ROBERT WALTERS

Robert Walters Technology

‘Lets Talk-Data Engineering’

Friday 20th August 2021

“An Introduction to Azure Data Factory”

with Paul Andrew (Avanade, Microsoft MVP)

MVP Microsoft® Most Valuable Professional

ROBERT WALTERS

Robert Walters Technology

‘Lets Talk-Data Engineering’

Daniel Bone

Recruitment Consultant with 3 years experience across IT / BI / Data

Founder of the ‘Lets Talk – Data Engineering’ group

Email: Daniel.Bone@robertwalters.com
Phone number: 07766850780
LinkedIn: https://www.linkedin.com/in/daniel-bone-01a3b4199/

ROBERT WALTERS

Robert Walters Technology

‘Lets Talk-Data Engineering’

- Head of Business Intelligence / Analytics
- BI / Data Architect
- BI Manager
- BI Developer
- Data Analyst
- Data Engineer
- Data Visualisation Developer
- Data Scientist
- Machine Learning Engineer
- BI / Data Consultant

Email: Daniel.Bone@robertwalters.com
Phone number: 07766850780
LinkedIn: https://www.linkedin.com/in/daniel-bone-01a3b4199/

ROBERT WALTERS

Robert Walters Technology

‘Lets Talk-Data Engineering’

Enjoy Paul’s Session!

A Introduction to Azure Data Factory

Integration Pipelines

Paul Andrew | Technical Architect in Azure CoE

Microsoft® MVP Most Valuable Professional

avanade

@MrPaulAndrew In/MrPaulAndrew MrPaulAndrew.com

A Introduction to Azure ~~Data Factory~~

Integration Pipelines

Paul Andrew | Technical Architect in Azure CoE

Microsoft® Most Valuable Professional MVP

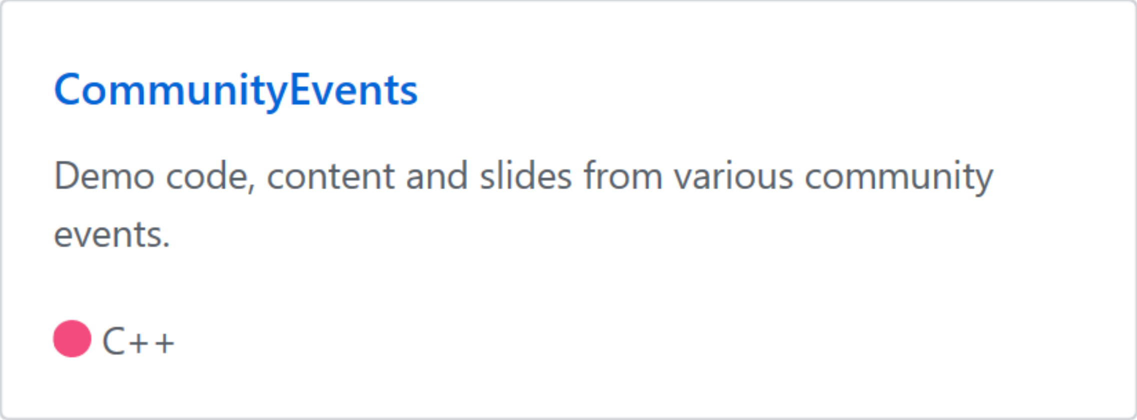
avanade

@MrPaulAndrew In/MrPaulAndrew MrPaulAndrew.com

Go deeper on a topic…

https://mrpaulandrew.com

https://github.com/mrpaulandrew



CommunityEvents

Demo code, content and slides from various community events.

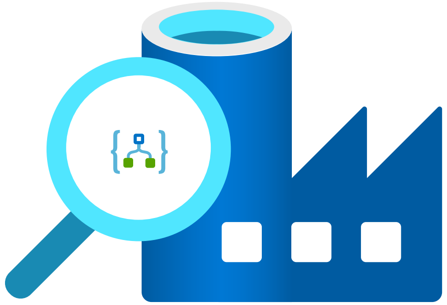
● C++

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

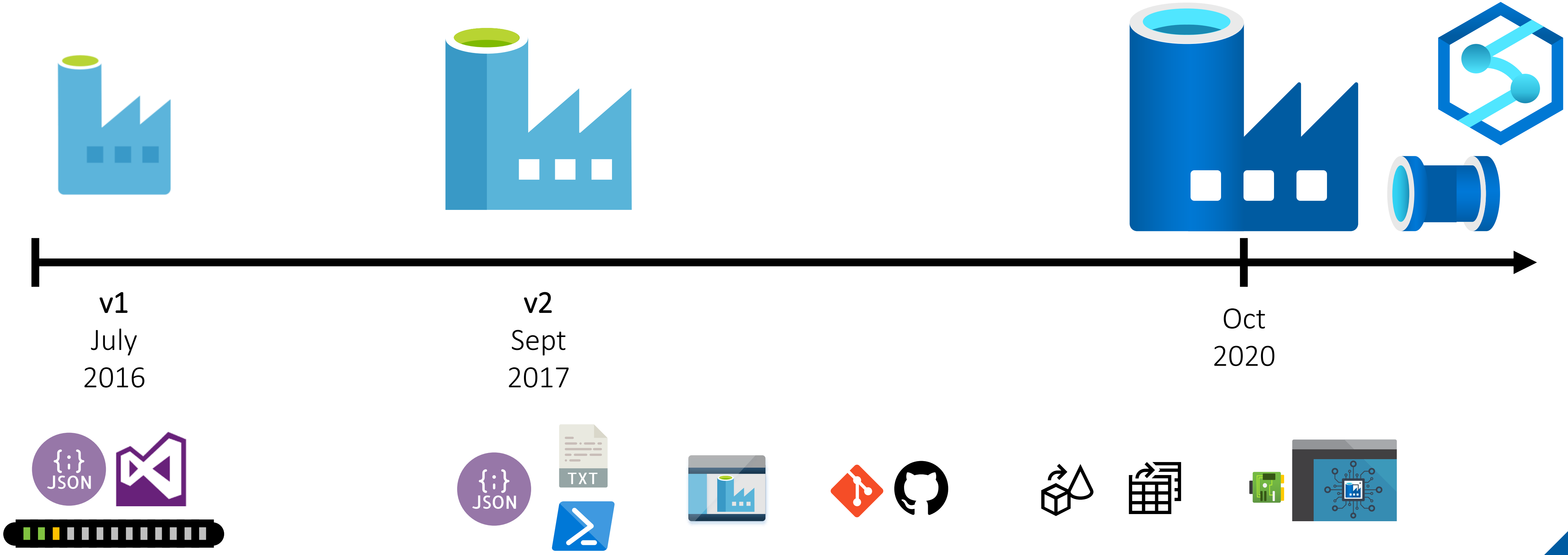
Agenda

- What is it and why use it?
- Data Factory Components
- Common Activities
- Execution Dependencies
- Integration Runtimes
    - Azure/Hosted/SSIS
- Data Factory Data Flows
- Source Control
- Deployments
- Monitoring & Logging
- Conclusions

Azure Data Factory – What is it? Why use it?



