ETL in Azure Made Easy

with Data Factory Data Flows

Paul Andrew | Principal Consultant & Solution Architect

Microsoft® Most Valuable Professional

@MrPaulAndrew

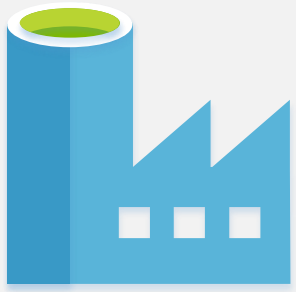
Microsoft Partner
Microsoft
Gold Cloud Platform
Gold Data Analytics
Gold Data Platform
Gold DevOps
Gold Datacenter

adatis

Extract Transform Load

ETL in Azure Made Easy

with Data Factory Mapping Data Flows



Paul Andrew | Principal Consultant & Solution Architect

Microsoft® Most Valuable Professional

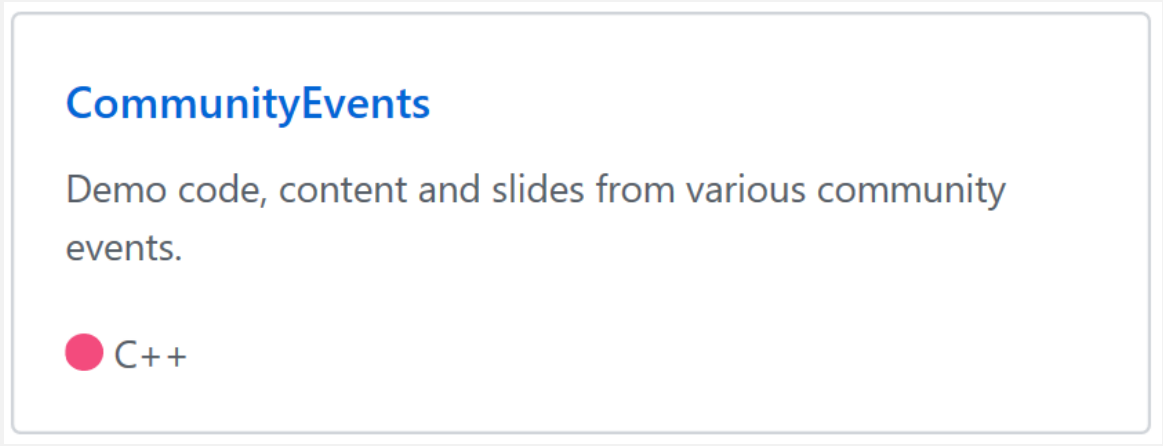
@MrPaulAndrew

Microsoft Partner
Microsoft
Gold Cloud Platform
Gold Data Analytics
Gold Data Platform
Gold DevOps
Gold Datacenter

adatis

GitHub

https://github.com/mrpaulandrew



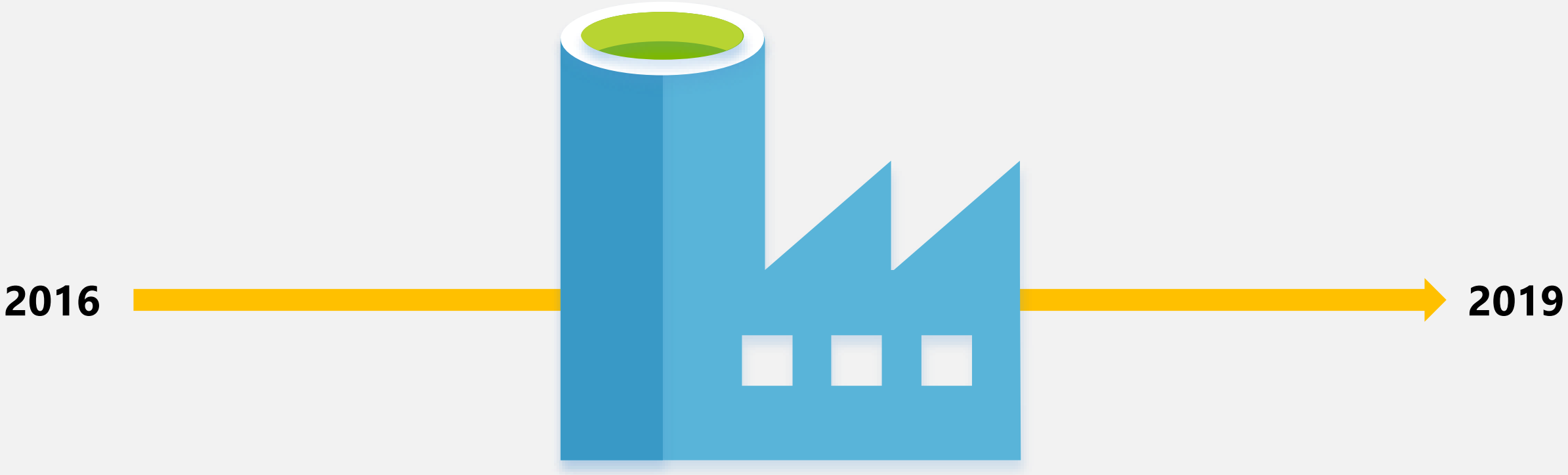
{Event/Location}-{Month}-{Year}

adatis

Azure Data Factory

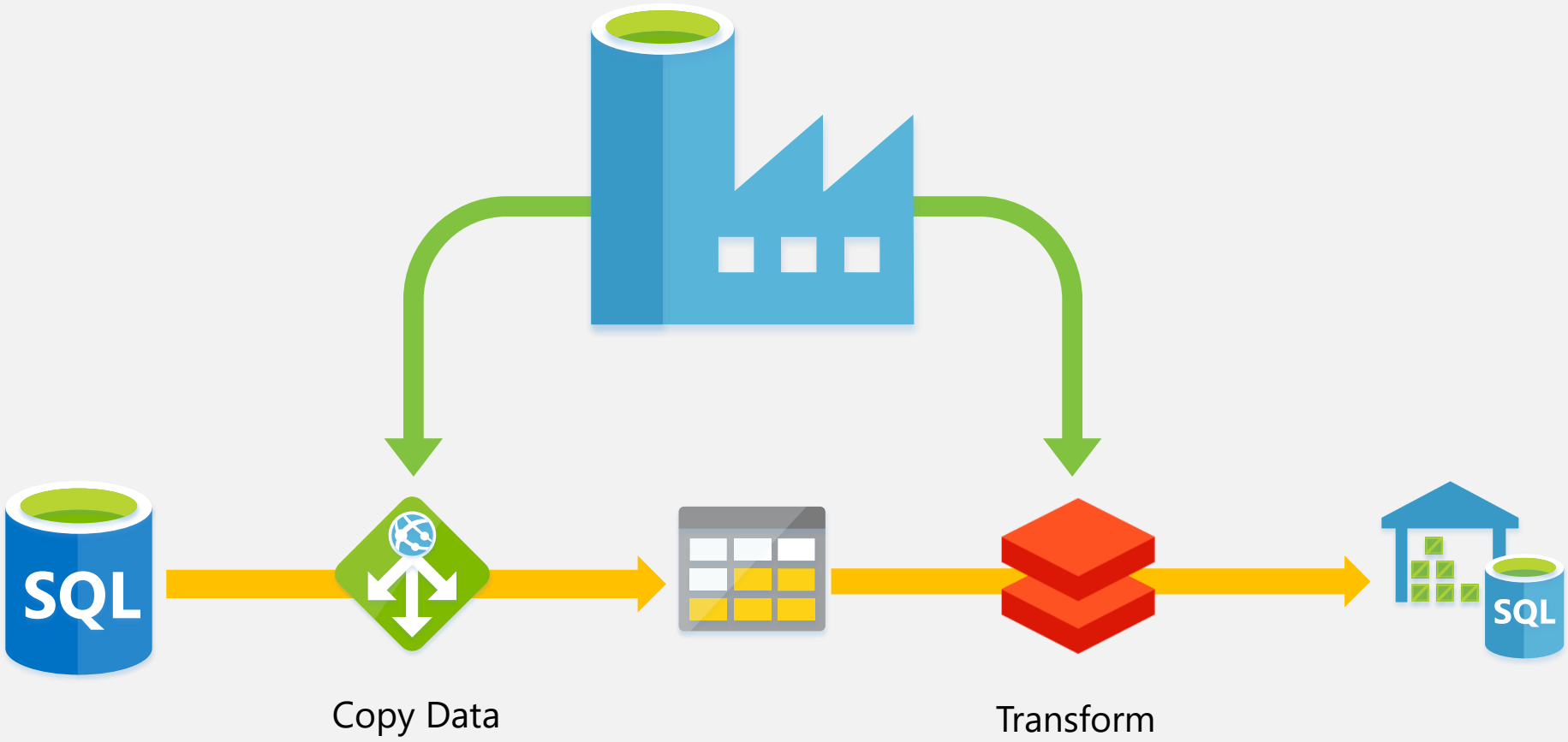
adatis

What is Azure Data Factory?



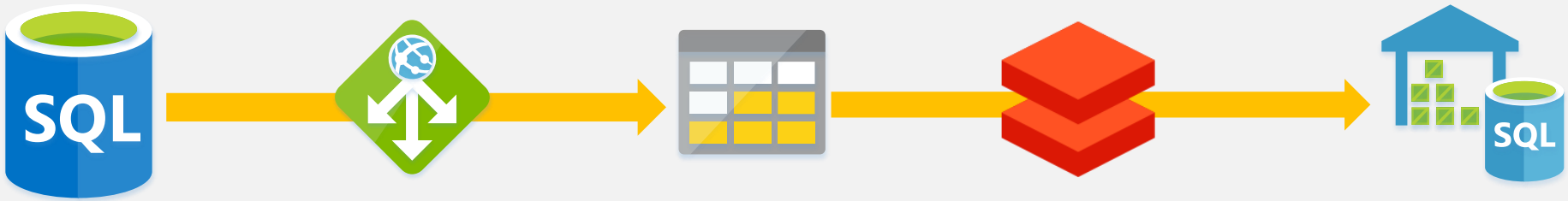
adatis

What is Azure Data Factory?



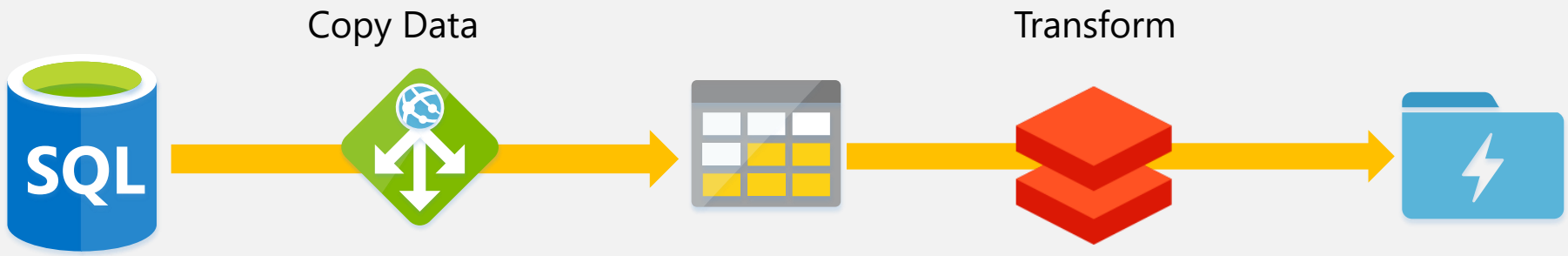
adatis

What is Azure Data Factory?

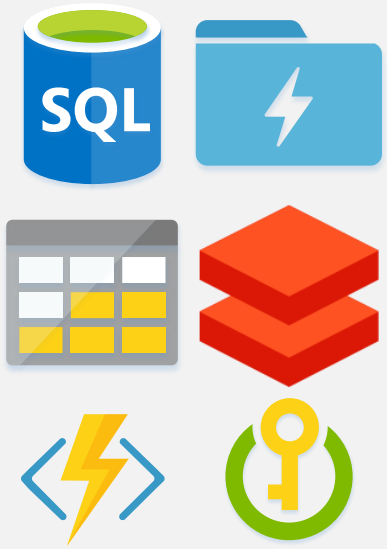


adatis

Data Factory Components

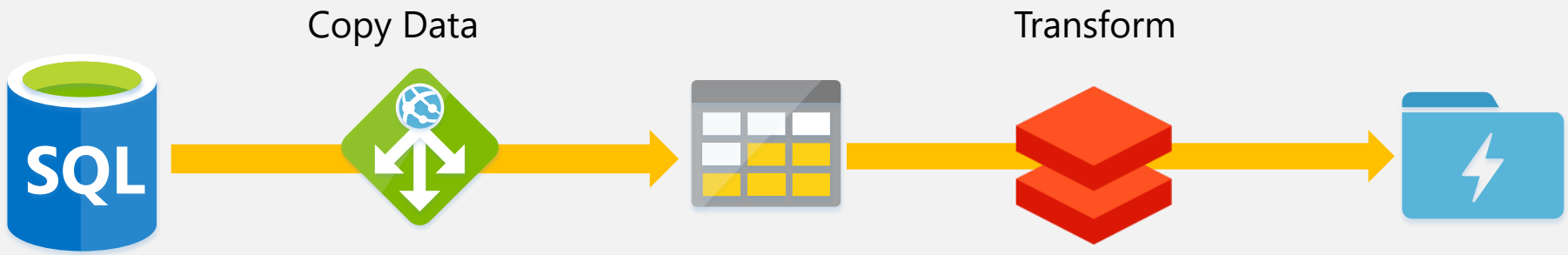


1 **Linked Services** – How do I connect? Like the SSIS connection manager.

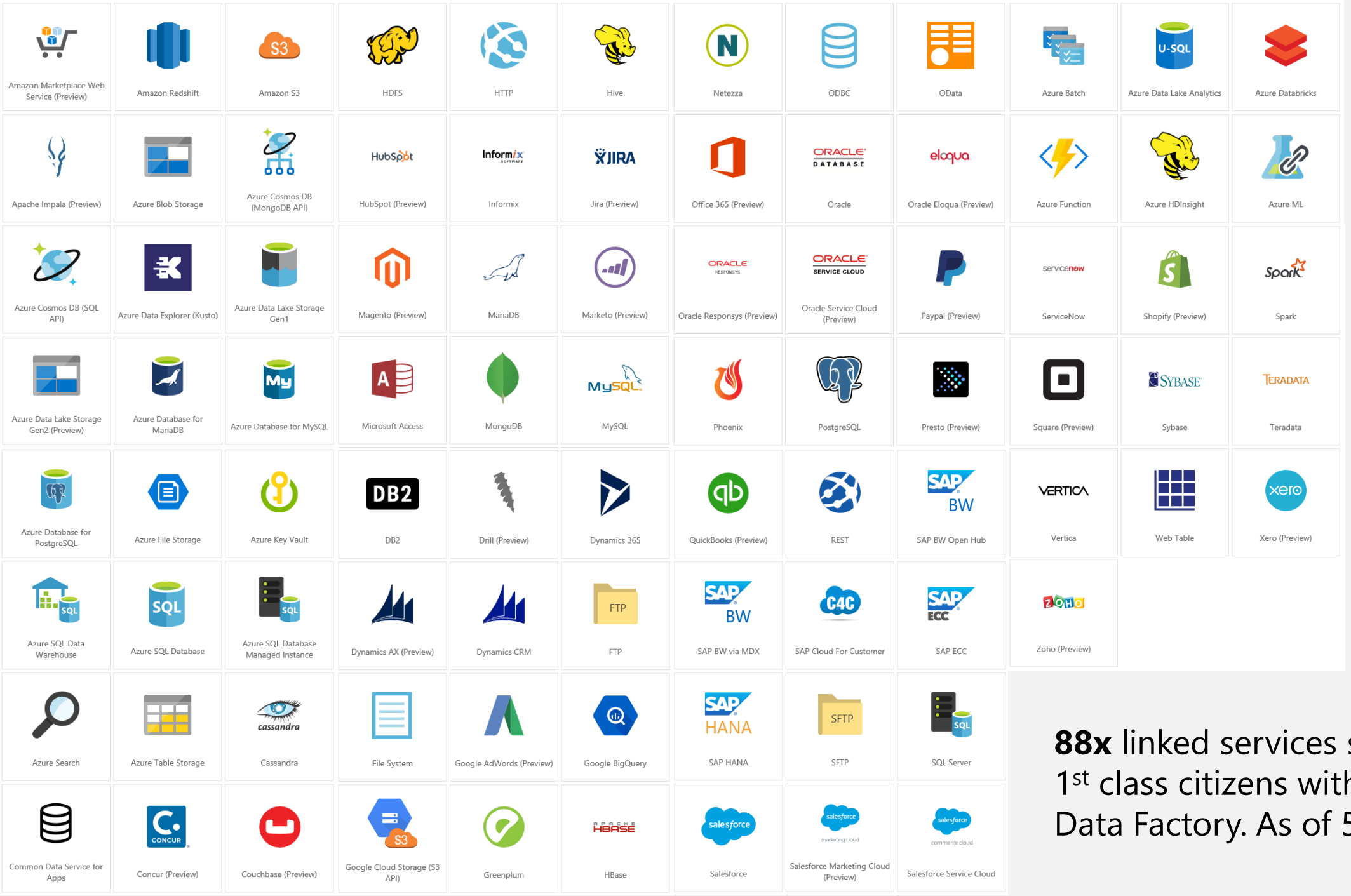


adatis

Data Factory Components

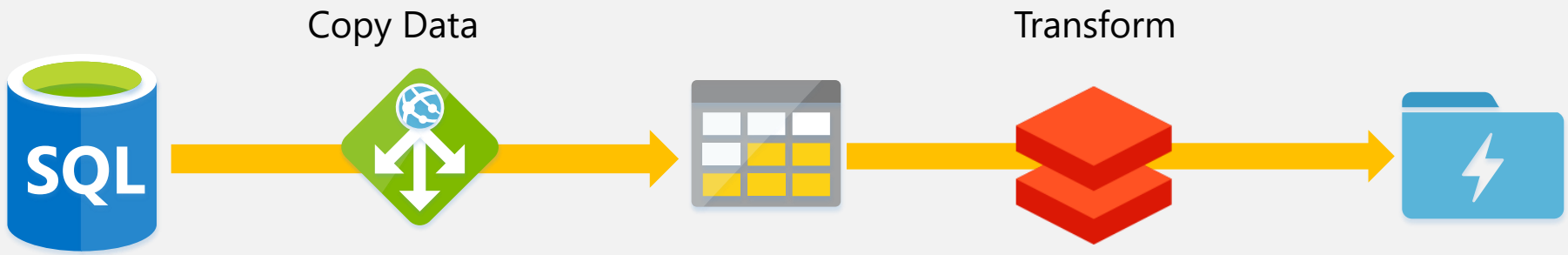


1 Linked Services



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

Data Factory Components



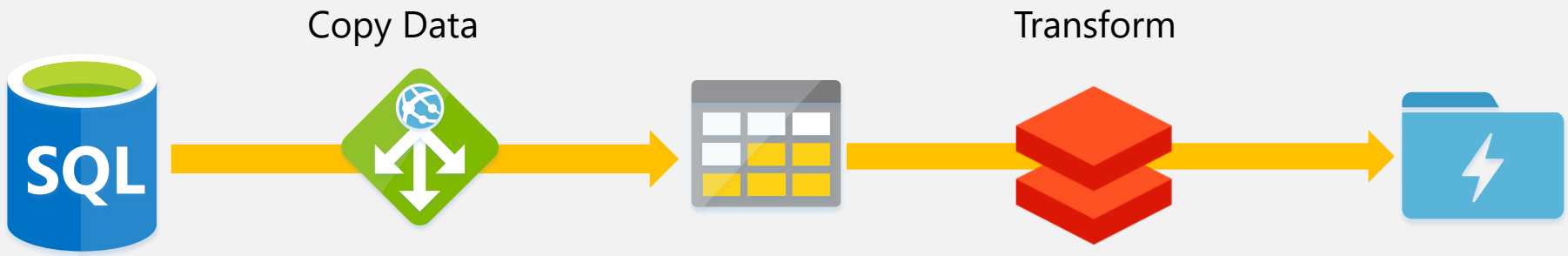
1. **Linked Services**
2. **Data Sets** – Where is my data? What format? What file path/table do I need?

