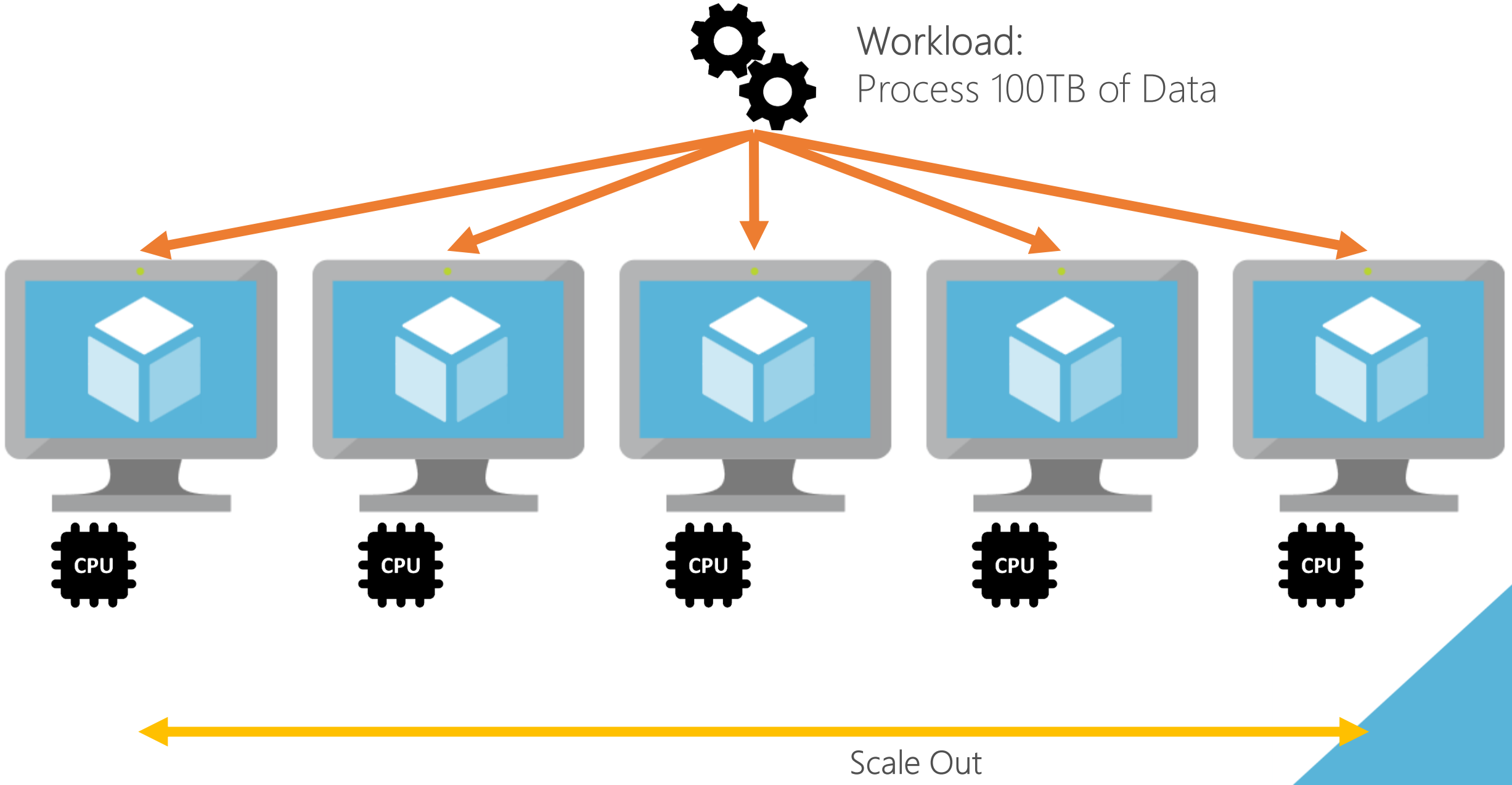
Scale Up vs Scale Out



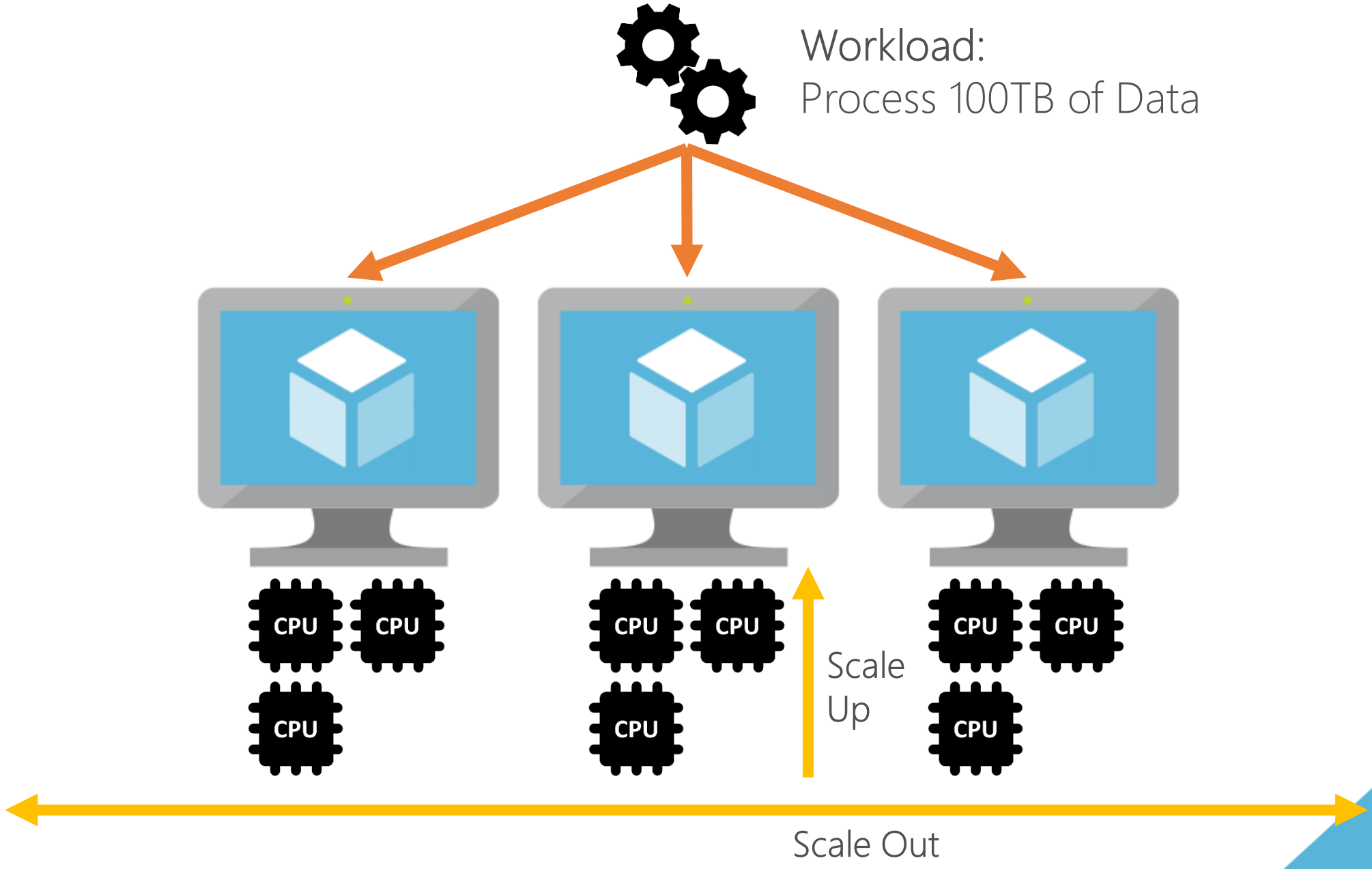
Scale Up and Scale Out



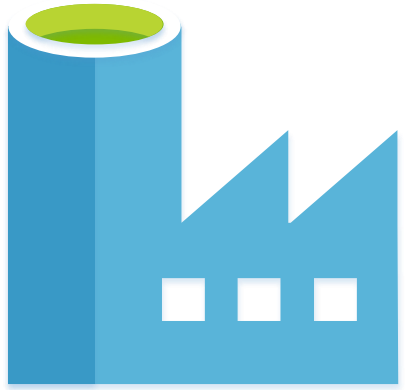
Scale Up and Scale Out



Scale Up and Scale Out

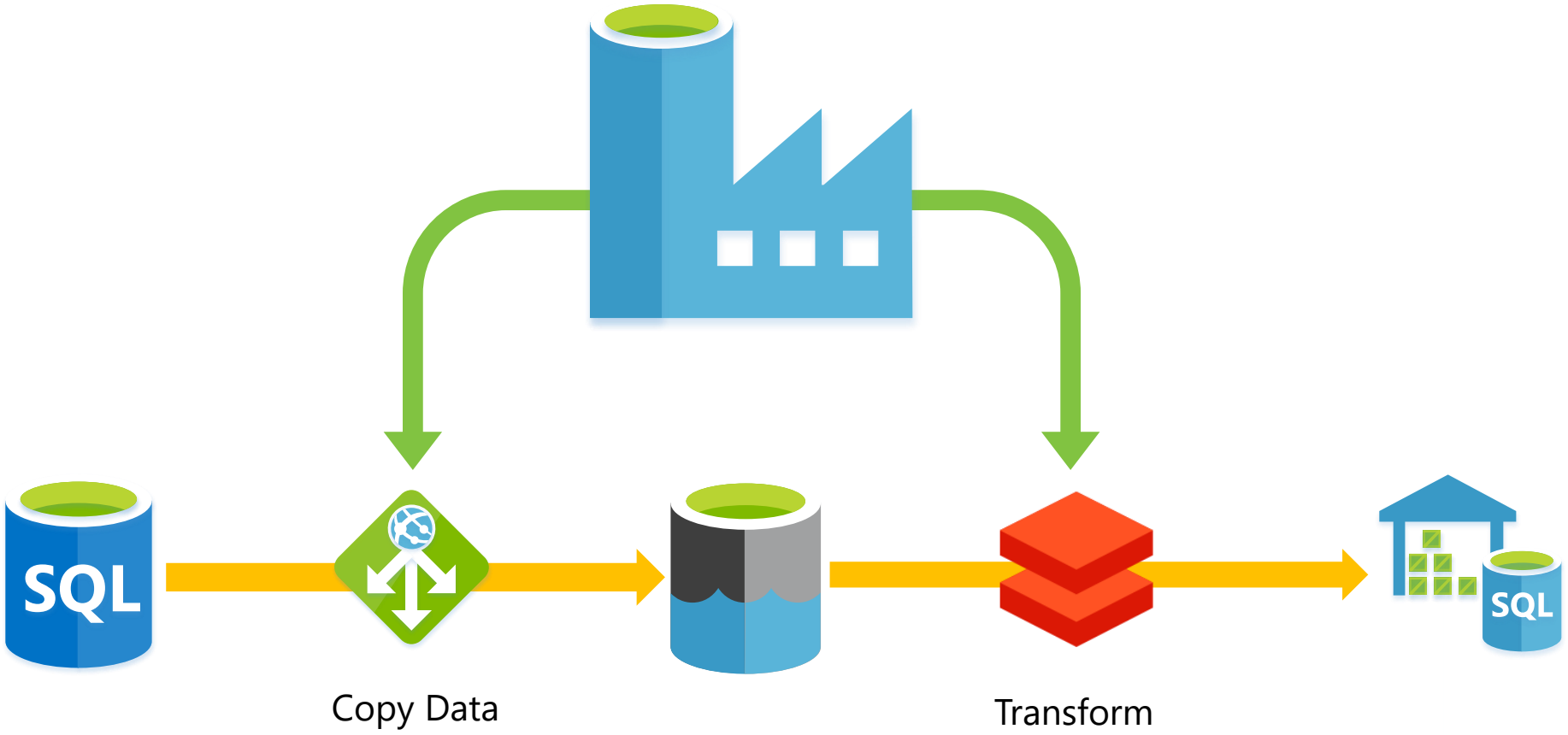


Azure Data Factory

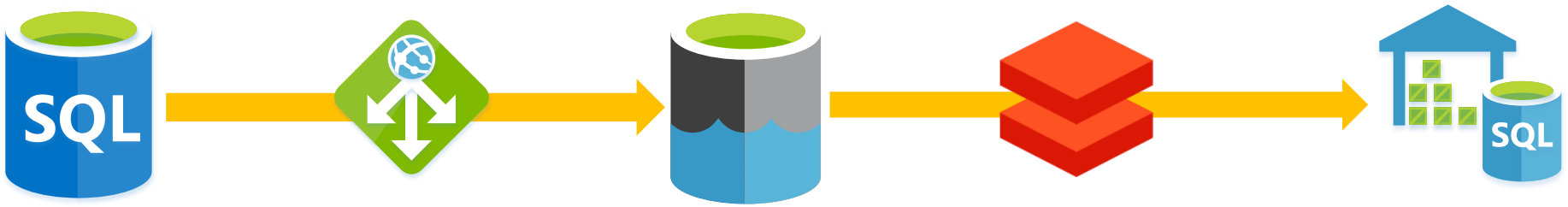


What is Azure Data Factory?

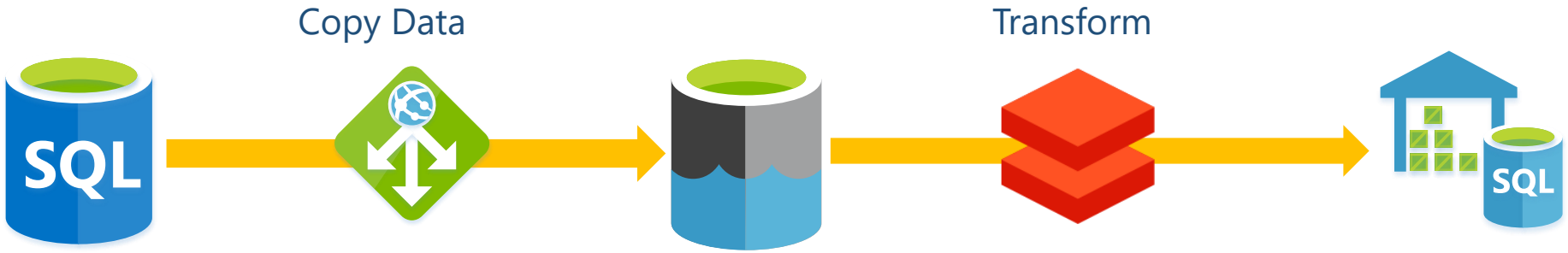
What is Azure Data Factory?



What is Azure Data Factory?



Data Factory Components



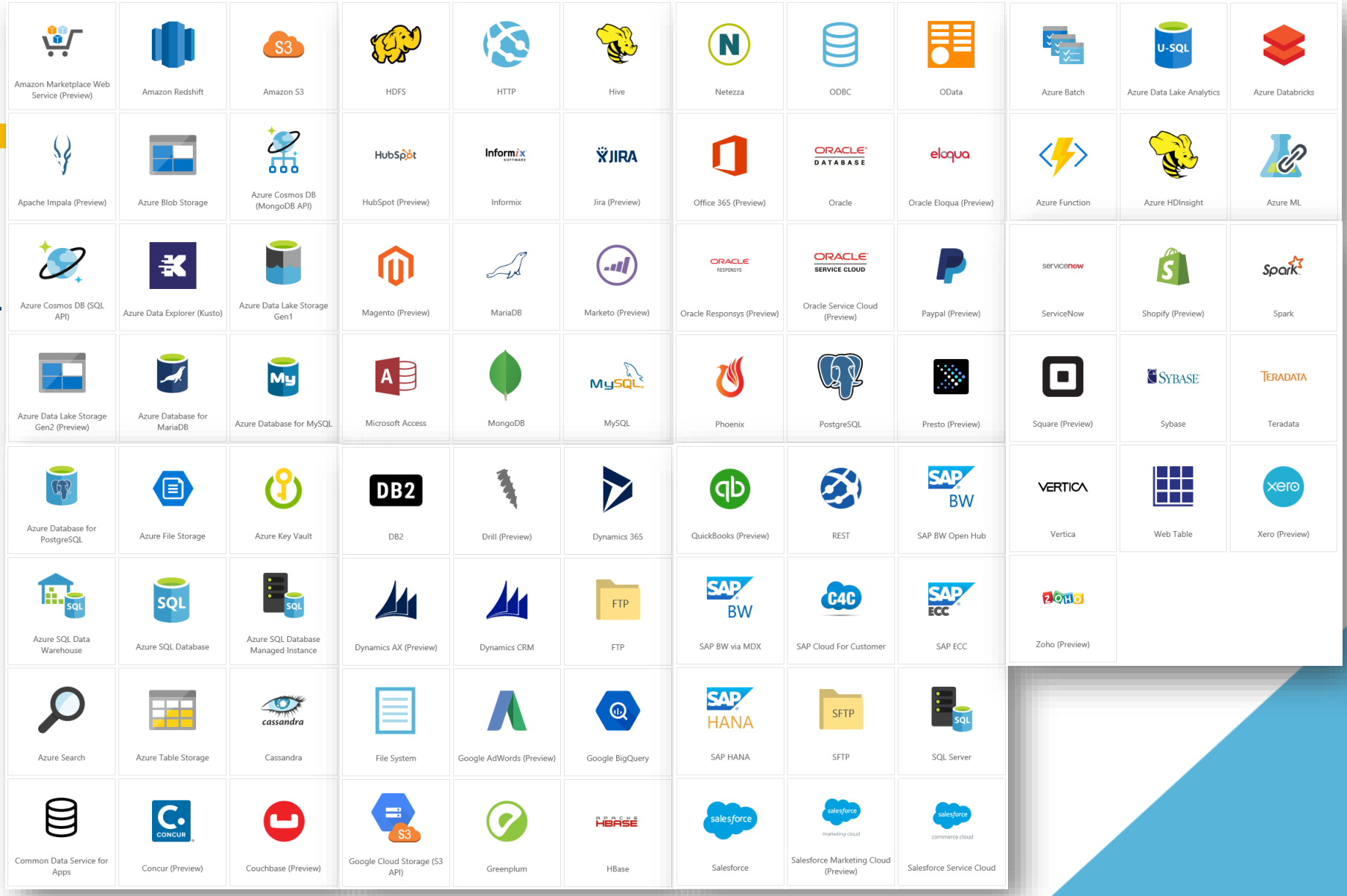
1 **Linked Services** – How and what to connect to. Like the SSIS connection manager.

SQLDBLinkedService

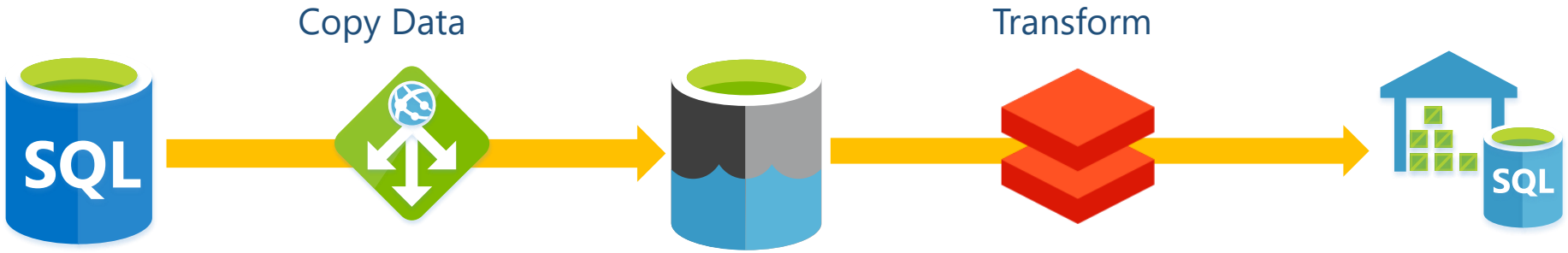
ConnectionString: *Server=MyServer;Database=myDataBase*
UserName: "MrPaulAndrew"
Password: **********

Data Factory Components

1 Linked Services –

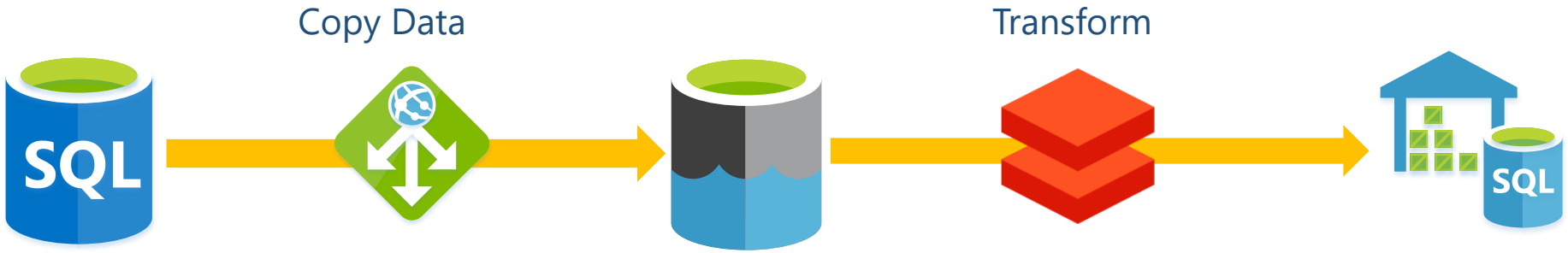


Data Factory Components



1 **Linked Services**

Data Factory Components

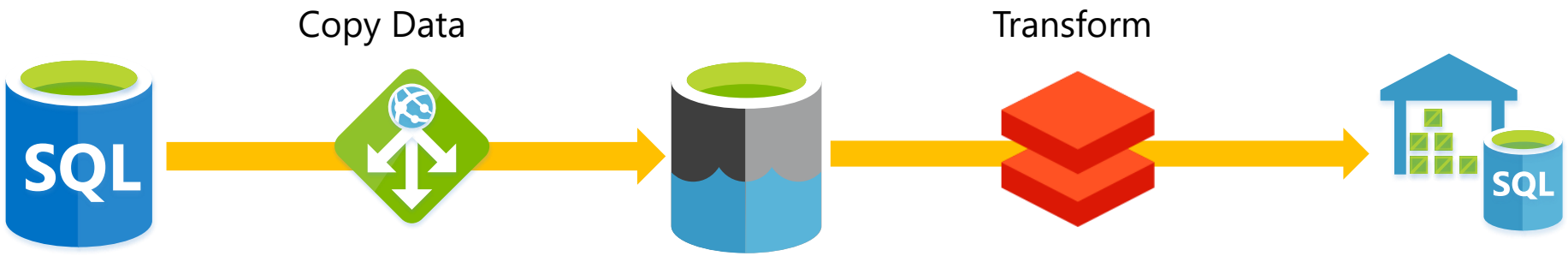


1. **Linked Services**
2. **Data Sets** – Where is my data? What format? What file path/table do I need?