A Quick History Lesson



v1
July
2016

v2
Sept
2017

Oct
2020

Time Slices
VS2015 Projects

Notepad
Development

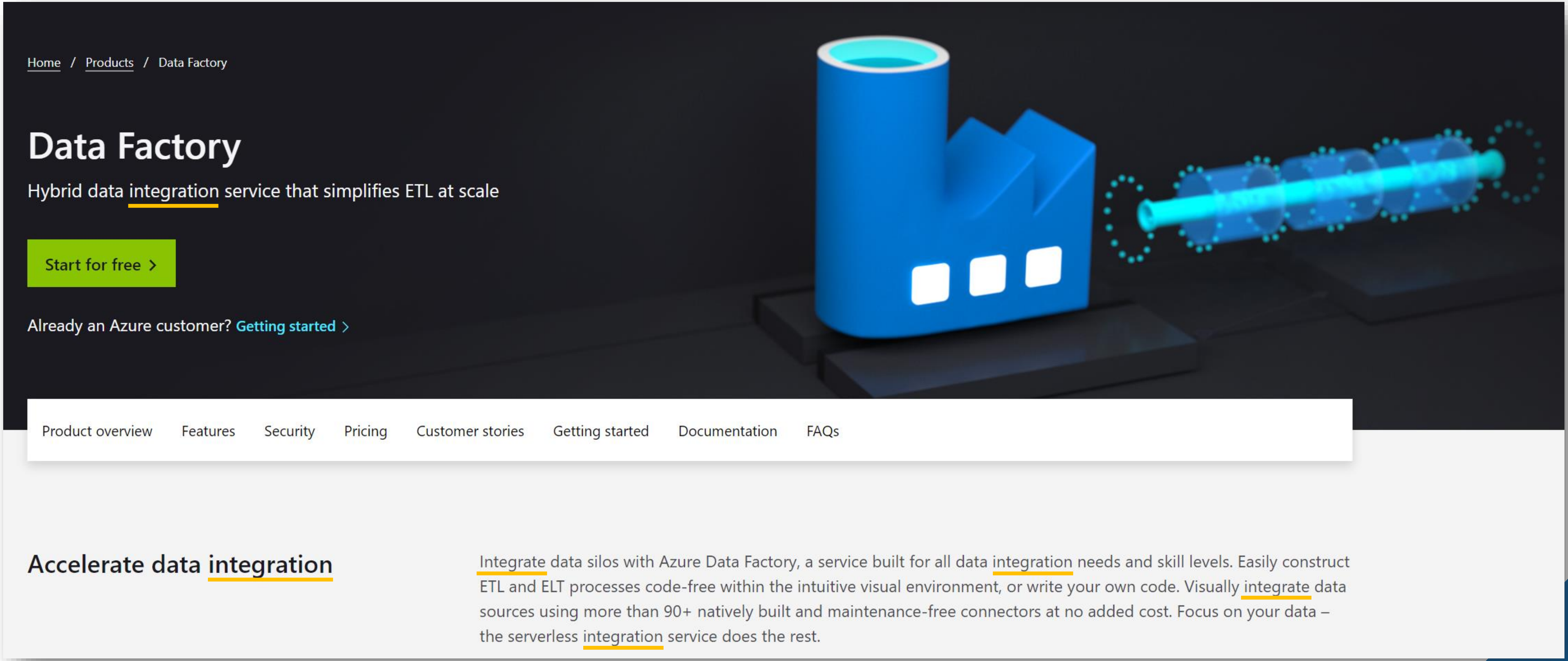
New UI

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 >

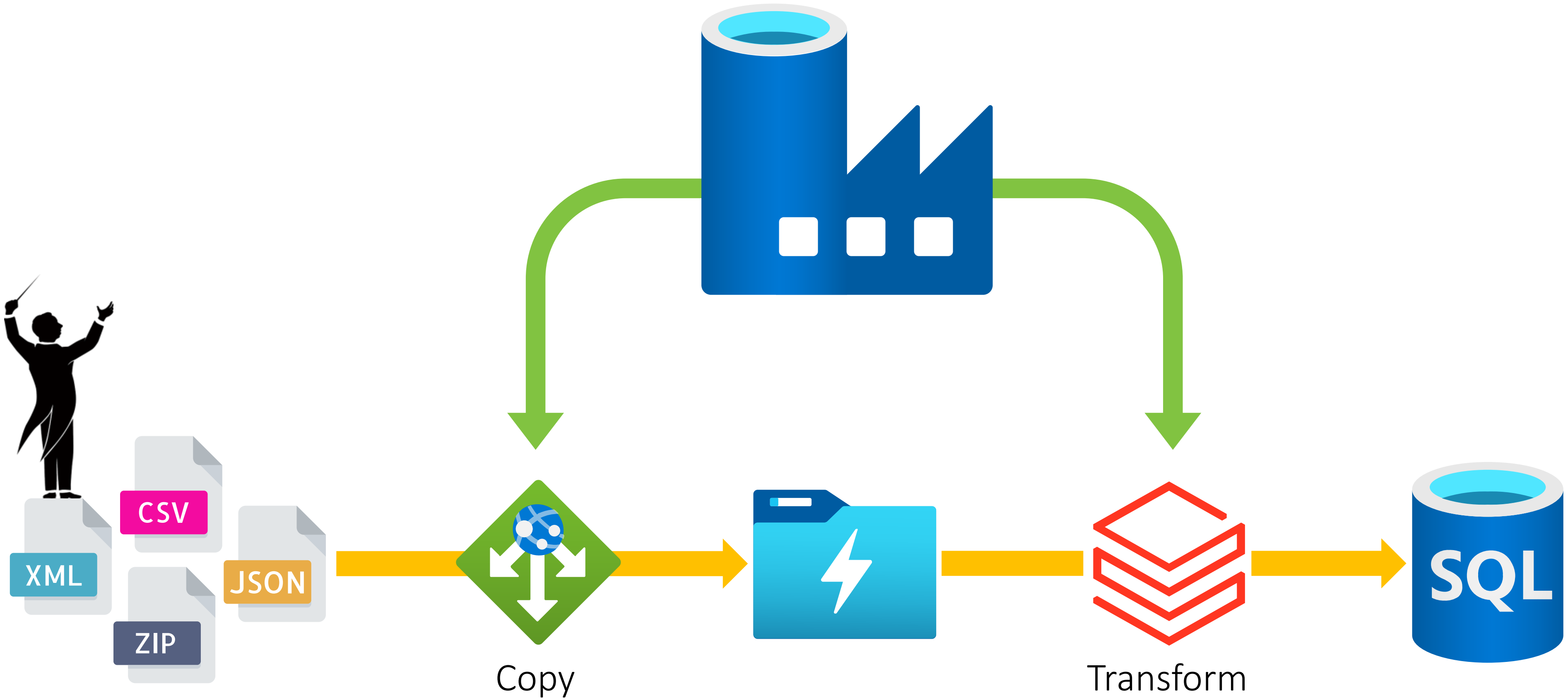
Product overview Features Security Pricing Customer stories Getting started Documentation FAQs

Accelerate data integration

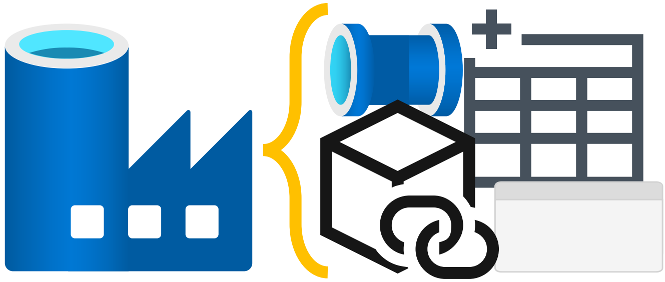
Integrate data silos with Azure Data Factory, a service built for all data integration needs and skill levels. Easily construct ETL and ELT processes code-free within the intuitive visual environment, or write your own code. Visually integrate data sources using more than 90+ natively built and maintenance-free 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

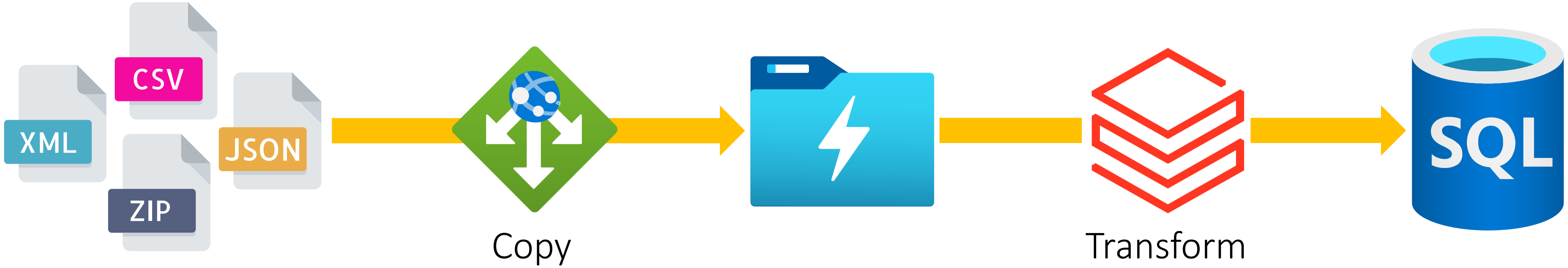
What is Azure Data Factory (ADF)?