dbo.DimCustomer

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

adatis

Data Factory Components



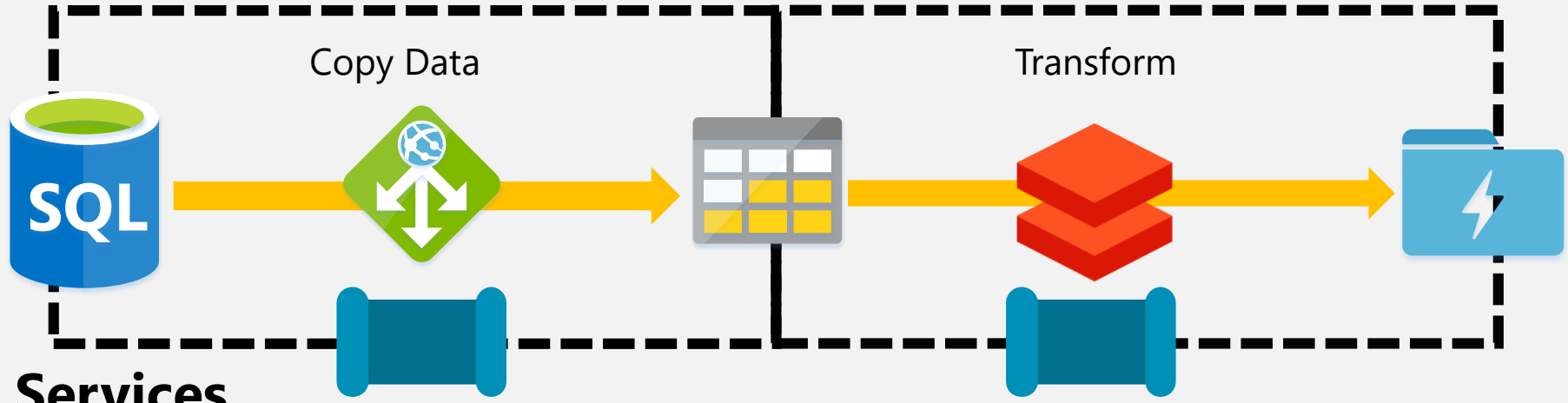
1. **Linked Services**
2. **Data Sets**
3. **Activities** – What do we want to happen? With what conditions?

Databricks Notebook Activity

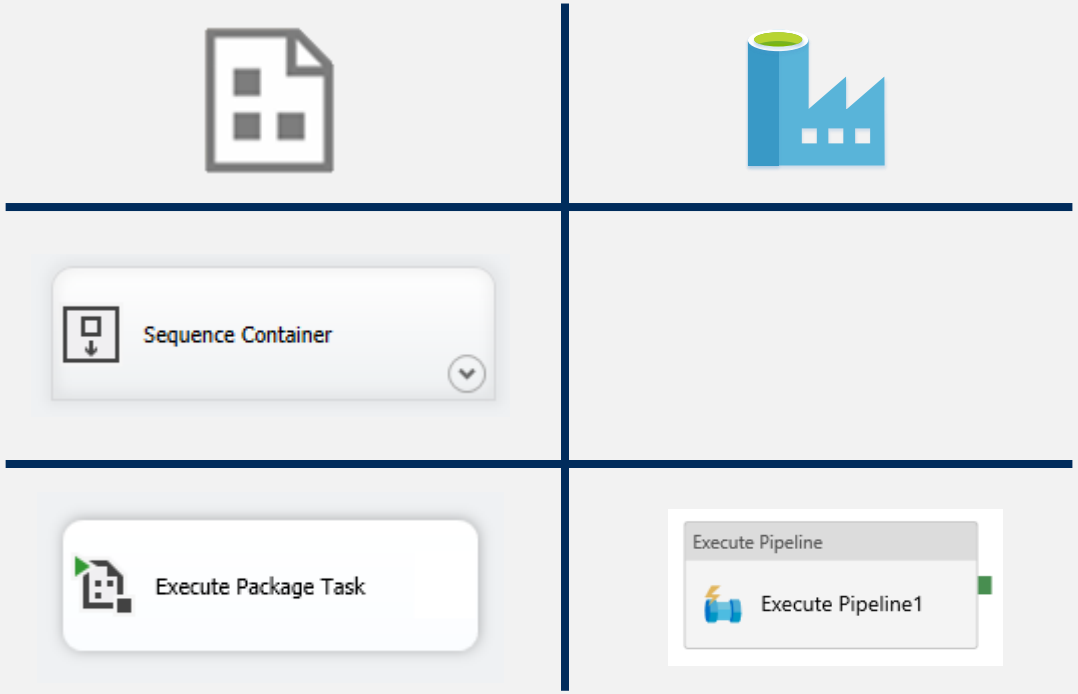
notebookPath: */Playground/Playing*
baseParameters: *Testing*
libraries[jar]: dbfs:/lib1.jar
linkedServiceName: *BricksOfData01*

adatis

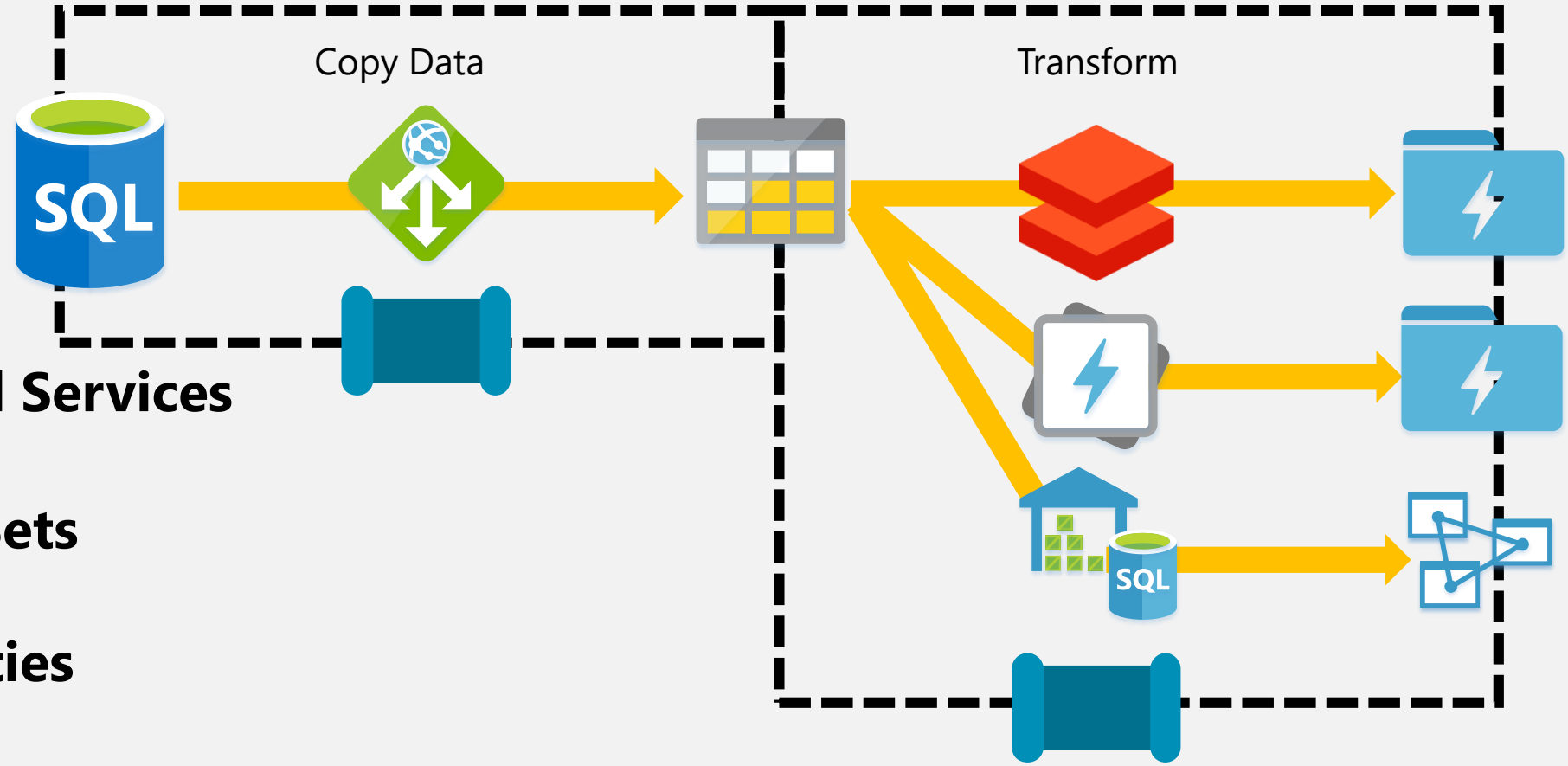
Data Factory Components



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



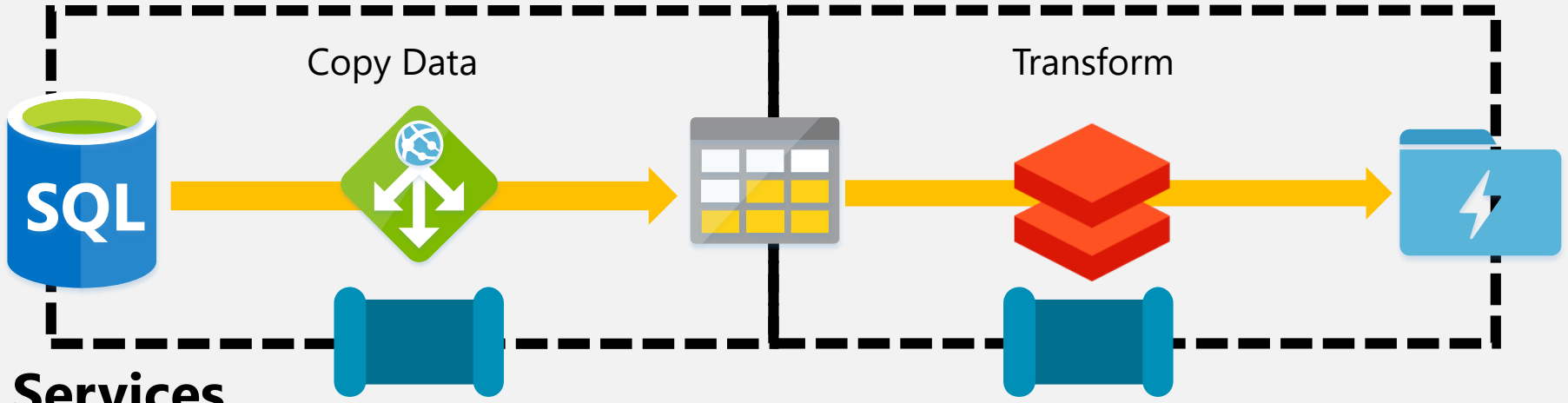
Data Factory Components



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

adatis

Data Factory Components

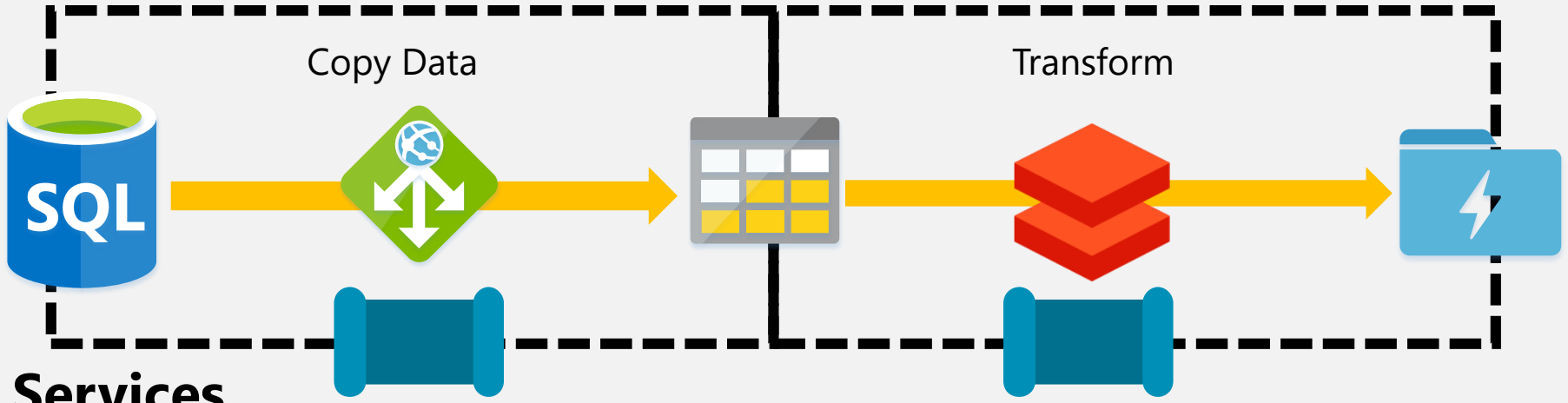


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers** – How are we going to tell our pipeline(s) to execute?

