

# Azure Data Factory

A Complete Introduction



Paul Andrew | Principal Consultant & Solution Architect



altius



@MrPaulAndrew



In/MrPaulAndrew



MrPaulAndrew.com

# Azure Data Factory

An ~~Complete~~ Introduction... (to the key things Paul thinks you need to know based on experience.)



Paul Andrew | Principal Consultant & Solution Architect



altius



@MrPaulAndrew

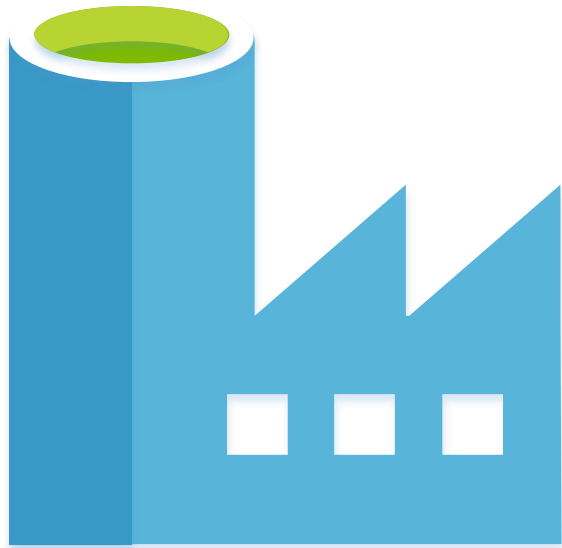


In/MrPaulAndrew



MrPaulAndrew.com

# Azure Data Factory

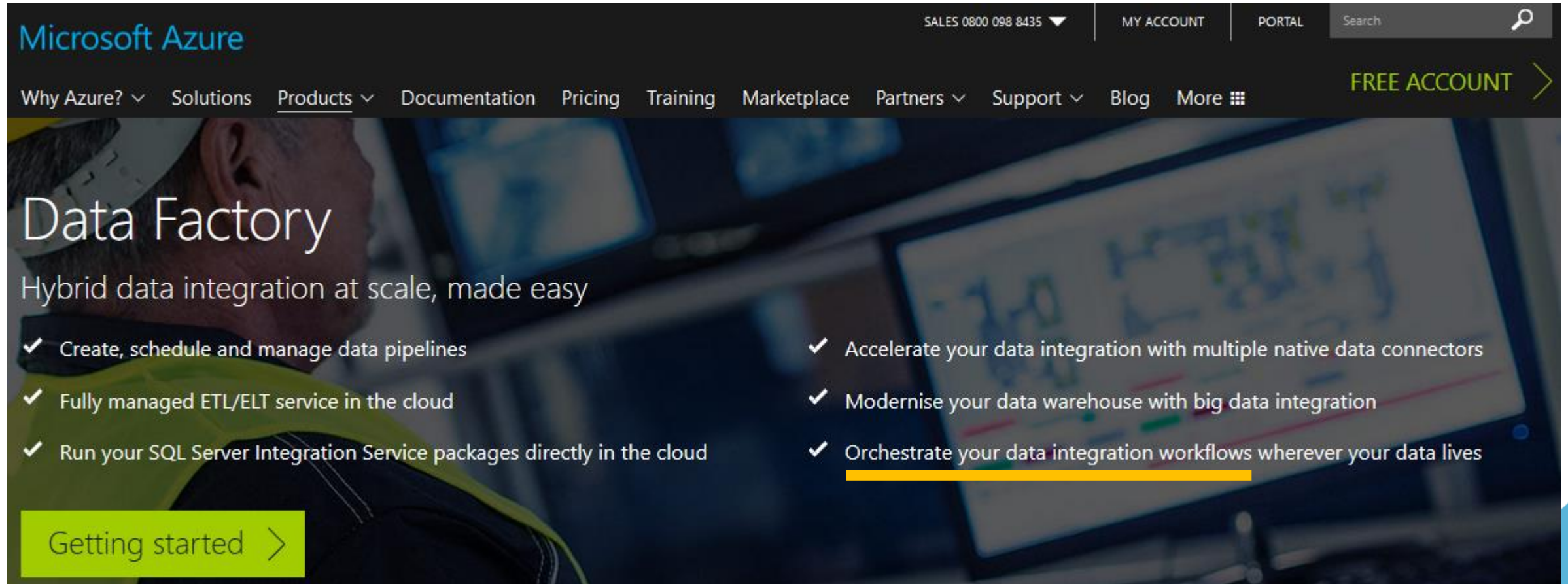


# What is Azure Data Factory?



# What is Azure Data Factory?

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



Microsoft Azure

SALES 0800 098 8435 ▼ | MY ACCOUNT | PORTAL | Search

Why Azure? ▾ Solutions Products ▾ Documentation Pricing Training Marketplace Partners ▾ Support ▾ Blog More ☰

**FREE ACCOUNT** >

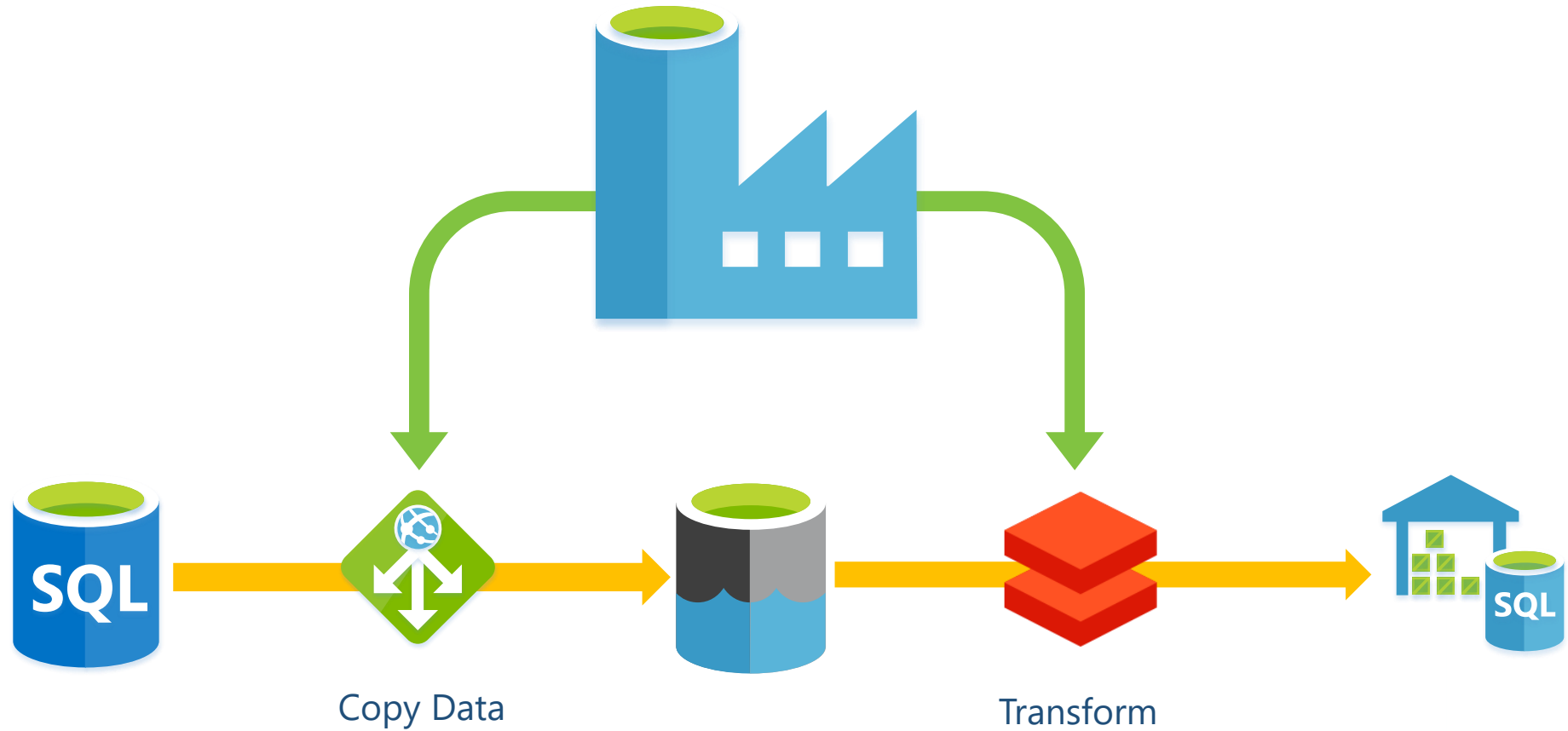
## Data Factory

Hybrid data integration at scale, made easy

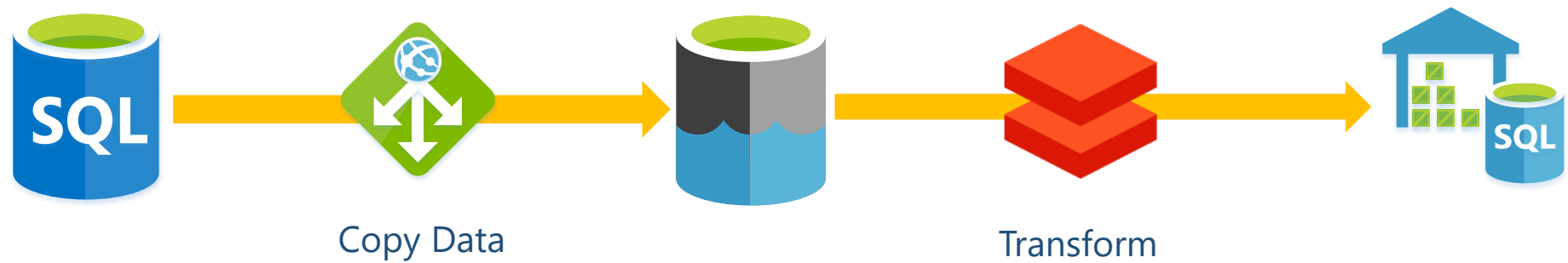
- ✓ Create, schedule and manage data pipelines
- ✓ Fully managed ETL/ELT service in the cloud
- ✓ Run your SQL Server Integration Service packages directly in the cloud
- ✓ Accelerate your data integration with multiple native data connectors
- ✓ Modernise your data warehouse with big data integration
- ✓ Orchestrate your data integration workflows wherever your data lives

Getting started >

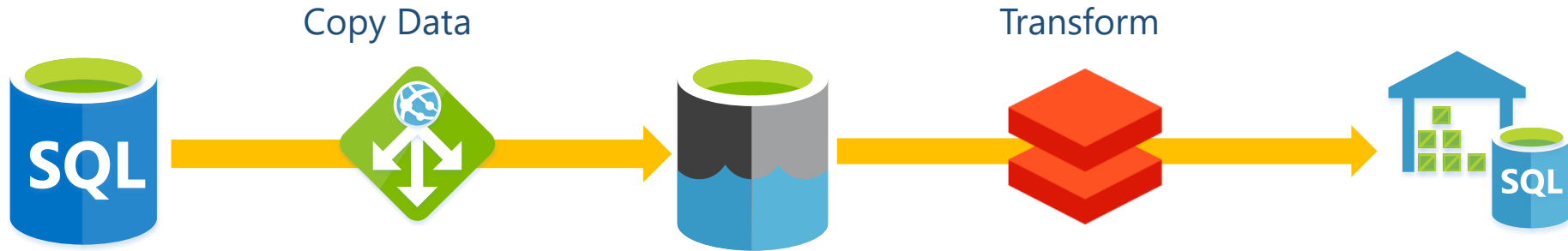
# What is Azure Data Factory?



# What is Azure Data Factory?



# Data Factory Components



1

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





# Data Factory Components



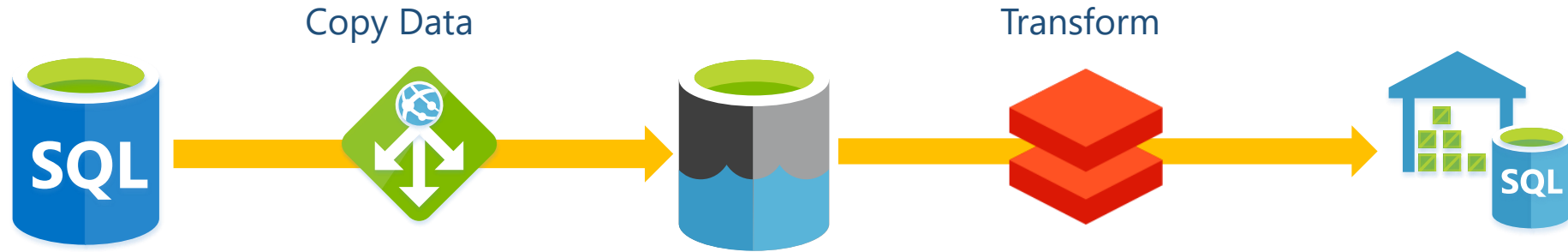
1

## Linked Services –



 Amazon Marketplace Web Service (Preview)	 Amazon Redshift	 Amazon S3	 HDFS	 HTTP	 Hive	 Netezza	 ODBC	 OData	 Azure Batch	 Azure Data Lake Analytics	 Azure Databricks
 Apache Impala (Preview)	 Azure Blob Storage	 Azure Cosmos DB (MongoDB API)	 HubSpot (Preview)	 Informix	 Jira (Preview)	 Office 365 (Preview)	 Oracle	 Oracle Eloqua (Preview)	 Azure Function	 Azure HDInsight	 Azure ML
 Azure Cosmos DB (SQL API)	 Azure Data Explorer (Kusto)	 Azure Data Lake Storage Gen1	 Magento (Preview)	 MariaDB	 Marketo (Preview)	 Oracle Responsys (Preview)	 Oracle Service Cloud (Preview)	 Paypal (Preview)	 ServiceNow	 Shopify (Preview)	 Spark
 Azure Data Lake Storage Gen2 (Preview)	 Azure Database for MariaDB	 Azure Database for MySQL	 Microsoft Access	 MongoDB	 MySQL	 Phoenix	 PostgreSQL	 Presto (Preview)	 Square (Preview)	 Sybase	 Teradata
 Azure Database for PostgreSQL	 Azure File Storage	 Azure Key Vault	 DB2	 Drill (Preview)	 Dynamics 365	 QuickBooks (Preview)	 REST	 SAP BW Open Hub	 Vertica	 Web Table	 Xero (Preview)
 Azure SQL Data Warehouse	 Azure SQL Database	 Azure SQL Database Managed Instance	 Dynamics AX (Preview)	 Dynamics CRM	 FTP	 SAP BW via MDX	 SAP Cloud For Customer	 SAP ECC	 Zoho (Preview)		
 Azure Search	 Azure Table Storage	 Cassandra	 File System	 Google AdWords (Preview)	 Google BigQuery	 SAP HANA	 SFTP	 SQL Server			
 Common Data Service for Apps	 Concur (Preview)	 Couchbase (Preview)	 Google Cloud Storage (S3 API)	 Greenplum	 HBase	 Salesforce	 Salesforce Marketing Cloud (Preview)	 Salesforce Service Cloud			

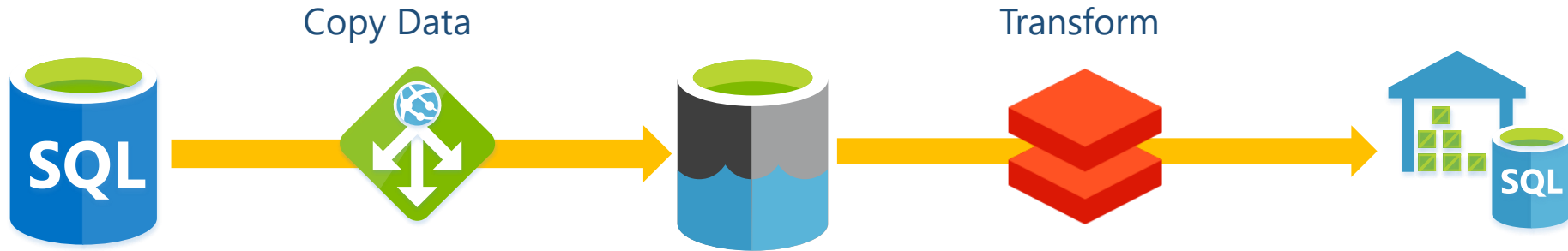
# Data Factory Components



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?

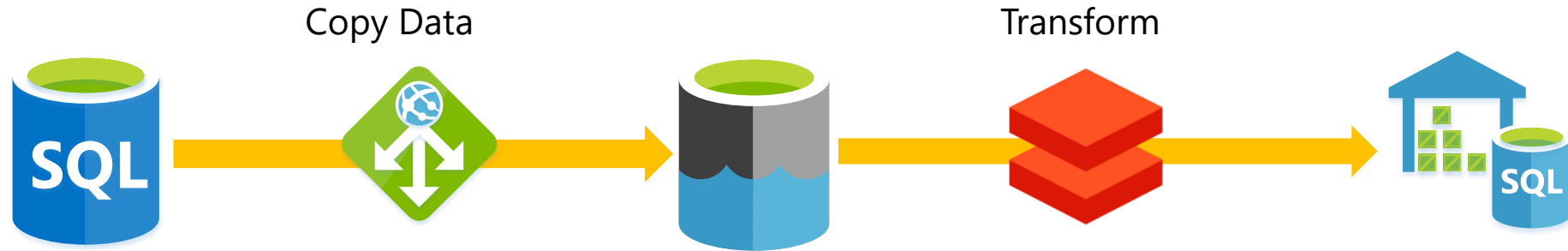


dbo.DimCustomer



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

# Data Factory Components



1

## Linked Services

2

## Data Sets

3

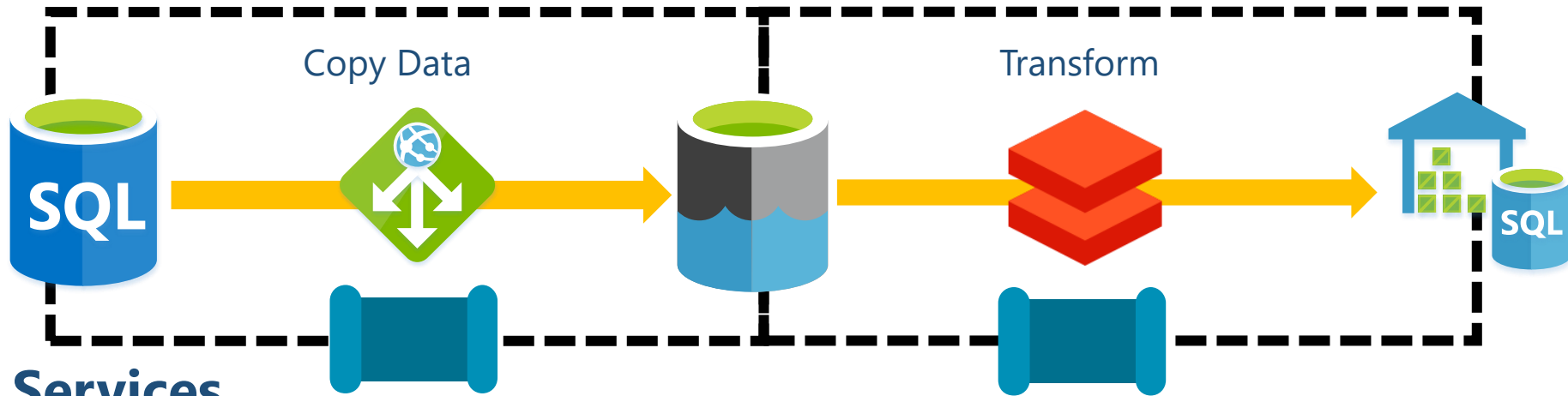
**Activities** – What do we want to happen?  
With what conditions?