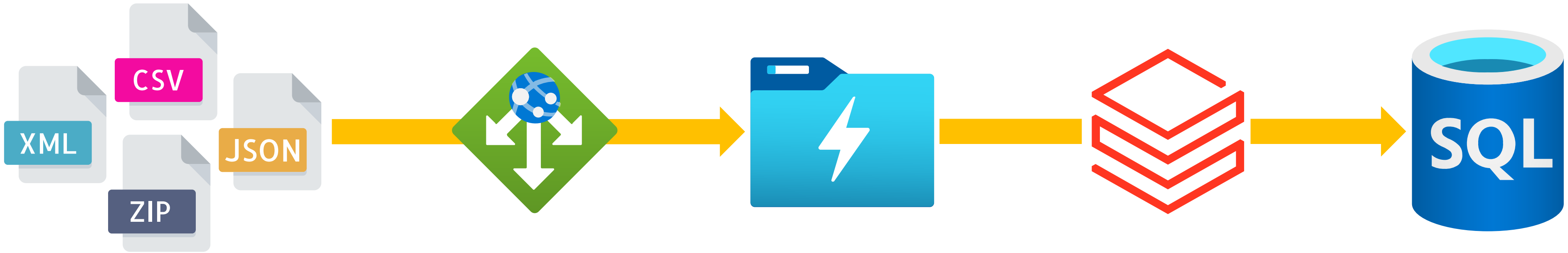
Data Factory Components



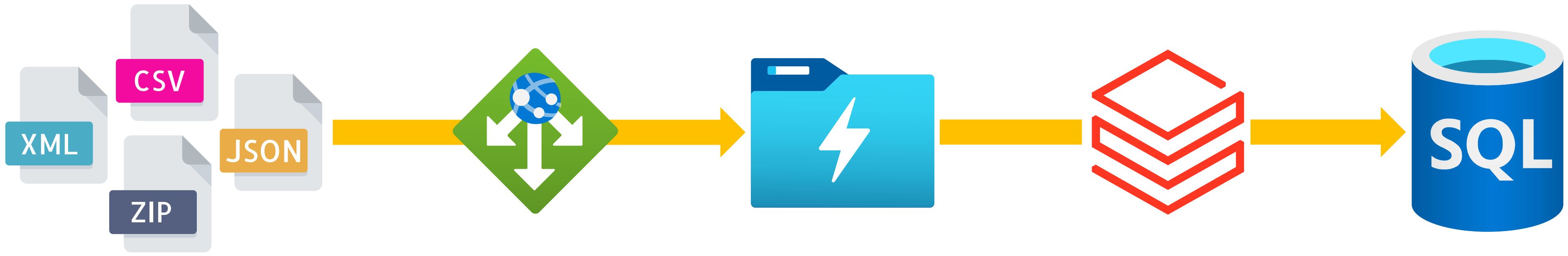
Data Factory Components



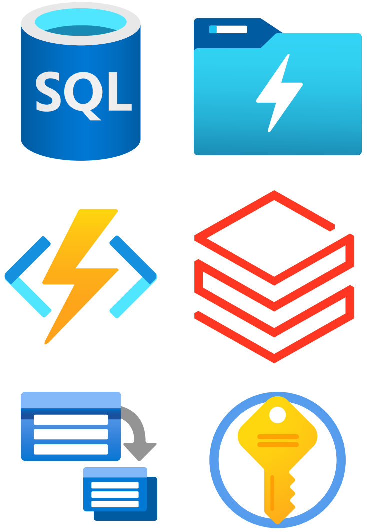
Data Factory Components



Data Factory Components



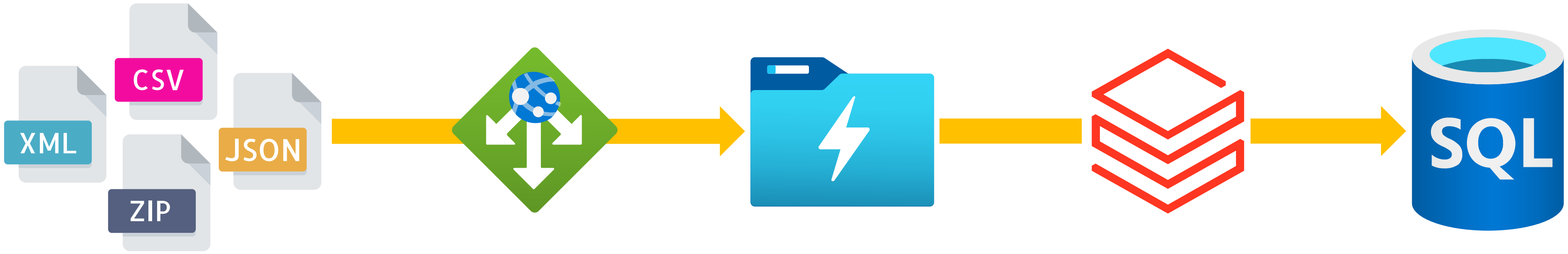
1 **Linked Services** – What to interact with and how?



SQLDBLinkedService

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

Data Factory Components



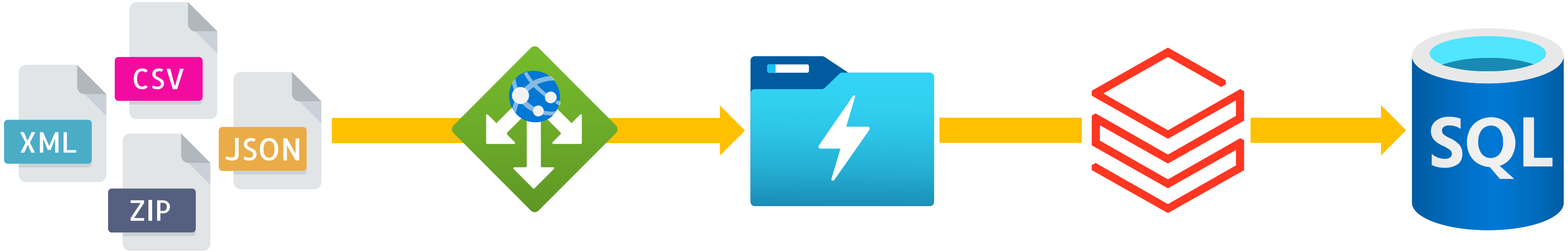
1 **Linked Services**

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

[dbo].[SalesOrders]

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

Data Factory Components



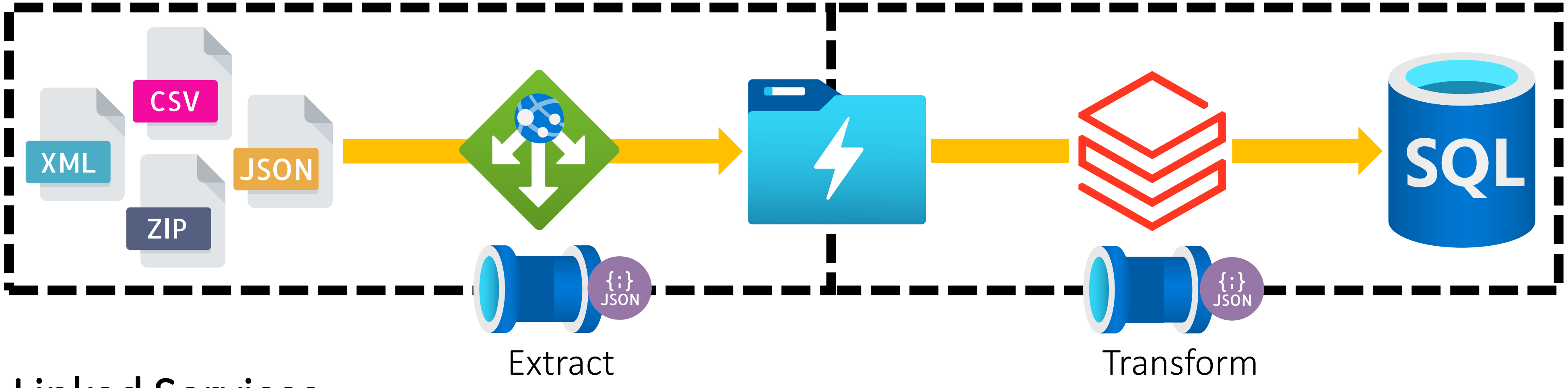
1. **Linked Services**
2. **Datasets**
3. **Activities** – What do we want to happen when we invoke a Linked Service?
   With what conditions?