SQL dbo.DimCustomer

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

Data Factory Components

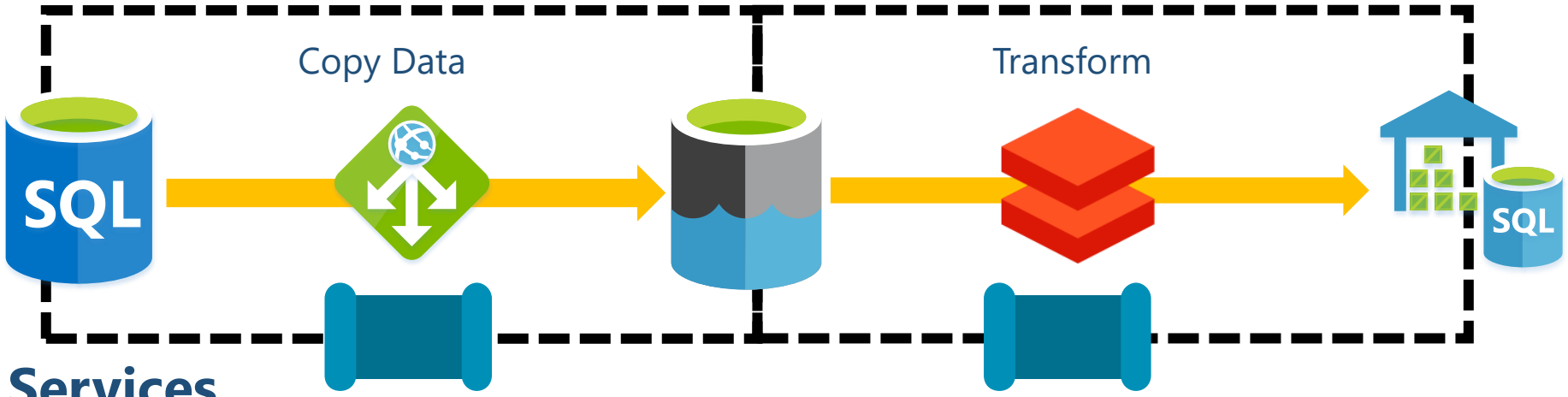


1. **Linked Services**
2. **Data Sets**
3. **Activities** – What do we want to happen? With what conditions?

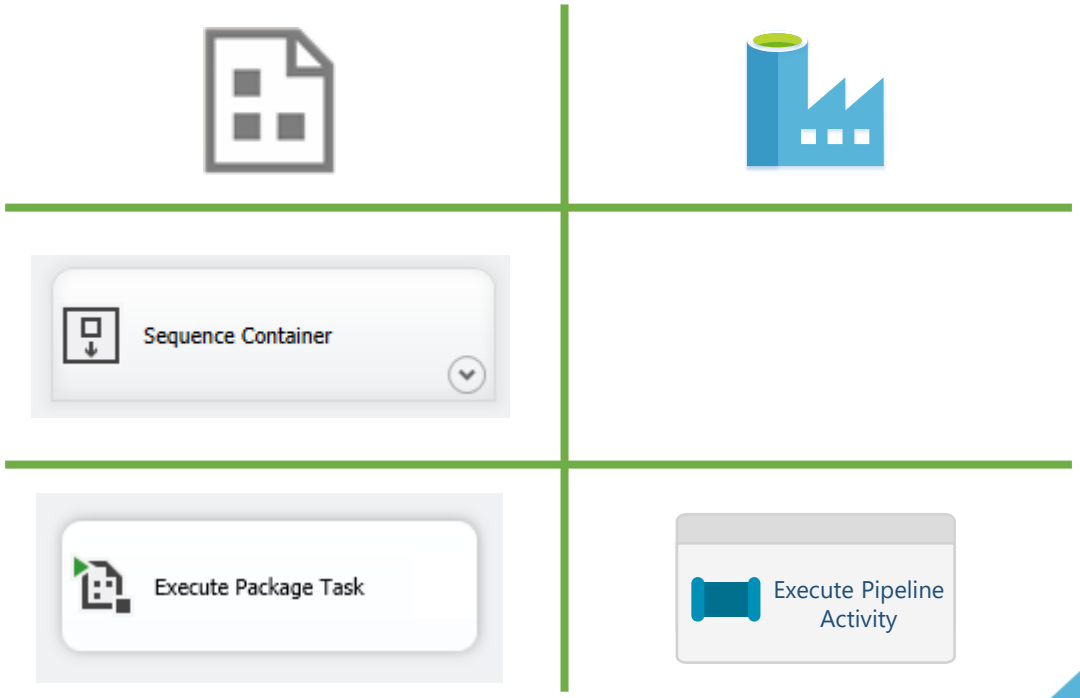
Databricks Notebook Activity

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

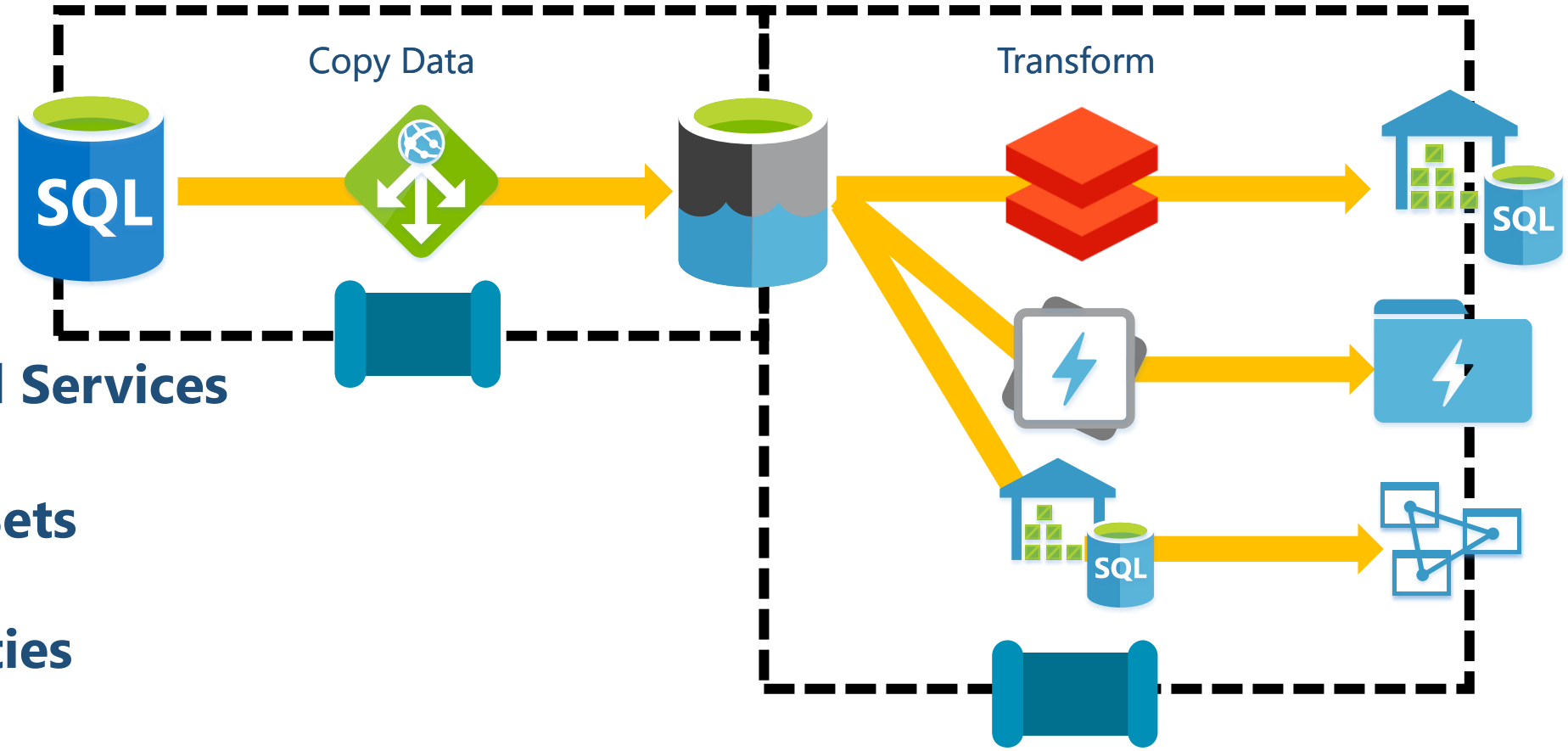
Data Factory Components



1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines** – What groups of work do I want to do?

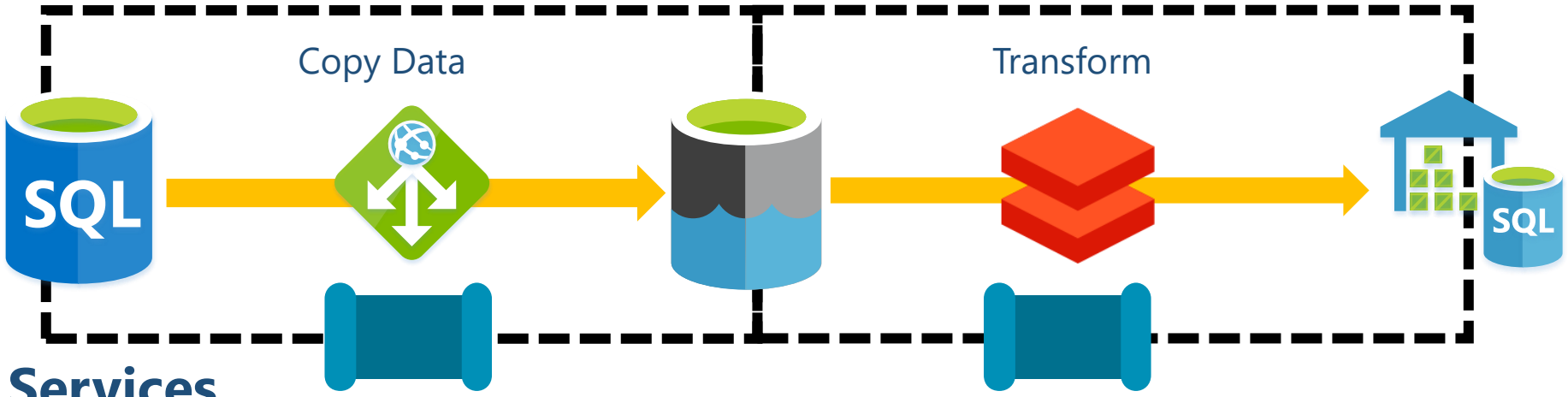


Data Factory Components



1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines** – What groups of work do I want to do?

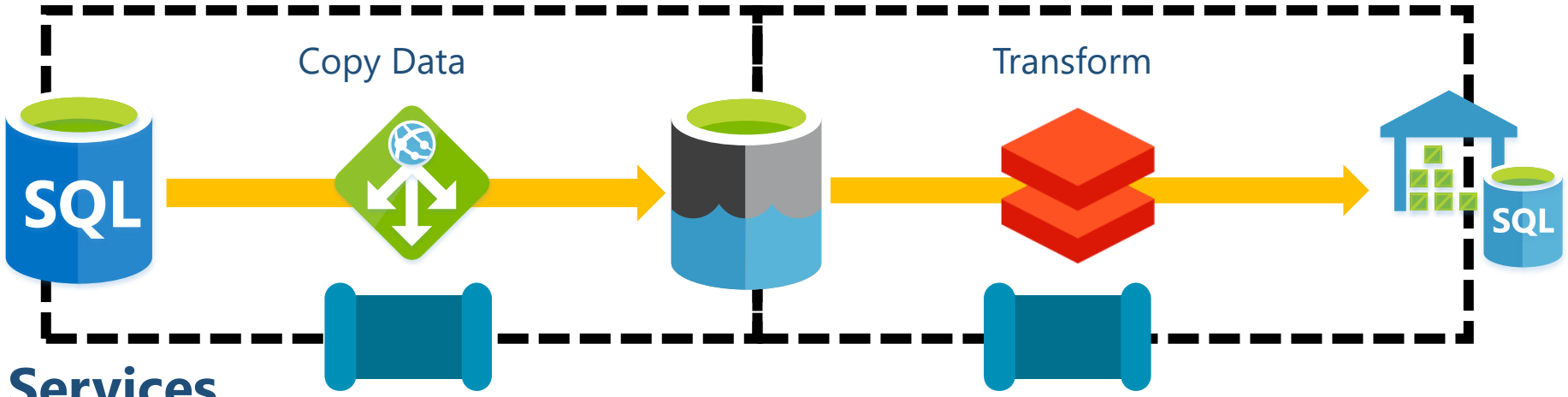
Data Factory Components



1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers** – How are we going to tell our pipeline(s) to execute?

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

Data Factory Components

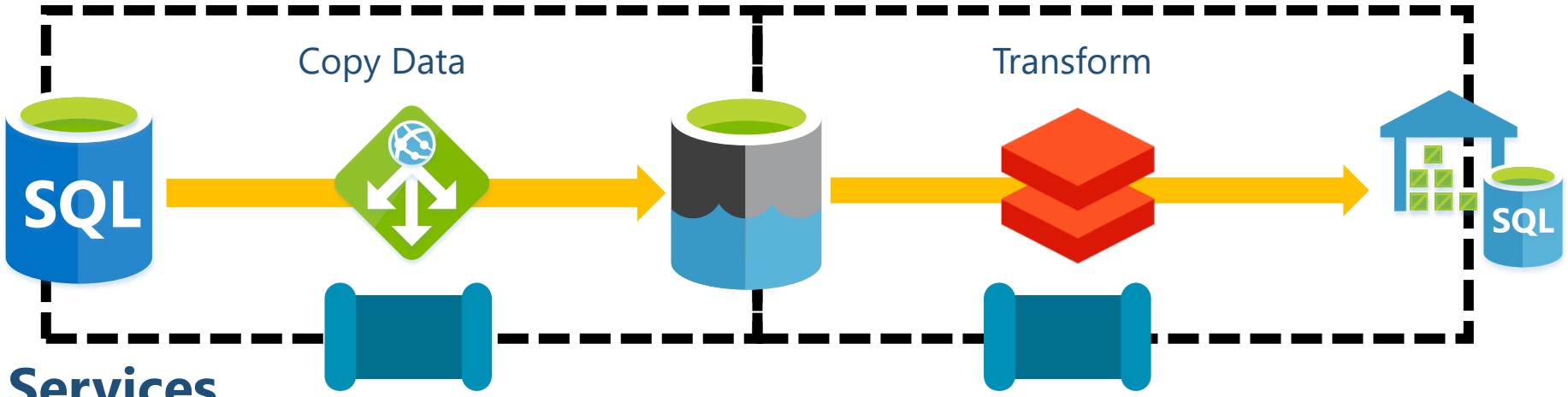


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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

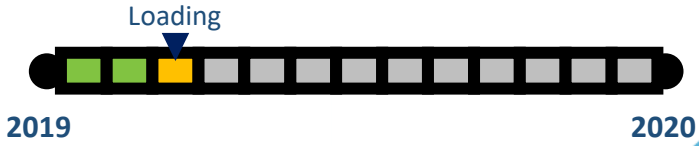
```
Invoke-AzureRmDataFactoryV2Pipeline
-DataFactoryName $dataFactoryName
-ResourceGroupName $resourceGroupName
-PipelineName $pipelineName
```

Data Factory Components

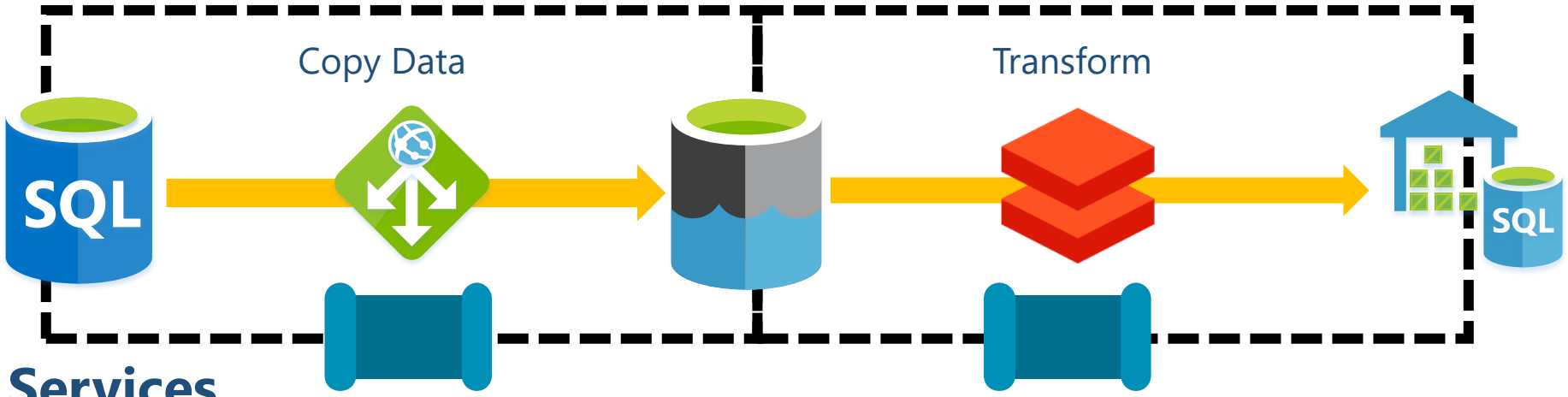


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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



Data Factory Components

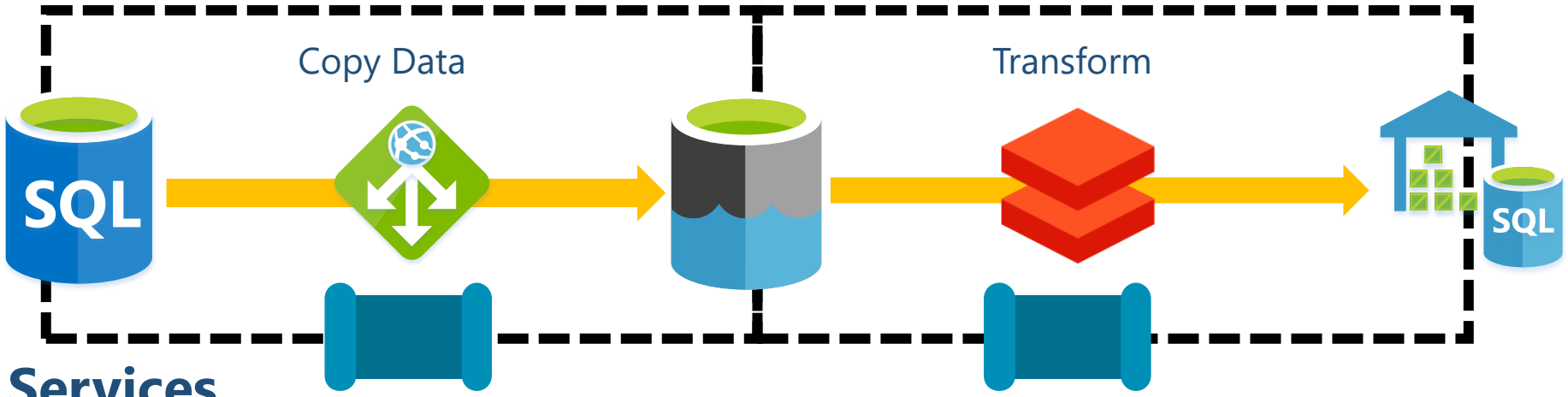


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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

- Every 1 minute.
- UTC

Data Factory Components

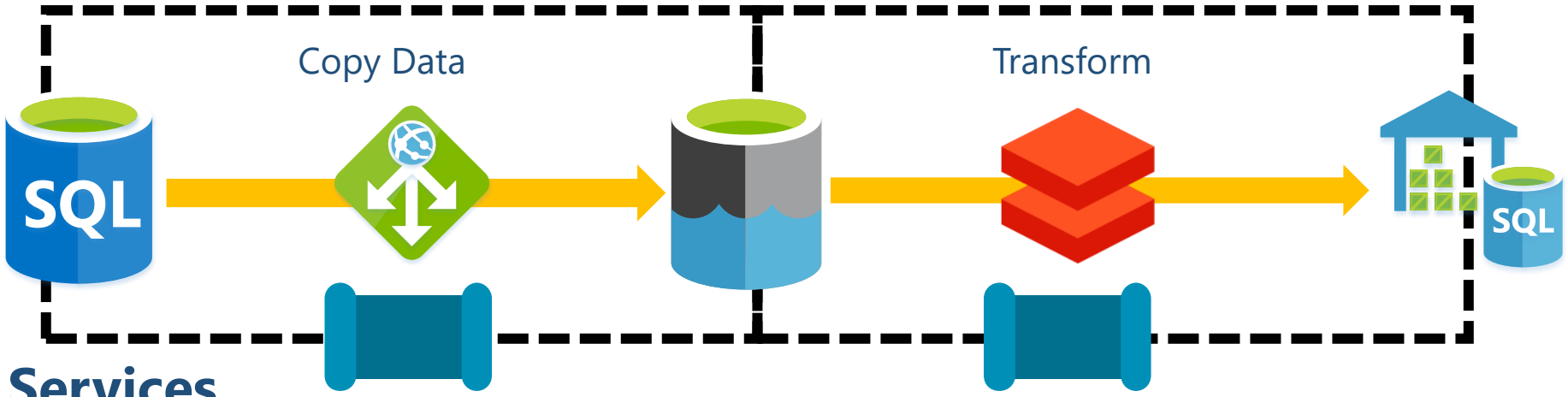


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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

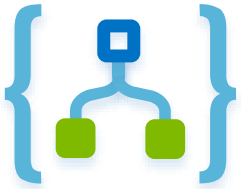
{Path} Created
{Path} Deleted

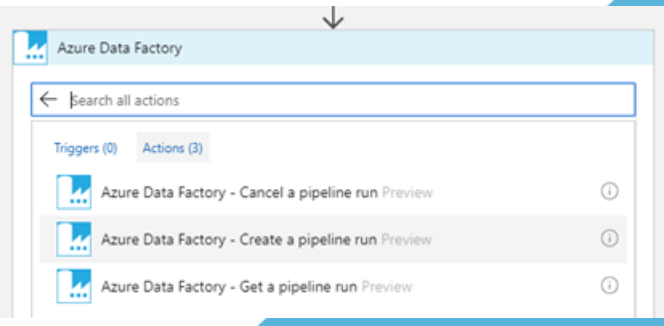
Data Factory Components



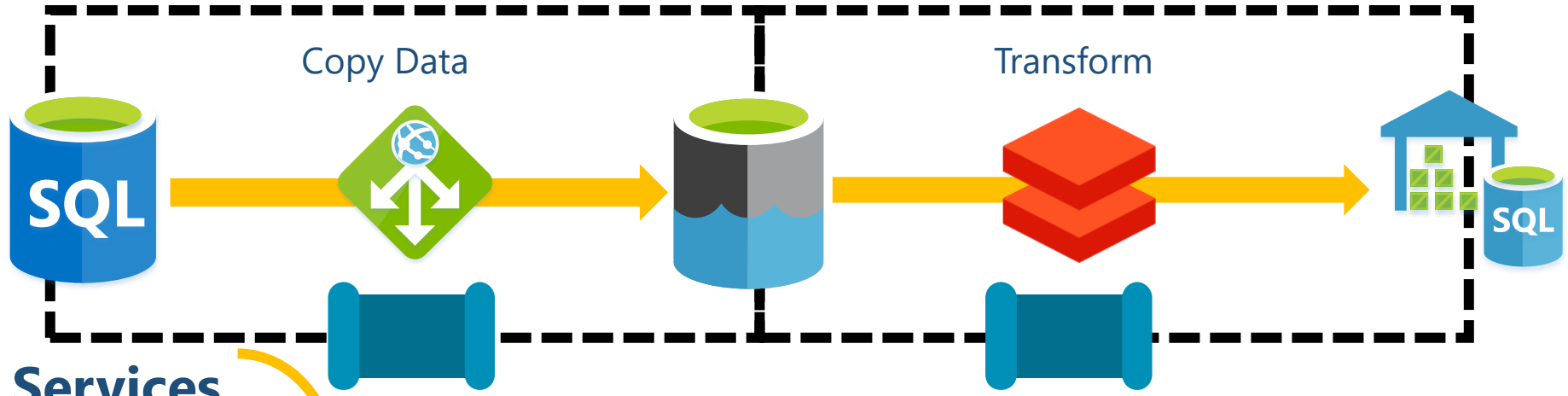
1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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