### 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?



Sequence Container

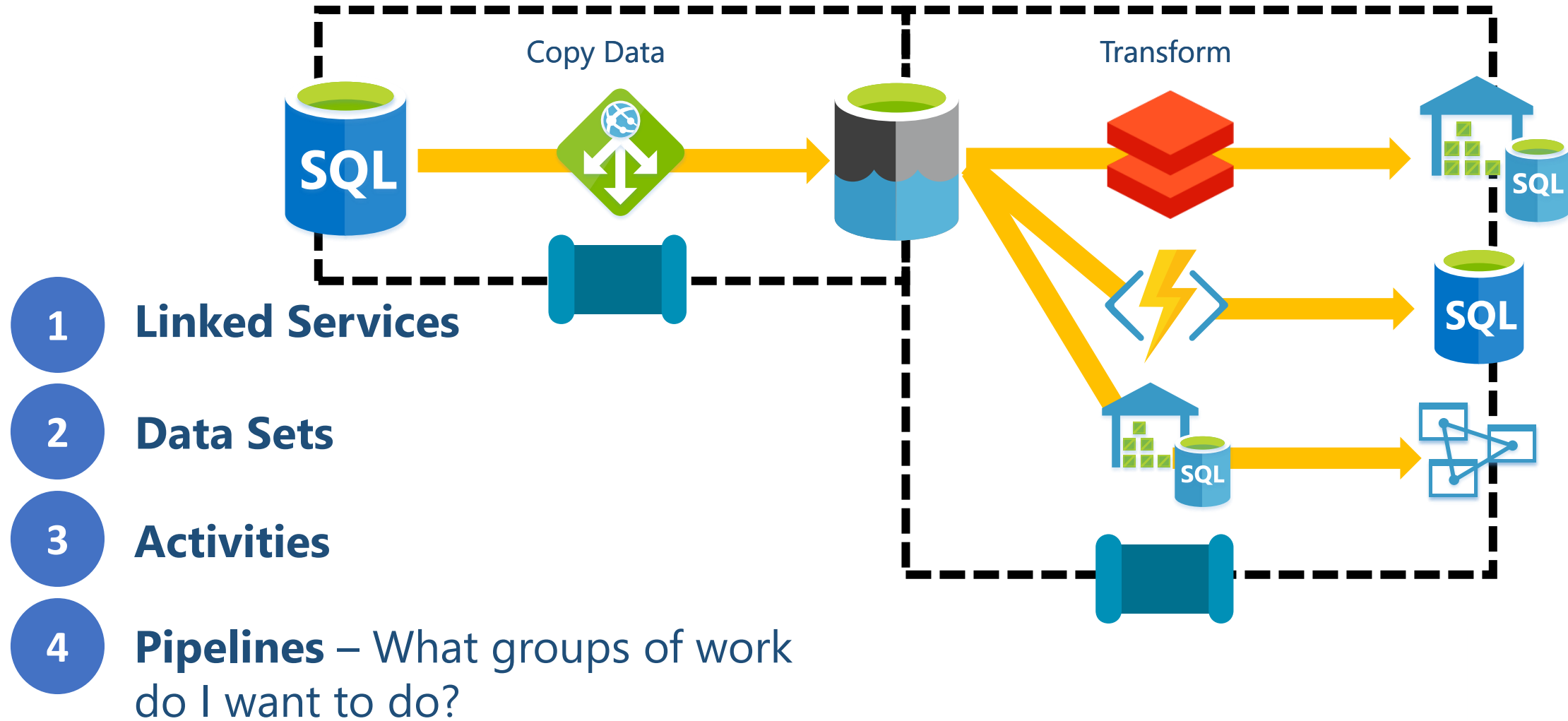


Execute Package Task

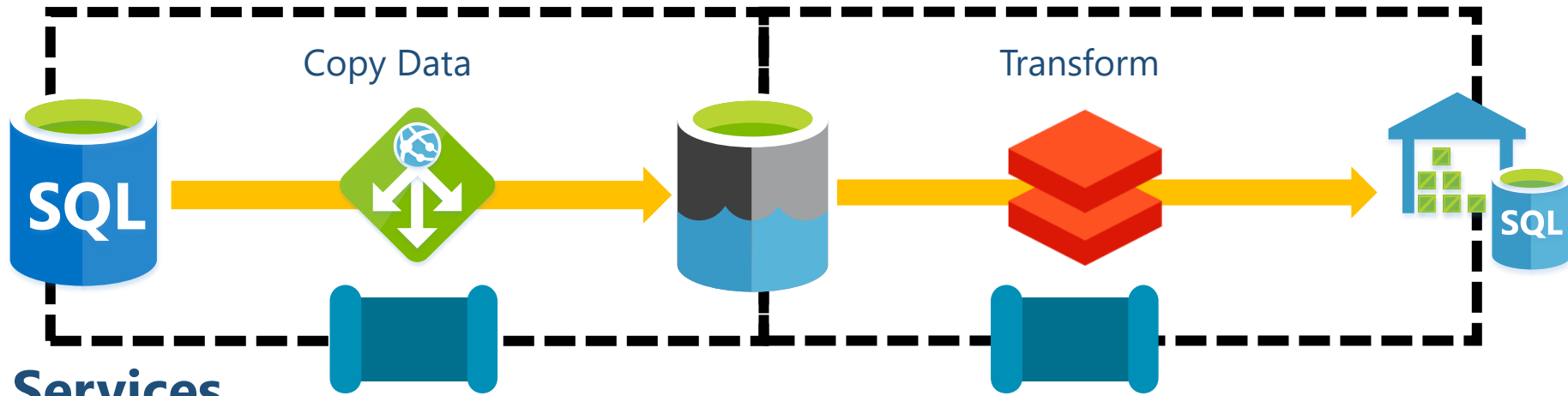


Execute Pipeline Activity

# Data Factory Components



# Data Factory Components



1

**Linked Services**

2

**Data Sets**

3

**Activities**

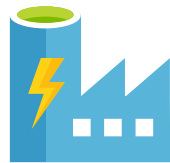
4

**Pipelines**

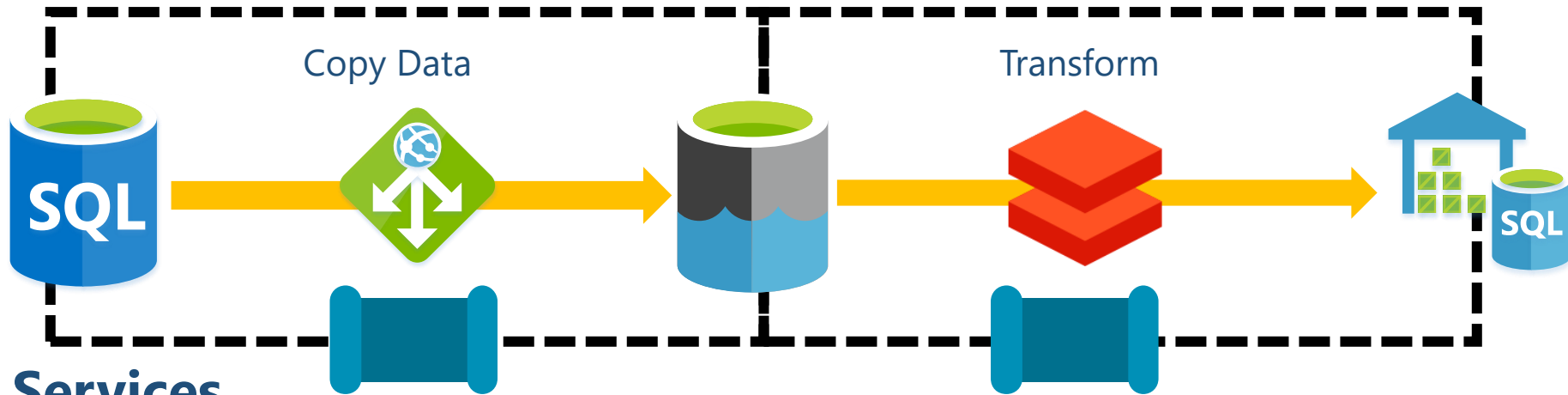
5

**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

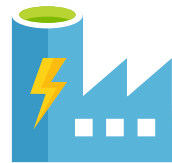
Activities

4

Pipelines

5

Triggers



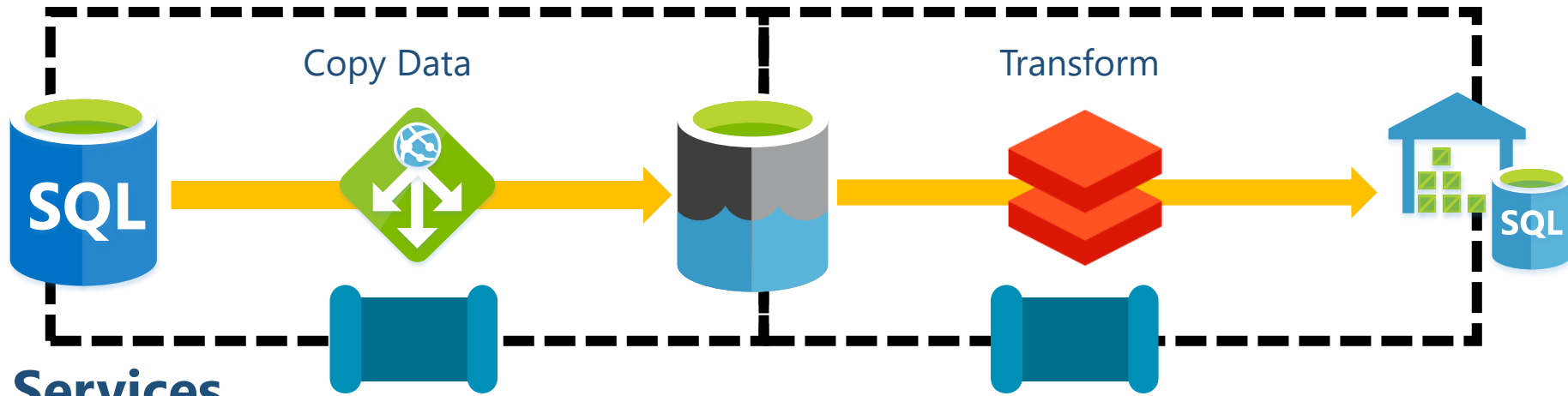
- **Manual**
- Tumbling Windows
- Scheduled
- Blob File Events
- Logic App Calls



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



# Data Factory Components



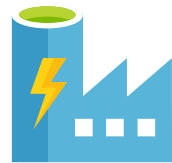
1 Linked Services

2 Data Sets

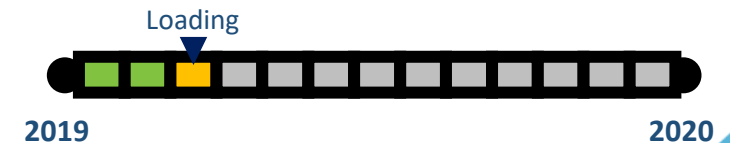
3 Activities

4 Pipelines

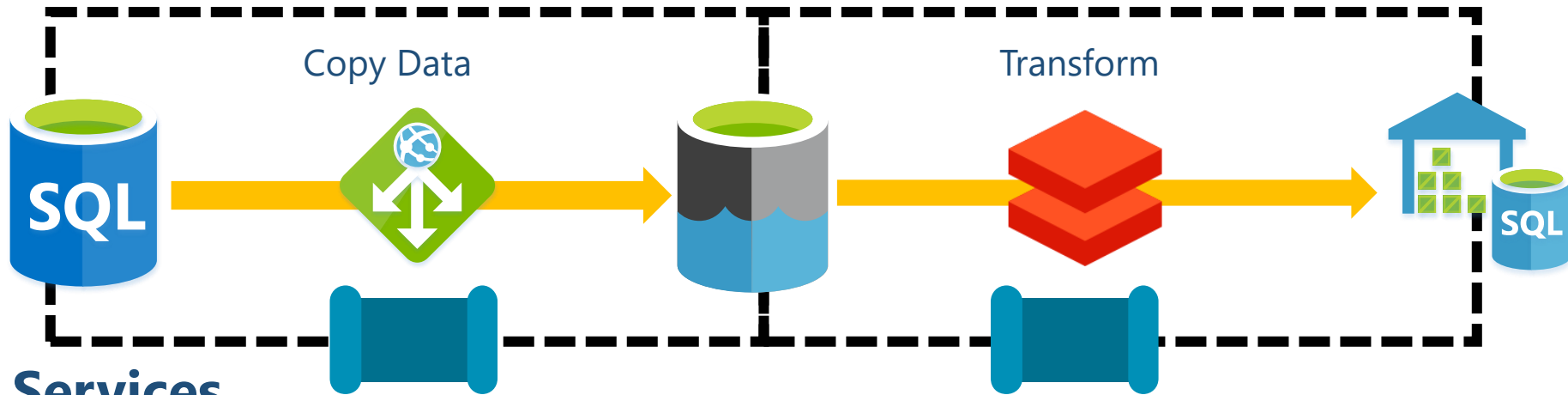
5 Triggers



- 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

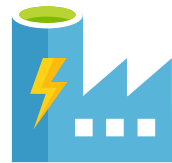
Activities

4

Pipelines

5

Triggers

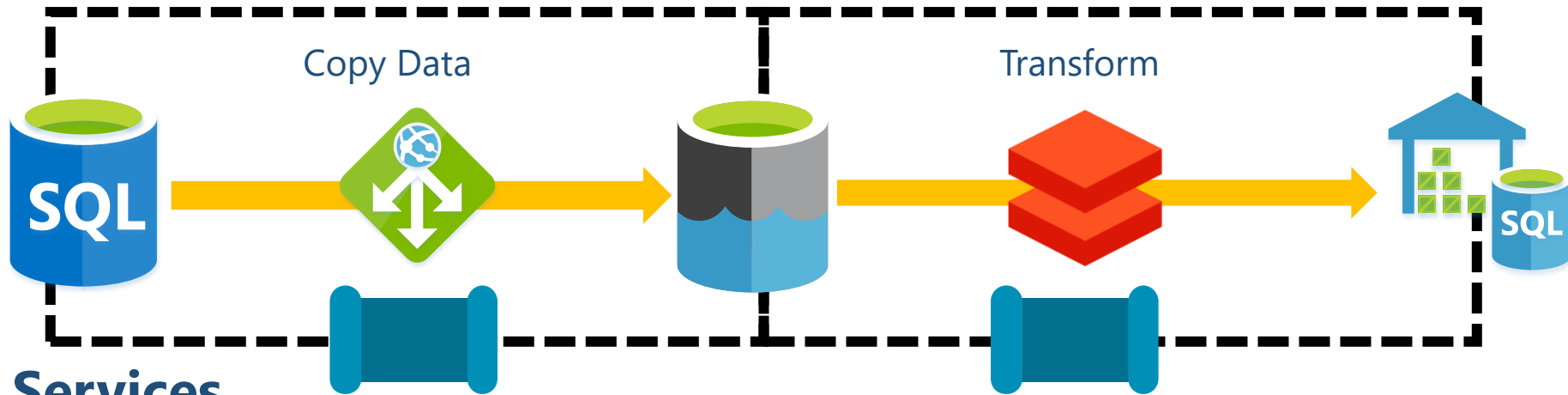


- Manual via UI
- Tumbling Windows
- **Scheduled**
- Blob File Events
- Logic App Calls



- Every 1 minute.
- UTC

# Data Factory Components



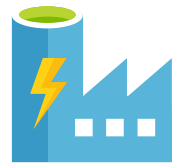
1 Linked Services

2 Data Sets

3 Activities

4 Pipelines

5 Triggers

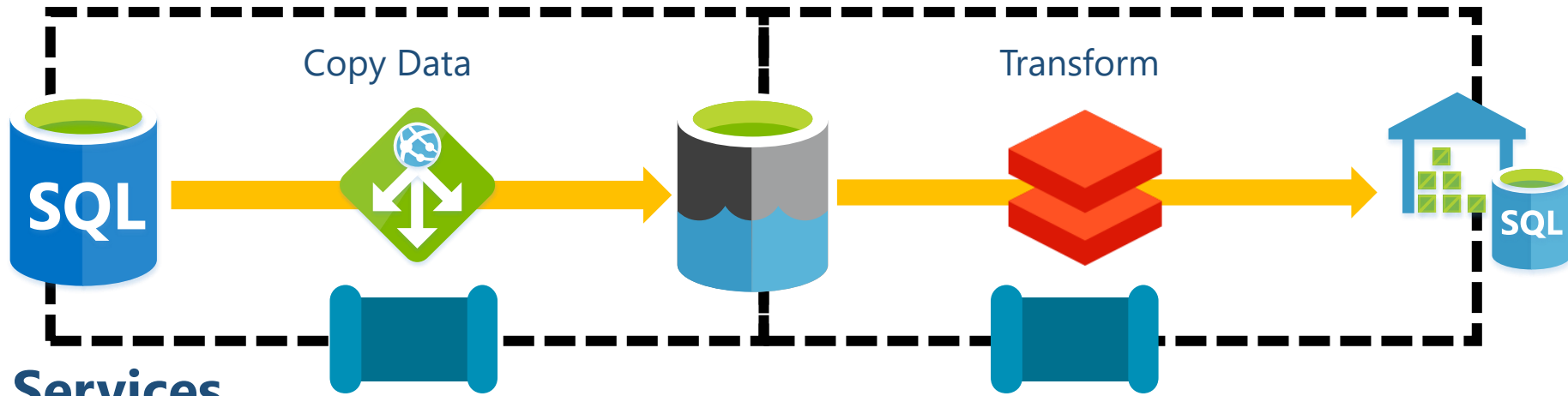


- Manual via UI
- Tumbling Windows
- Scheduled
- **Blob File Events**
- Logic App Calls



{Path} Created  
{Path} Deleted

# Data Factory Components



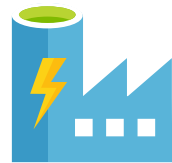
1 Linked Services

2 Data Sets

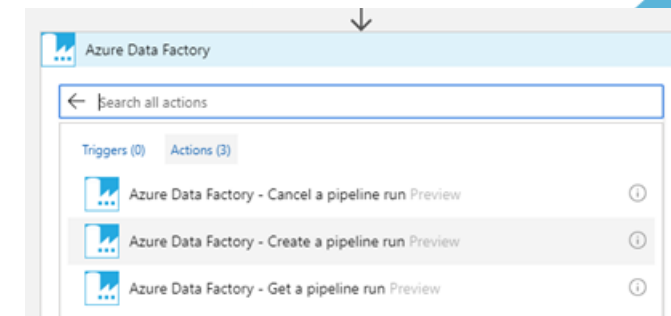
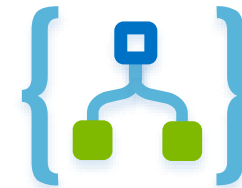
3 Activities