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

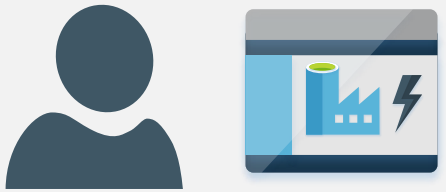
adatis

Data Factory Components



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

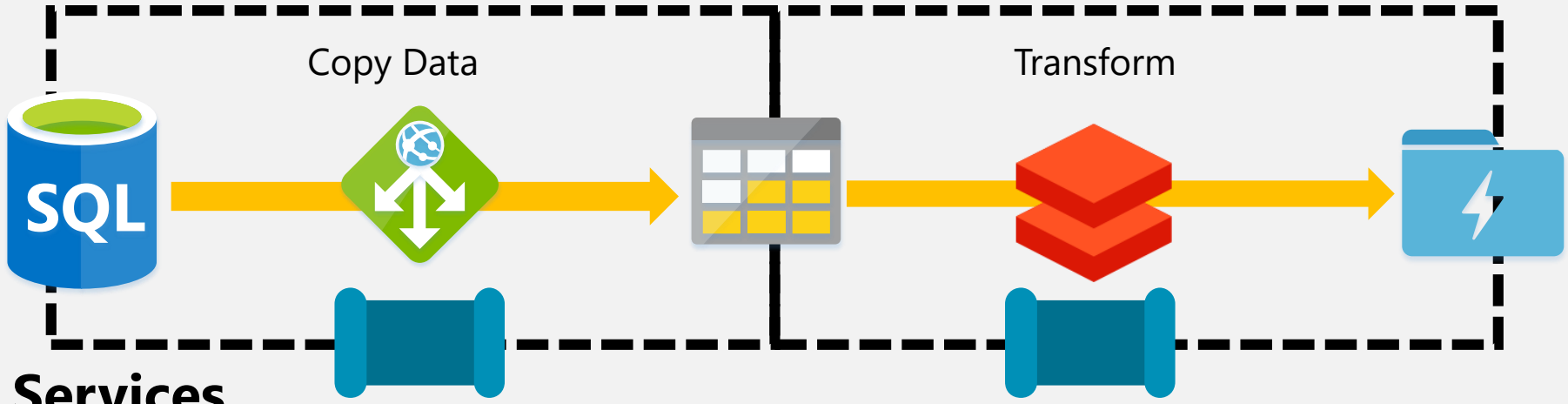
- **Manual**
- Tumbling Windows
- Scheduled
- Blob File Events
- Logic App Calls



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

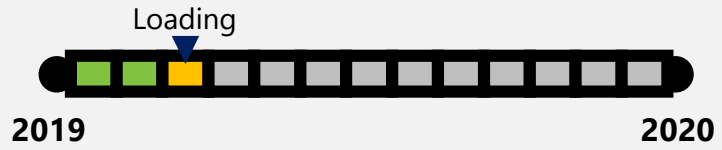
adatis

Data Factory Components



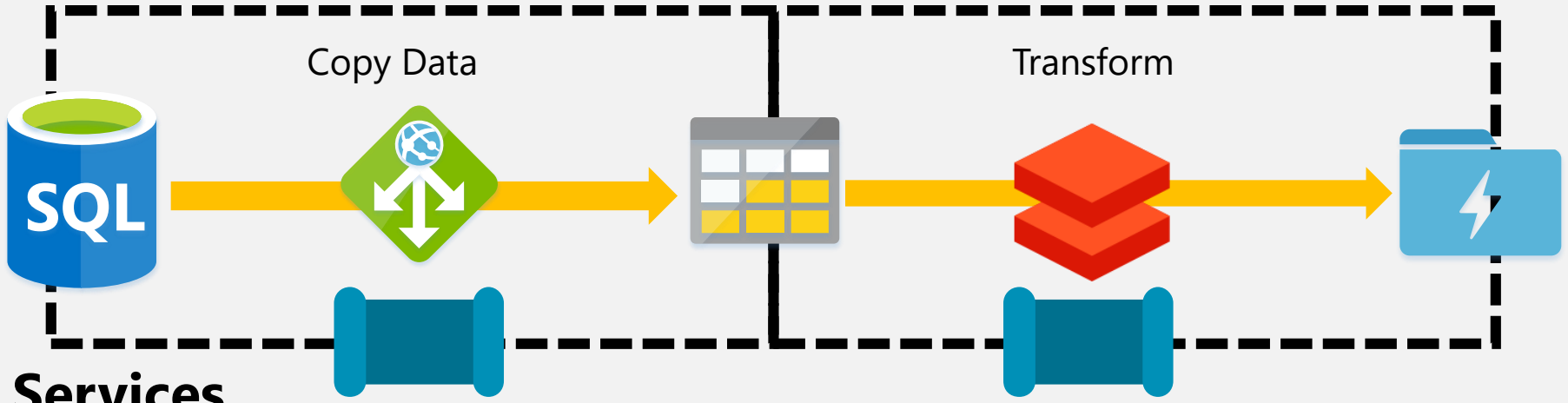
1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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



adatis

Data Factory Components



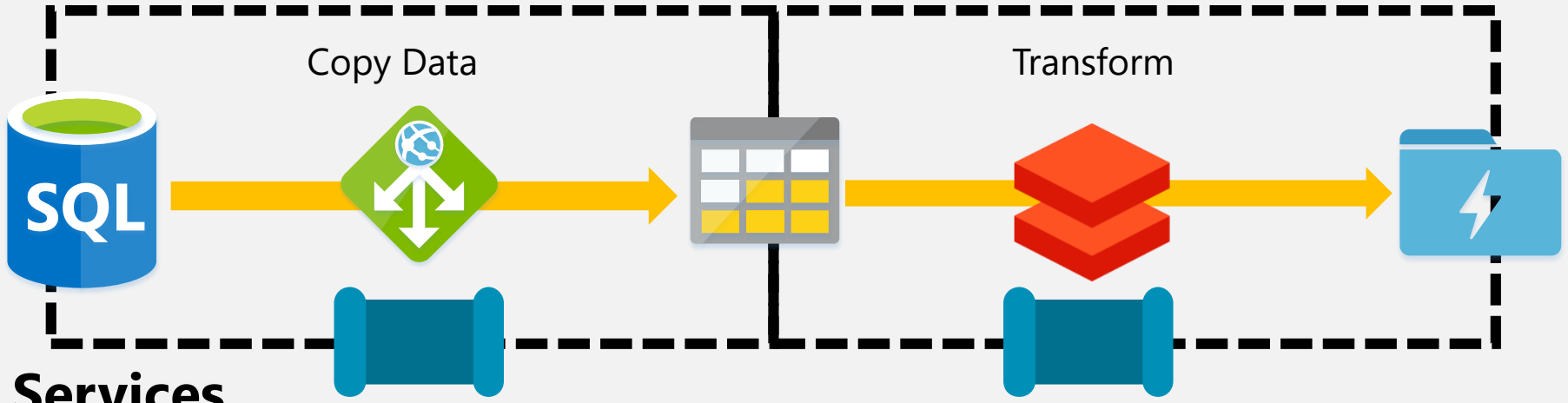
1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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

- Every 1 minute.
- UTC

adatis

Data Factory Components



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

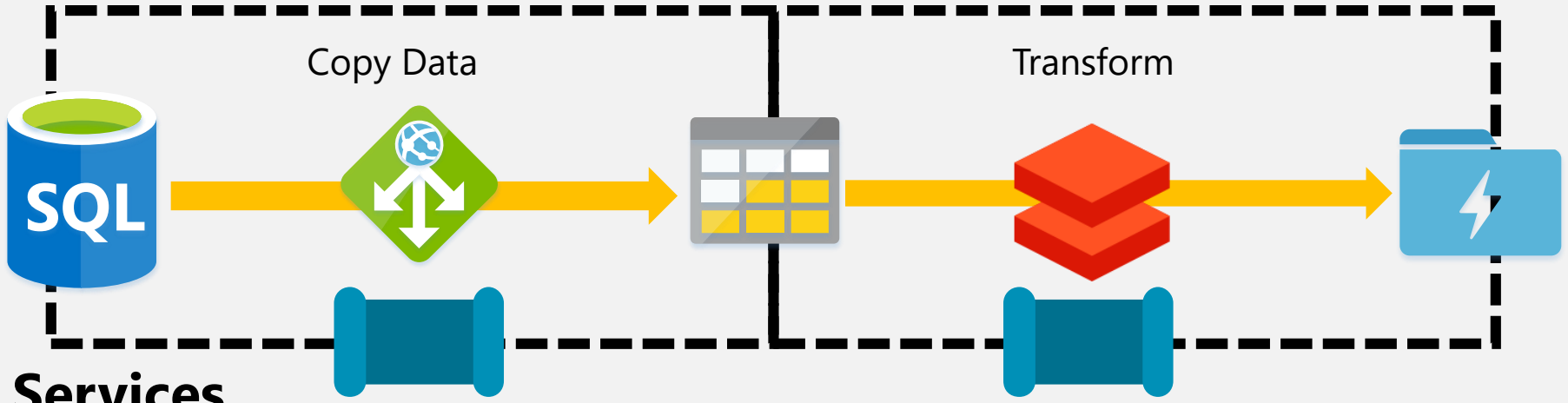
- Manual via UI
- Tumbling Windows
- Scheduled
- **Blob File Events**
- Logic App Calls

CSV

- {Path} Created
- {Path} Deleted

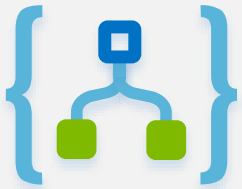
adatis

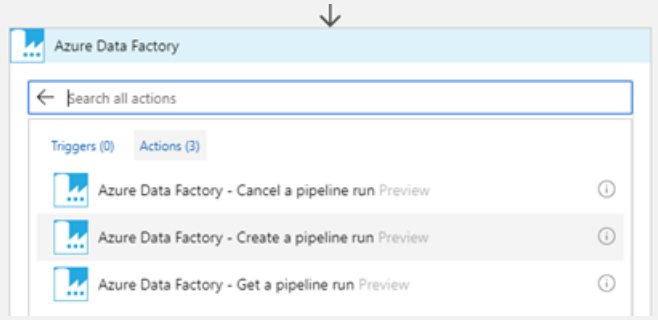
Data Factory Components



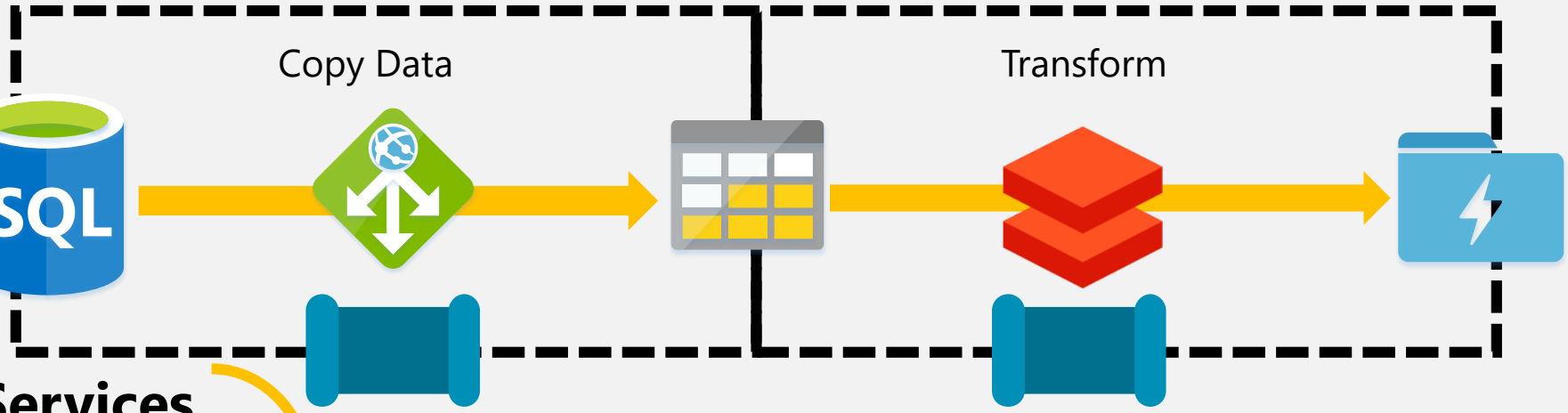
1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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





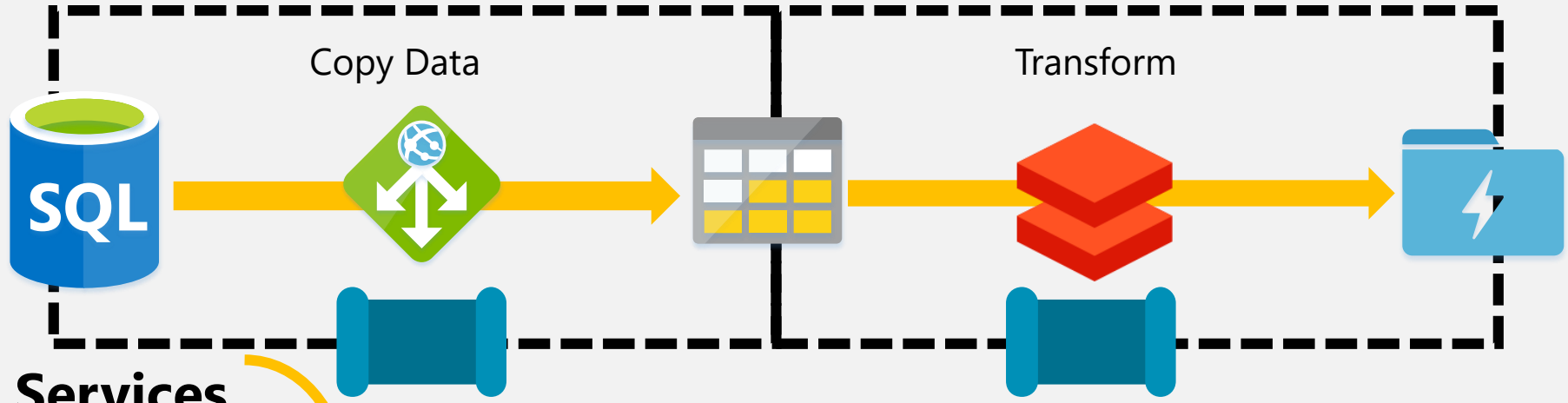
Data Factory Components



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

adatis

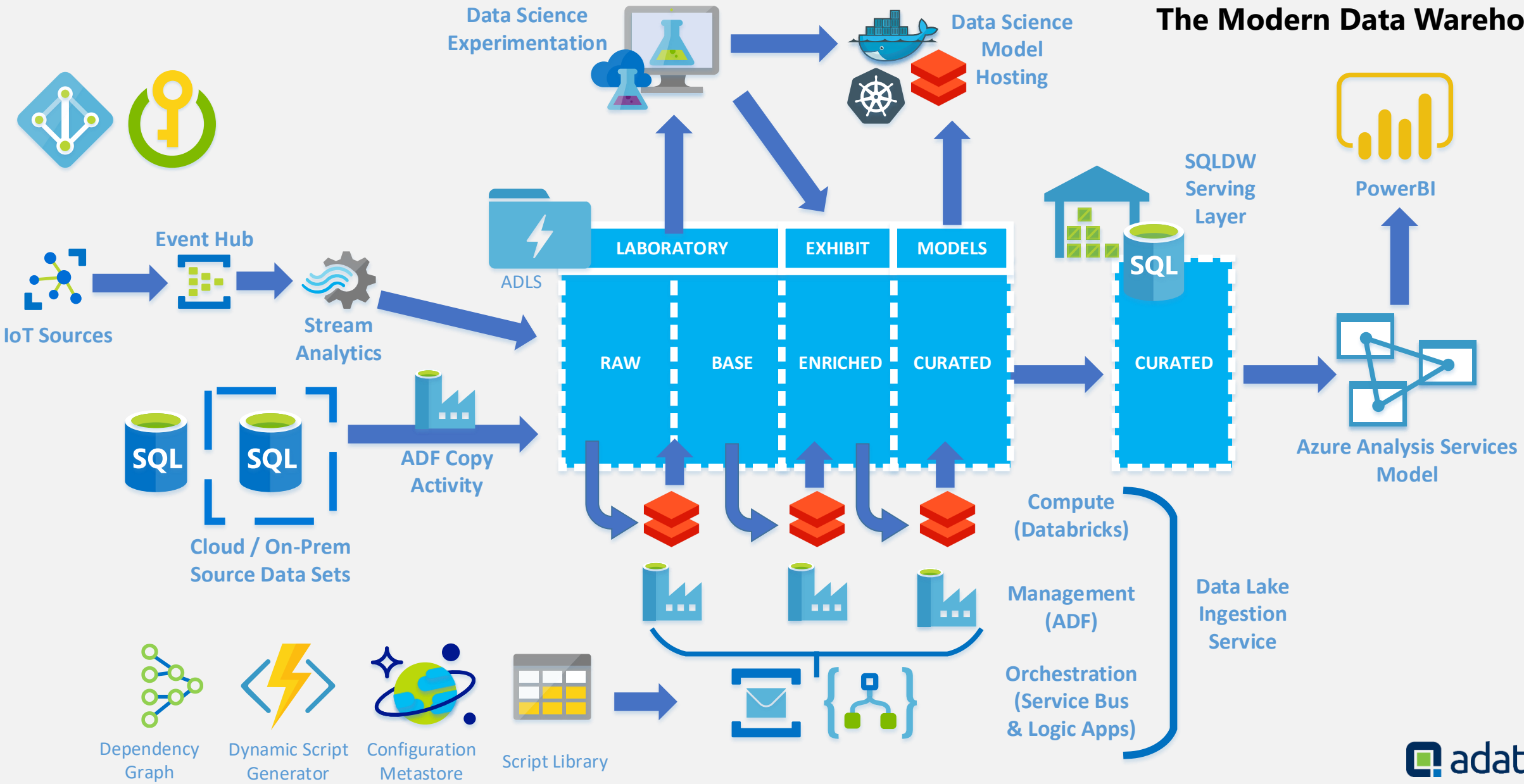
Data Factory Control Flow Components



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

adatis

Why use Azure Data Factory?