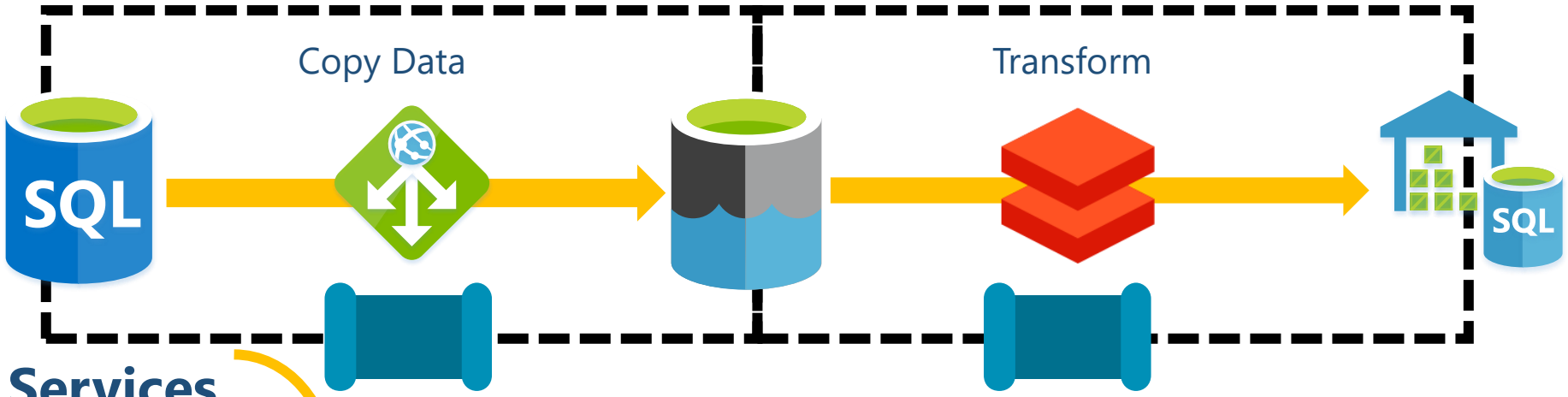


Data Factory Components



1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

Data Factory Control Flow Components



1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

Integration Runtimes

1 **Azure** Integration Runtime

Movement Hours

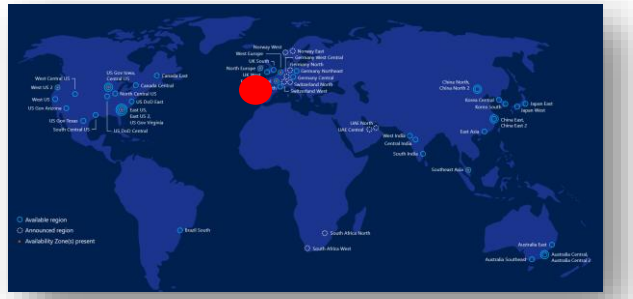
Activity Orchestration

Flexible Region

2 **SSIS** Integration Runtime

SSIS Package Execution

Specified Region



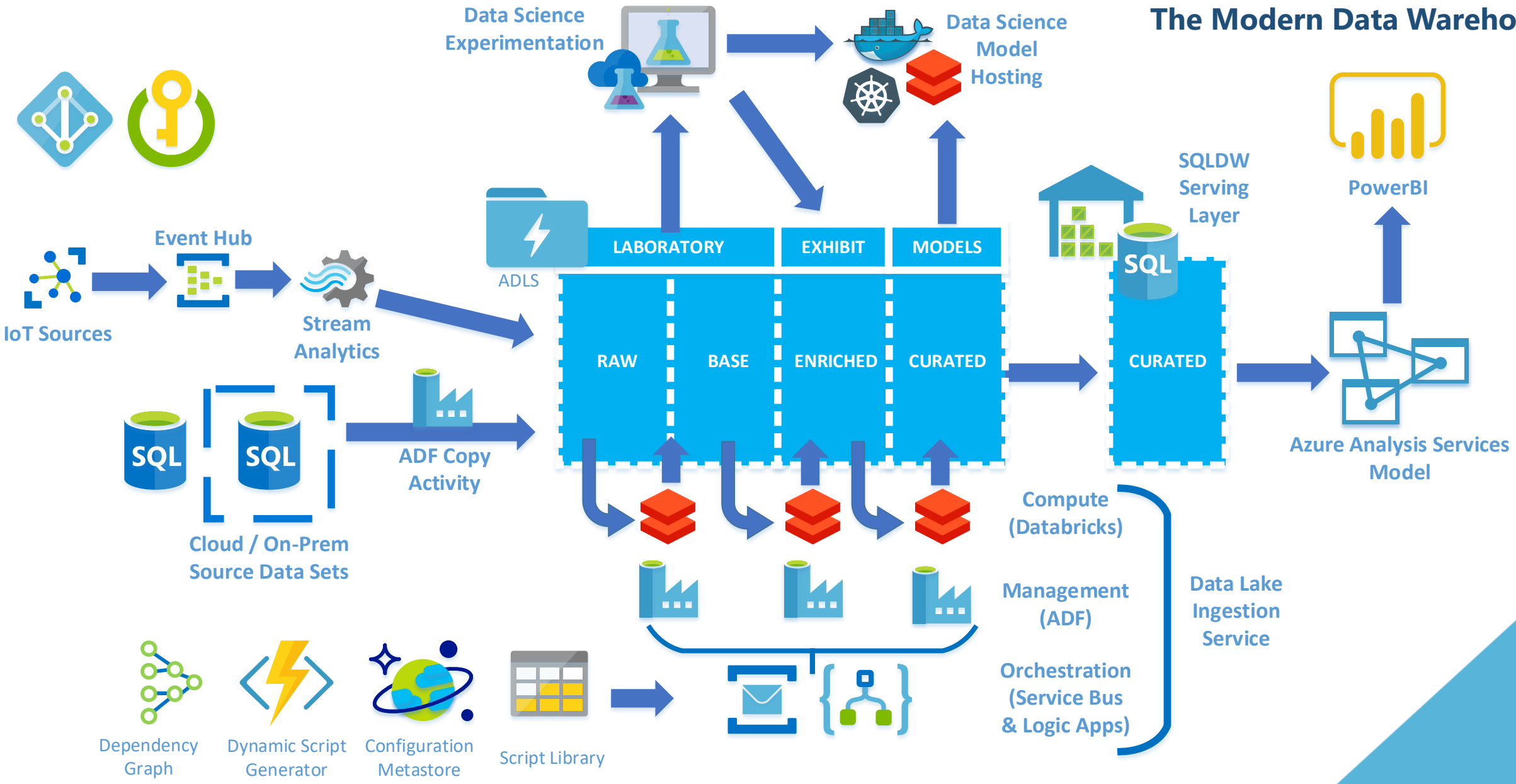
3 **Self Hosted** Integration Runtime

Gateway Access

Activity Orchestration

Virtual Machine

Why use Azure Data Factory?

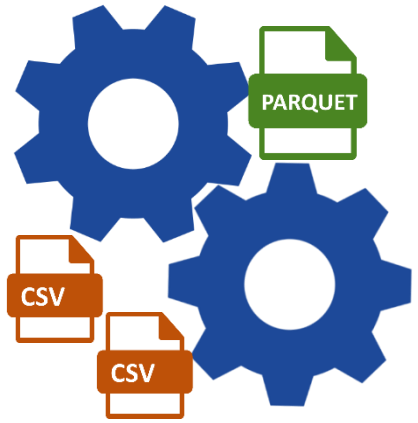


Data Factory What & Why - Recap

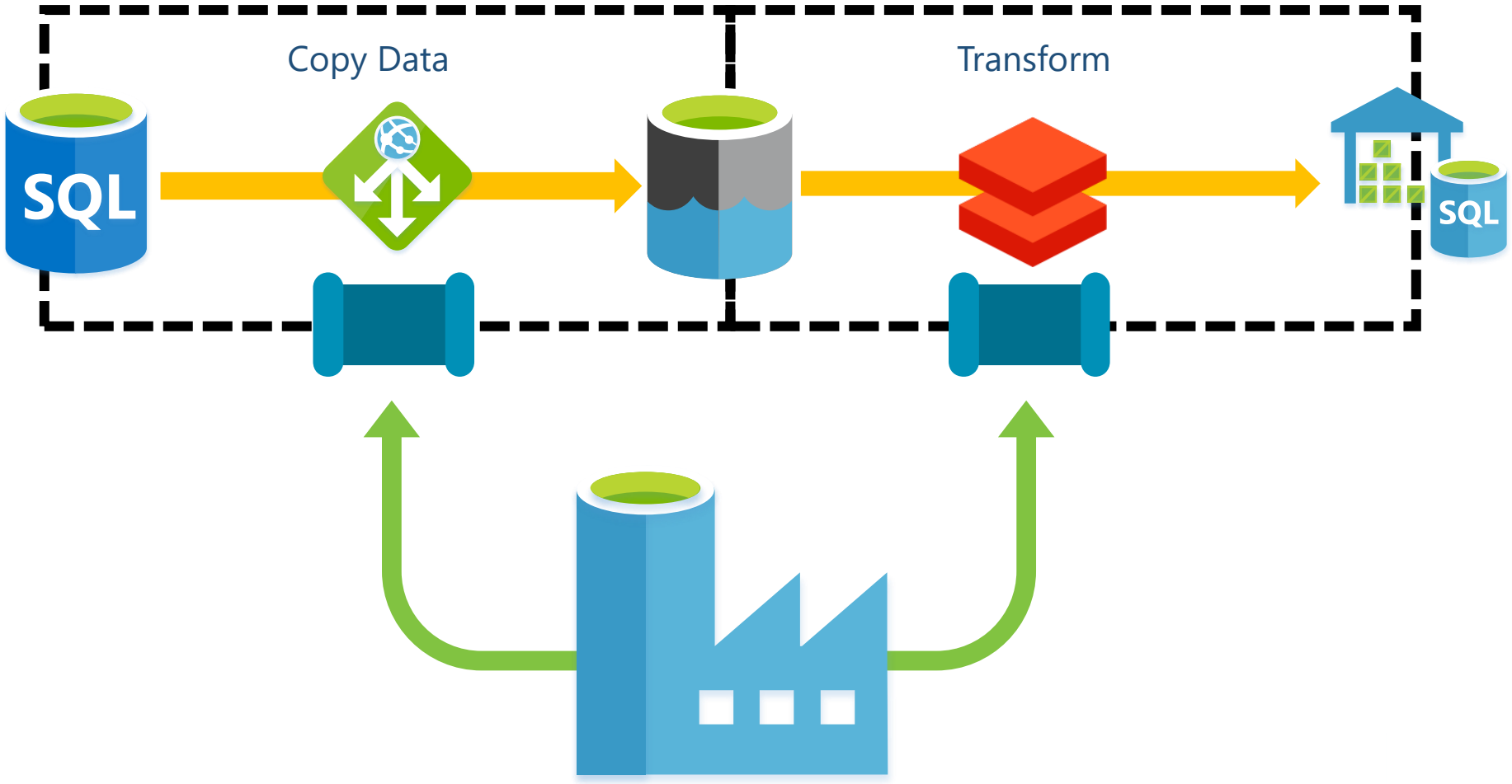
1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

1. **Azure** Integration Runtime
2. **SSIS** Integration Runtime
3. **Self Hosted** Integration Runtime

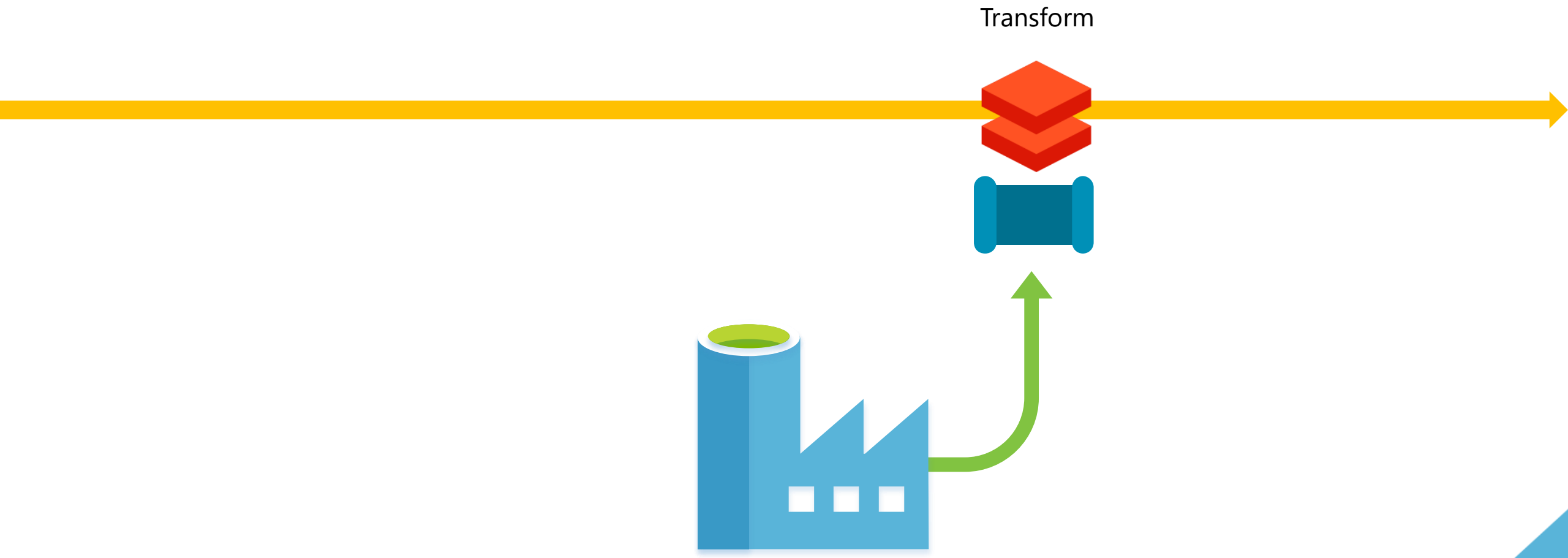
Data Transformation in Azure



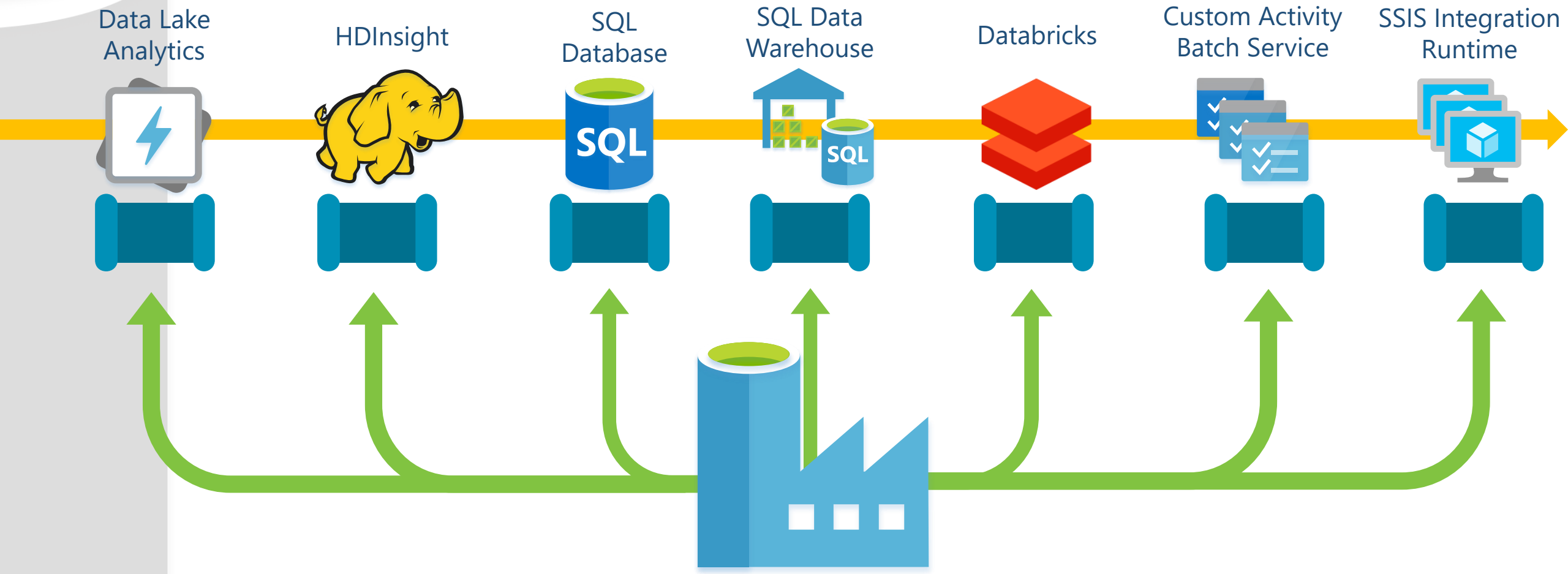
Data Factory Control Flow Components



Data Transformation in Azure



Data Transformation in Azure



Future Uncertain

Expensive Clusters

Only Scales Up

Requires at least 1TB

Learn Spark, Python/Scala

Custom Apps on IaaS

IaaS VMs 20min start

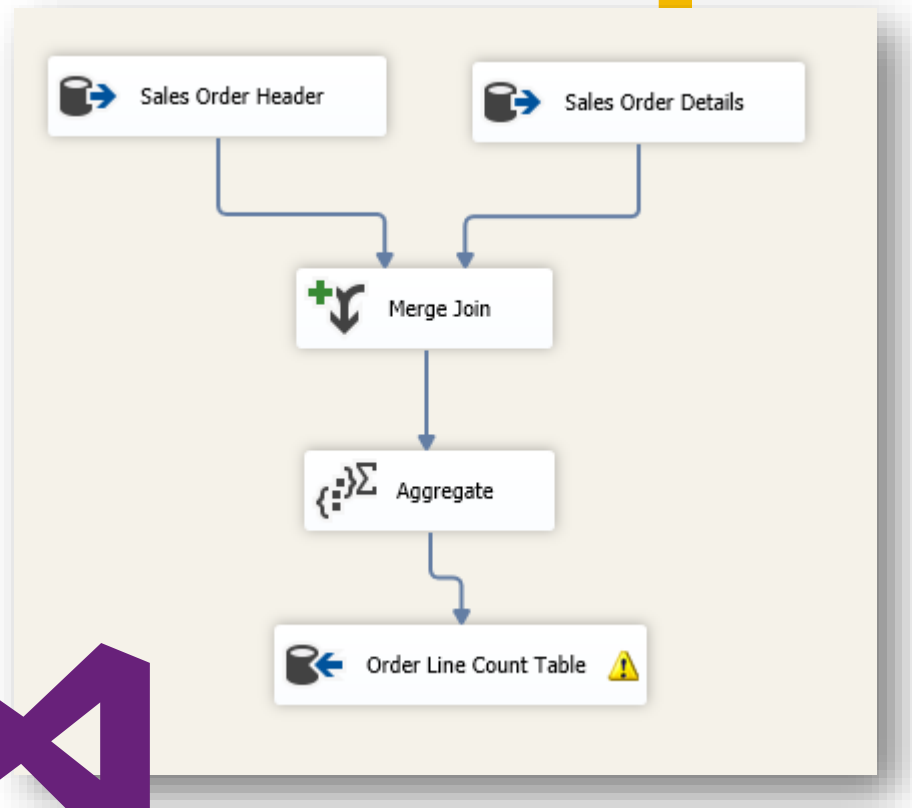
Data Transformation in Azure with SSIS

SSIS Package

8x Packages per Node

10x Nodes

SSIS Integration Runtime



Visual Studio – SQL Server Data Tools (SSDT)

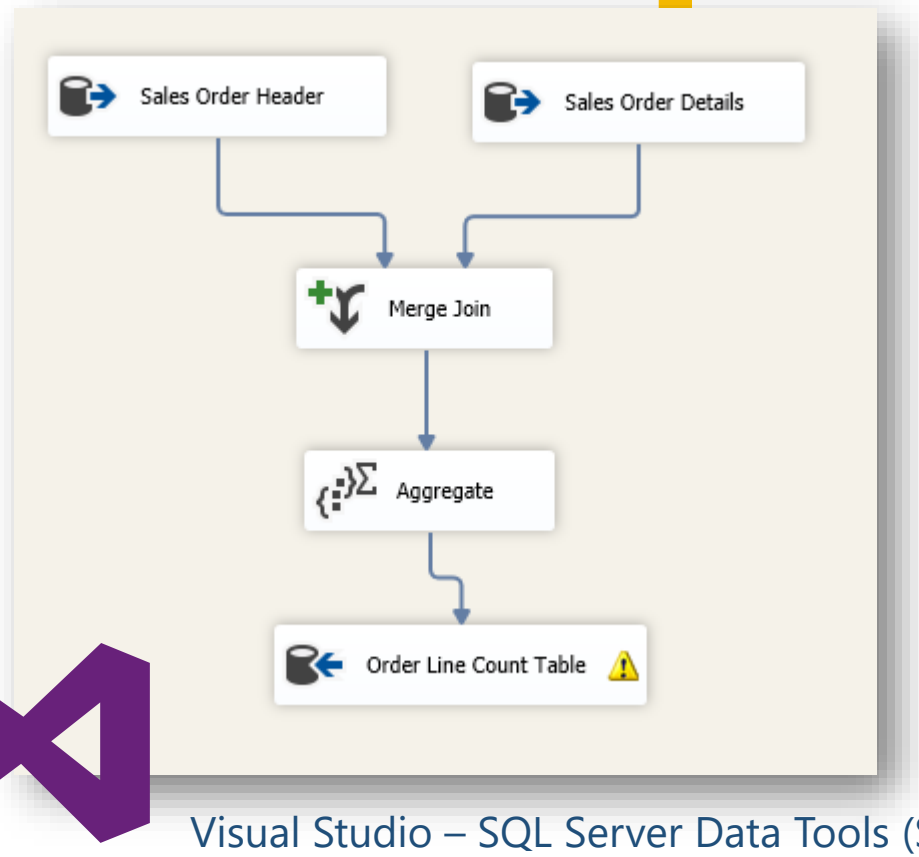
Data Factory

Logical or Managed Instance

SSISDB

Data Transformation in Azure with SSIS

SSIS Package



Visual Studio – SQL Server Data Tools (SSDT)

