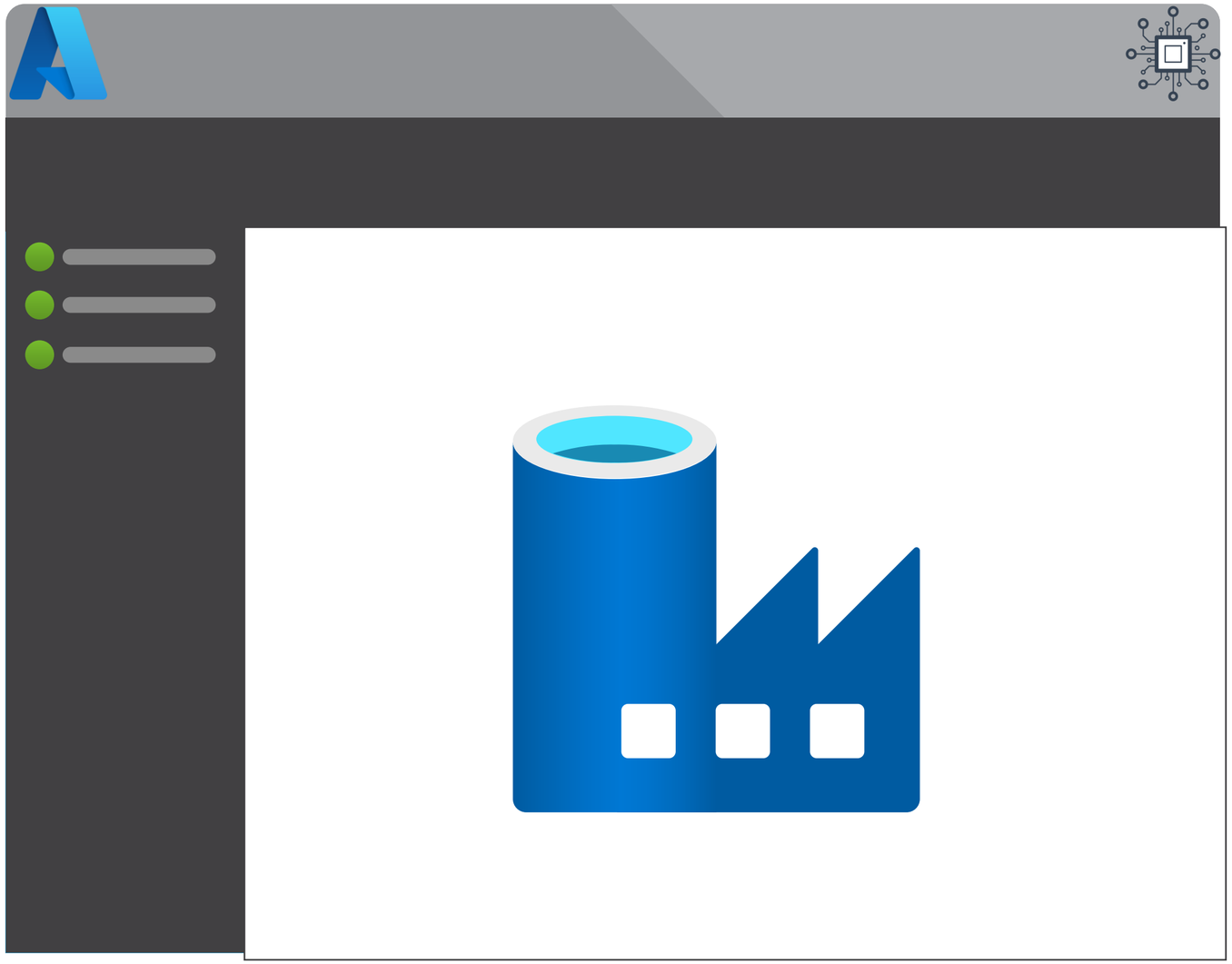
# An Introduction to Azure Data Factory

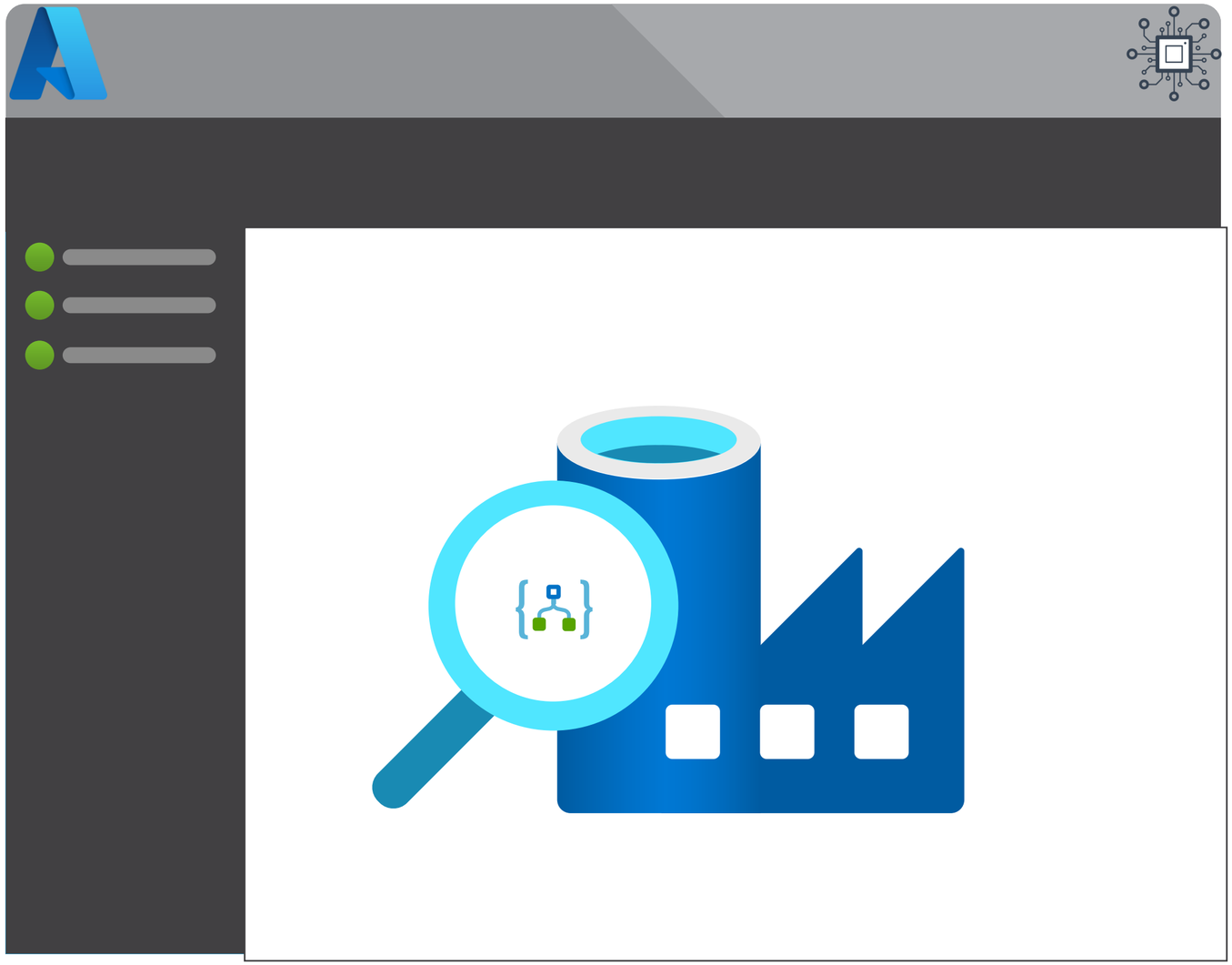
## Pipeline Fundamentals



- The History of Azure Orchestration
- Synapse Analytics vs Data Factory
- Integration Components
- Common Activities
- Execution Dependencies

# An Introduction to Azure Data Factory

## Pipeline Fundamentals



- The History of Azure Orchestration
- Synapse Analytics vs Data Factory
- Integration Components
- Common Activities
- Execution Dependencies

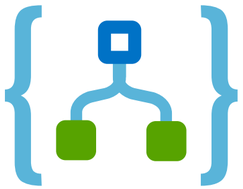
# A Quick History Lesson

SQL Server
SQL Agent

SQLDB
(PaaS)

Automation

Logic Apps



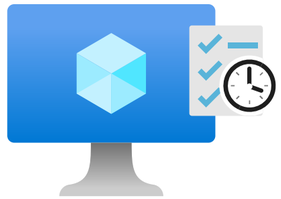
Functions

SQL Managed
Instance

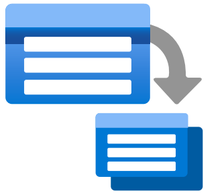
DTU Jobs

Elastic Job Agent

SQL Server
Virtual Machine



Virtual Machine
Job Schedule



Batch

# A Quick History Lesson

**v1**
July
2016

**v2**
Sept
2017

Oct
2020

Time Slices
VS2015 Projects

Notepad
Development

New UI

Source
Control

Data
Flows

Private
Endpoints

# What is Azure Data Factory (ADF)?

Home / Products / Data Factory

## Data Factory

Hybrid data integration service that simplifies ETL at scale

Start for free >

Already an Azure customer? Getting started >

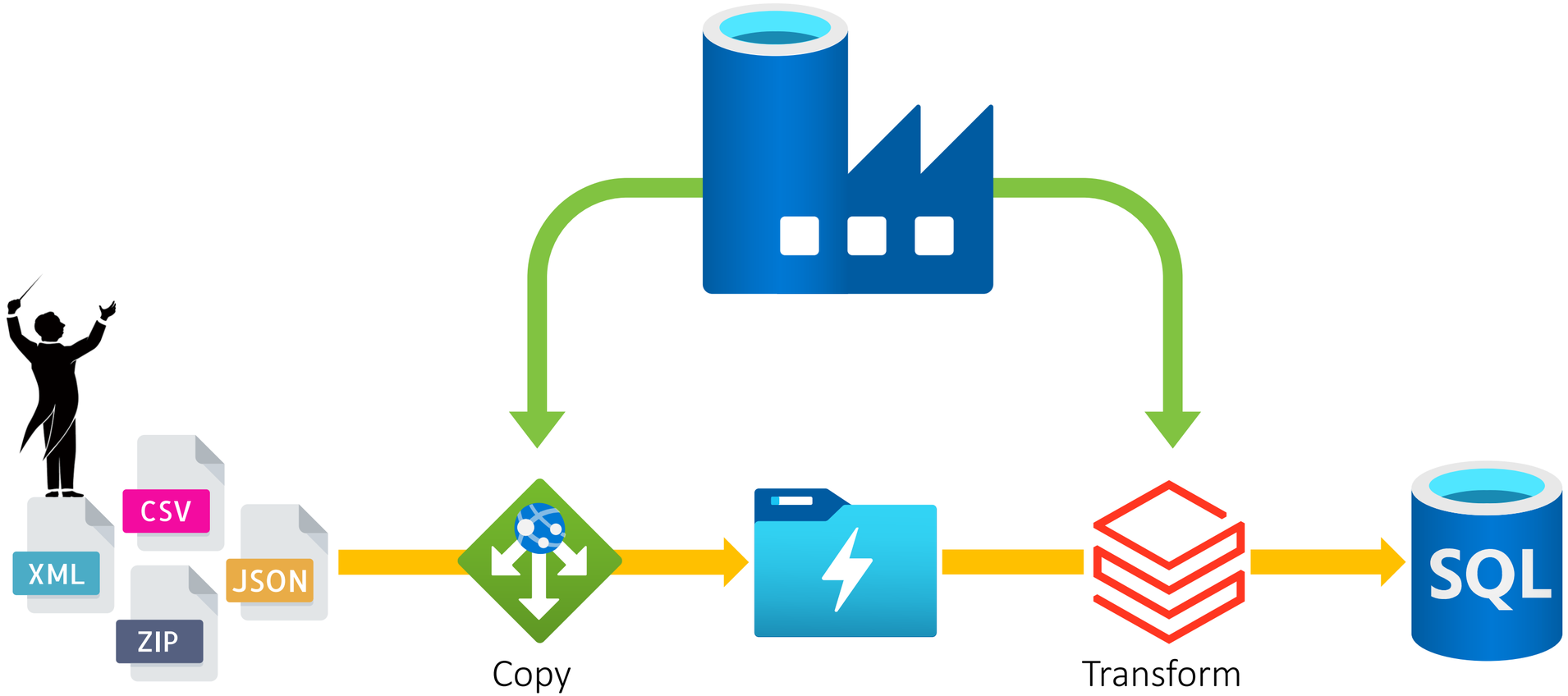
Product overview | Features | Security | Pricing | Customer stories | Getting started | Documentation | FAQs

### Accelerate data integration

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

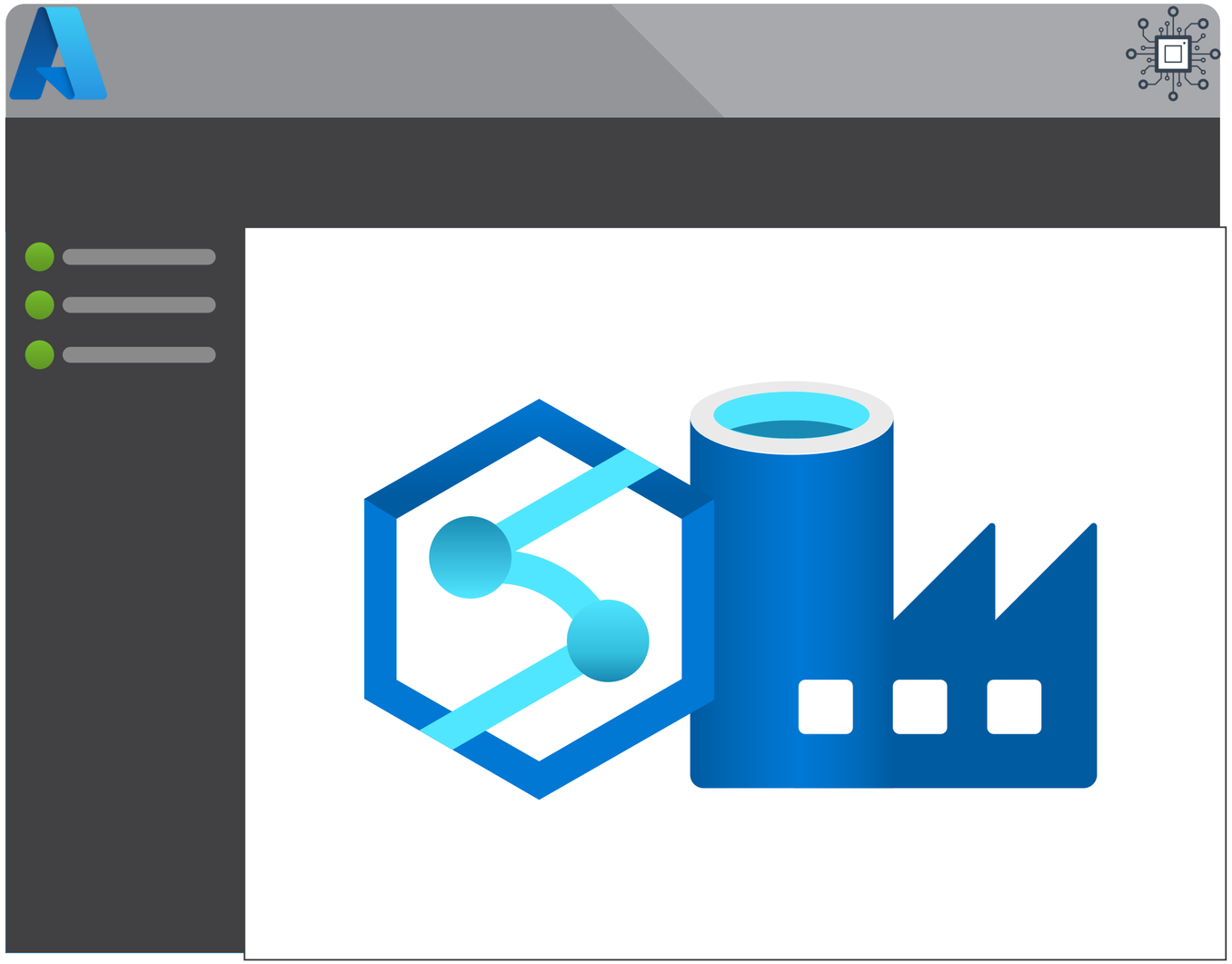
https://azure.microsoft.com/en-gb/services/data-factory

# What is Azure Data Factory (ADF)?



# An Introduction to Azure Data Factory

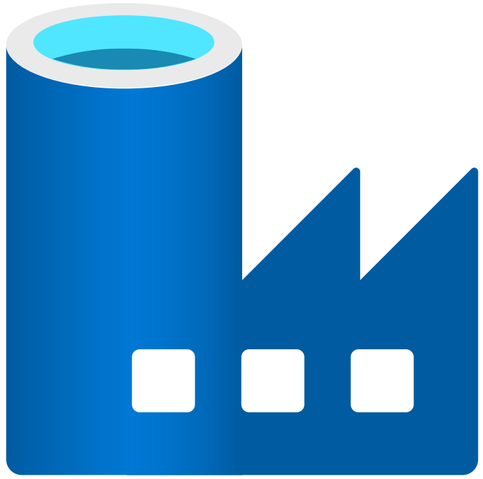
## Pipeline Fundamentals

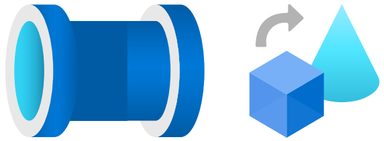


- The History of Azure Orchestration
- Synapse Analytics vs Data Factory
- Integration Components
- Common Activities
- Execution Dependencies

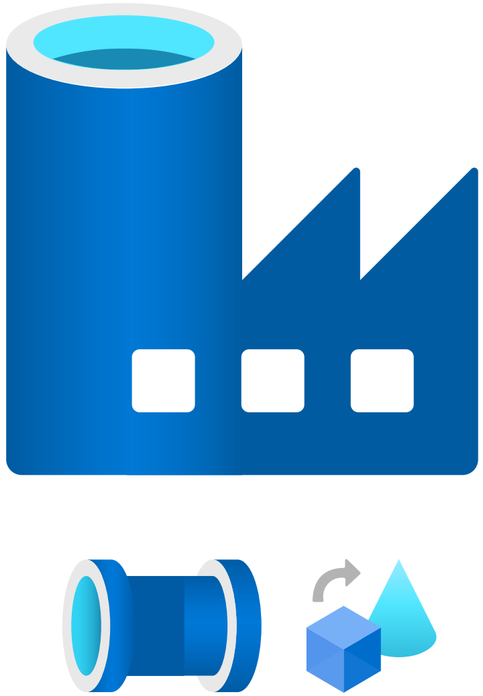
# Synapse Analytics vs Data Factory

https://docs.microsoft.com/en-us/azure/synapse-analytics/data-integration/concepts-data-factory-differences



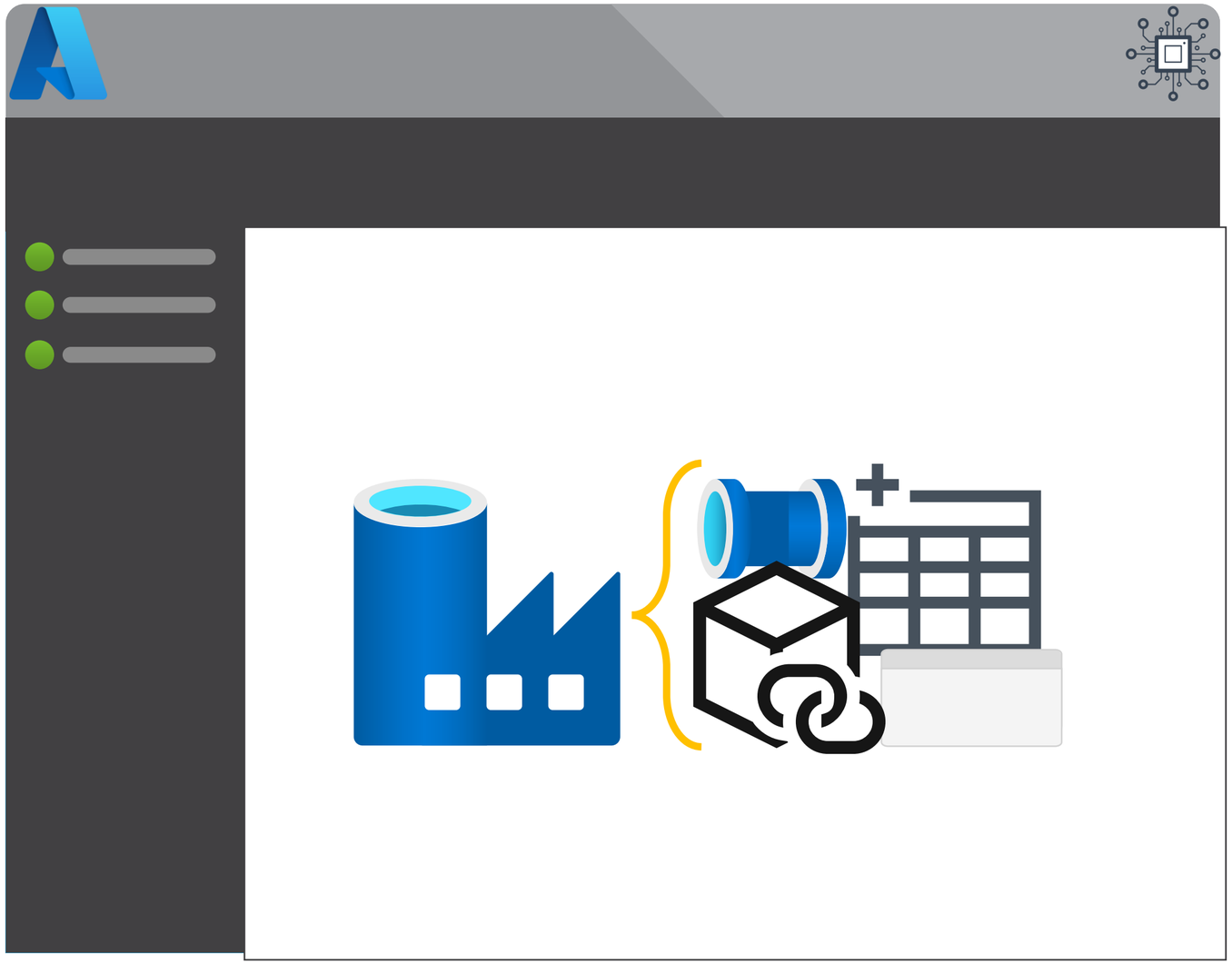


# Synapse Analytics vs Data Factory



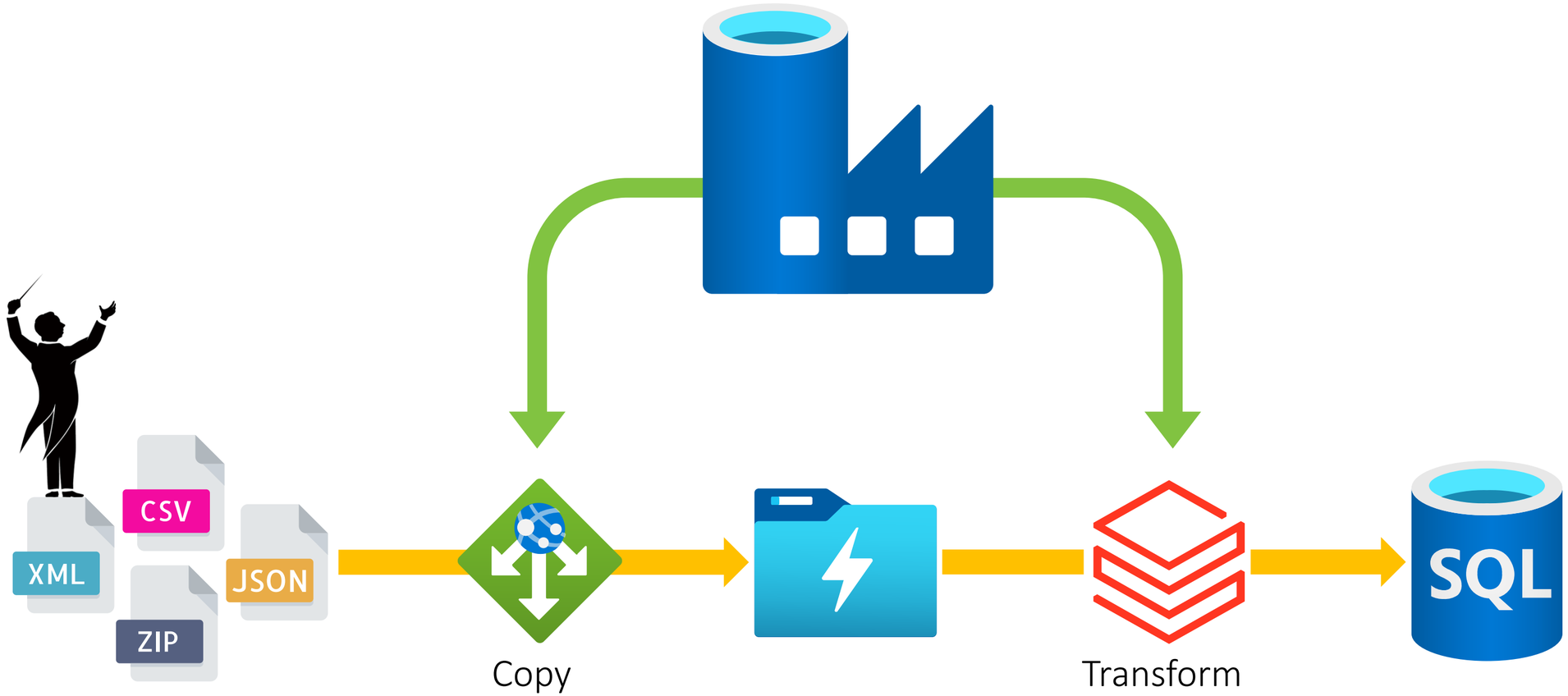
# An Introduction to Azure Data Factory

## Pipeline Fundamentals

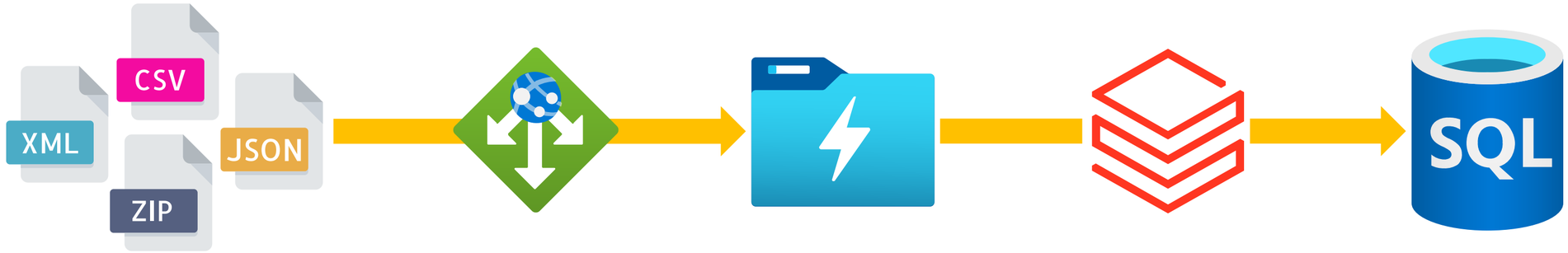


- The History of Azure Orchestration
- Synapse Analytics vs Data Factory
- Integration Components
- Common Activities
- Execution Dependencies

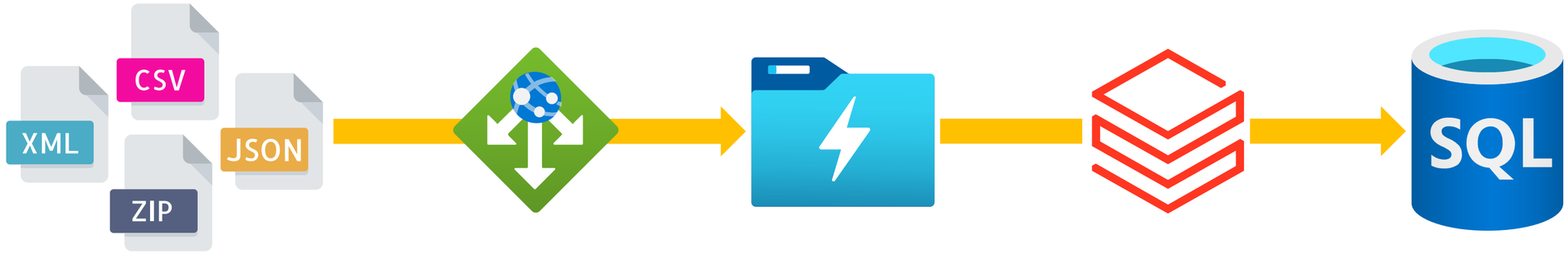
# Data Factory Components



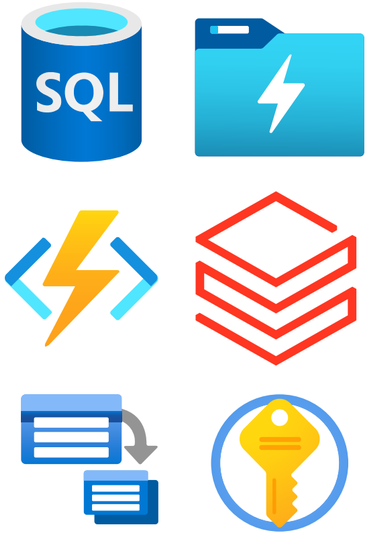
# Data Factory Components



# Data Factory Components




1 **Linked Services** – What to interact with and how?

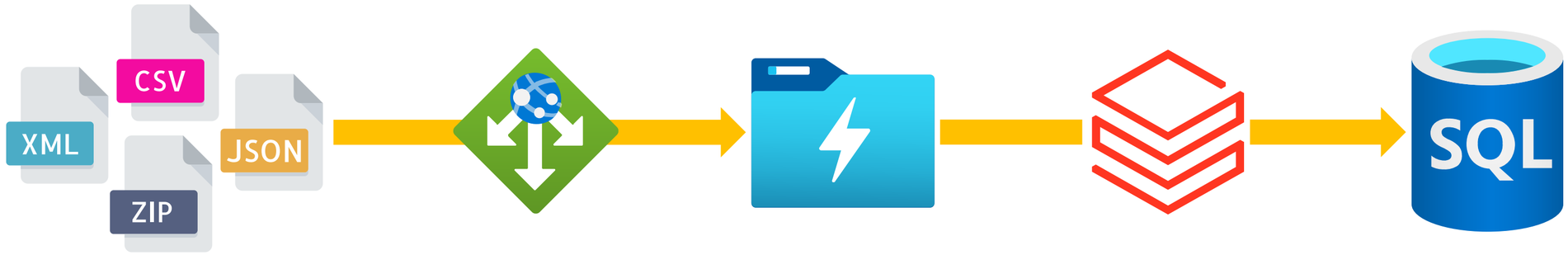



{;} JSON

SQLDBLinkedService

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

# Data Factory Components

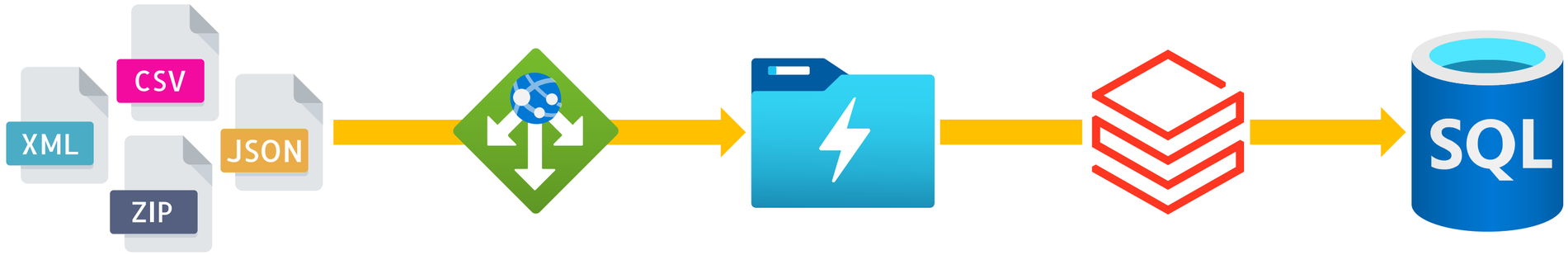



1. **Linked Services**
2. **Datasets** – Where is my data? What format? What file path/table do I need?

SQL [dbo].[SalesOrders]

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

# Data Factory Components



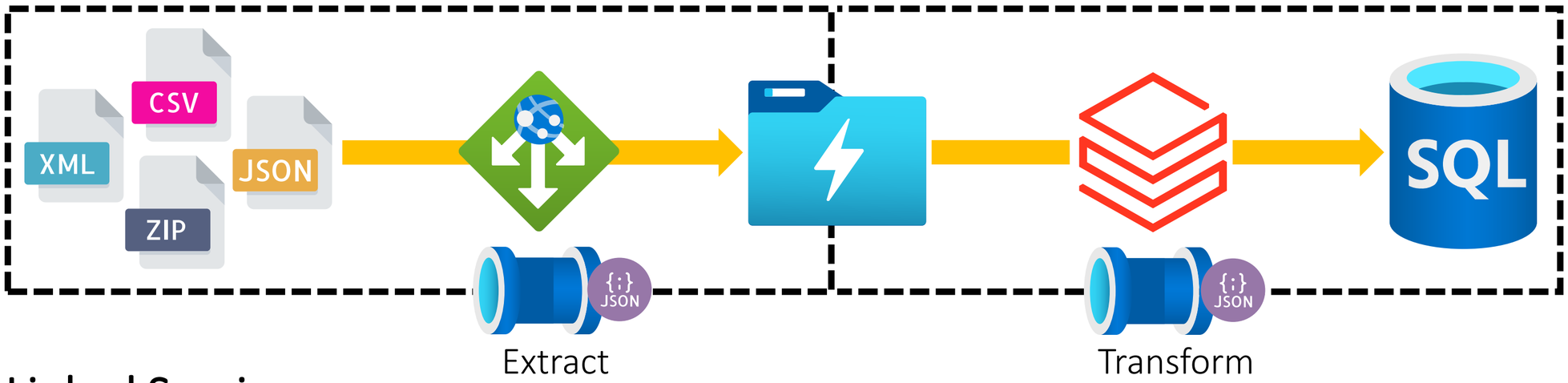

1. **Linked Services**
2. **Datasets**
3. **Activities** – What do we want to happen when we invoke a Linked Service?
   With what conditions?

{:} JSON

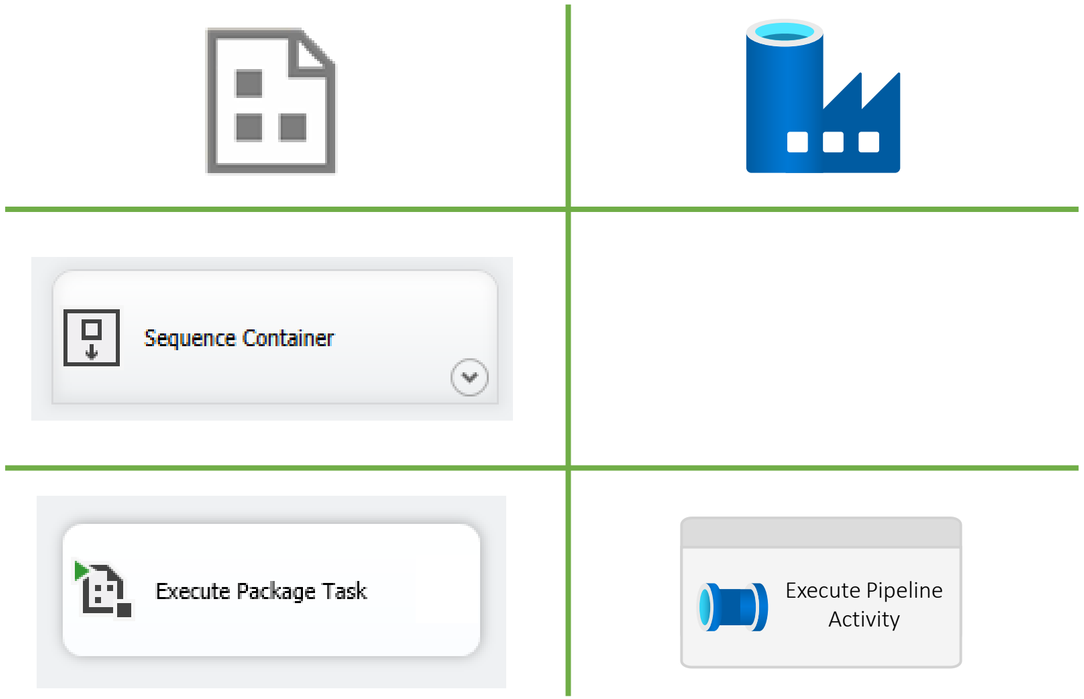
**Databricks Notebook Activity**

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

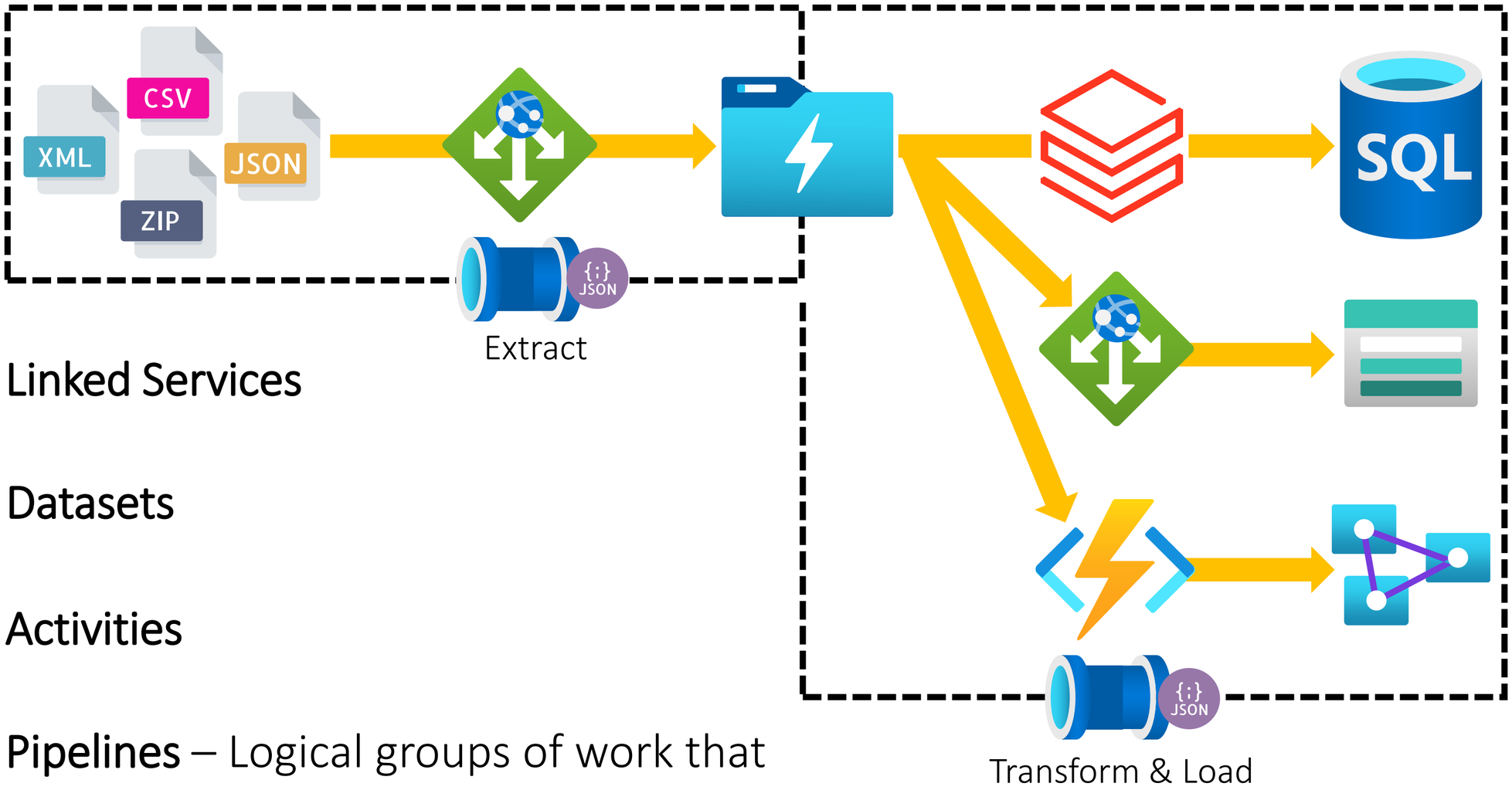
# Data Factory Components




1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines** – Logical groups of work that can be executed.