4 Pipelines

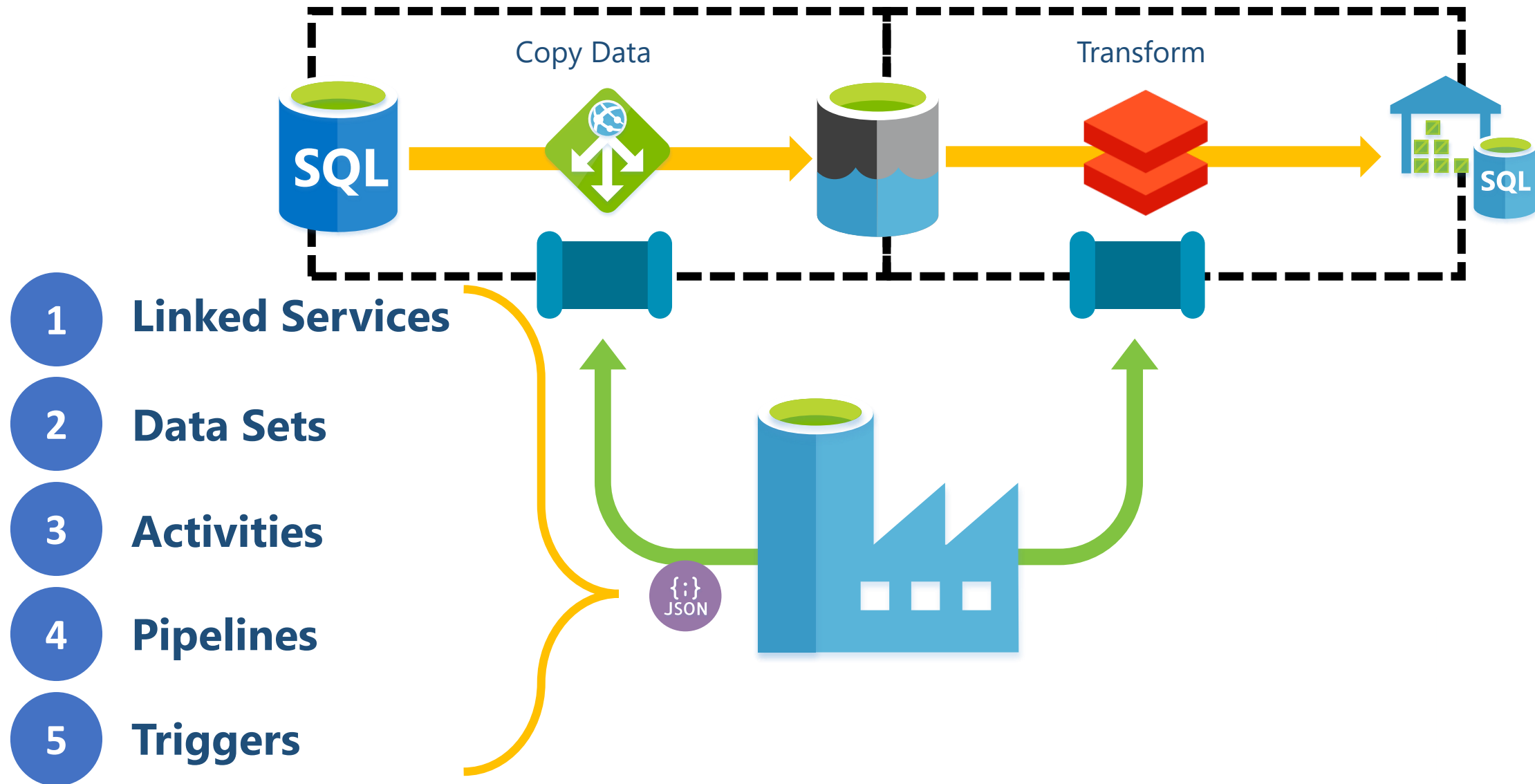
5 Triggers



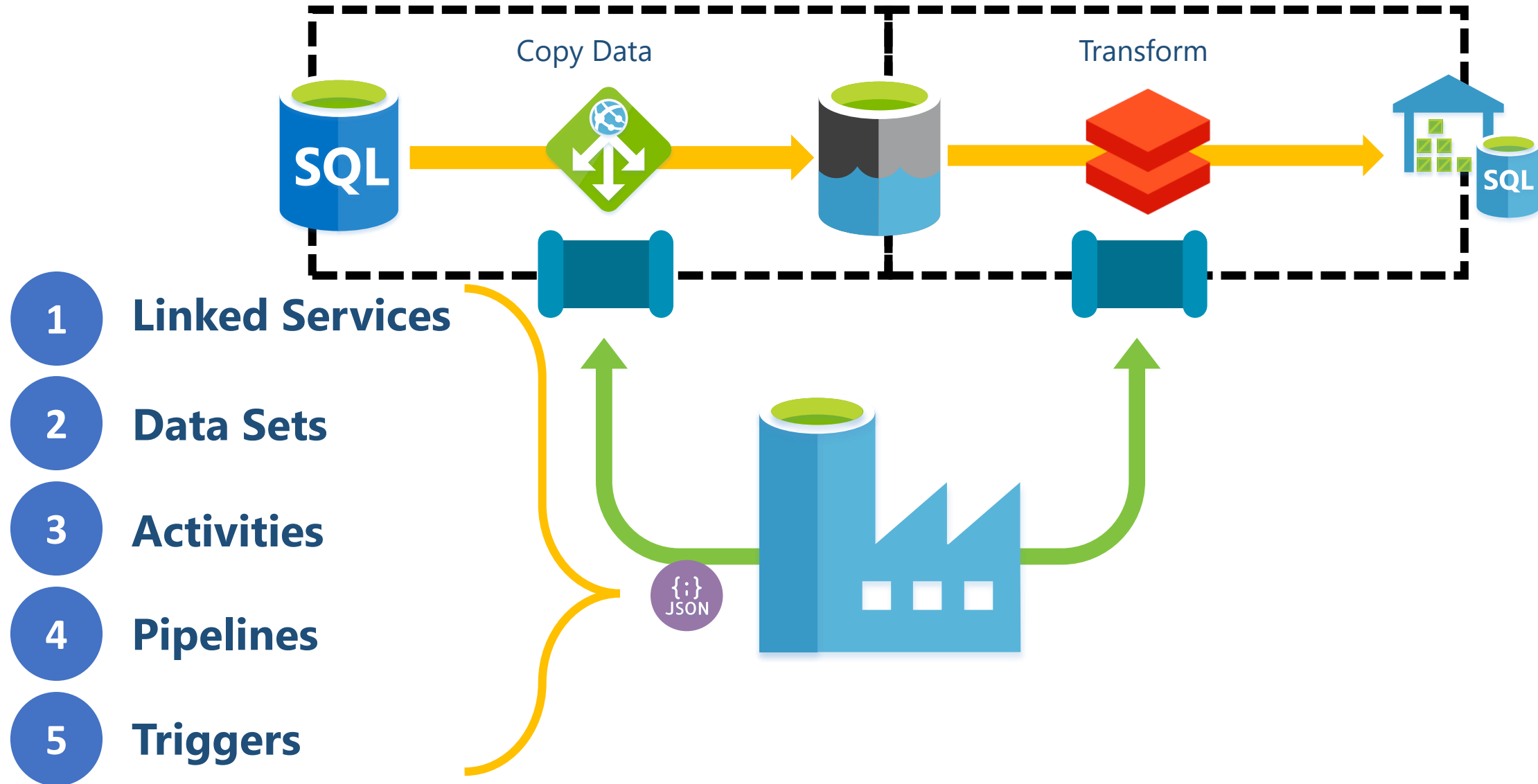
- Manual via UI
- Tumbling Windows
- Scheduled
- Blob File Events
- **Logic App Calls**



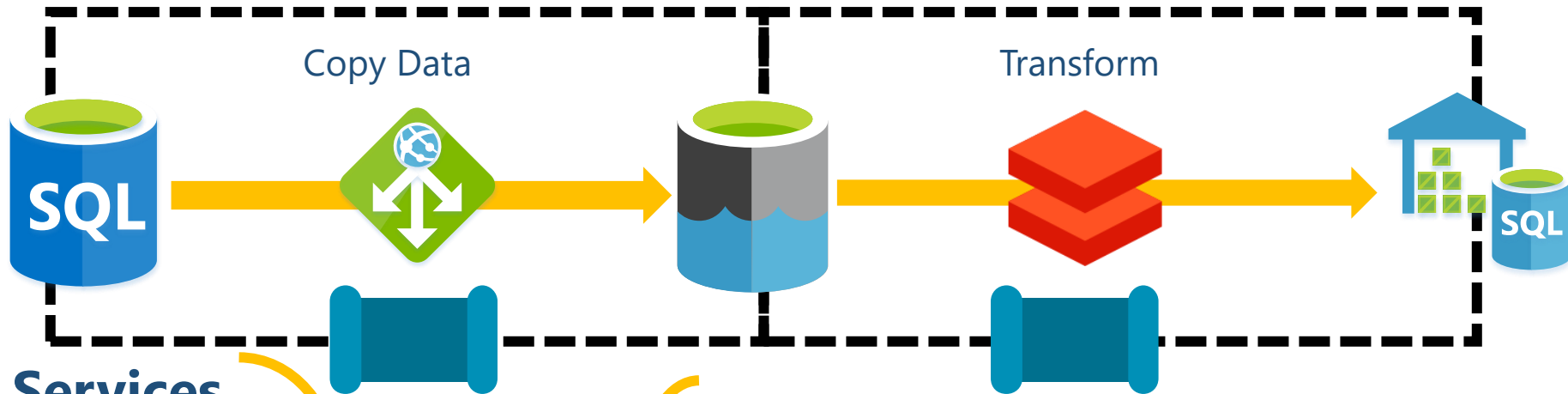
# Data Factory Components



# Data Factory Control Flow Components



# Data Factory Control Flow Components



1 ✓ **Linked Services**

2 ✓ **Data Sets**

3 ✓ **Activities**

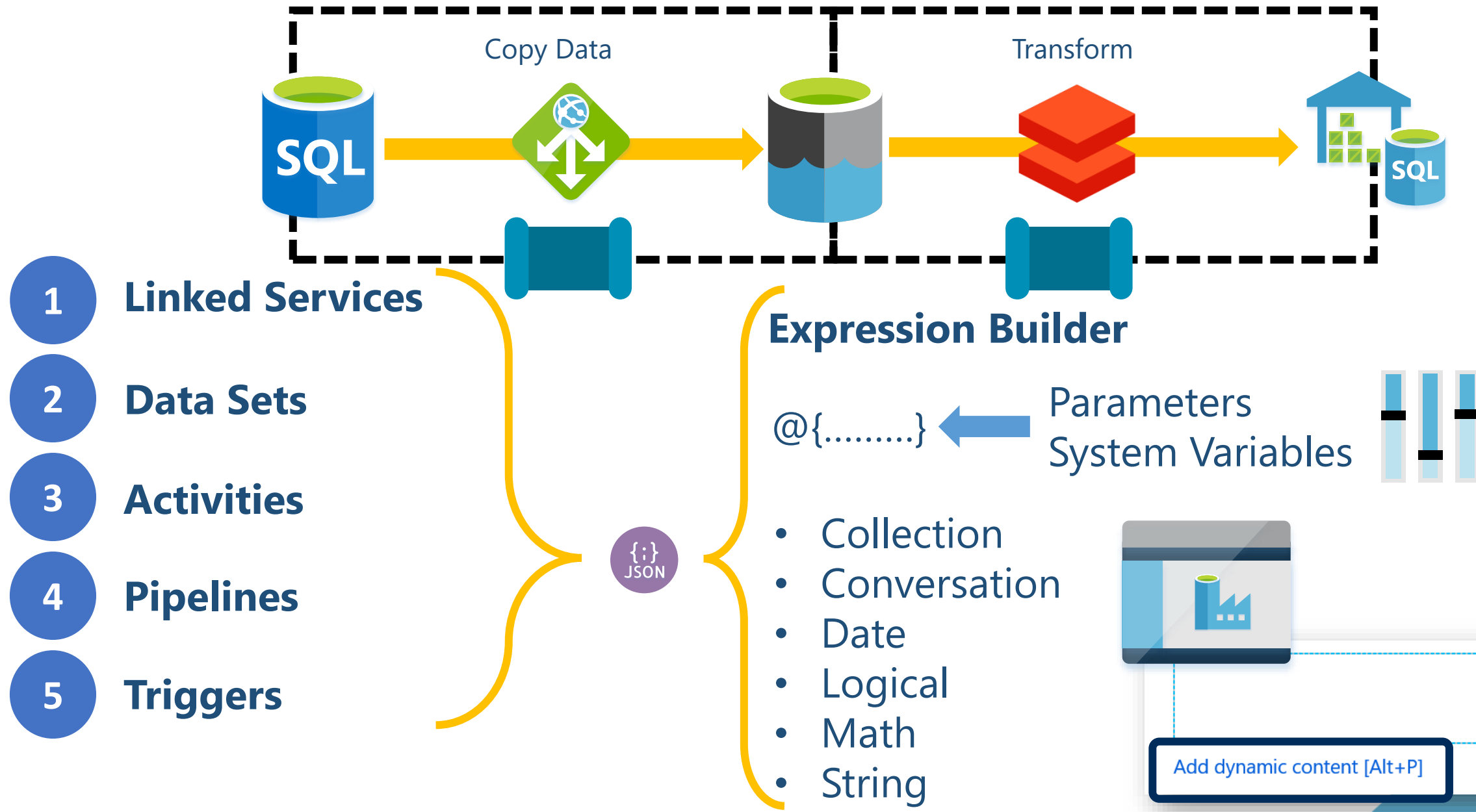
4 ✓ **Pipelines**

5 ✗ **Triggers**

{:}  
JSON

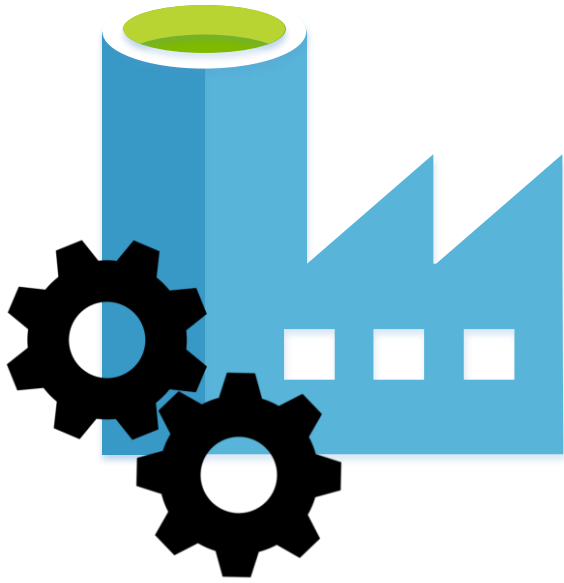
```
{
  "name": "GenericSQLDB",
  "type": "Microsoft.DataFactory/factories/linkedservices",
  "properties": {
    "parameters": {
      "ServerInstance": {
        "type": "String"
      },
      "DatabaseName": {
        "type": "String"
      },
      "SQLUser": {
        "type": "String"
      },
      "SQLPassword": {
        "type": "String"
      }
    },
    "type": "AzureSqlDatabase",
    "typeProperties": {
      "connectionString": "Integrated Security=False;Encrypt=True;ConnectionTimeout=30;
Data Source=@{linkedService().ServerInstance};
InitialCatalog=@{linkedService().DatabaseName};
UserID=@{linkedService().SQLUser};
Password=@{linkedService().SQLPassword}"
    }
  }
}
```

# Data Factory Control Flow Components





# Integration Runtimes



# Integration Runtimes



1

**Azure**  
Integration Runtime

Data Movements

Activity  
Orchestration







2

**SSIS**  
Integration Runtime

SSIS Package  
Execution







3

**Self Hosted**  
Integration Runtime

Gateway Access

Activity  
Orchestration





# Integration Runtimes







1

**Azure**  
Integration Runtime

Movement Hours

Activity Orchestration







2

**SSIS**  
Integration Runtime

SSIS Package Execution







3

**Self Hosted**  
Integration Runtime

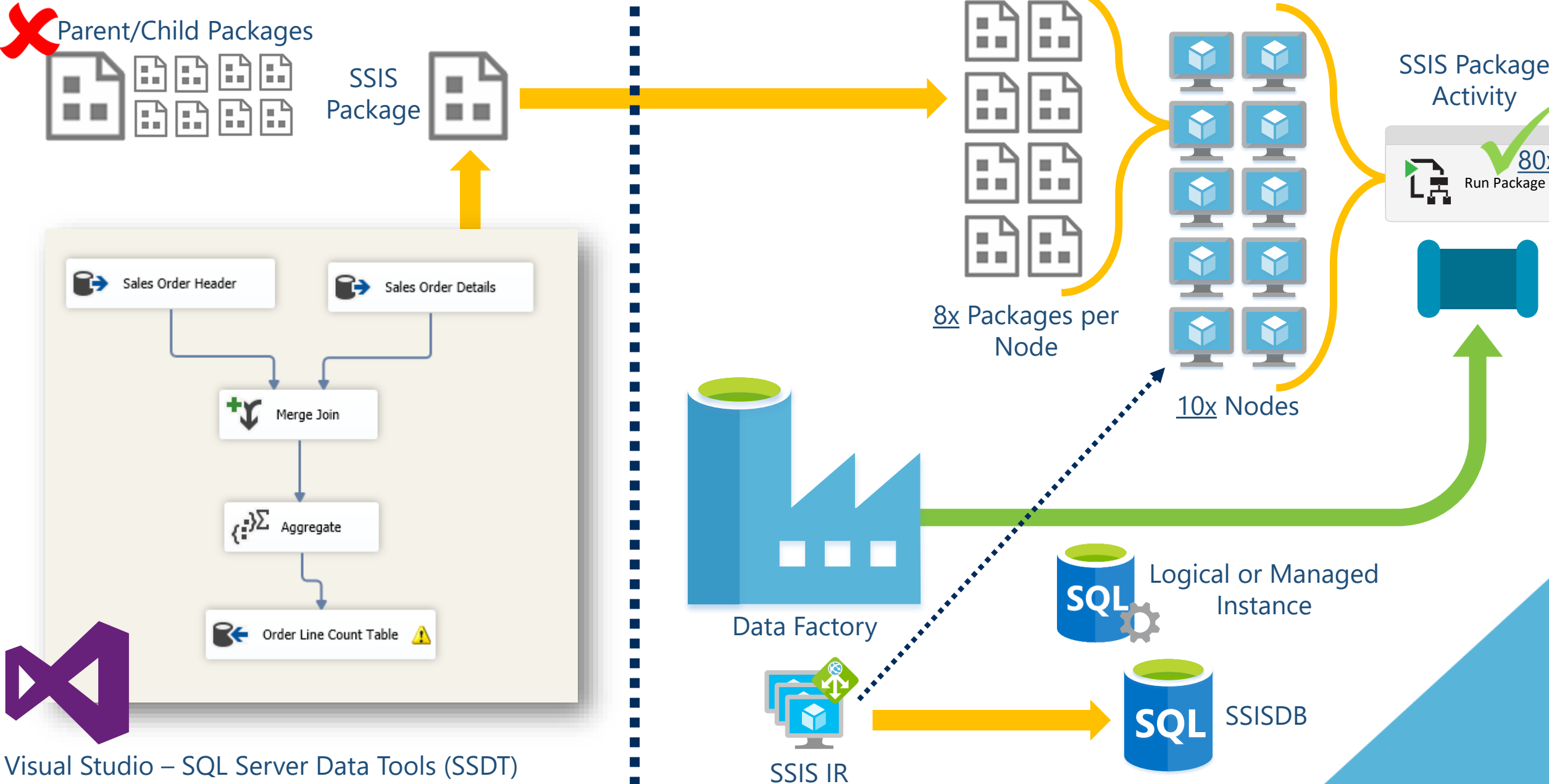
Gateway Access

Activity Orchestration





# Running SSIS Packages in Azure Data Factory



# Integration Runtimes



1

**Azure**  
Integration Runtime

Movement Hours



Activity  
Orchestration







2

**SSIS**  
Integration Runtime

SSIS Package  
Execution









3

**Self Hosted**  
Integration Runtime

Gateway Access



Activity  
Orchestration





# Integration Runtimes







1

**Azure**  
Integration Runtime

Movement Hours

Activity Orchestration







2

**SSIS**  
Integration Runtime

SSIS Package Execution







3

**Self Hosted**  
Integration Runtime

Gateway Access

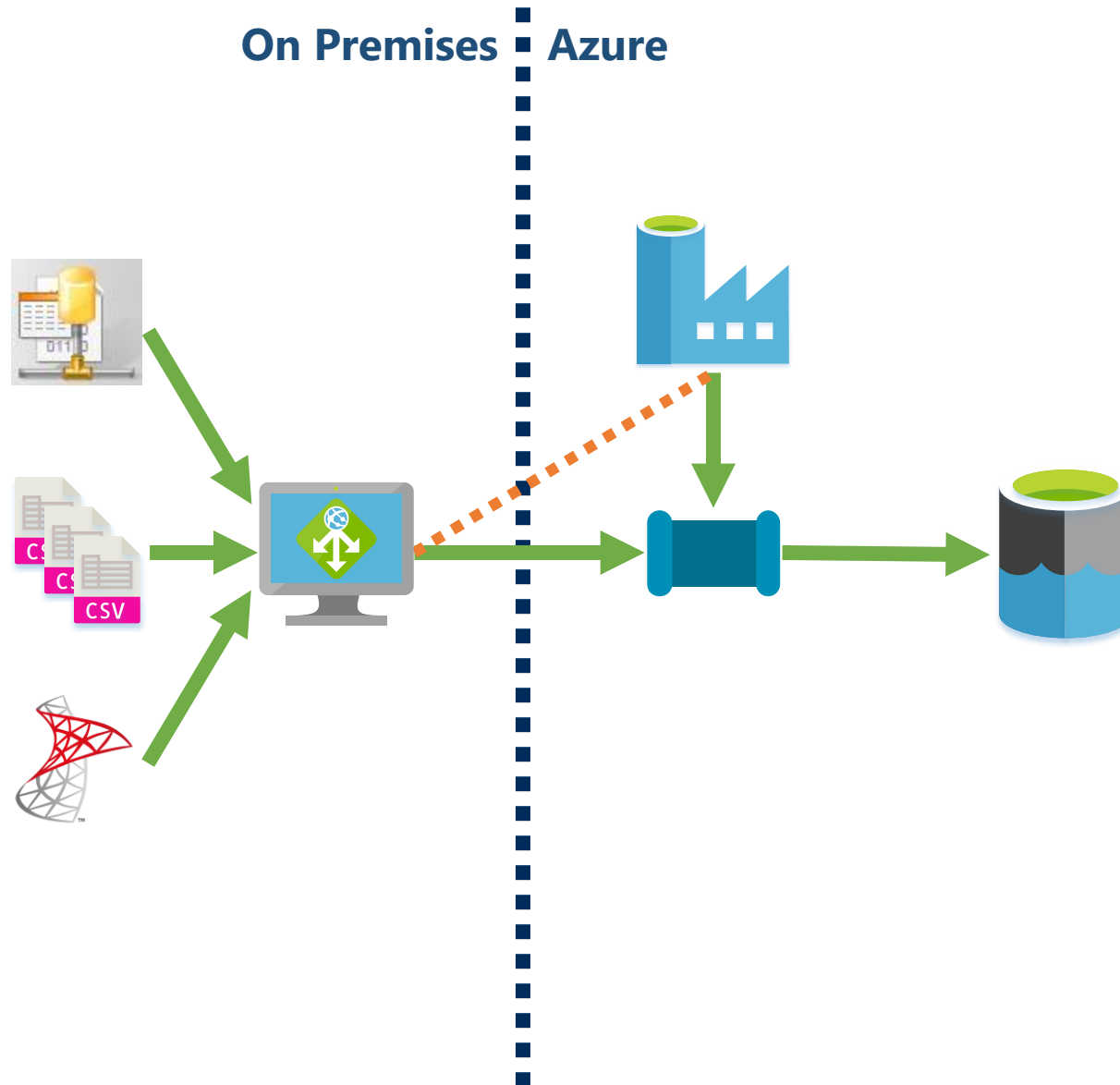
Activity Orchestration



Virtual Machine



# Single Hosted IR



### Integration runtime setup

**Settings** Nodes Auto update Sharing

Install integration runtime on Windows machine or add further nodes using the Authentication Key.