8x Packages per Node

10x Nodes

SSIS Integration Runtime

Express Route

VNet

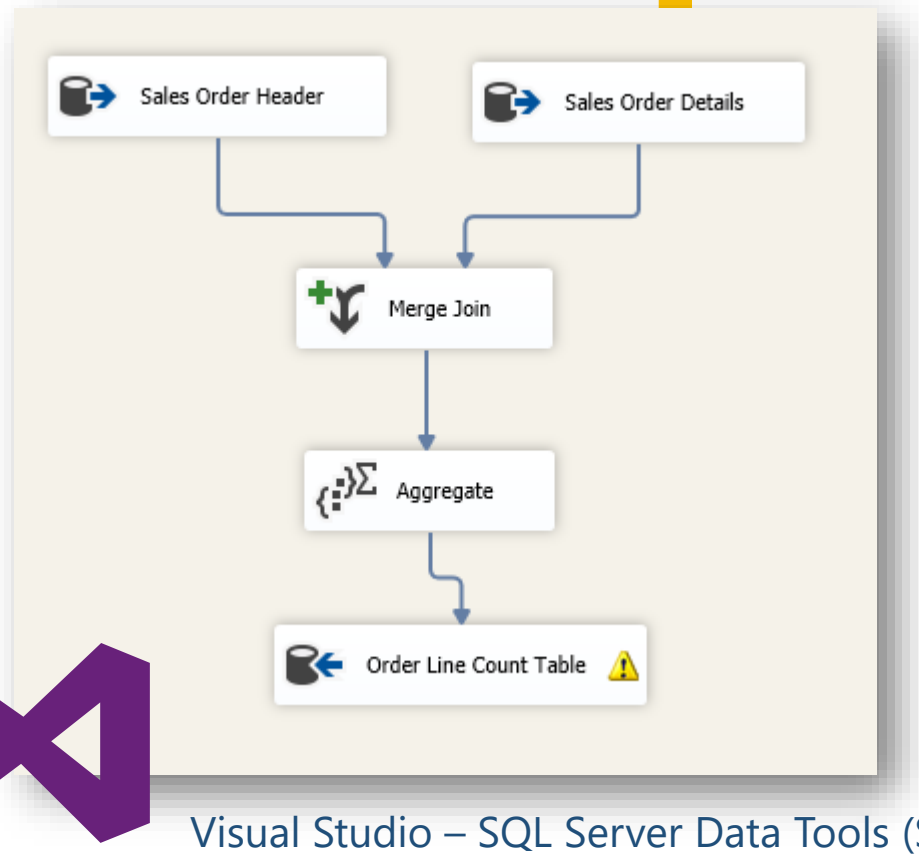
Data Factory

Logical or Managed Instance

SSISDB

Data Transformation in Azure with SSIS

SSIS Package



Visual Studio – SQL Server Data Tools (SSDT)

1x Package per Node

D64_v3
64 Cores
256GB RAM
£1,700 a month

1x Node

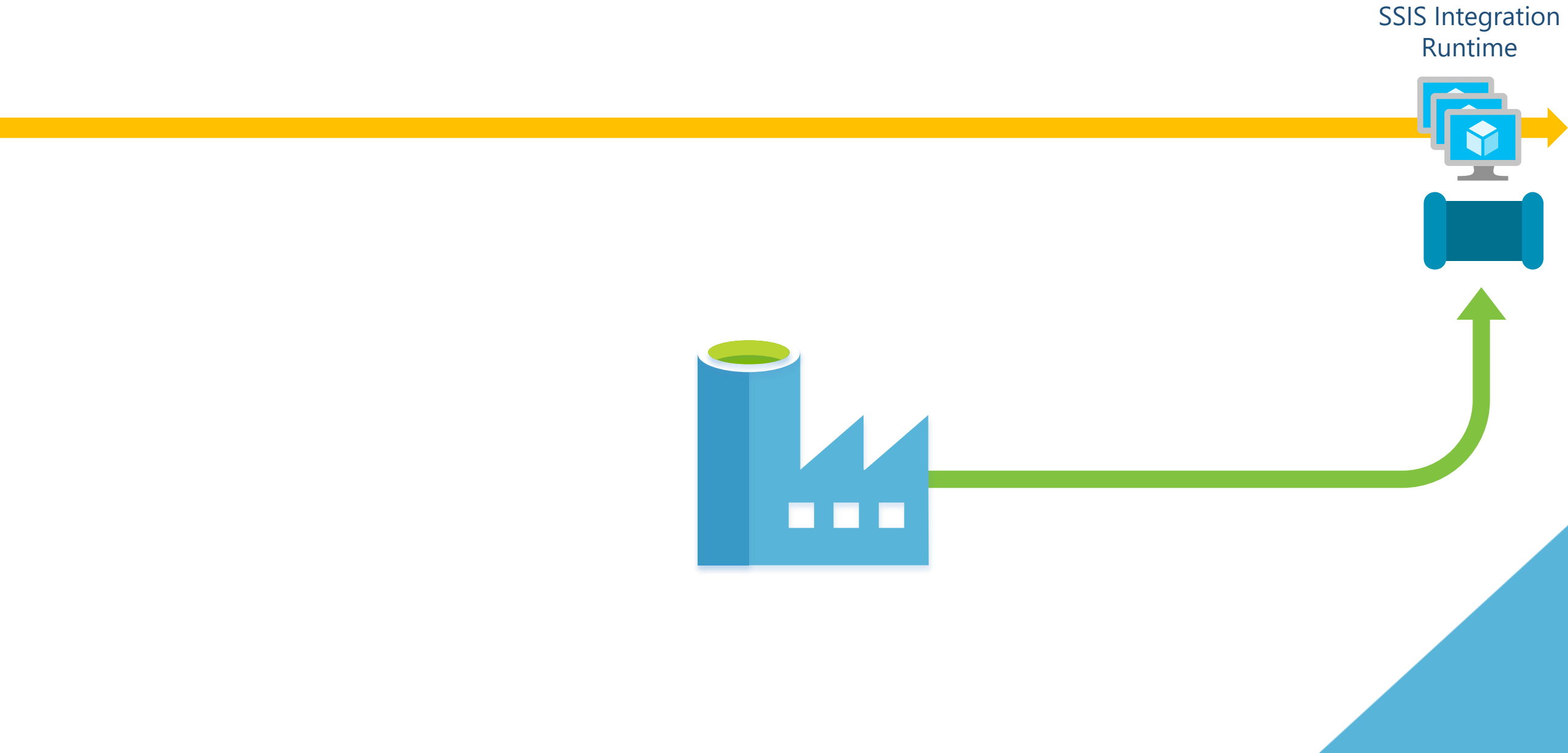
SSIS Integration Runtime

Data Factory

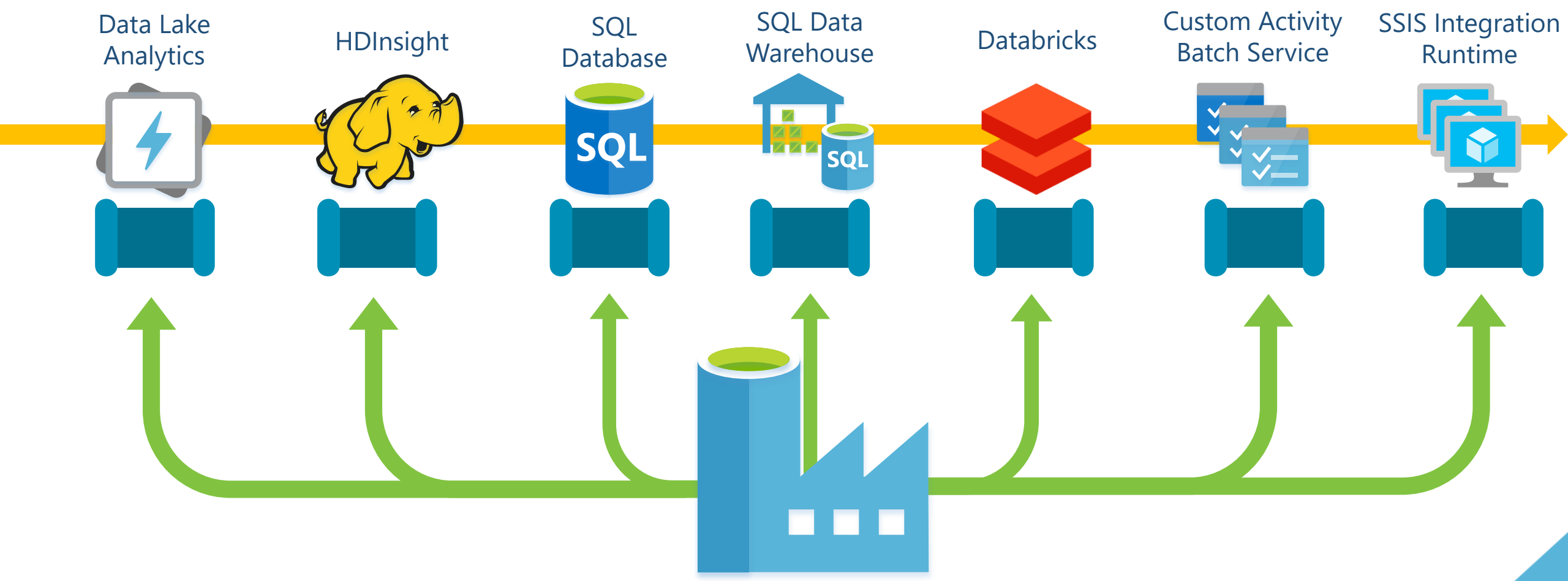
Logical or Managed Instance

SSISDB

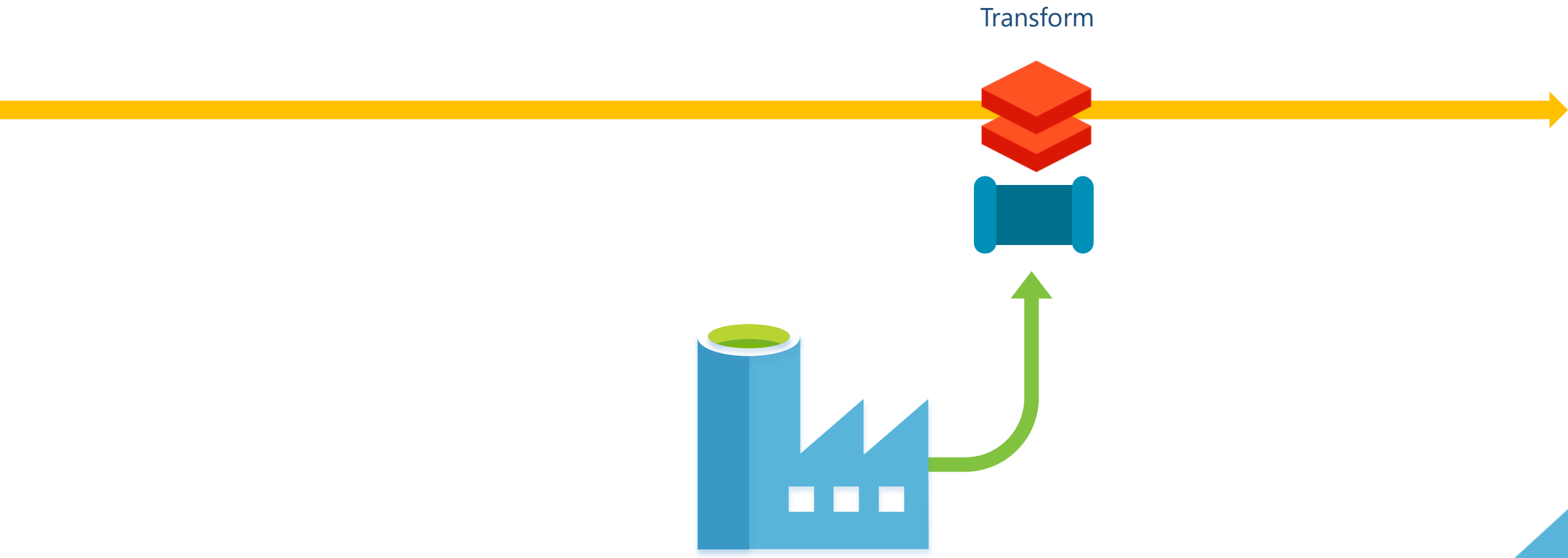
Data Transformation in Azure



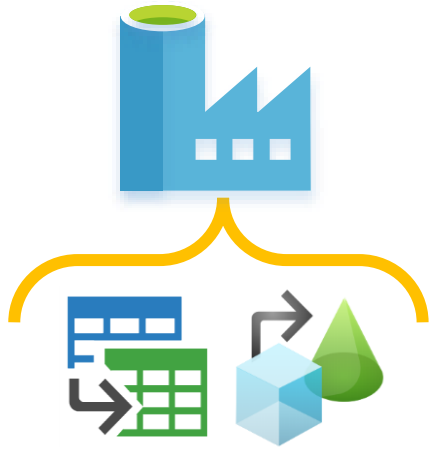
Data Transformation in Azure



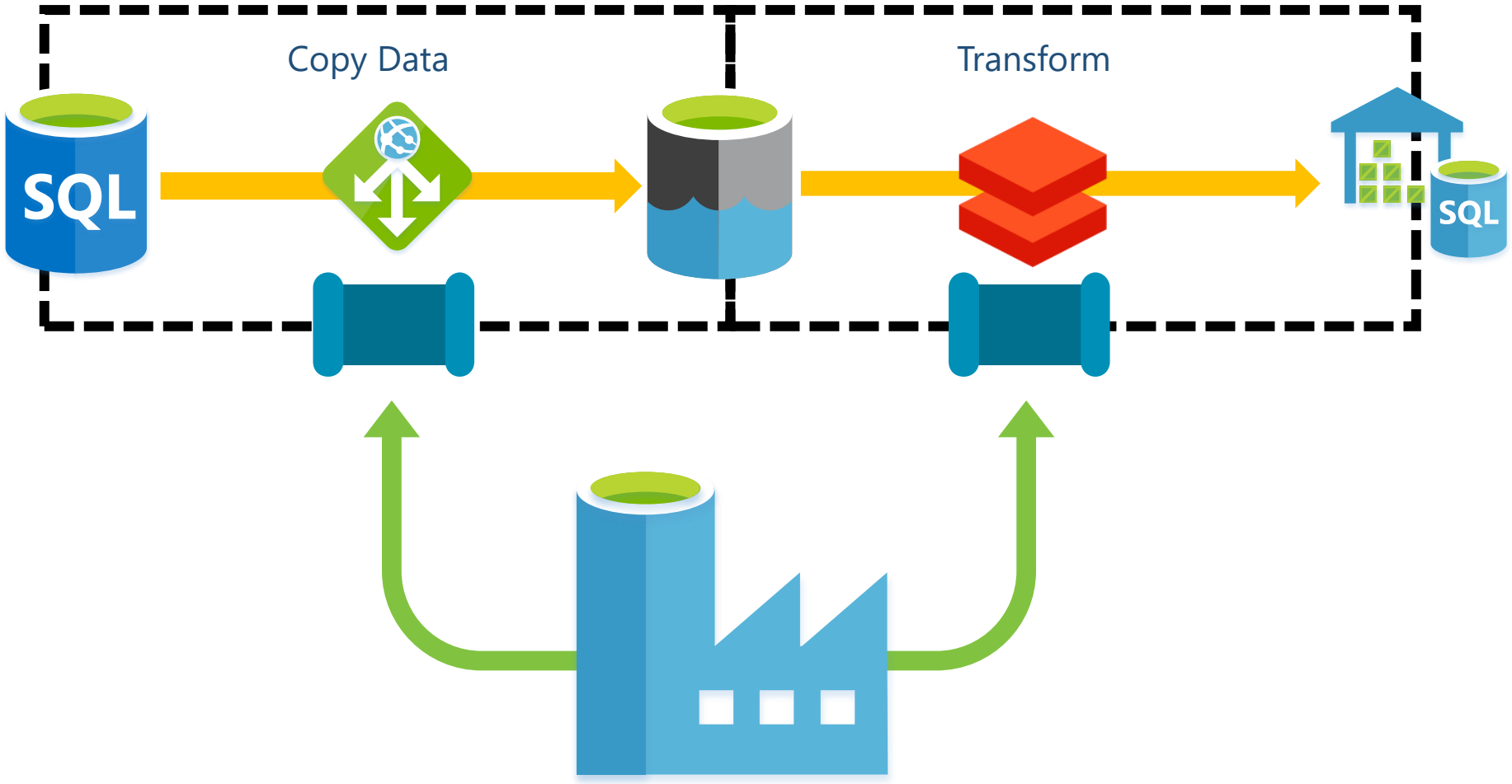
Data Transformation in Azure



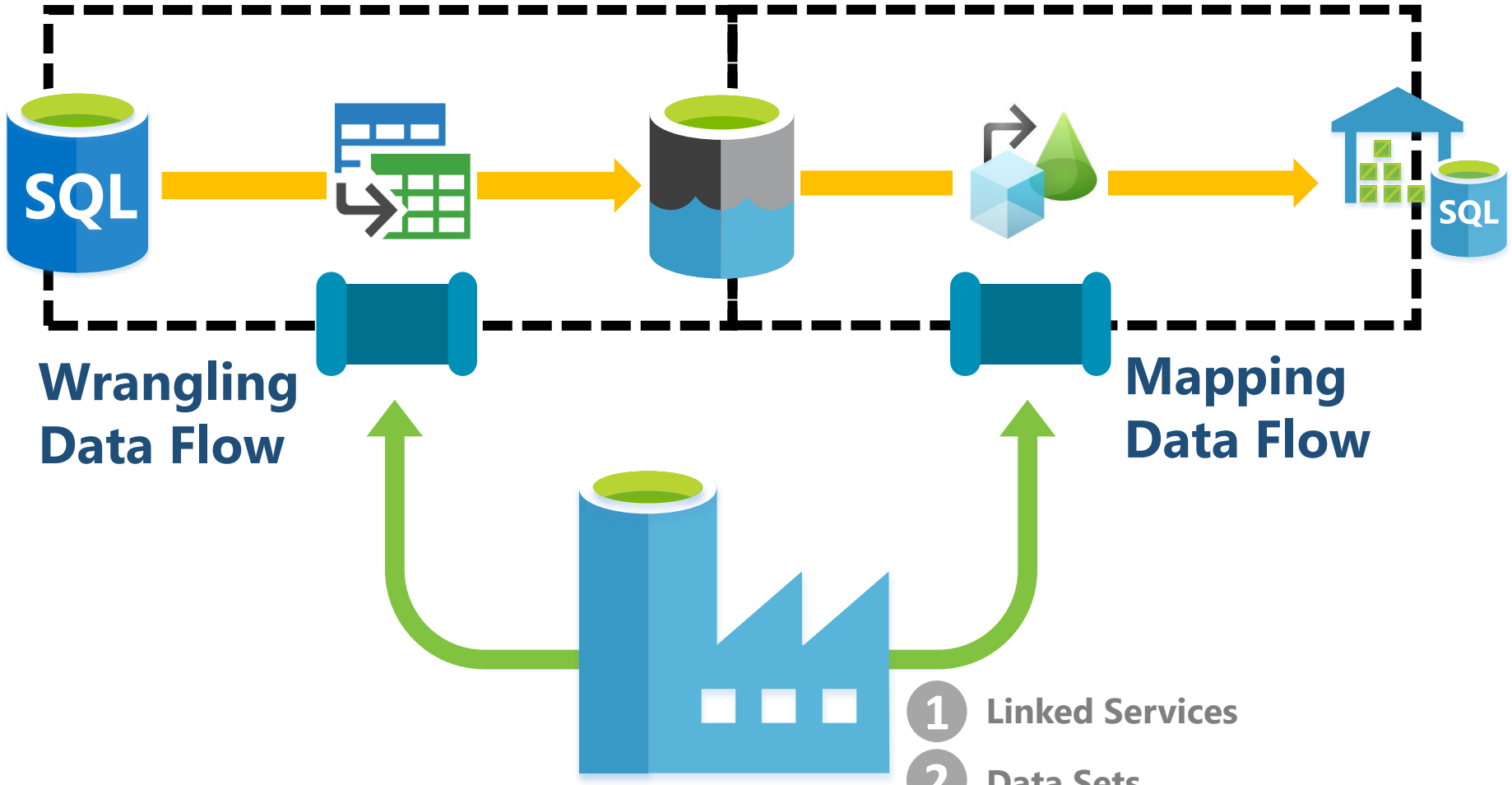
Data Factory Data Flows



Data Factory Control Flow Components

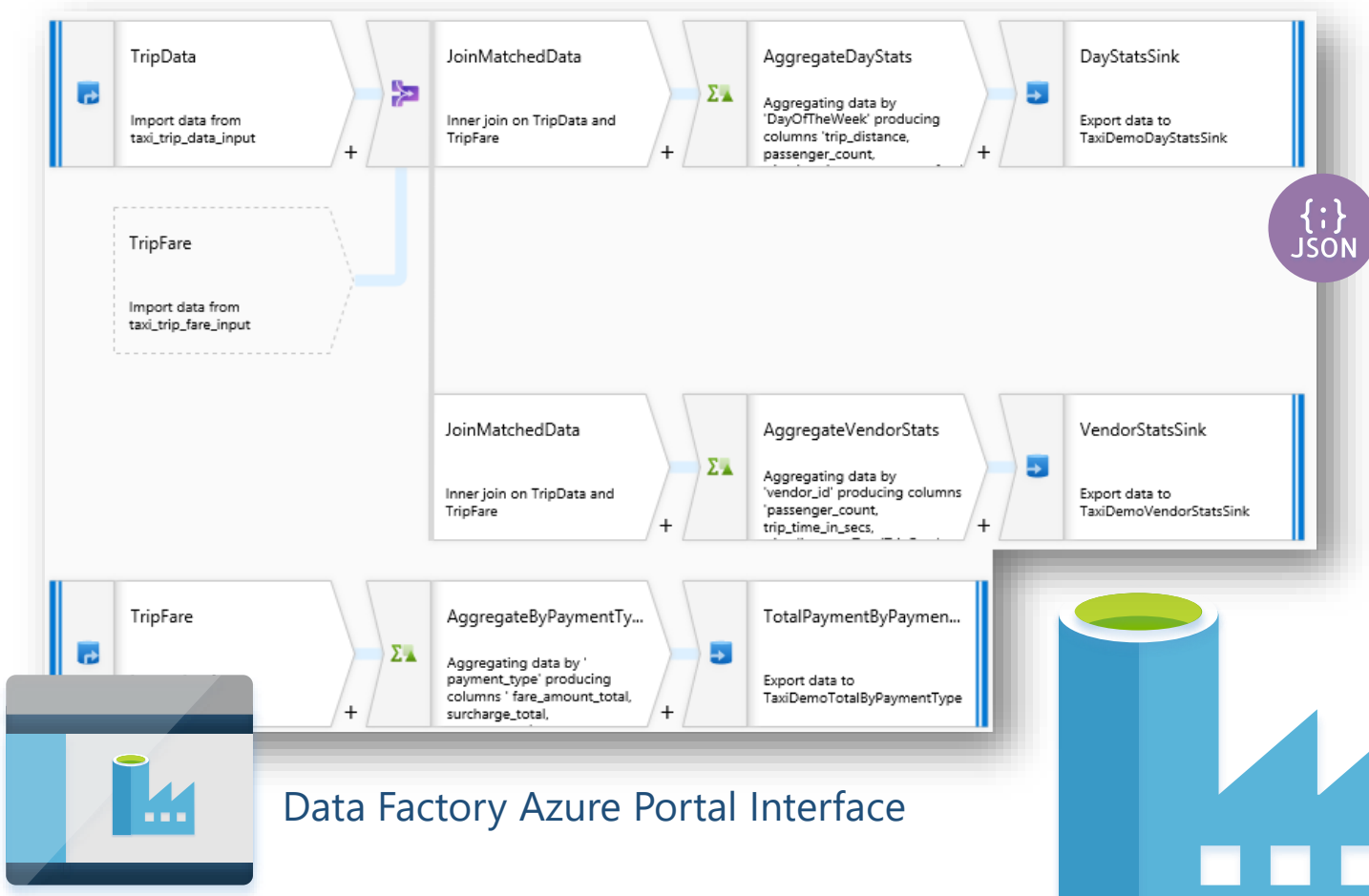


Data Factory Data Flows



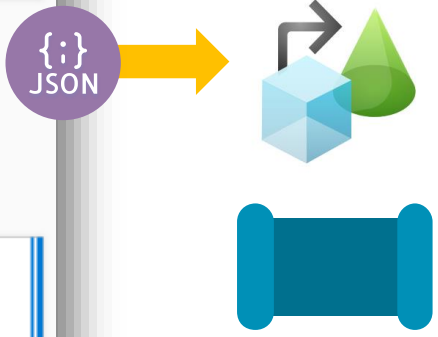
Mapping Data Flows

What is a Mapping Data Flow?