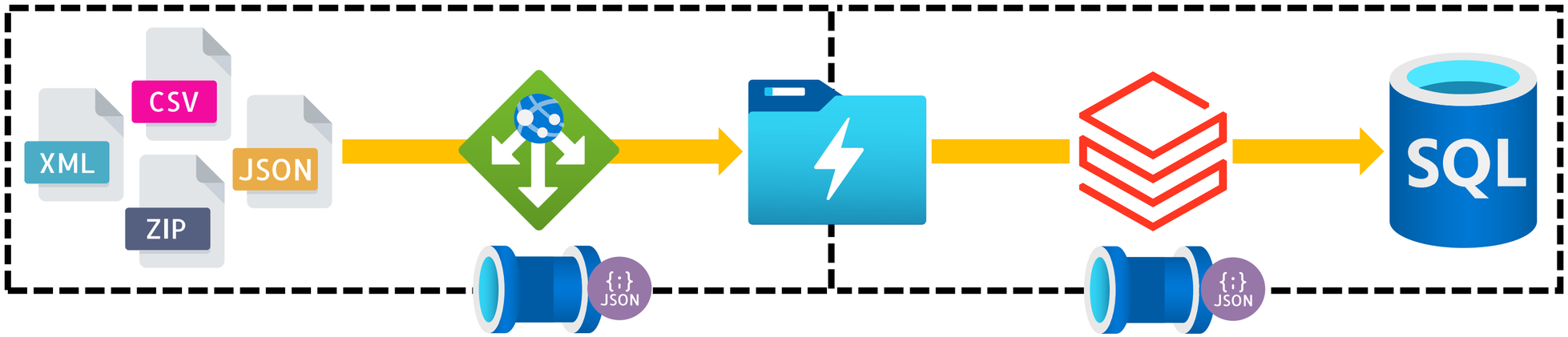


# Data Factory Components

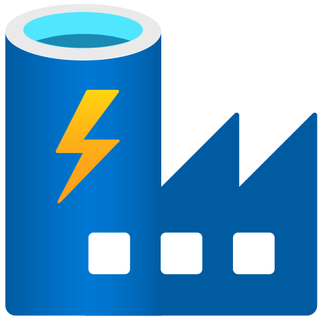



1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines** – Logical groups of work that can be executed.

# Data Factory Components

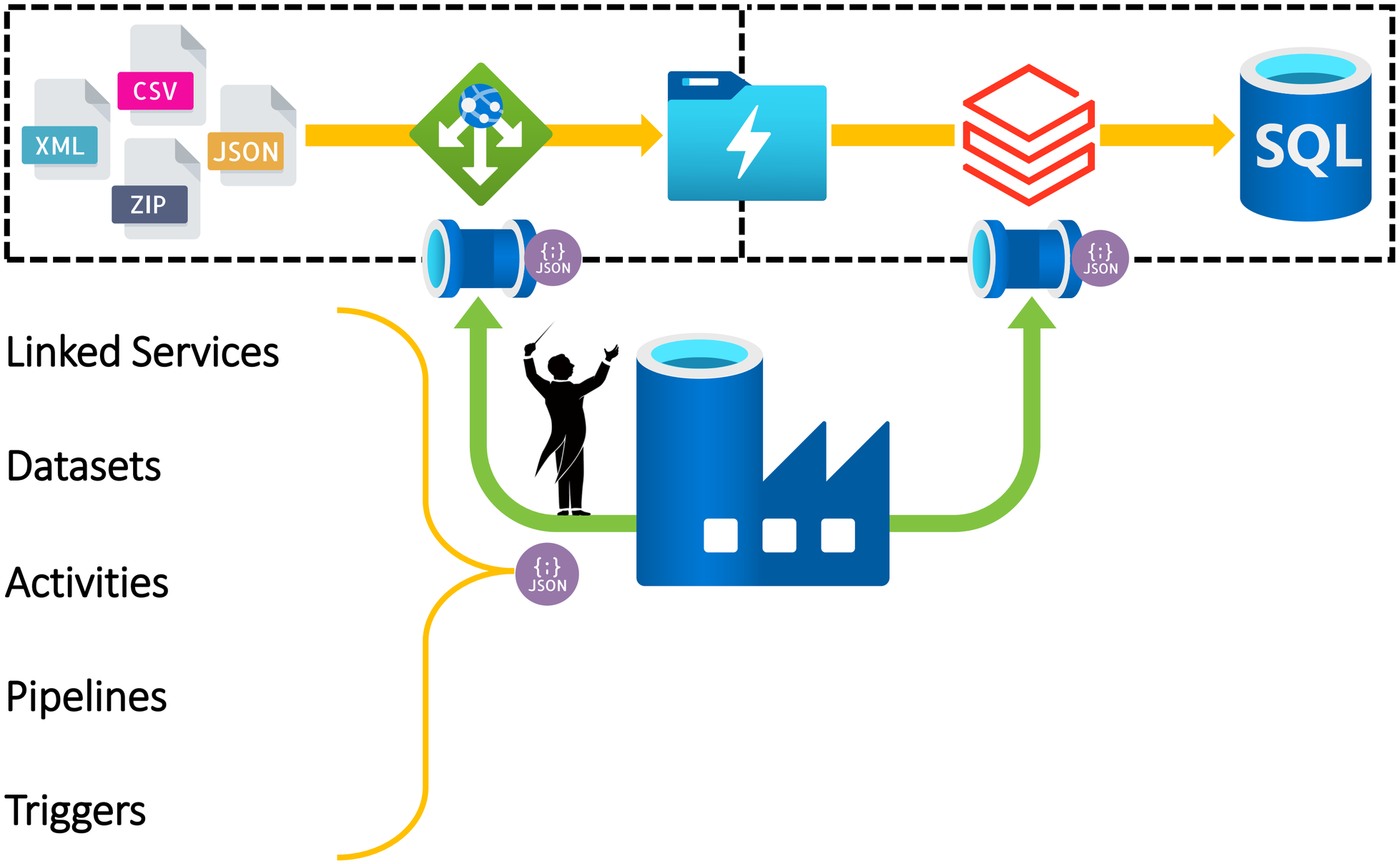



1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines**
5. **Triggers** – Telling our when pipelines to run.



- Manually
- Programmatically
- Schedule
- Tumbling Windows
- Storage Events
- Custom Events

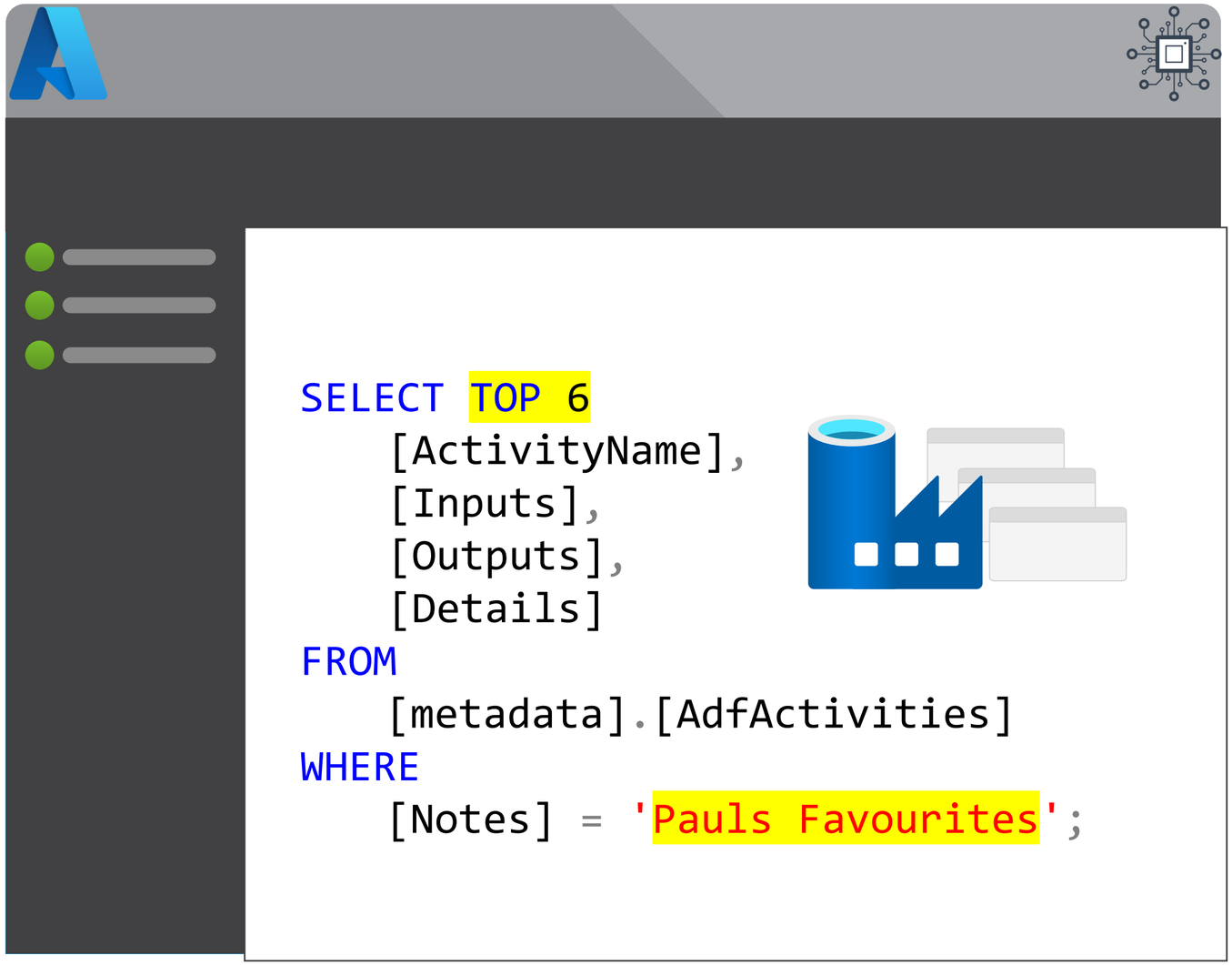
# Data Factory Components




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

# An Introduction to Azure Data Factory

## Pipeline Fundamentals

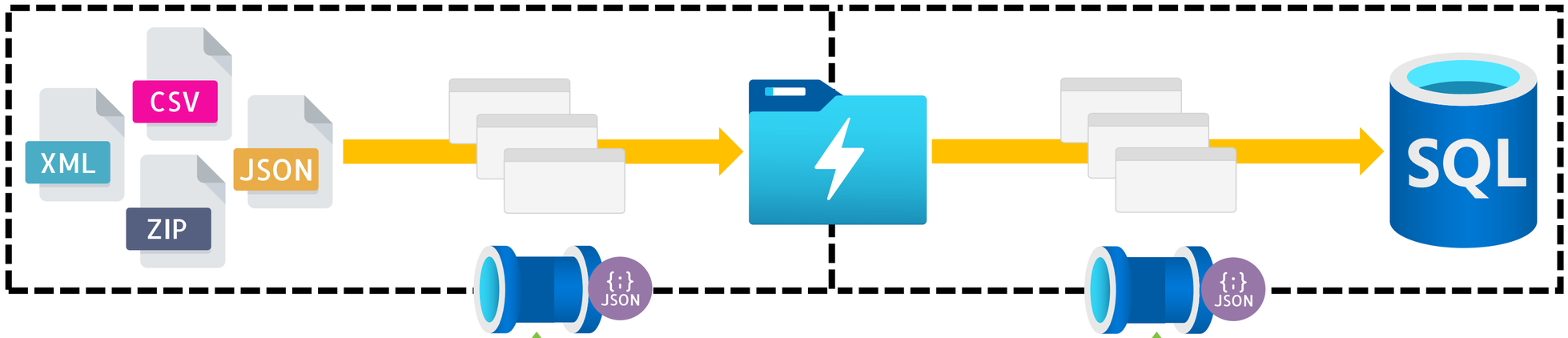


```
SELECT TOP 6
    [ActivityName],
    [Inputs],
    [Outputs],
    [Details]
FROM
    [metadata].[AdfActivities]
WHERE
    [Notes] = 'Pauls Favourites';
```

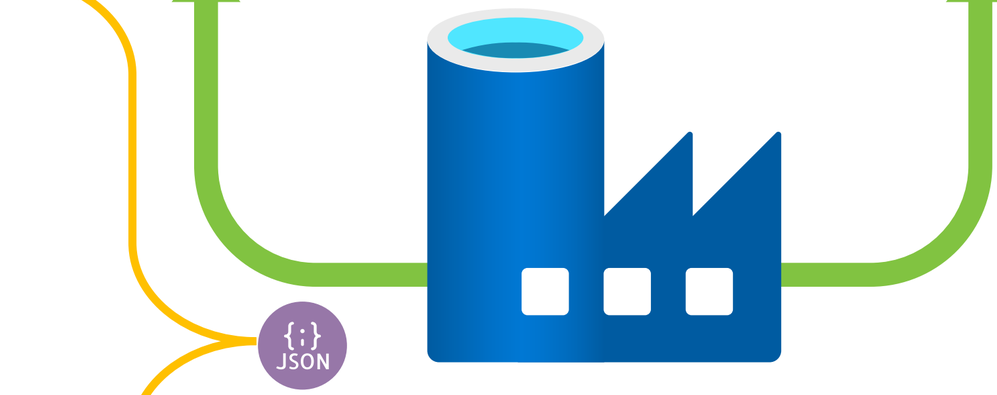


- The History of Azure Orchestration
- Synapse Analytics vs Data Factory
- Integration Components
- Common Activities
- Execution Dependencies

# Data Factory Common Activities

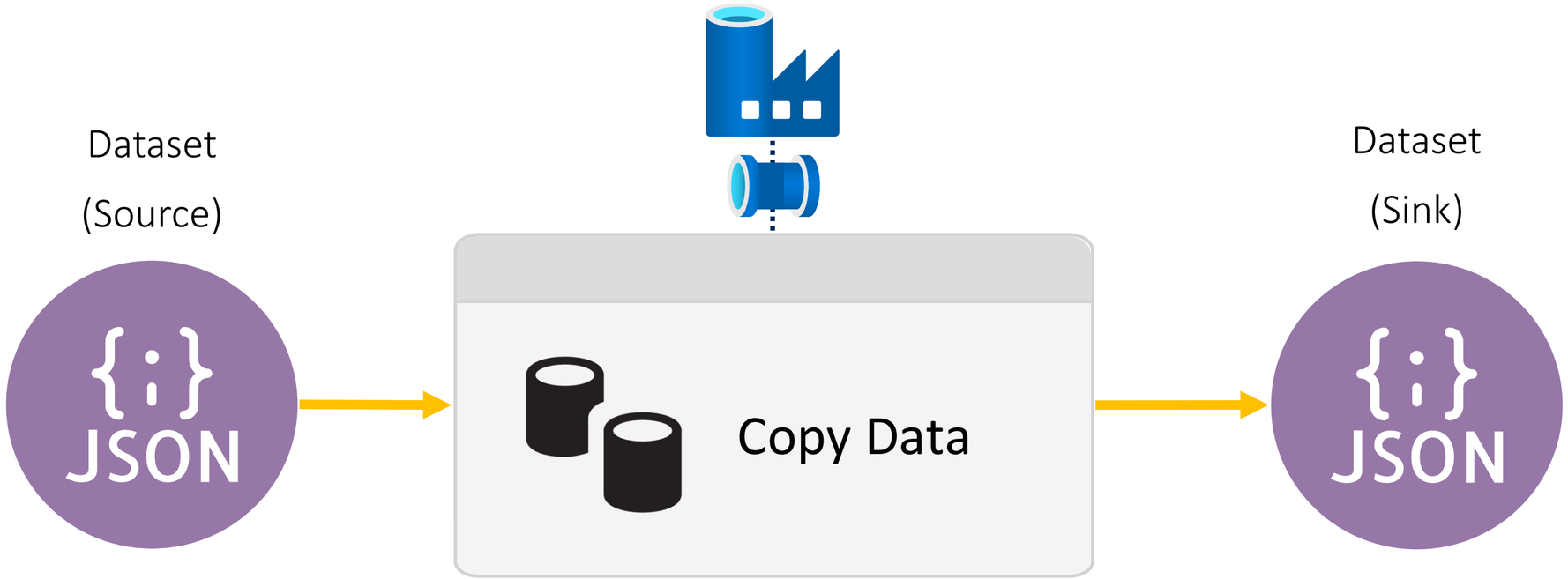



1. Linked Services
2. Datasets
3. Activities
4. Pipelines
5. Triggers



# Copy

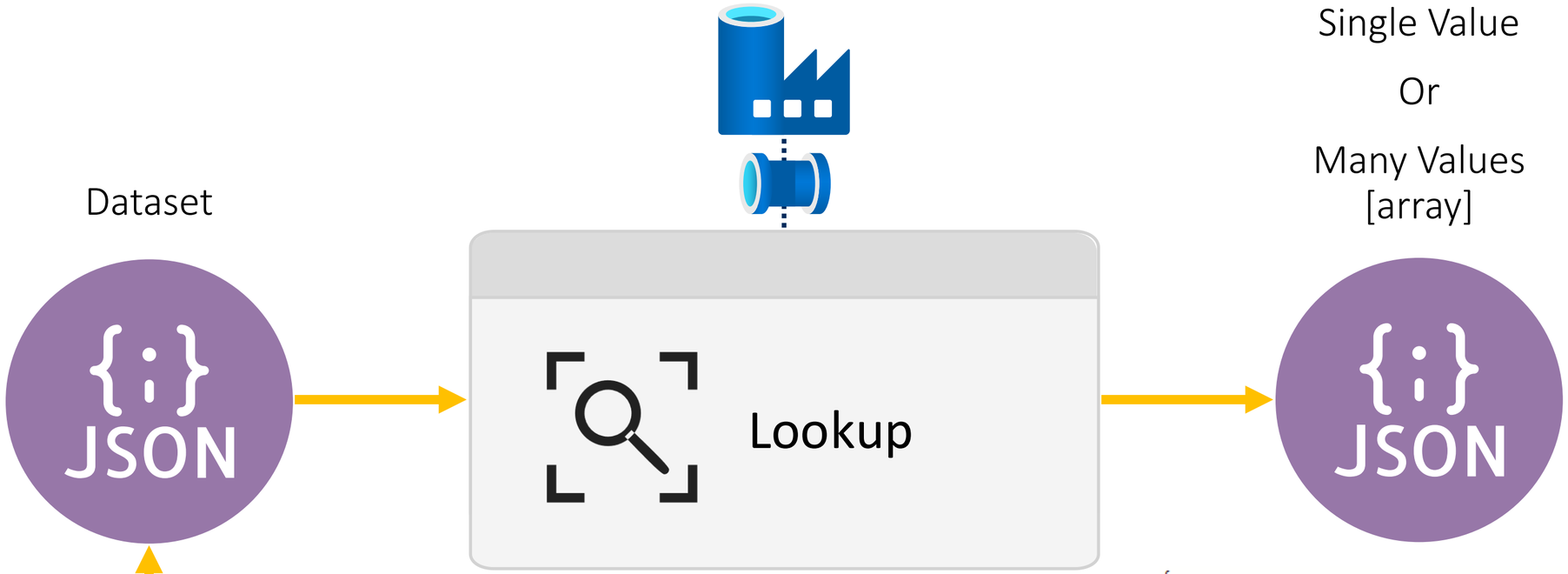
## Getting your data from A to B (not a Move operation)




- Auto Scaling
- Transactional Restarts
- Handle Zip Compression
- Attribute Mapping and Schema Drift
- Handle Failed Rows
- Add Custom Attributes
- Parse Excel & JSON Files

# Lookup

## Get value(s) to support other control flow activities



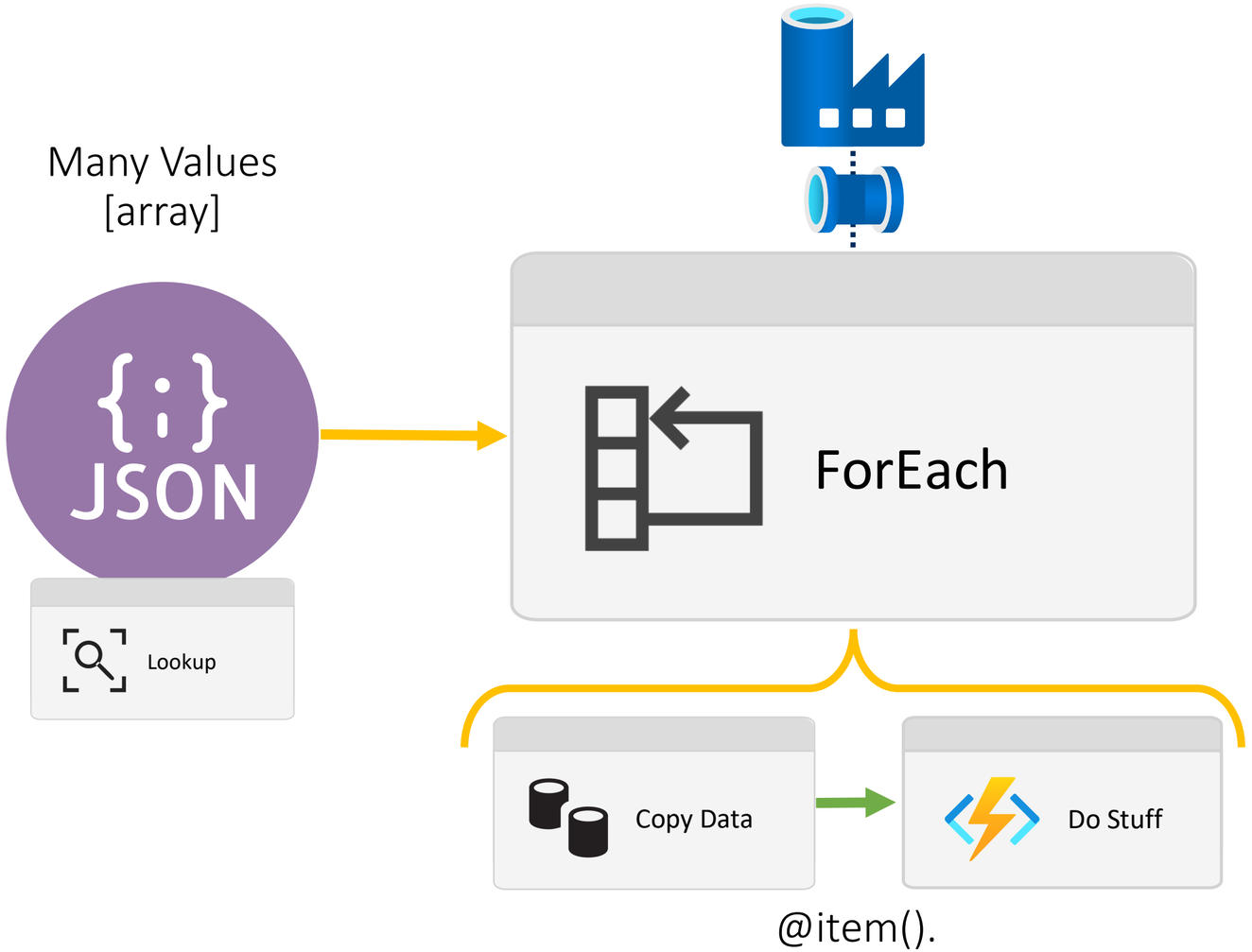
```
SELECT
    [SourceDIR],
    [TargetDIR],
    [FileName]
FROM
    [dbo].[FileList]
```

```
{
    "count": 3,
    "value": [
        {
            "SourceDIR": "ADFRoot\\ForUpload\\People\\",
            "TargetDIR": "RAW",
            "FileName": "Address.csv"
        },
        {
            "SourceDIR": "ADFRoot\\ForUpload\\People\\",
            "TargetDIR": "RAW",
            "FileName": "Gender.csv"
        },
        {
            "SourceDIR": "ADFRoot\\ForUpload\\People\\",
            "TargetDIR": "RAW",
            "FileName": "Ids.csv"
        }
    ]
}
```

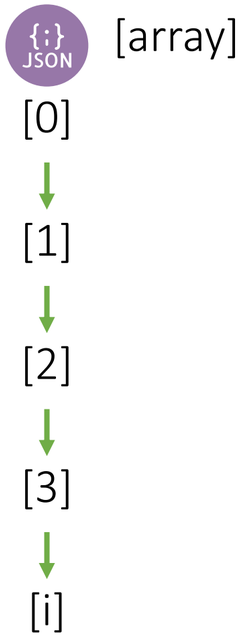
https://docs.microsoft.com/en-us/azure/data-factory/control-flow-lookup-activity

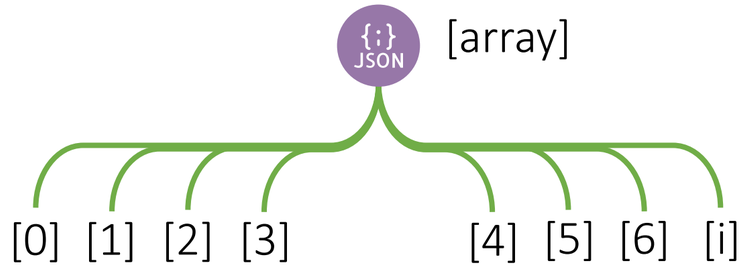
# For Each

## Iterating over other control flow activities




IsSequential: true



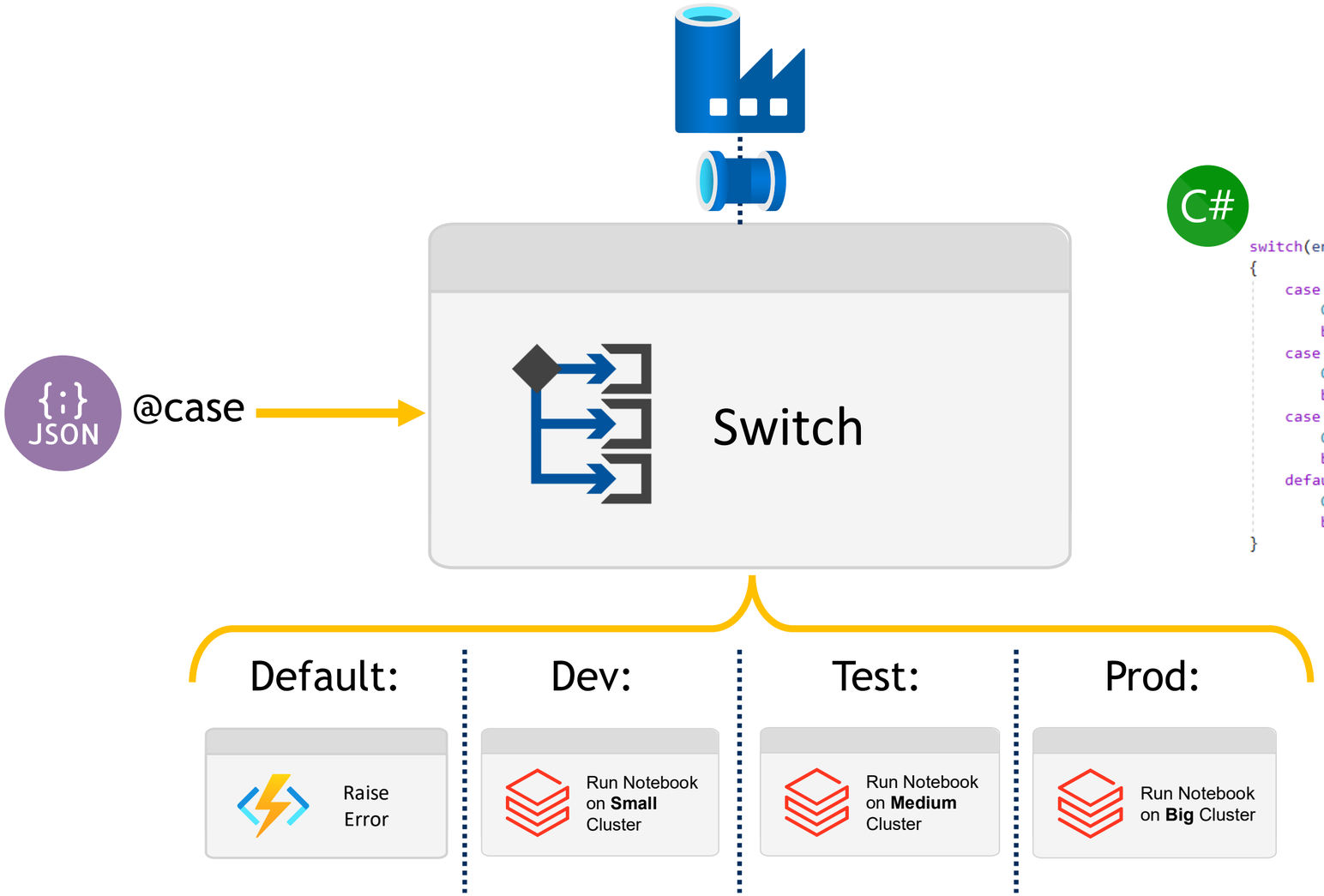

Batch Count Default: 20

Batch Count Max: 50

https://docs.microsoft.com/en-us/azure/data-factory/control-flow-for-each-activity

# Switch

## Execute other control flow components based on a provided condition



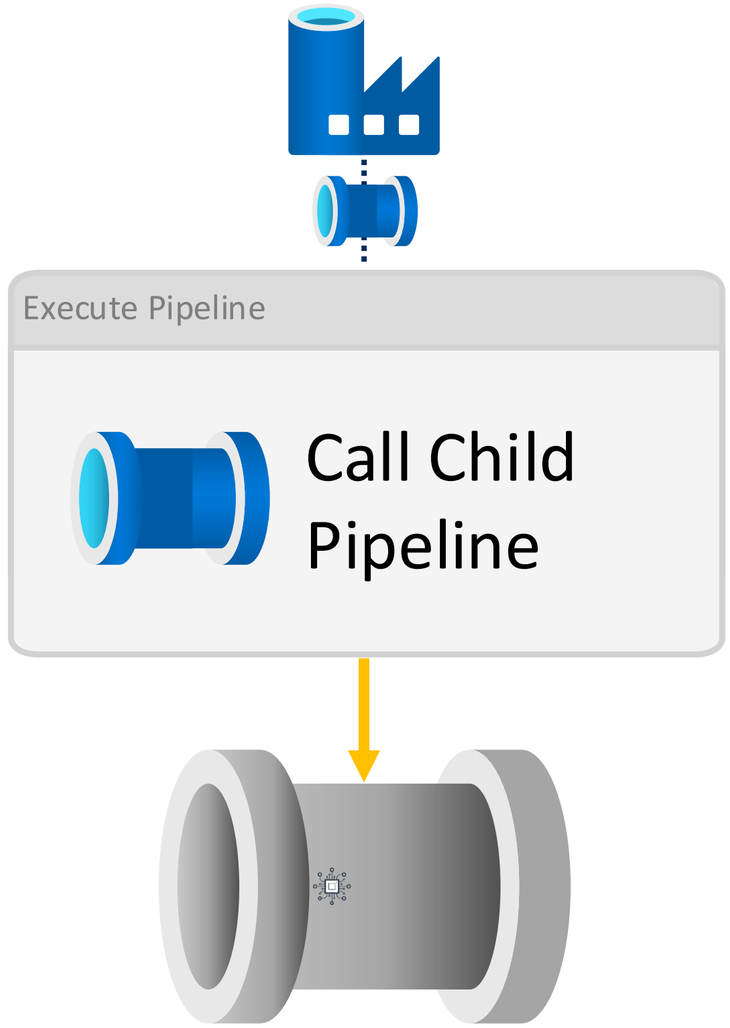

C#

```
switch(environmentValue)
{
    case "Dev":
        Console.WriteLine("Running in development envirnoment.");
        break;
    case "Test":
        Console.WriteLine("Running in testing envirnoment.");
        break;
    case "Prod":
        Console.WriteLine("Running in production envirnoment.");
        break;
    default:
        Console.WriteLine("Envirnoment unknown.");
        break;
}
```

https://mrpaulandrew.com/2020/01/22/using-the-azure-data-factory-switch-activity/

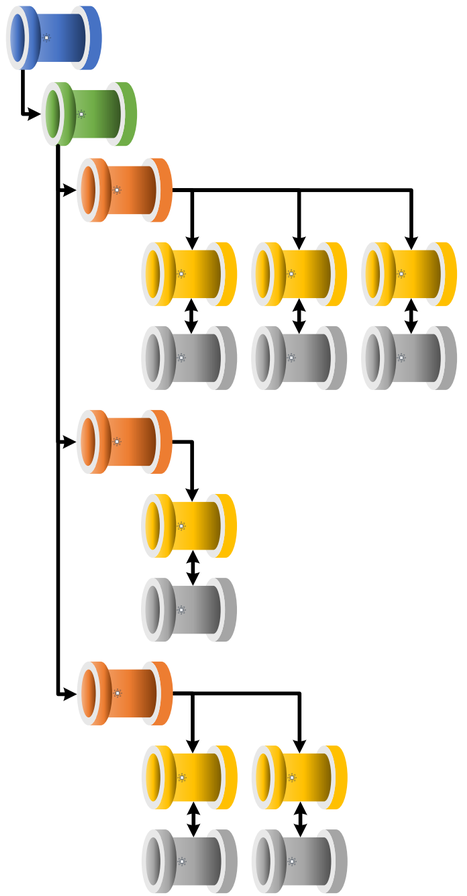
# Execute Pipeline

## Chaining pipeline executions via an activity



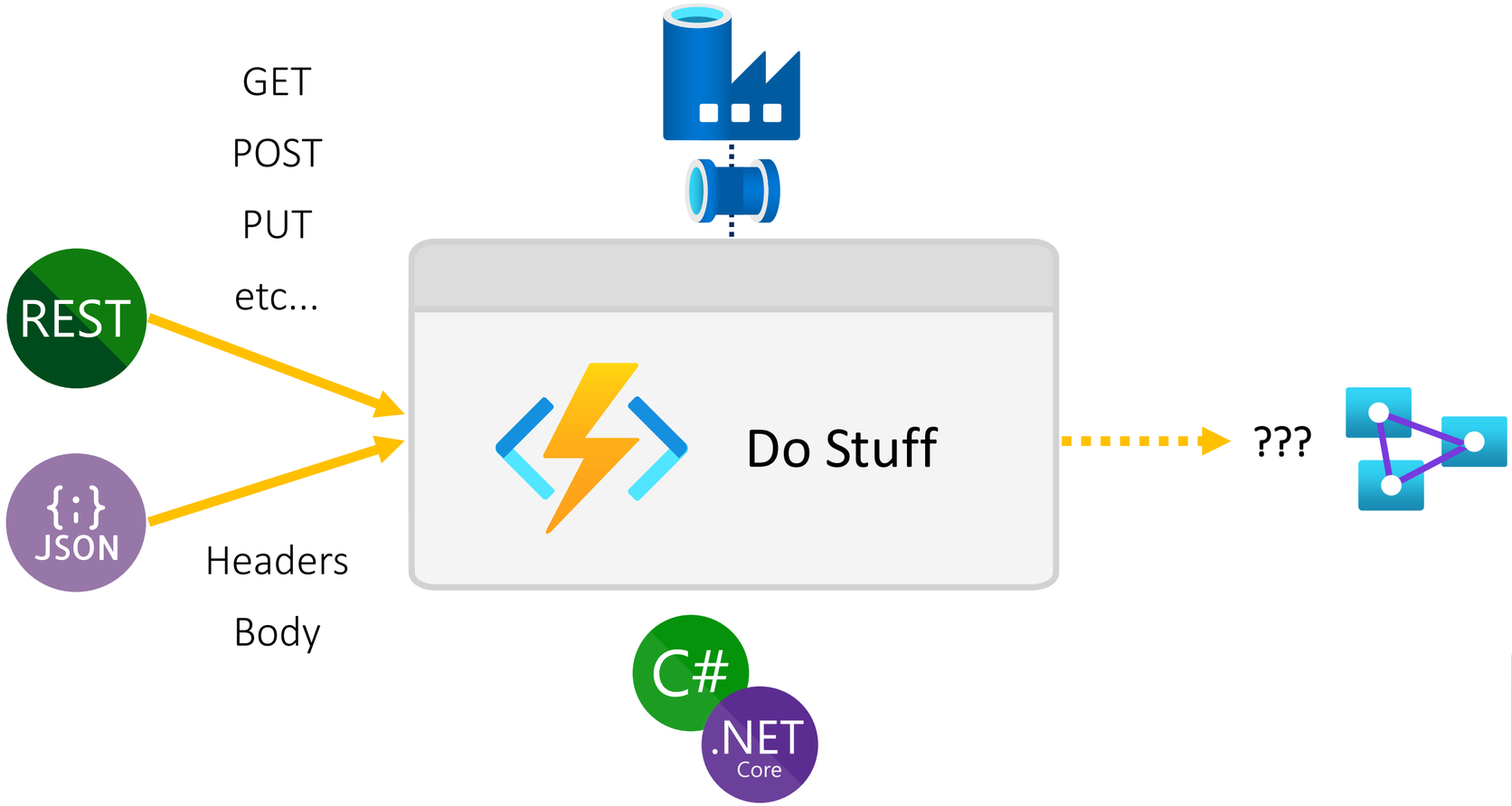

**Pipeline Hierarchies Generation Control**

https://mrpaulandrew.com/2019/09/25/azure-data-factory-pipeline-hierarchies-generation-control