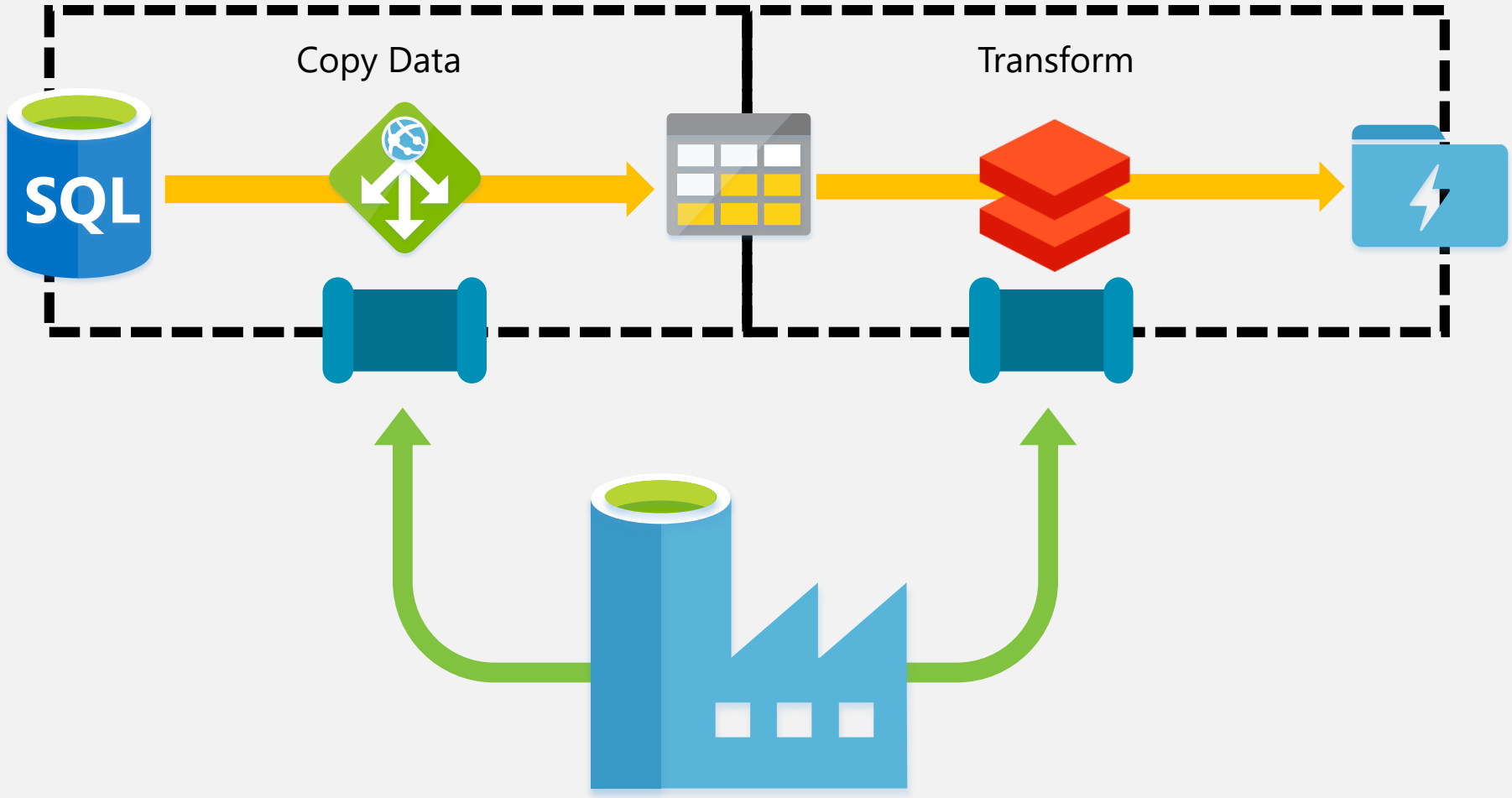


adatis

Data Transformation in Azure

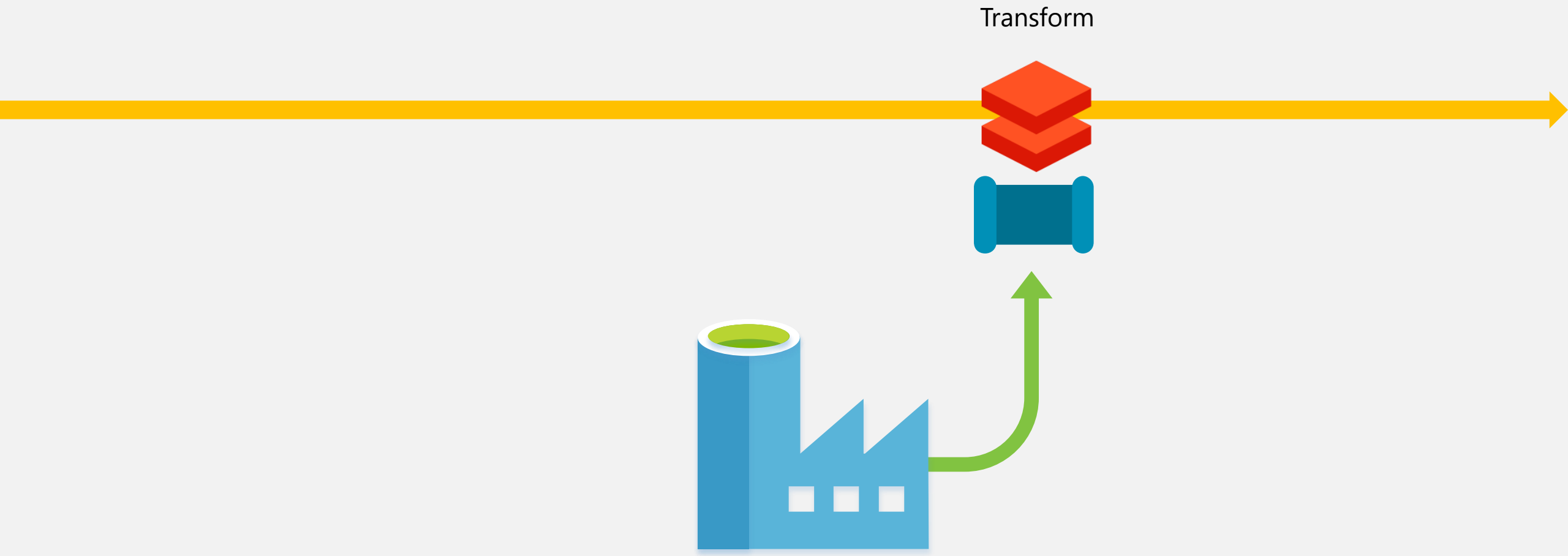
adatis

Data Factory Control Flow Components



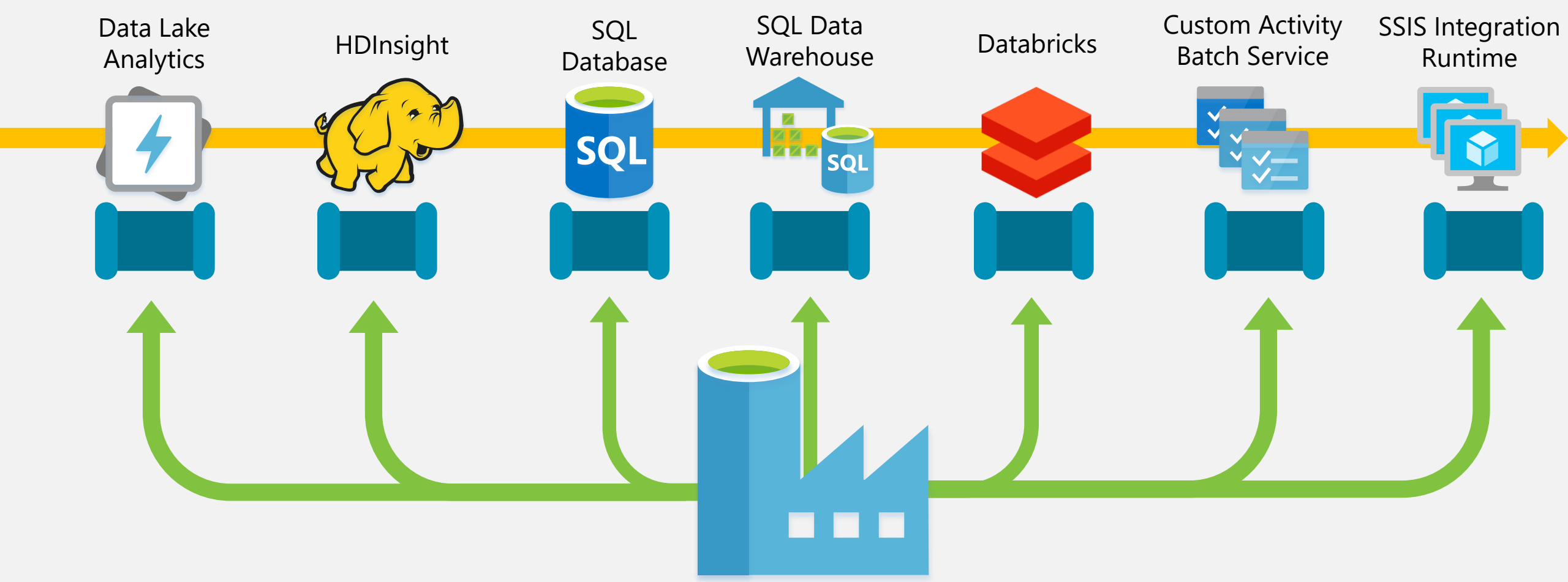
adatis

Data Transformation in Azure



adatis

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

adatis

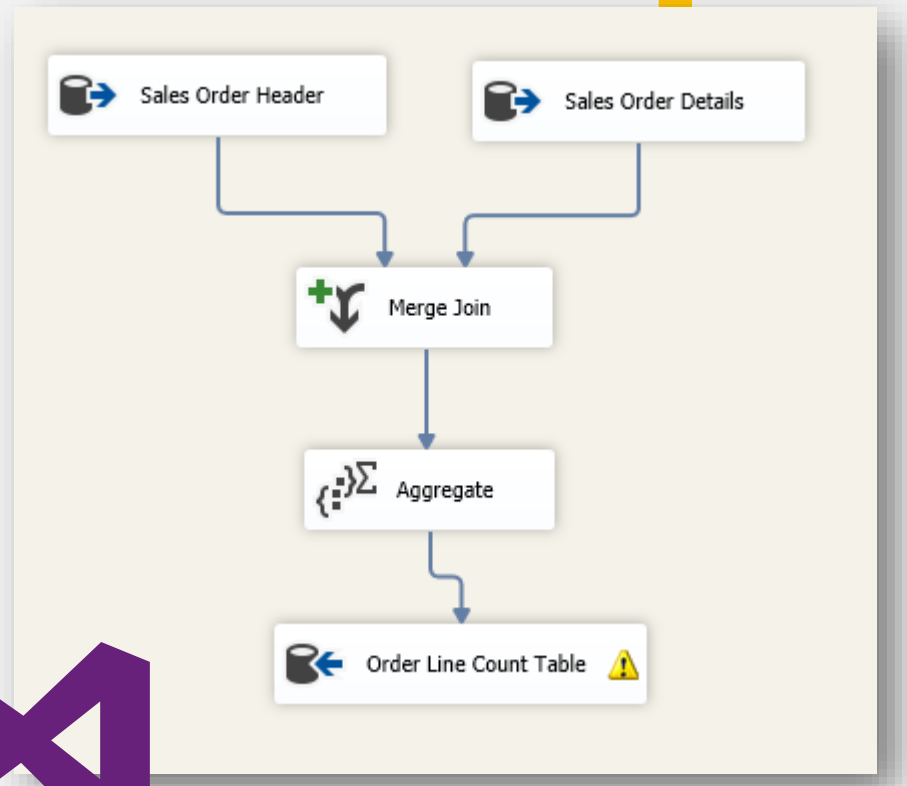
Data Transformation in Azure with SSIS

SSIS Package

8x Packages per Node

10x Nodes

SSIS Integration Runtime



Visual Studio – SQL Server Data Tools (SSDT)