{:}
JSON

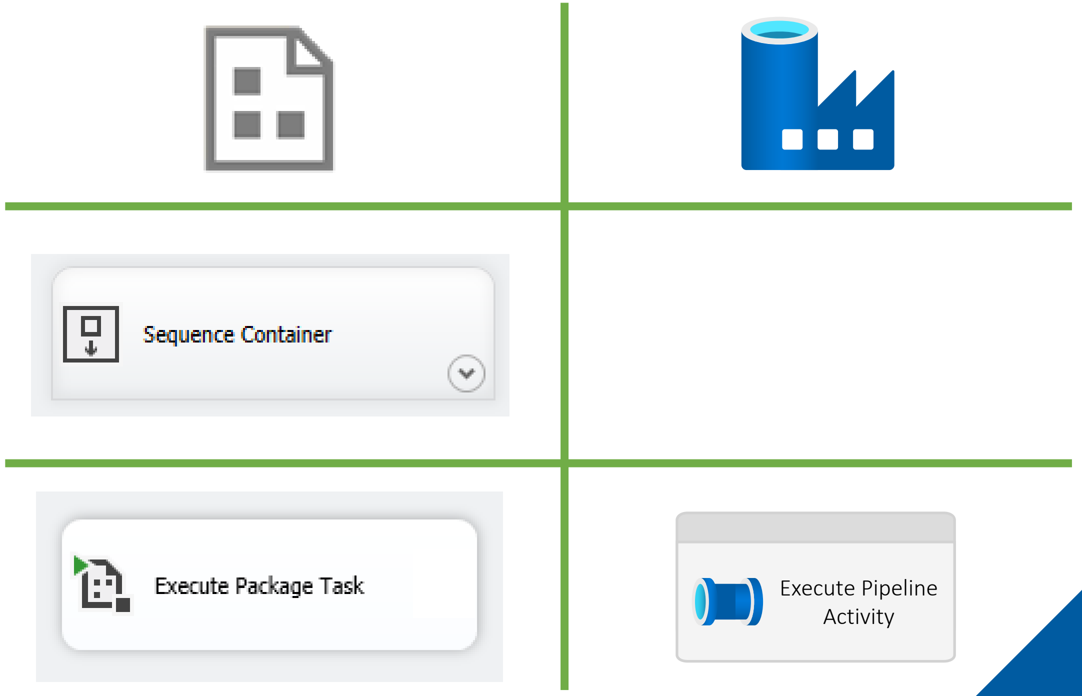
Databricks Notebook Activity

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

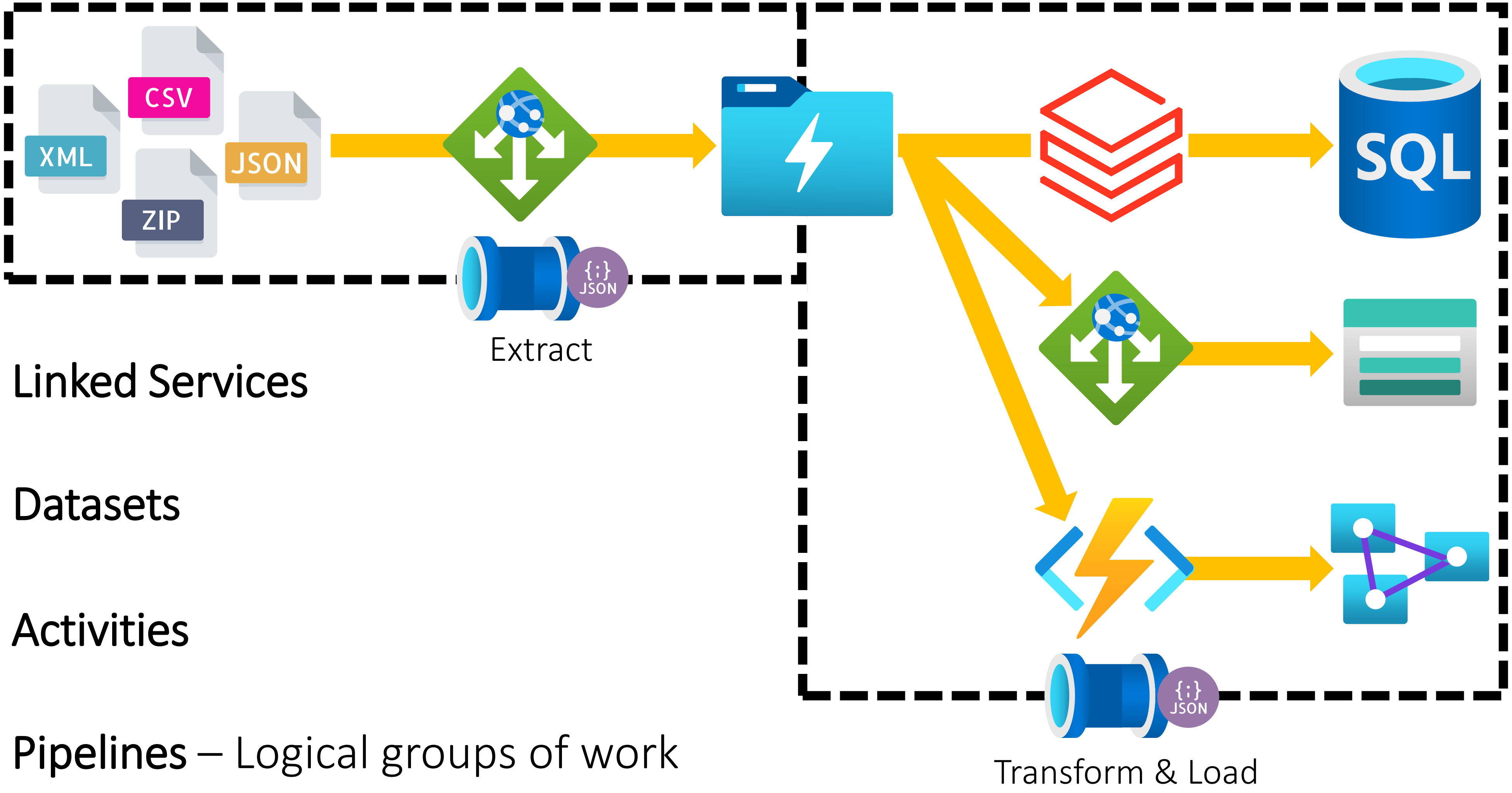
Data Factory Components



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

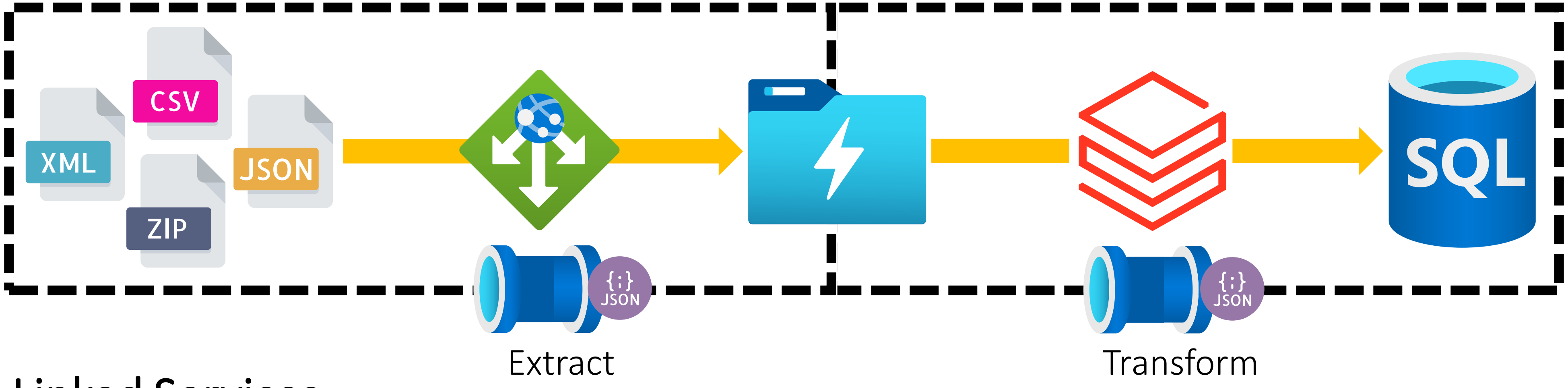


Data Factory Components



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**

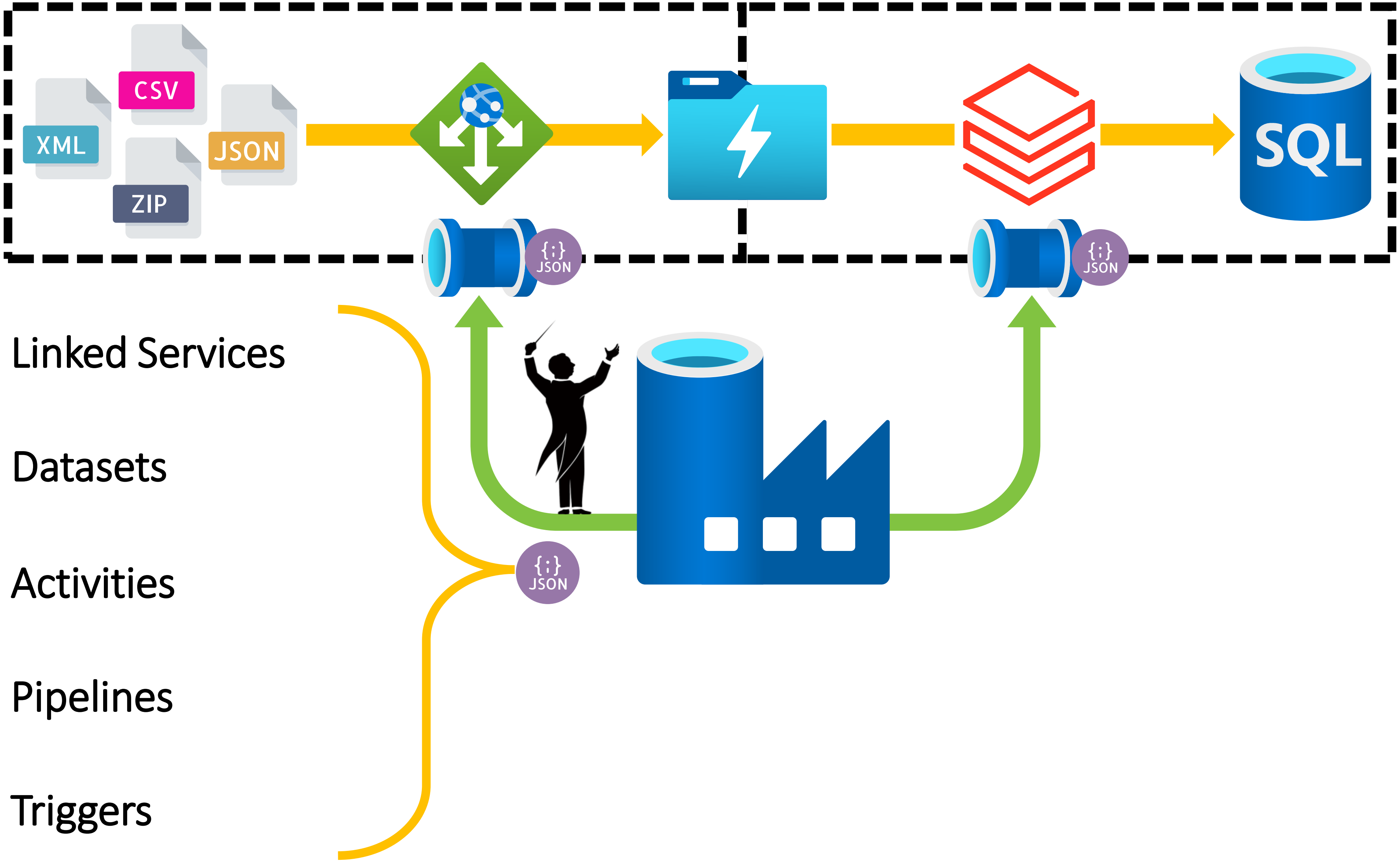
3 **Activities**

4 **Pipelines**

5 **Triggers** – Telling our when pipelines to run.