# Azure Function

Extend Data Factory with custom serverless code executions via REST calls




Web

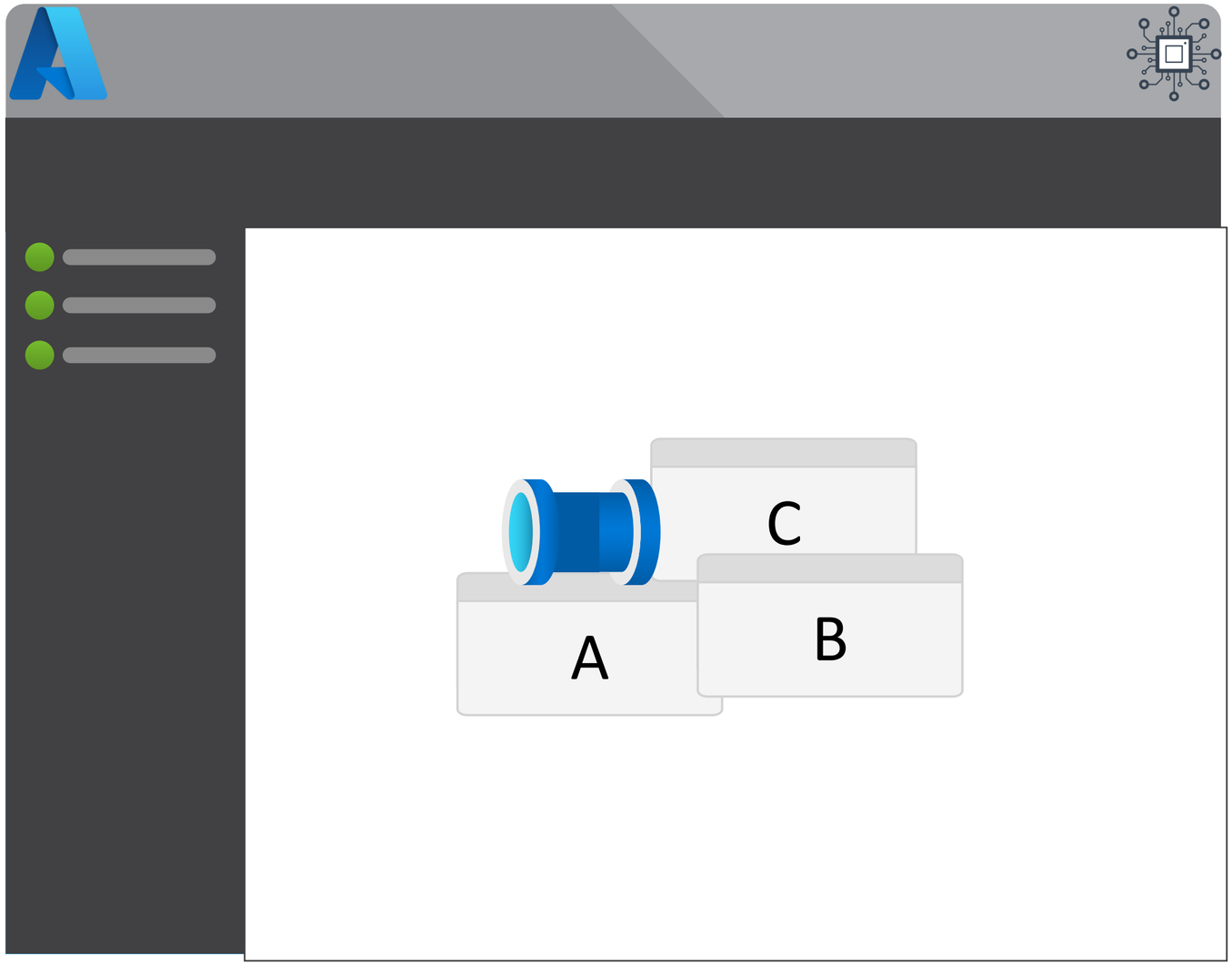
HTTP Do Stuff

Web Hook

Do Stuff

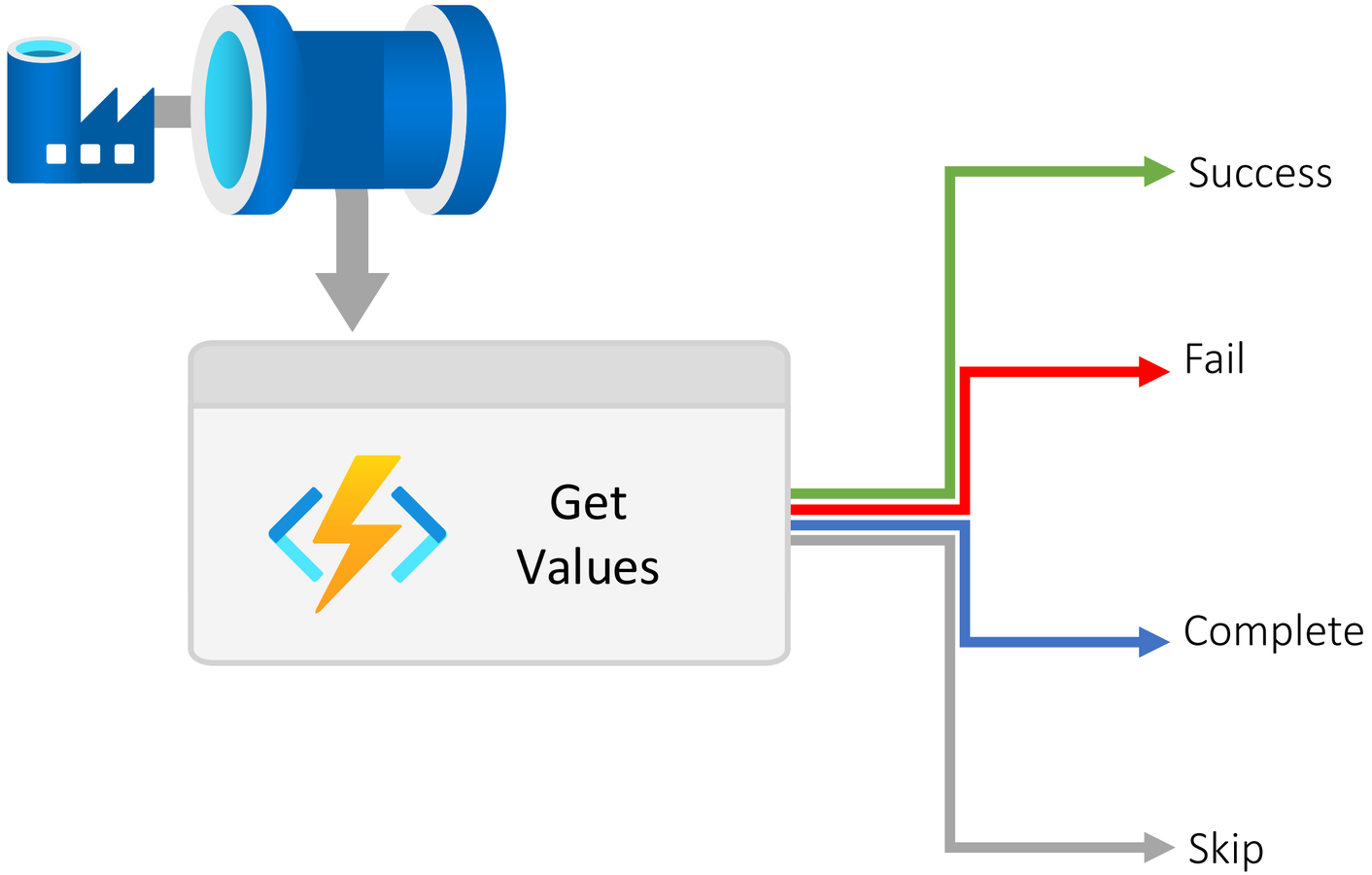
# An Introduction to Azure Data Factory

## Pipeline Fundamentals

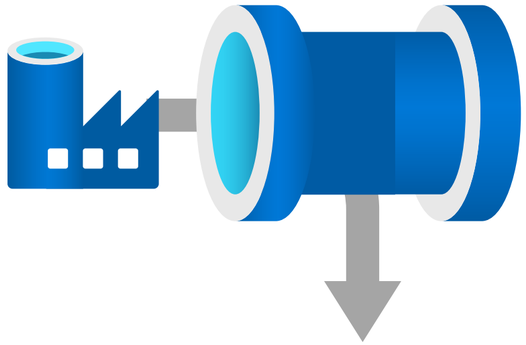


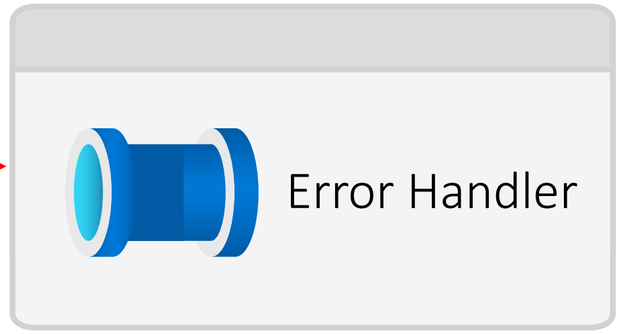

- The History of Azure Orchestration
- Synapse Analytics vs Data Factory
- Integration Components
- Common Activities
- Execution Dependencies

# Execution Dependency Options

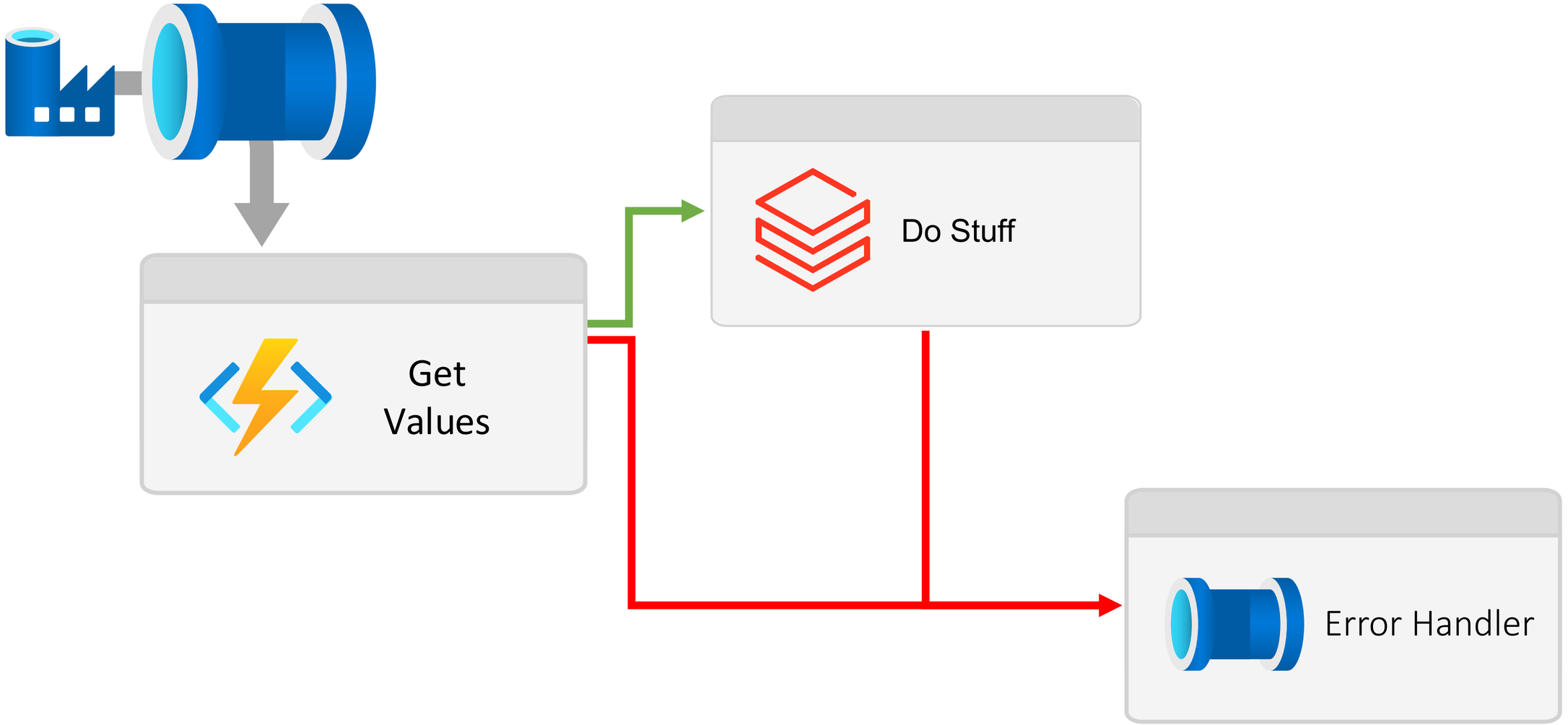


# Execution On Failure

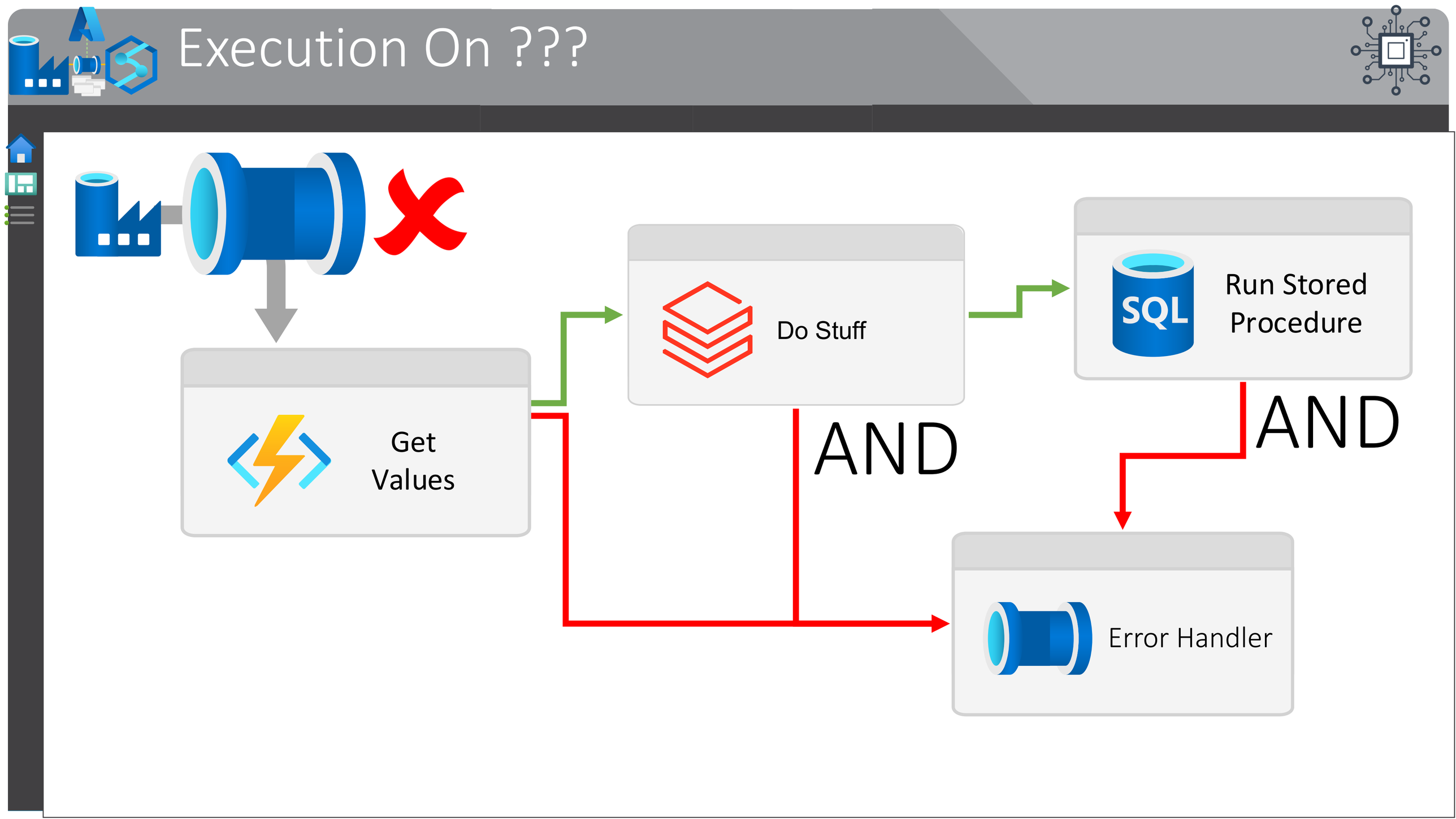




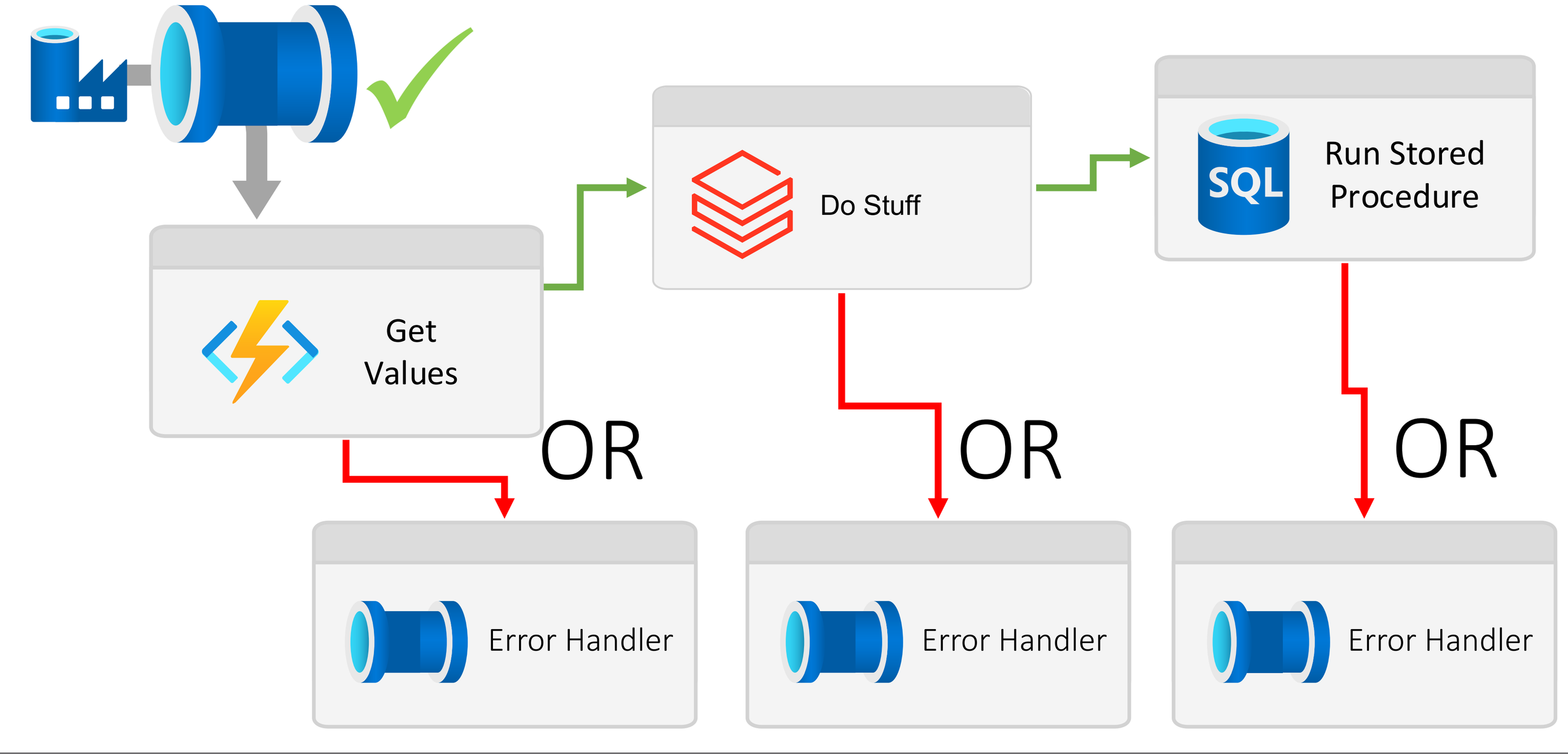
# Execution On Failure or On Success



# Execution On ???

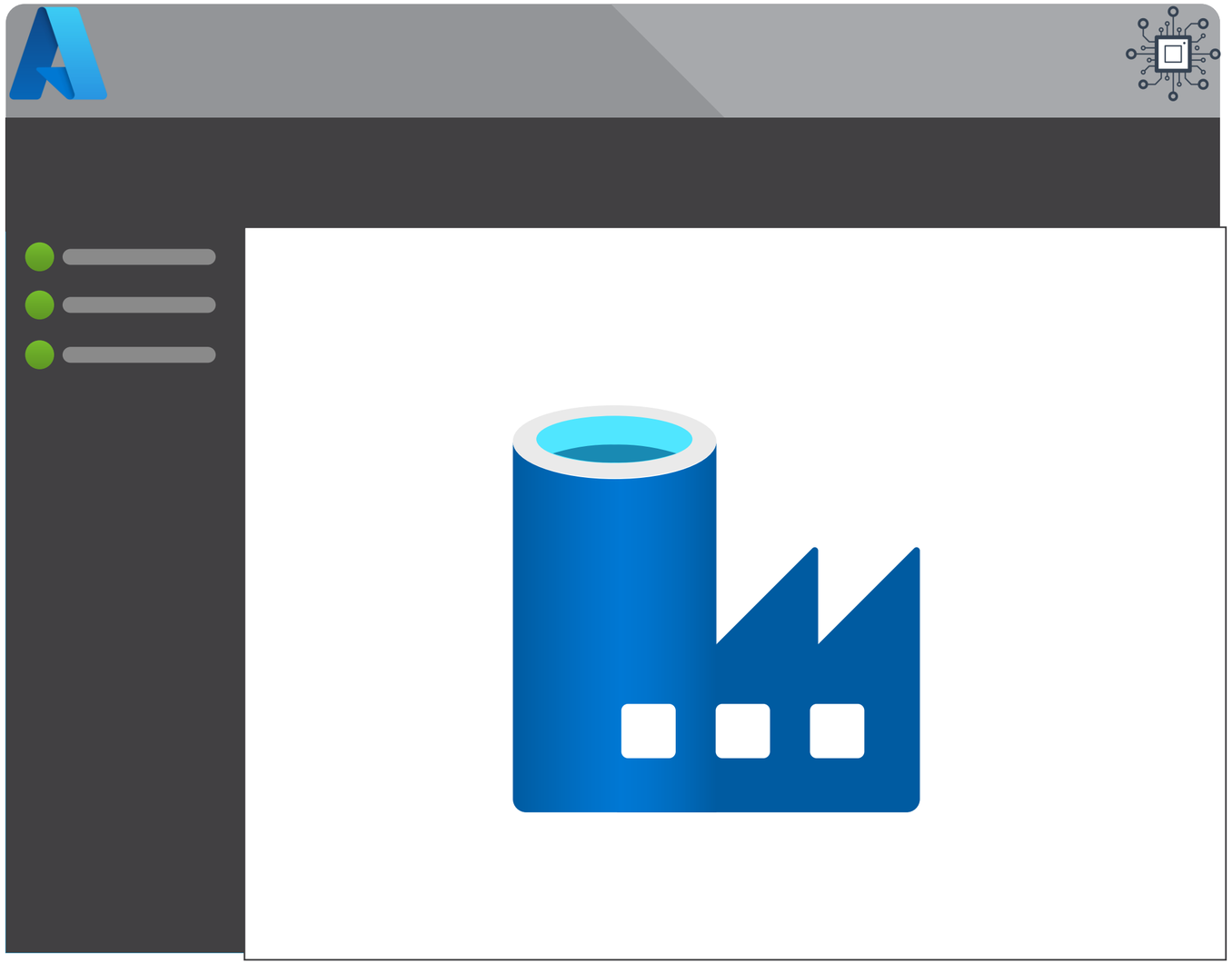


# Execution On Failure or On Success



# An Introduction to Azure Data Factory

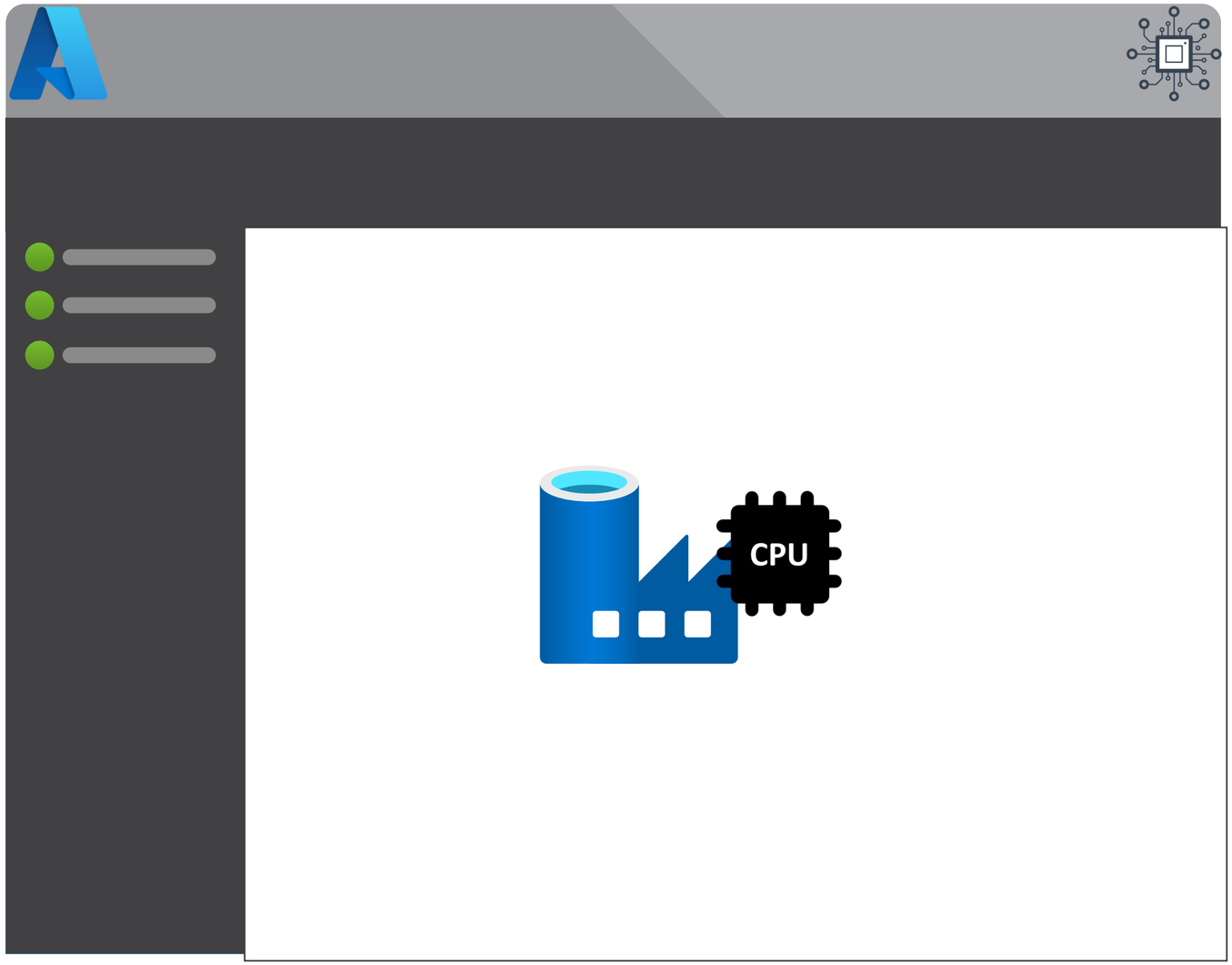
## Pipeline Fundamentals ✓



- The History of Azure Orchestration
- Synapse Analytics vs Data Factory
- Integration Components
- Common Activities
- Execution Dependencies