Name: integrationRuntime2

Option 1: Express setup  
[Click here to launch the express setup for this computer](#)

Option 2: Manual setup  
Step 1: [Download and install integration runtime](#)  
Step 2: Use this key to register your integration runtime

NAME	AUTHENTICATION KEY		
Key1	IR@764c300b-fb33-4d0d-b662-8714248132ec@PaulsFunFactoryV2@eu@		
Key2	IR@764c300b-fb33-4d0d-b662-8714248132ec@PaulsFunFactoryV2@eu@		

### Microsoft Integration Runtime Configuration Manager

Home Settings Diagnostics Update Help

✓ Self-hosted node is connected to the cloud service

Data Factory: PaulsFunFactoryV2  
Integration Runtime: WorkLaptop  
Node: AUK-PA394

Stop Service

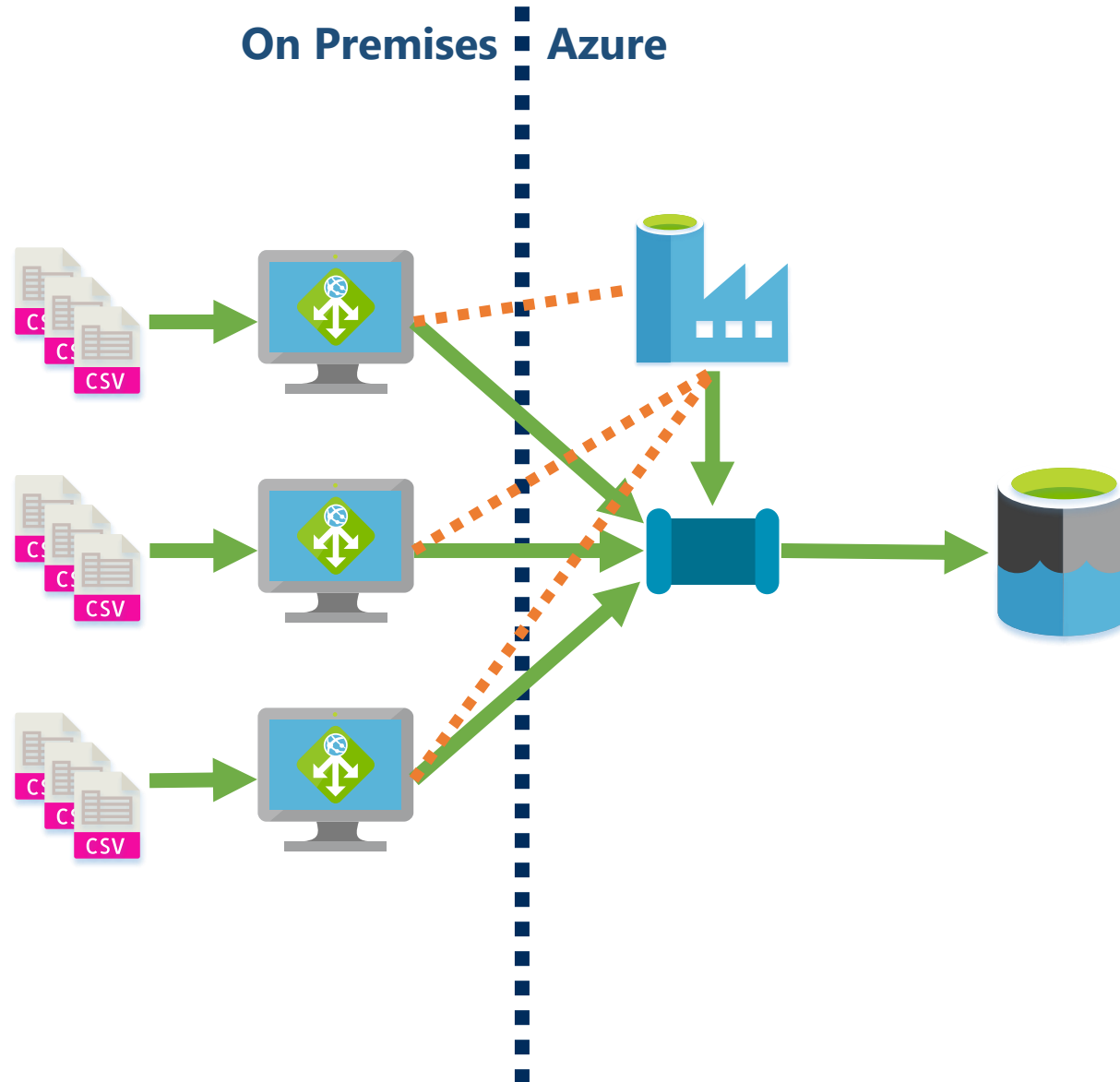
Data Source Credential ⓘ

Credential store: On-premises  
Credential status: In sync  
Last backup time: N/A

Generate Backup Import Backup

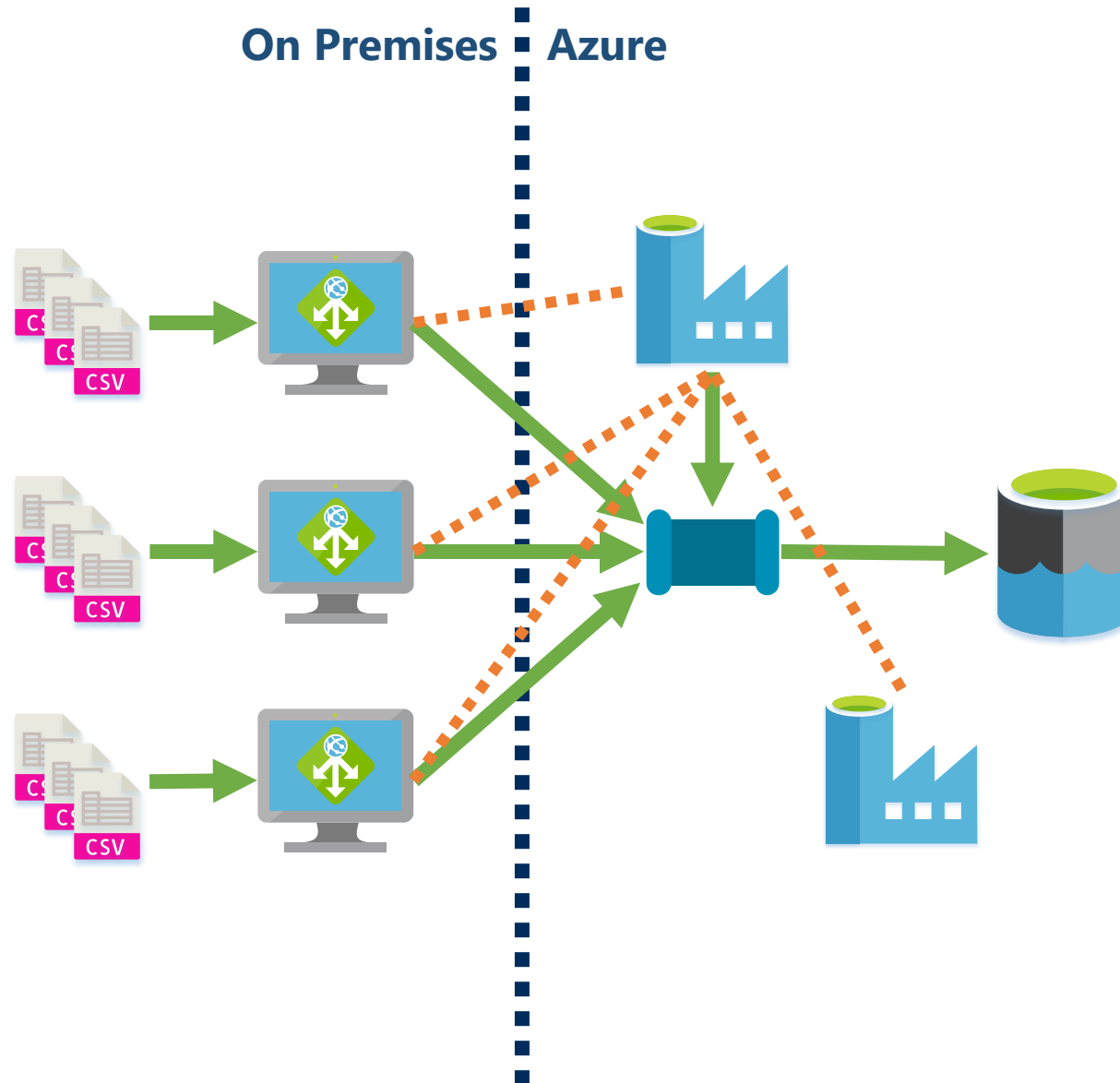
✓ Connected to the cloud service (Data Factory V2)

# Multiple Hosted IR's (Failover & Load Balancing)

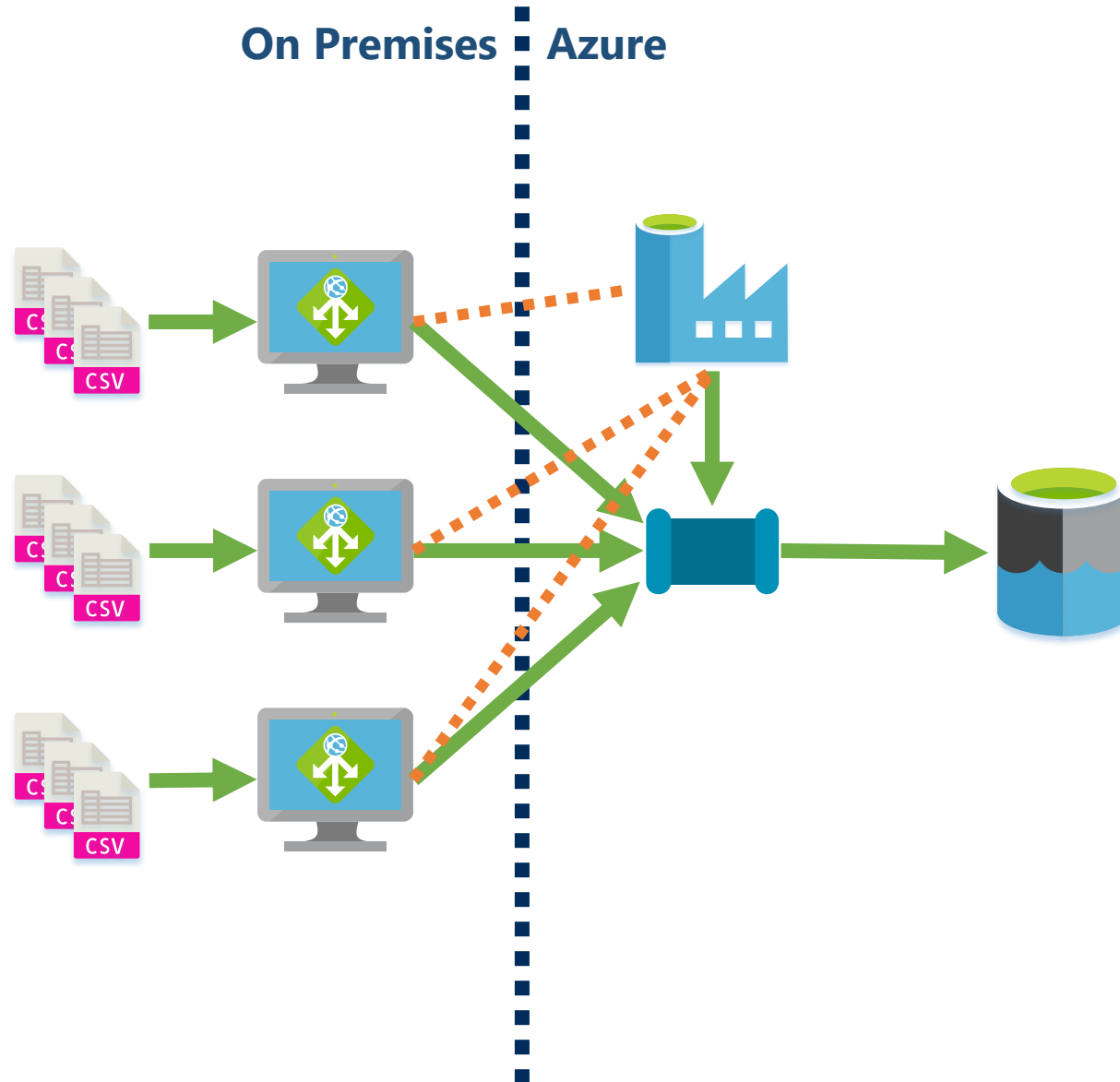




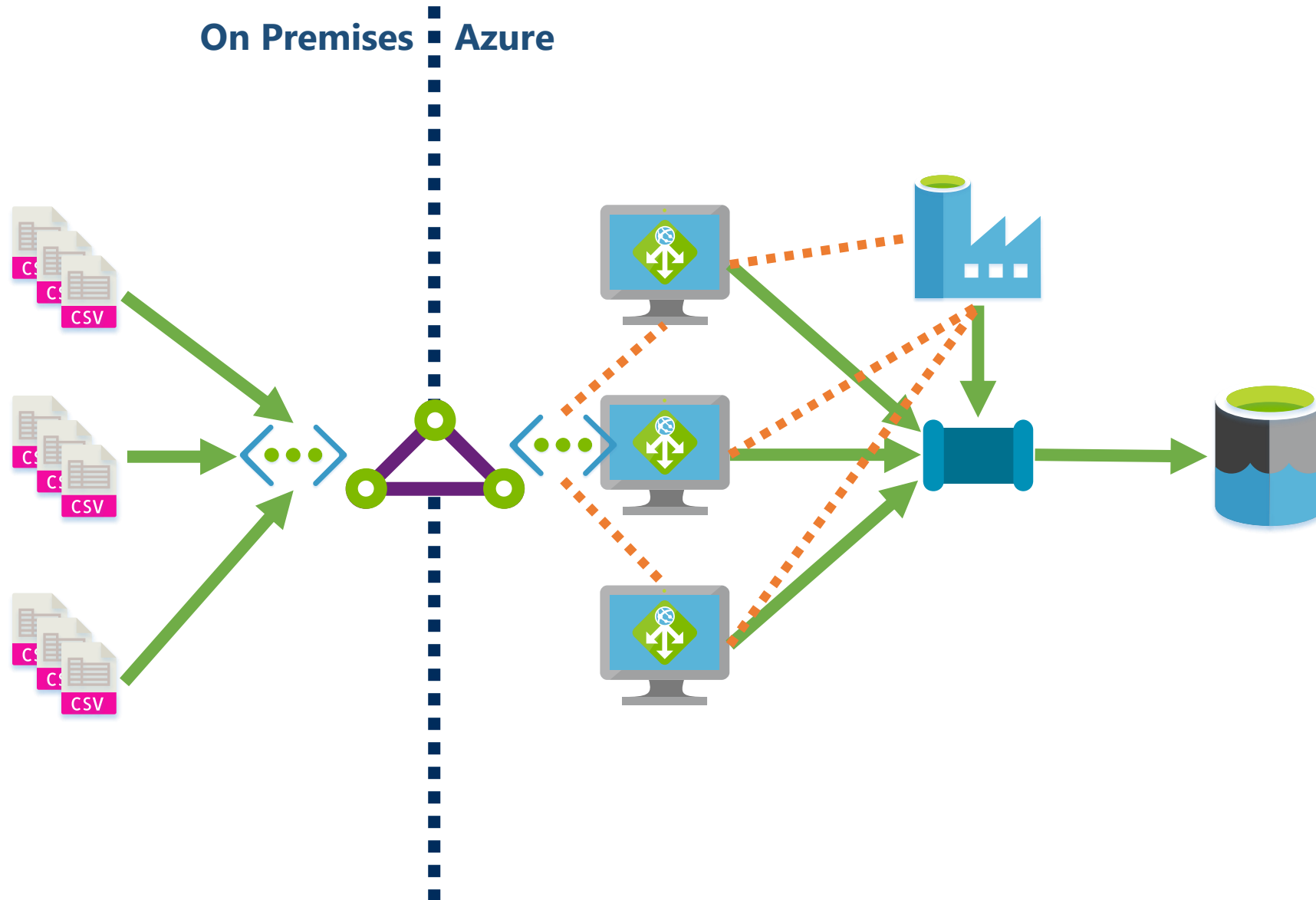
# Hosted IR Linked to Multiple Data Factory's