Data Factory

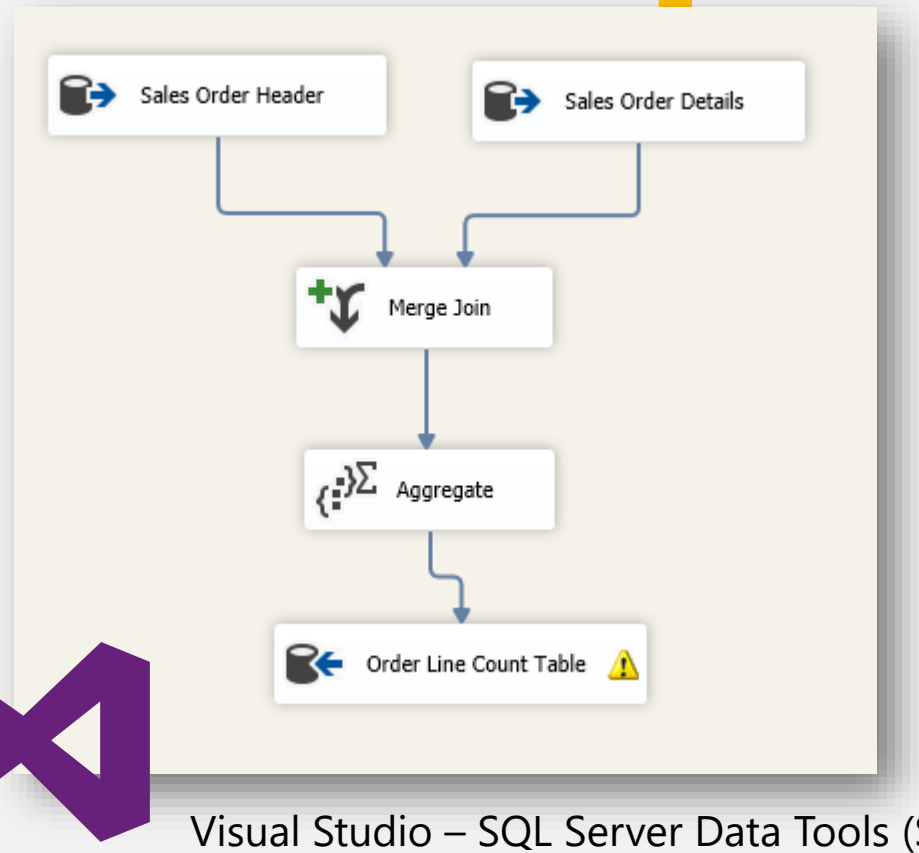
Logical or Managed Instance

SSISDB

adatis

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

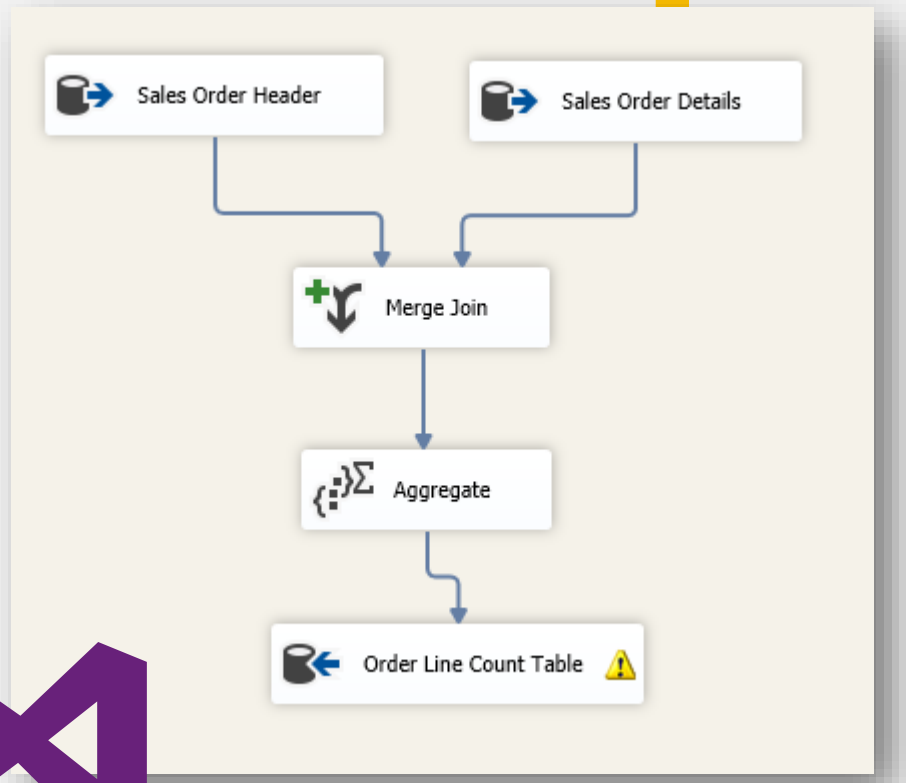
Logical or Managed Instance

SSISDB

adatis

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

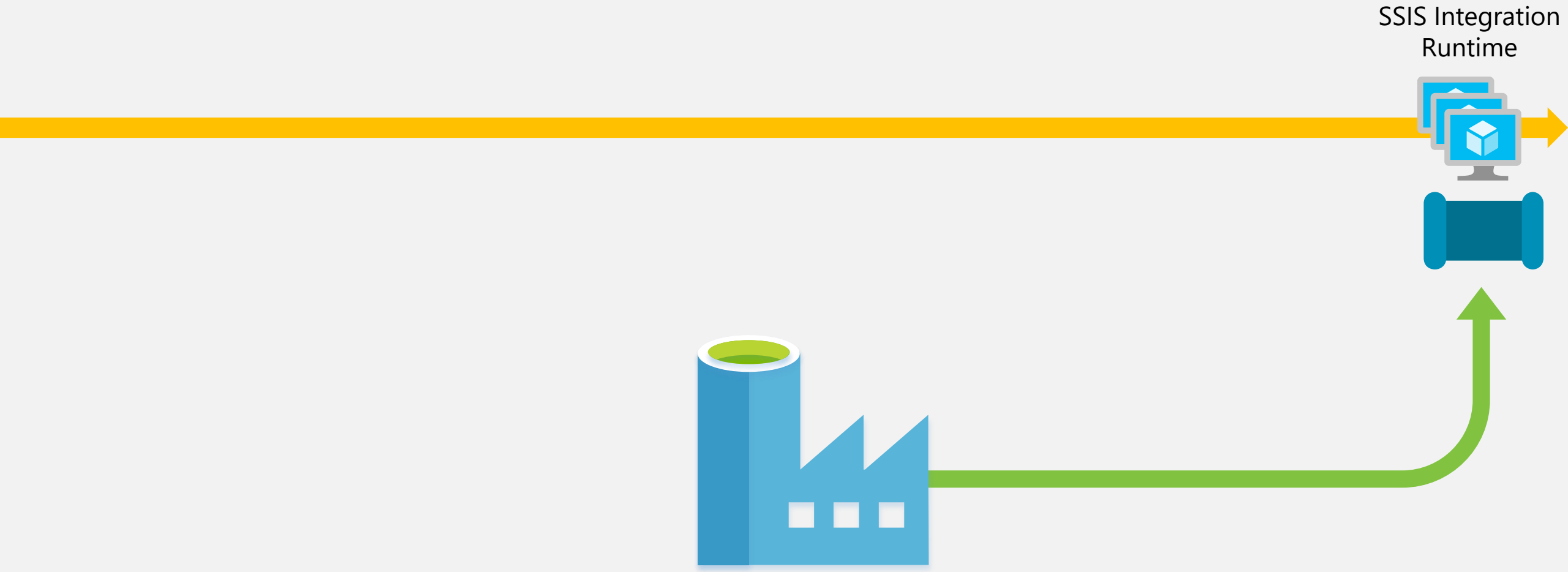
Data Factory

Logical or Managed Instance

SSISDB

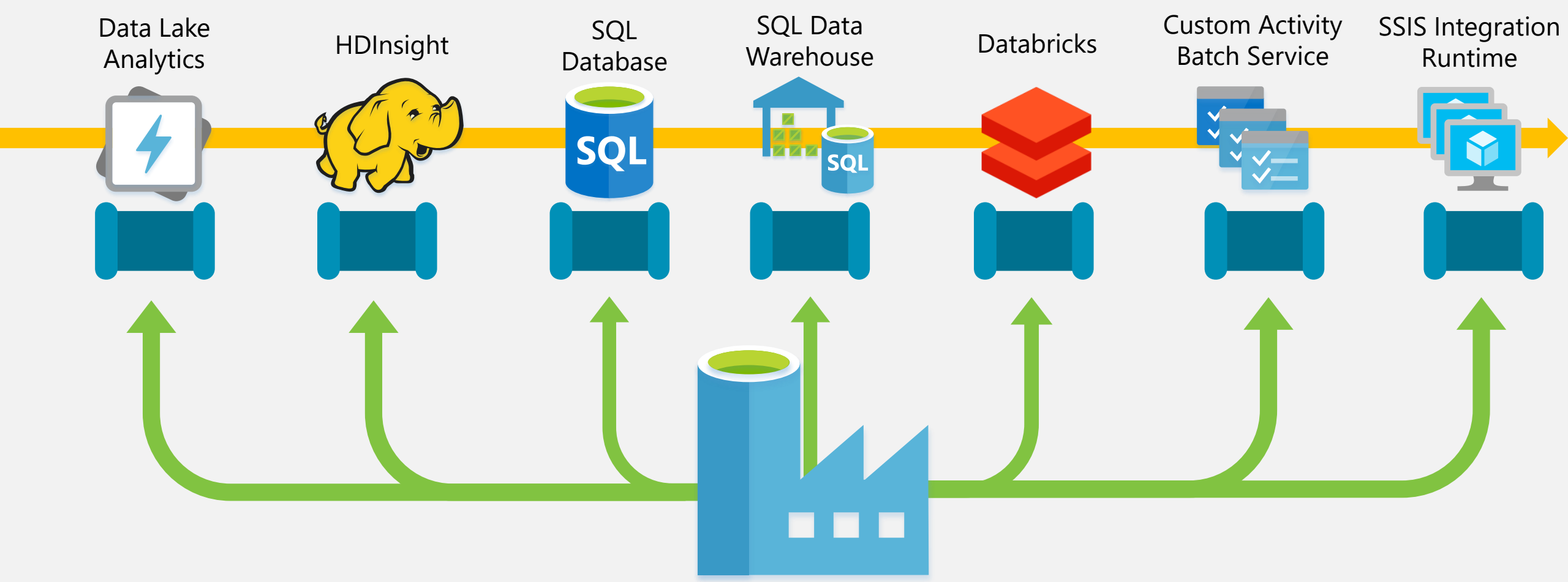
adatis

Data Transformation in Azure



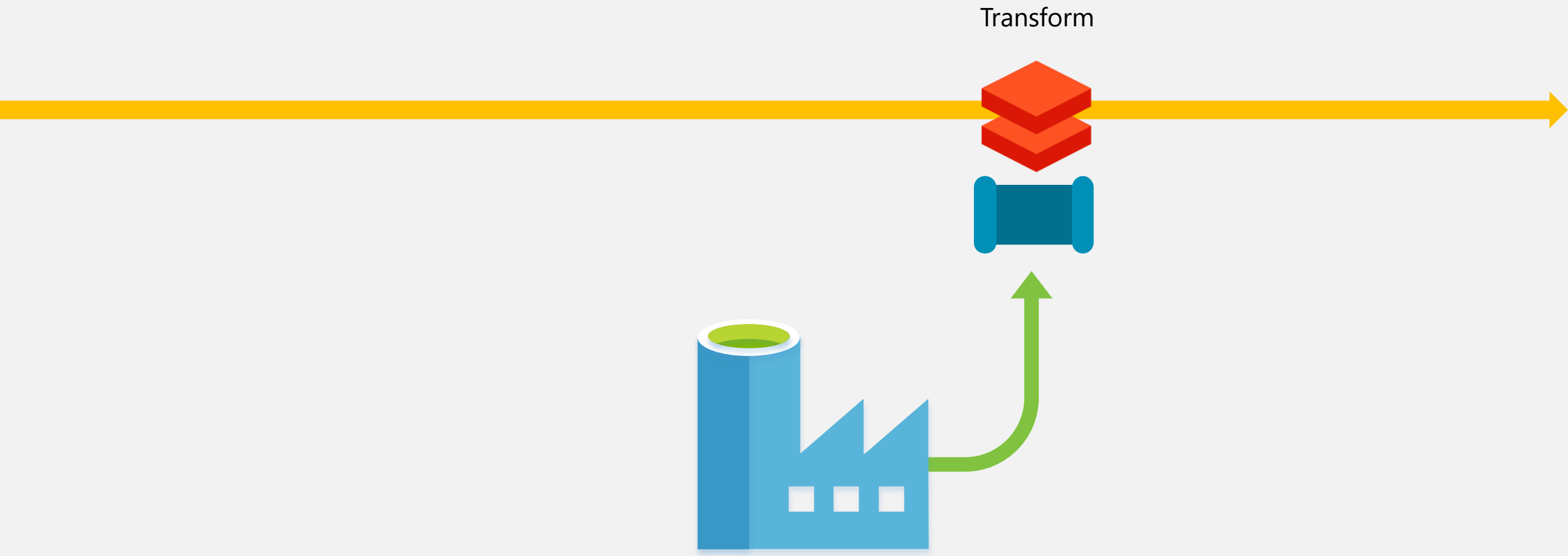
adatis

Data Transformation in Azure



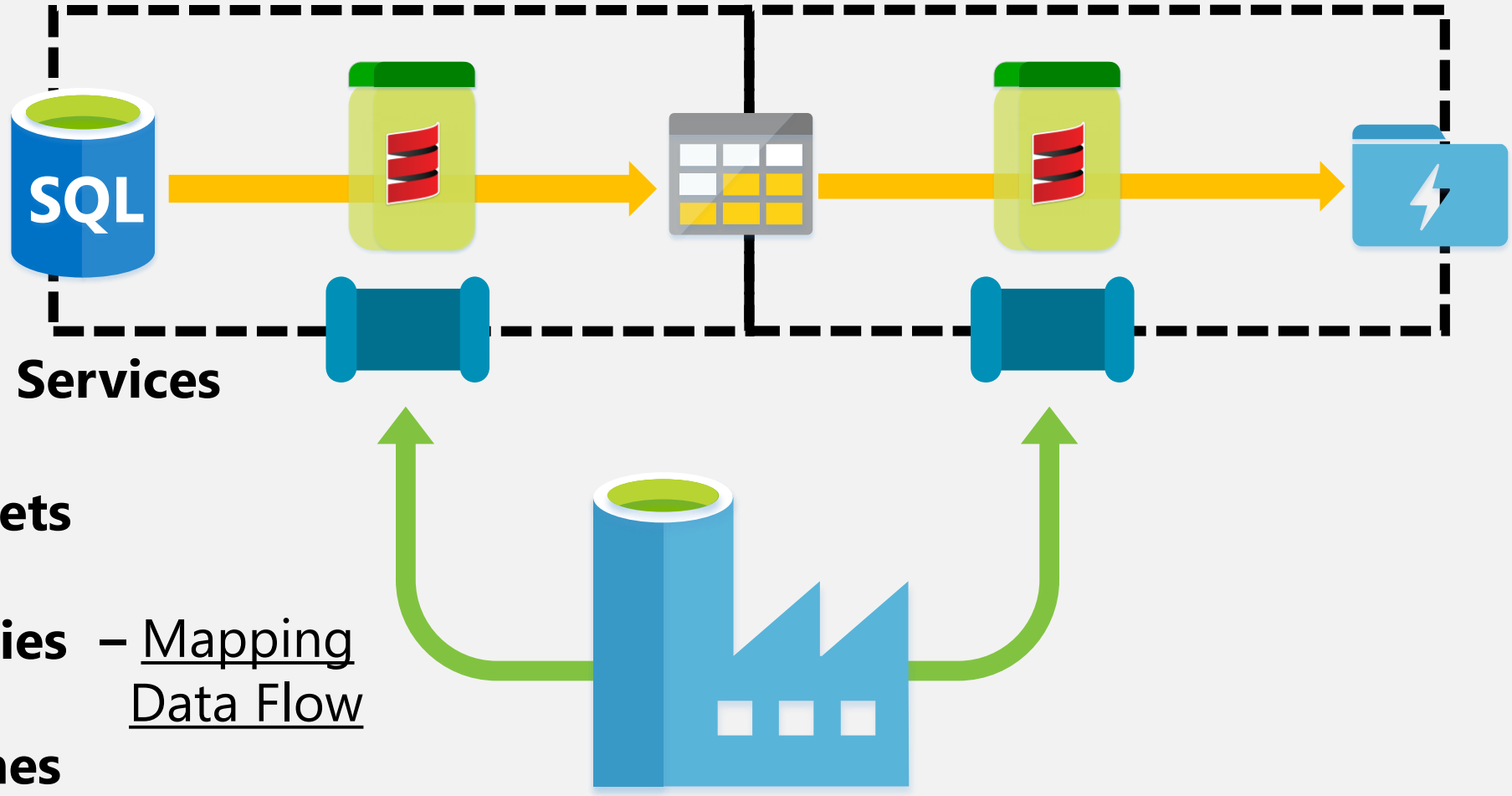
adatis

Data Transformation in Azure



adatis

Data Factory Control Flow Components



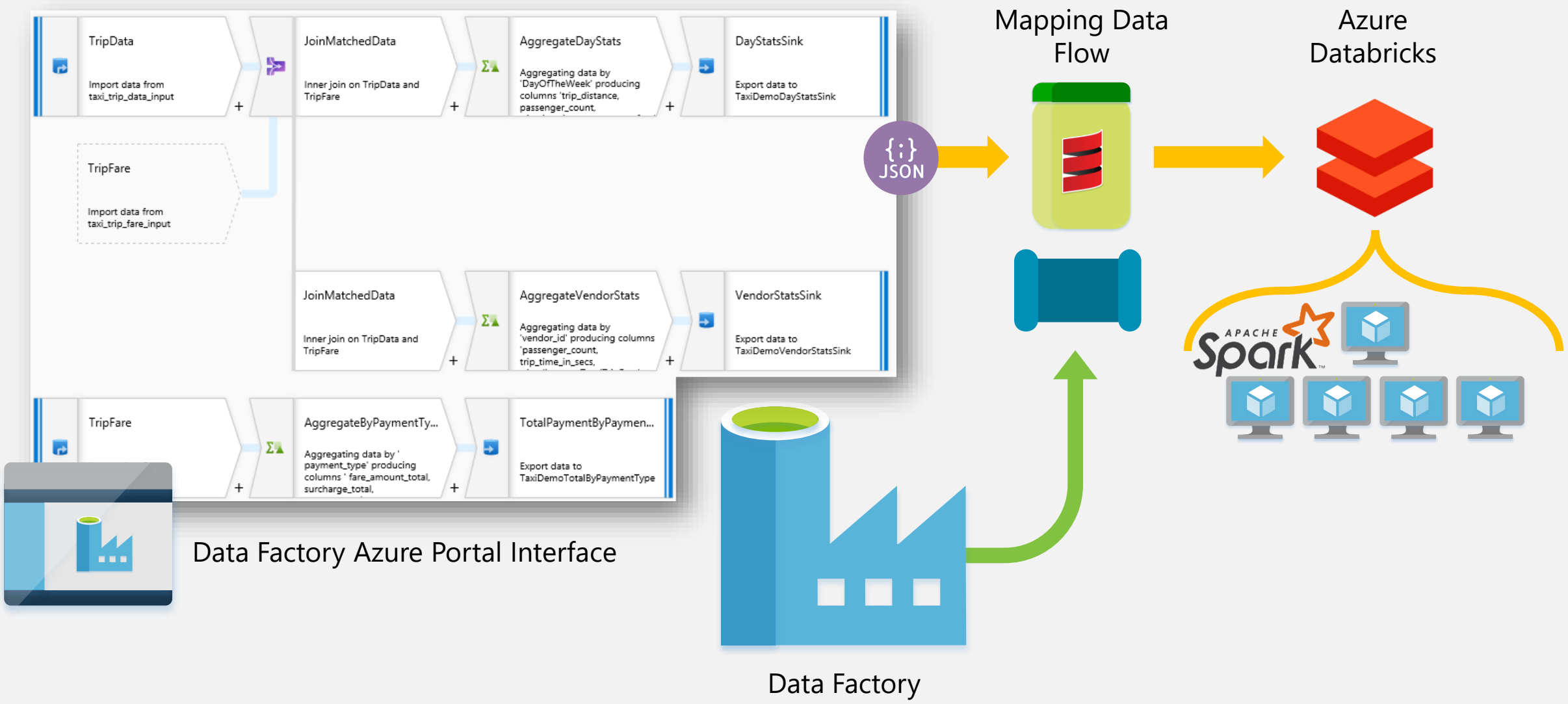
1. **Linked Services**
2. **Data Sets**
3. **Activities** – Mapping Data Flow
4. **Pipelines**
5. **Triggers**

adatis

Mapping Data Flows

adatis

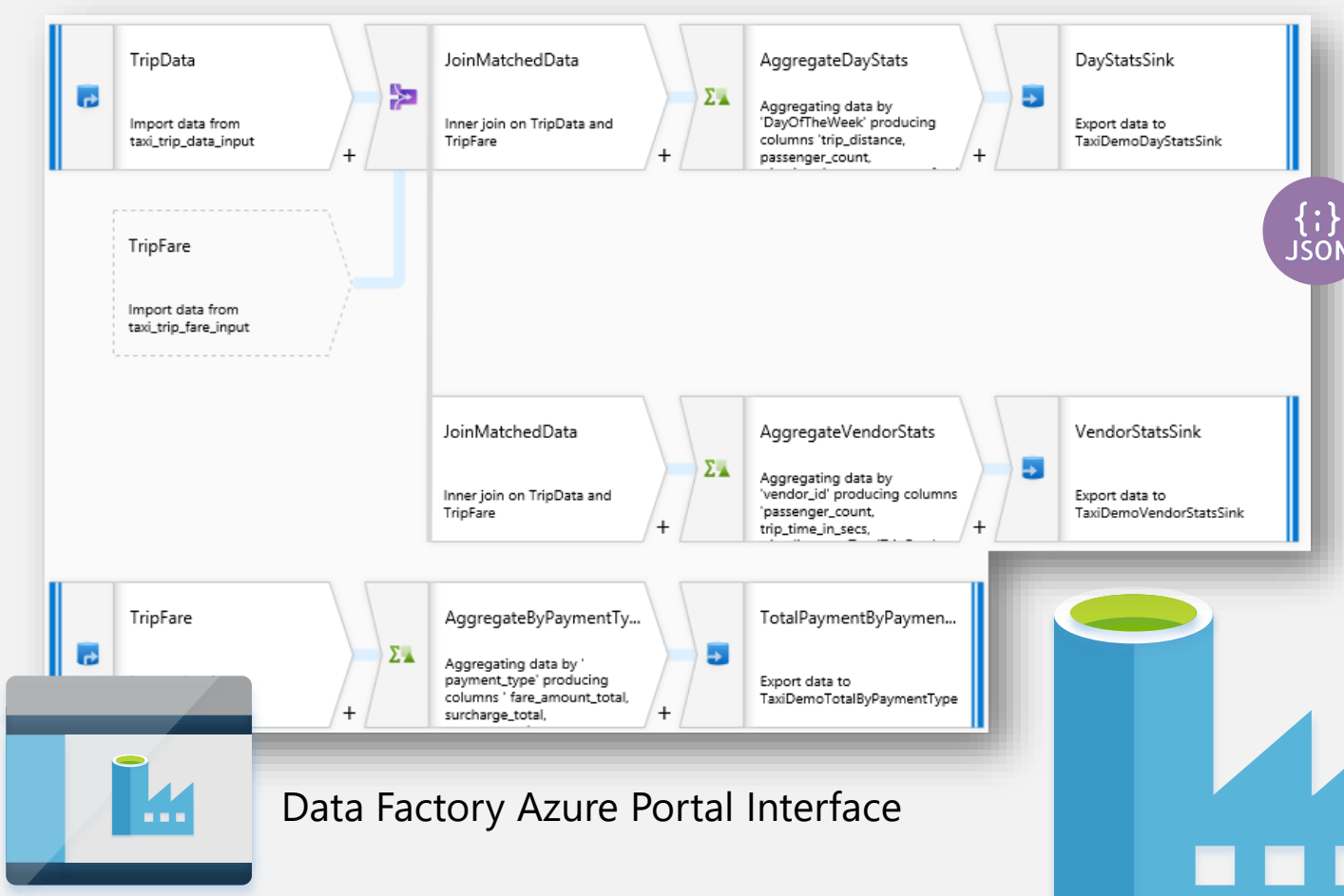
What is a Mapping Data Flow?



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

adatis

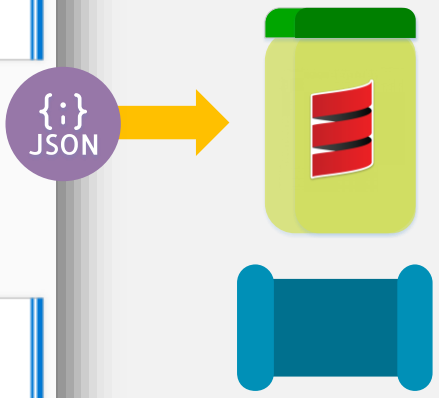
What is a Mapping Data Flow?