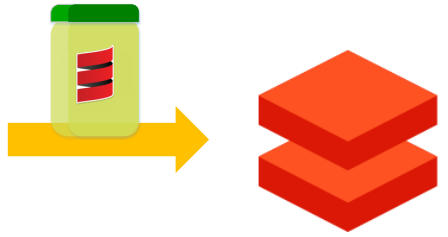


Data Factory Azure Portal Interface

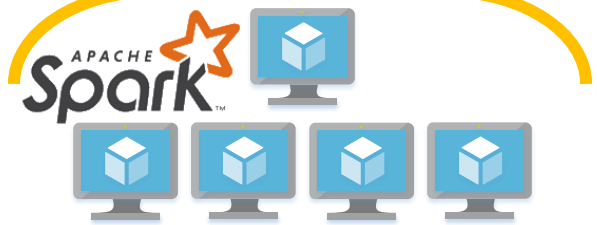
Data Factory



Mapping Data Flow

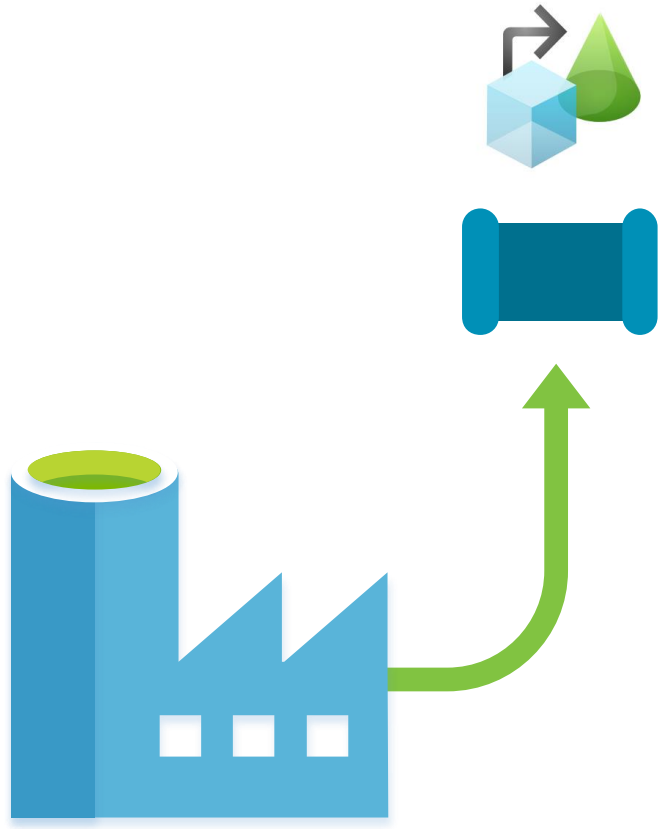


Azure Databricks

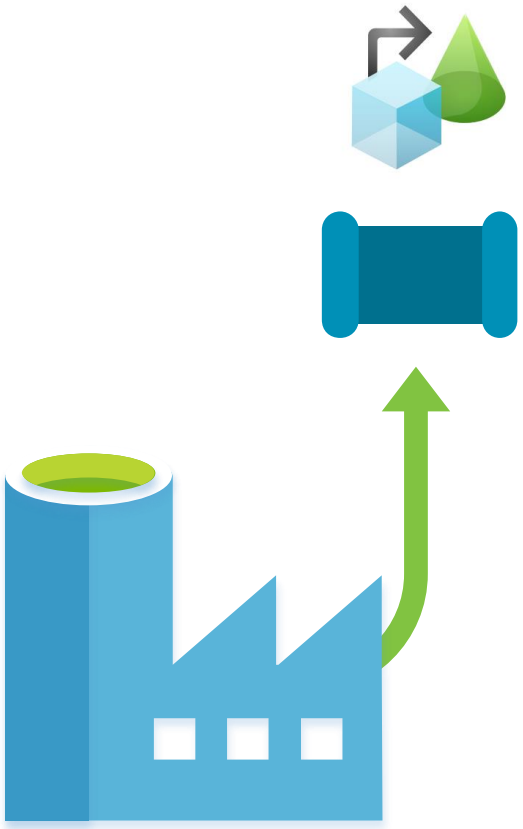


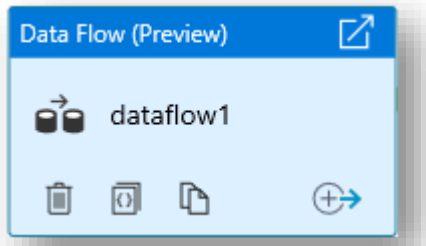
https://mrpaulandrew.com/2018/10/22/what-is-azure-data-factory-data-flow/

Mapping Data Flows

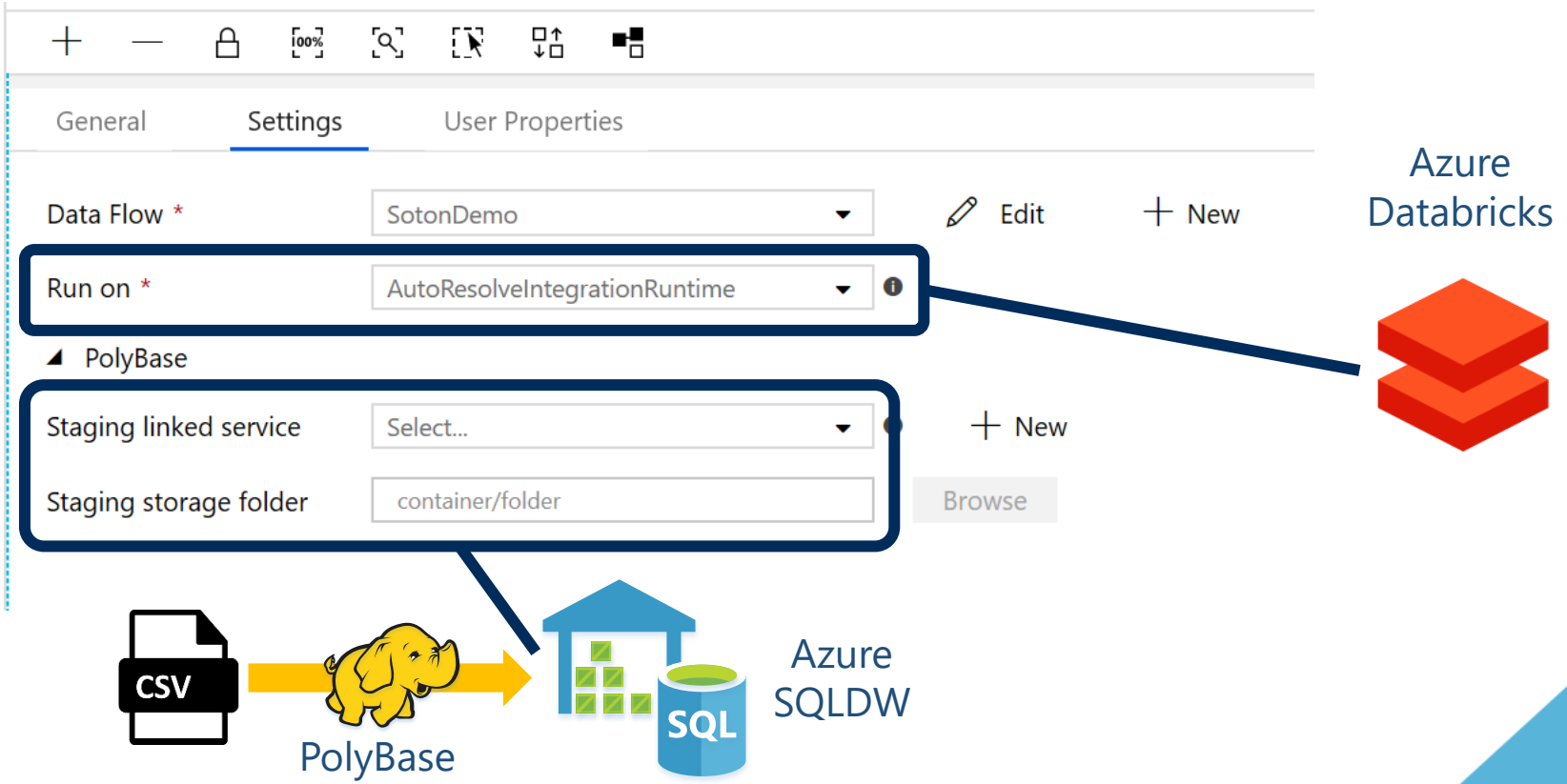


Mapping Data Flows – Settings & Concepts





New Control Flow Activity



Integration Runtimes

1 **Azure** Integration Runtime

Movement Hours

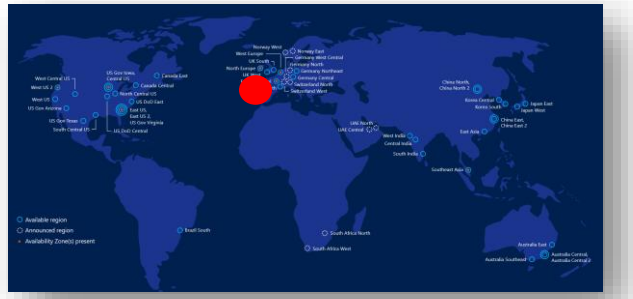
Activity Orchestration

Flexible Region

2 **SSIS** Integration Runtime

SSIS Package Execution

Specified Region



3 **Self Hosted** Integration Runtime

Gateway Access

Activity Orchestration

Virtual Machine

Integration Runtimes – Mapping Data Flow Cluster

1

Azure
Integration Runtime

Movement Hours

Activity
Orchestration

Flexible Region

Data Flow run time

Compute Type * General Purpose

Core count * 4 (+ 4 Driver cores)

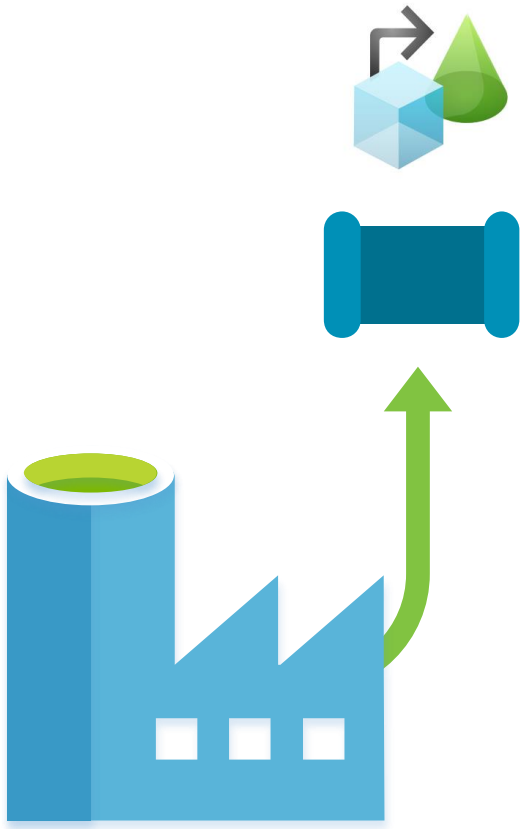
Time to live (in minutes)

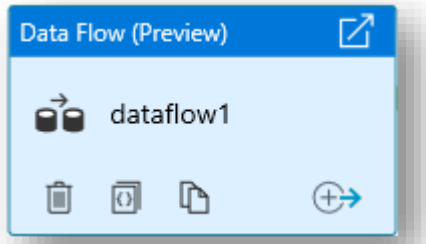
Time to live feature is coming soon

=

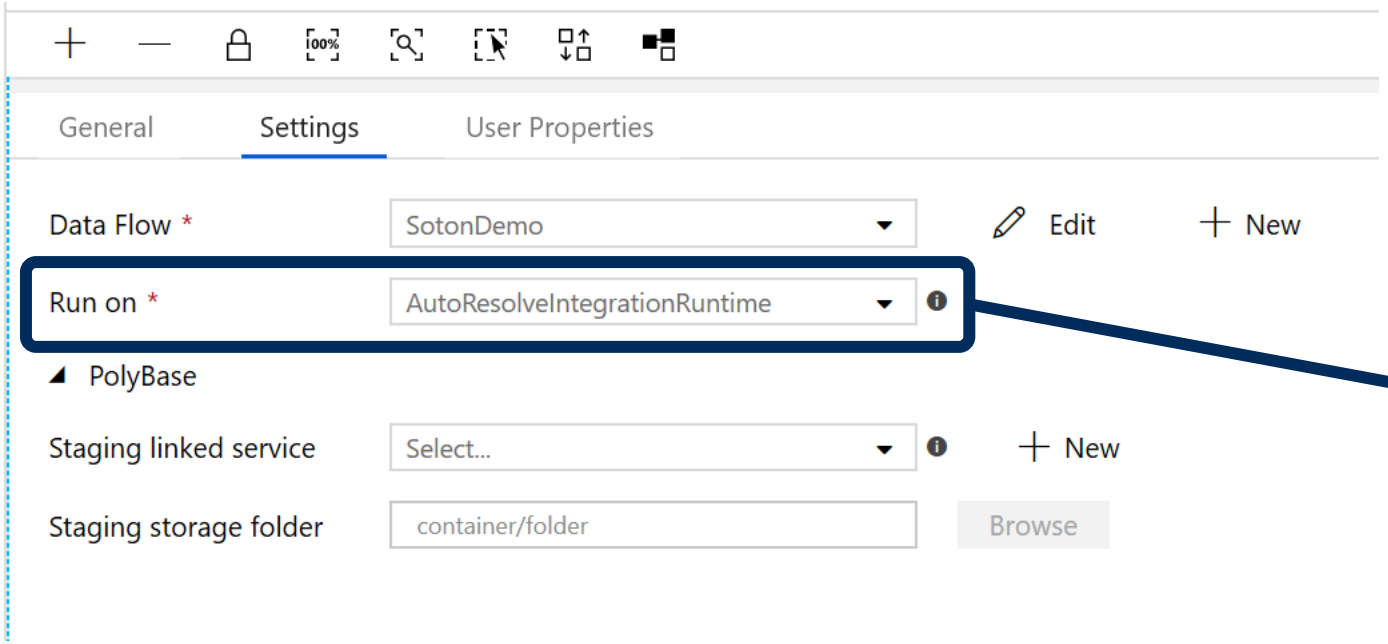
- General Purpose
- Memory Optimised
- Compute Optimised