# Using a Hosted IR with Express Route



# Using a Hosted IR with Express Route



# Recap: Data Factory

**1 Linked Services**

**2 Data Sets**

**3 Activities**

**4 Pipelines**

**5 Triggers**

**1**

**Azure**  
Integration Runtime

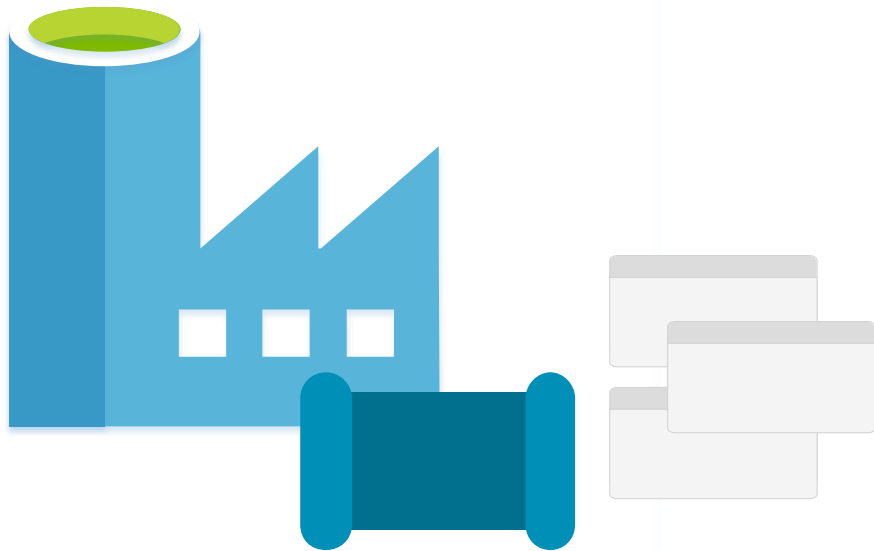
**2**

**SSIS**  
Integration Runtime

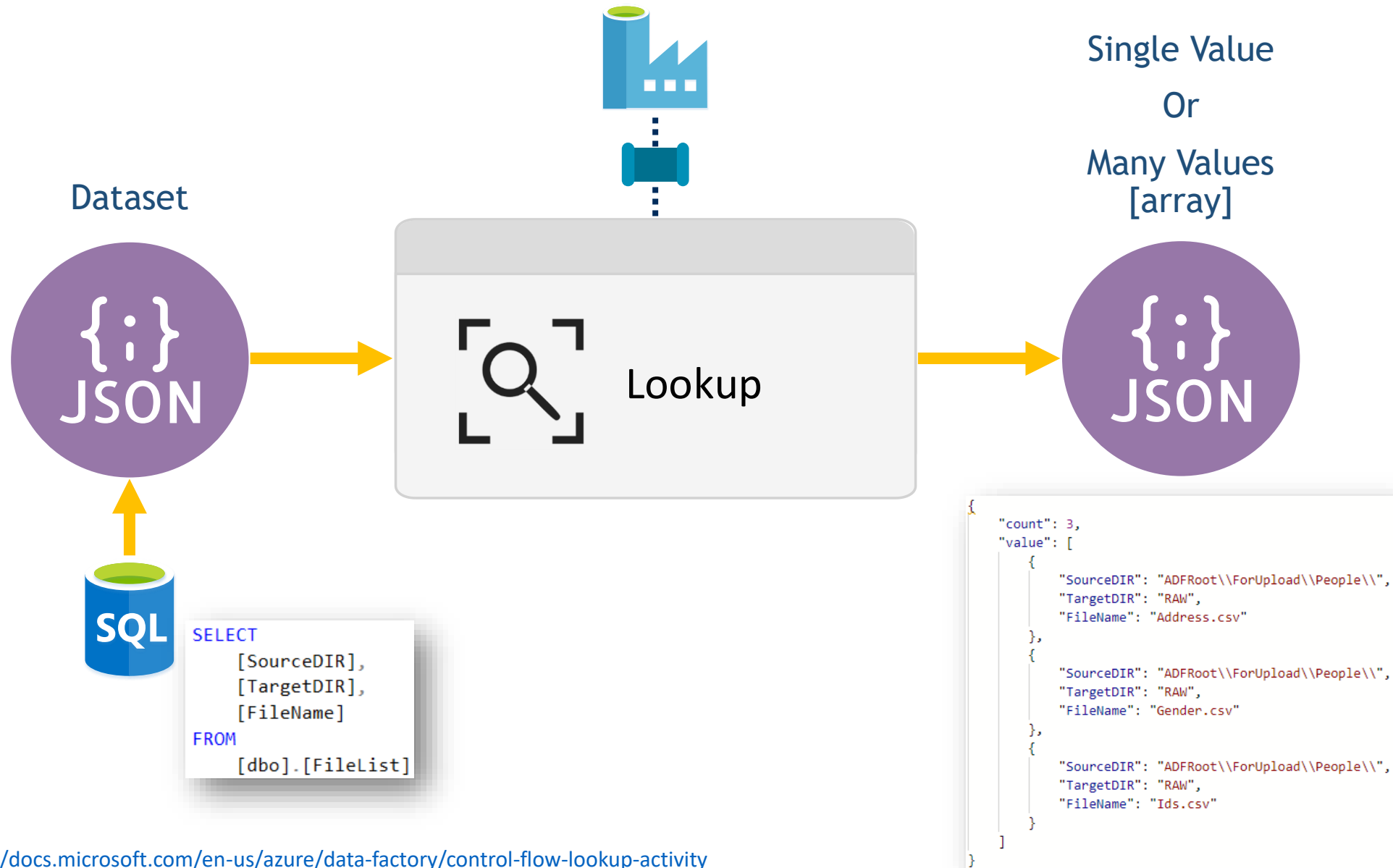
**3**

**Self Hosted**  
Integration Runtime

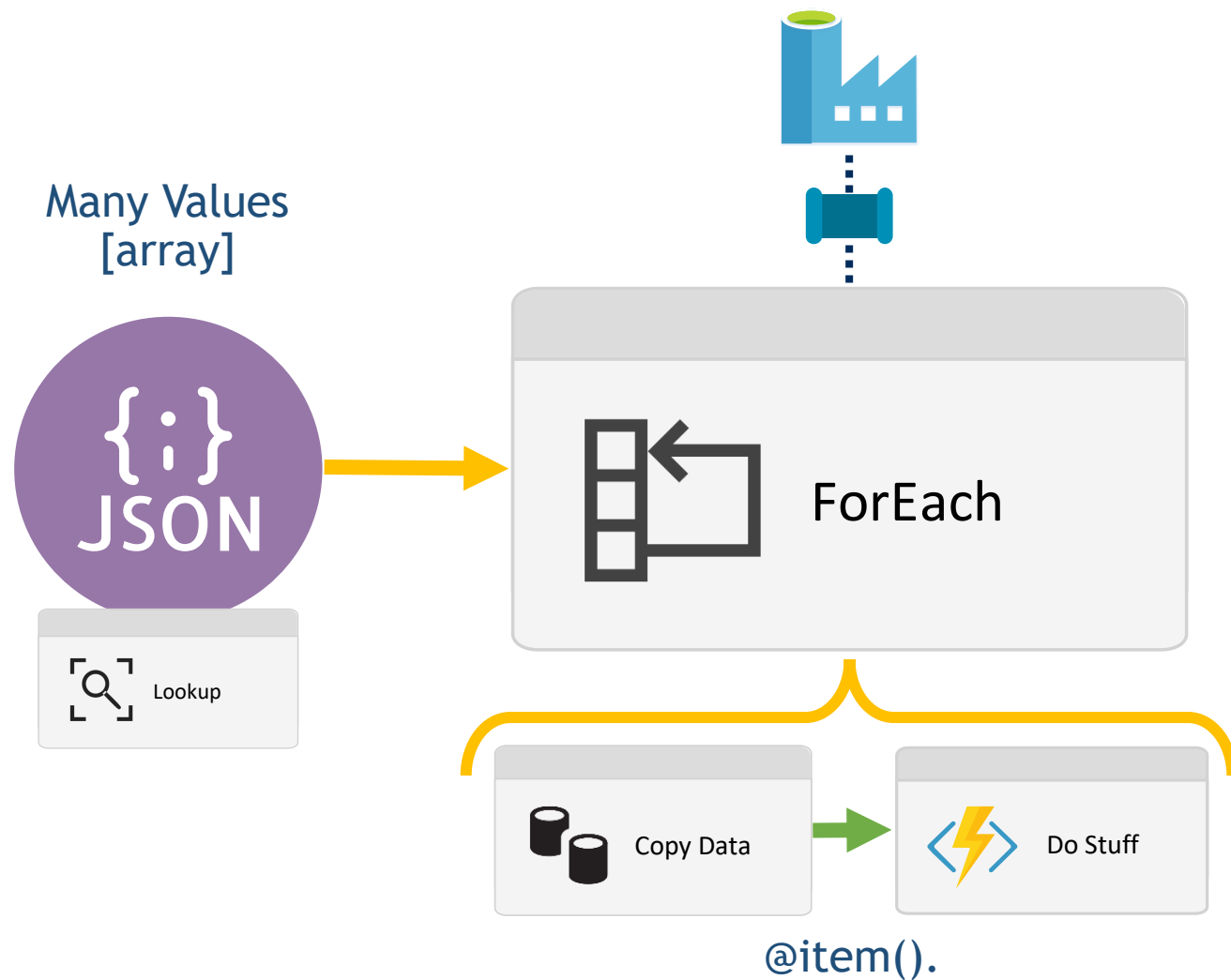
# Data Factory Key Activities



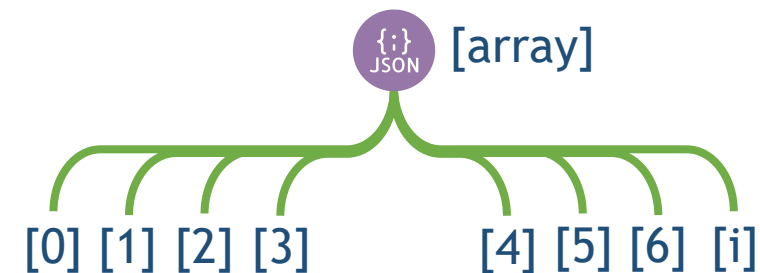
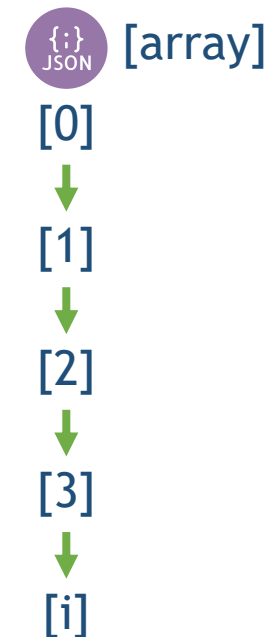
# Lookup Activity



# ForEach Activity



IsSequential:  
true



Batch Count Default: 20

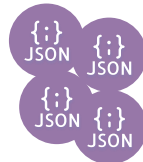
Batch Count Max: 50

# Custom Activity

## Linked Services

Azure Batch

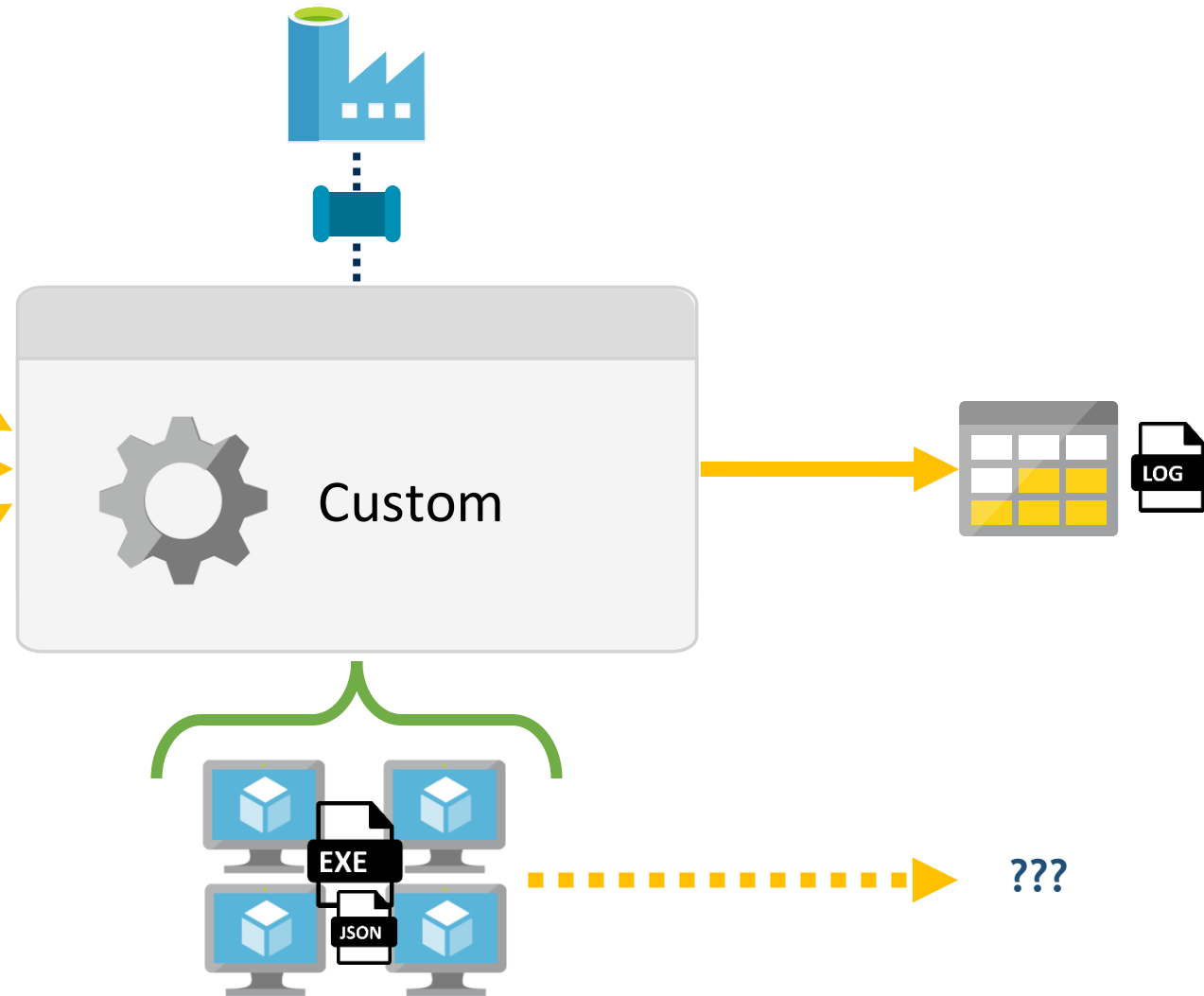
Azure Blob Storage



## References Objects

Datasets: []

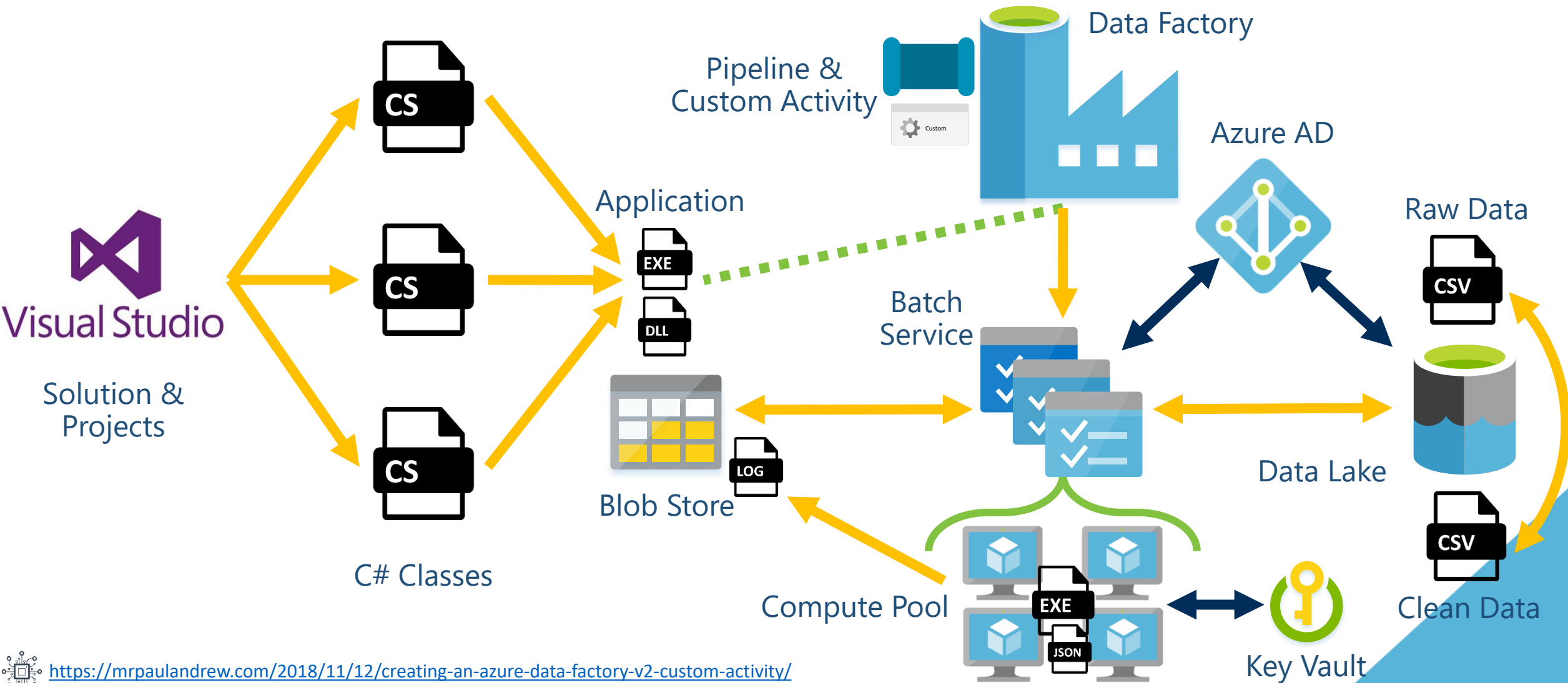
Linked Services: []





# Building a Custom Activity

A .Net Console App Executed Using Azure Batch Service.



# Recap: Data Factory – Useful Activities

1

## **Lookup**

Get value(s) from lots of places.

2

## **ForEach**

Iterations, sequentially or in parallel.

3

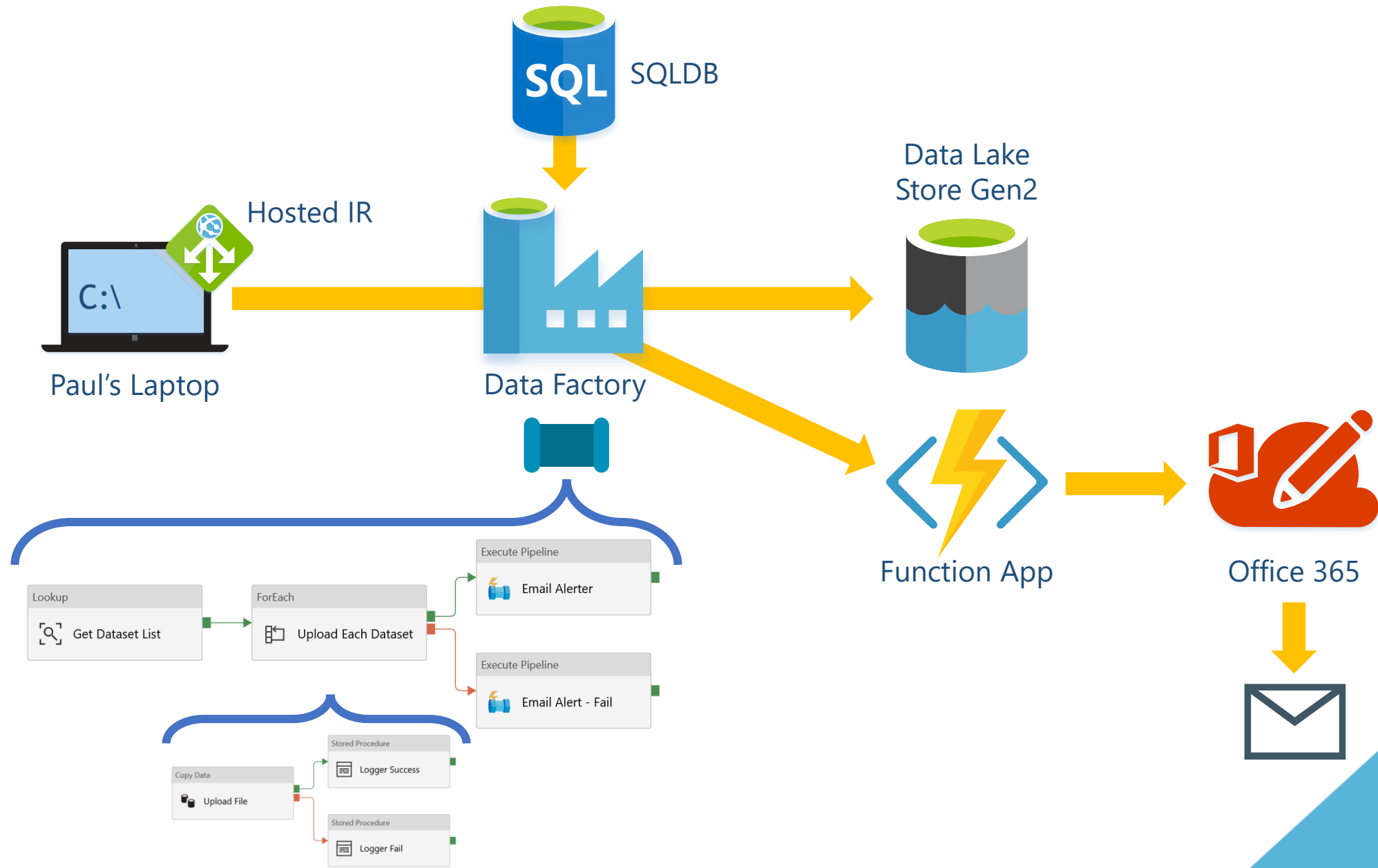
## **Custom**

Bespoke code executed by Azure Batch compute pools.