# An Introduction to Azure Data Factory

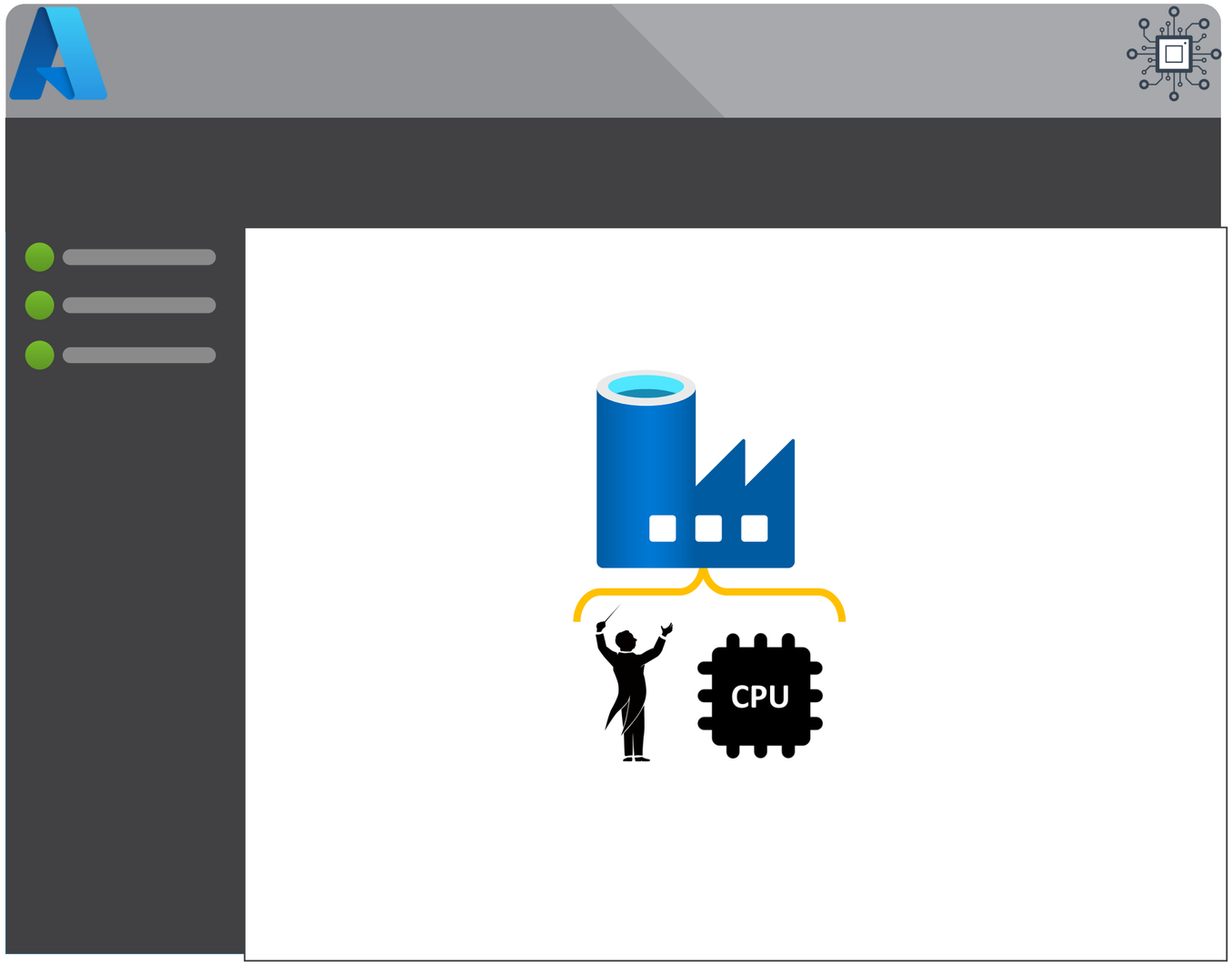
## Integration Runtimes




- Compute Types
  - Azure
  - Hosted
  - SSIS
- Patterns & Configuration

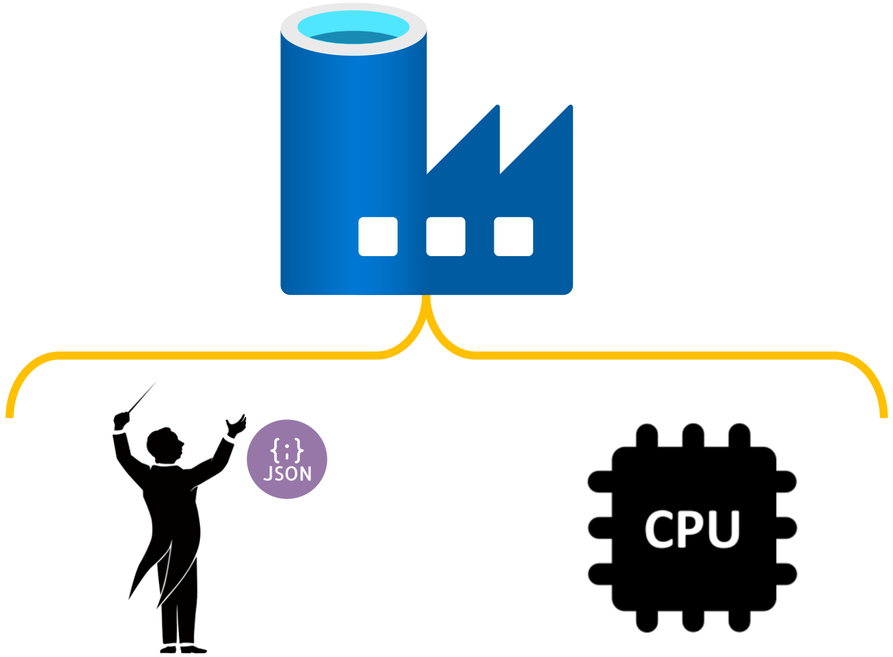
# An Introduction to Azure Data Factory

## Integration Runtimes




- Compute Types
  - Azure
  - Hosted
  - SSIS
- Patterns & Configuration

# What is an Integration Runtime?

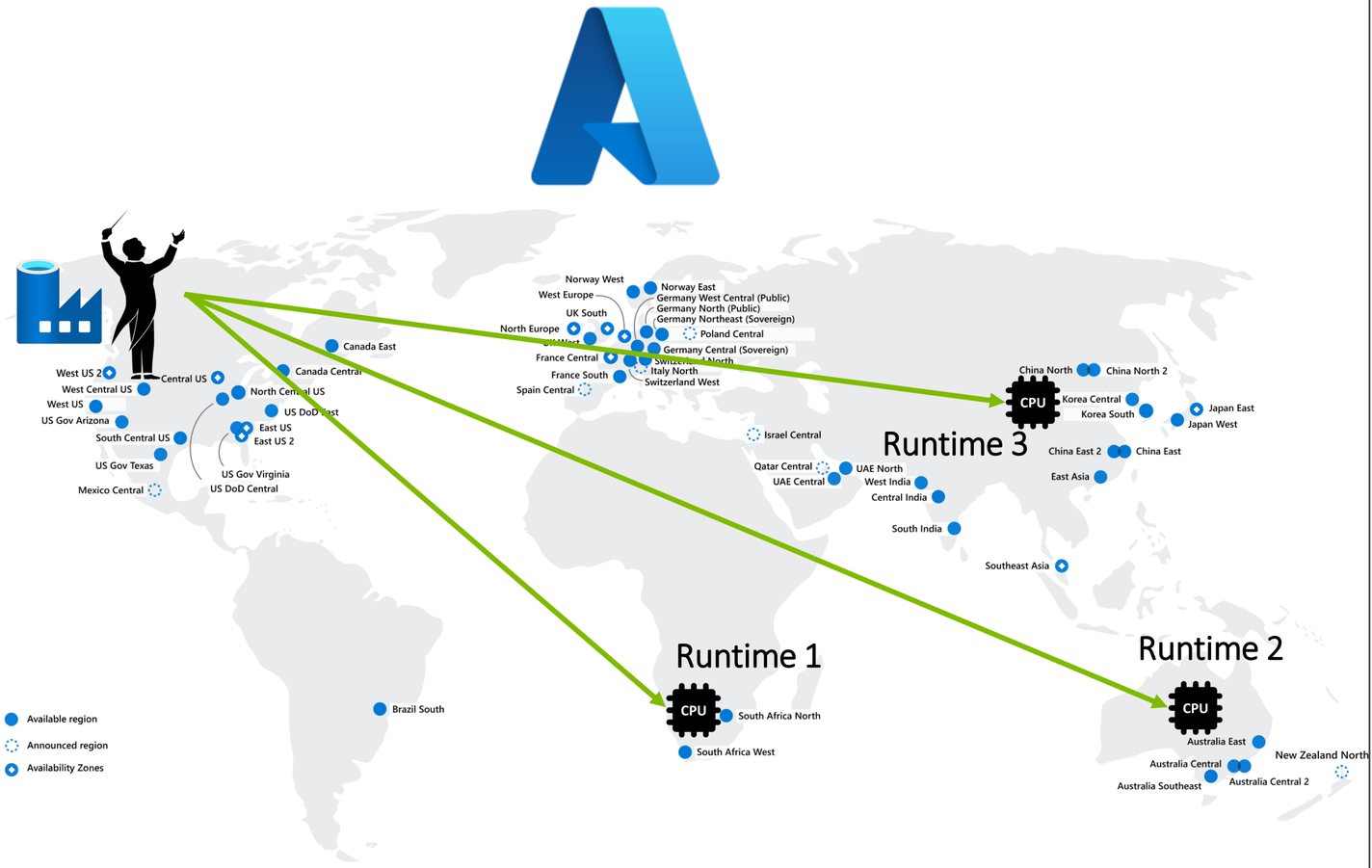



Orchestrator

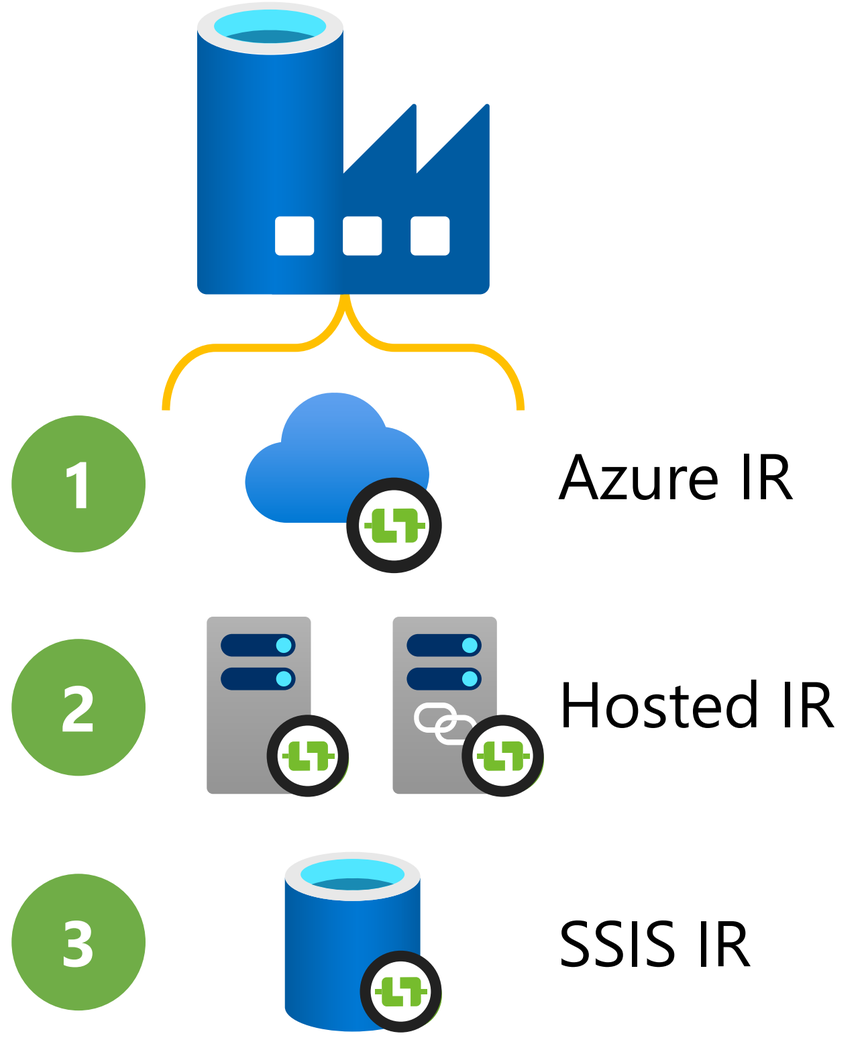
Runtime

Fixed Region

Flexible Region

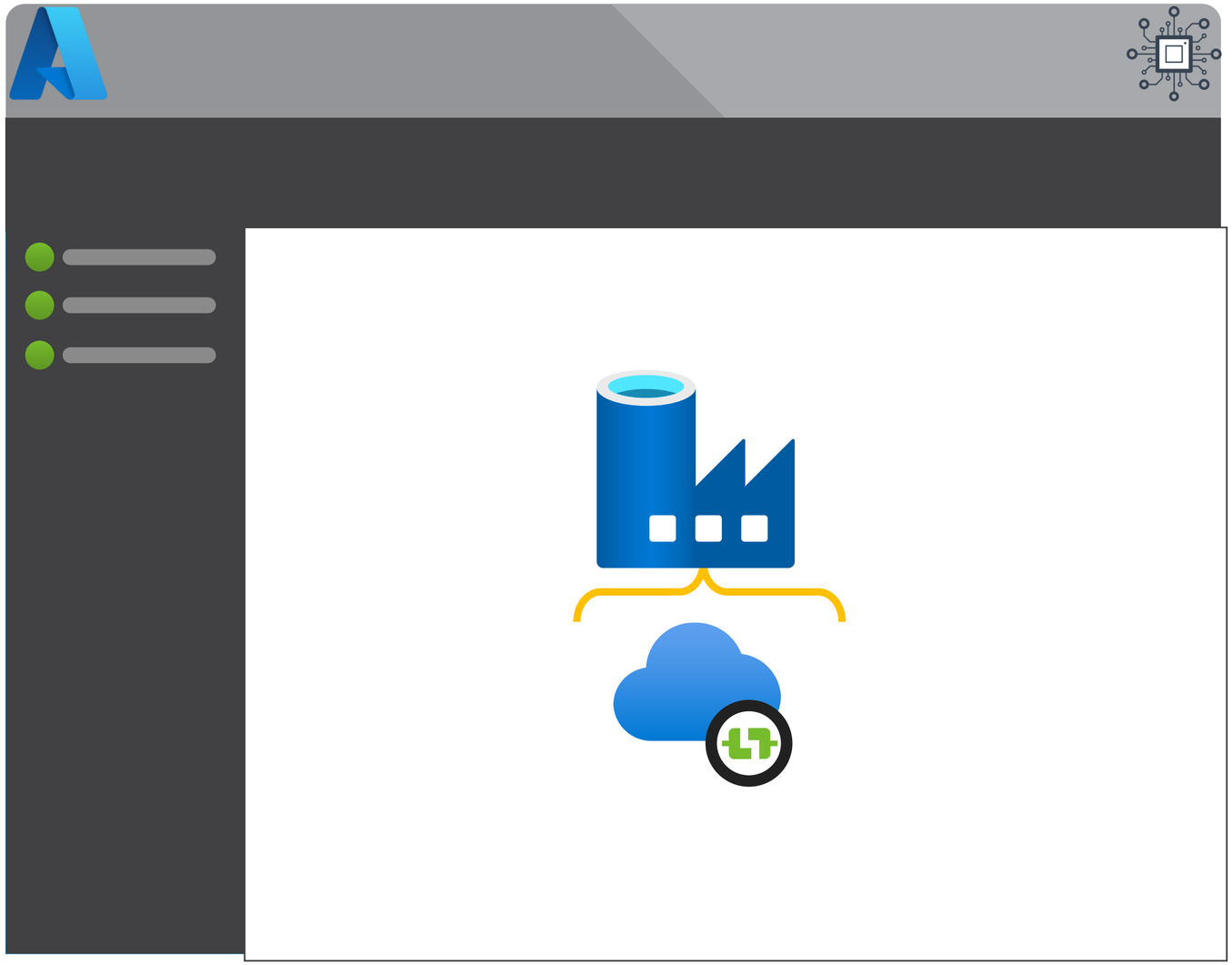


# What can an Integration Runtime do?



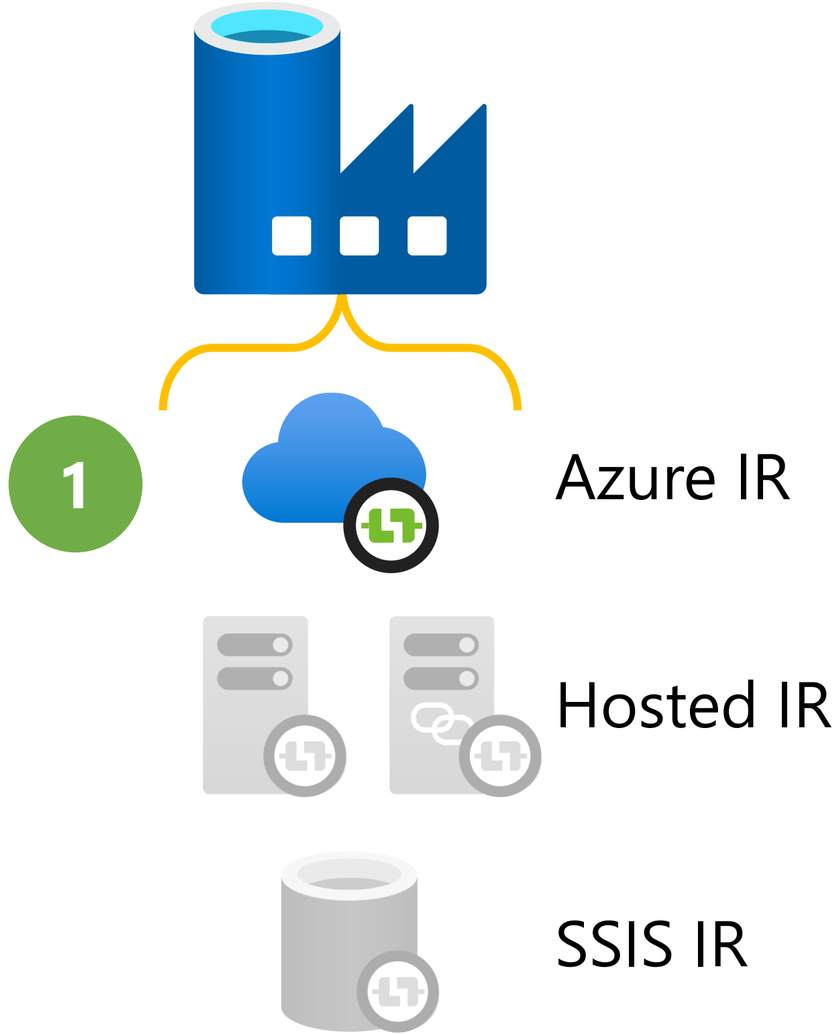
# An Introduction to Azure Data Factory

## Integration Runtimes

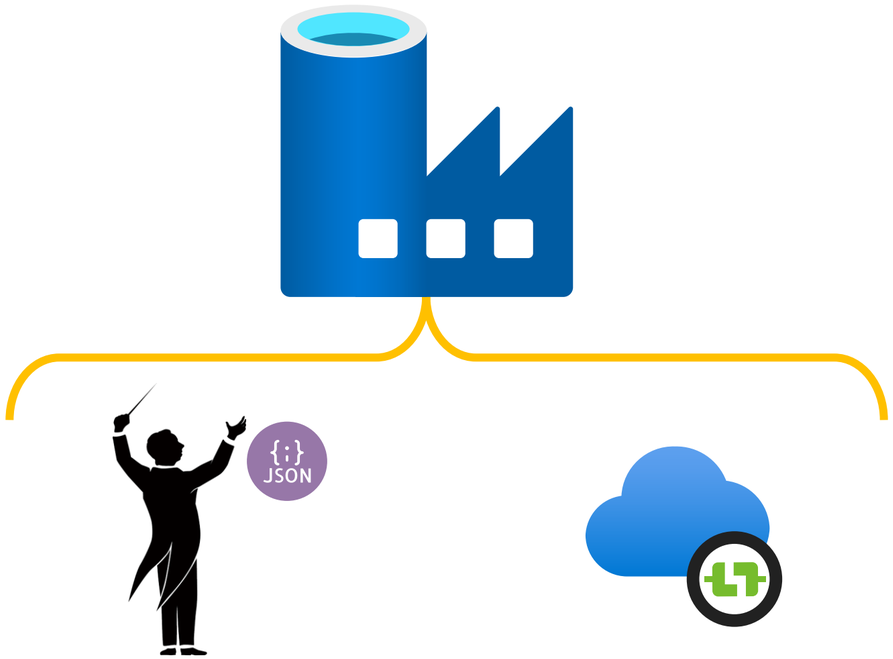


- Compute Types
  - Azure
  - Hosted
  - SSIS
- Patterns & Configuration

# What can an Integration Runtime do?



# What is an Integration Runtime?




Orchestrator

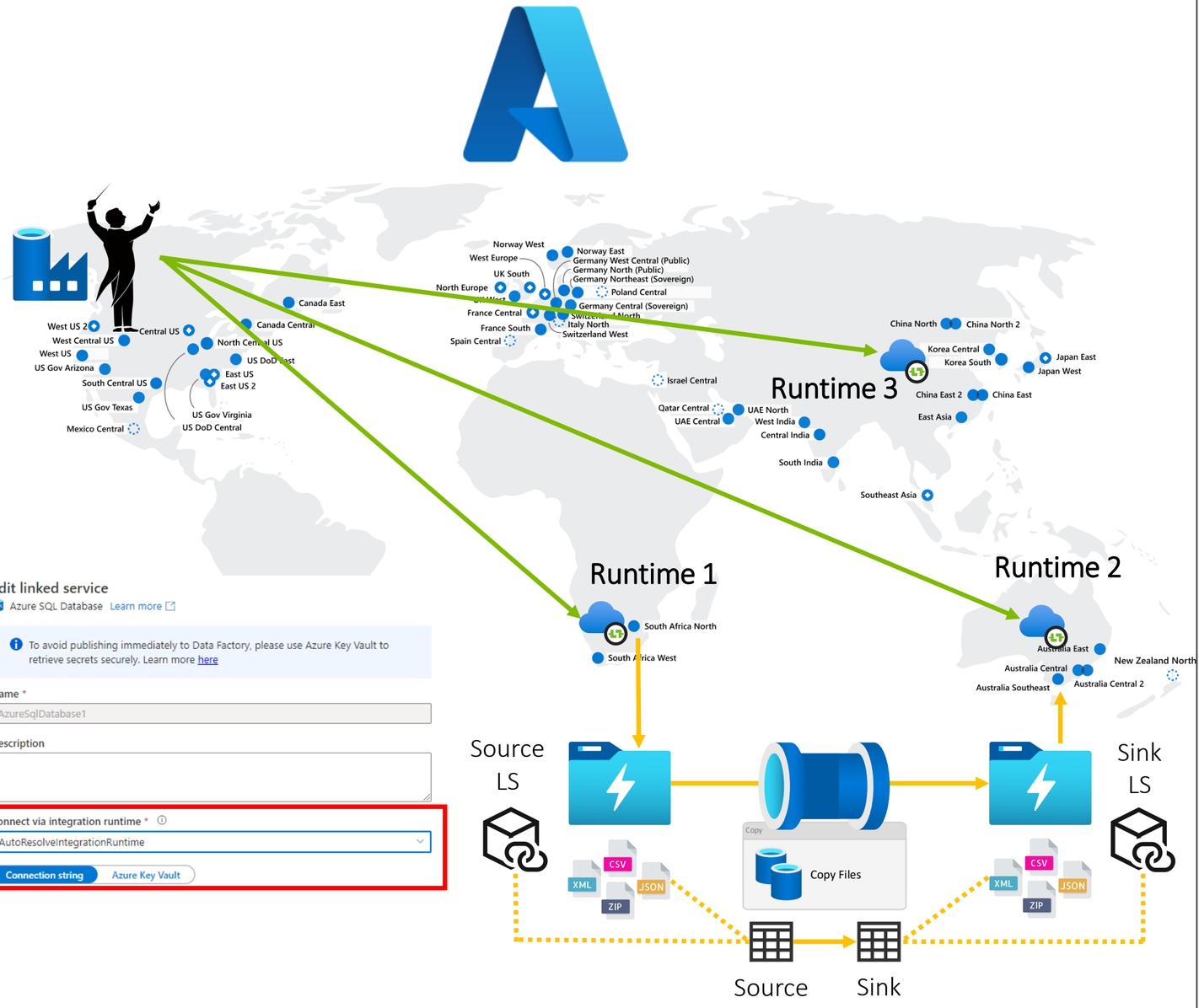
Fixed Region



Runtime

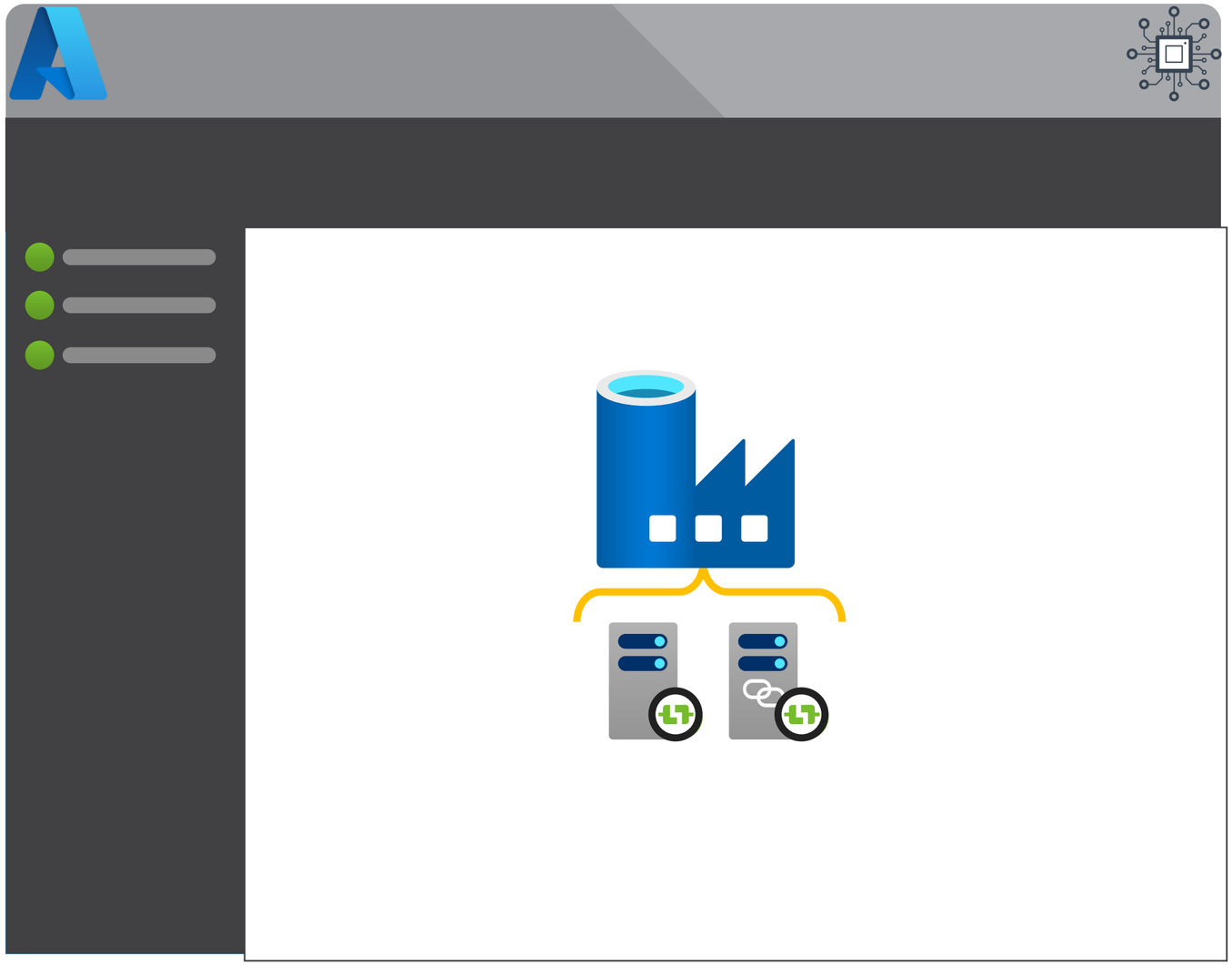
Flexible Region

*AutoResolveIntegrationRuntime*



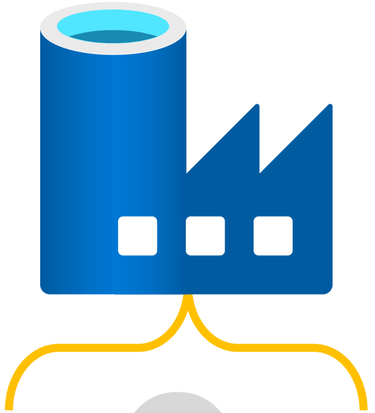
# An Introduction to Azure Data Factory

## Integration Runtimes



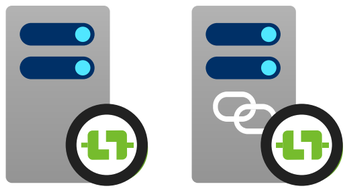
- Compute Types
  - Azure
  - Hosted
  - SSIS
- Patterns & Configuration

# What can an Integration Runtime do?



Azure IR

2



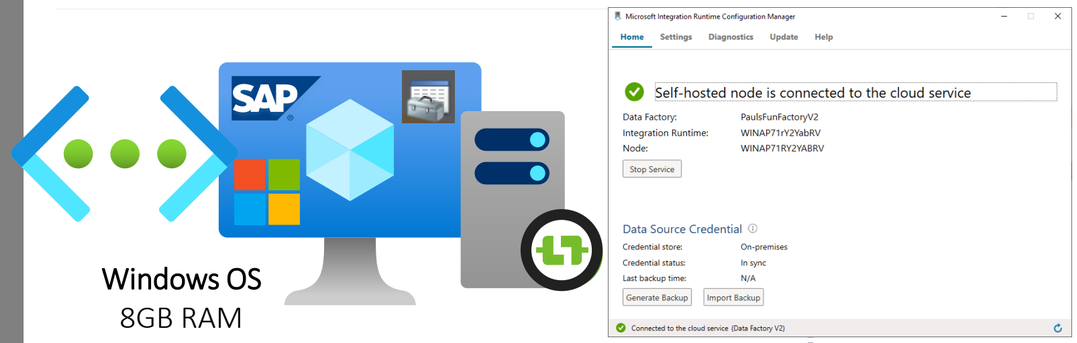
Hosted IR

SSIS IR

# Hosted Integration Runtime

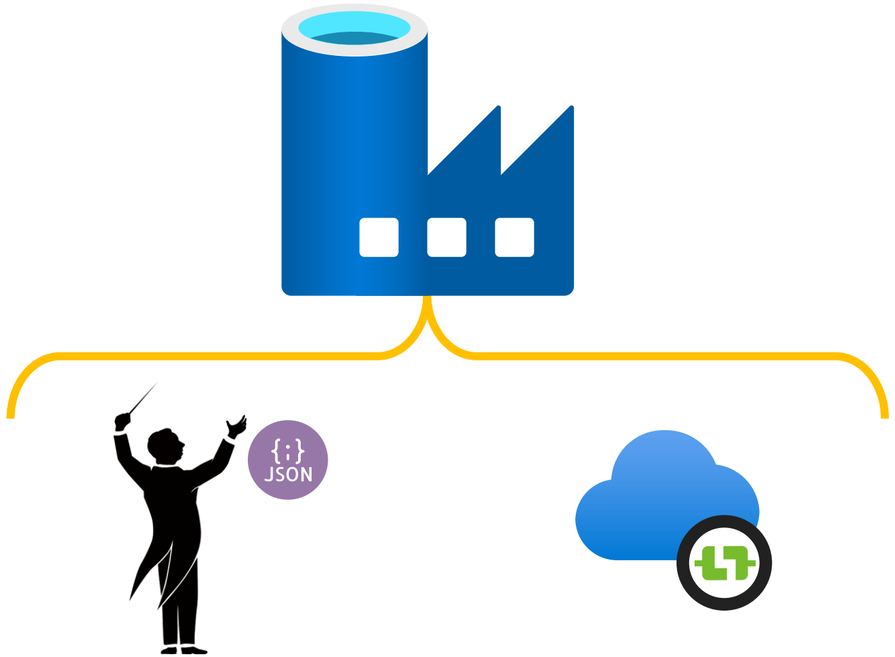
Windows OS
32GB RAM
6 CPU Cores
(12 Logical Cores)
80GB Storage

Data Sources




Windows OS
8GB RAM
4 CPU Cores
80GB Storage

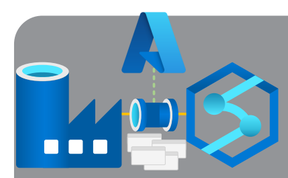
Runtime
Self Hosted

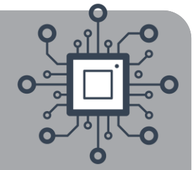


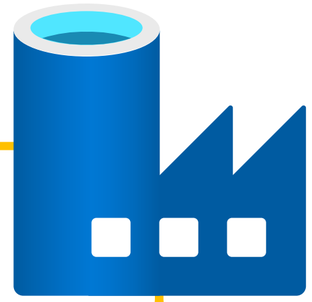

Orchestrator
Fixed Region

Runtime
Flexible Region

# Hosted Integration Runtime



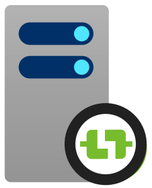




SAP

CSV XML JSON ZIP

Microsoft SQL Server



Runtime

Self Hosted

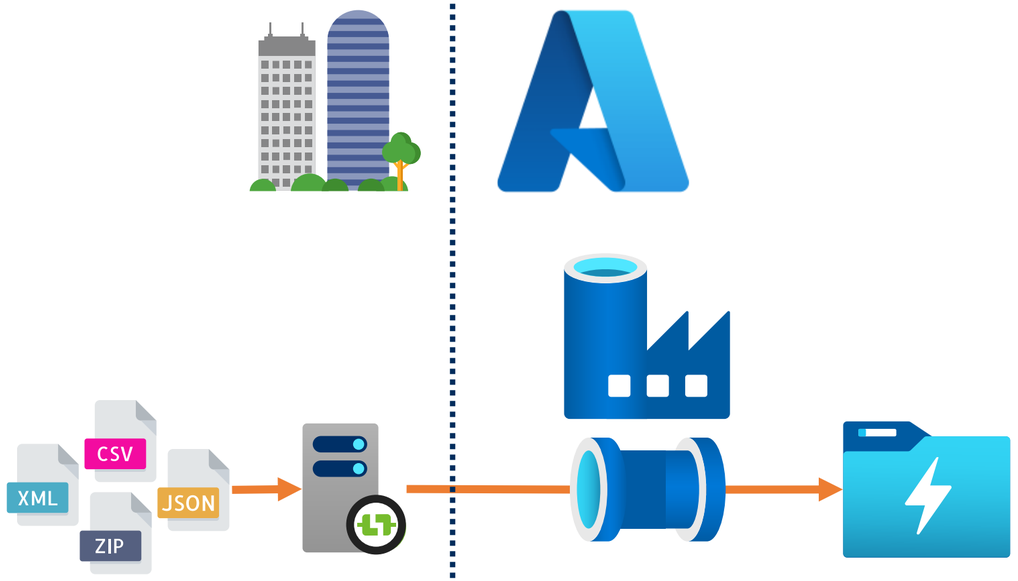
Orchestrator

Fixed Region

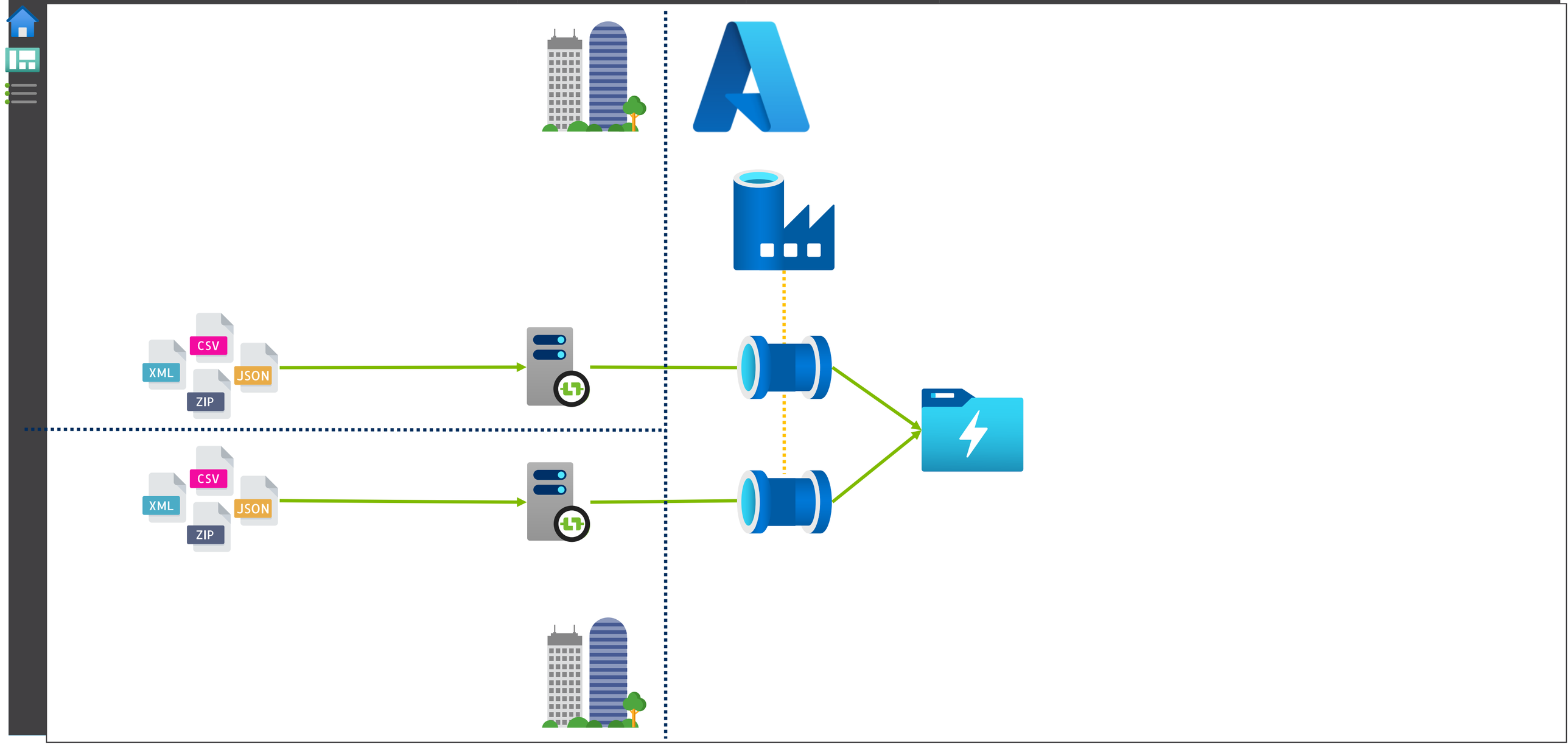
Runtime

Flexible Region

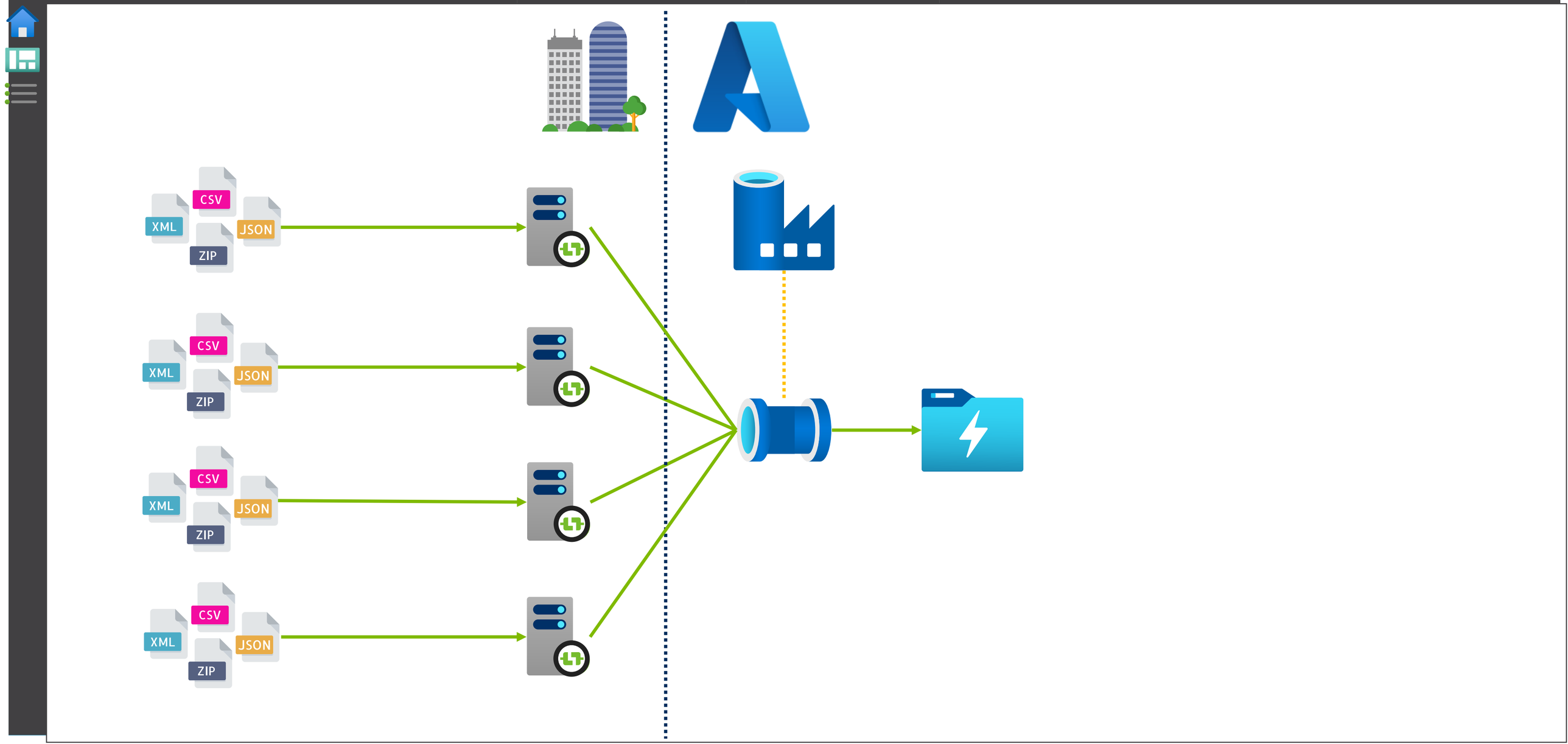
# Single Hosted IR



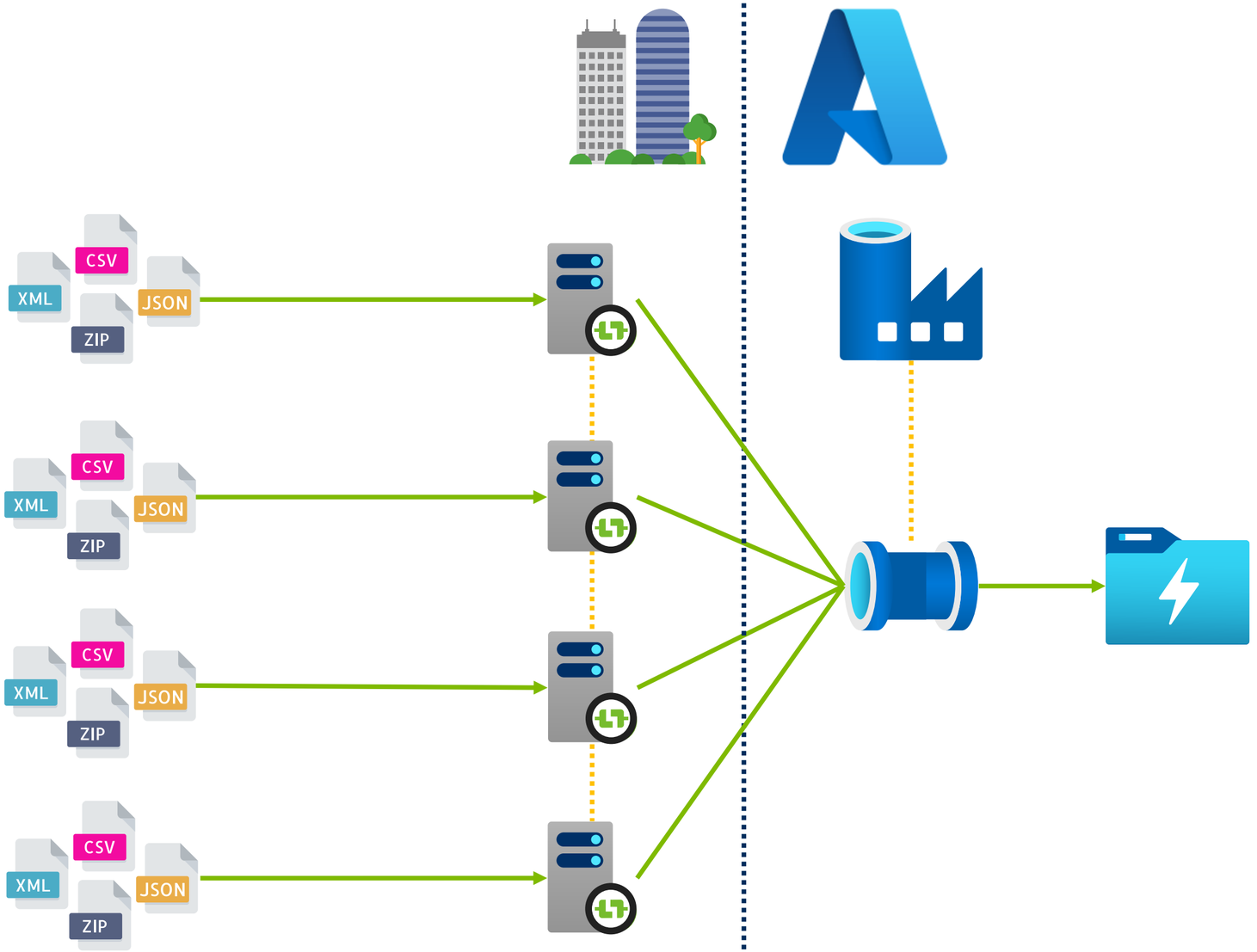
# Multiple Hosted IR's at Different Sites



# Multiple Hosted IR's as Separate Endpoints

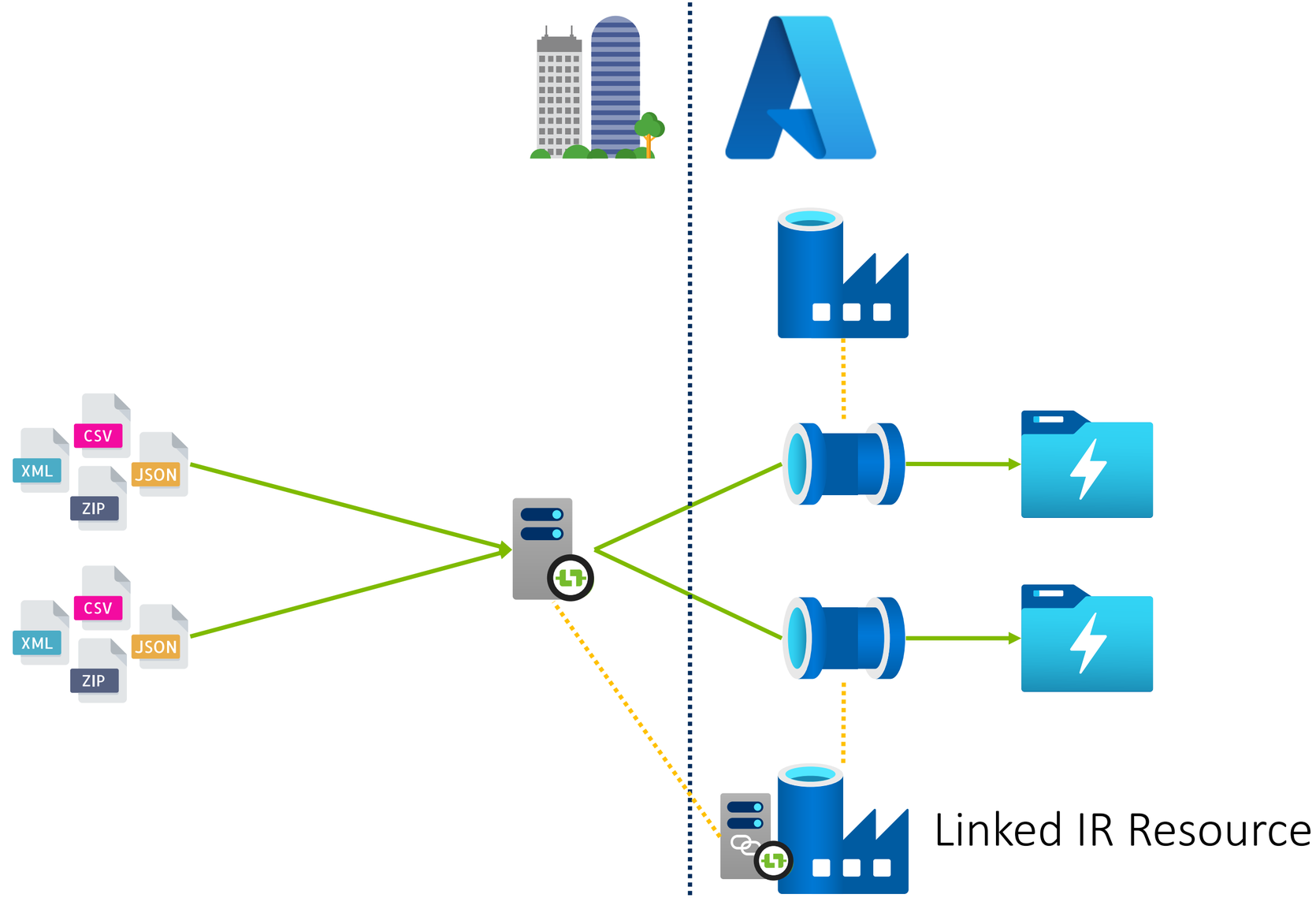


# Single Hosted IR with Multiple Nodes

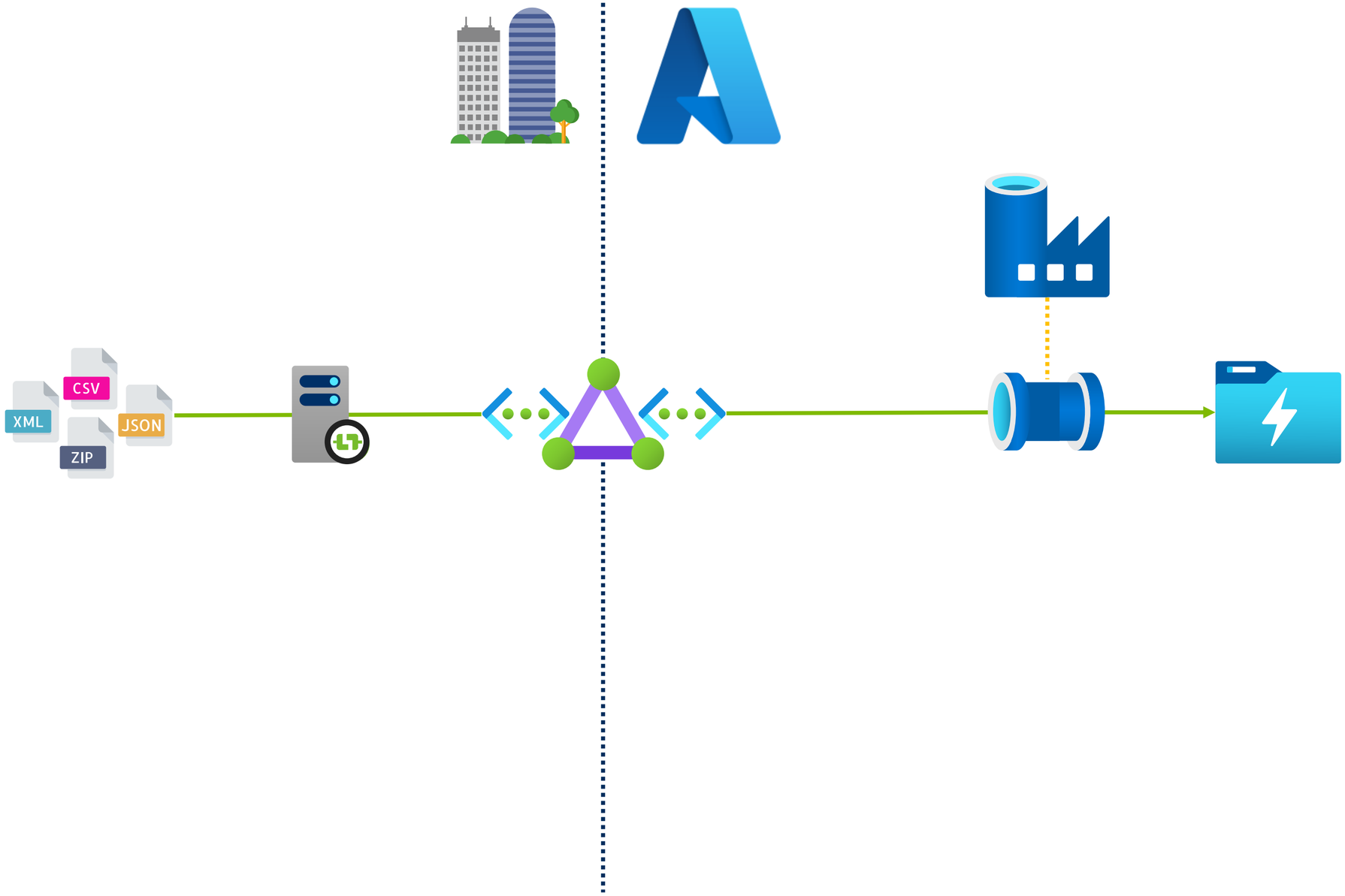



* Max of 4 Nodes per IR Endpoint

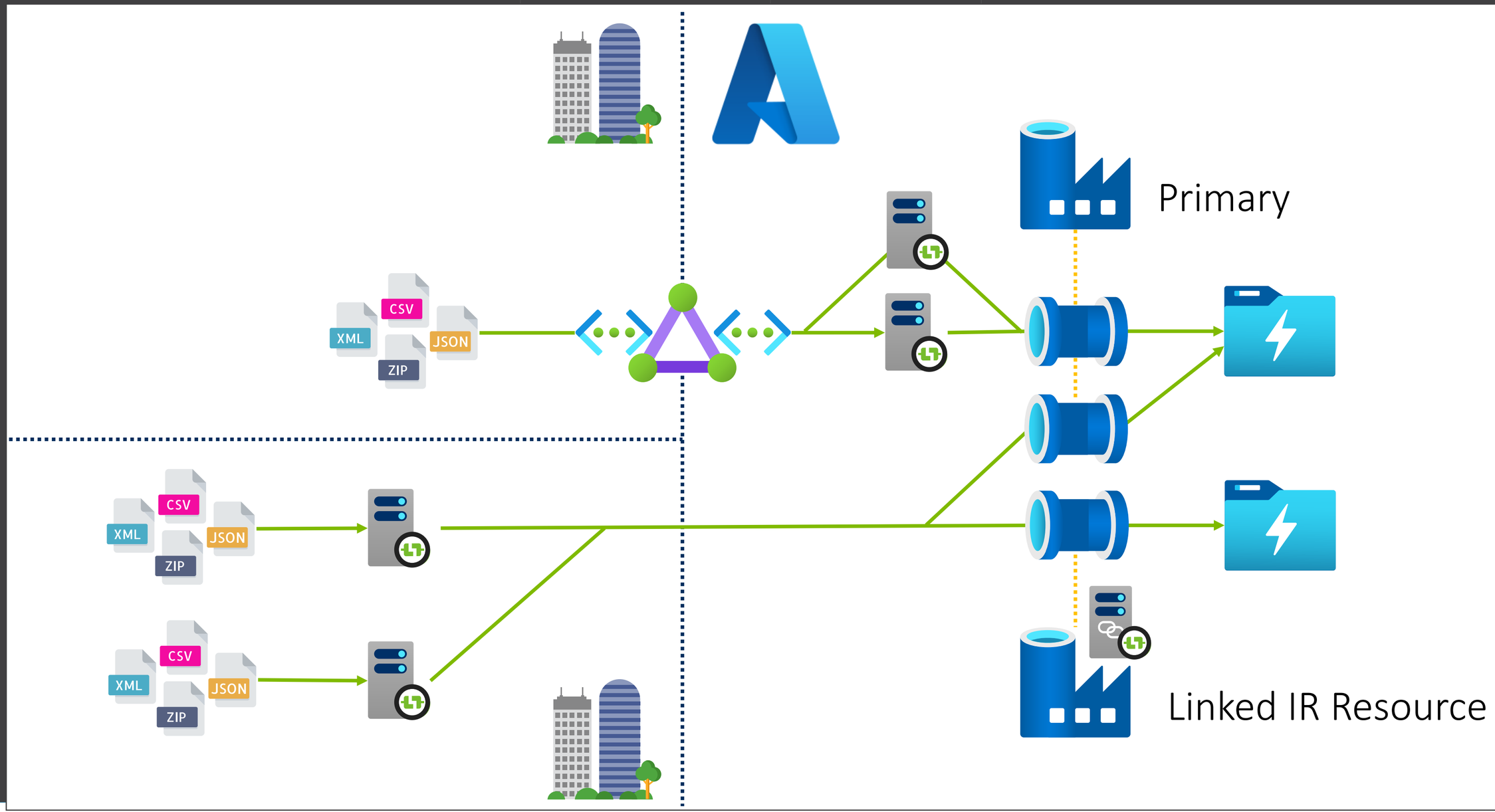
# Single Hosted IR Linked to Multiple Data Factories