- Manually
- Programmatically
- Schedule
- Tumbling Windows
- File Event

Data Factory Components

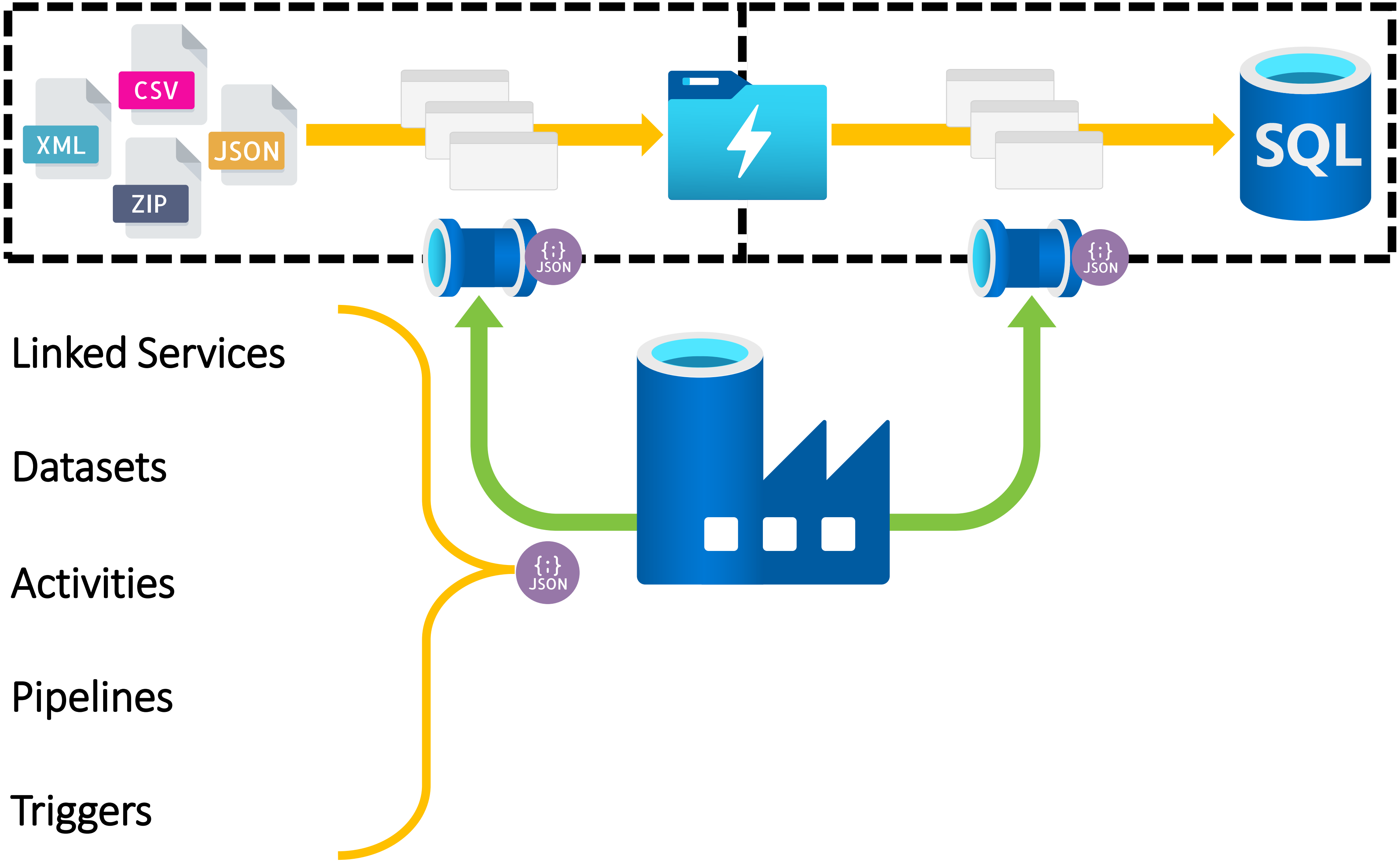


1. Linked Services
2. Datasets
3. Activities
4. Pipelines
5. Triggers

Common Activities

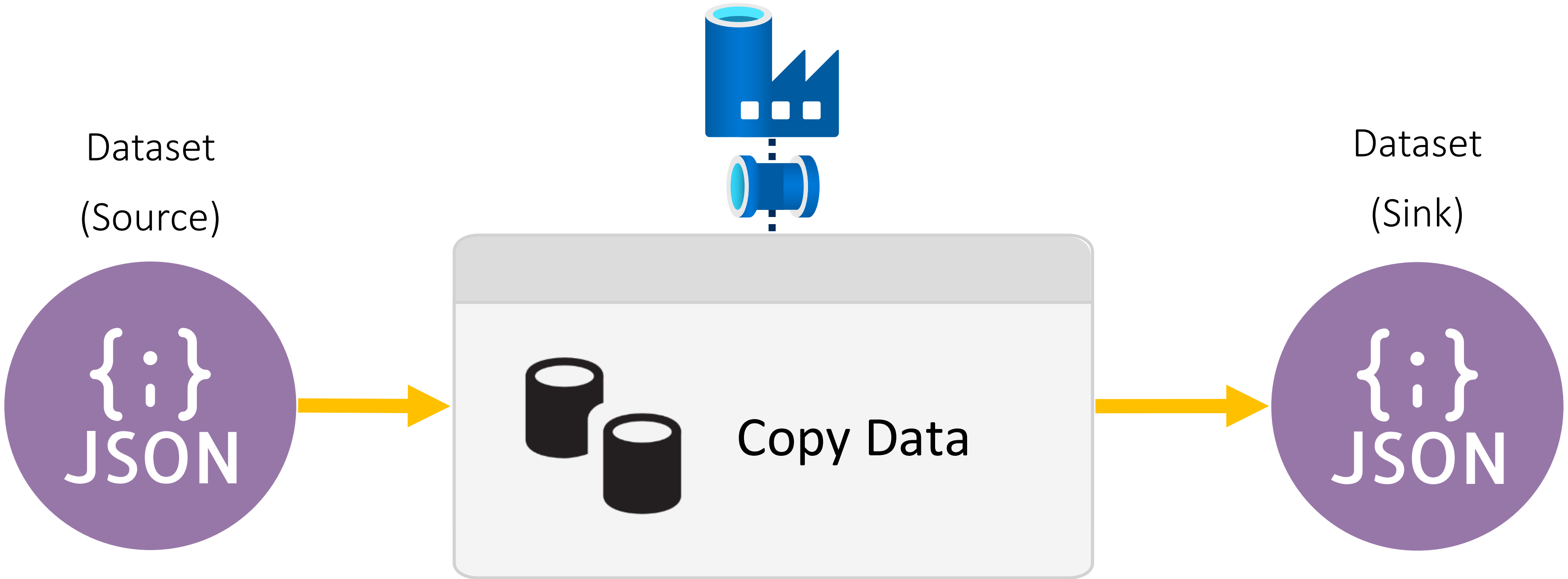
```
SELECT TOP 5
    [ActivityName],
    [Inputs],
    [Outputs],
    [Details]
FROM
    [metadata].[AdfActivities]
WHERE
    [Notes] = 'Pauls Favourites';
```

Data Factory Common Activities



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

Copy



Dataset

(Source)