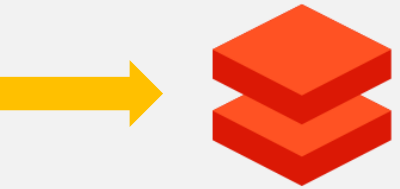
Data Factory Azure Portal Interface

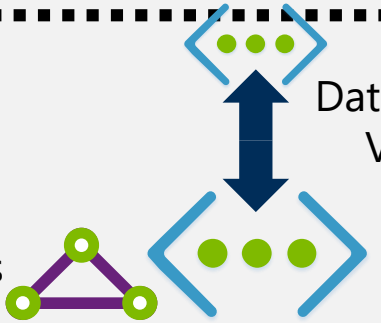
Data Factory

Mapping Data Flow



Azure Databricks





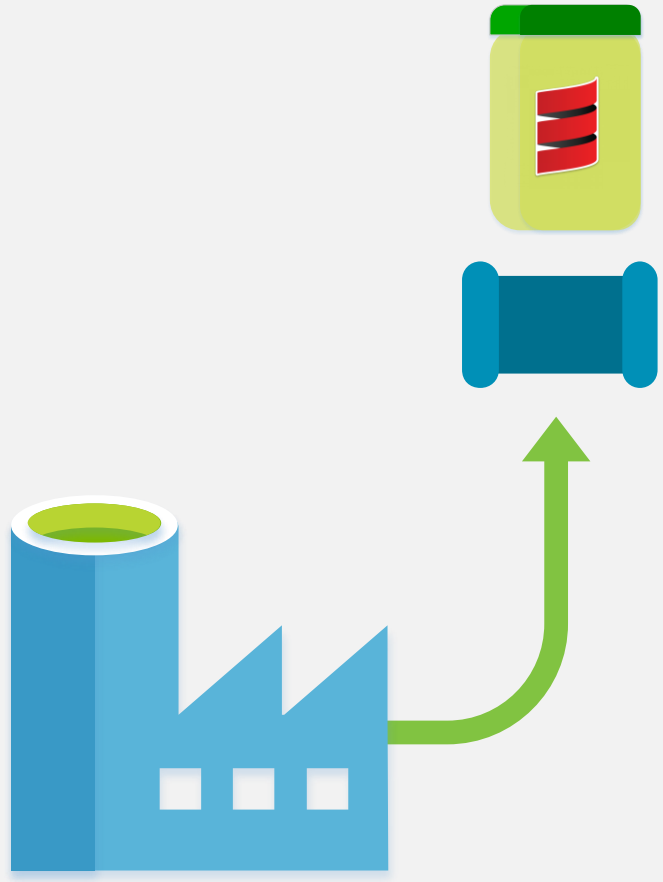
Databricks VNet

Express Route

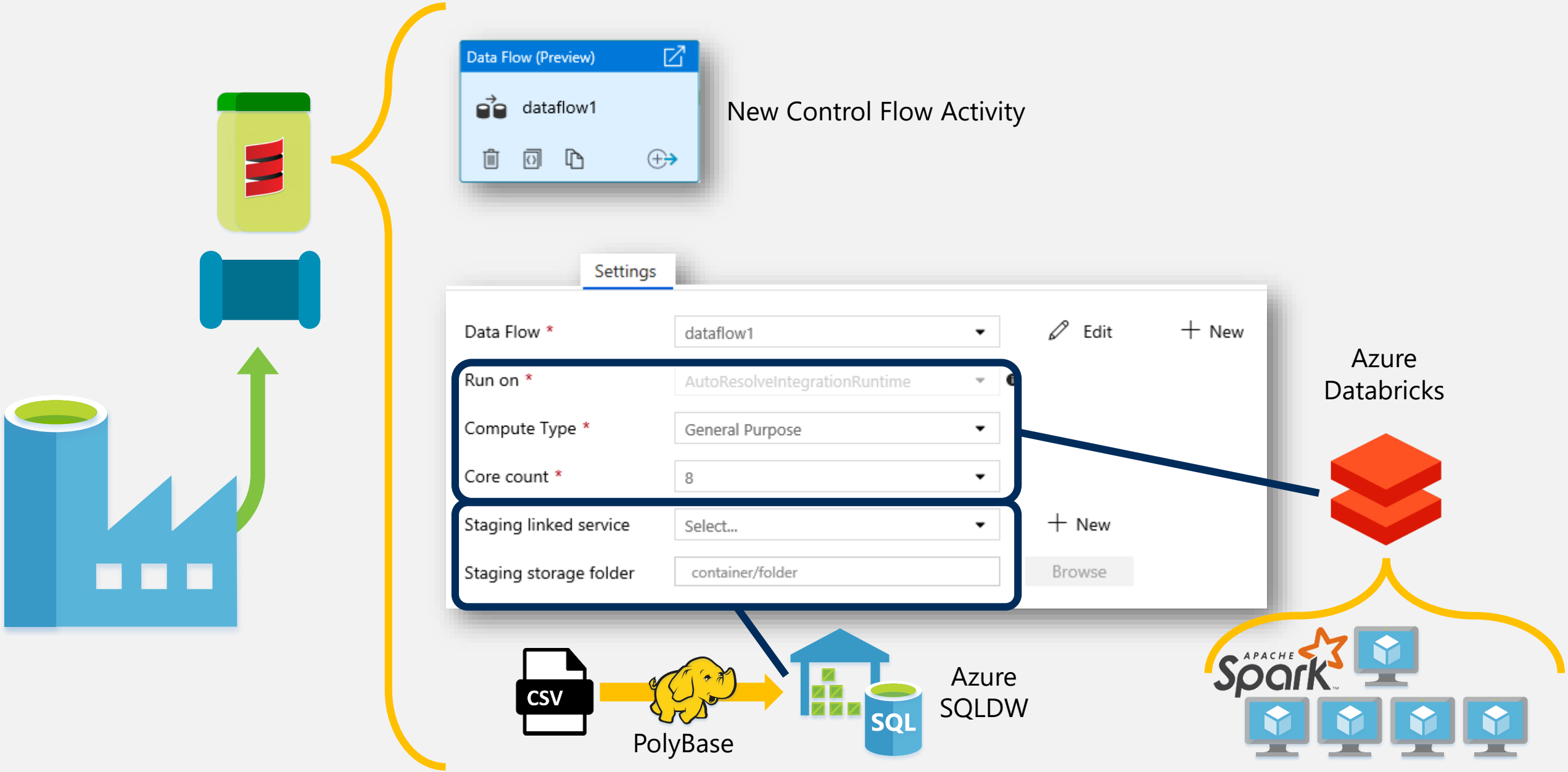
Production VNet

adatis

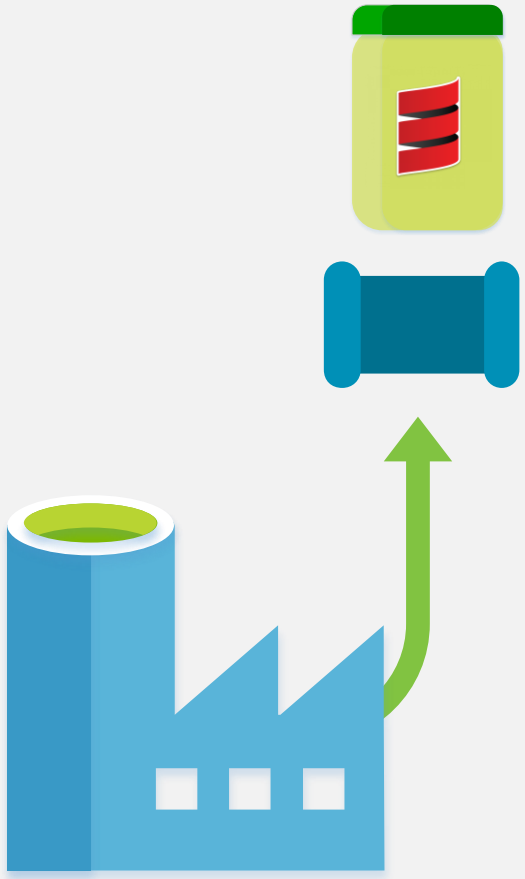
Mapping Data Flows



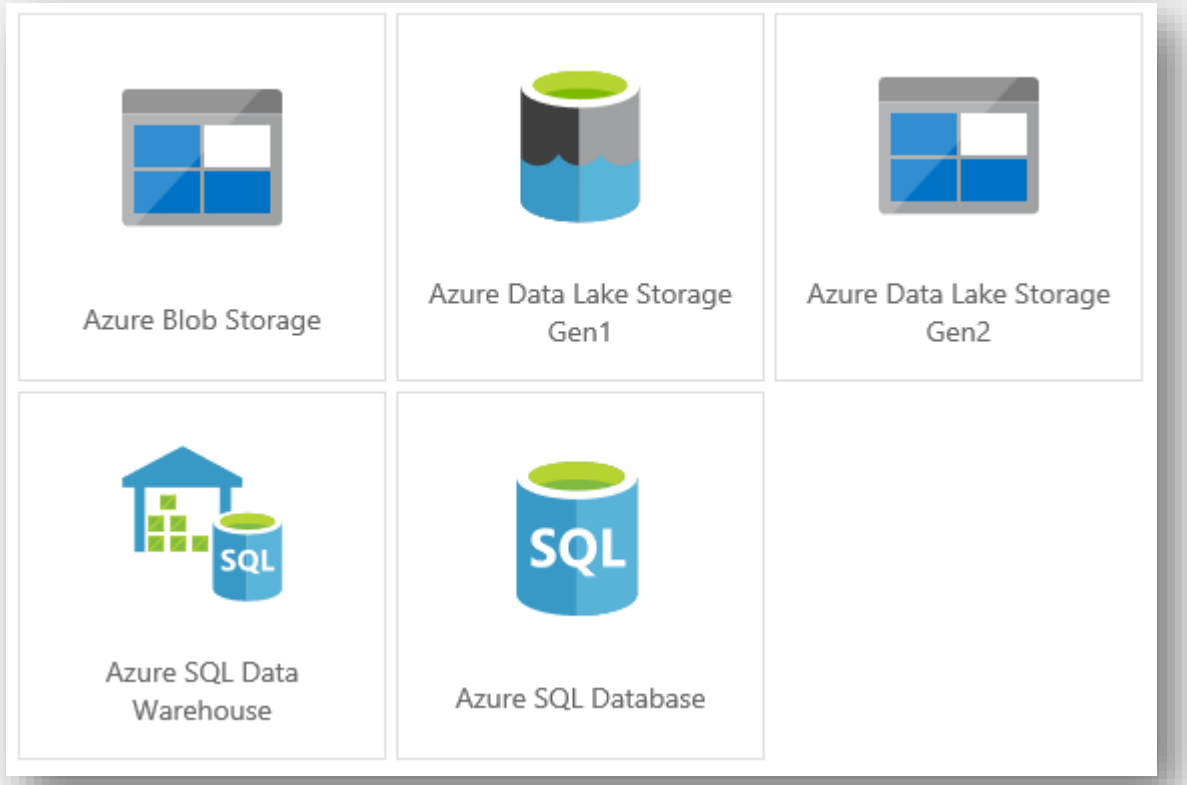
Mapping Data Flows – Settings & Concepts



Mapping Data Flows – Settings & Concepts

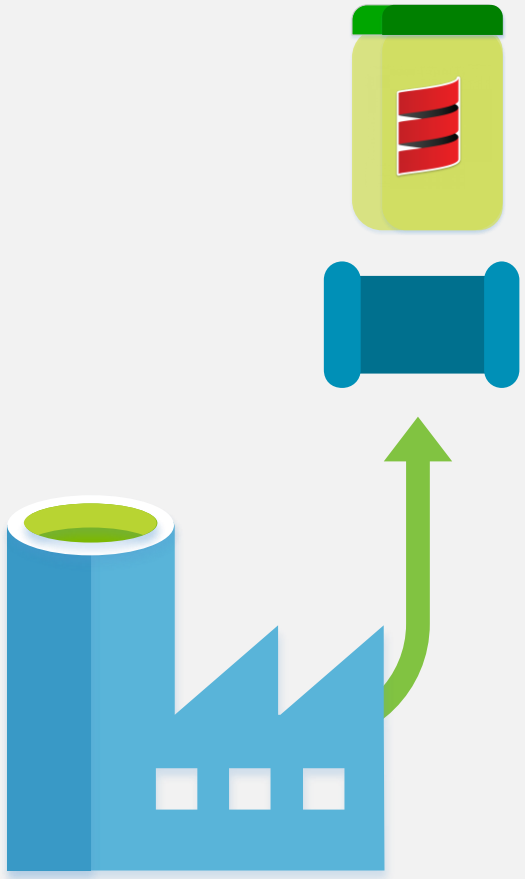


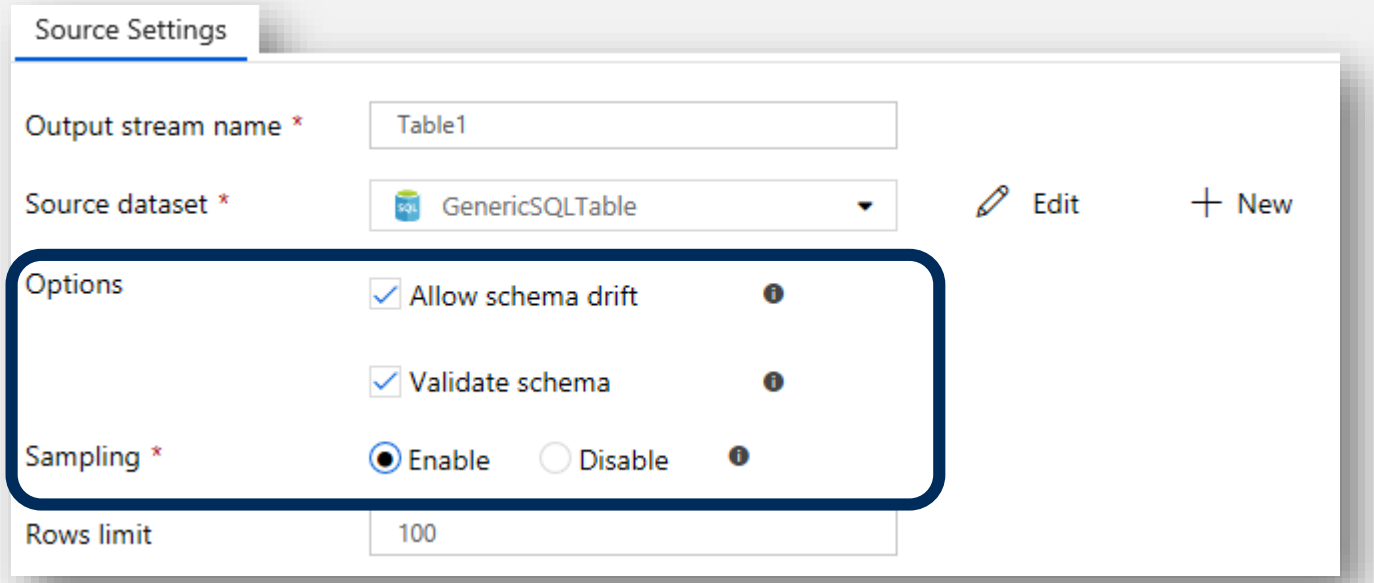
Currently Available:



adatis

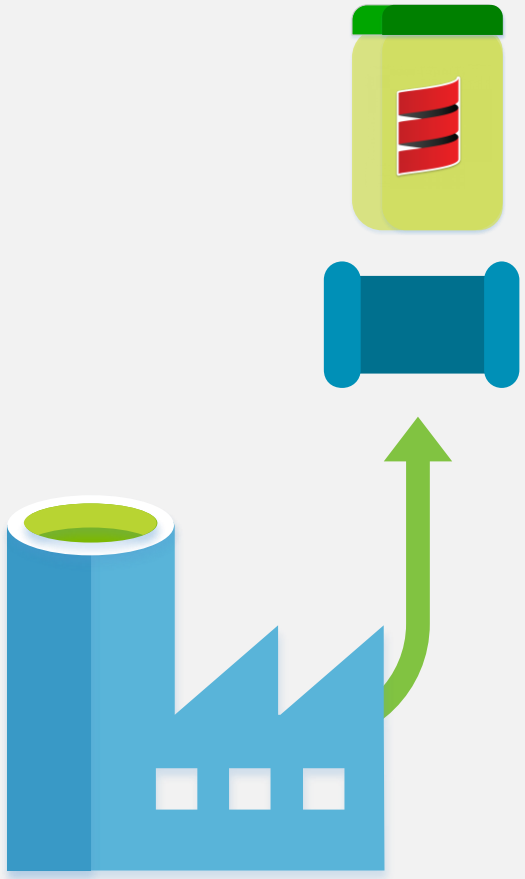
Mapping Data Flows – Settings & Concepts