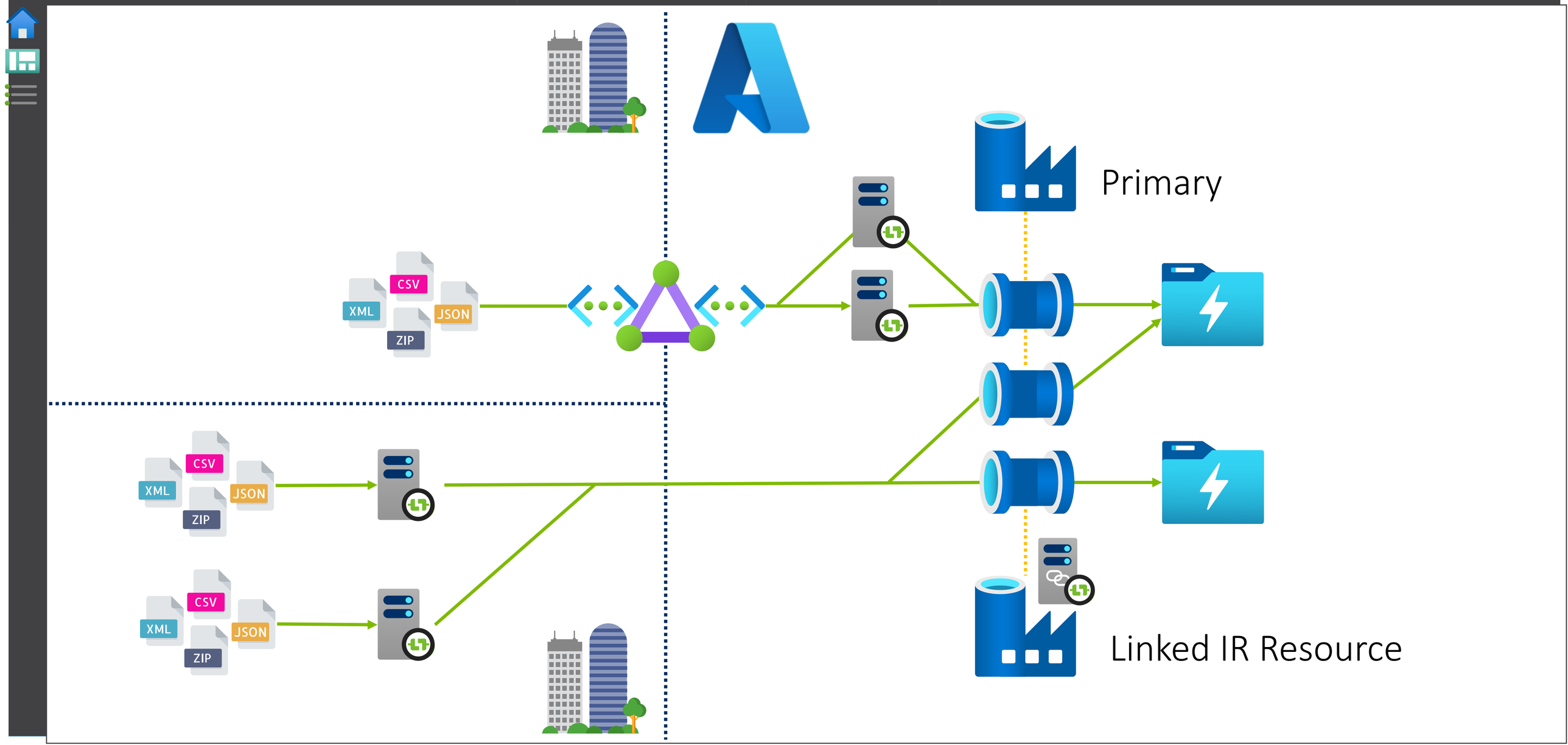
# Hosted IR’s with Express Route Connections



# Hosted IR Combinations



# Horizontal vs Vertical Environment Processing



# Hosted IR's vs Environments

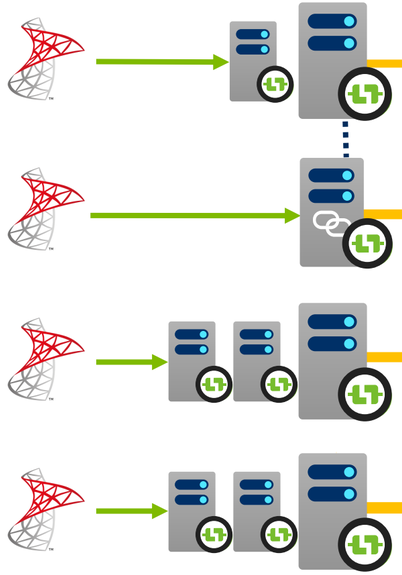
Dev

Test

PreProd

Prod

SQL Server Instances



25x Sites
4x Environments

2x IR's for Dev
1x Linked IR for Test (from Dev)

3x IR's for PreProd
3x IR's for Prod

8x Virtual Machines per Site
200x Virtual Machines Total

# Hosted IR's vs Environments

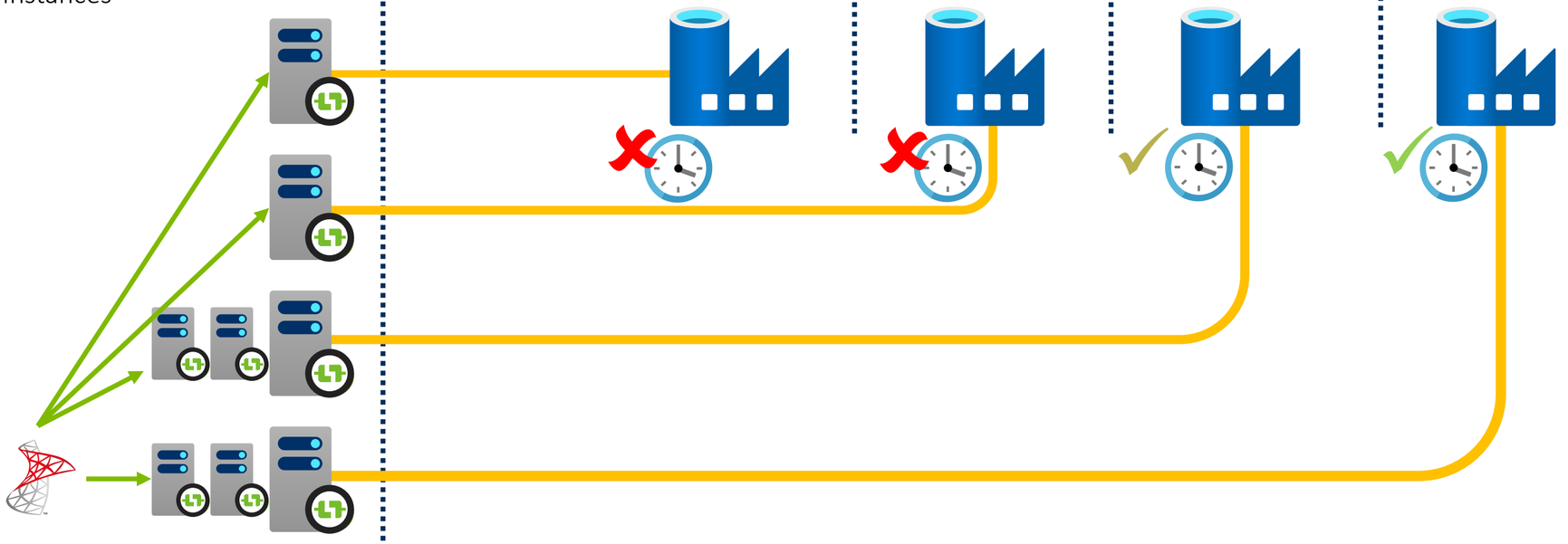
Dev

Test

PreProd

Prod

SQL Server Instances



25x Sites
4x Environments

2x IR's for Dev
1x Linked IR for Test (from Dev)

3x IR's for PreProd
3x IR's for Prod

8x Virtual Machines per Site
200x Virtual Machines Total

# Hosted IR's vs Environments

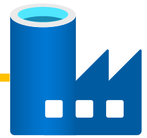
Dev

Test

PreProd

Prod

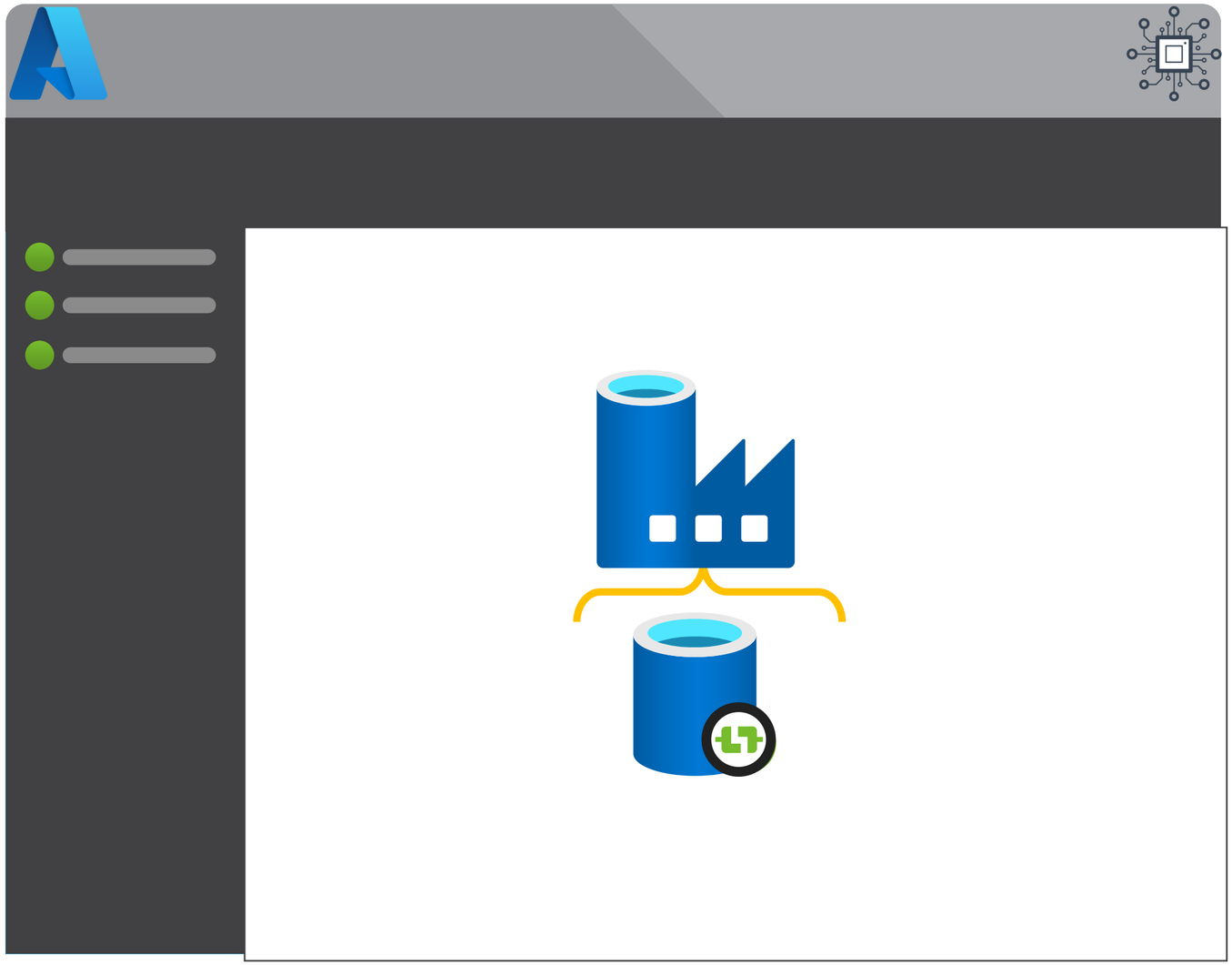
SQL Server Instances



Common

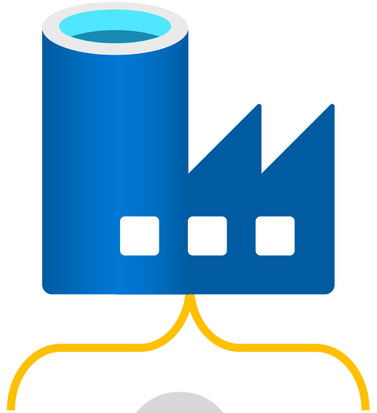
# An Introduction to Azure Data Factory

## Integration Runtimes

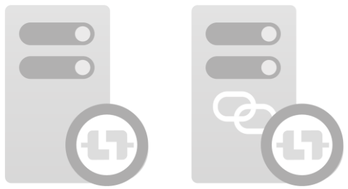


- Compute Types
  - Azure
  - Hosted
  - SSIS
- Patterns & Configuration

# What can an Integration Runtime do?



Azure IR



Hosted IR

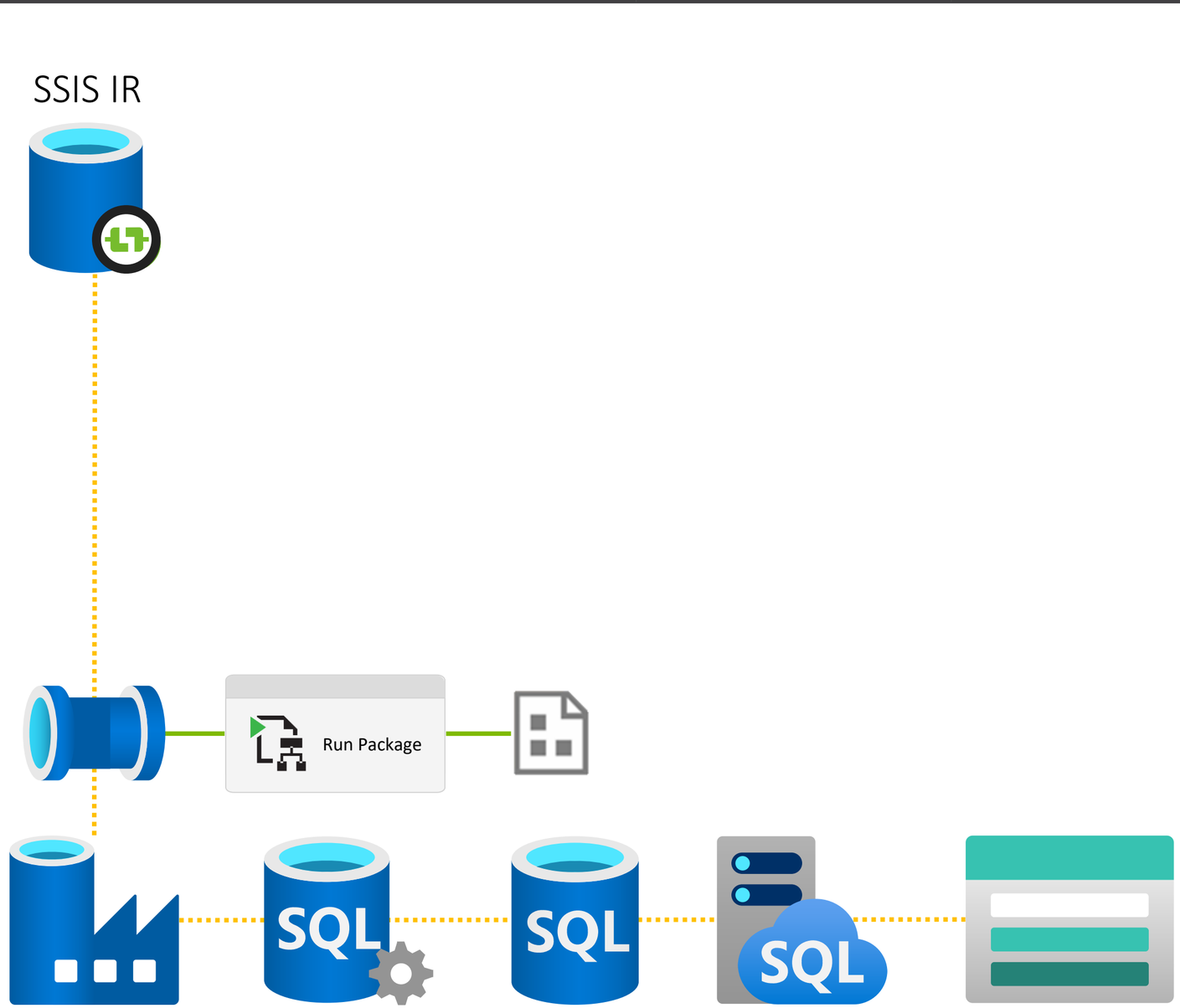
3

SSIS IR

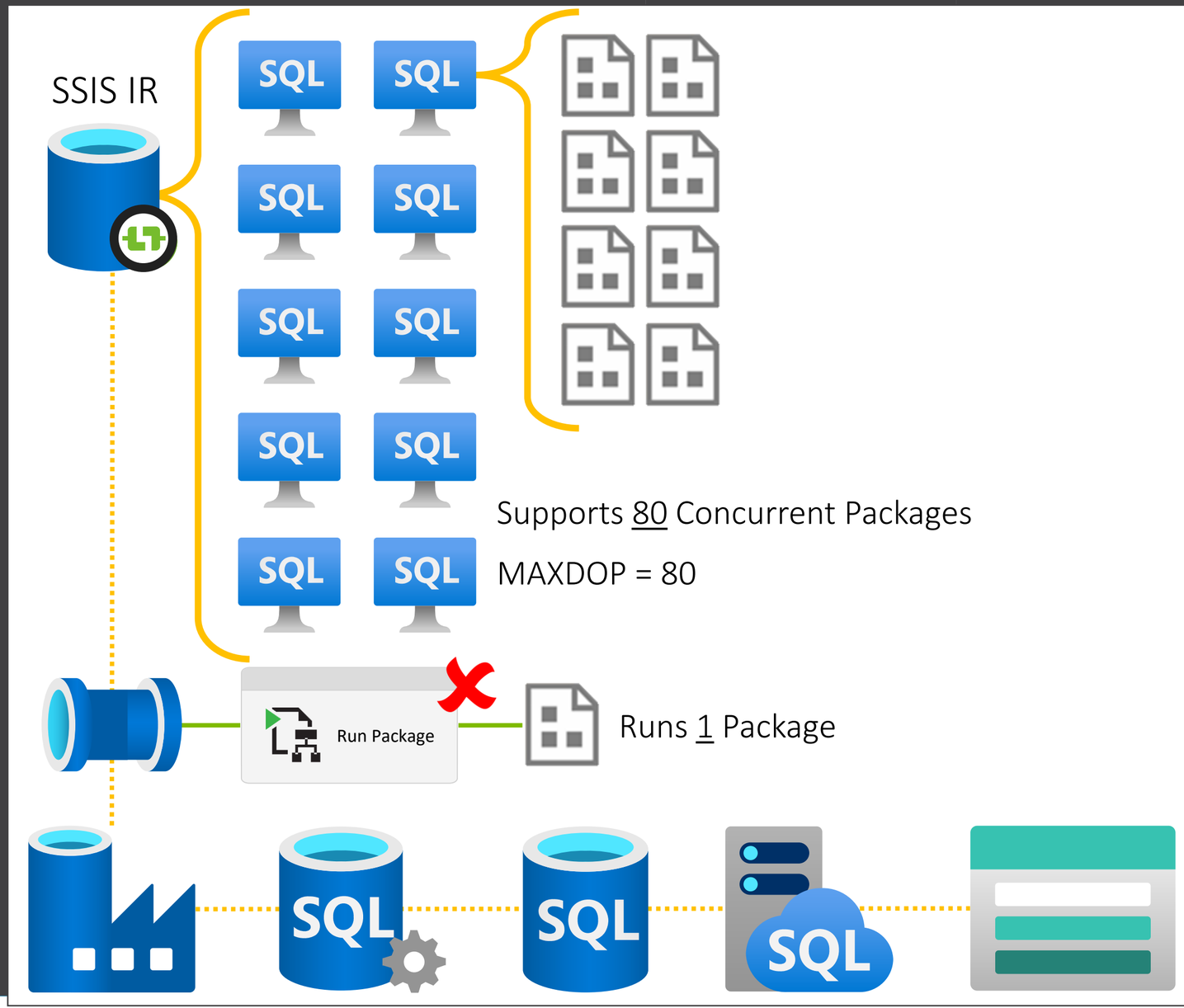
# Running an SSIS Package in Azure

SSIS IR

# Running an SSIS Package in Azure

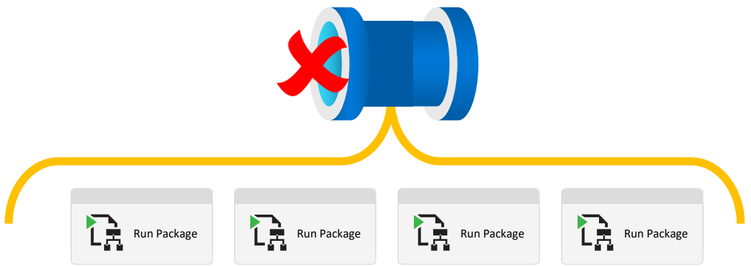


# Problem: Using All Of The SSIS IR Compute



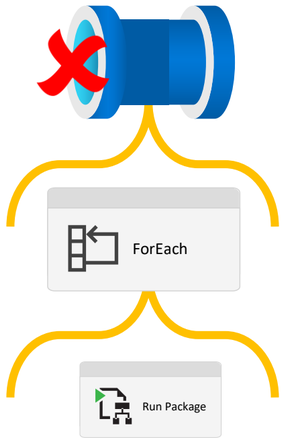
Parent Package

Child Packages x80




Pipeline x1

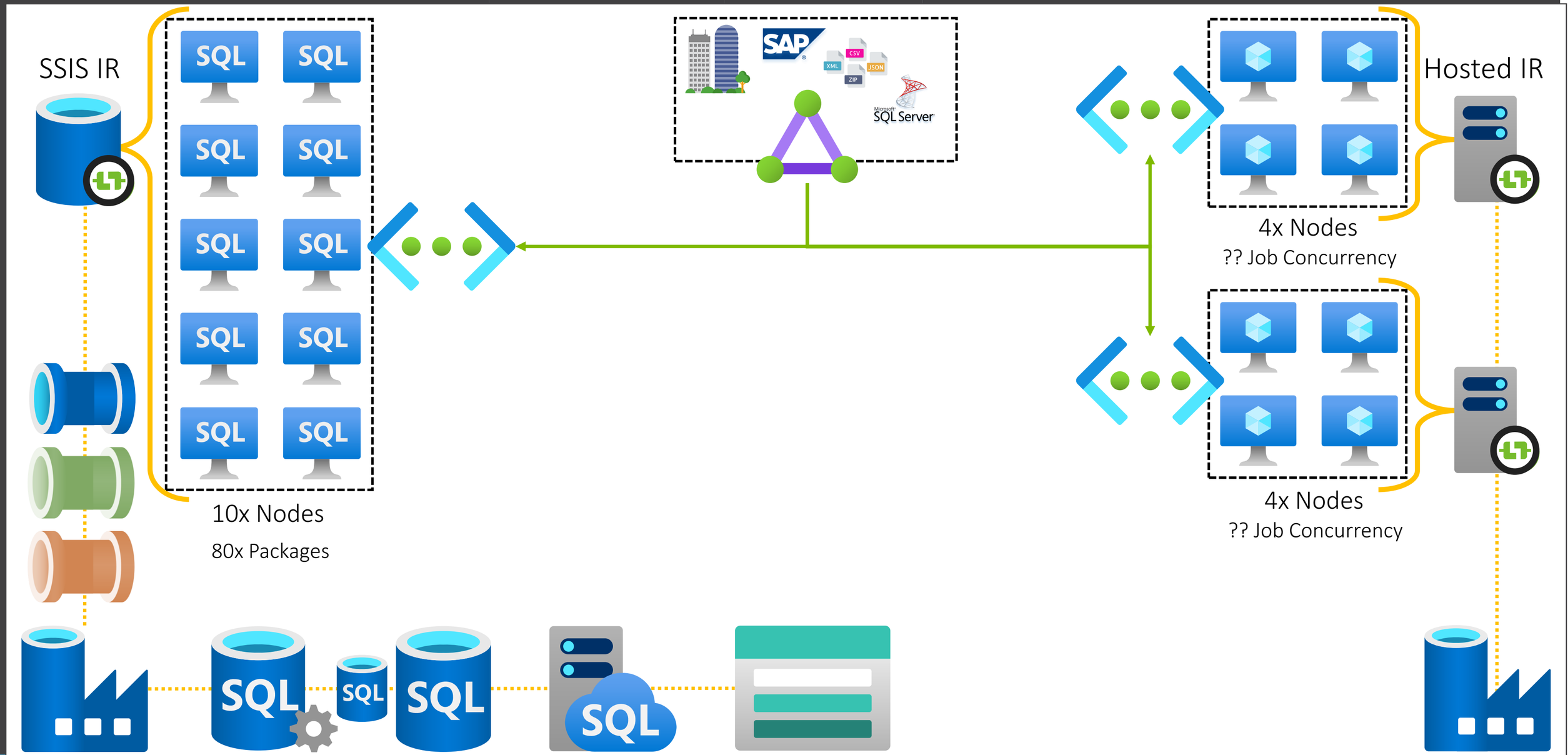
Activities x80




Pipeline x1

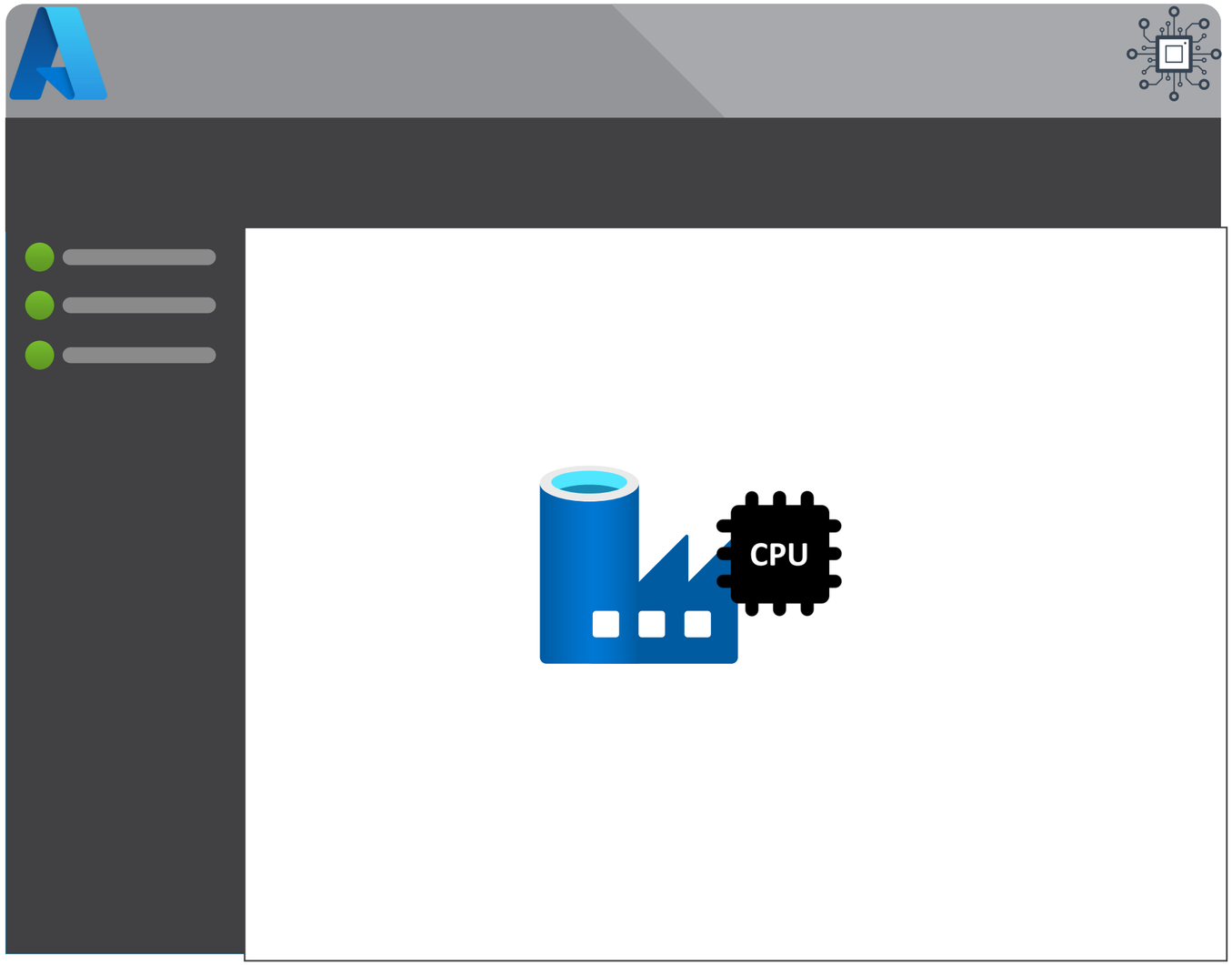
ForEach
Max Batch
(50)