Copy Data

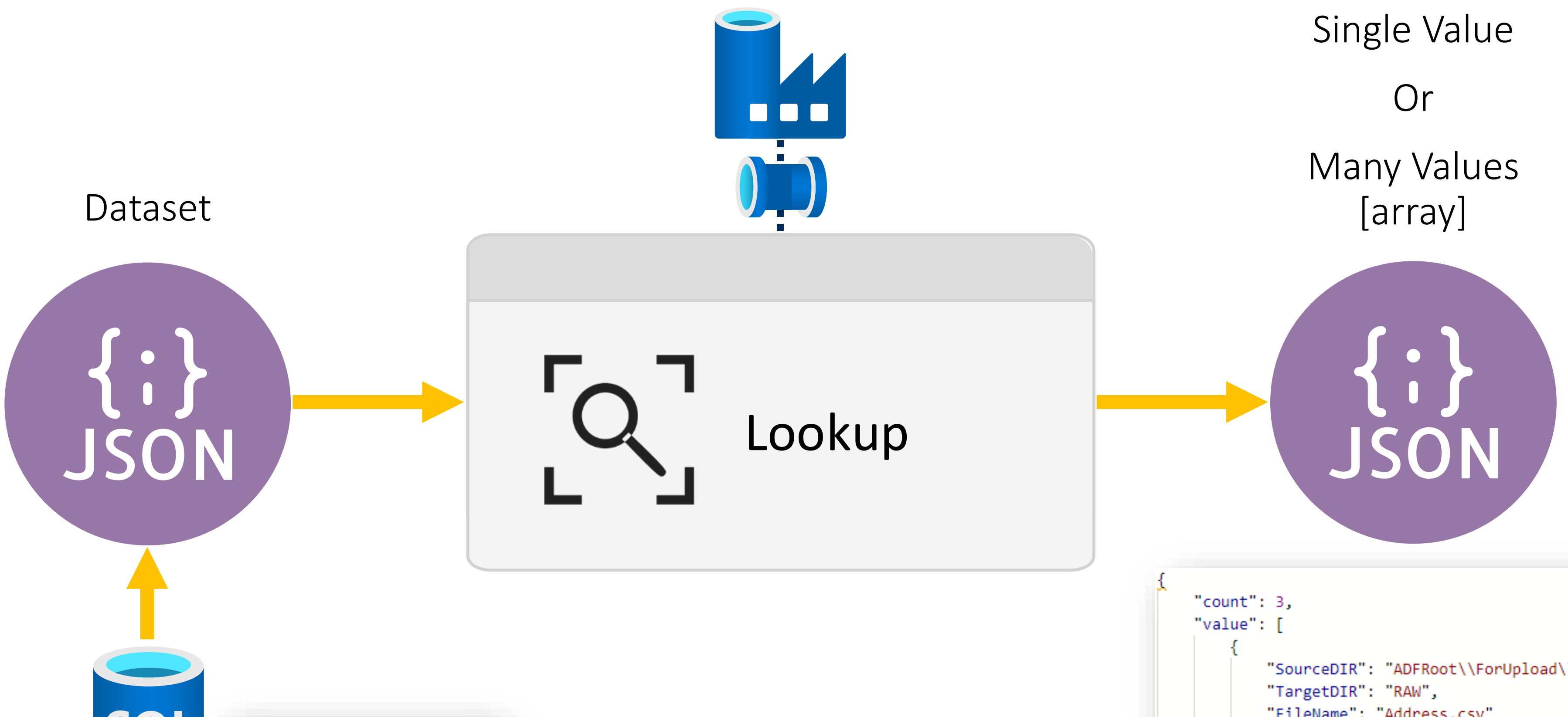
Dataset

(Sink)

- 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 to support other control flow activities

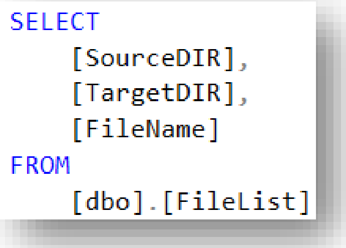


Dataset

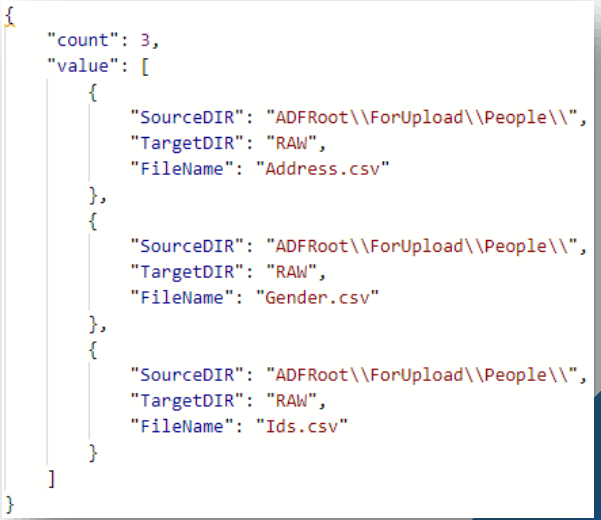
Single Value

Or

Many Values
[array]