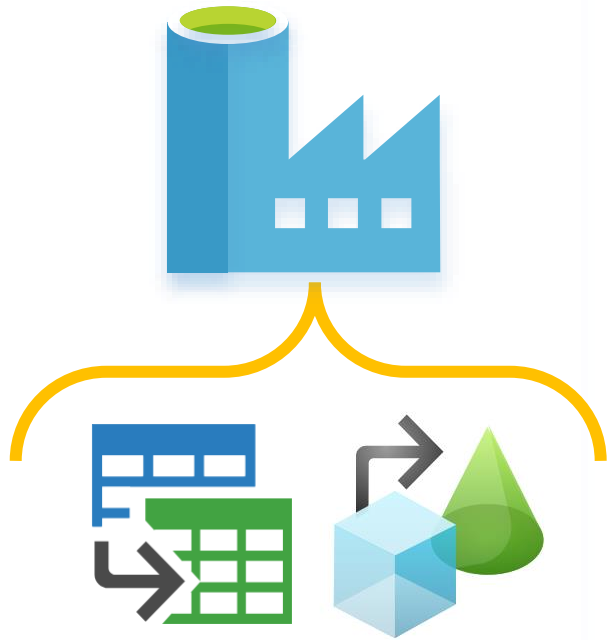
# Simple Dynamic Copy Pipeline



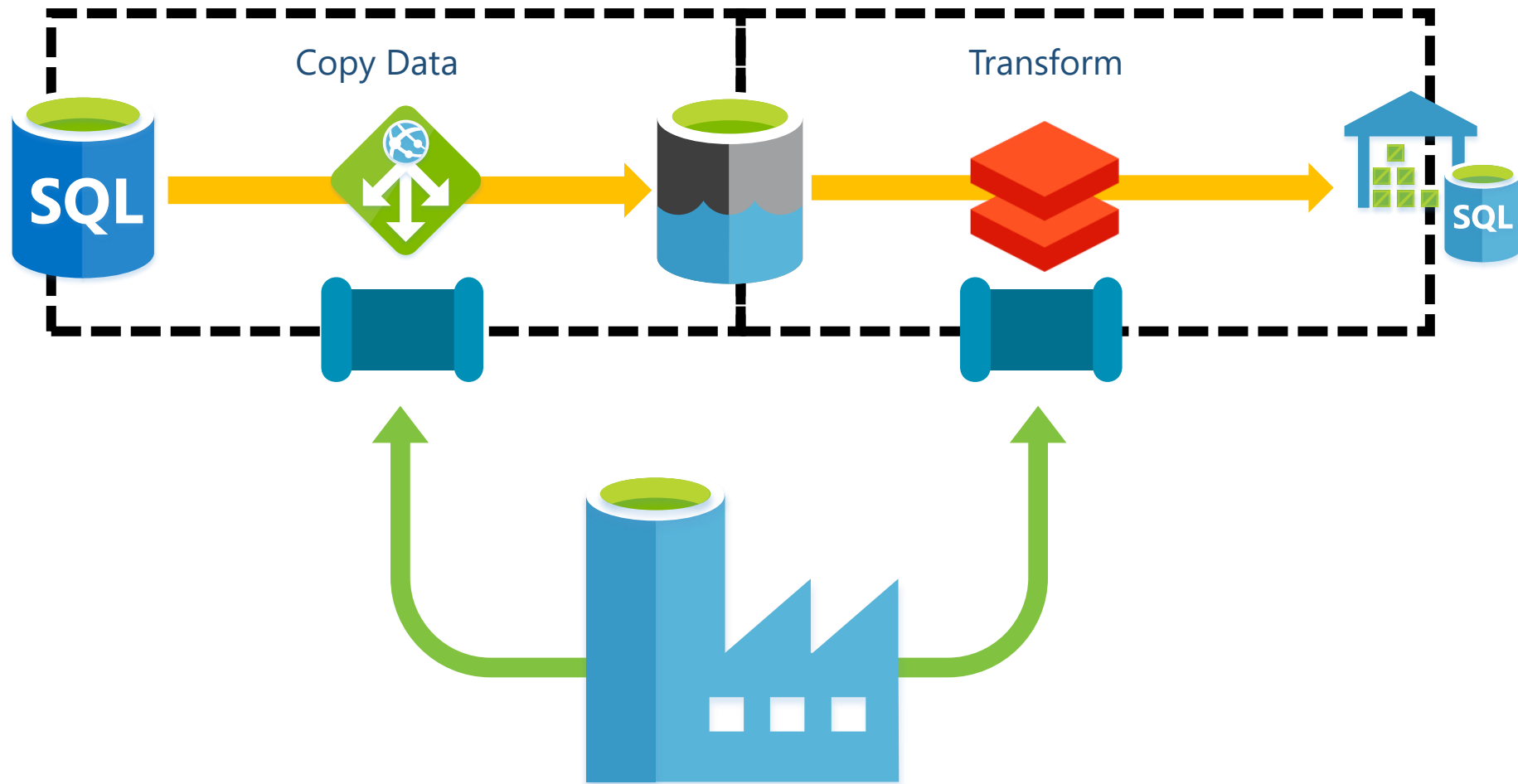
# Demo Architecture



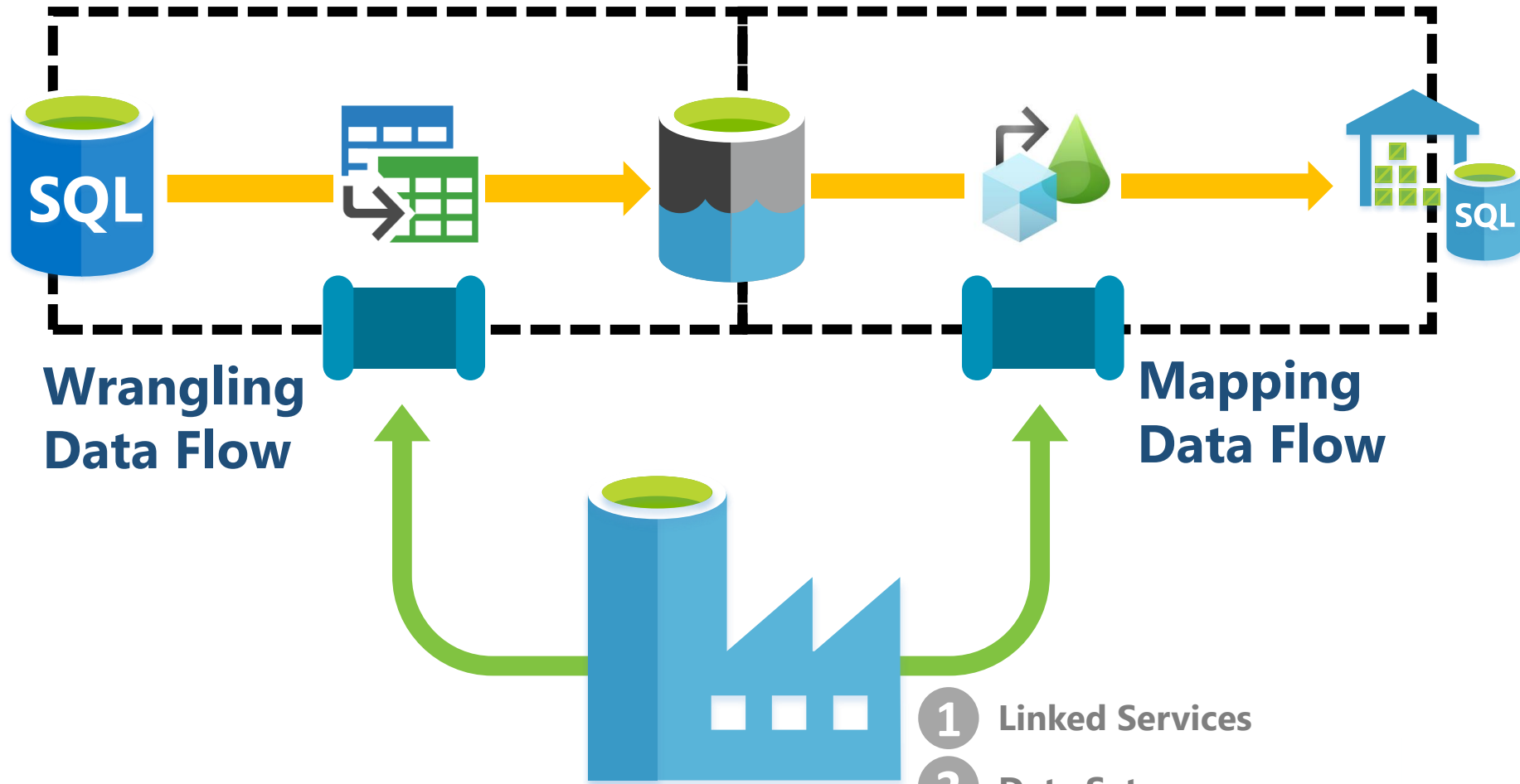
# Data Factory Data Flows



# Data Factory Control Flow Components



# Data Factory Data Flows



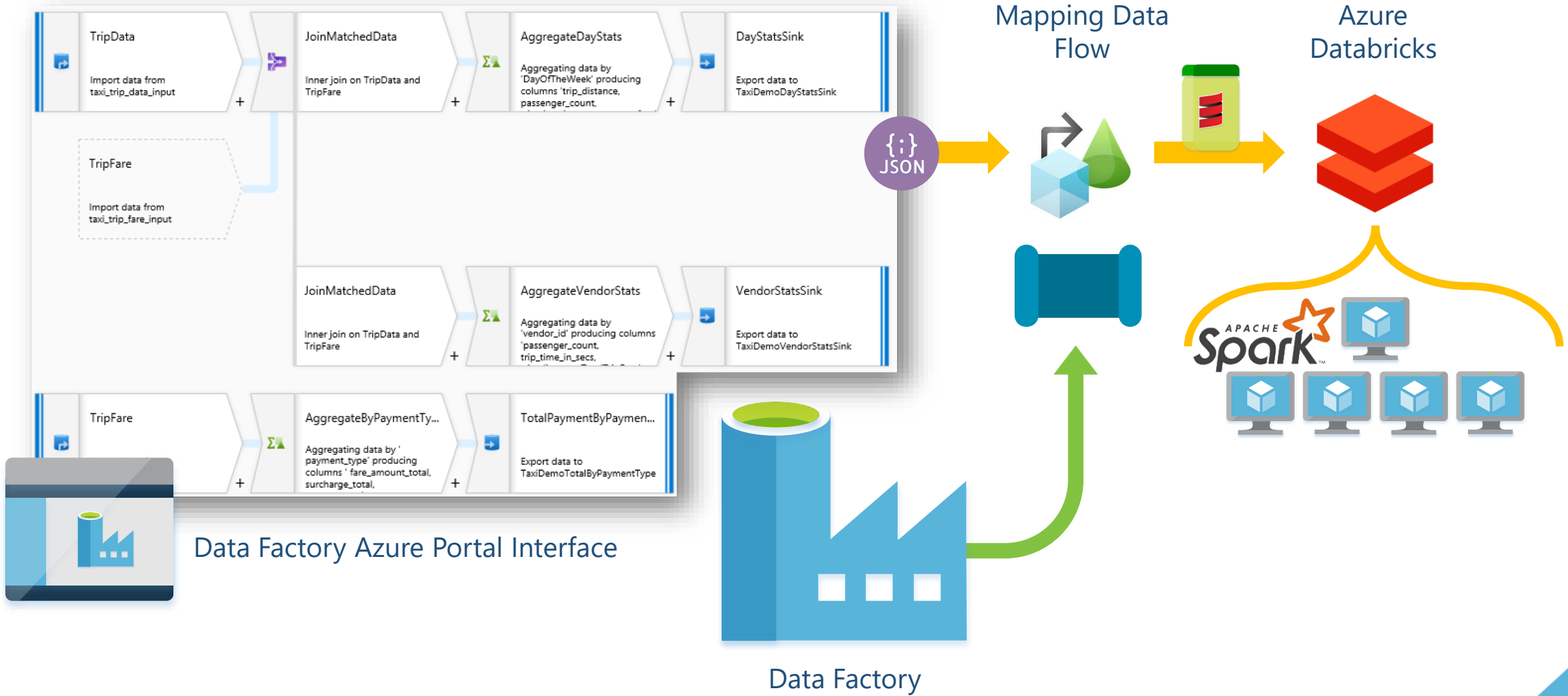
- 1 Linked Services
- 2 Data Sets
- 3 Activities
- 4 Pipelines
- 5 Triggers

# Mapping Data Flows

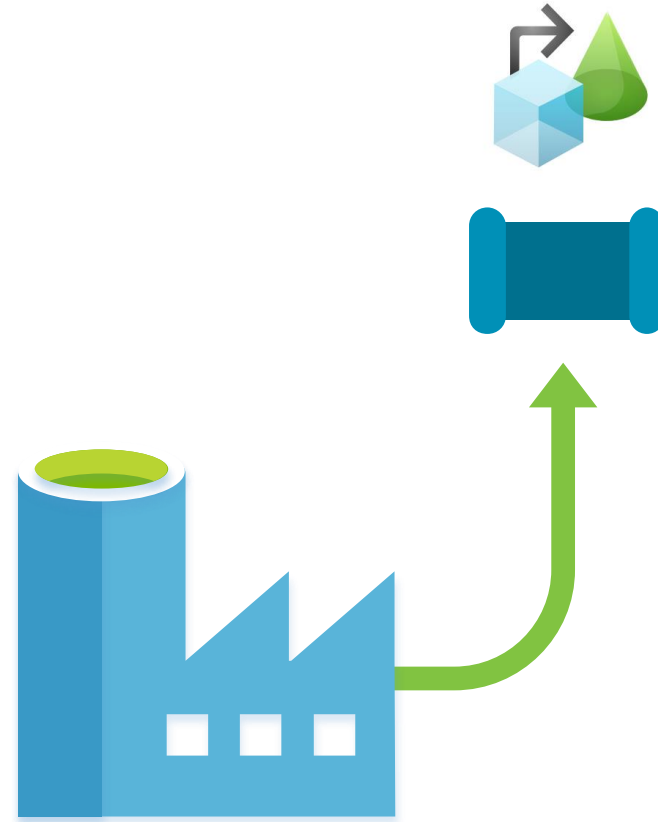




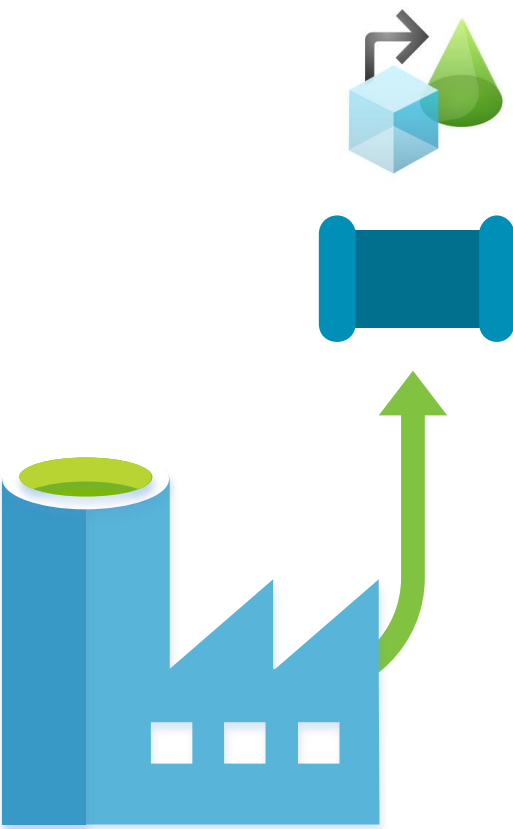
# What is a Mapping Data Flow?



# Mapping Data Flows



# Mapping Data Flows – Settings & Concepts



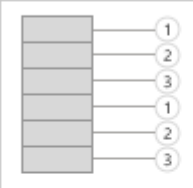
sink1  
Add sink dataset

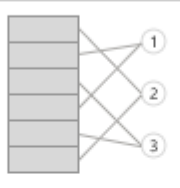
SinkSettingsMappingOptimizeInspectData Preview

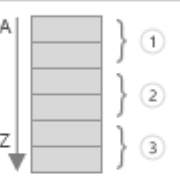
Partition option \*

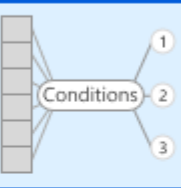
☐ Use current partitioning☐ Single partition☒ Set Partitioning

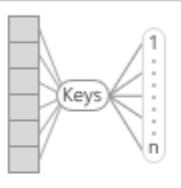
Partition type \*

  
Round Robin

  
Hash

  
Dynamic Range

  
Fixed Range

  
Key

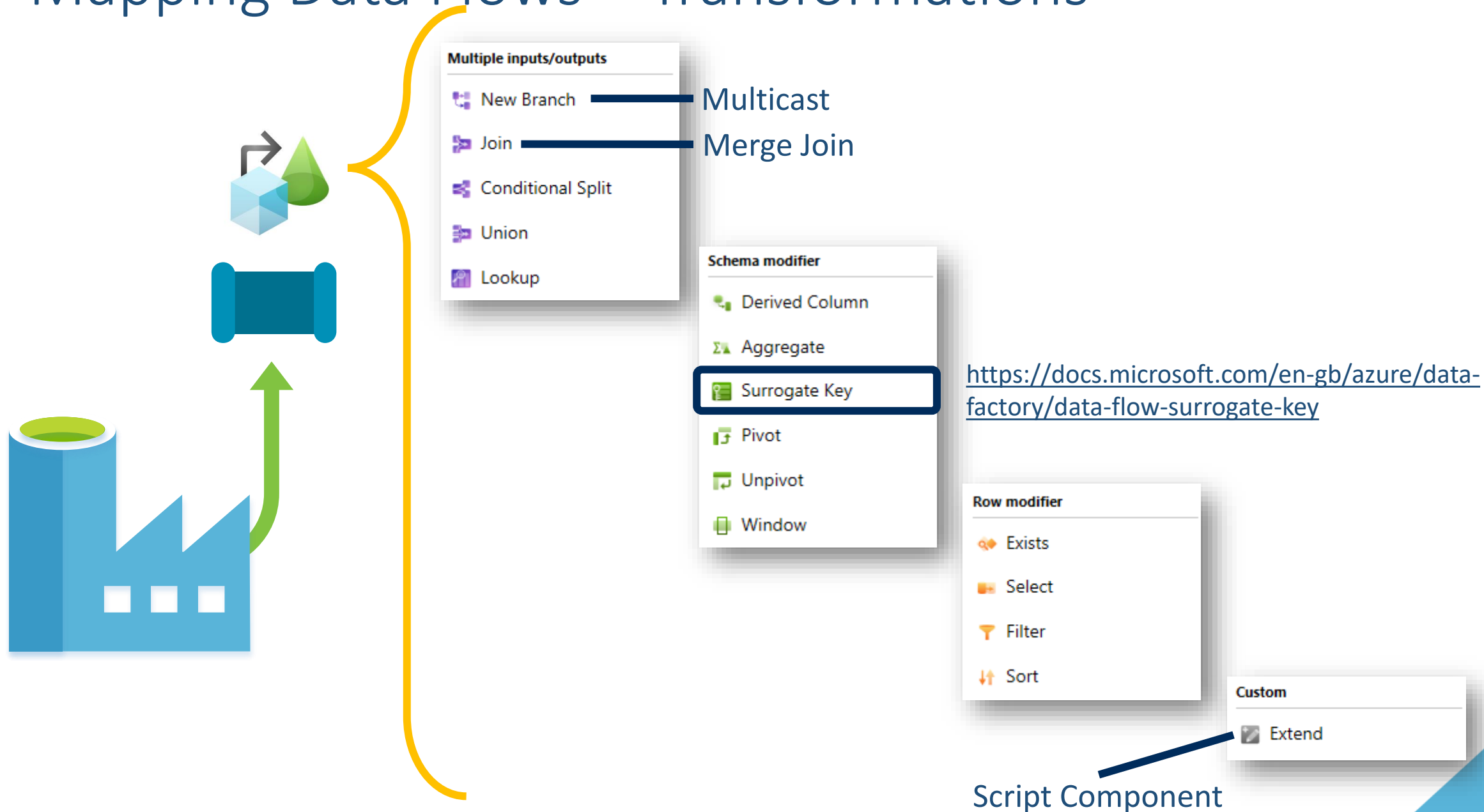
Number of partitions \*

Condition to partition \*

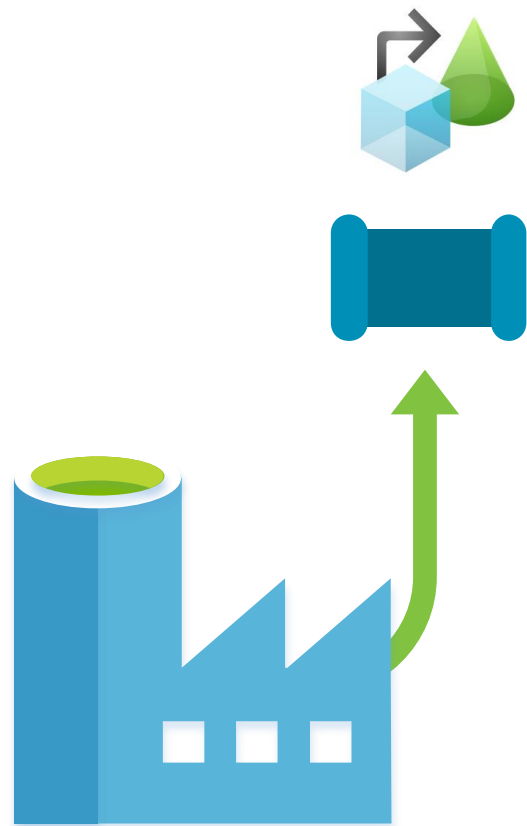
Condition

ANY + -

# Mapping Data Flows – Transformations



# Mapping Data Flows – Expression Builder



**Visual Expression Builder**

Currently working on: year

Filter...