# What are the advantages of using this pattern?



# An Introduction to Azure Data Factory

## Integration Runtimes

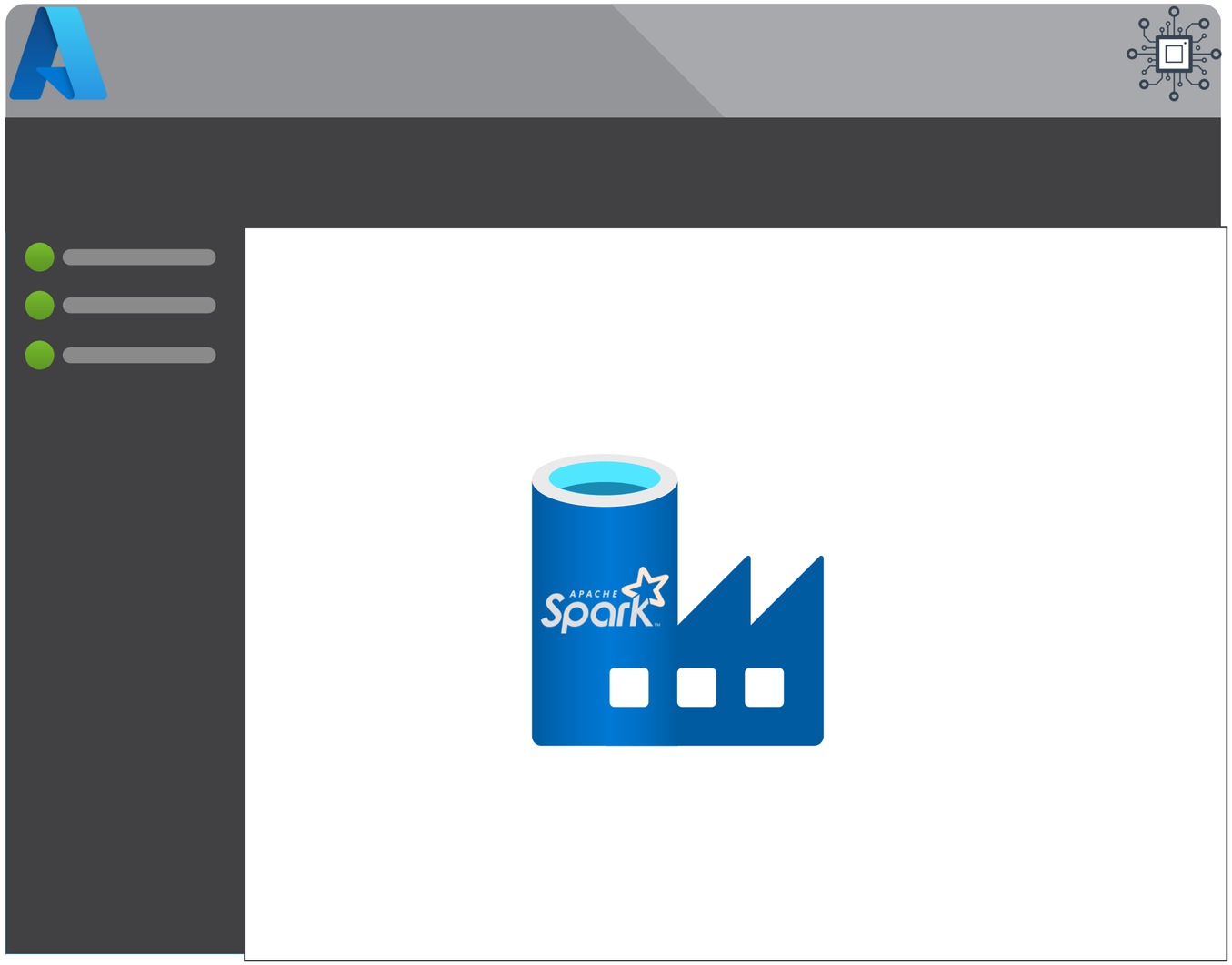



- Compute Types
  - Azure
  - Hosted
  - SSIS
- Patterns & Configuration

# Demo

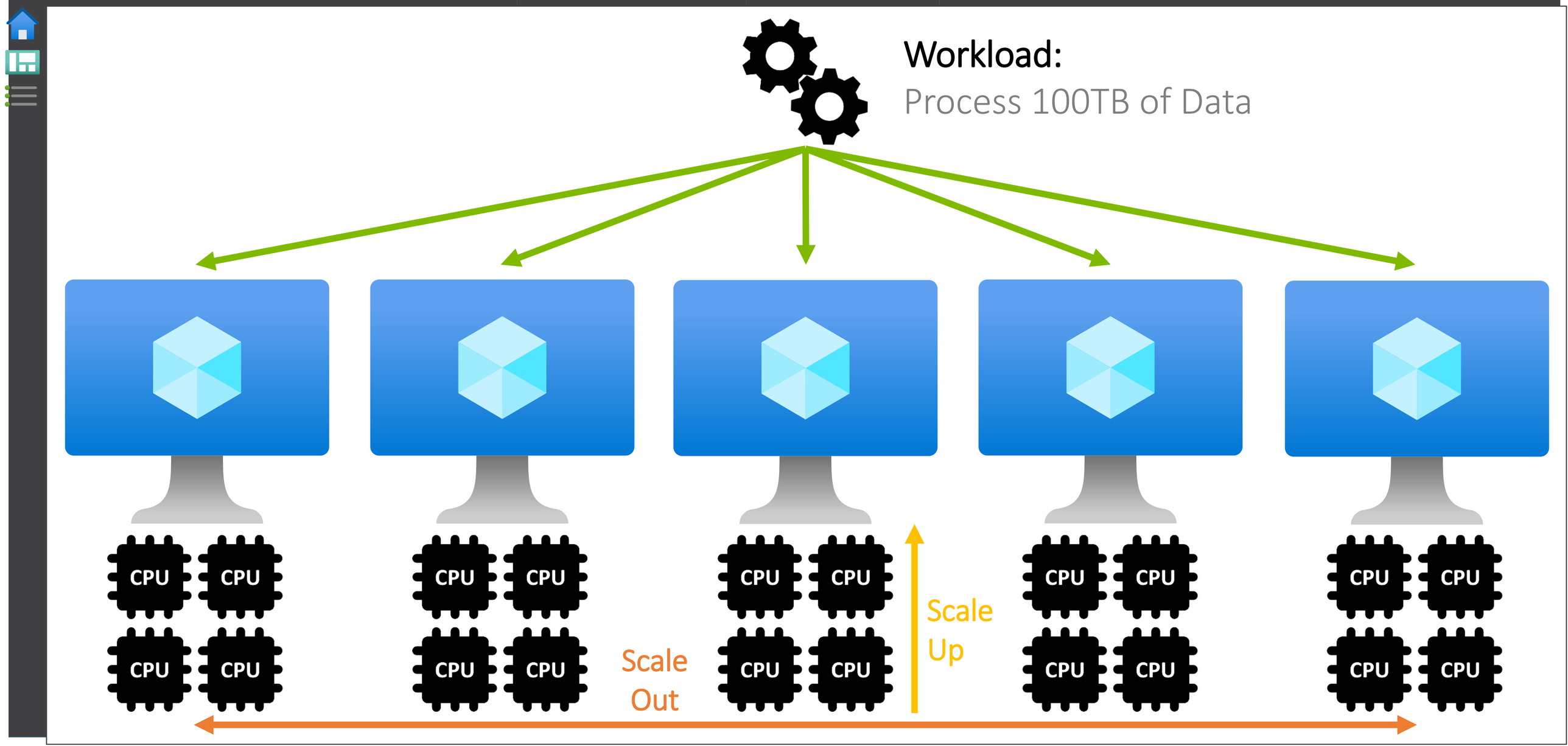
# An Introduction to Azure Data Factory

## Data Transformation

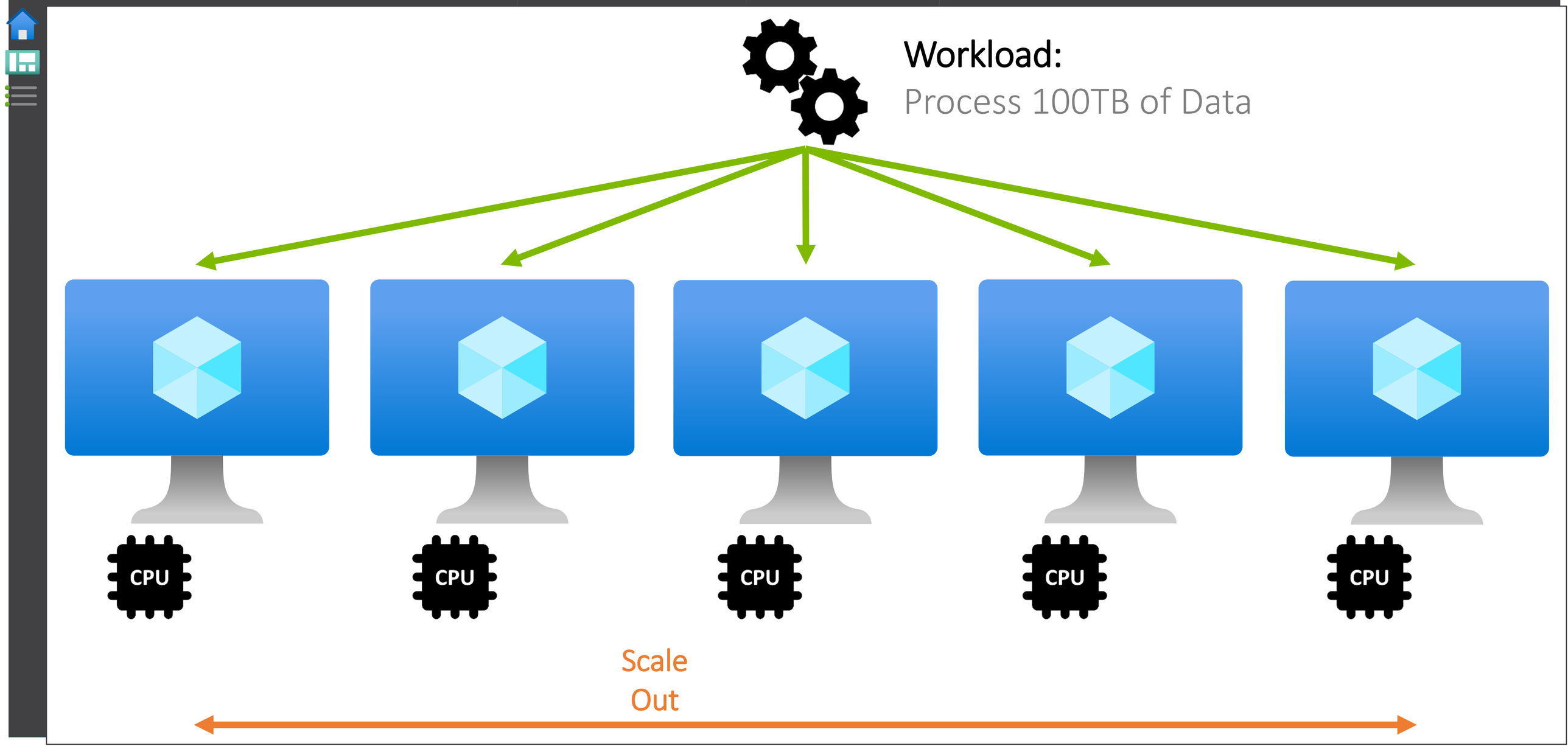



- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

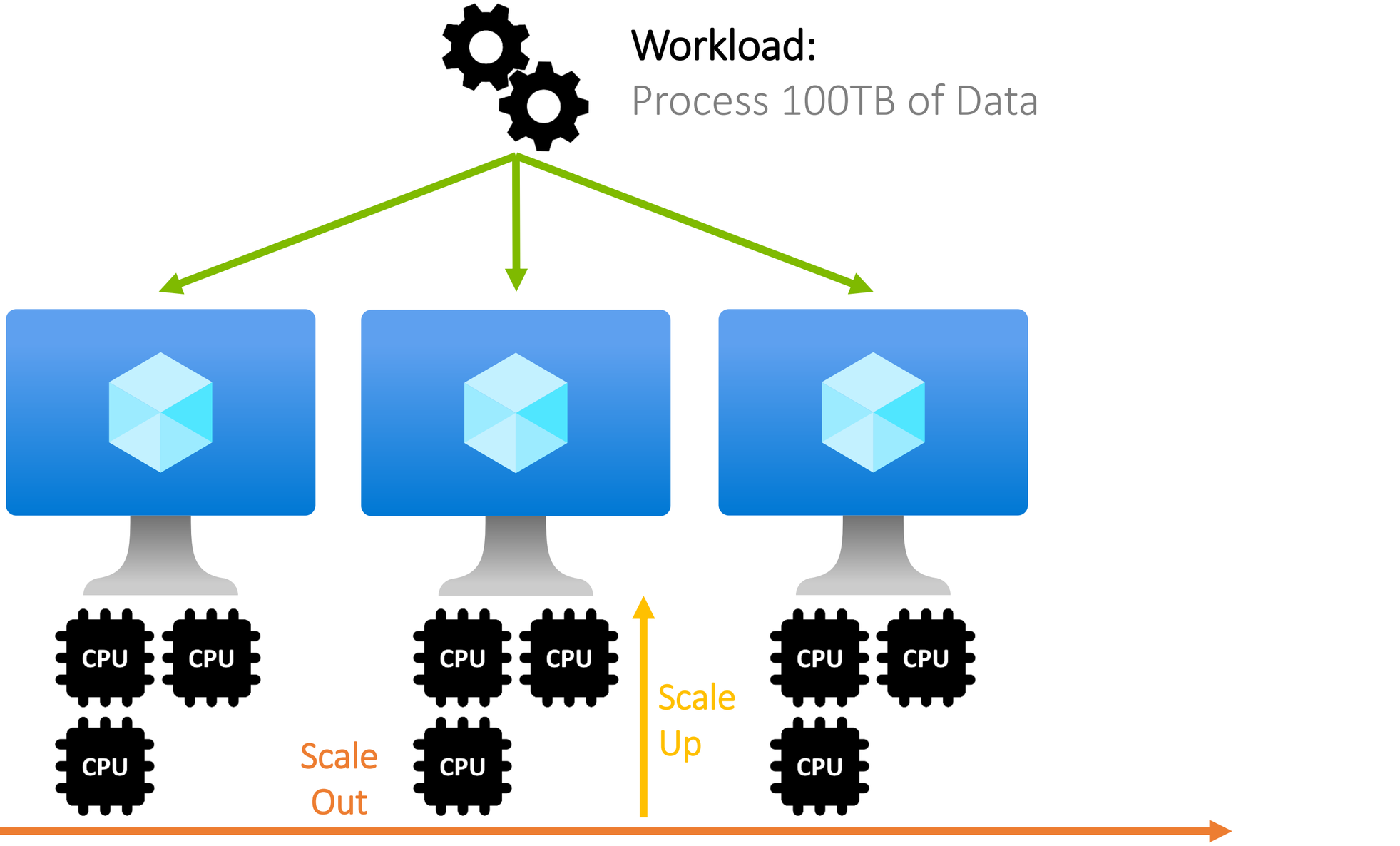
# Scaling Up and/or Scaling Out



# Scaling Up and/or Scaling Out

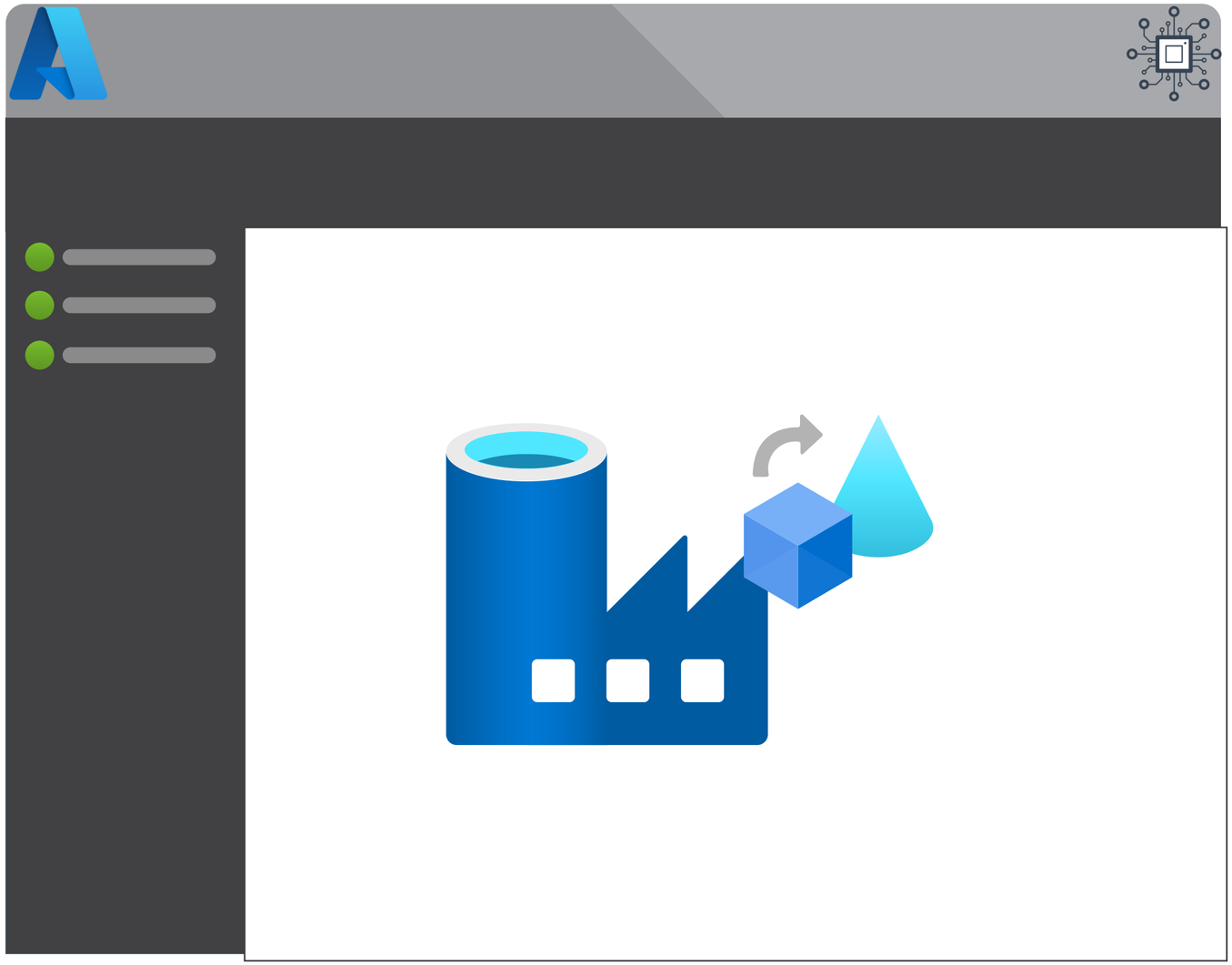


# Scaling Up and/or Scaling Out



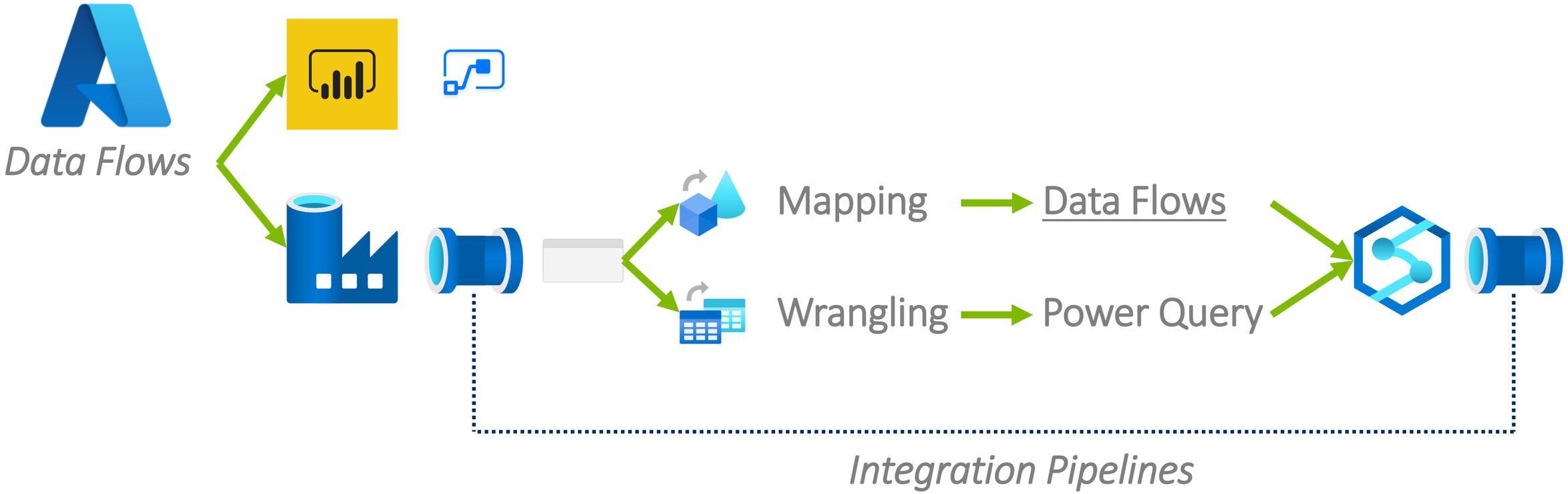
# An Introduction to Azure Data Factory

## Data Transformation

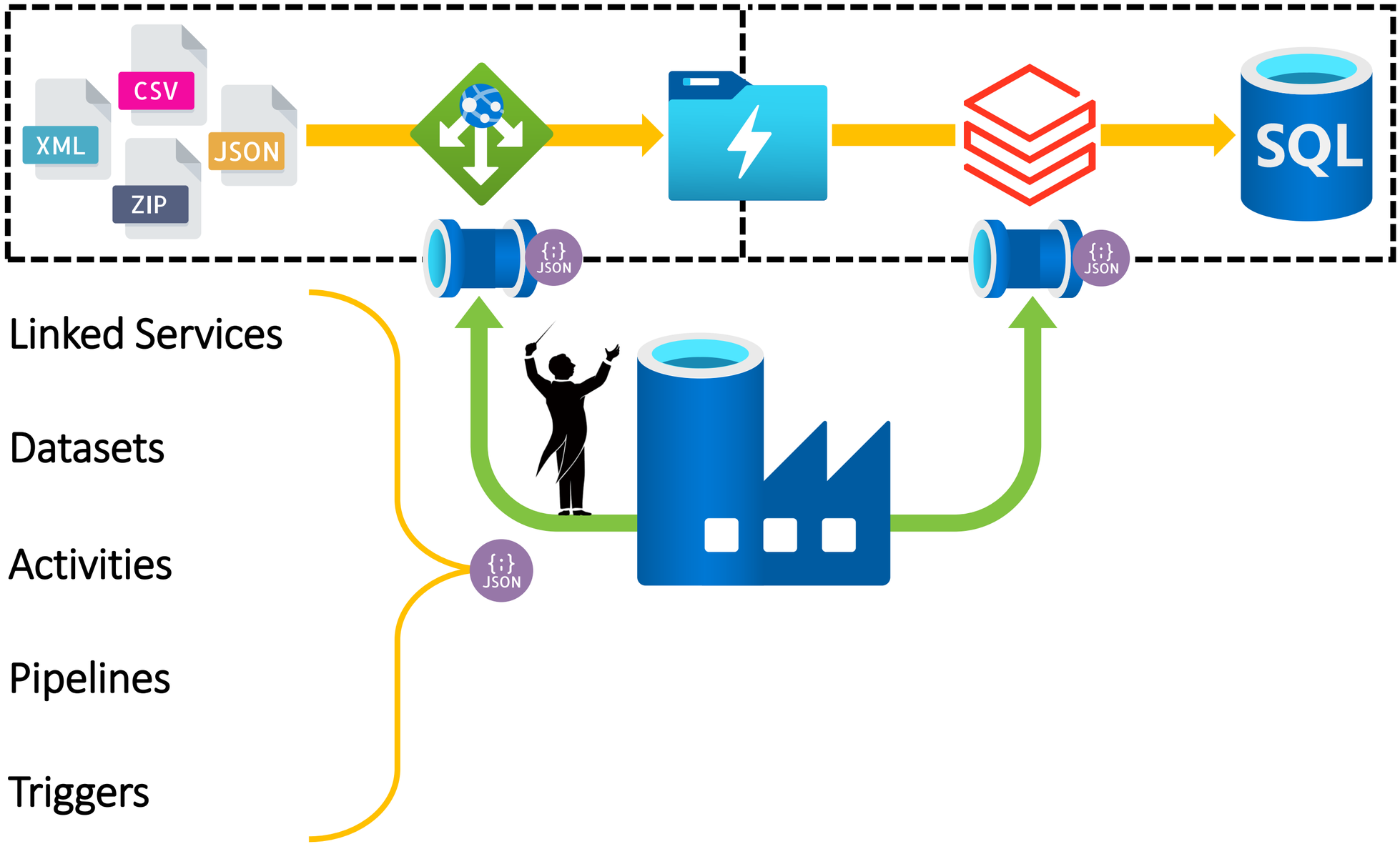


- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

# Terminology Clarification

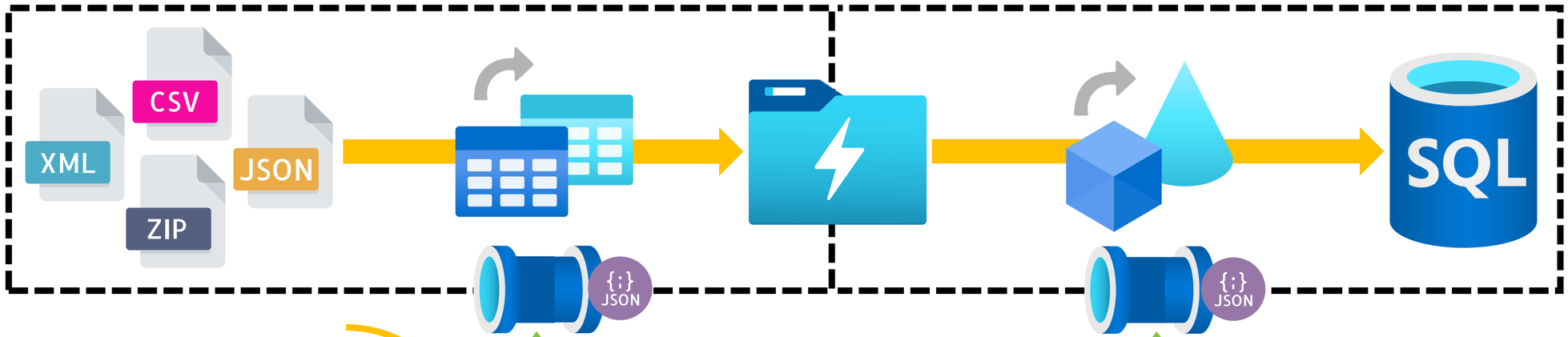


# Control Flow Components




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

# ~~Control~~ Data Flow Components

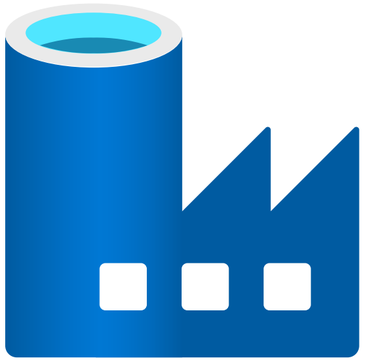



1. Linked Services
2. Datasets
3. Activities
4. Pipelines
5. Triggers

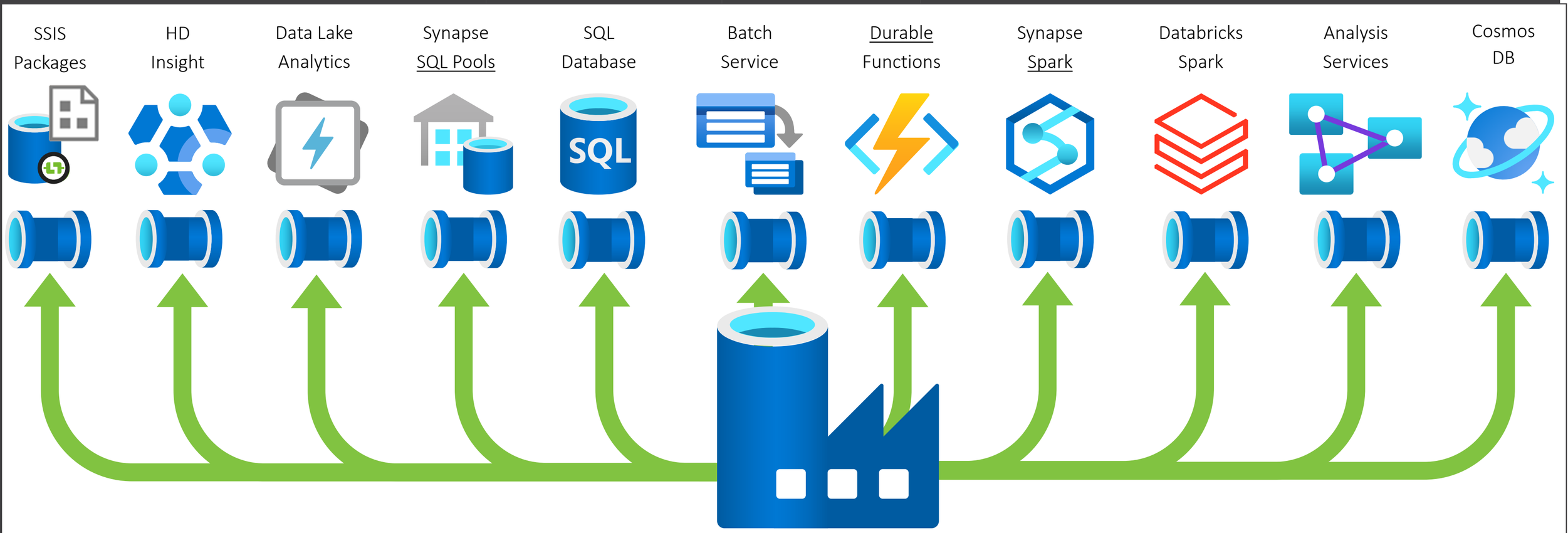
Mapping → Data Flows

Wrangling → Power Query

# Data Transformation in Azure

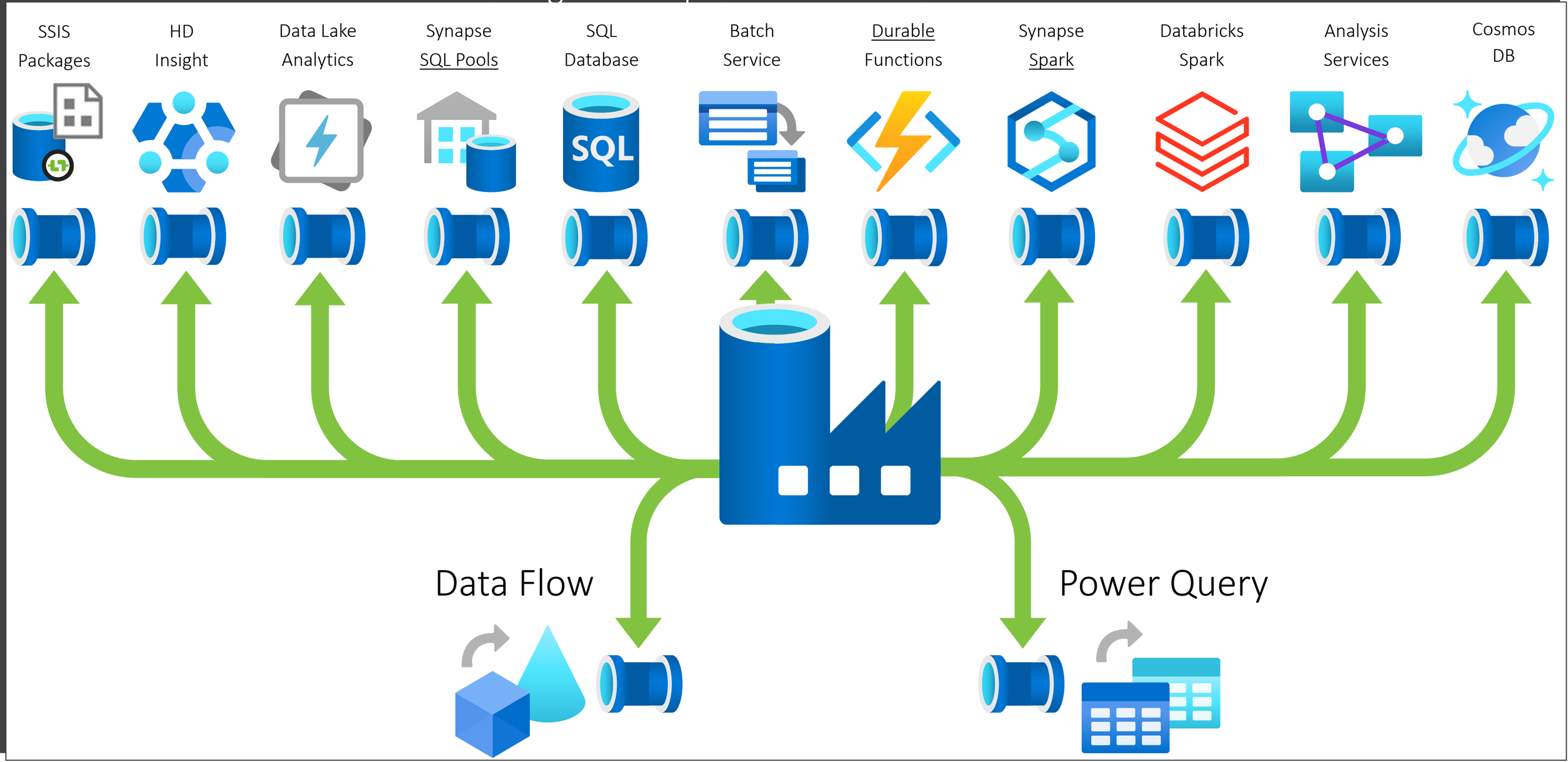


# Other Data Transformation Services in Azure



# Other Data Transformation Services in Azure

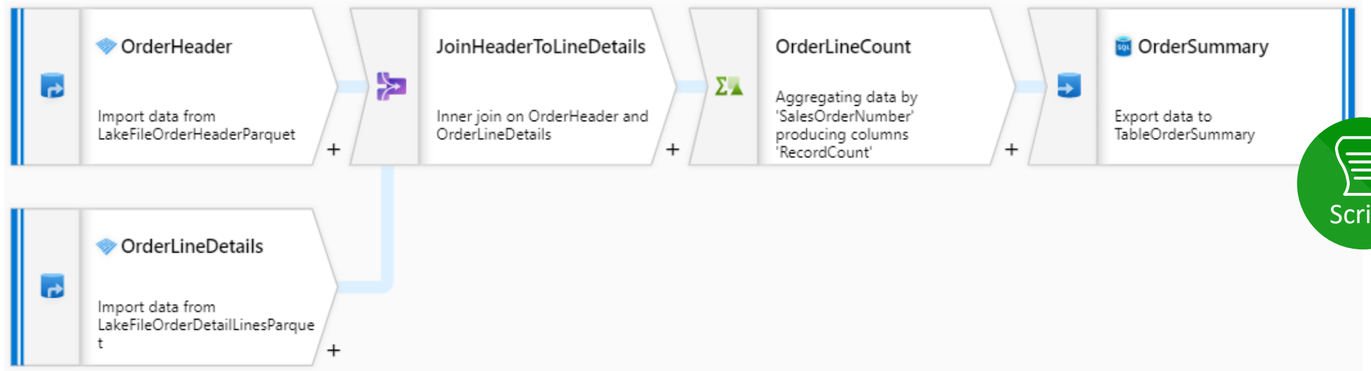
## When Should We Use These Integration Pipeline Transformation Activities?

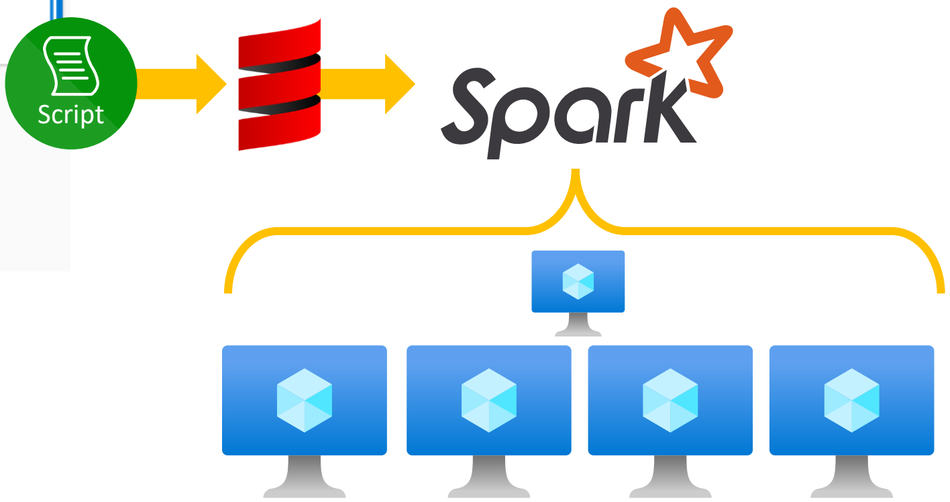


# What is a ~~Mapping~~ Data Flow?

Control Flow

Data Flow

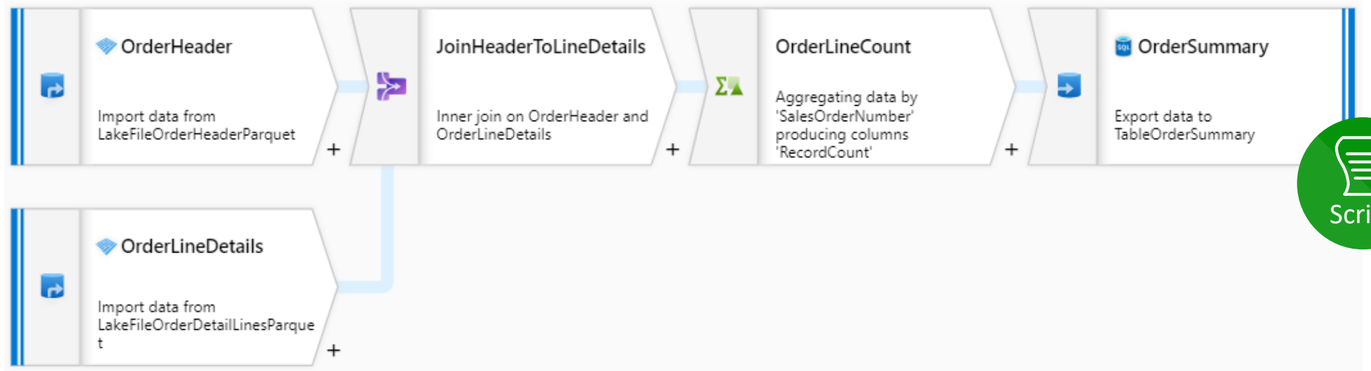


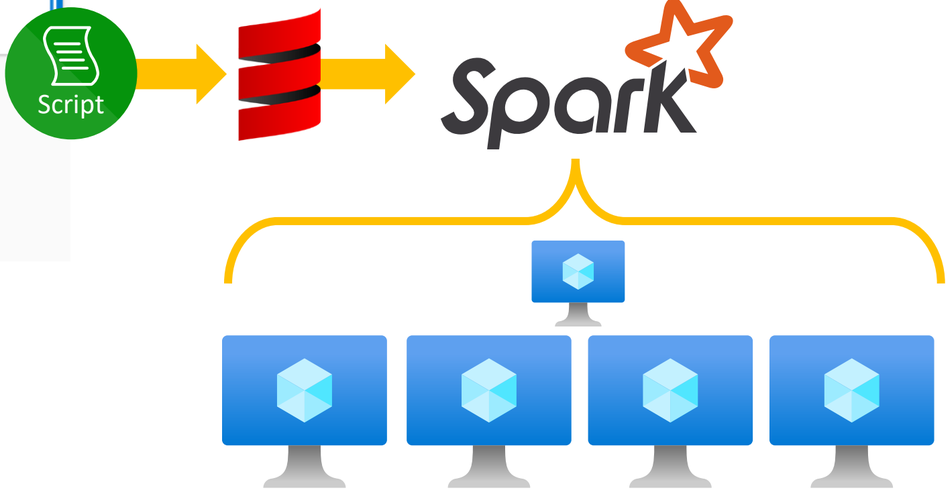



# Q: What is a ~~Mapping~~ Data Flow?

Control Flow

Data Flow




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

# Data Flows – Inputs & Outputs

Source & Sink

Linked Services

Source Types

Dataset

Inline

# Data Flows – Transformations

New Branch

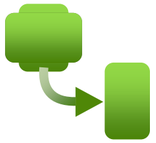
Join

Conditional Split

Exists

Union

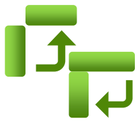
Lookup


Derived Column

Select

Aggregate

Surrogate Key


Pivot/Unpivot

Window

Rank

External Call

Cast

Flatten

Parse

Stringify

Filter

Sort

Alter Row

Assert

Flowlet

Key

Input & Output Modifiers

Schema Modifiers

Formatters

Row Modifiers

# Data Flows – Transformations

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



New Branch



Join



Conditional Split



Exists



Union



Lookup

## Components

| Operation / Activity | Description | SSIS equivalent | SQL Server equivalent |
|---|---|---|---|
| New branch | Create a new flow branch with the same data | Multicast (+icon) | `SELECT INTO`<br>`SELECT OUTPUT` |
| Join | Join data from two streams based on a condition | Merge join | `INNER/LEFT/RIGHT JOIN,`<br>`CROSS/FULL OUTER JOIN` |
| Conditional Split | Route data into different streams based on conditions | Conditional Split | `SELECT INTO WHERE condition1`<br>`SELECT INTO WHERE condition2`<br>`CASE ... WHEN` |
| Union | Collect data from multiple streams | Union All | `SELECT col1a UNION (ALL)`<br>`SELECT col1b` |
| Lookup | Lookup additional data from another stream | Lookup | *Subselect, function,*<br>`LEFT/RIGHT JOIN` |
| Derived Column | Compute new columns based on the existing once | Derived Column | `SELECT Column1 * 1.09 as NewColumn` |
| Aggregate | Calculate aggregation on the stream | Aggregate | `SELECT Year(DateOfBirth) as YearOnly,`<br>`MIN(), MAX(), AVG()`<br>`GROUP BY Year(DateOfBirth)` |
| Surrogate Key | Add a surrogate key column to output stream from a specific value | Script Component | `SELECT ROW_NUMBER()`<br>`OVER(ORDER BY name ASC) AS Row#,`<br>`name`<br>`FROM sys.databases` |



Filter



Sort



Alter Row

Key

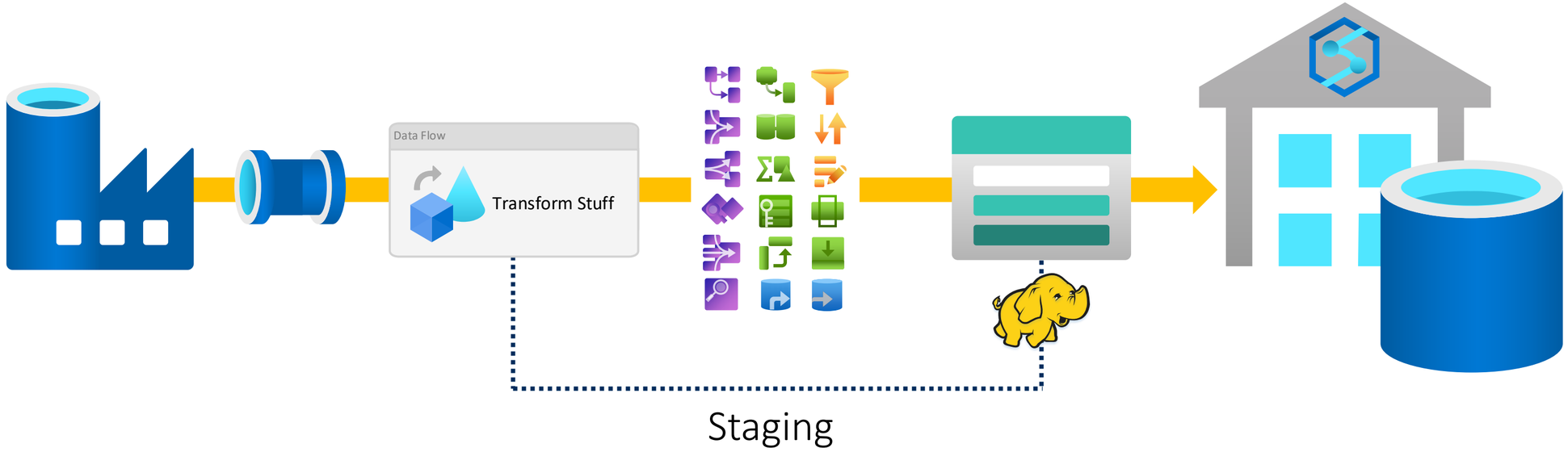
Input & Output Modifiers

Schema Modifiers

Formatters

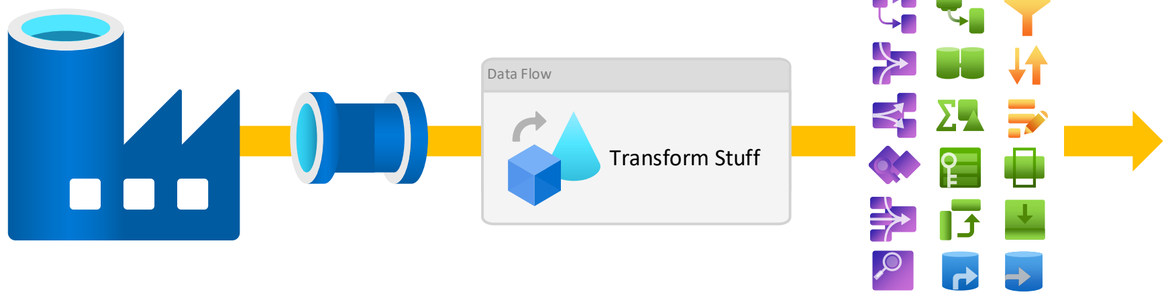
Row Modifiers

# Data Flows – Data Warehouse Loading (PolyBase)




Staging

# Data Flows – Expression Builder

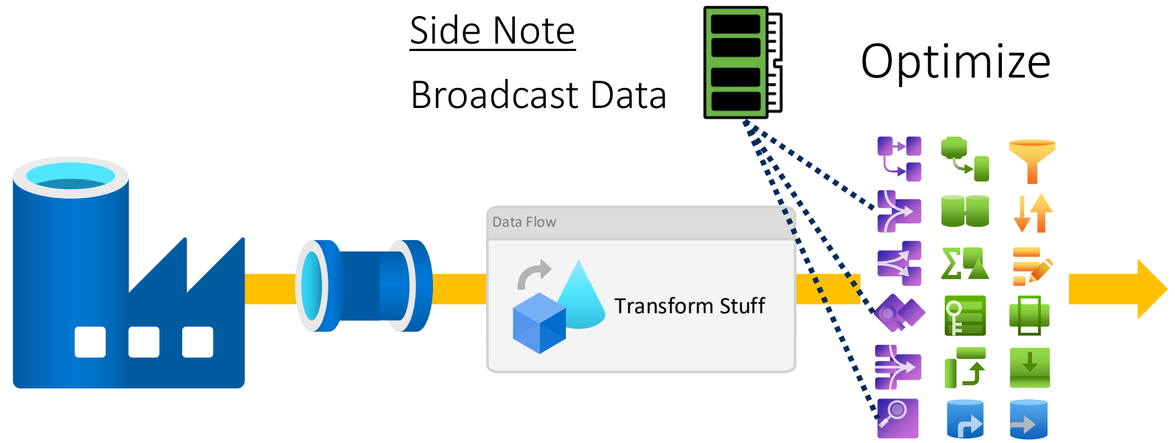


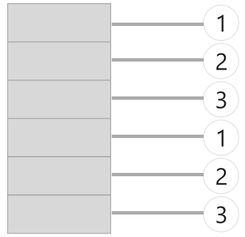

Expression Builder

- Count If
- Sum If
- Regex Extract
- Regex Match
- Regex Split
- R-Like

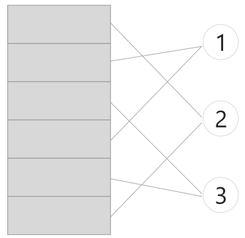
https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-expression-functions