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

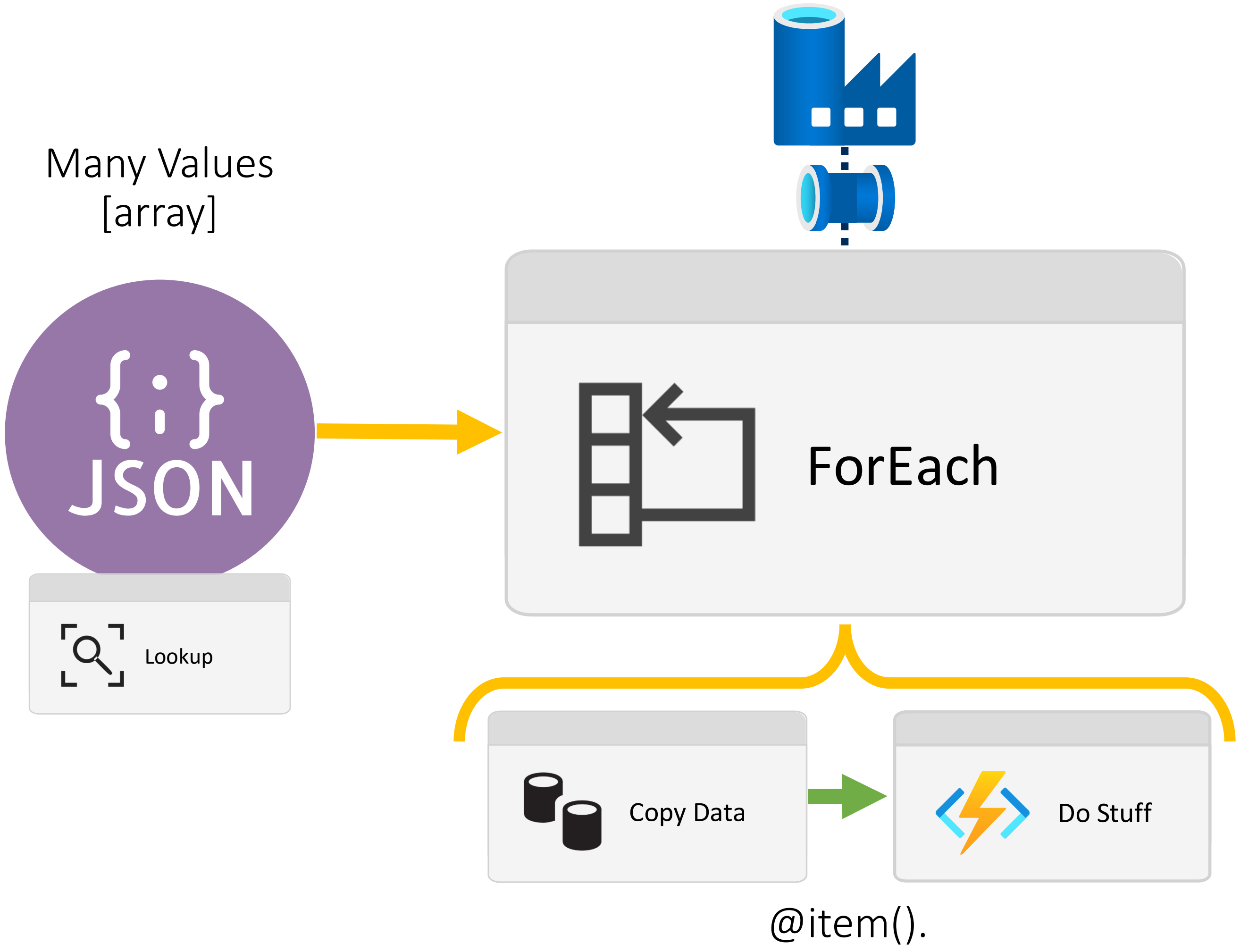


```
{
    "count": 3,
    "value": [
        {
            "SourceDIR": "ADFRoot\\ForUpload\\People\\",
            "TargetDIR": "RAW",
            "FileName": "Address.csv"
        },
        {
            "SourceDIR": "ADFRoot\\ForUpload\\People\\",
            "TargetDIR": "RAW",
            "FileName":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

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

ForEach

Scaling Out Control Flow Activities



Many Values
[array]

JSON

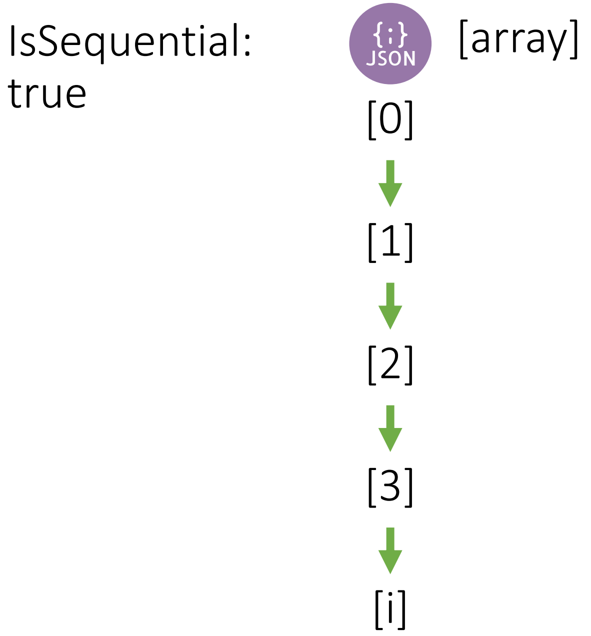
Lookup

ForEach

Copy Data

Do Stuff

@item().



IsSequential:
true

JSON [array]

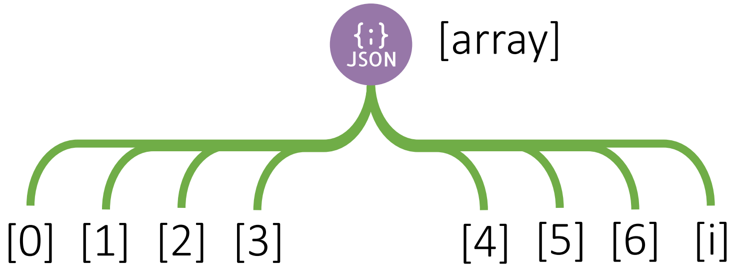
[0]

[1]

[2]

[3]

[i]



JSON [array]

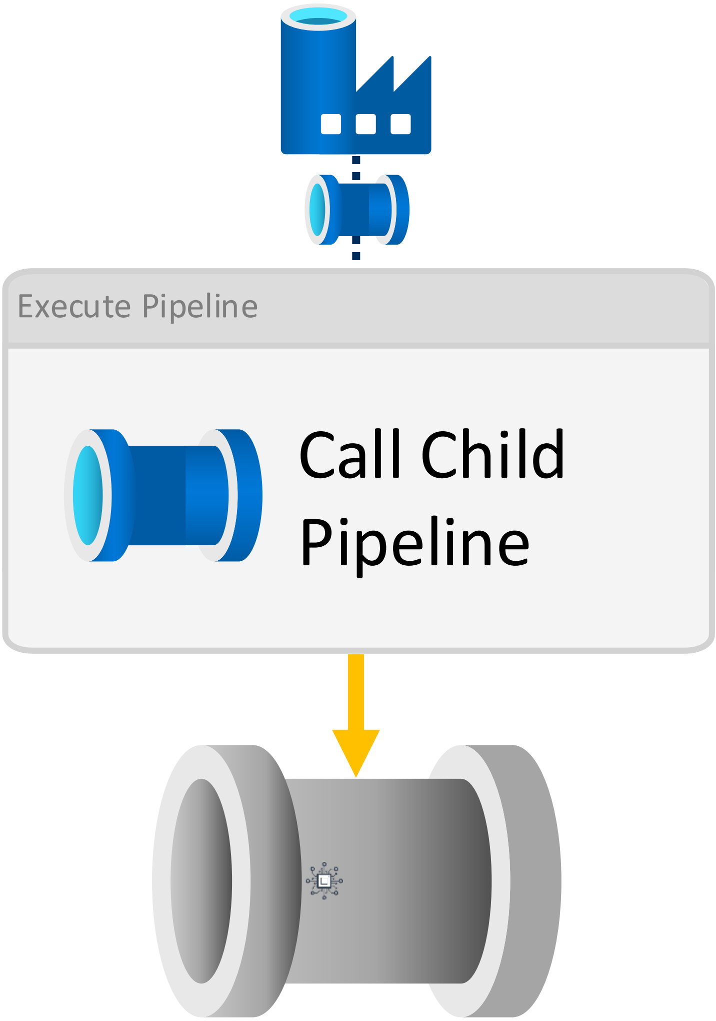
[0] [1] [2] [3] [4] [5] [6] [i]

Batch Count Default: 20