Mapping Data Flows – Settings & Concepts



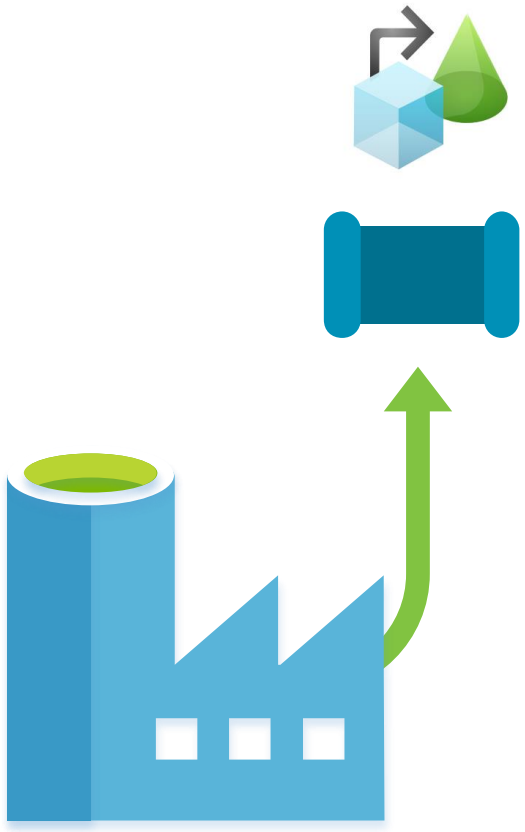


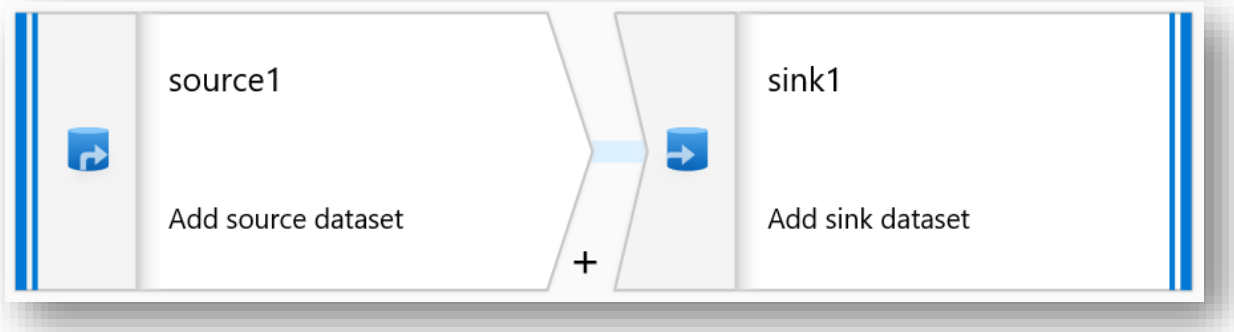
New Control Flow Activity



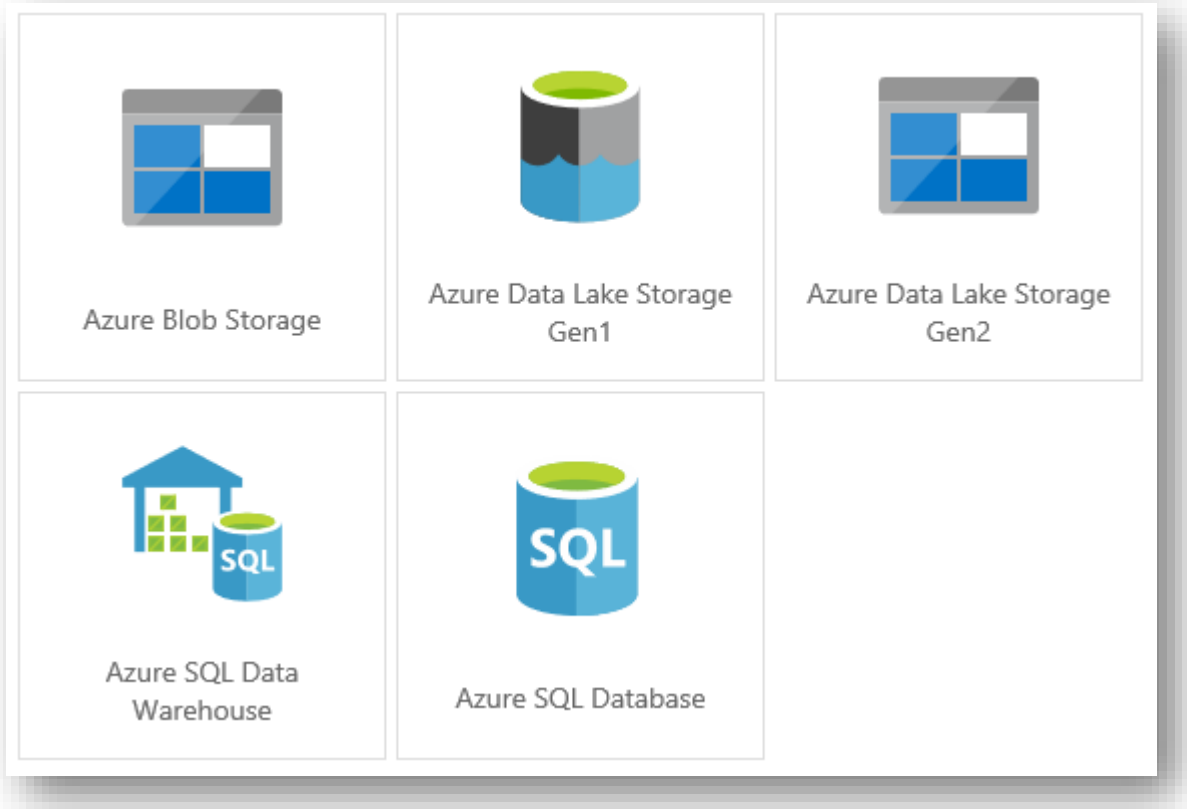
Azure Databricks

Mapping Data Flows – Settings & Concepts

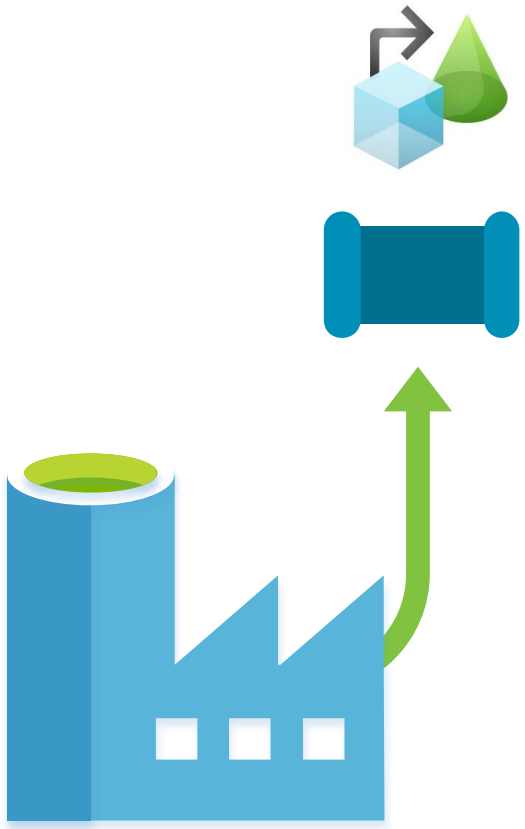


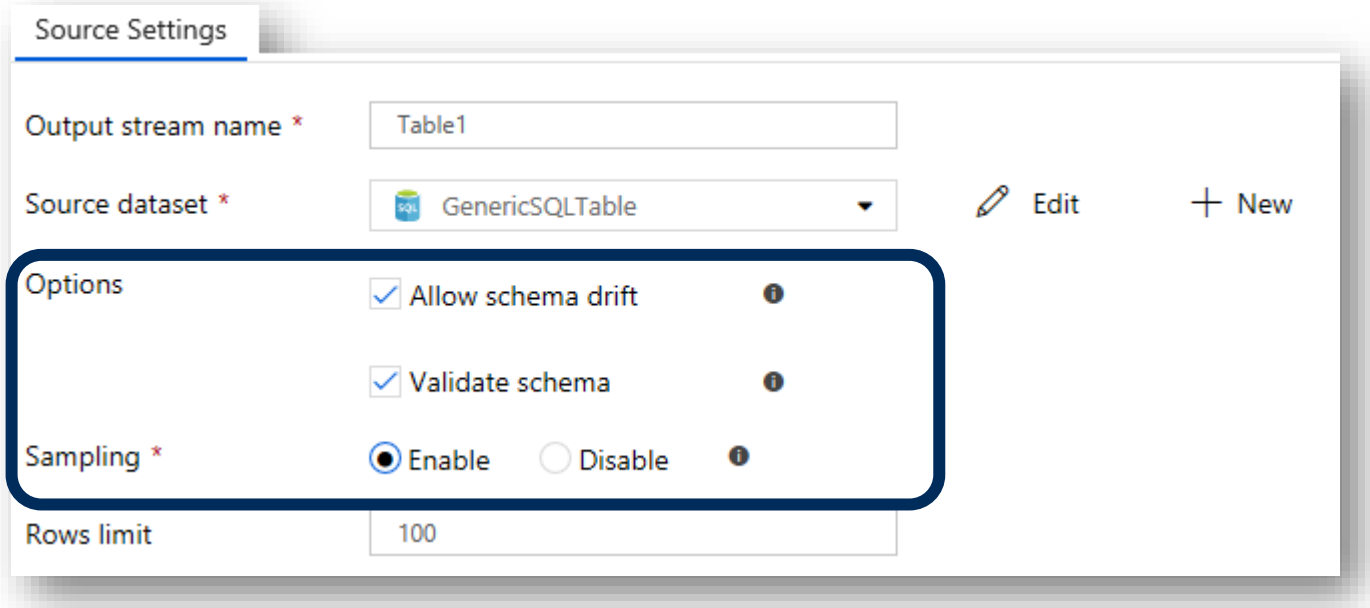


Currently Available:

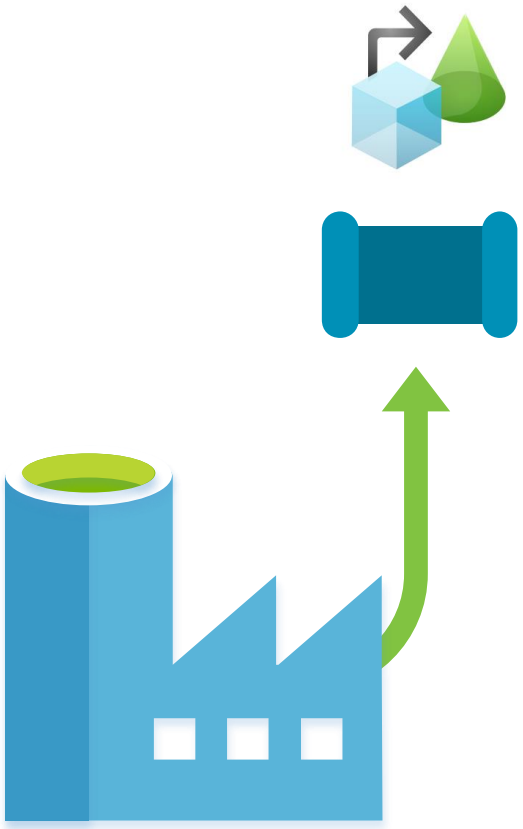


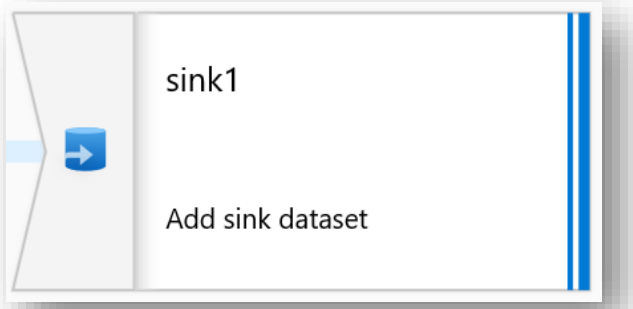
Mapping Data Flows – Settings & Concepts

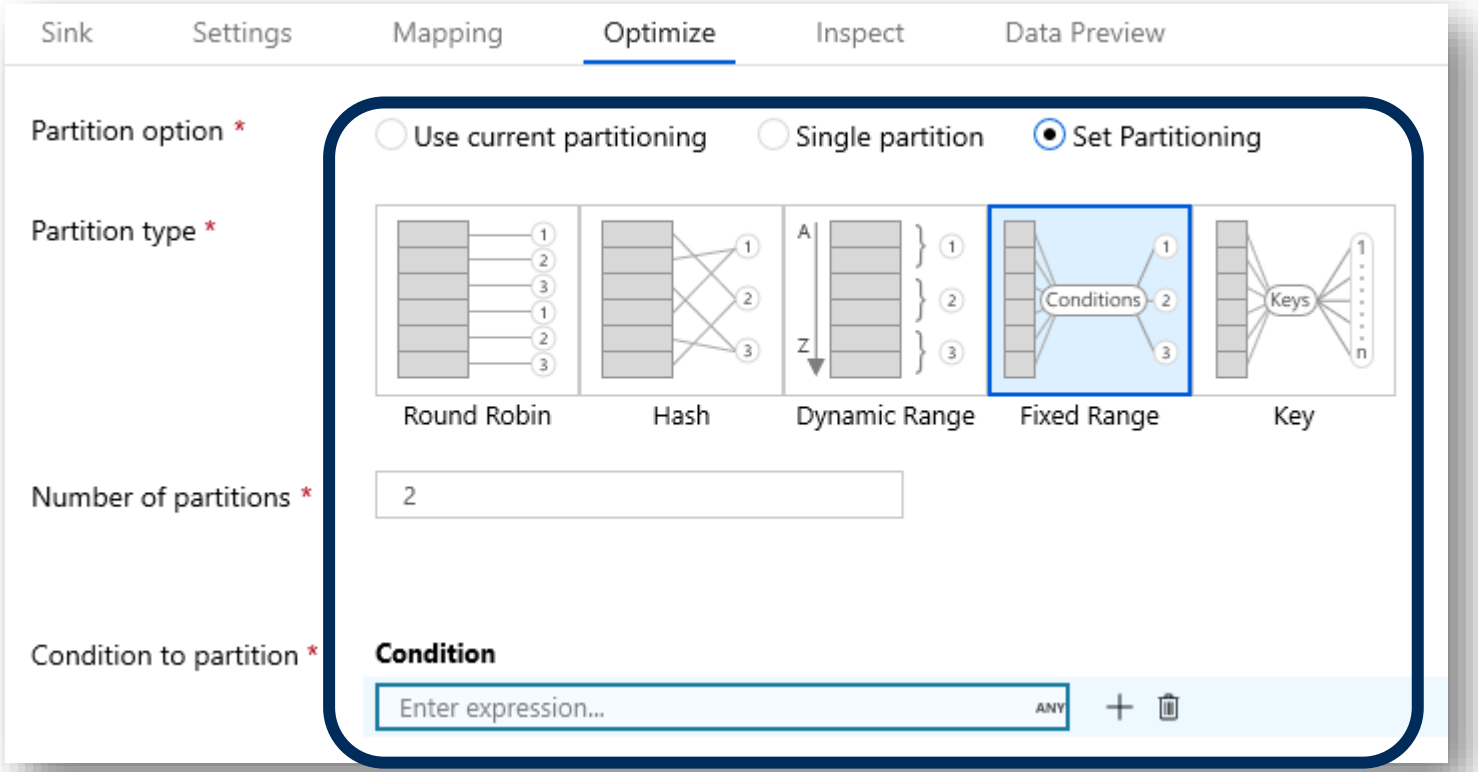




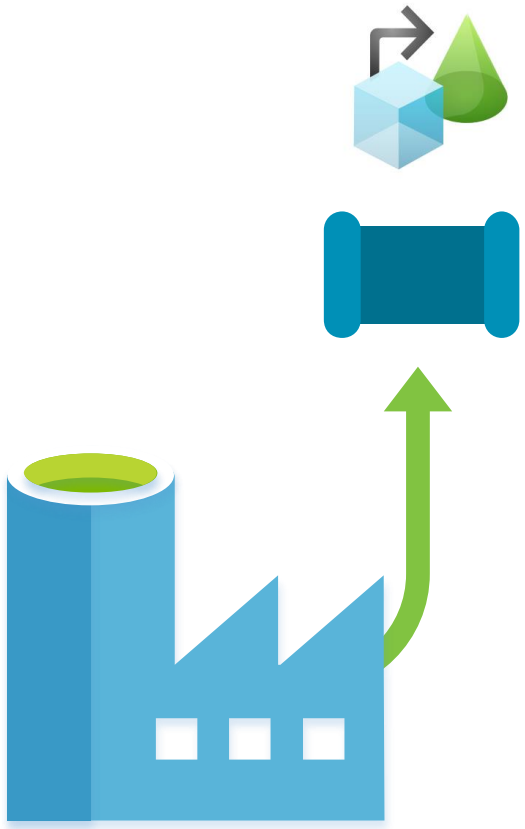
Mapping Data Flows – Settings & Concepts







Mapping Data Flows – Transformations



Multiple inputs/outputs

- New Branch — Multicast
- Join — Merge Join
- Conditional Split
- Union
- Lookup

Schema modifier

- Derived Column
- Aggregate
- Surrogate Key
- Pivot
- Unpivot
- Window

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

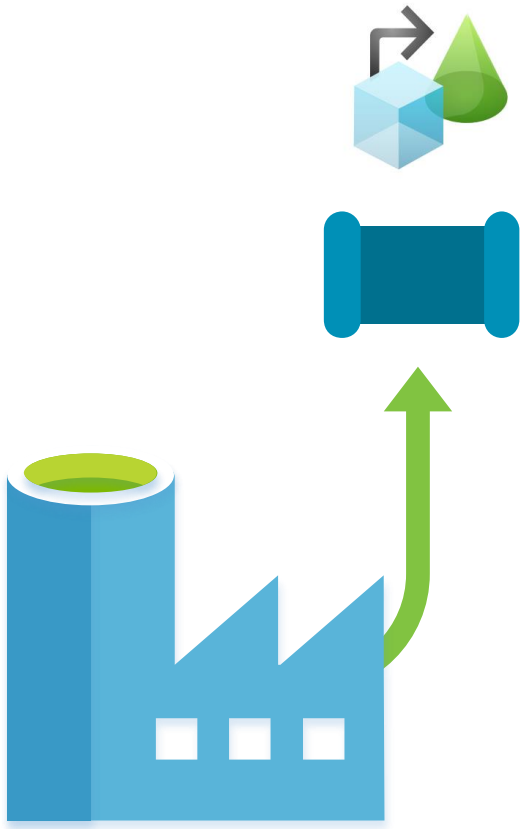
Row modifier

- Exists
- Select
- Filter
- Sort

Custom

- Extend — Script Component

Mapping Data Flows – Expression Builder



Visual Expression Builder

Currently working on: year

Filter...

All | String | Math | Date | Logical | Input

abc md5(ANY expression)

123 nextSequence()

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

regexMatch(abc string, abc regex to match)

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

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

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

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

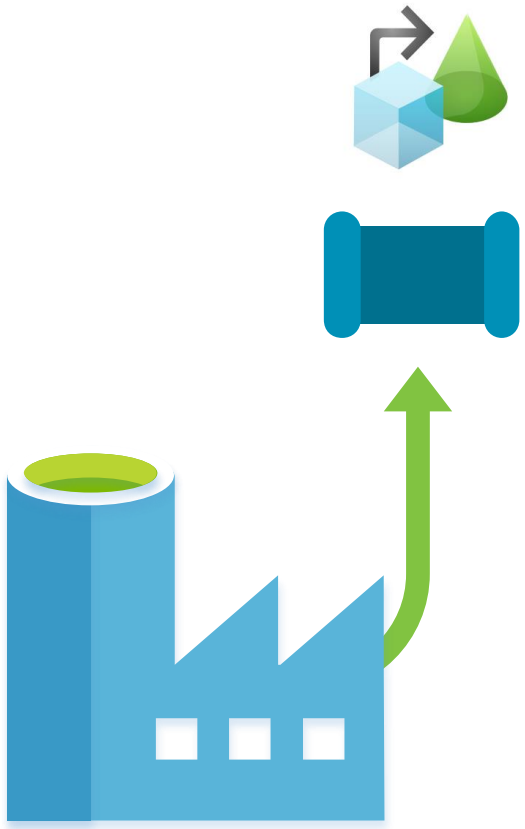
+ - * / ||

Data preview

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

Output: year 123	title abc
-	-

Mapping Data Flows – Debug Mode



ADWAnalysis | ADWAnalysisB...

Debug | Saved | Validate | Source Settings | Cluster ● ForDataFlow

OrderHeader
Columns:
22 total

Join1
Inner join on OrderHeader and OrderDetails

Aggregate1
Aggregating data by 'SalesOrderNumber' producing columns 'OrderLineCount'

sink1
Export data to ADWOrderLineCountTable

OrderDetails
Import data from ADWSalesOrderDetail

Add Source

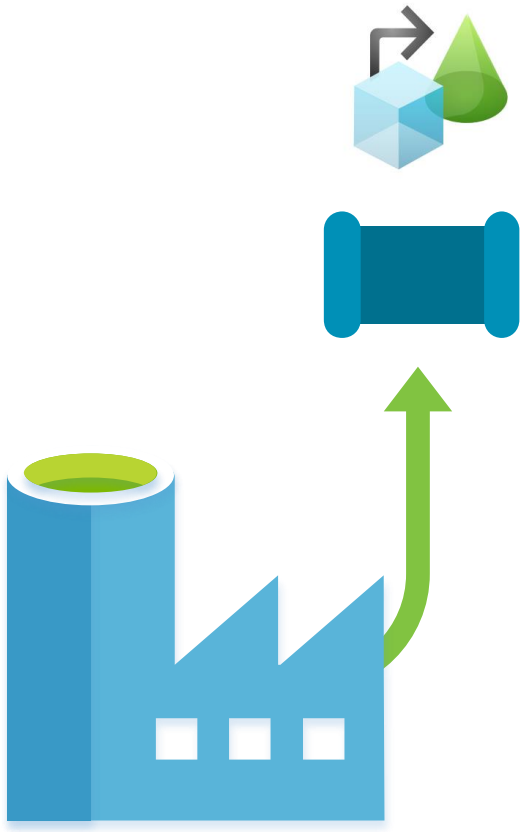
Only gives you a General Purpose cluster

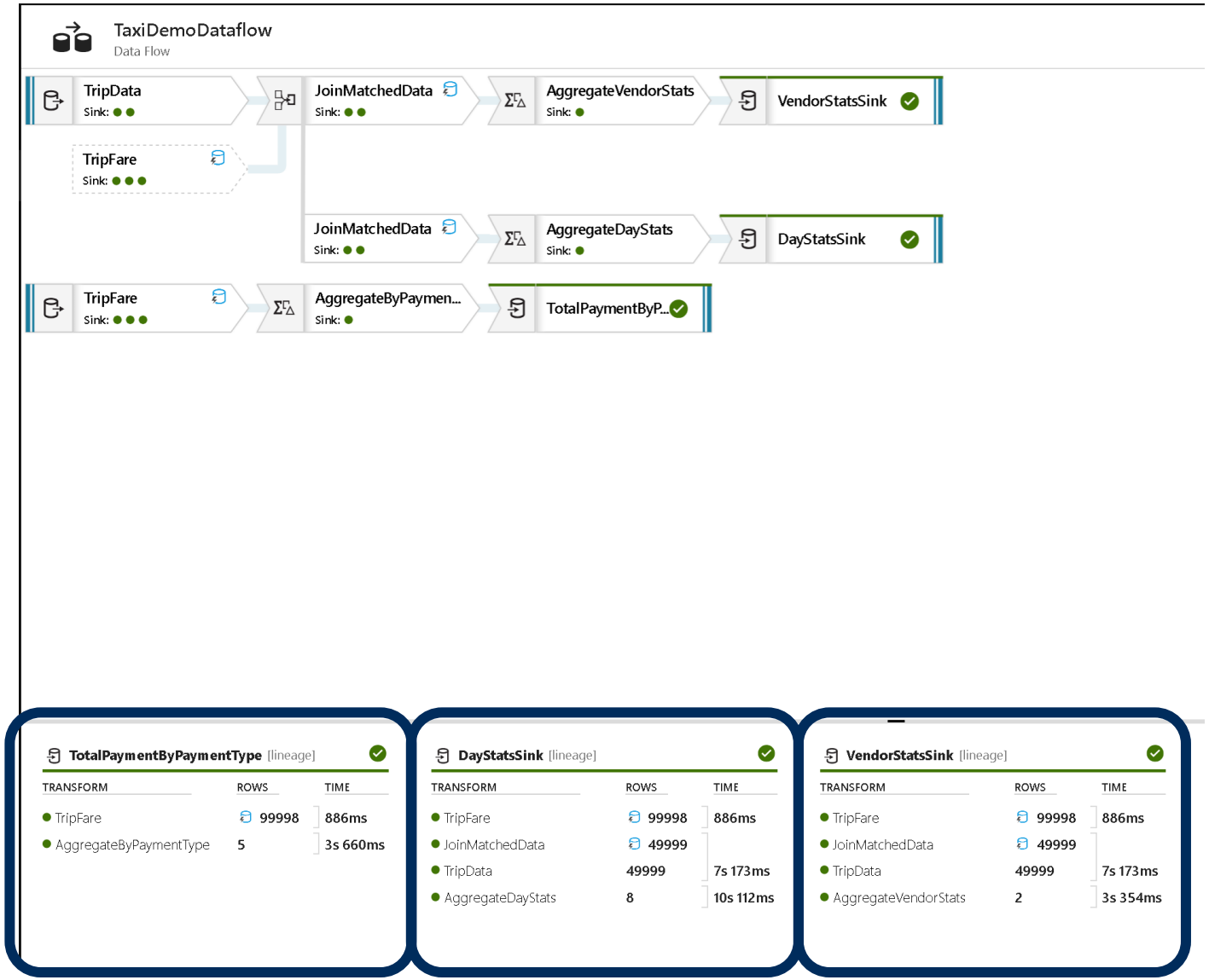
Source Settings | Define schema | Optimize | Inspect | Data Preview

	Updated	New	Unchanged	Total
Number of rows	N/A	N/A	N/A	32

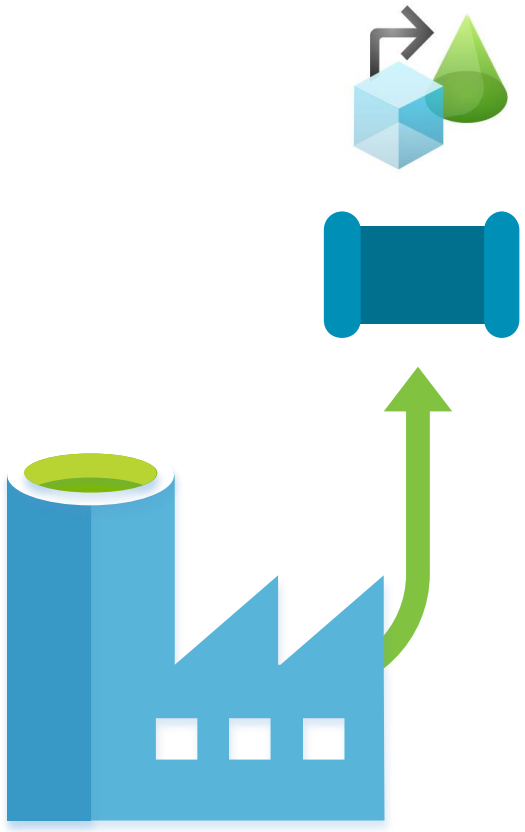
SalesOrderID	RevisionNumber	OrderDate	DueDate	ShipDate
71774	2	06/01/2008 00:06:00	06/13/2008 00:06:00	06/08/2008
71776	2	06/01/2008 00:06:00	06/13/2008 00:06:00	06/08/2008

Mapping Data Flows – Monitoring





Mapping Data Flows

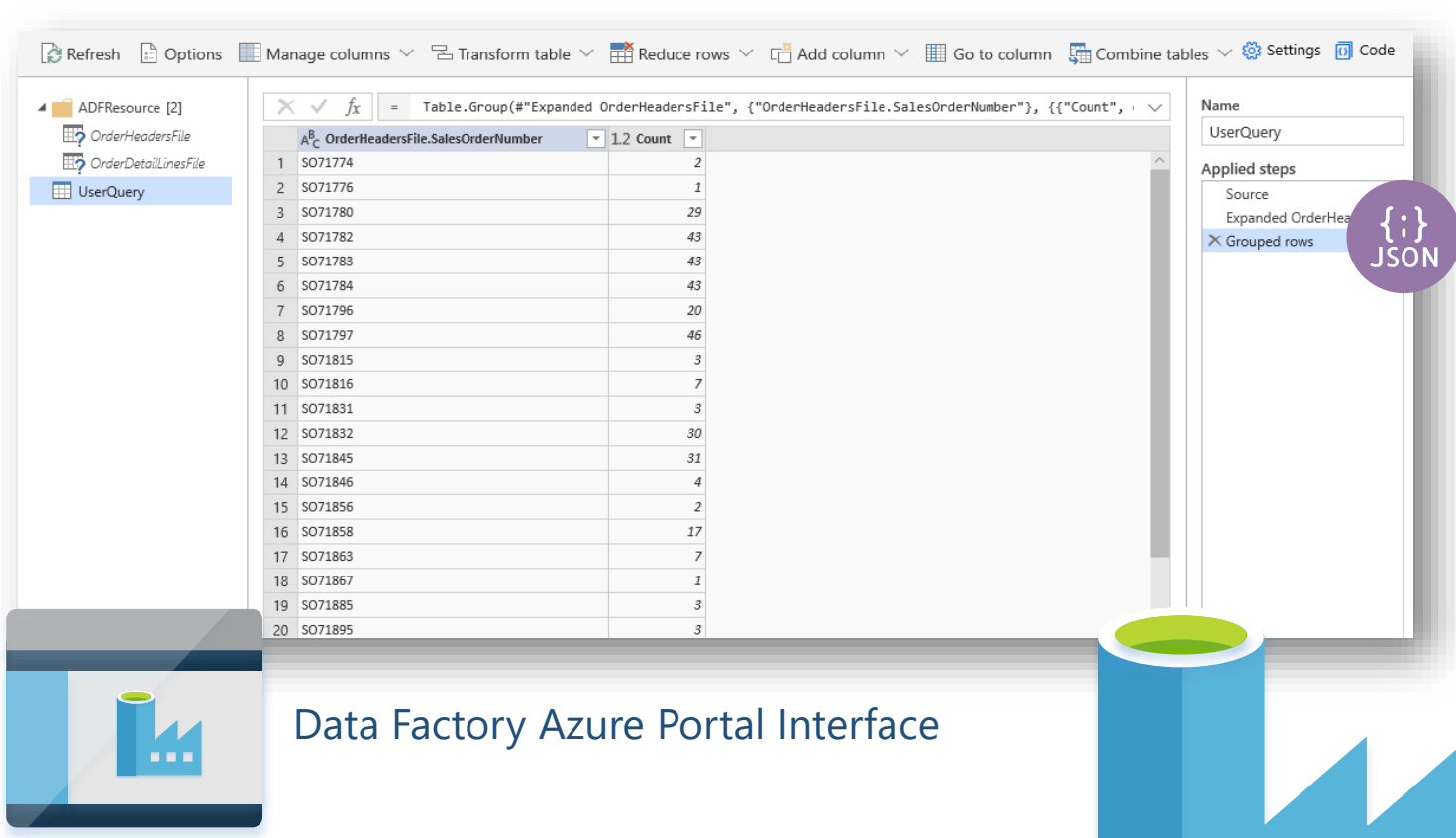


1. **Activity**
   https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-overview
2. **Source & Sink**
   https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-schema-drift
3. **Transformations**
   https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-aggregate
4. **Expression Builder**
   https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-expression-functions
5. **Debug Mode**
   https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-debug-mode
6. **Monitoring**
   https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-monitoring

Mark Kromer
https://github.com/kromerm/adfdataflowdocs

Wrangling Data Flows

What is a Wrangling Data Flow?