# Data Flows – Data Distribution

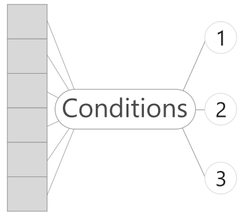


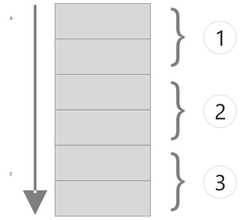

Round Robin



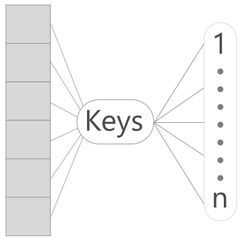

Hashed




Conditional

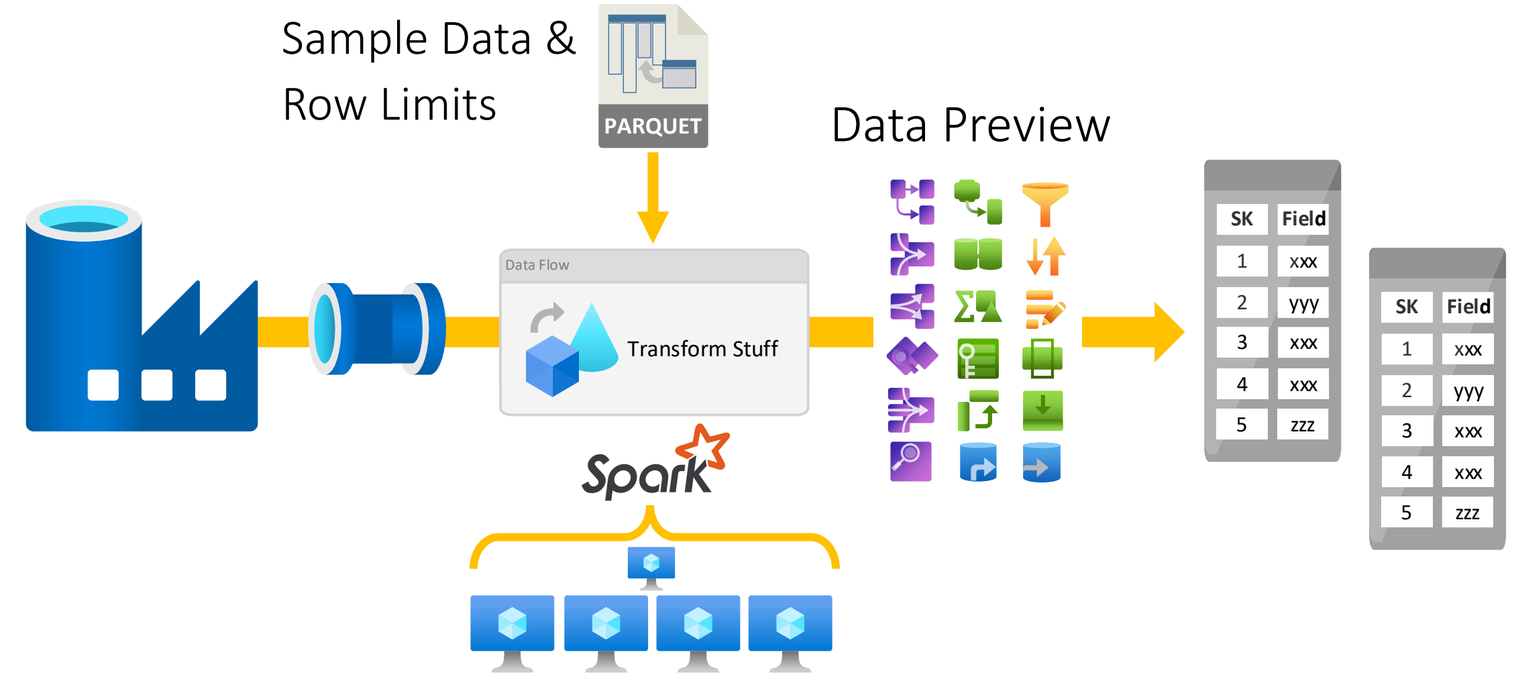



Dynamic Ranges



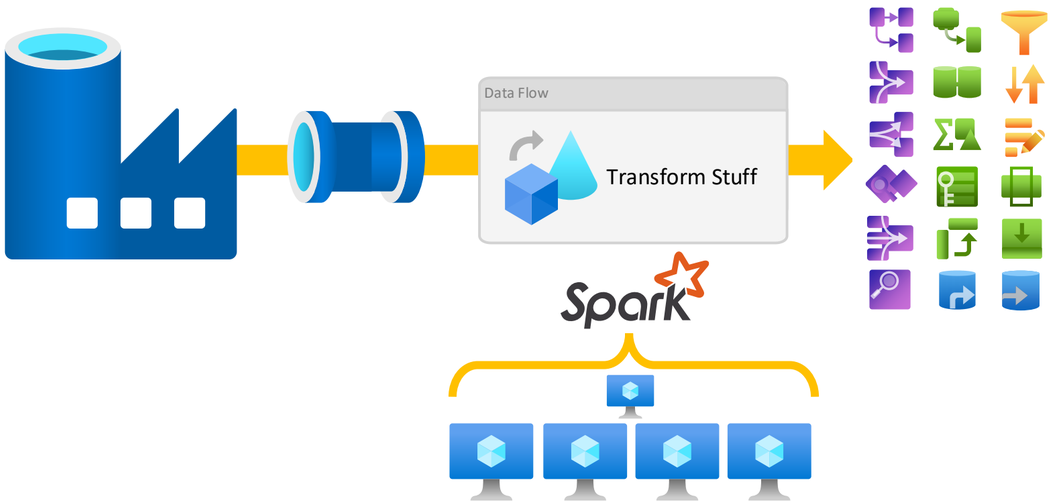

Key Based

# Data Flows – Debugging

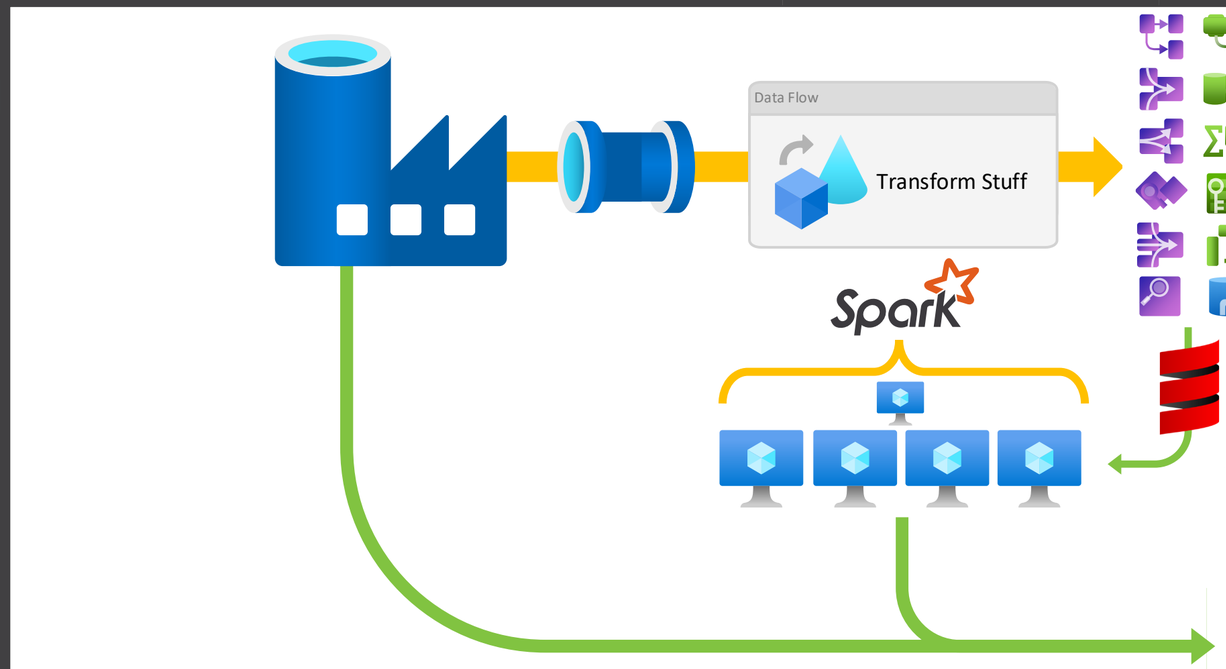


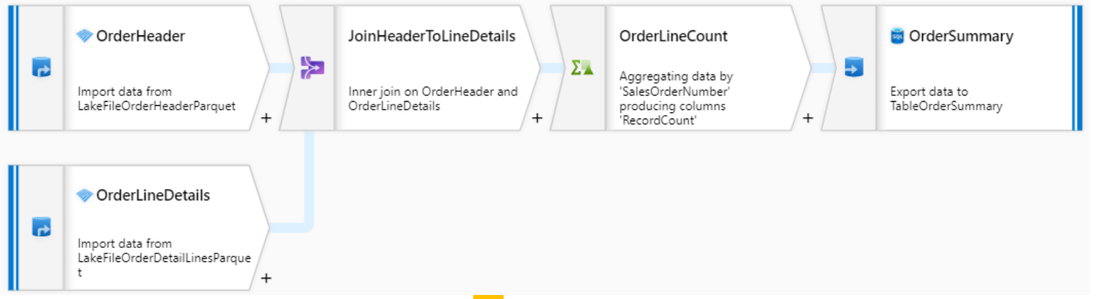

Enable Data Flow Debug Mode

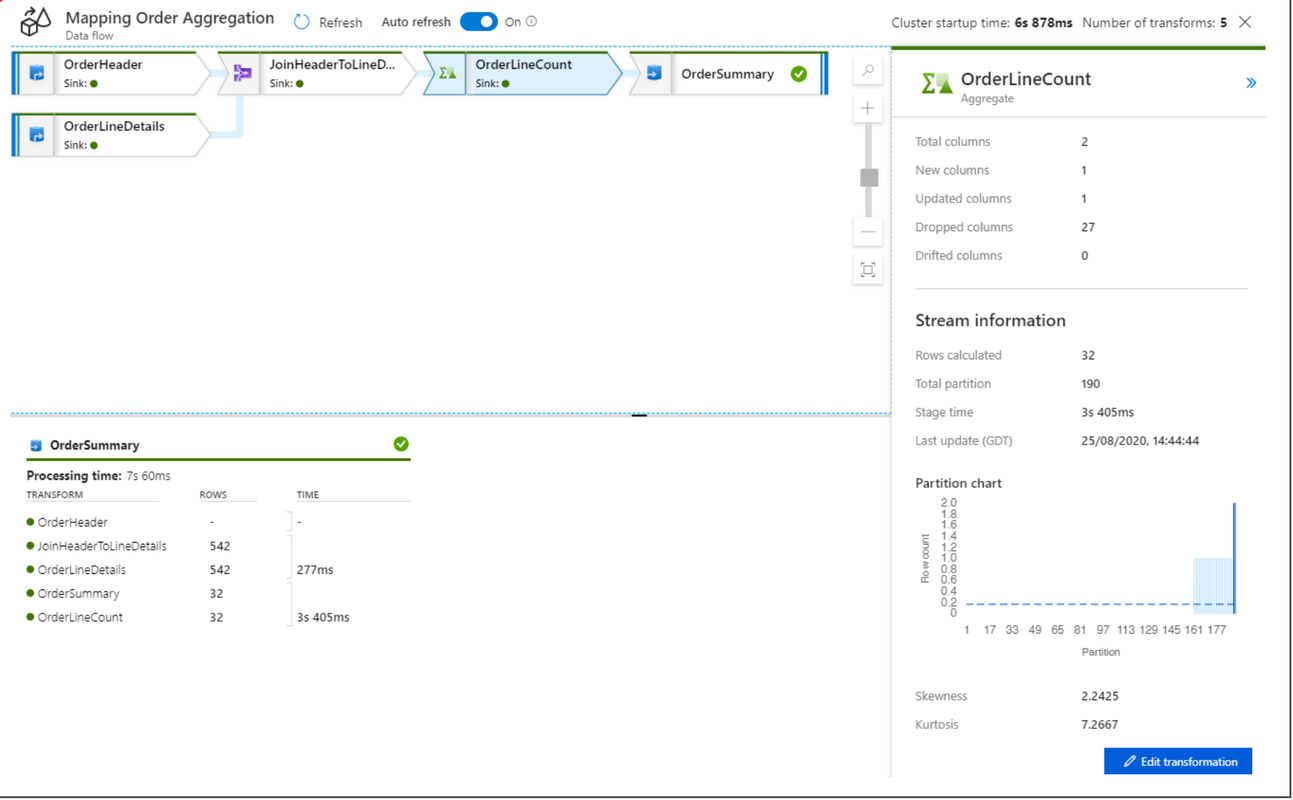
# Data Flows – Monitoring



# Data Flows – Monitoring

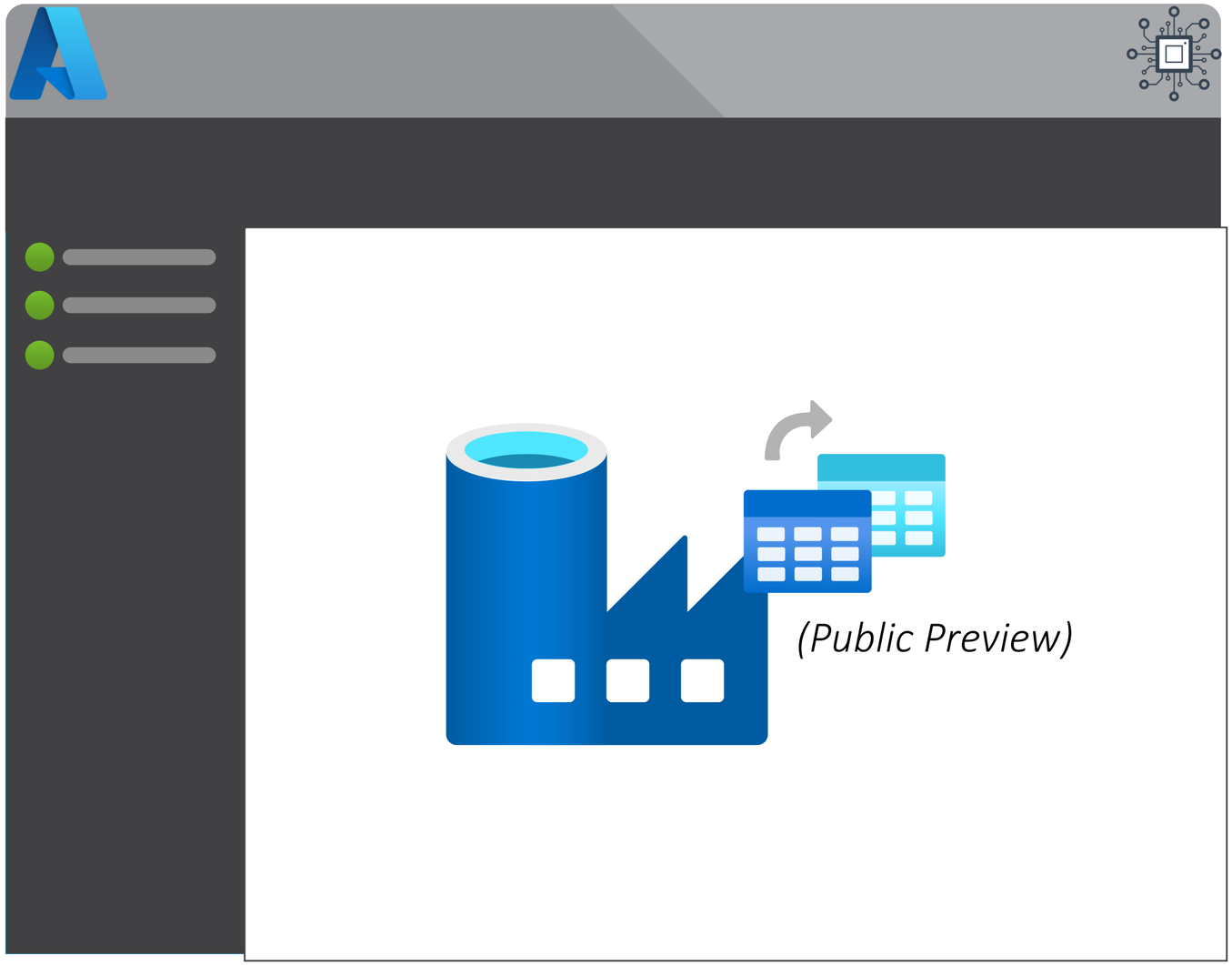


# An Introduction to Azure Data Factory

## Data Transformation

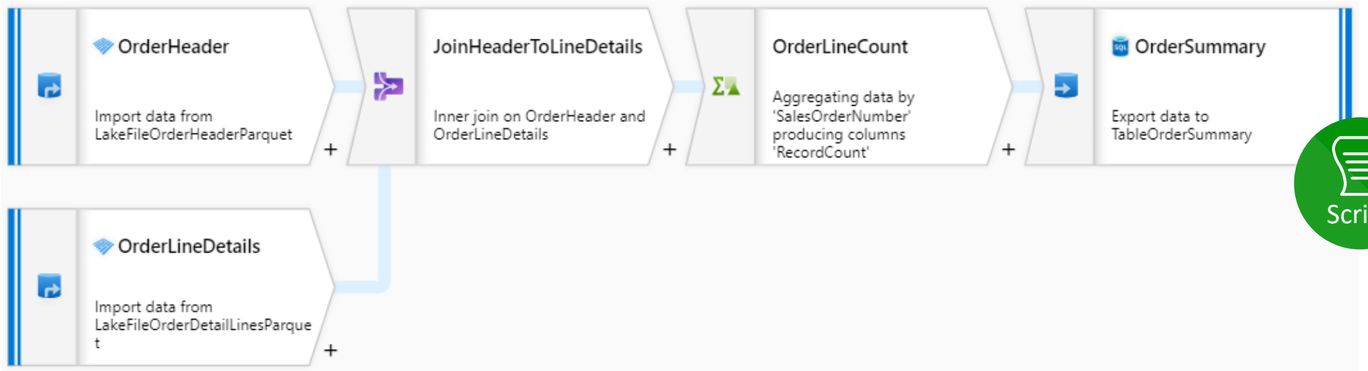


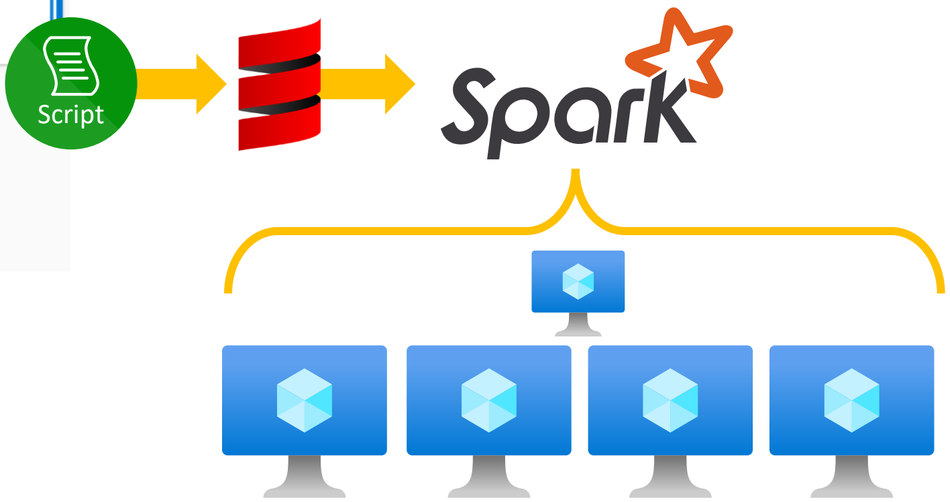

- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

# What is a Data Flow?

Control Flow

Data Flow



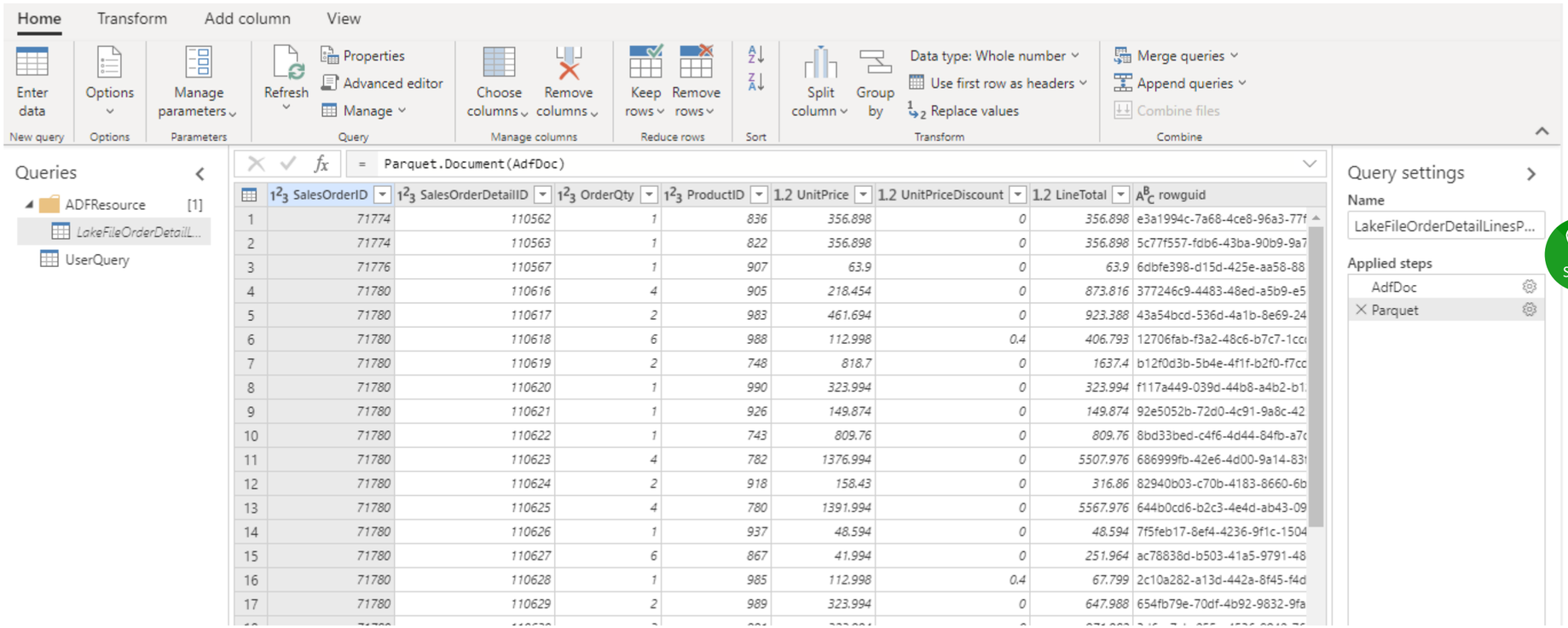



# What is a Power Query Activity?

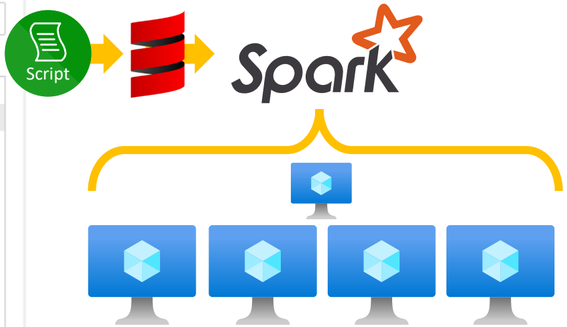
# What is a Power Query Activity?

Control Flow

Power Query



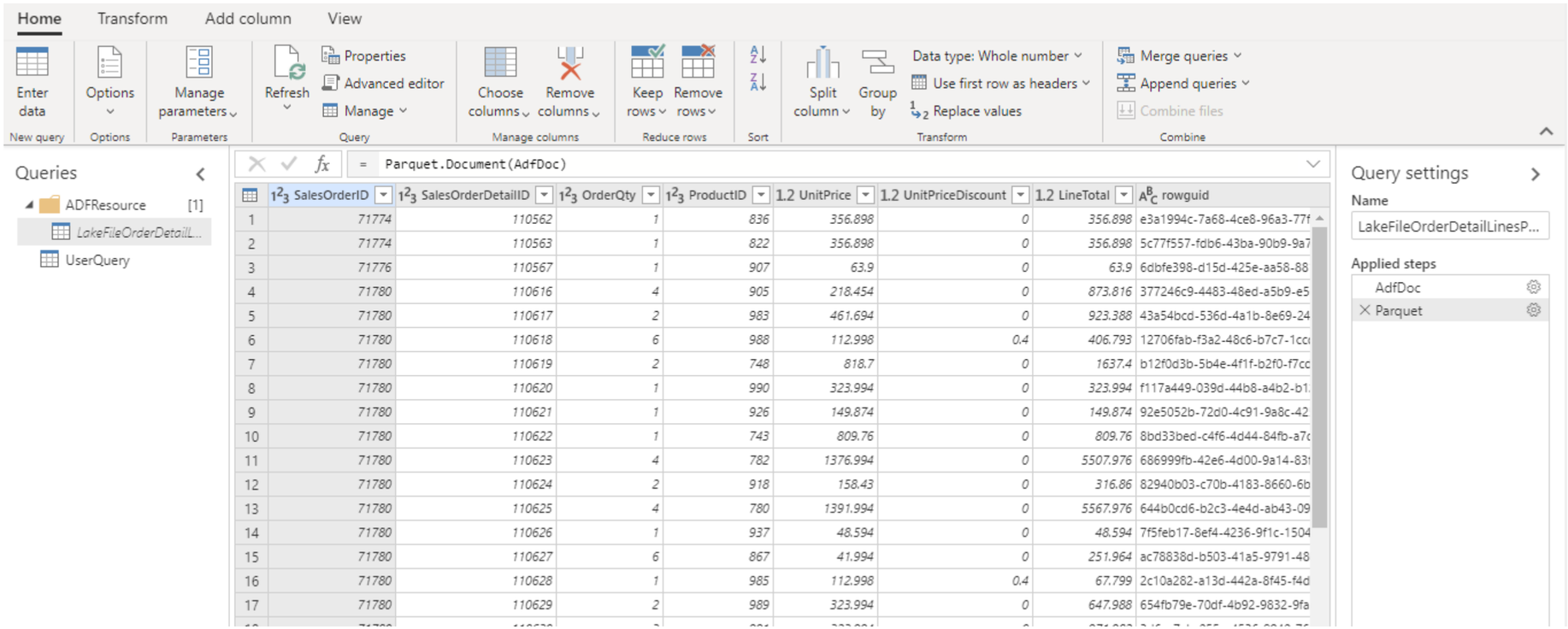

| | SalesOrderID | SalesOrderDetailID | OrderQty | ProductID | UnitPrice | UnitPriceDiscount | LineTotal | rowguid |
|---|---|---|---|---|---|---|---|---|
| 1 | 71774 | 110562 | 1 | 836 | 356.898 | 0 | 356.898 | e3a1994c-7a68-4ce8-96a3-77f |
| 2 | 71774 | 110563 | 1 | 822 | 356.898 | 0 | 356.898 | 5c77f557-fdb6-43ba-90b9-9a7 |
| 3 | 71776 | 110567 | 1 | 907 | 63.9 | 0 | 63.9 | 6dbfe398-d15d-425e-aa58-88 |
| 4 | 71780 | 110616 | 4 | 905 | 218.454 | 0 | 873.816 | 377246c9-4483-48ed-a5b9-e5 |
| 5 | 71780 | 110617 | 2 | 983 | 461.694 | 0 | 923.388 | 43a54bcd-536d-4a1b-8e69-24 |
| 6 | 71780 | 110618 | 6 | 988 | 112.998 | 0.4 | 406.793 | 12706fab-f3a2-48c6-b7c7-1cc |
| 7 | 71780 | 110619 | 2 | 748 | 818.7 | 0 | 1637.4 | b12f0d3b-5b4e-4f1f-b2f0-f7cc |
| 8 | 71780 | 110620 | 1 | 990 | 323.994 | 0 | 323.994 | f117a449-039d-44b8-a4b2-b1 |
| 9 | 71780 | 110621 | 1 | 926 | 149.874 | 0 | 149.874 | 92e5052b-72d0-4c91-9a8c-42 |
| 10 | 71780 | 110622 | 1 | 743 | 809.76 | 0 | 809.76 | 8bd33bed-c4f6-4d44-84fb-a7c |
| 11 | 71780 | 110623 | 4 | 782 | 1376.994 | 0 | 5507.976 | 686999fb-42e6-4d00-9a14-83 |
| 12 | 71780 | 110624 | 2 | 918 | 158.43 | 0 | 316.86 | 82940b03-c70b-4183-8660-6b |
| 13 | 71780 | 110625 | 4 | 780 | 1391.994 | 0 | 5567.976 | 644b0cd6-b2c3-4e4d-ab43-09 |
| 14 | 71780 | 110626 | 1 | 937 | 48.594 | 0 | 48.594 | 7f5feb17-8ef4-4236-9f1c-1504 |
| 15 | 71780 | 110627 | 6 | 867 | 41.994 | 0 | 251.964 | ac78838d-b503-41a5-9791-48 |
| 16 | 71780 | 110628 | 1 | 985 | 112.998 | 0.4 | 67.799 | 2c10a282-a13d-442a-8f45-f4d |
| 17 | 71780 | 110629 | 2 | 989 | 323.994 | 0 | 647.988 | 654fb79e-70df-4b92-9832-9fa |



# What can a Power Query Activity do?

Control Flow

Power Query



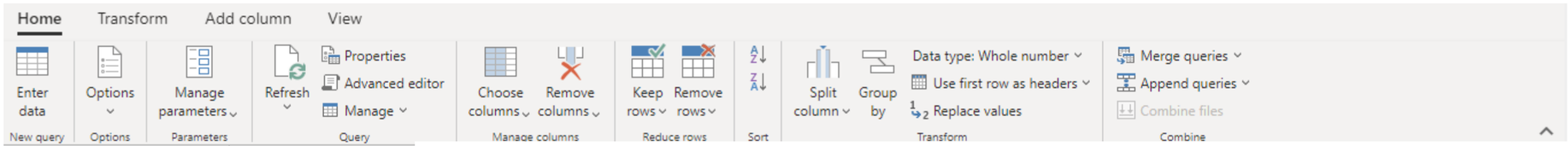

| | SalesOrderID | SalesOrderDetailID | OrderQty | ProductID | UnitPrice | UnitPriceDiscount | LineTotal | rowguid |
|---|---|---|---|---|---|---|---|---|
| 1 | 71774 | 110562 | 1 | 836 | 356.898 | 0 | 356.898 | e3a1994c-7a68-4ce8-96a3-77f |
| 2 | 71774 | 110563 | 1 | 822 | 356.898 | 0 | 356.898 | 5c77f557-fdb6-43ba-90b9-9a7 |
| 3 | 71776 | 110567 | 1 | 907 | 63.9 | 0 | 63.9 | 6dbfe398-d15d-425e-aa58-88 |
| 4 | 71780 | 110616 | 4 | 905 | 218.454 | 0 | 873.816 | 377246c9-4483-48ed-a5b9-e5 |
| 5 | 71780 | 110617 | 2 | 983 | 461.694 | 0 | 923.388 | 43a54bcd-536d-4a1b-8e69-24 |
| 6 | 71780 | 110618 | 6 | 988 | 112.998 | 0.4 | 406.793 | 12706fab-f3a2-48c6-b7c7-1cc |
| 7 | 71780 | 110619 | 2 | 748 | 818.7 | 0 | 1637.4 | b12f0d3b-5b4e-4f1f-b2f0-f7cc |
| 8 | 71780 | 110620 | 1 | 990 | 323.994 | 0 | 323.994 | f117a449-039d-44b8-a4b2-b1. |
| 9 | 71780 | 110621 | 1 | 926 | 149.874 | 0 | 149.874 | 92e5052b-72d0-4c91-9a8c-42 |
| 10 | 71780 | 110622 | 1 | 743 | 809.76 | 0 | 809.76 | 8bd33bed-c4f6-4d44-84fb-a7c |
| 11 | 71780 | 110623 | 4 | 782 | 1376.994 | 0 | 5507.976 | 686999fb-42e6-4d00-9a14-83 |
| 12 | 71780 | 110624 | 2 | 918 | 158.43 | 0 | 316.86 | 82940b03-c70b-4183-8660-6b |
| 13 | 71780 | 110625 | 4 | 780 | 1391.994 | 0 | 5567.976 | 644b0cd6-b2c3-4e4d-ab43-09 |
| 14 | 71780 | 110626 | 1 | 937 | 48.594 | 0 | 48.594 | 7f5feb17-8ef4-4236-9f1c-1504 |
| 15 | 71780 | 110627 | 6 | 867 | 41.994 | 0 | 251.964 | ac78838d-b503-41a5-9791-48 |
| 16 | 71780 | 110628 | 1 | 985 | 112.998 | 0.4 | 67.799 | 2c10a282-a13d-442a-8f45-f4d |
| 17 | 71780 | 110629 | 2 | 989 | 323.994 | 0 | 647.988 | 654fb79e-70df-4b92-9832-9fa |

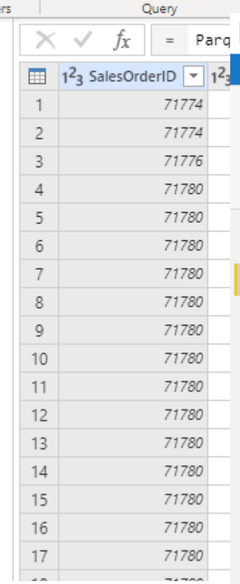
# What can a Power Query Activity do?

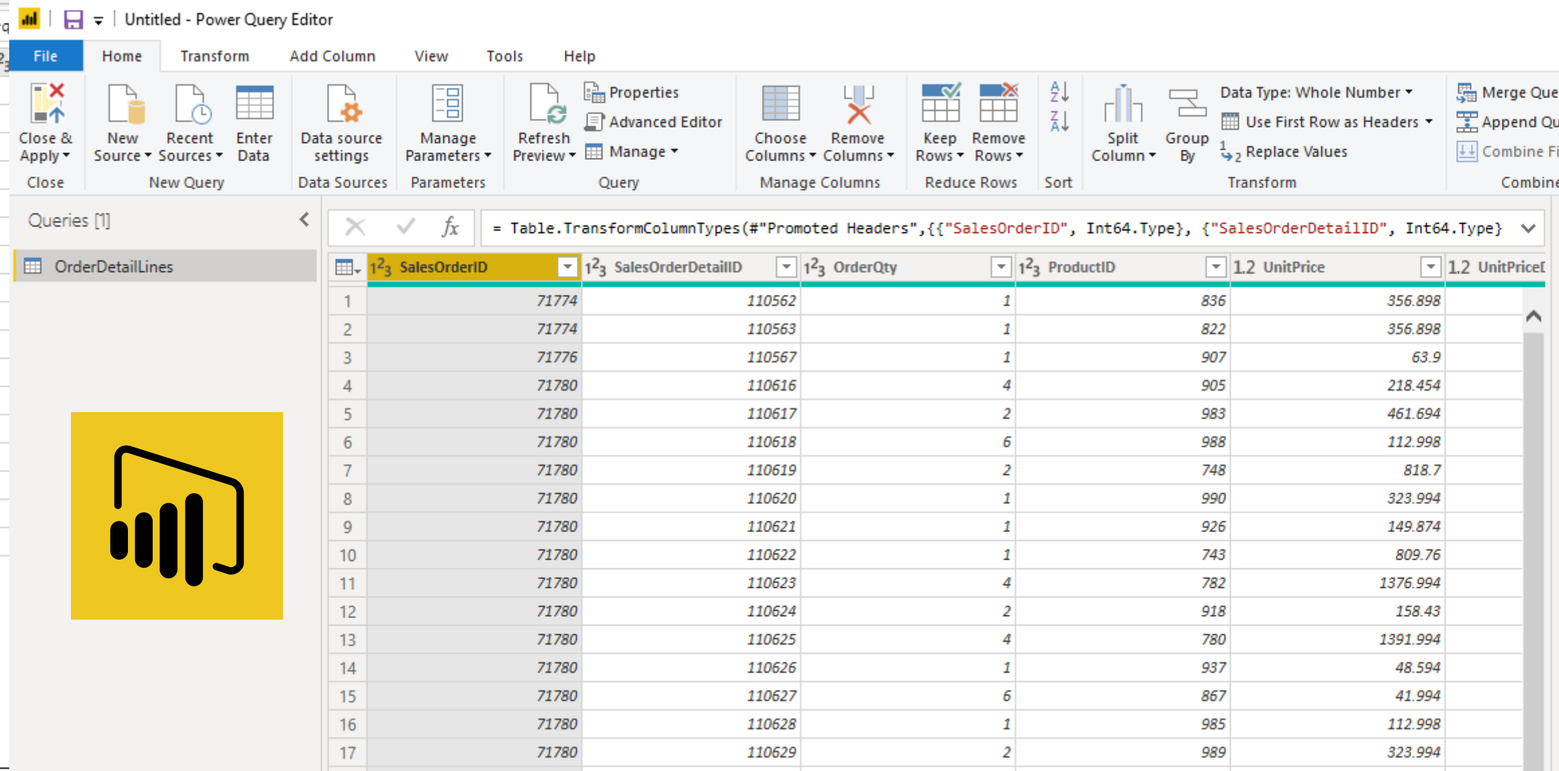
## Home

Control Flow

Power Query



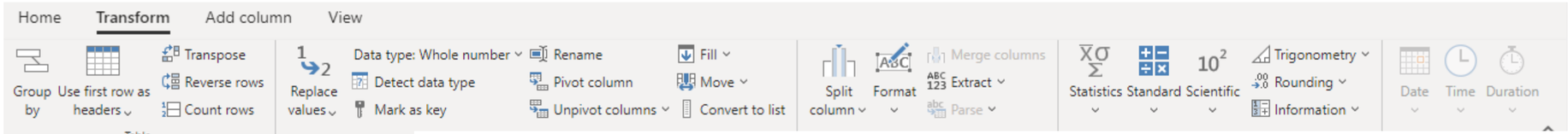



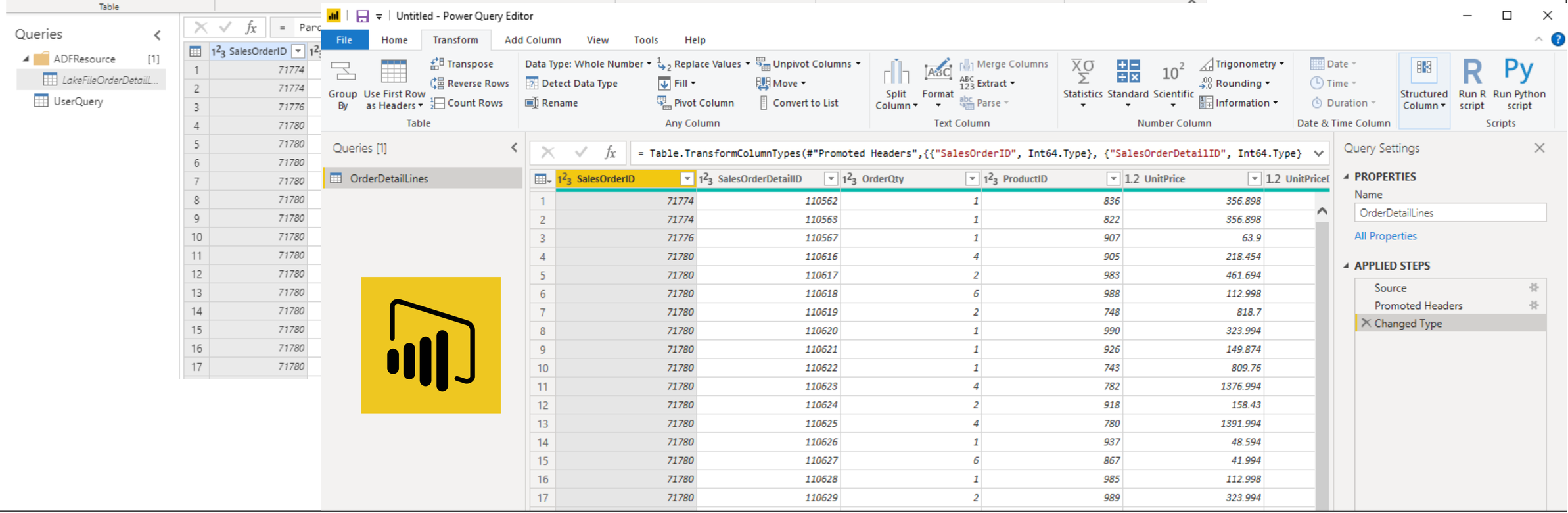
# What can a Power Query Activity do?