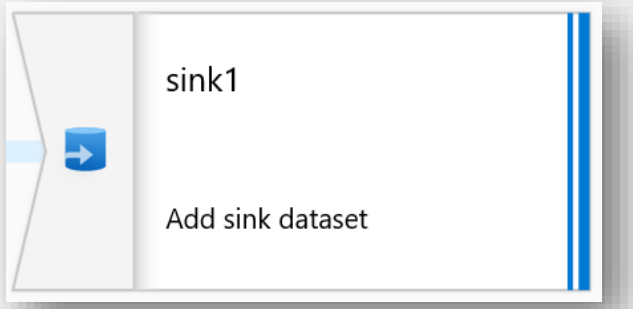


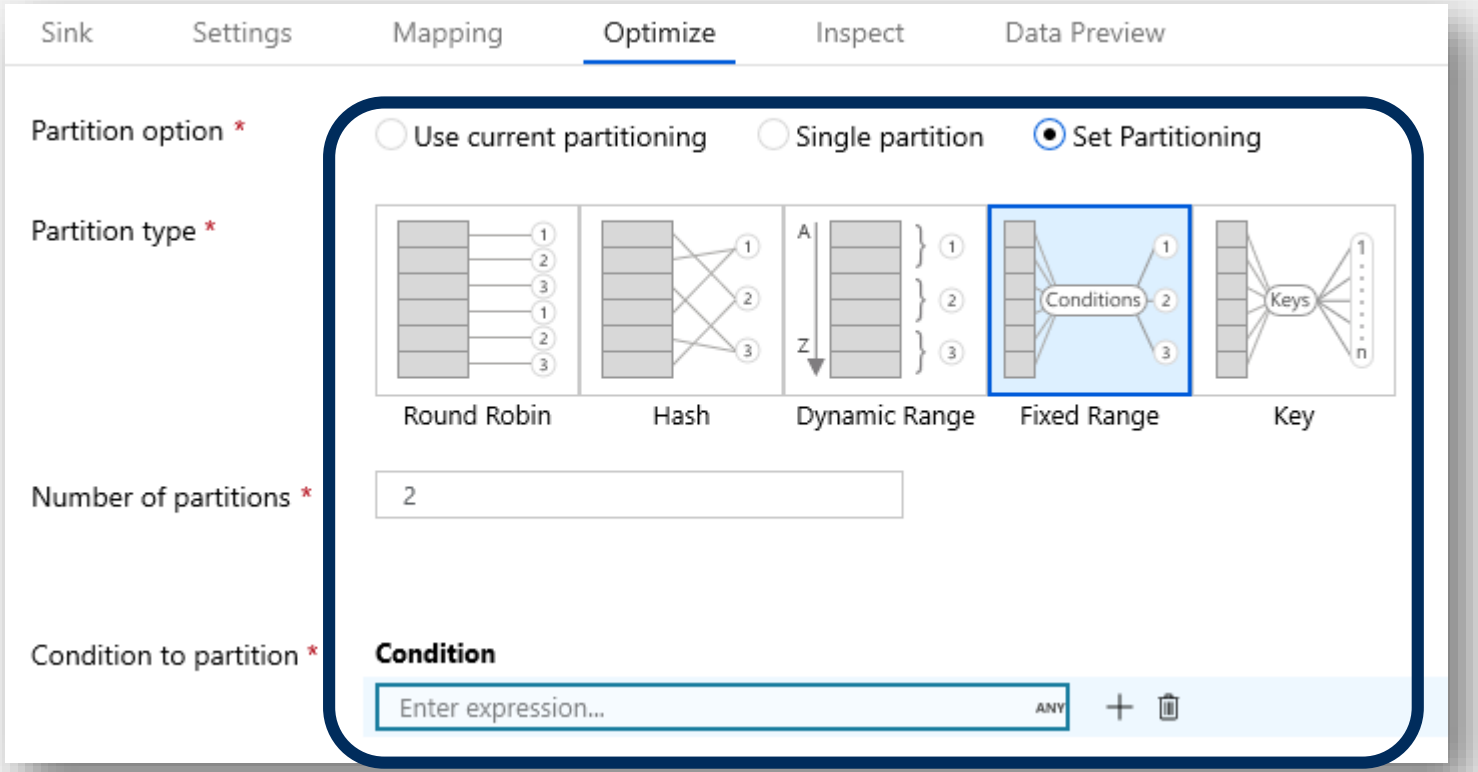


adatis

Mapping Data Flows – Settings & Concepts

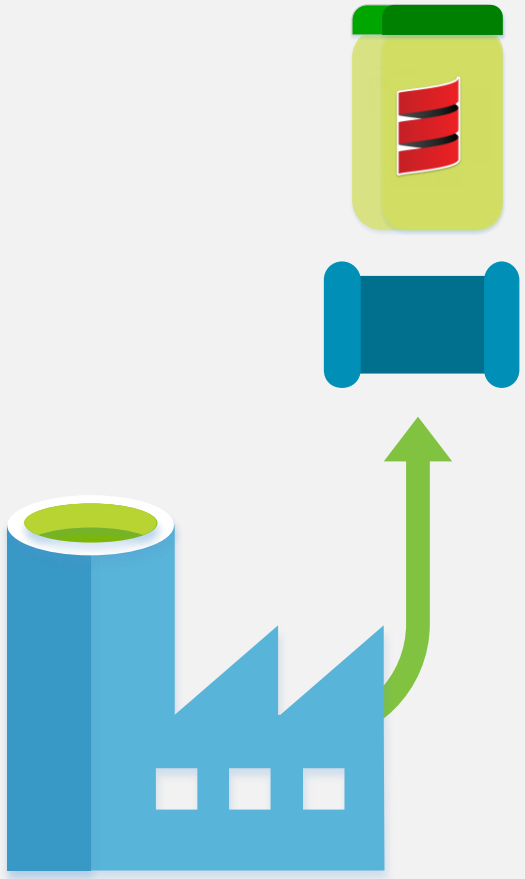






adatis

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

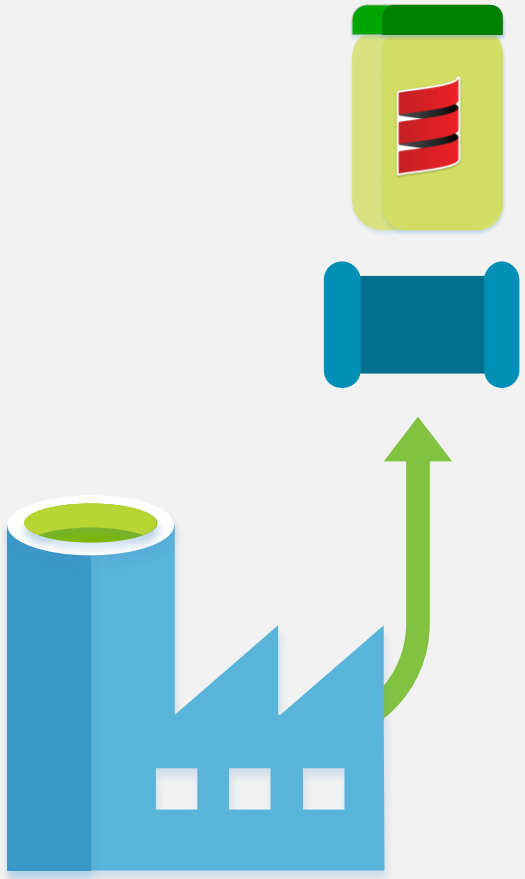
- Exists
- Select
- Filter
- Sort

Custom

- Extend — Script Component

adatis

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'

+ - * / ||

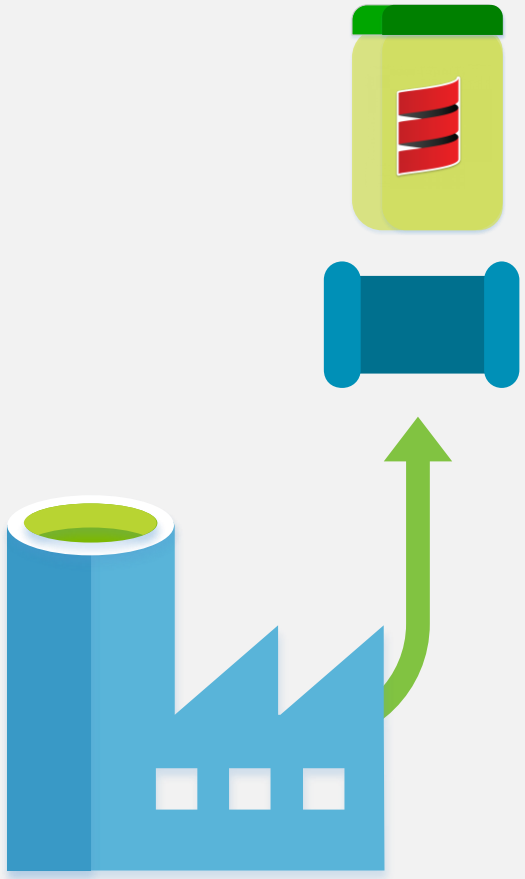
Data preview

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

Output: year 123	title abc
-	-

adatis

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

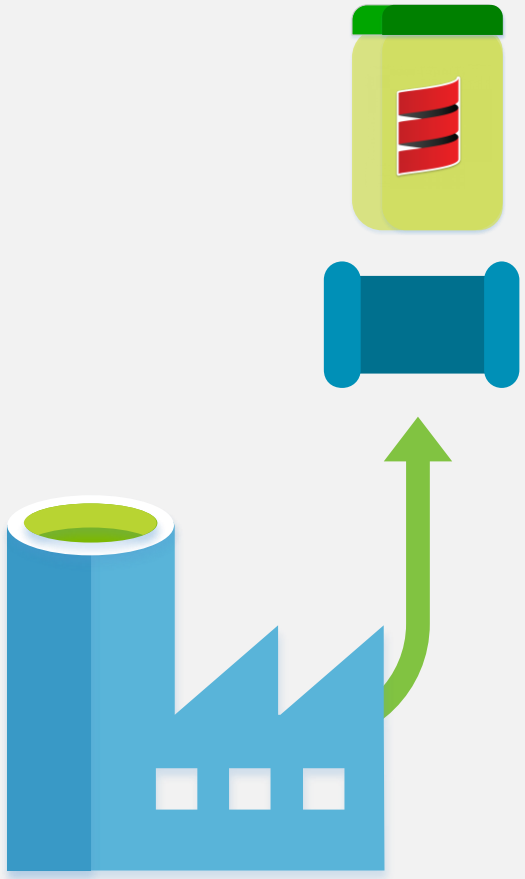
Source Settings | Define schema | Optimize | Inspect | Data Preview

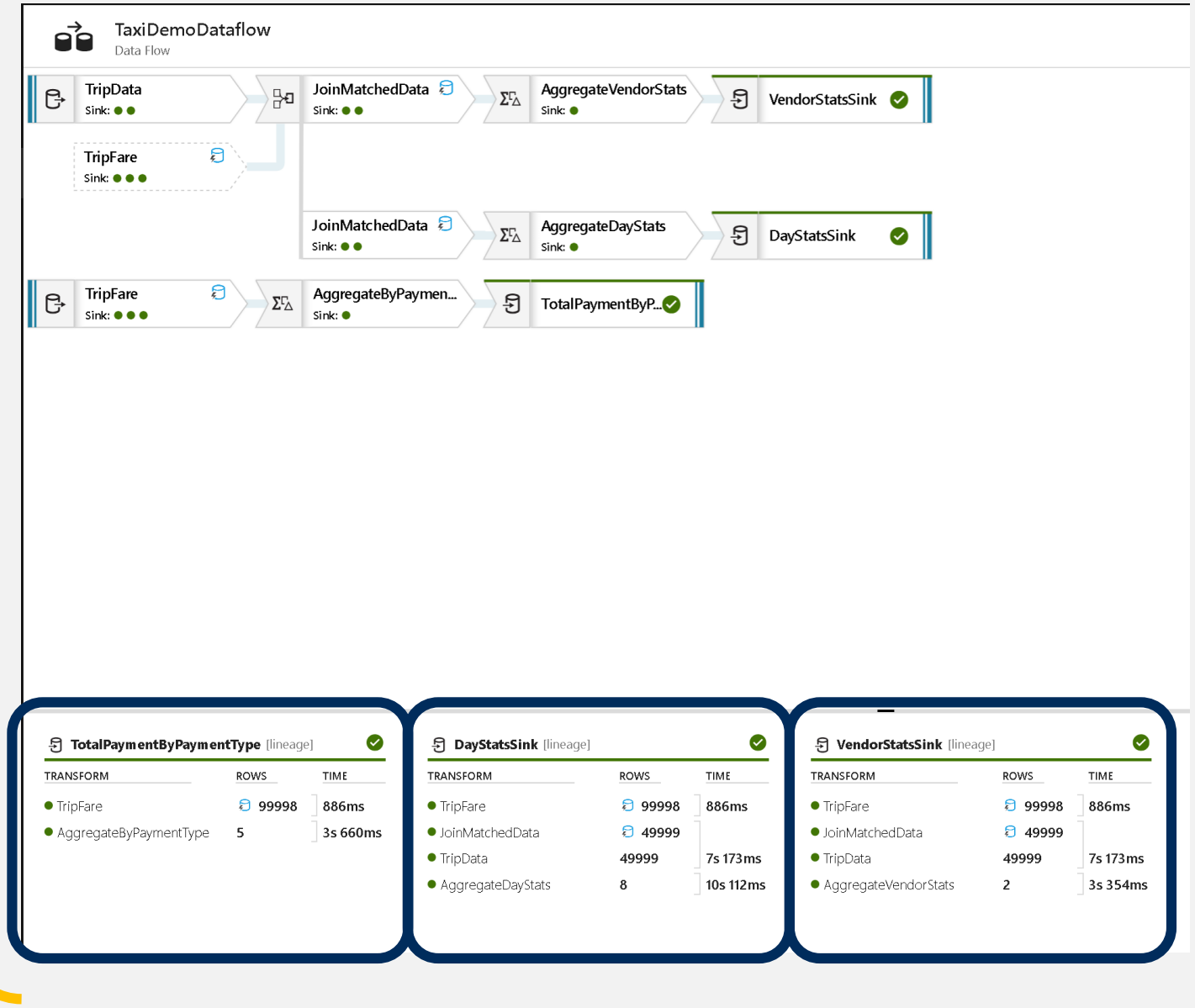
	Updated*	New*	Unchanged	Total
Number of rows	N/A	N/A	N/A	32

SalesOrderID 123	RevisionNumber abc	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

adatis

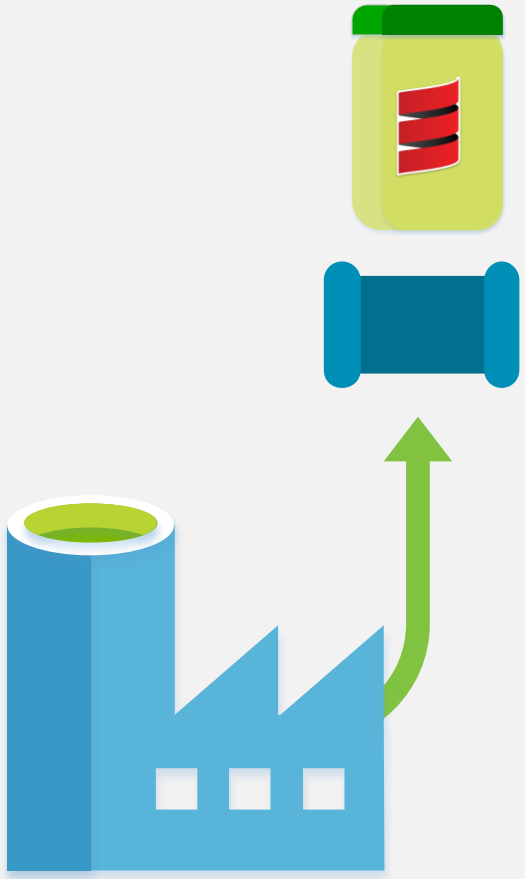
Mapping Data Flows – Monitoring



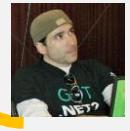


adatis

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



https://github.com/kromerm/adfdataflowdocs/tree/master/videos

adatis

Demo

Note to self; start clusters!