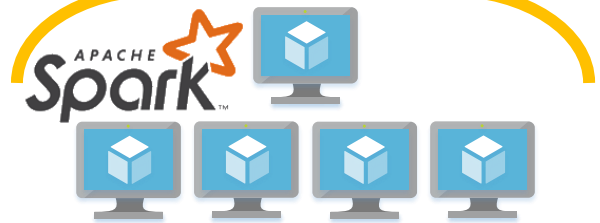


Data Factory Azure Portal Interface

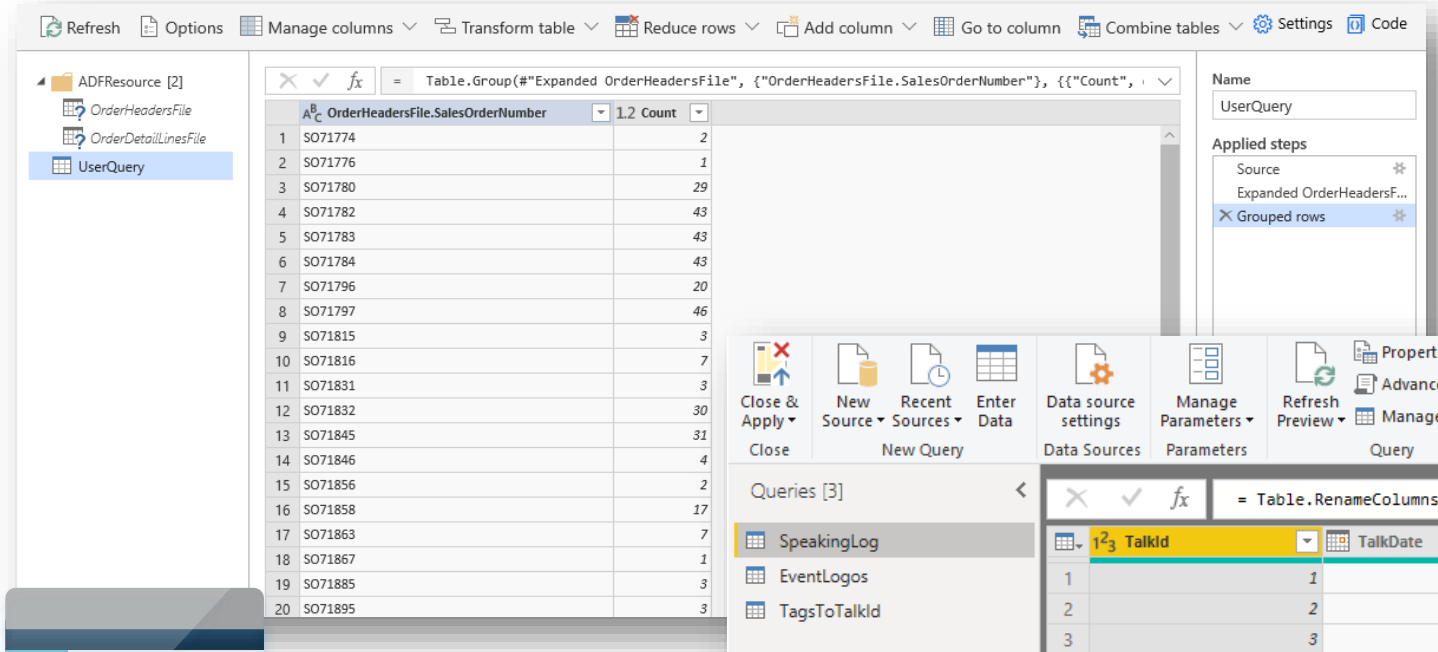
Data Factory

Wrangling Data Flow

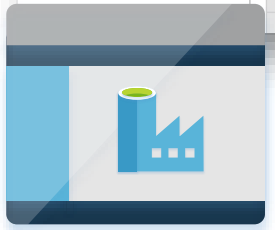
Azure Databricks



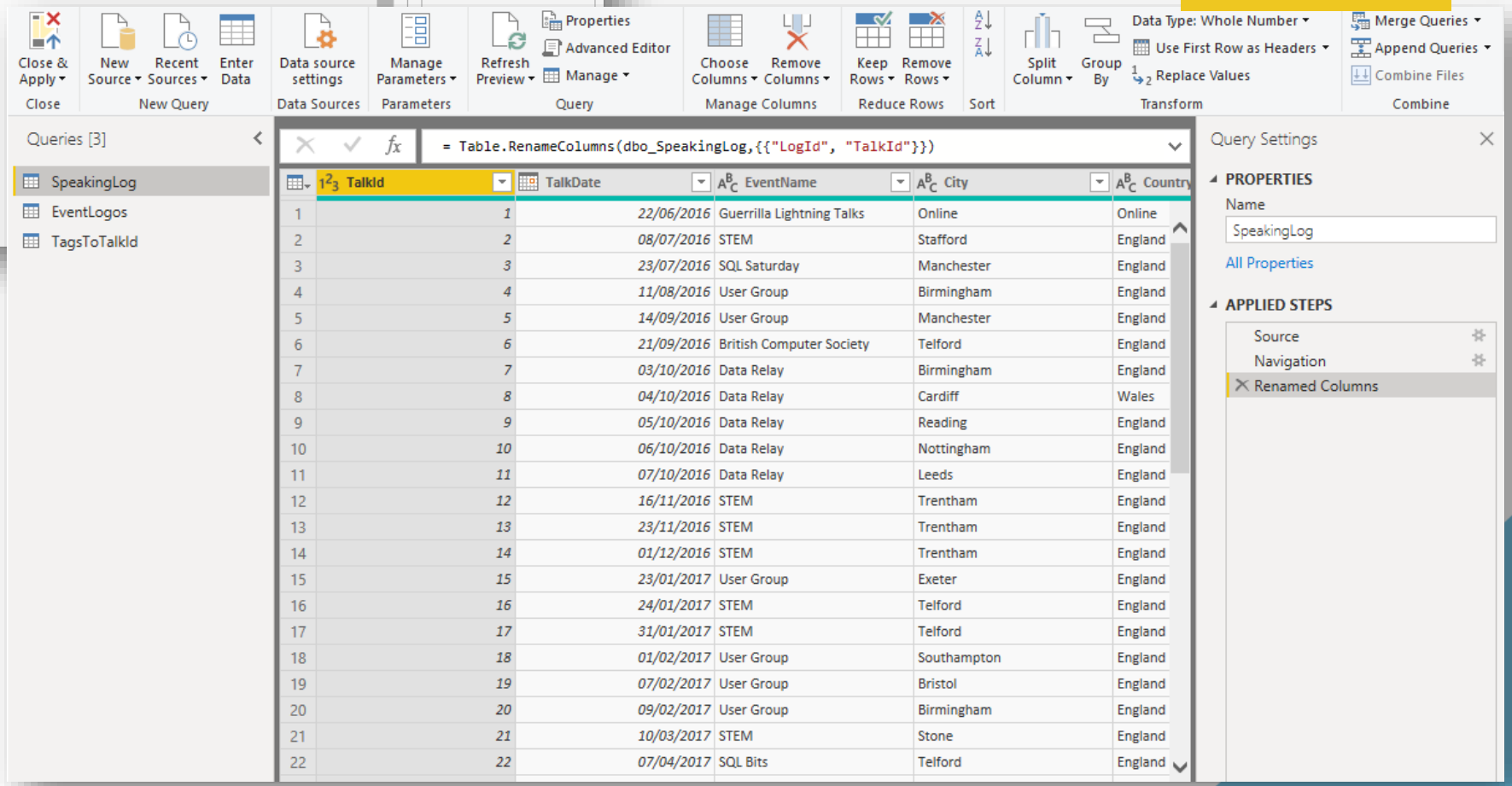
What is a Wrangling Data Flow?



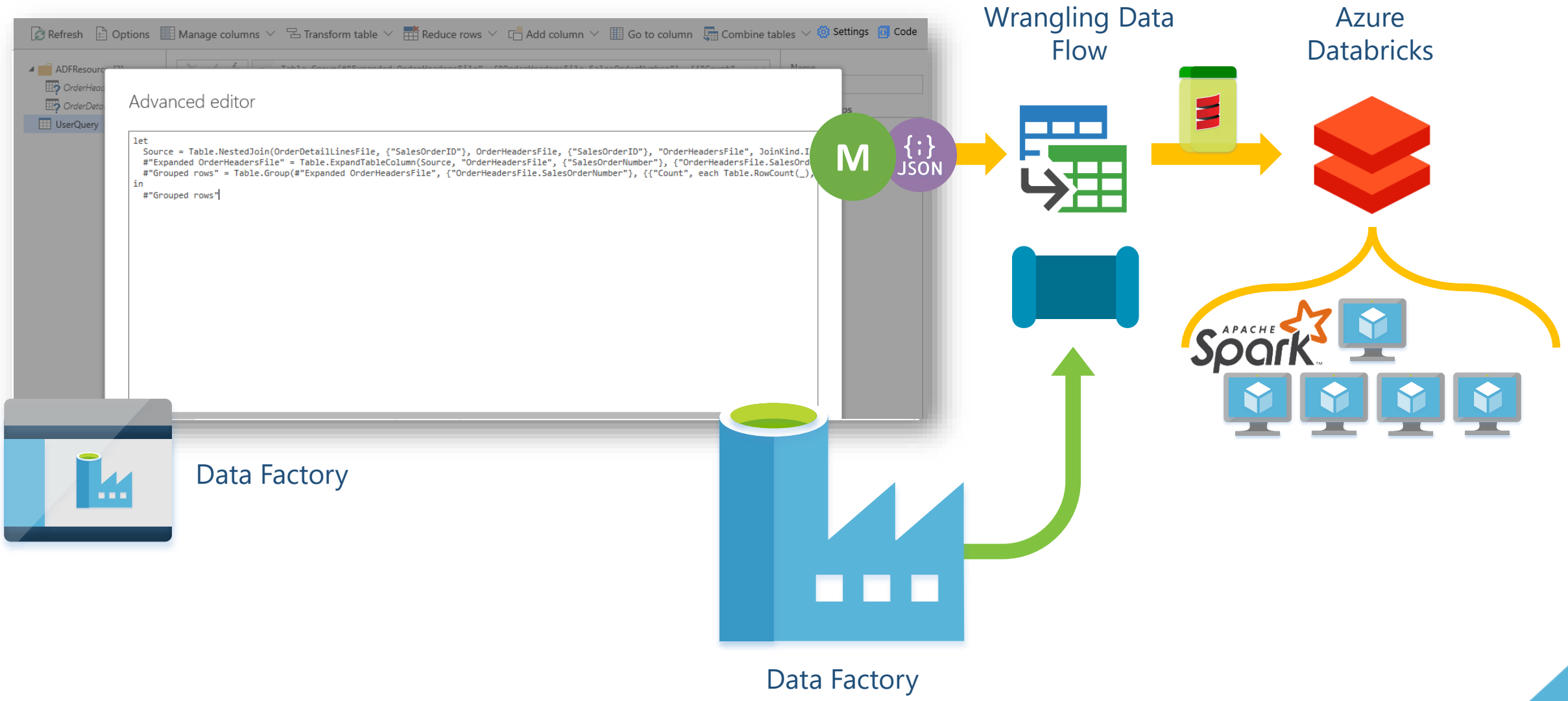
Power BI Desktop



Data Factory



What is a Wrangling Data Flow?

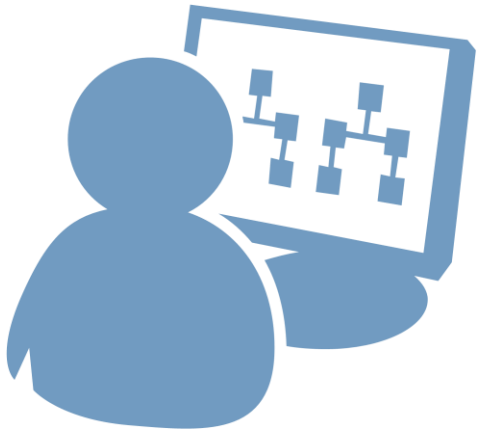


Note to self - start clusters!

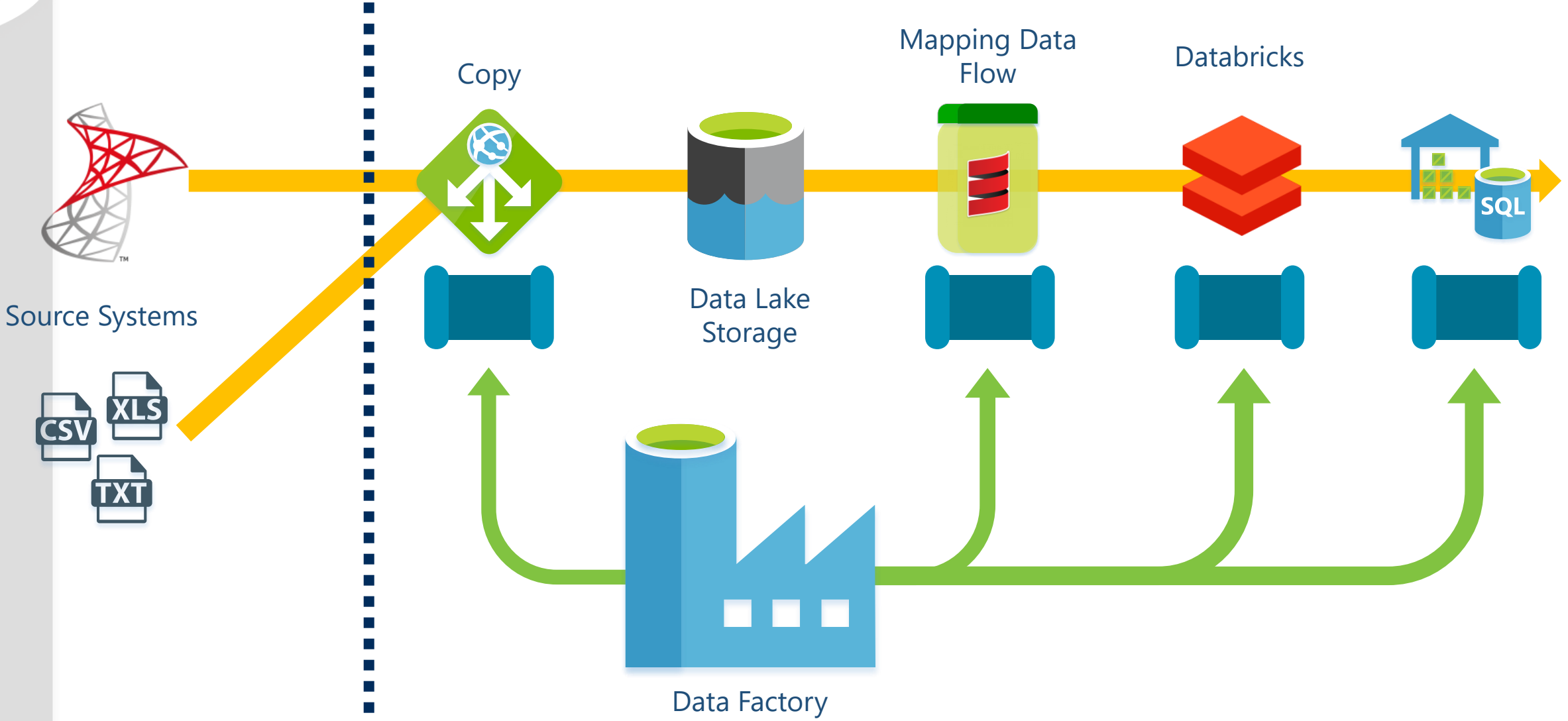
Demo Summary

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

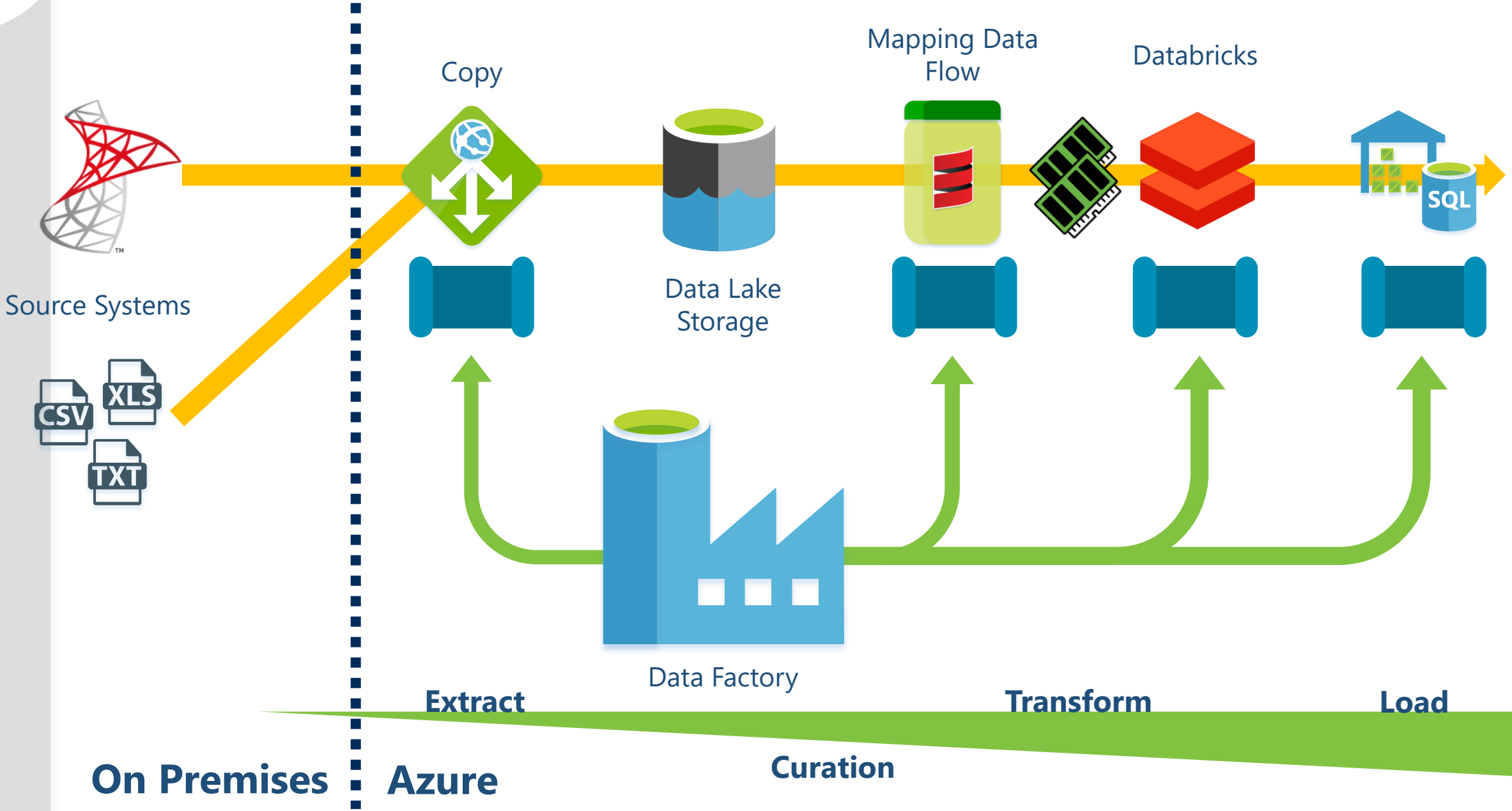
Design Patterns



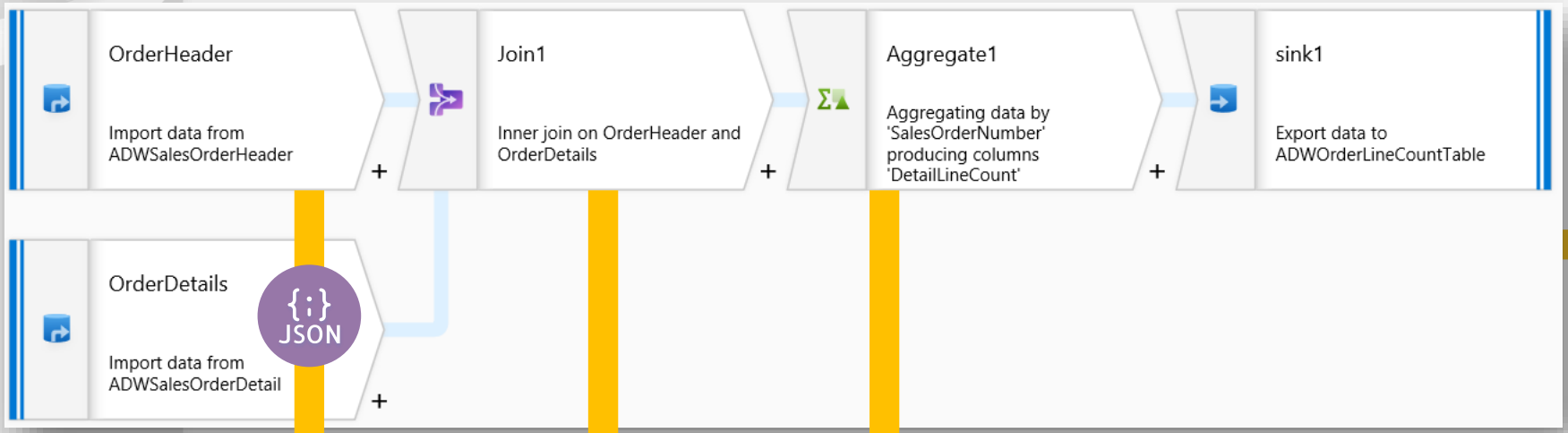
Mapping Data Flow Future Design Patterns ???