## Transform

Control Flow

Power Query



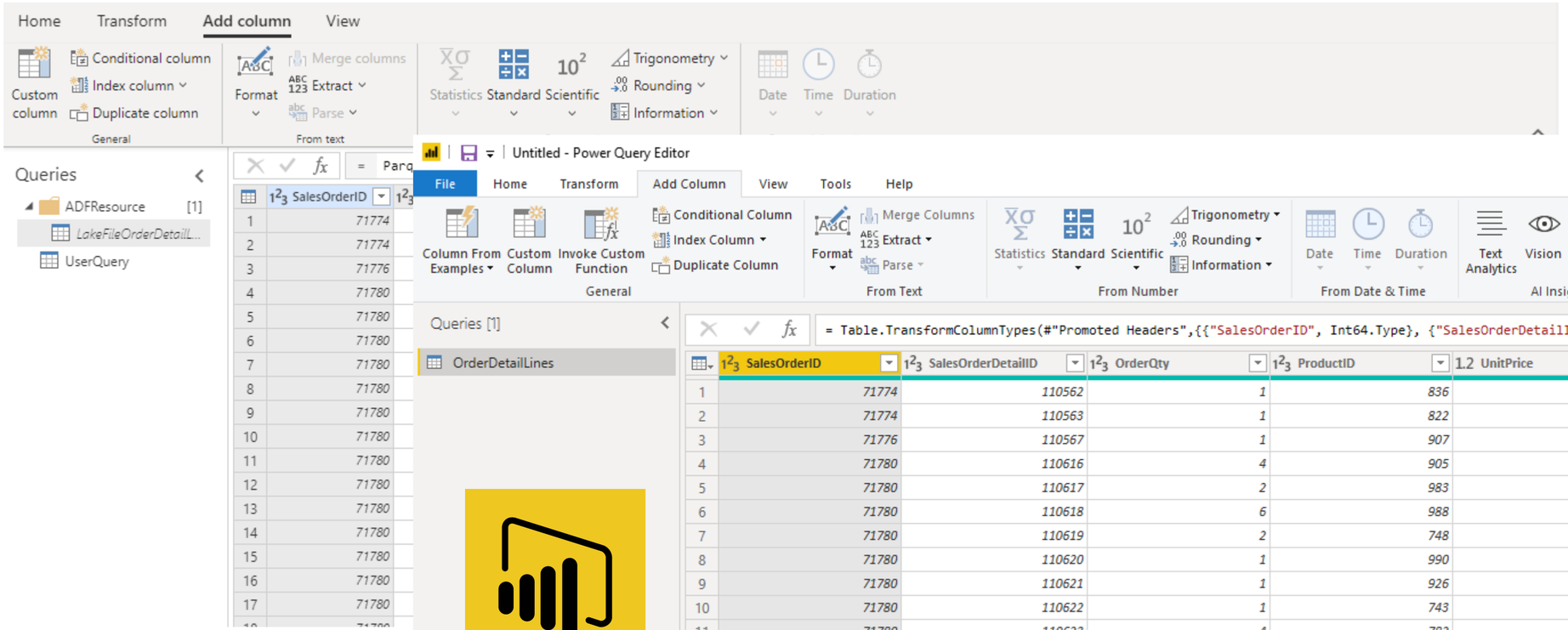



# What can a Power Query Activity do?

## Add Column

Control Flow

Power Query

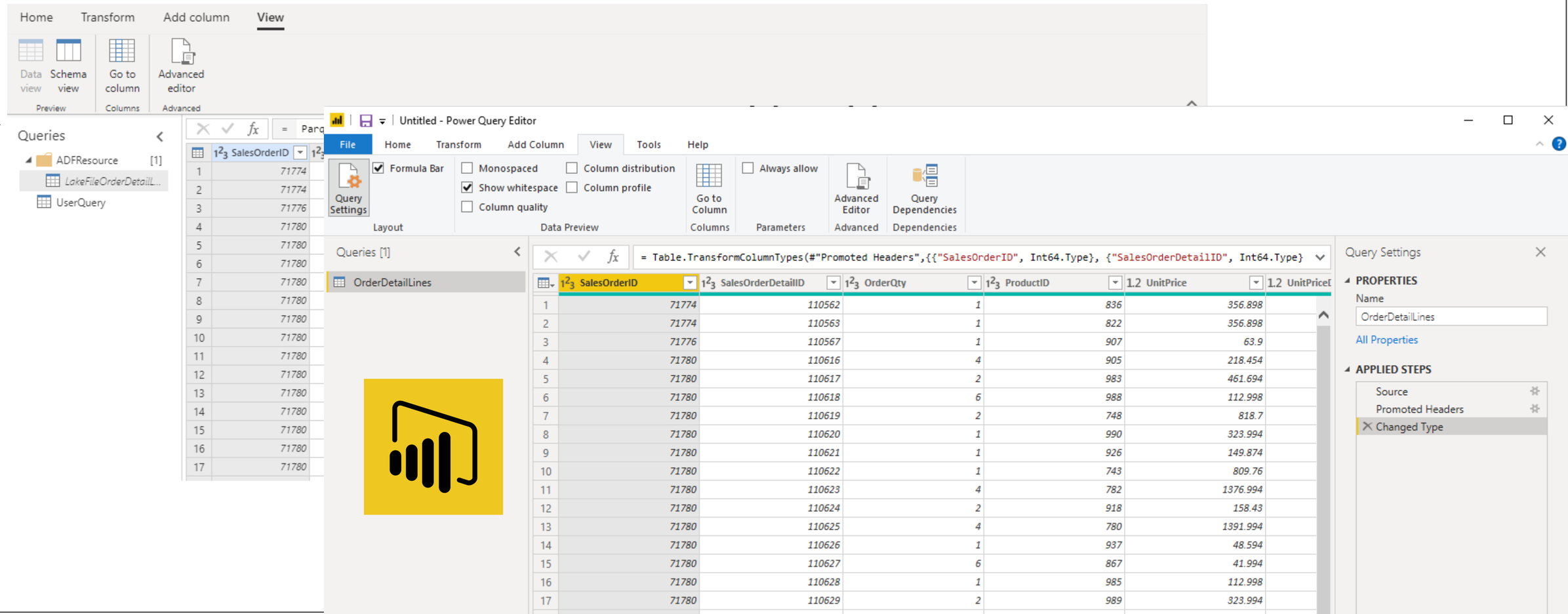


# What can a Power Query Activity do?

## View

Control Flow

Power Query



# What can a Power Query Activity do?

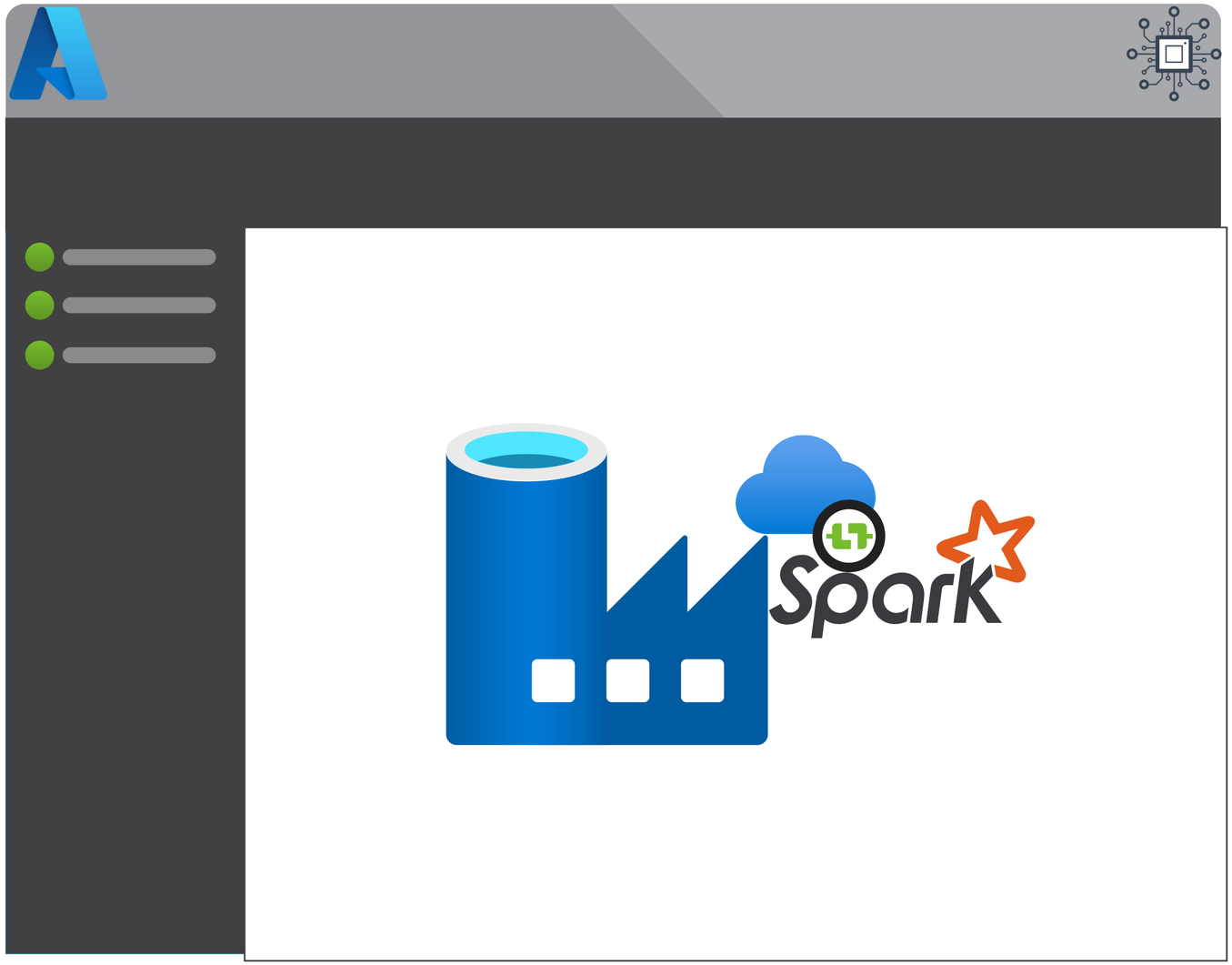
## View

Control Flow

Power Query

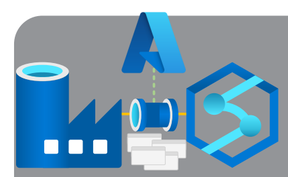
# An Introduction to Azure Data Factory

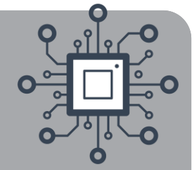
## Data Transformation



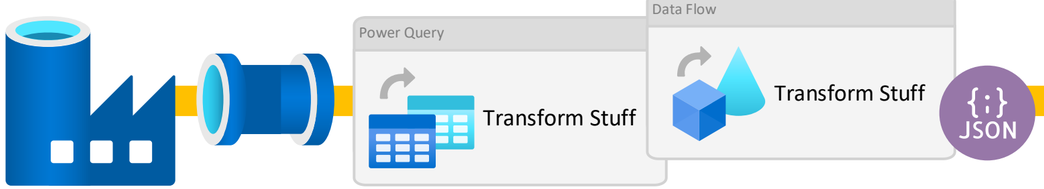

- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

# Spark Configuration





Control Flow



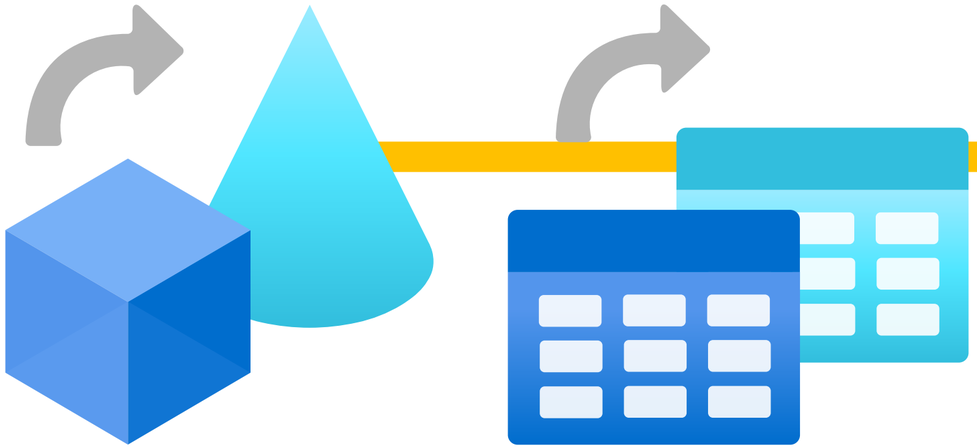
Default Azure IR

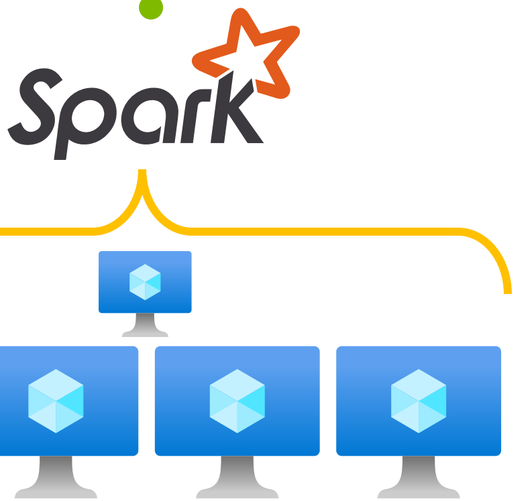
- General Purpose
- 4x Worker Nodes
- 1 Hour

Data Flow

or

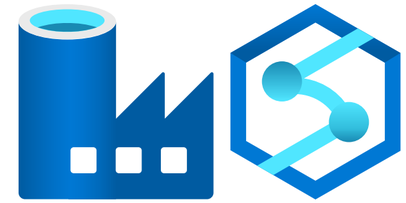
Power Query



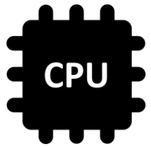



- Compute Type
- Number of Worker Nodes
- Cluster Time to Live

# Data Flow Compute – IR's vs Spark



Orchestrator




Integration Runtime (IR)

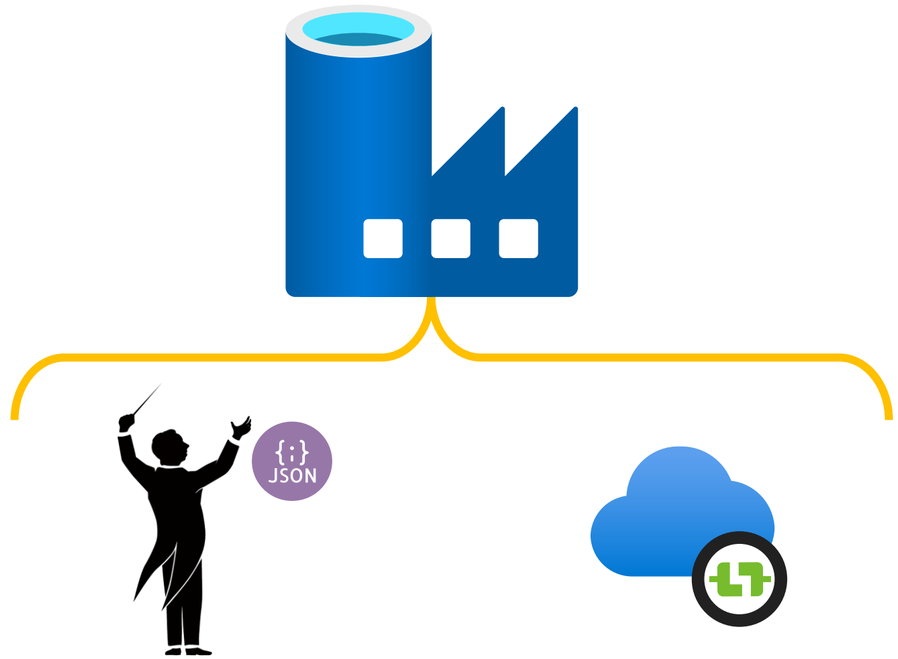
Azure

Hosted

Linked (Hosted)

SSIS

# What is an Integration Runtime?



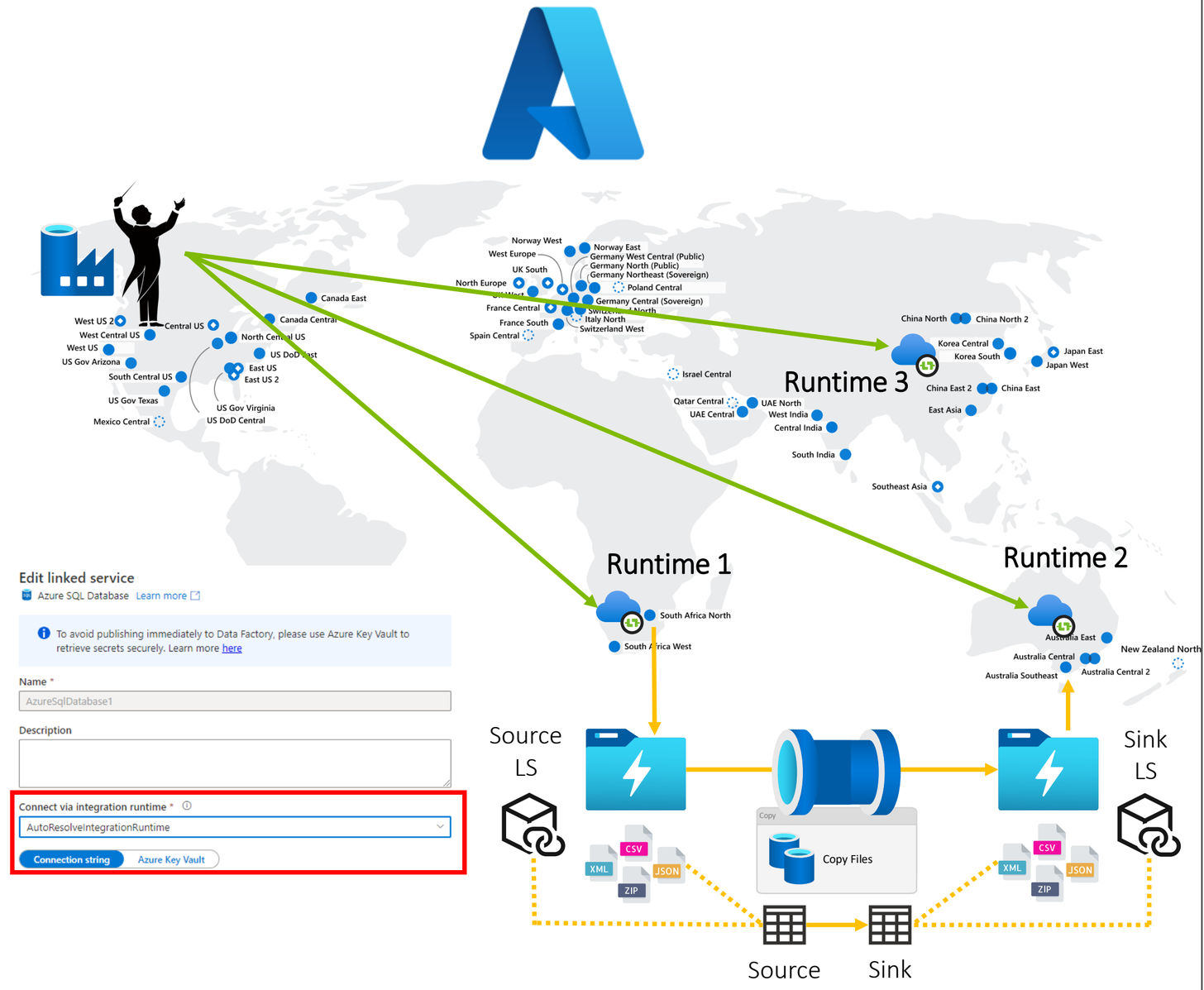

Orchestrator

Fixed Region

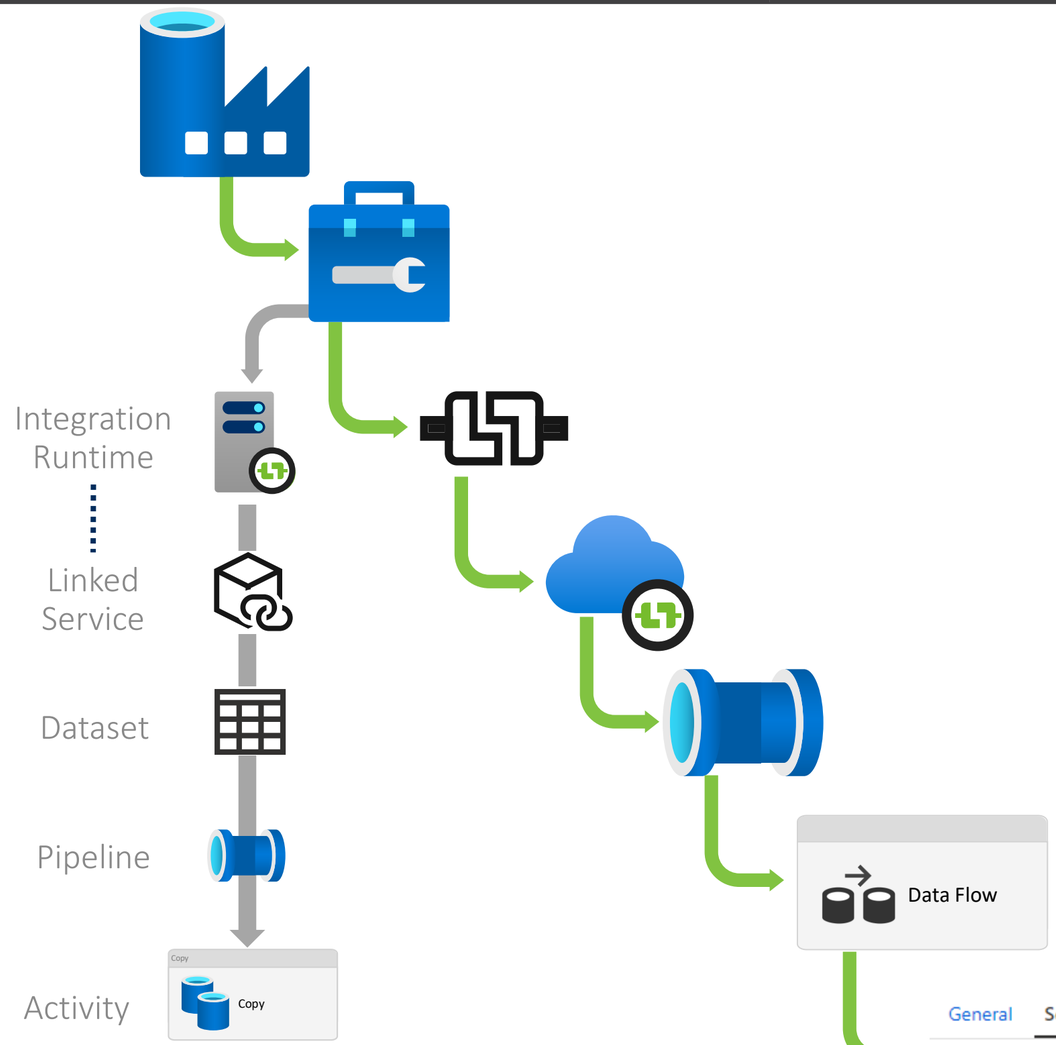
Runtime

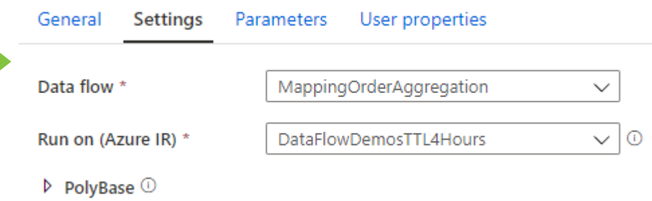
Flexible Location

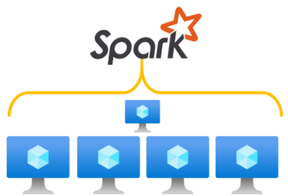
*AutoResolveIntegrationRuntime*



# Setting the Data Flow Cluster (IR Configuration)




Data Factory

Manage

Integration Runtimes

Azure IR

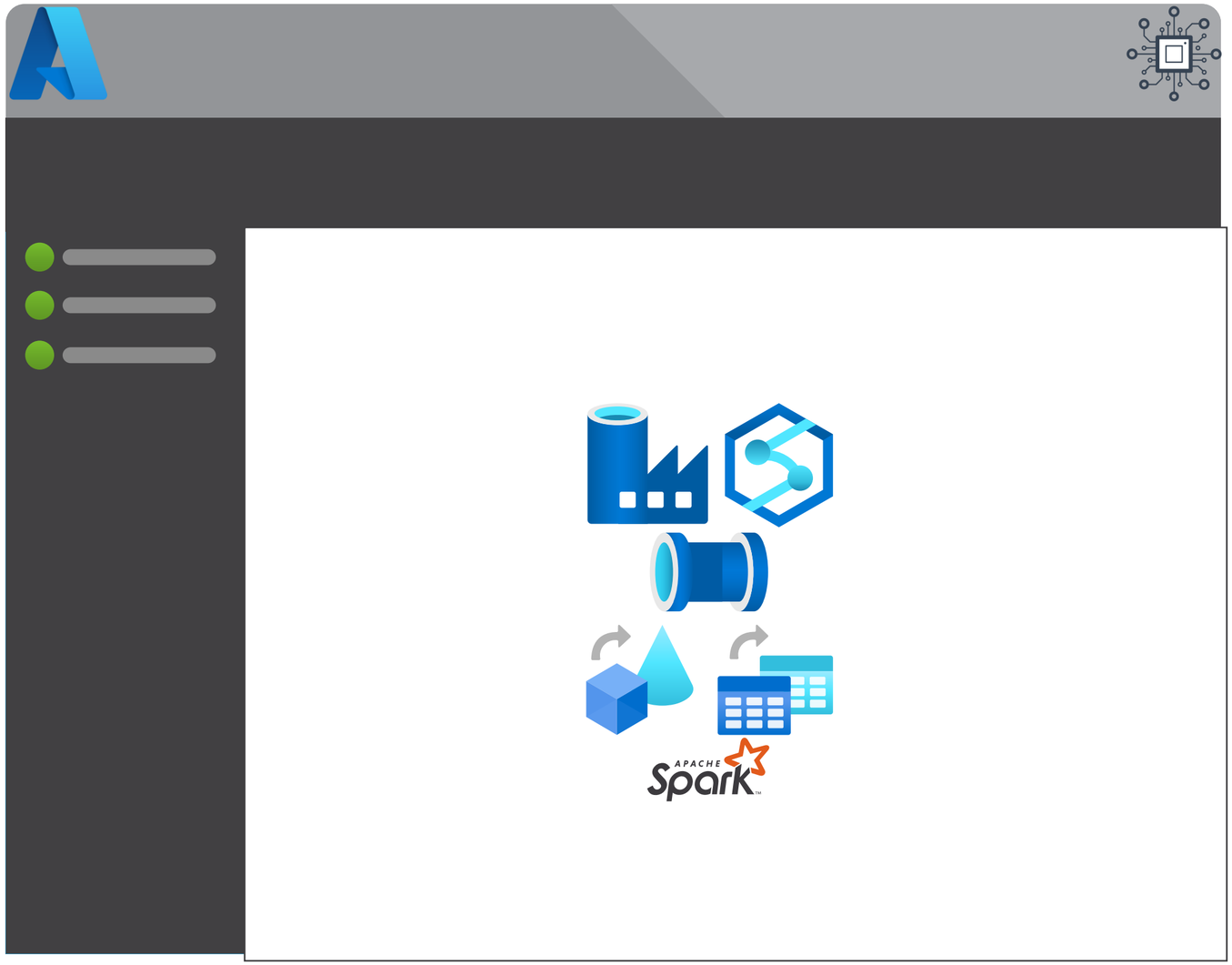
Pipeline

Data Flow Activity

Settings

# An Introduction to Azure Data Factory

## Data Transformation



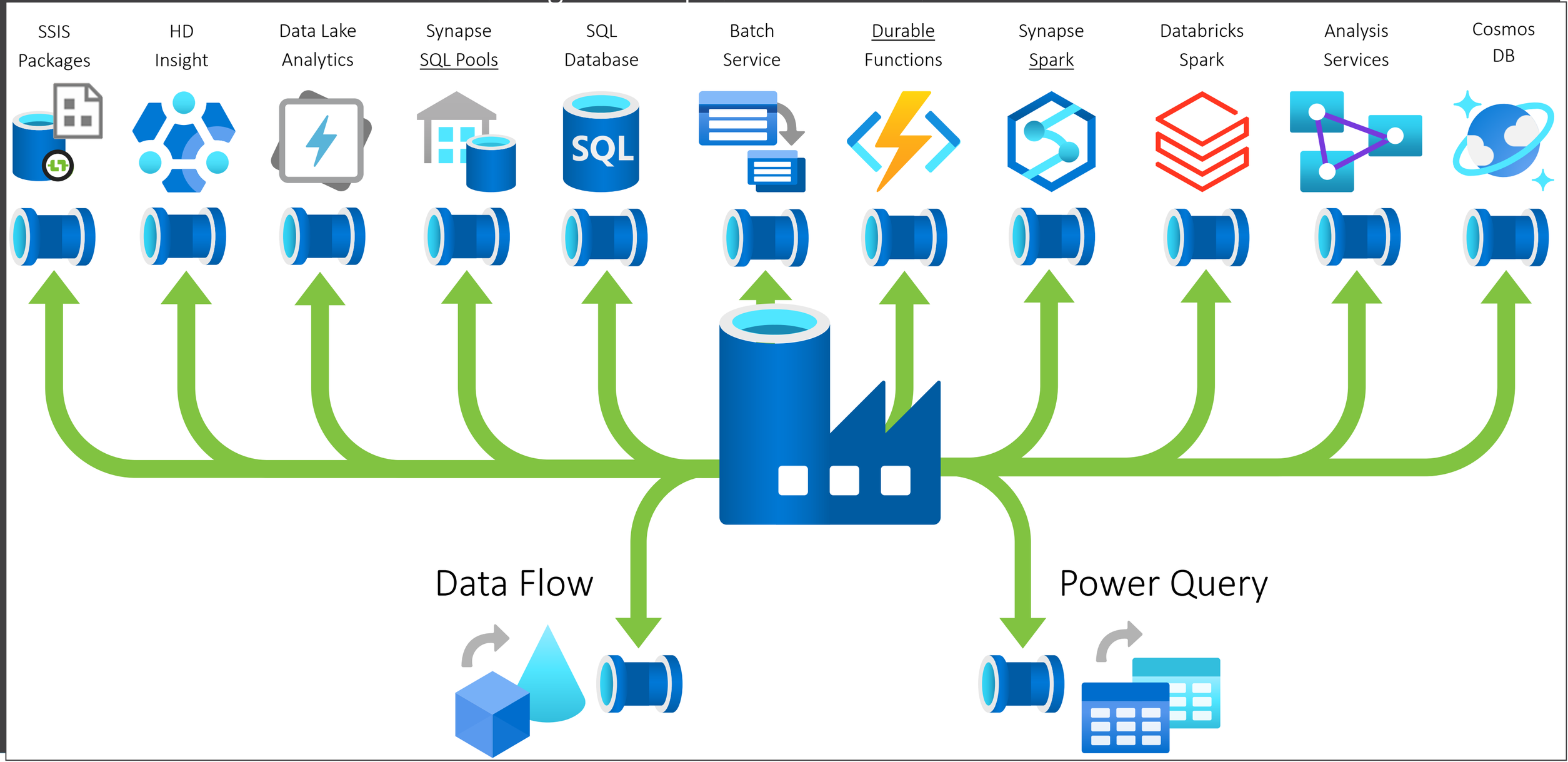

- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases

# Data Transformation Resources in Azure Comparison

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

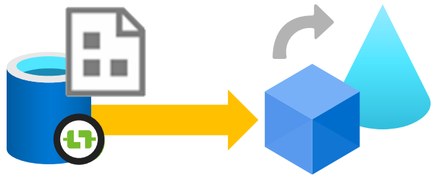
# Other Data Transformation Services in Azure

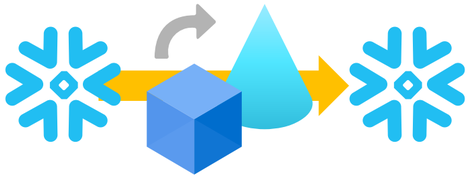
## When Should We Use These Integration Pipeline Transformation Activities?



# Use Cases

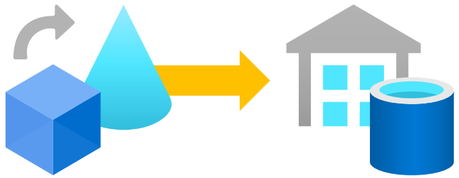
SSIS Package rebuild and skills migration.



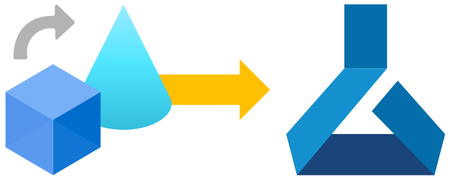


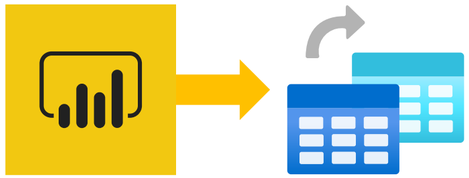
Inline dataset transformations.

Warehouse data distribution & loading.



Data model dataset preparation.

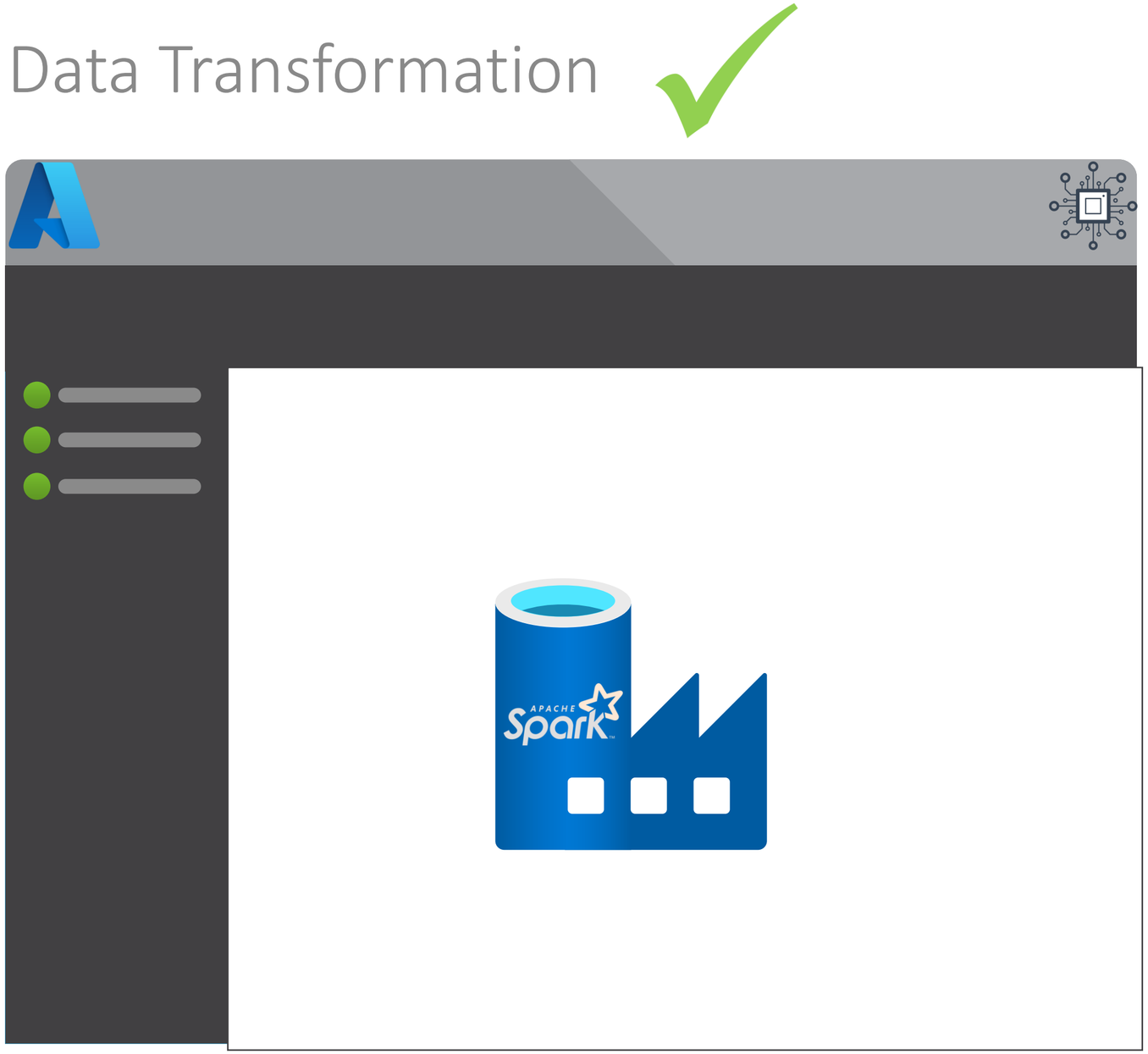




Power Query industrialisation.

# An Introduction to Azure Data Factory

## Data Transformation




- Data Flows
- Power Query Injection
- Spark Configuration
- Use Cases