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)

✕

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'

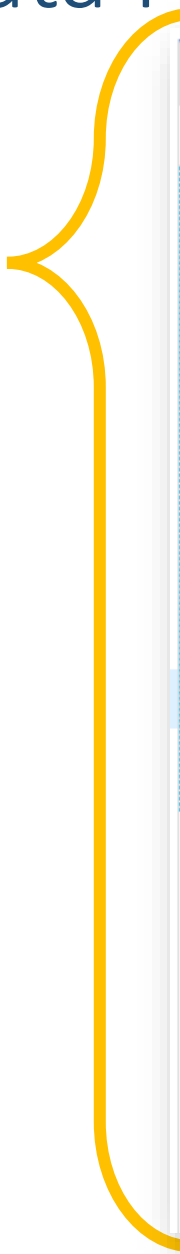

+ - \* / ||

**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** X **ADWAnalysisB...** X

☒ **Debug** Saved **Validate** Source Settings Cluster ForDataFlow

**OrderHeader**  
Columns: 22 total

**OrderDetails**  
Import data from ADWSalesOrderDetail

**Join1**  
Inner join on OrderHeader and OrderDetails

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

**sink1**  
Export data to ADWOrderLineCountTable

**Add Source**

General Purpose cluster

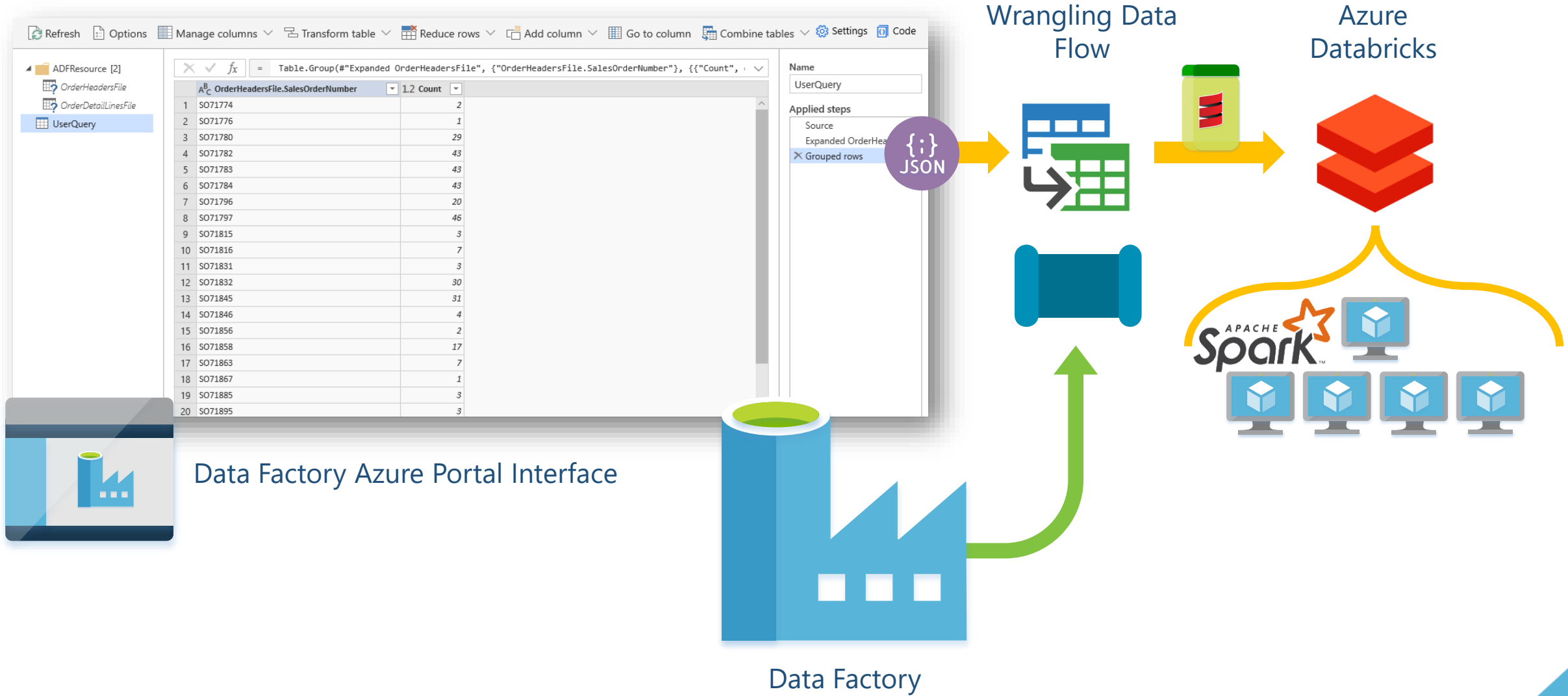
Source Settings Define schema Optimize Inspect **Data Preview**

	Updated*	New*	Unchanged	Total
Number of rows	N/A	N/A	N/A	32
<b>SalesOrderID</b> 123	<b>RevisionNumber</b> abc	<b>OrderDate</b> 🕒	<b>DueDate</b> 🕒	<b>ShipDate</b>
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

# Wrangling Data Flows



# What is a Wrangling Data Flow?





# What is a Wrangling Data Flow?

Refresh

Options

Manage columns

Transform table

Reduce rows

Add column

Go to column

Combine tables

Settings

Code

ADFRsource [2]

OrderHeadersFile

OrderDetailLinesFile

UserQuery

fx

=

Table.Group(#"Expanded OrderHeadersFile", {"OrderHeadersFile.SalesOrderNumber"}, {"Count",

	OrderHeadersFile.SalesOrderNumber	Count
1	SO71774	2
2	SO71776	1
3	SO71780	29
4	SO71782	43
5	SO71783	43
6	SO71784	43
7	SO71796	20
8	SO71797	46
9	SO71815	3
10	SO71816	7
11	SO71831	3
12	SO71832	30
13	SO71845	31
14	SO71846	4
15	SO71856	2
16	SO71858	17
17	SO71863	7
18	SO71867	1
19	SO71885	3
20	SO71895	3

Name

UserQuery

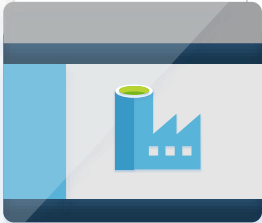
Applied steps

Source

Expanded OrderHeadersF...

Grouped rows

Power BI Desktop



Data Factory

Close & Apply

New Source

Recent Sources

Enter Data

Data source settings

Manage Parameters

Refresh Preview

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

Combine

Queries [3]

SpeakingLog

EventLogos

TagsToTalkId

fx

=

Table.RenameColumns(dbo\_SpeakingLog,{{"LogId", "TalkId"}})

	TalkId	TalkDate	EventName	City	Country
1	1	22/06/2016	Guerrilla Lightning Talks	Online	Online
2	2	08/07/2016	STEM	Stafford	England
3	3	23/07/2016	SQL Saturday	Manchester	England
4	4	11/08/2016	User Group	Birmingham	England
5	5	14/09/2016	User Group	Manchester	England
6	6	21/09/2016	British Computer Society	Telford	England
7	7	03/10/2016	Data Relay	Birmingham	England
8	8	04/10/2016	Data Relay	Cardiff	Wales
9	9	05/10/2016	Data Relay	Reading	England
10	10	06/10/2016	Data Relay	Nottingham	England
11	11	07/10/2016	Data Relay	Leeds	England
12	12	16/11/2016	STEM	Trentham	England
13	13	23/11/2016	STEM	Trentham	England
14	14	01/12/2016	STEM	Trentham	England
15	15	23/01/2017	User Group	Exeter	England
16	16	24/01/2017	STEM	Telford	England
17	17	31/01/2017	STEM	Telford	England
18	18	01/02/2017	User Group	Southampton	England
19	19	07/02/2017	User Group	Bristol	England
20	20	09/02/2017	User Group	Birmingham	England
21	21	10/03/2017	STEM	Stone	England
22	22	07/04/2017	SQL Bits	Telford	England

Query Settings

PROPERTIES

Name

SpeakingLog

All Properties

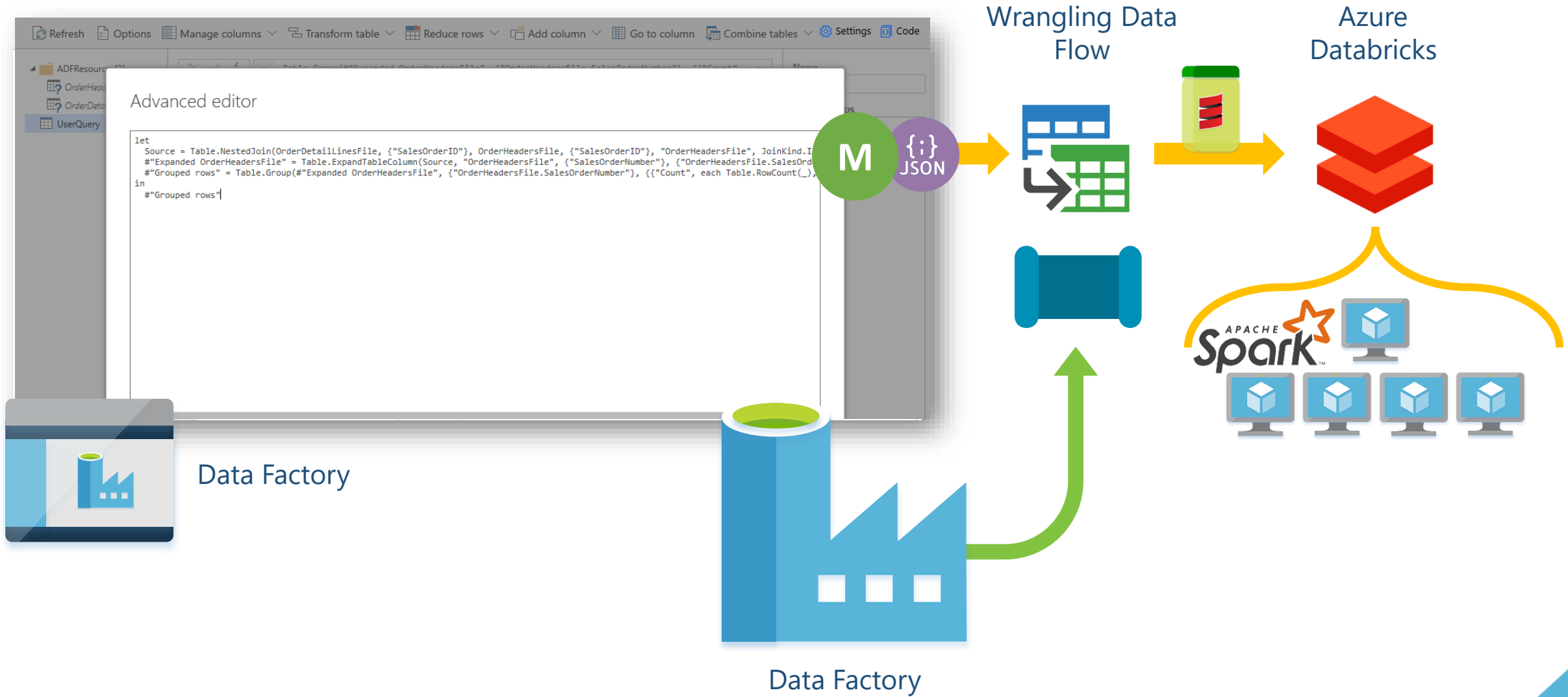
APPLIED STEPS

Source

Navigation


Rename Columns


# What is a Wrangling Data Flow?





# Data Flow - Cluster Configuration

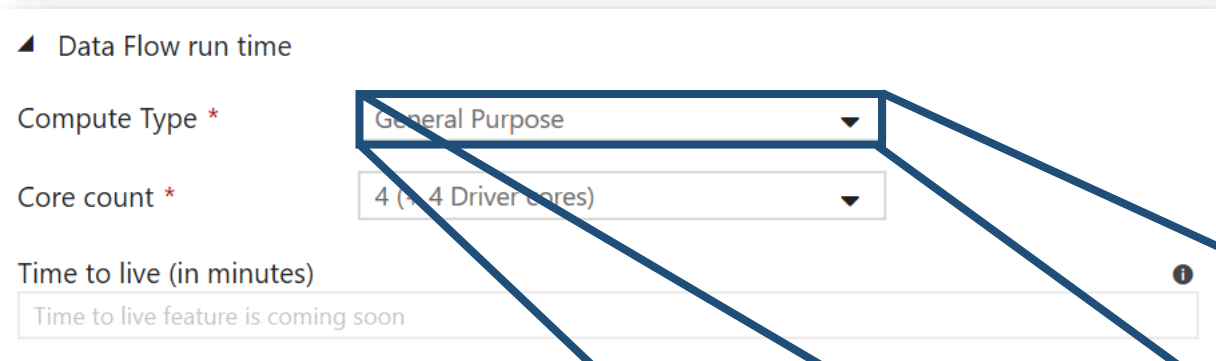
## Integration Runtimes

 **1 Azure**  
Integration Runtime

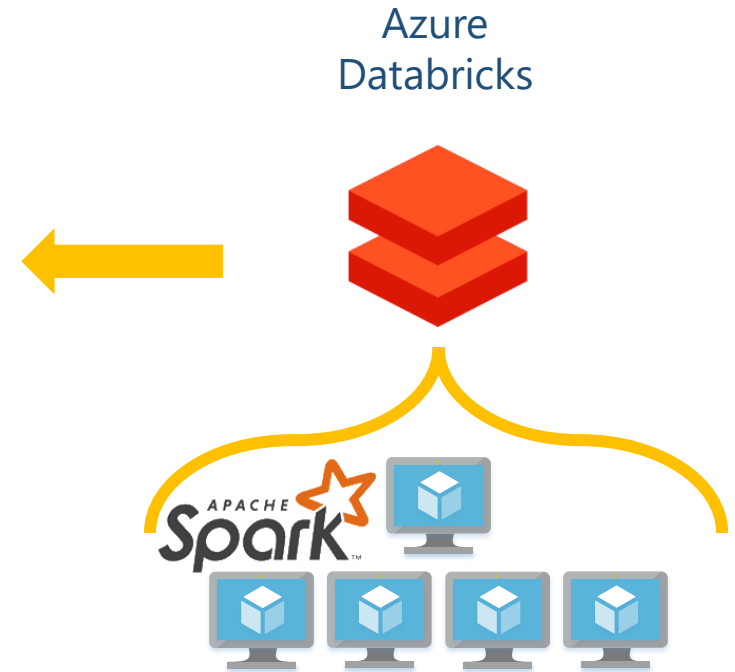
Movement Hours 

Activity Orchestration 

Flexible Region 

  
▲ Data Flow run time  
Compute Type \*   
Core count \*   
Time to live (in minutes)

- General Purpose
- Memory Optimised
- Compute Optimised



# Recap: Data Factory – Data Flows

1

## **Mapping**

Similar to SSIS Data Flows in appearance.

2

## **Wrangling**

Similar to Power BI Power Query.

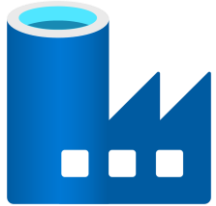
3

## **Data Flow Cluster Config**

Via the Data Factory Azure IR

# Data Factory DevOps – CI/CD





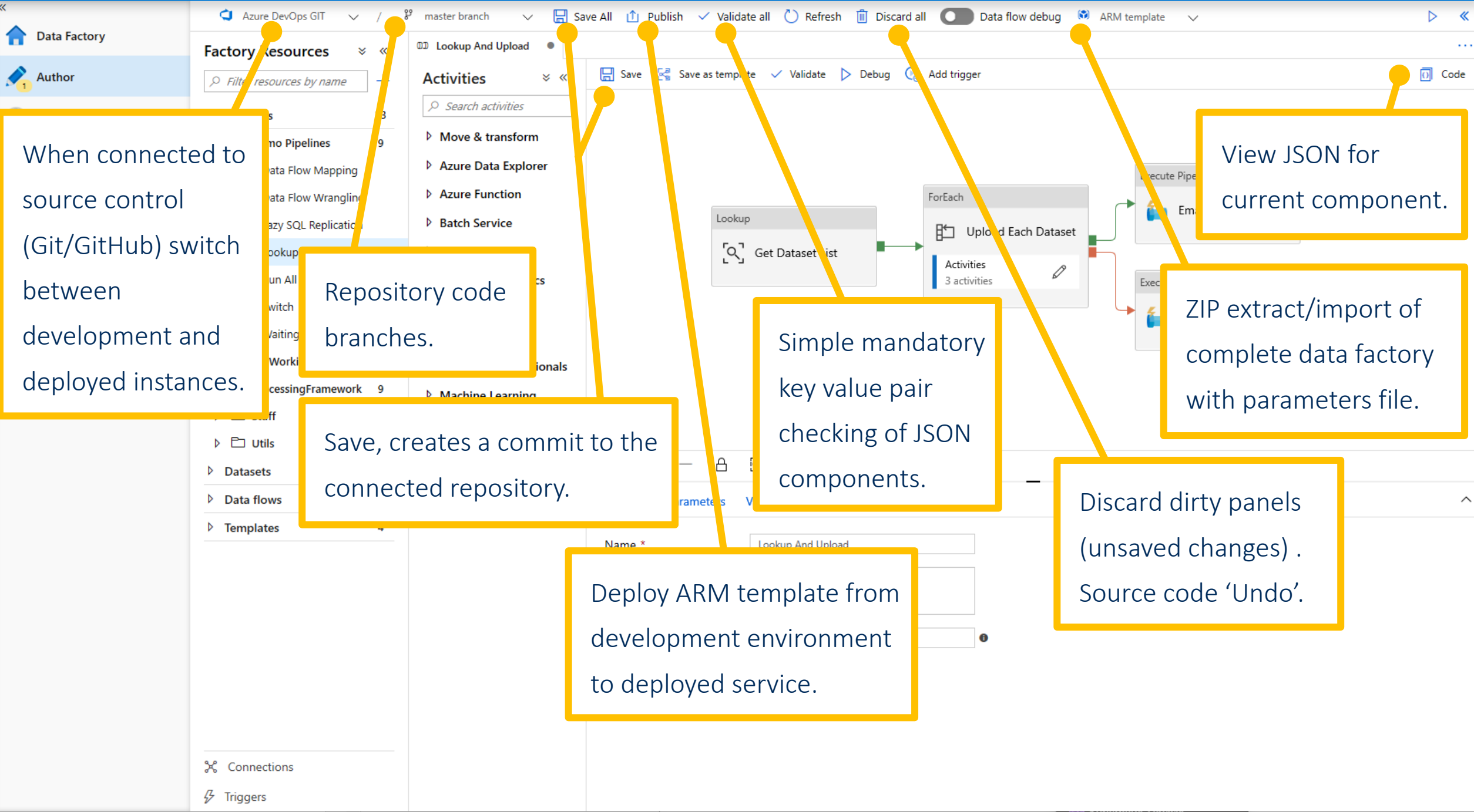
PaulsFunFactory



Documentation



Author & Monitor



- Data Factory
- Author
- Monitor

- Factory Resources
- Filter resources by name
- Pipelines 23
    - Demo Pipelines 9
      - Data Flow Mapping
      - Data Flow Wrangling
      - Lazy SQL Replication
      - Lookup And Upload
      - Run All SSIS Packages
      - Switch
      - WaitingPipeline
      - Working Progress 2
      - ProcessingFramework 9
      - Stuff 4
      - Utils 1
    - Datasets 27
    - Data flows 6
    - Templates 4

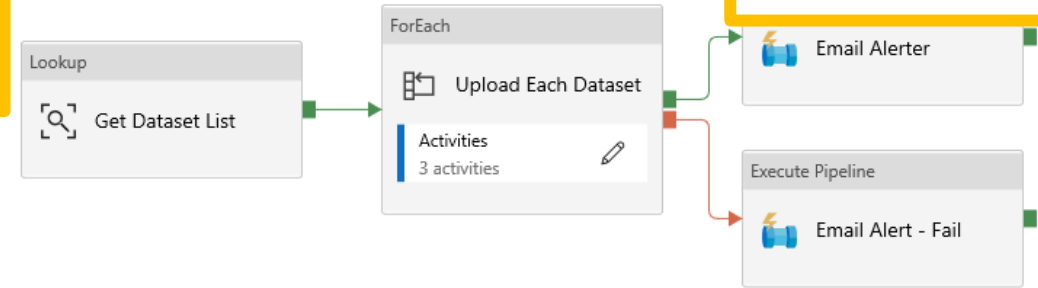
Lookup And Upload

Activities

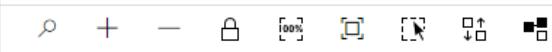
Search activities

Debug the Control Flow.  
Run the pipeline.

Save | Save as template | Validate | Debug | Add trigger



Debug the Data Flow.  
Get a cluster ready.