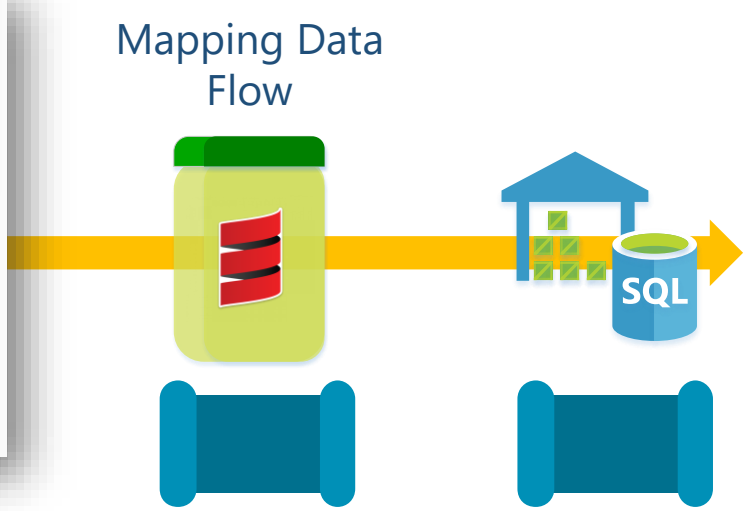


Mapping Data Flow Future Design Patterns ???



Mapping Data Flow Future Design Patterns ???





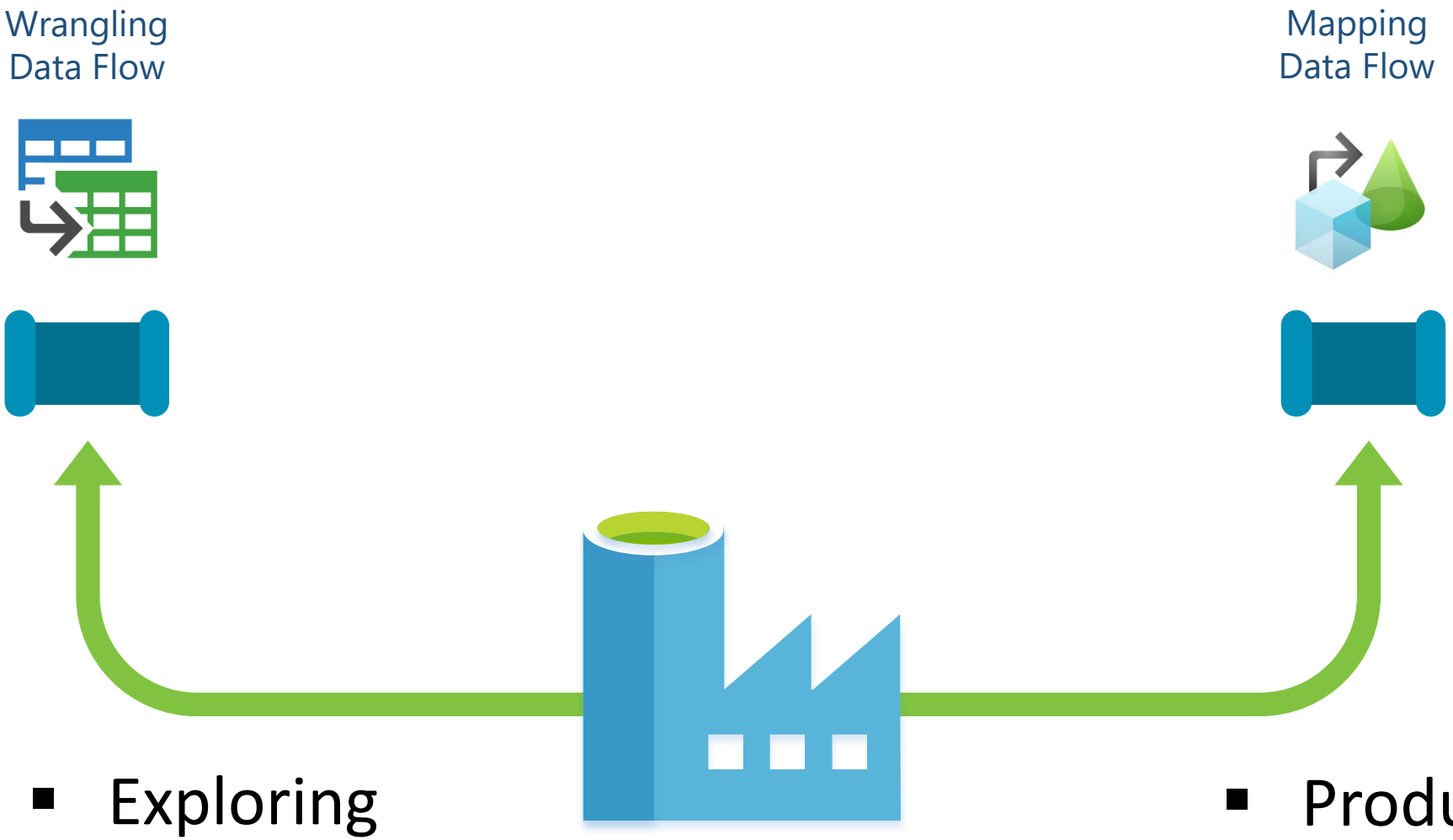

```
"fileName": {
    "value": "@dataset().FileName",
    "type": "Expression"
},
"folderPath": {
    "value": "@dataset().SourceDIR",
    "type": "Expression"
}
```

JSON

JSON

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

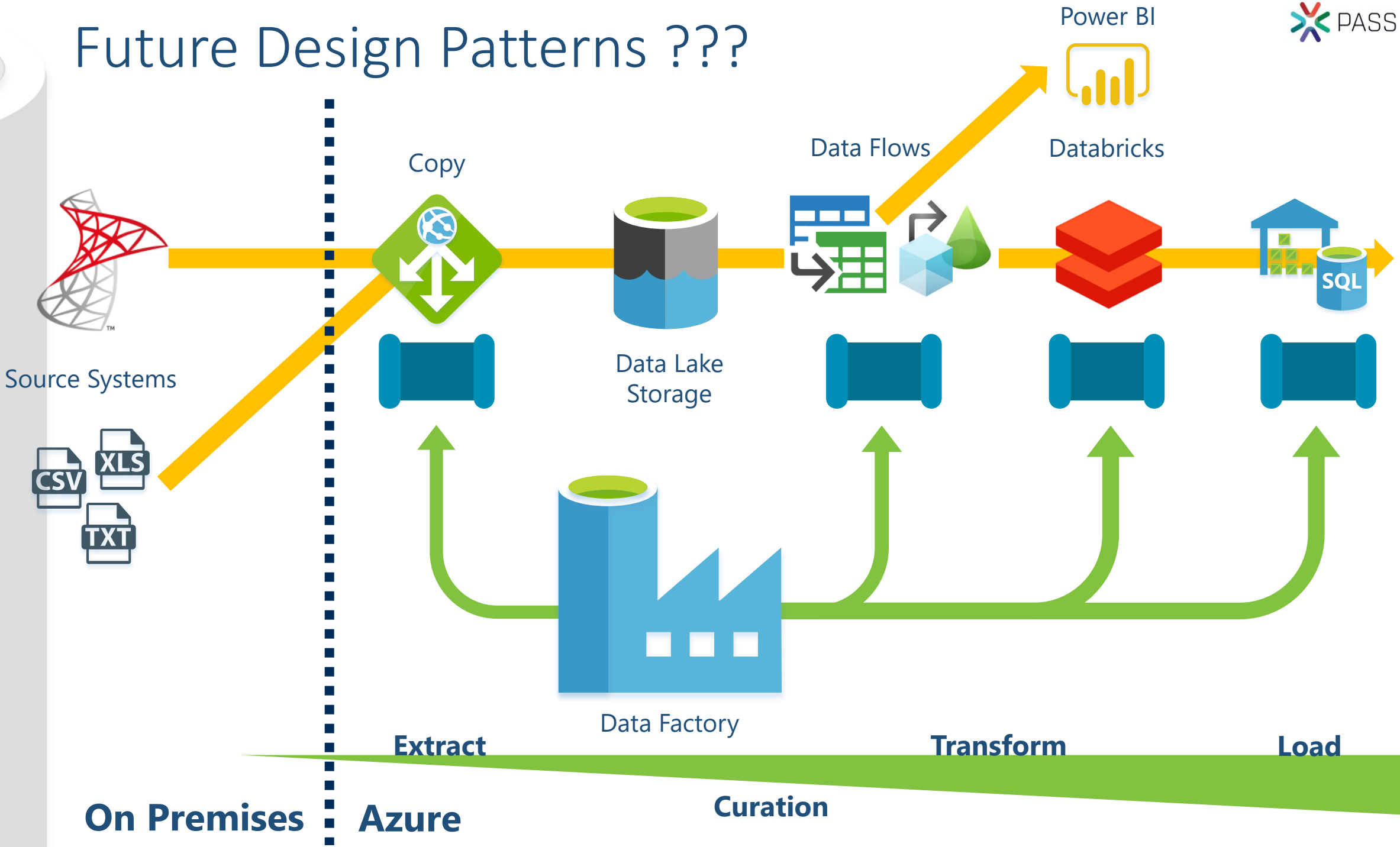
What else?



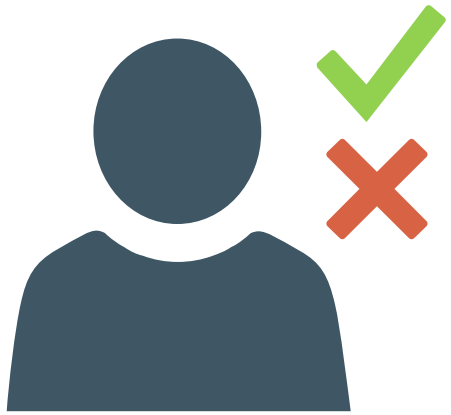
- Exploring
- Experimenting
- Analysing

- Productionising
- Engineering
- Warehousing

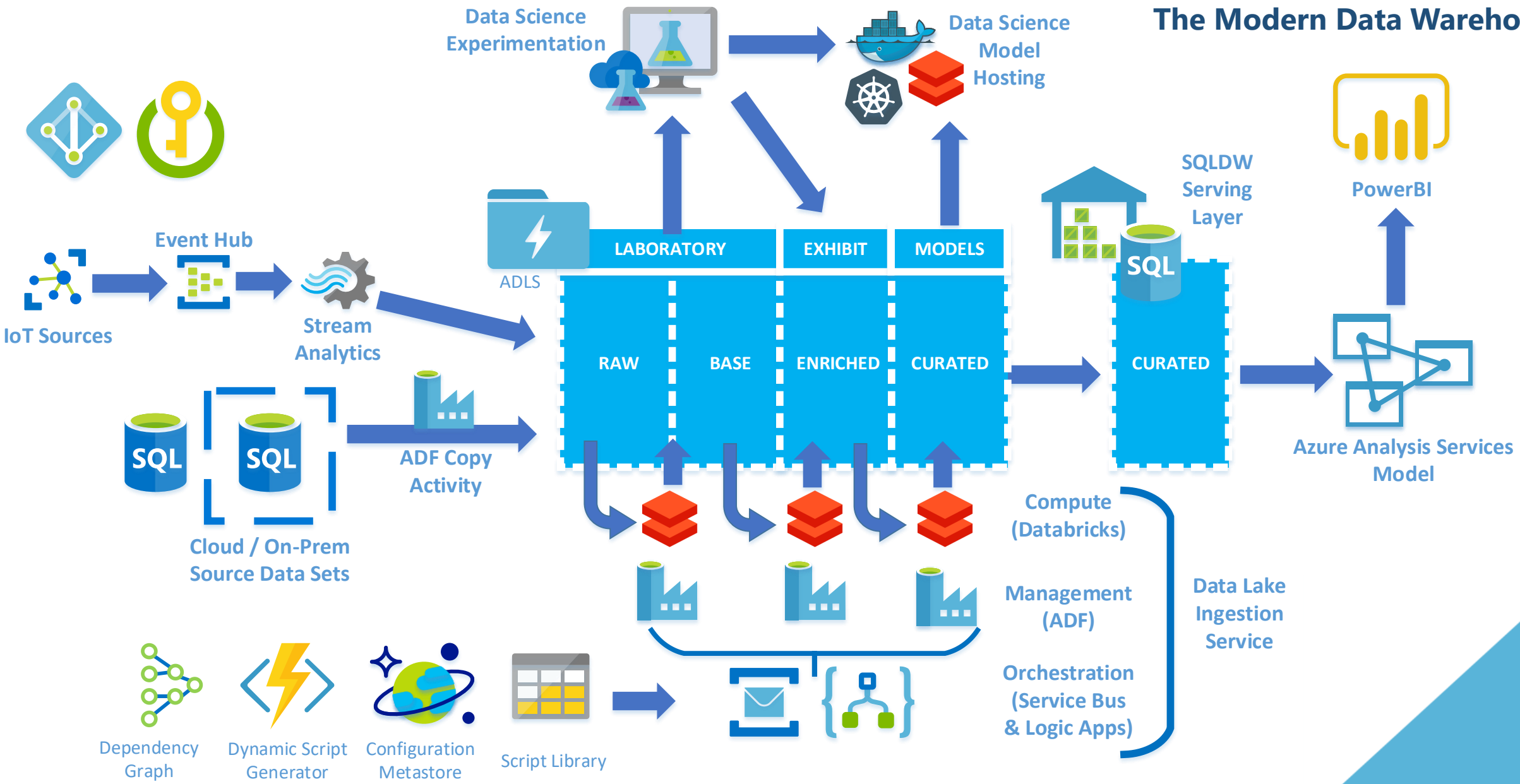
Future Design Patterns ???



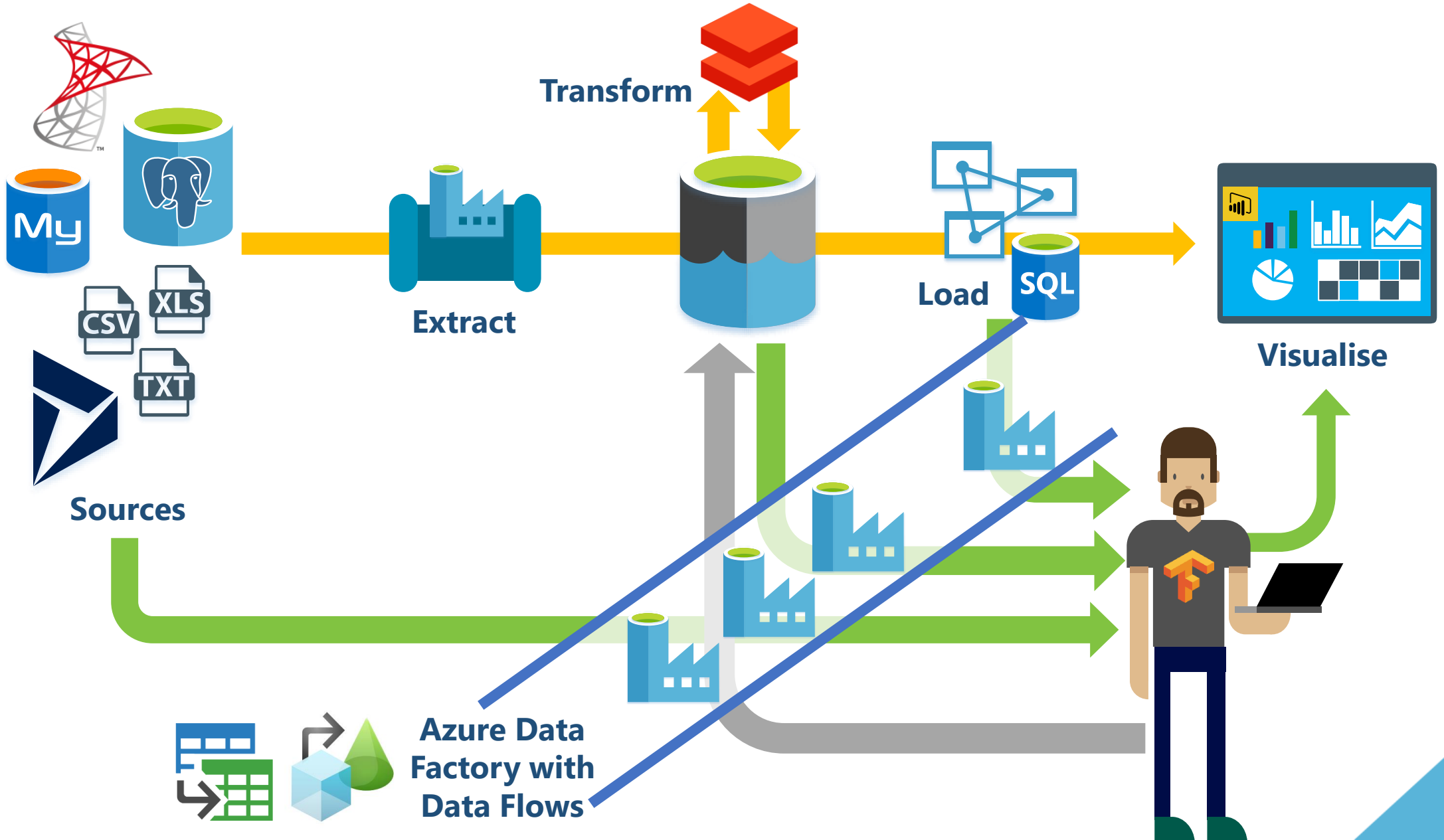
Conclusions



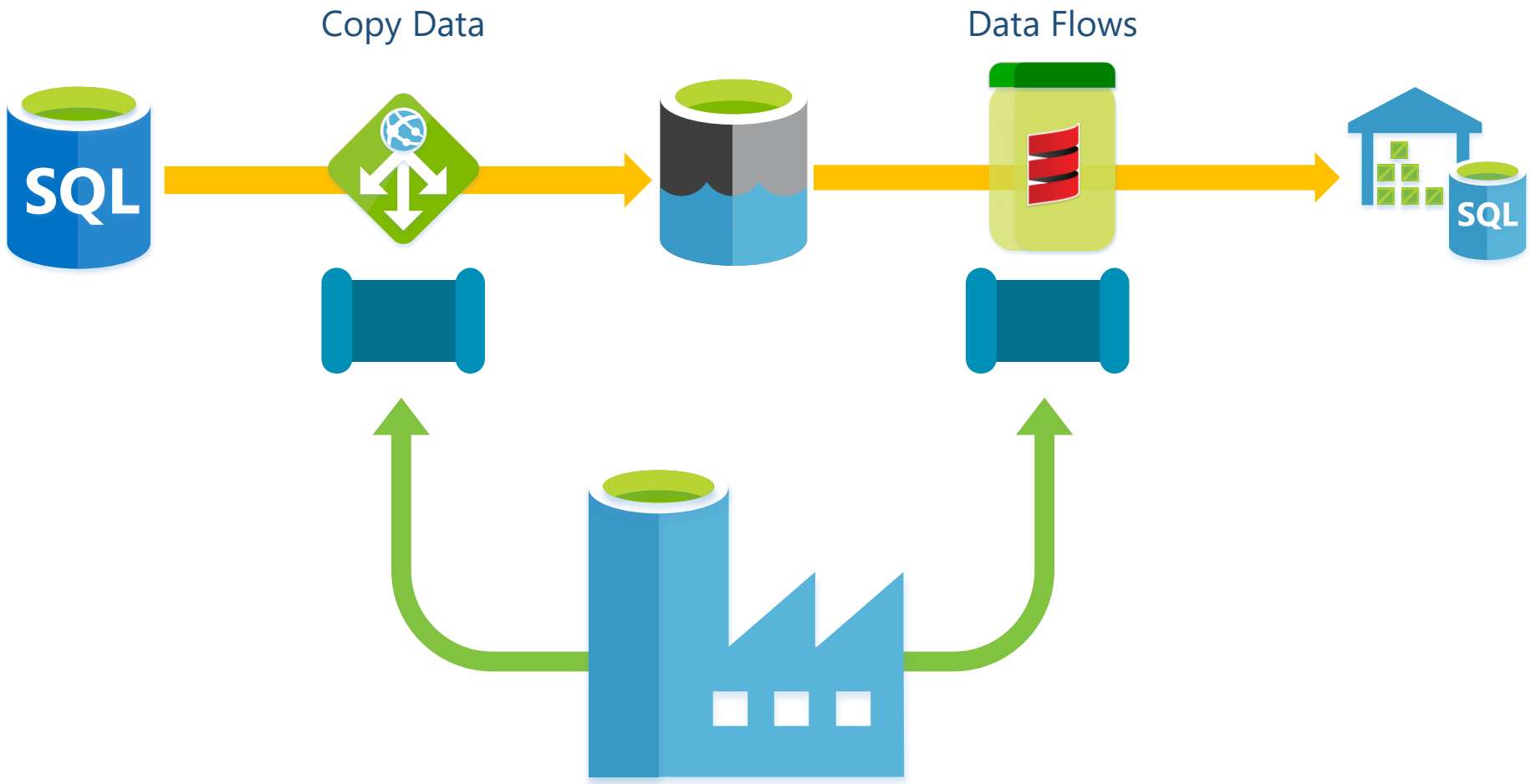
Why use Azure Data Factory?



Data Factory for the Data Scientist



What is Azure Data Factory?



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

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

Thanks for Listening

Paul Andrew

@MrPaulAndrew

Microsoft® MVP Most Valuable Professional

altius

Email: paul@mrpaulandrew.com
Blog: mrpaulandrew.com
GitHub: github.com/mrpaulandrew

PDF Slides