adatis

Demo Summary

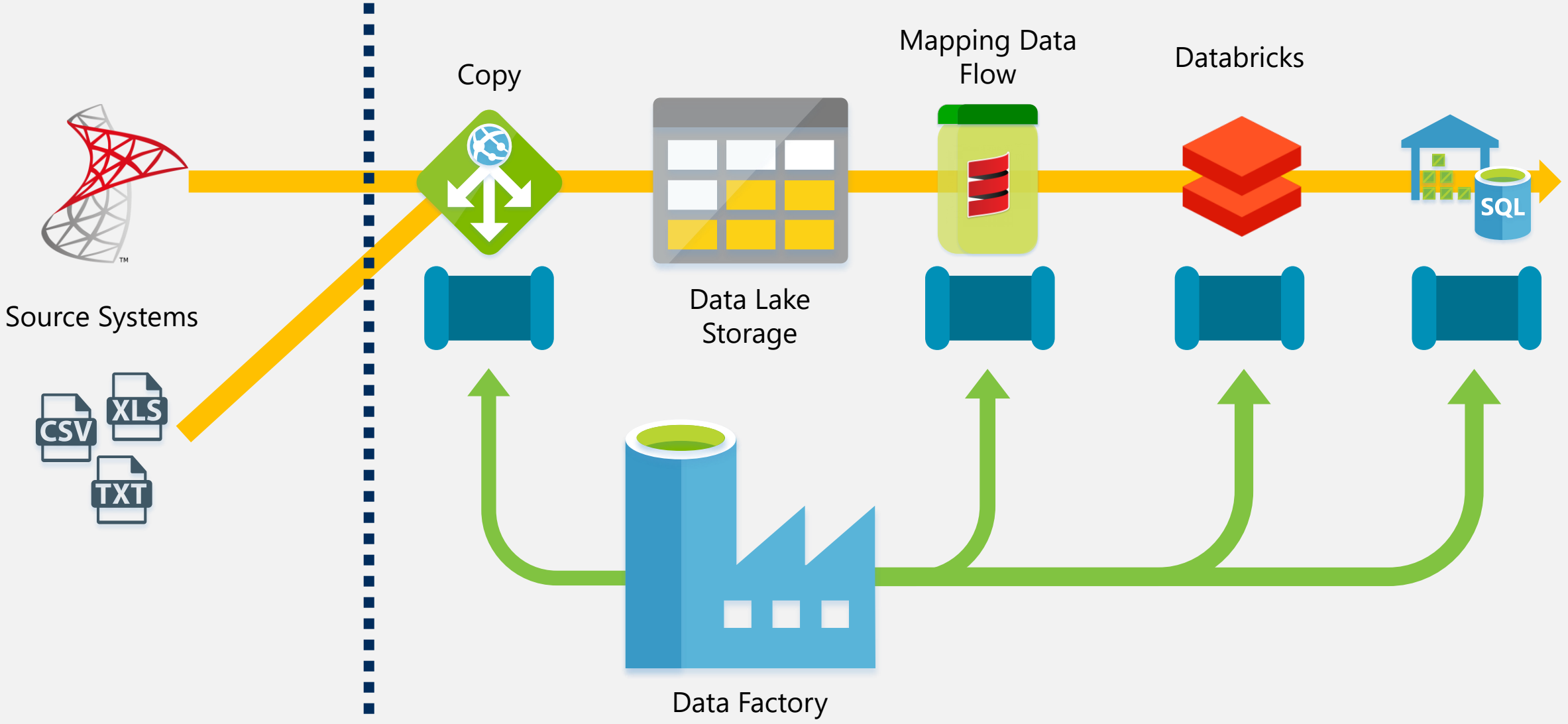
Transformation Method	Graphical Development	Scales Out	Scales Up	Cloud Native Tech
T-SQL (SQLDB)	✗	✗	✓	✗
SSIS	✓	✗	✓	✗
Scala (Databricks)	✗	✓	✓	✓
Mapping Data Flow	✓	✓	✓	✓

adatis

Design Patterns

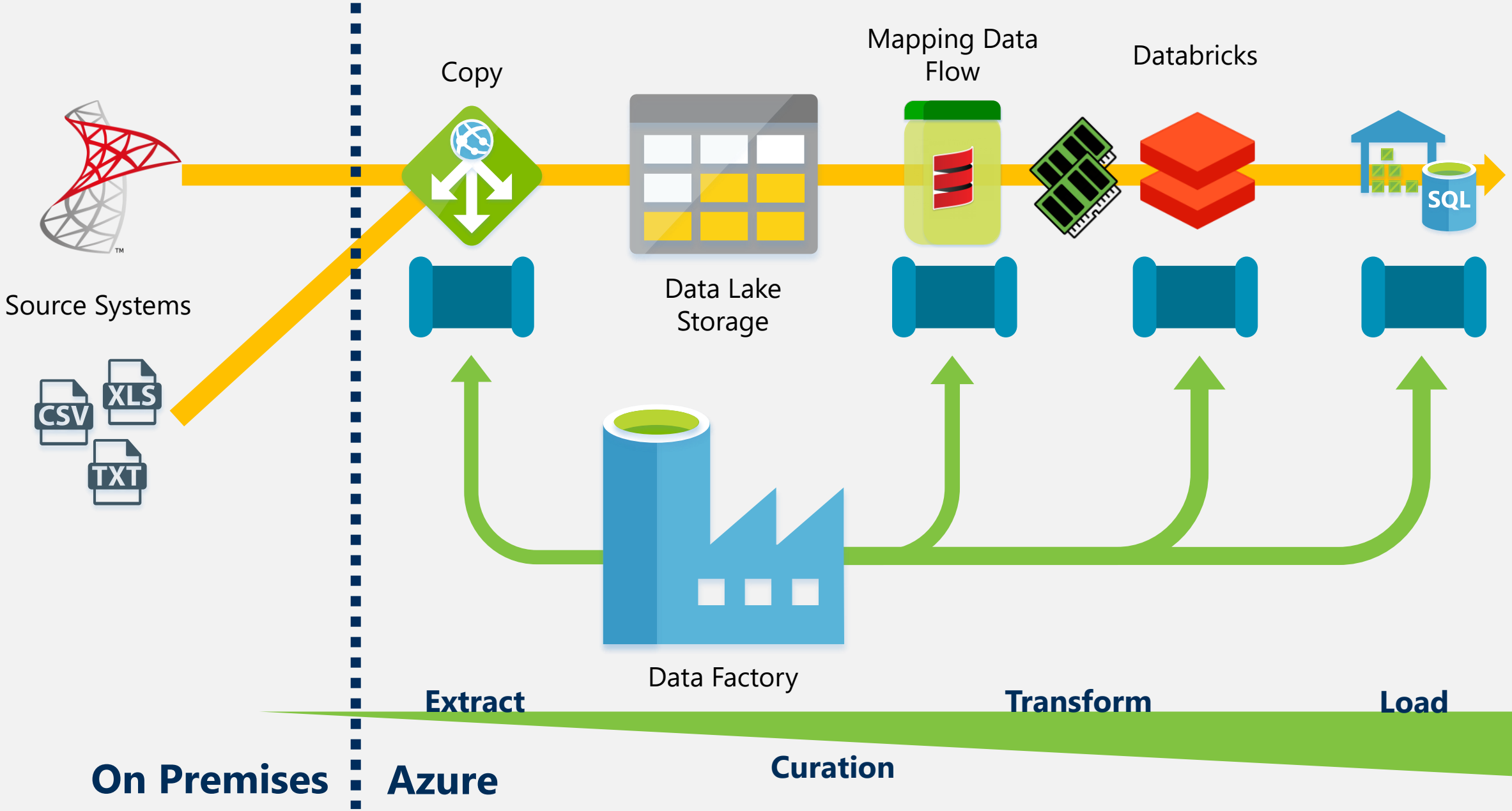
adatis

Mapping Data Flow Future Design Patterns ???

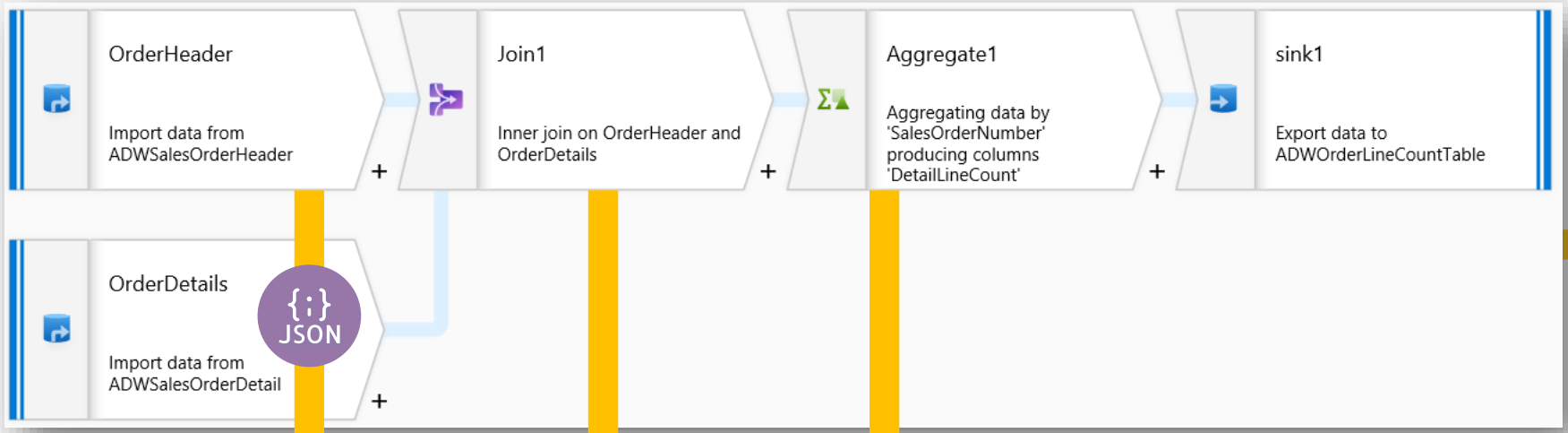


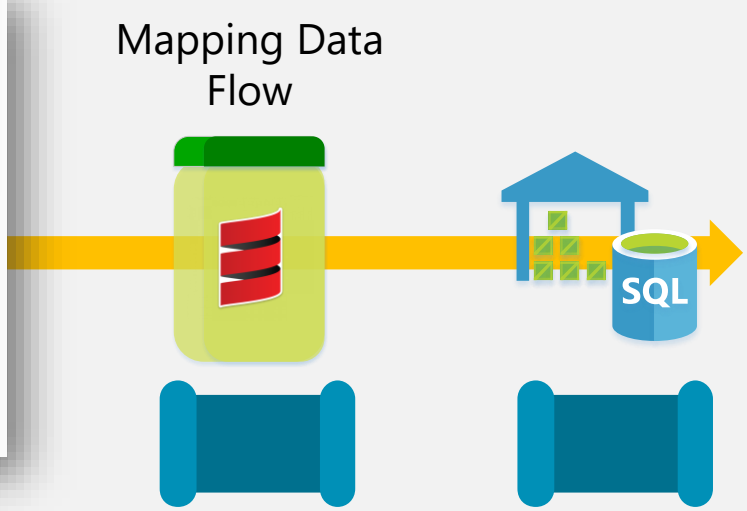
adatis

Mapping Data Flow Future Design Patterns ???

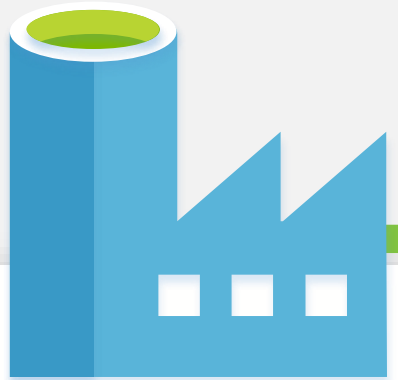


Mapping Data Flow Future Design Patterns ???






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

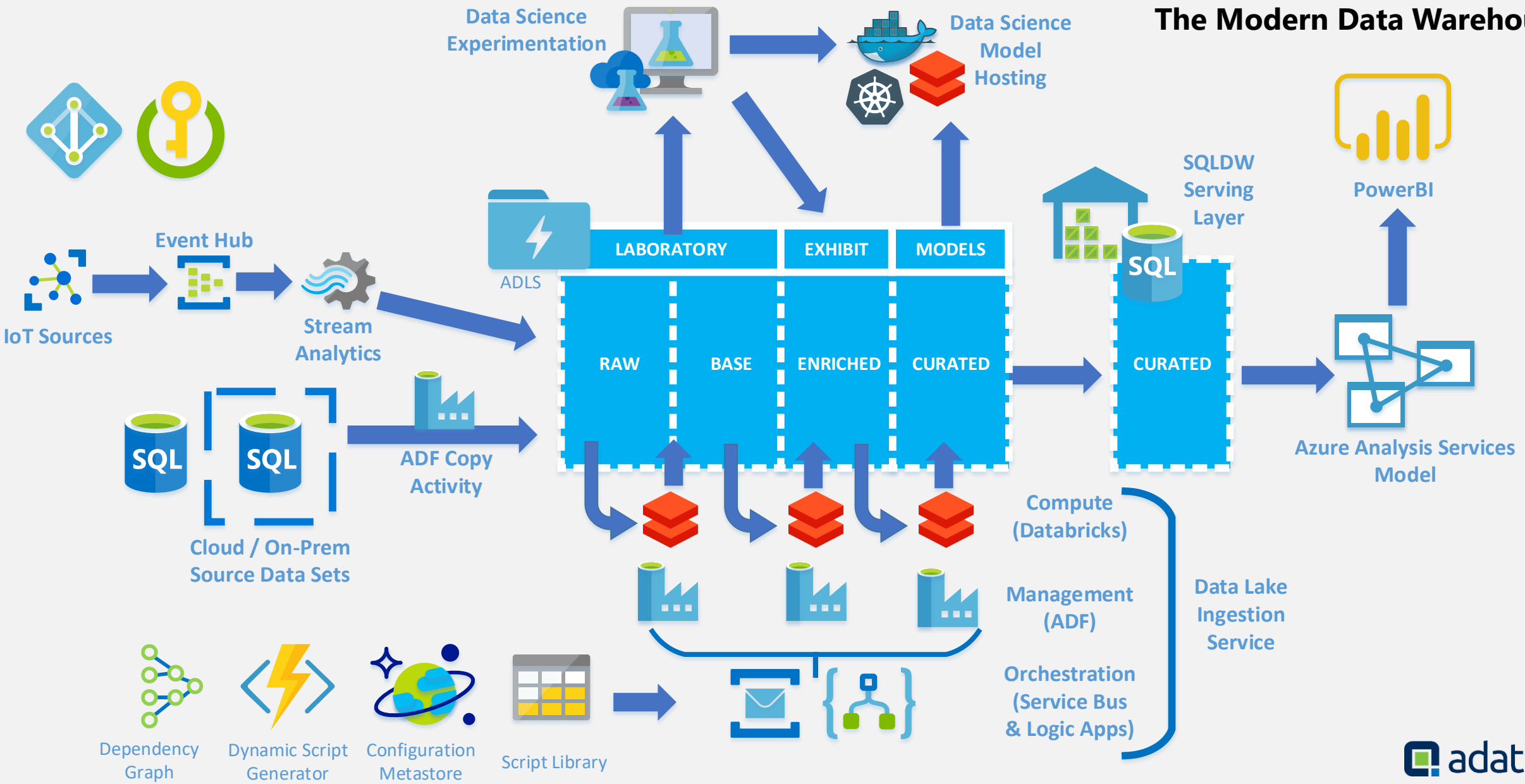


```
"transformations": [
    {
        "name": "Join1",
        "script": "OrderHeader, OrderDetail join(OrderHeader@SalesOrderID == Orde etail@SalesOrderID,\n\tjoinType:'inner',\n\tbroadcast: 'none')~> Join1"
    },
    {
        "name": "Aggregate1",
        "script": "Join1 aggregate(groupBy(SalesOrderNumber),\n\tDetailLineCount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

adatis

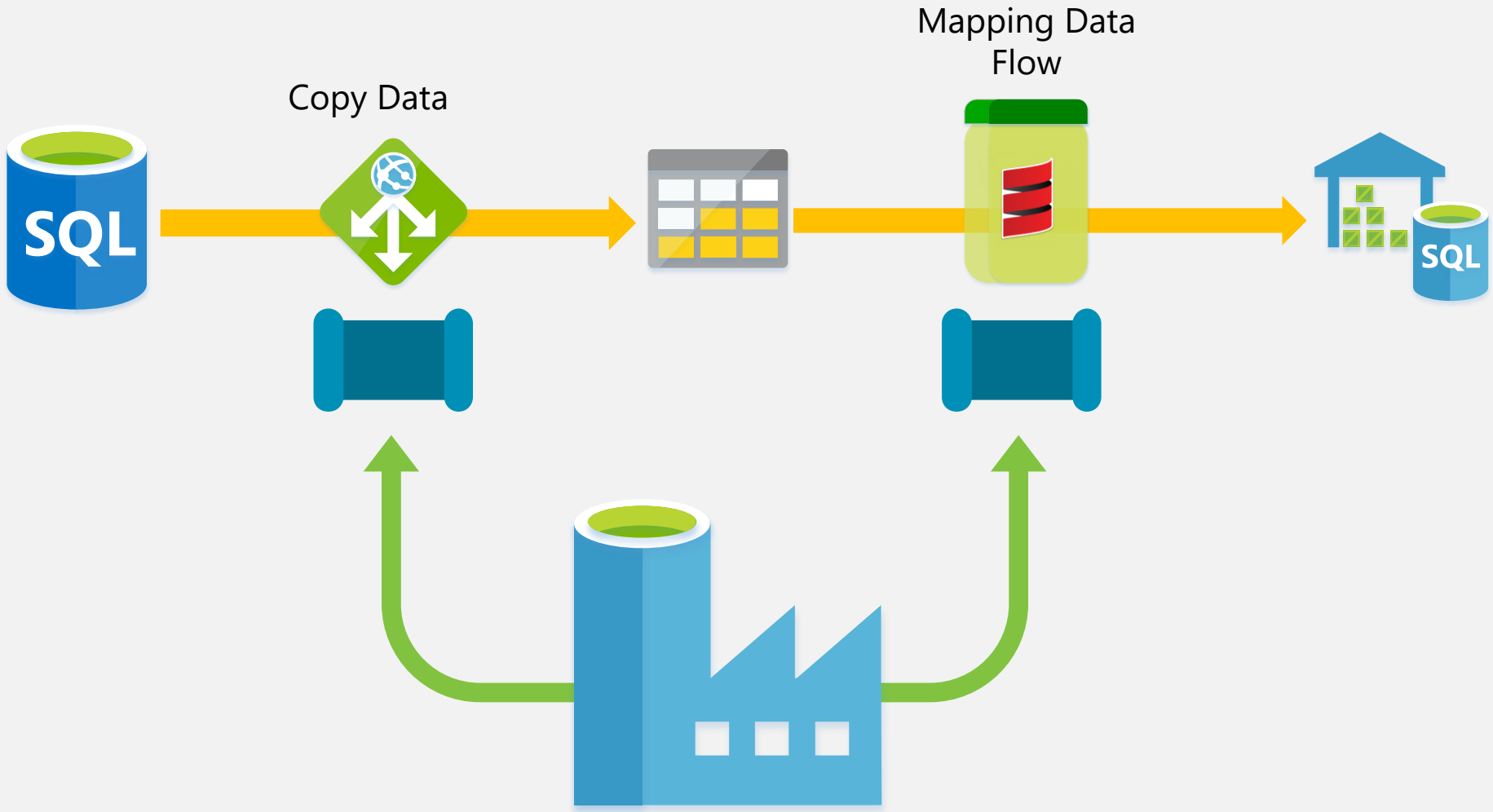
Conclusion

Why use Azure Data Factory?



adatis

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 an easy developer interface for both.

Thanks for Listening

Paul Andrew

@MrPaulAndrew

Microsoft® Most Valuable Professional

adatis

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

PDF Slides