General | Parameters | Variables | Output

Name \* | Lookup And Upload

Description | Simple dynamic demo pipeline

Concurrency |

Annotations | + New



Factory Resources

Filter resources by name

Pipelines23

Demo Pipelines9

Data Flow Mapping

Data Flow Wrangling

Lazy SQL Replication

Lookup And Upload

Run All SSIS Packages

Switch

WaitingPipeline

Working Progress2

ProcessingFramework9

Stuff4

Utils1

Datasets27

Data flows6

Templates4

Lookup And Upload

Activities

Search activities

Databricks

Data Lake Analytics

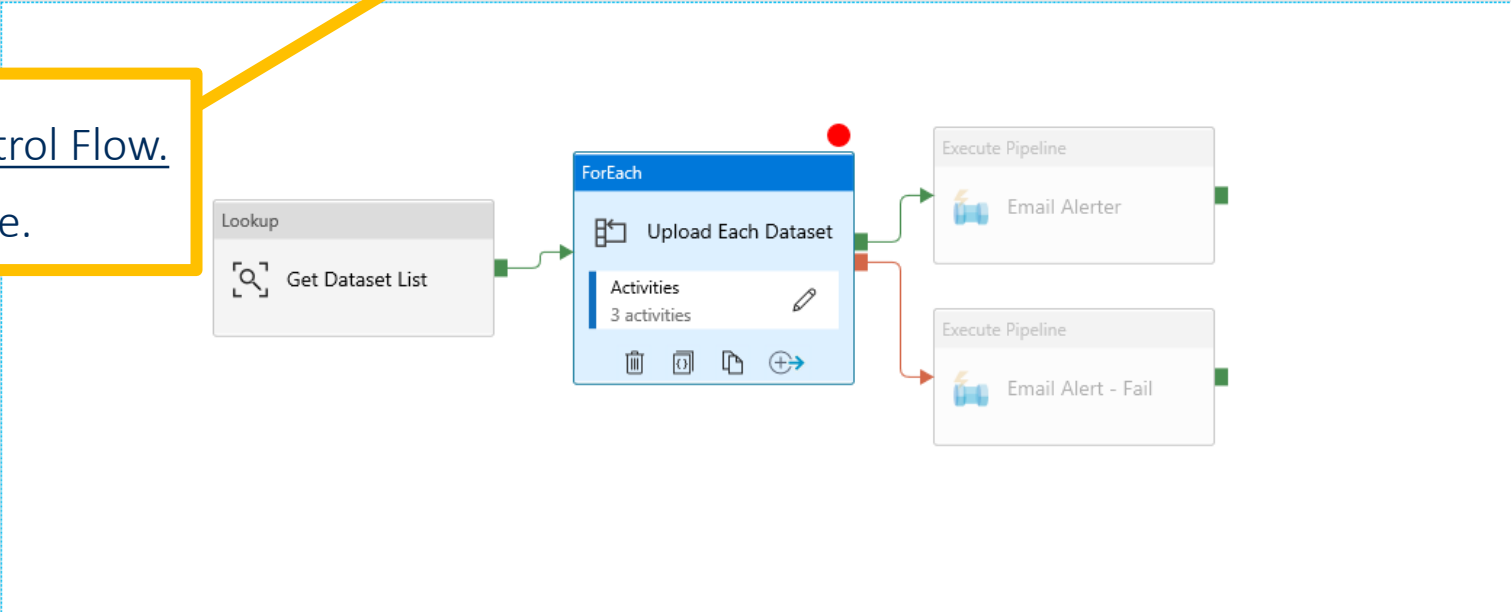
General

HDInsight

Iteration & conditionals

Machine Learning

Debug the Control Flow.  
Run the pipeline.



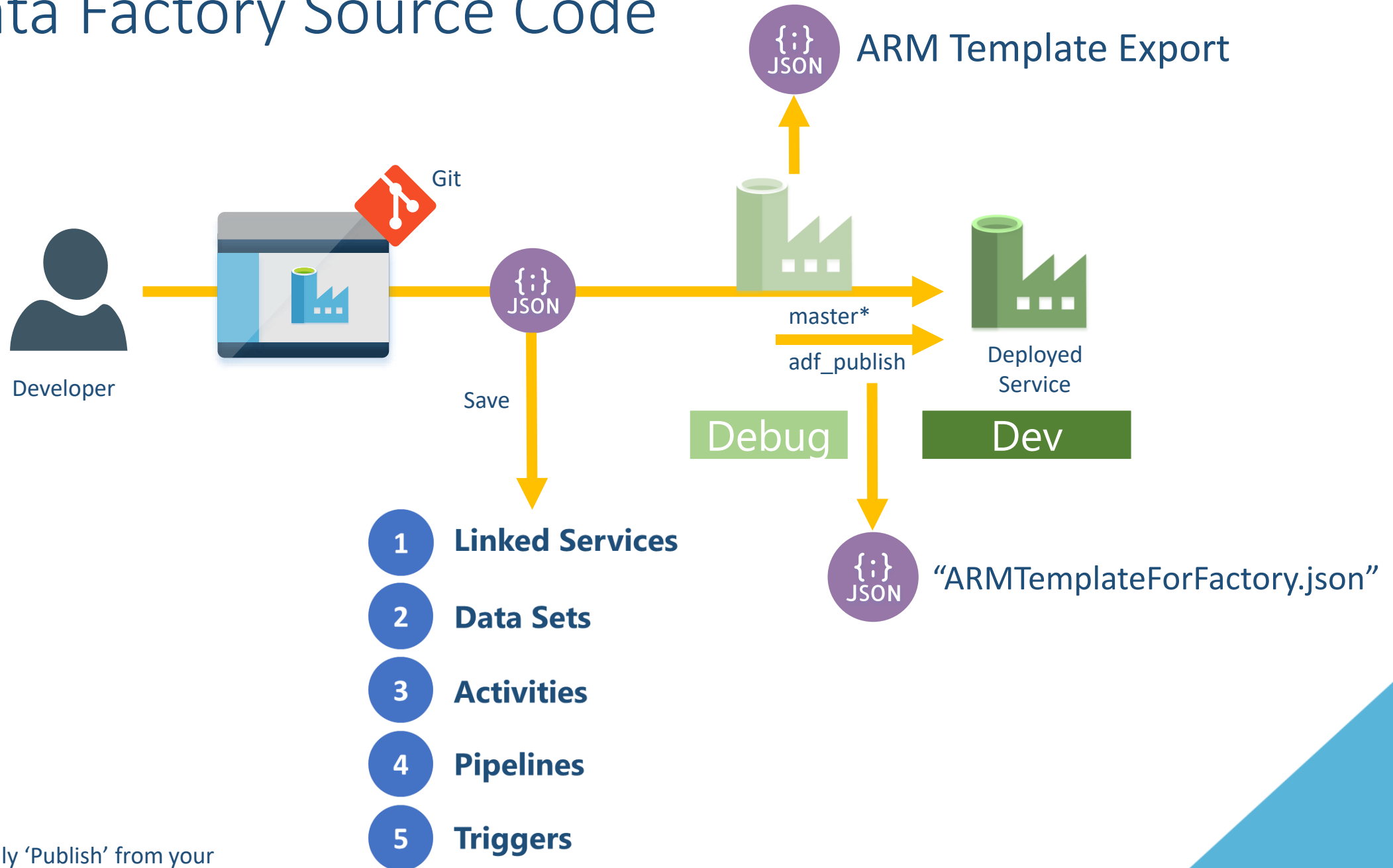
General Settings Activities (3) User properties

Name \*Upload Each Dataset

Description

Learn more

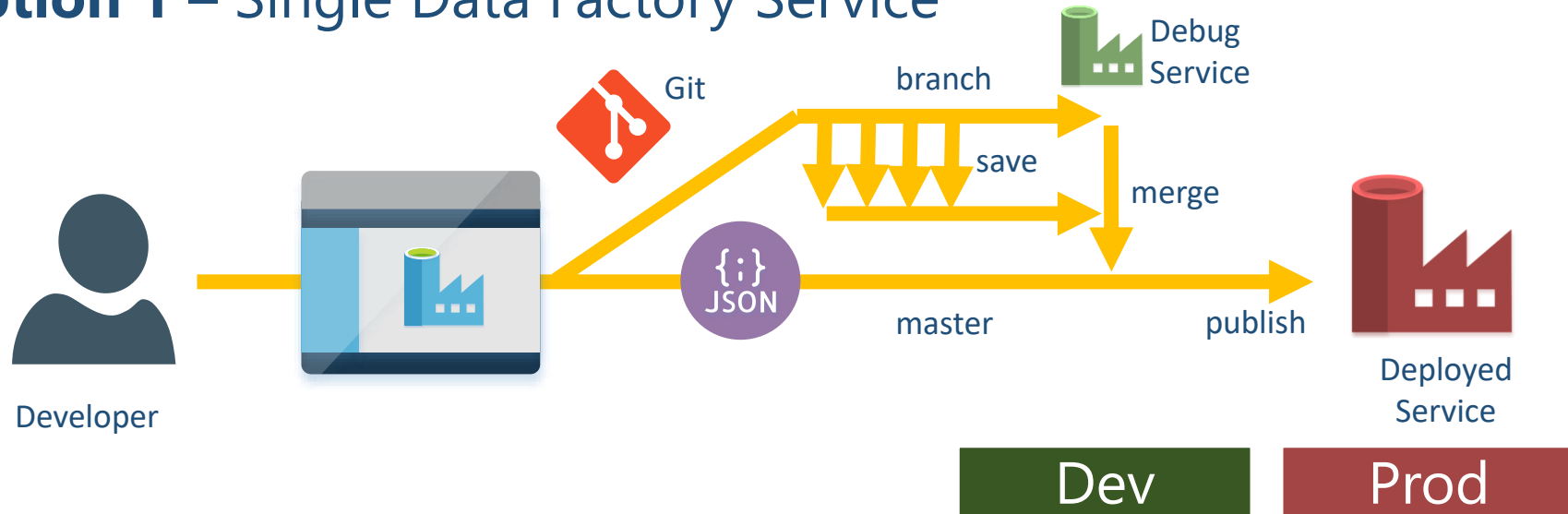
# Data Factory Source Code



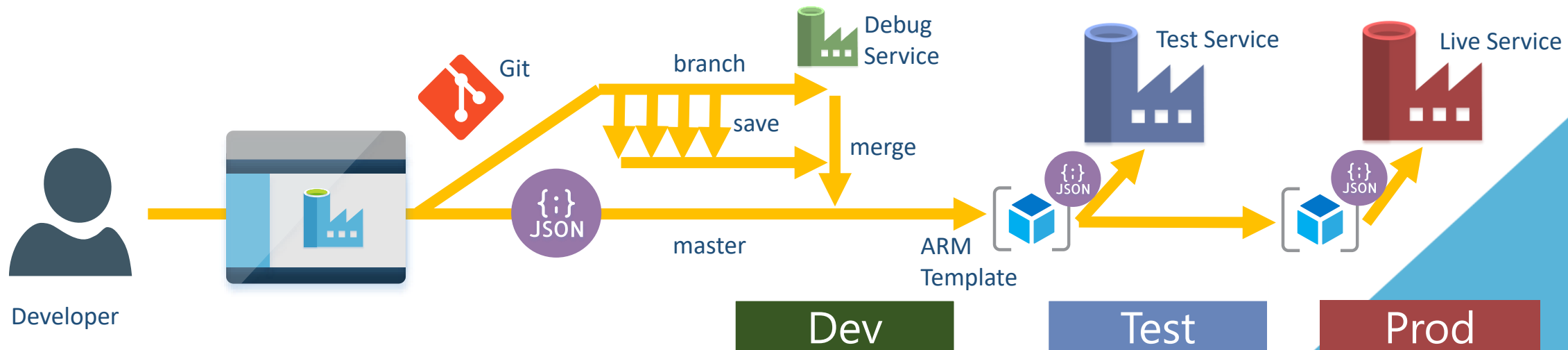
\* Can only 'Publish' from your default code branch.

# Data Factory Deployments

## Option 1 – Single Data Factory Service



## Option 2 – ARM Templates for Multiple Data Factory Services



# Recap: Data Factory – DevOps

1

## **Development and Debugging**

Via the Portal UI

2

## **Source Code**

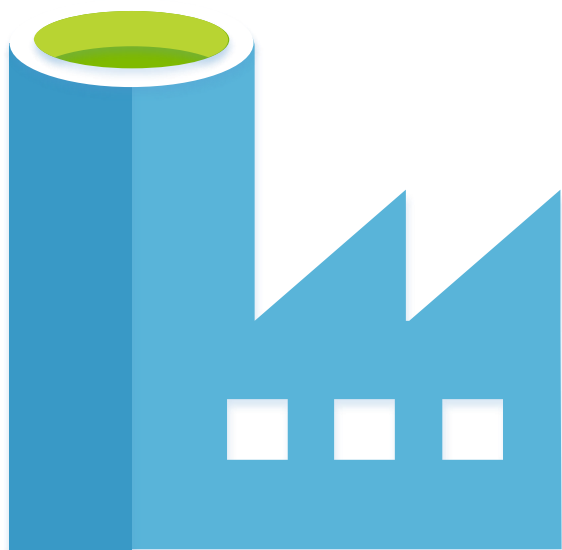
3x choices of JSON files!

3

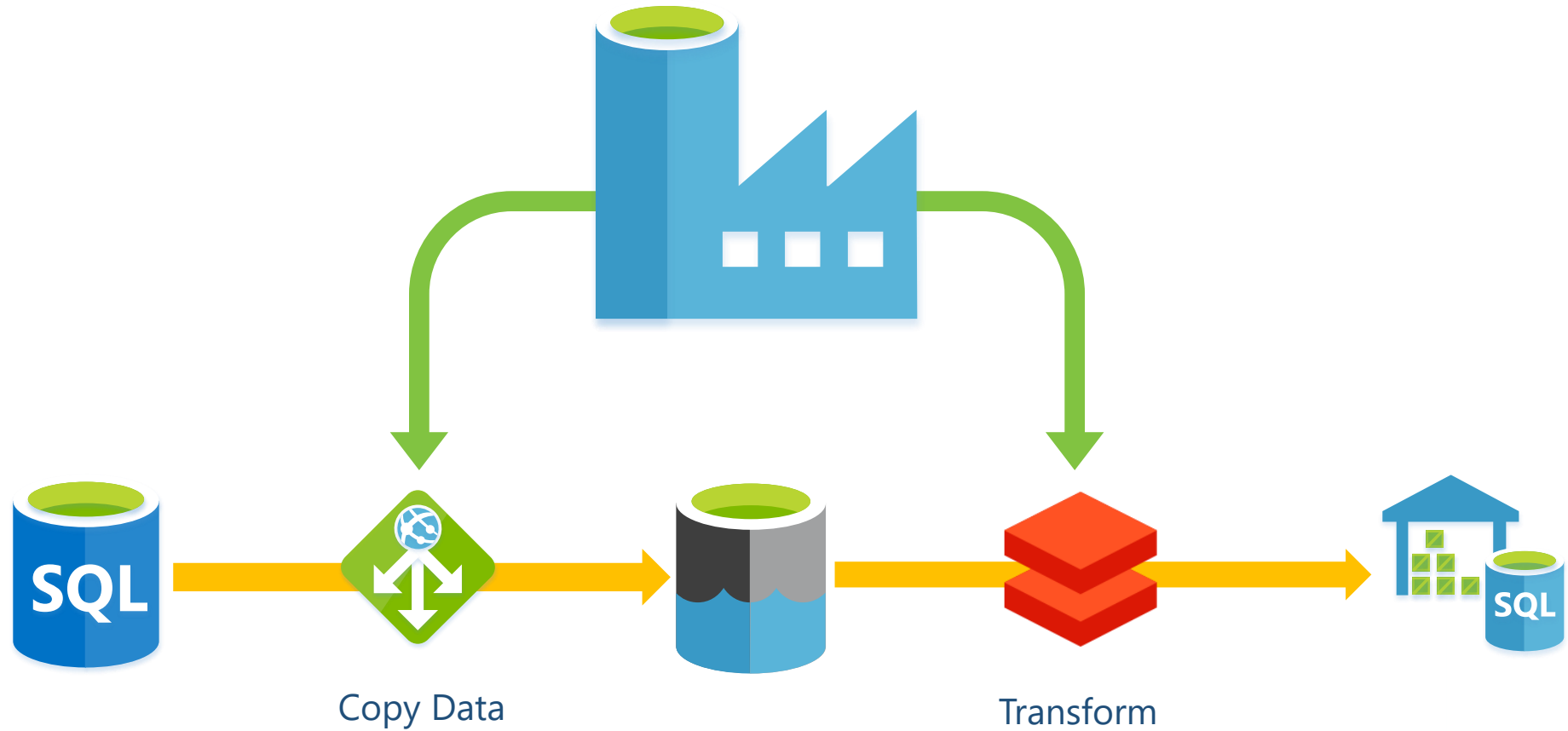
## **Deployment**

Several options depending on requirements.

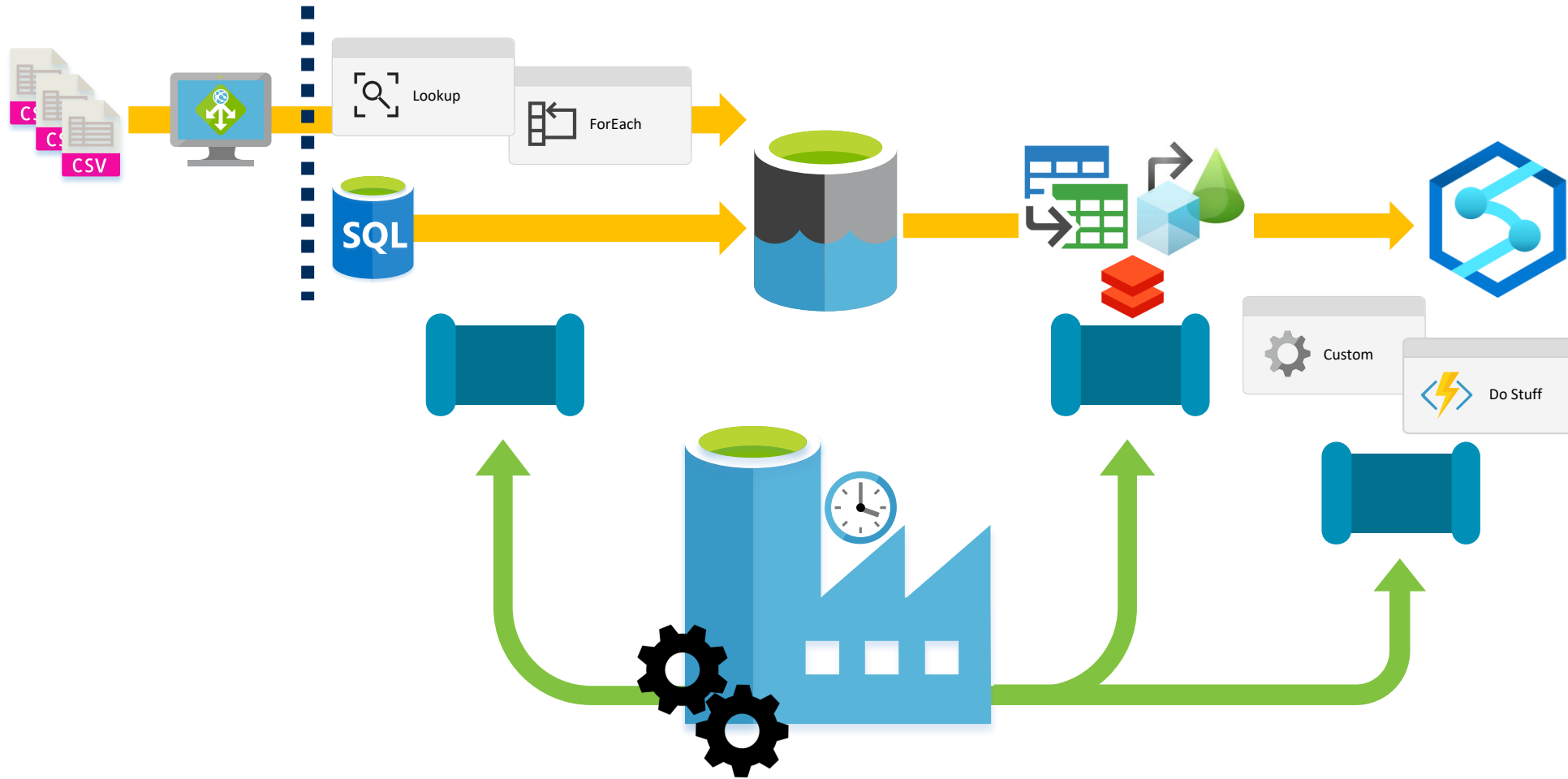
# Summary



# What is Azure Data Factory?



# What is Azure Data Factory?



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

# Thank you for listening...

Paul Andrew



altius

**Blog:** [mrpaulandrew.com](http://mrpaulandrew.com)

**Email:** [paul@mrpaulandrew.com](mailto:paul@mrpaulandrew.com)

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

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

**GitHub:** [github.com/mrpaulandrew](https://github.com/mrpaulandrew)



Slides:  
[Community Events](#)  
Repository