Batch Count Max: 50

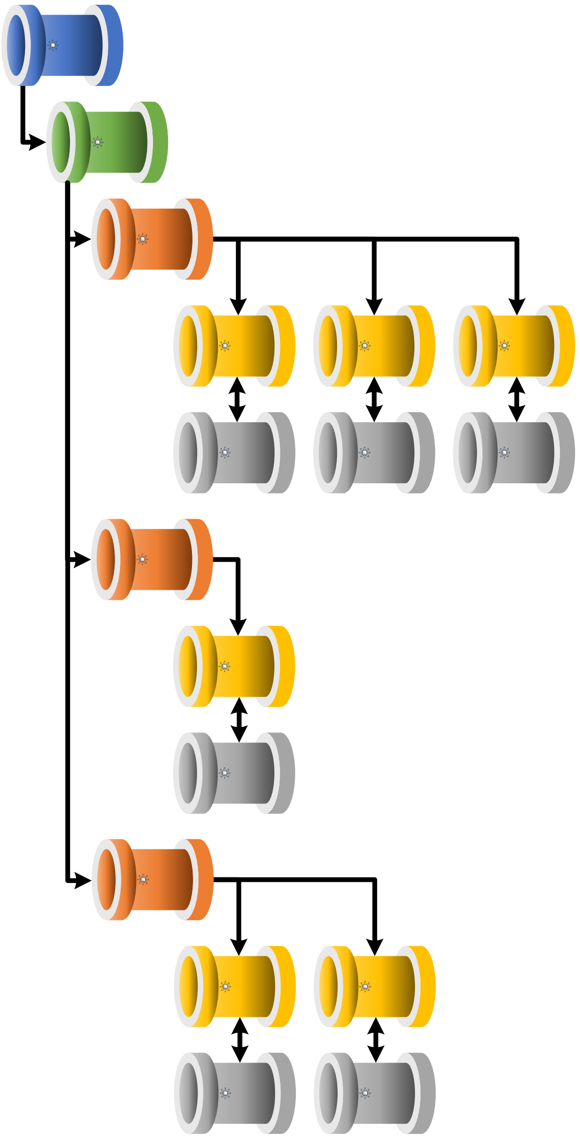
https://docs.microsoft.com/en-us/azure/data-factory/control-flow-for-each-activity

Execute Pipeline



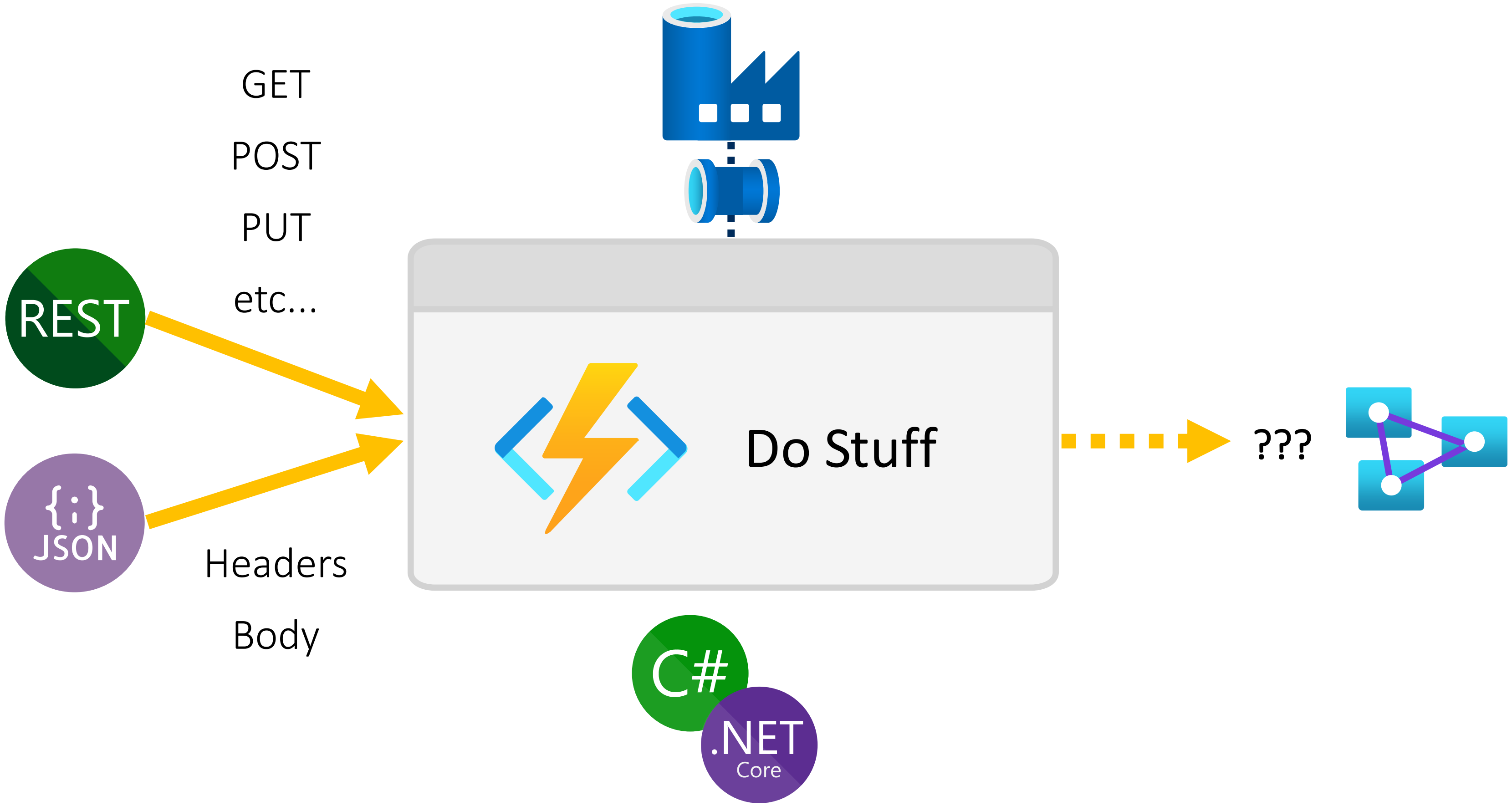
Pipeline Hierarchies Generation Control

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



Azure Function

Extend Data Factory with Rest Calls



GET

POST

PUT

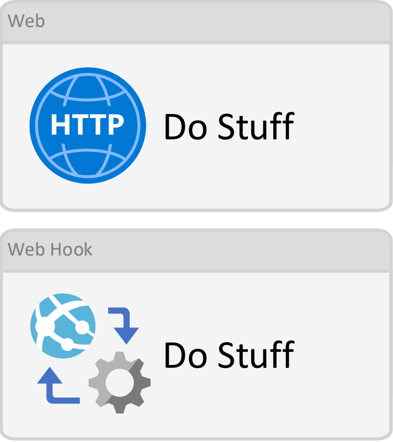
etc...

Headers

Body

Do Stuff

???



Web

Do Stuff

Web Hook

Do Stuff

Custom

Extend Data Factory with Custom Code

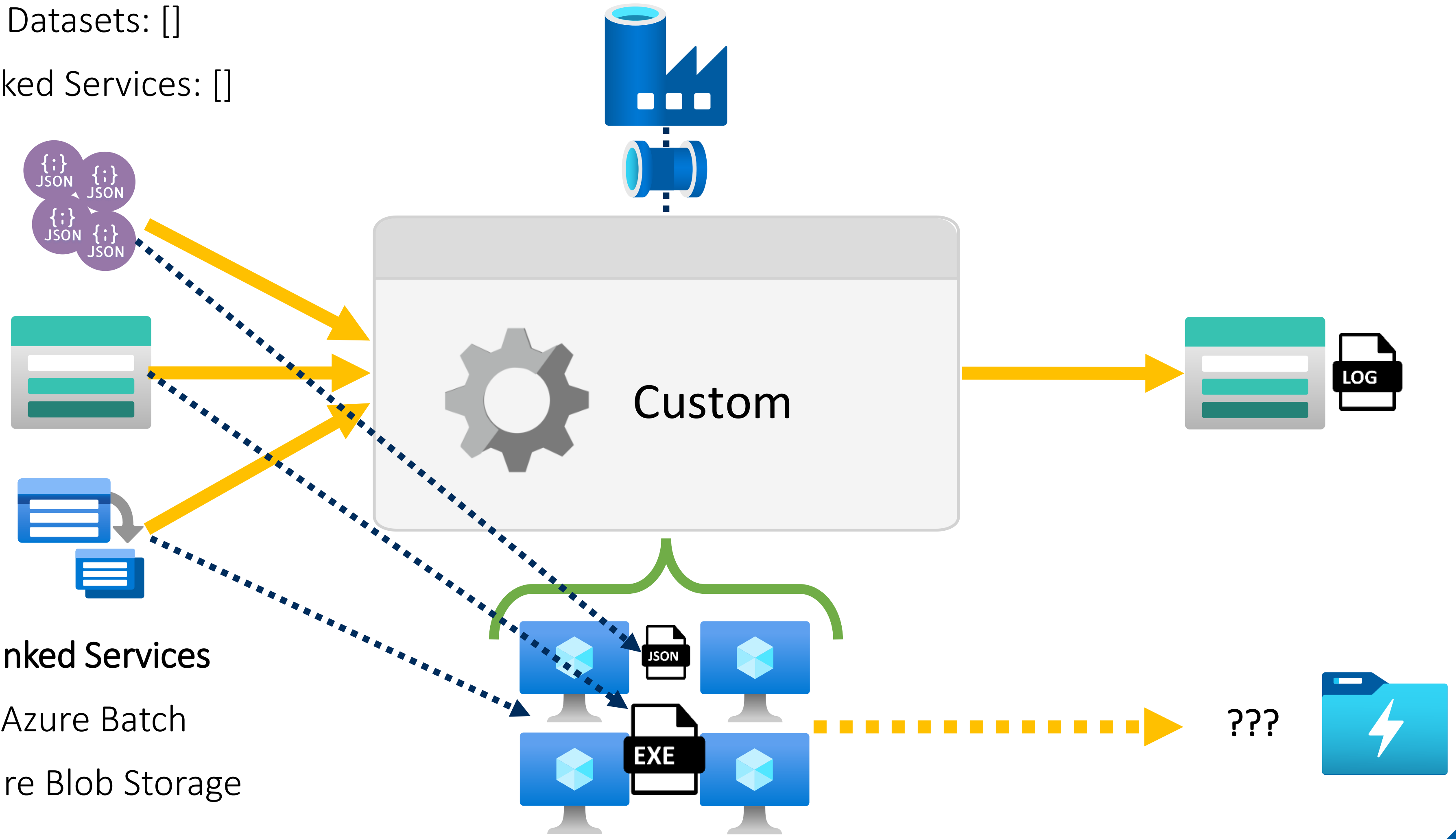
Creating a Custom Activity

https://mrpaulandrew.com/2018/11/12/creating-an-azure-data-factory-v2-custom-activity/

References Objects

Datasets: []

Linked Services: []

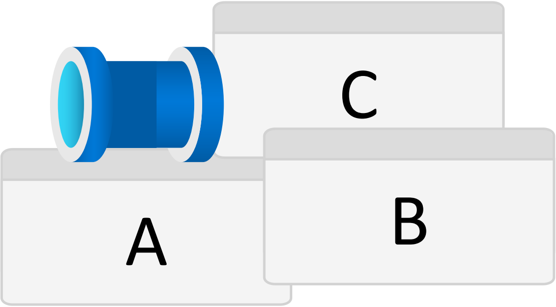


Linked Services

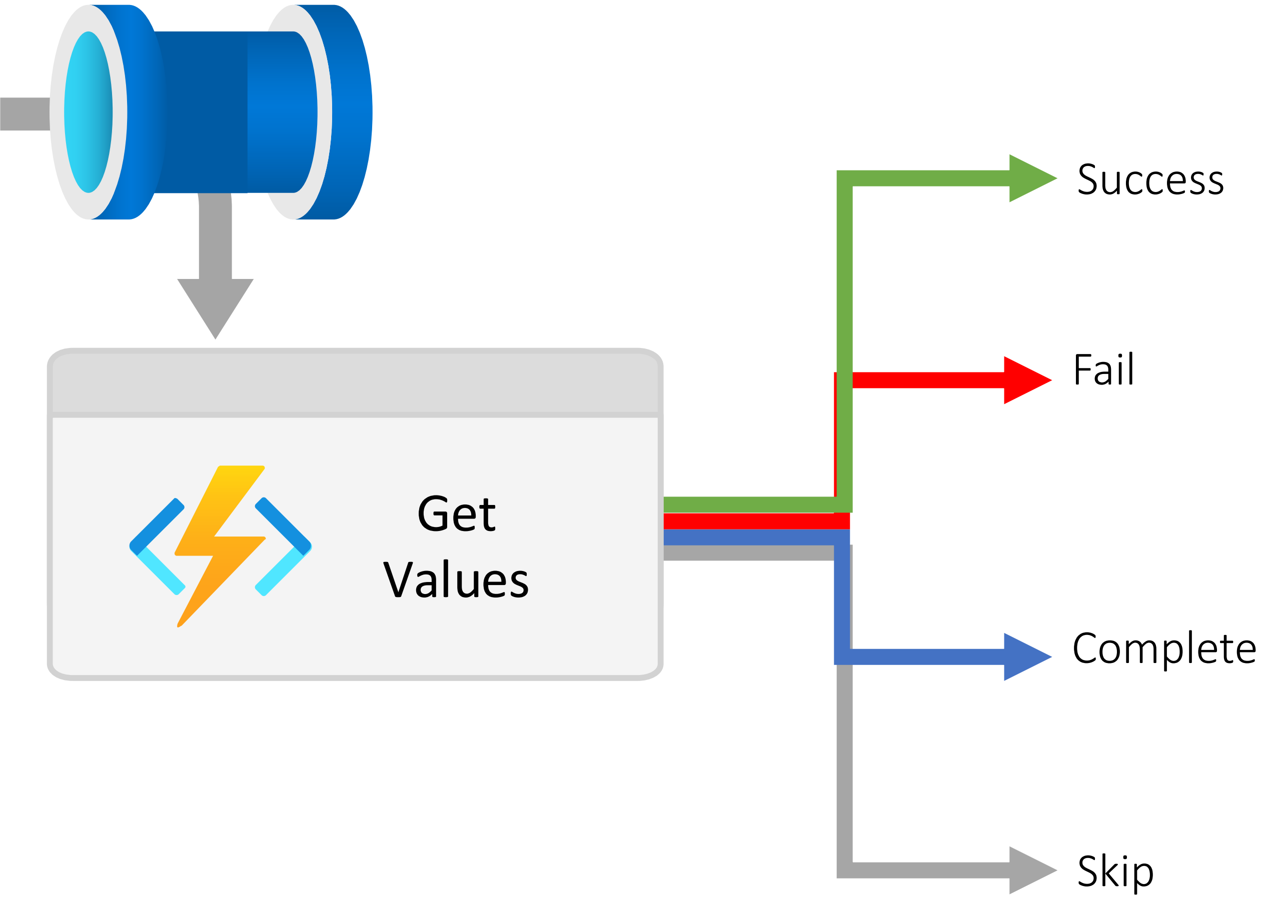
Azure Batch

Azure Blob Storage

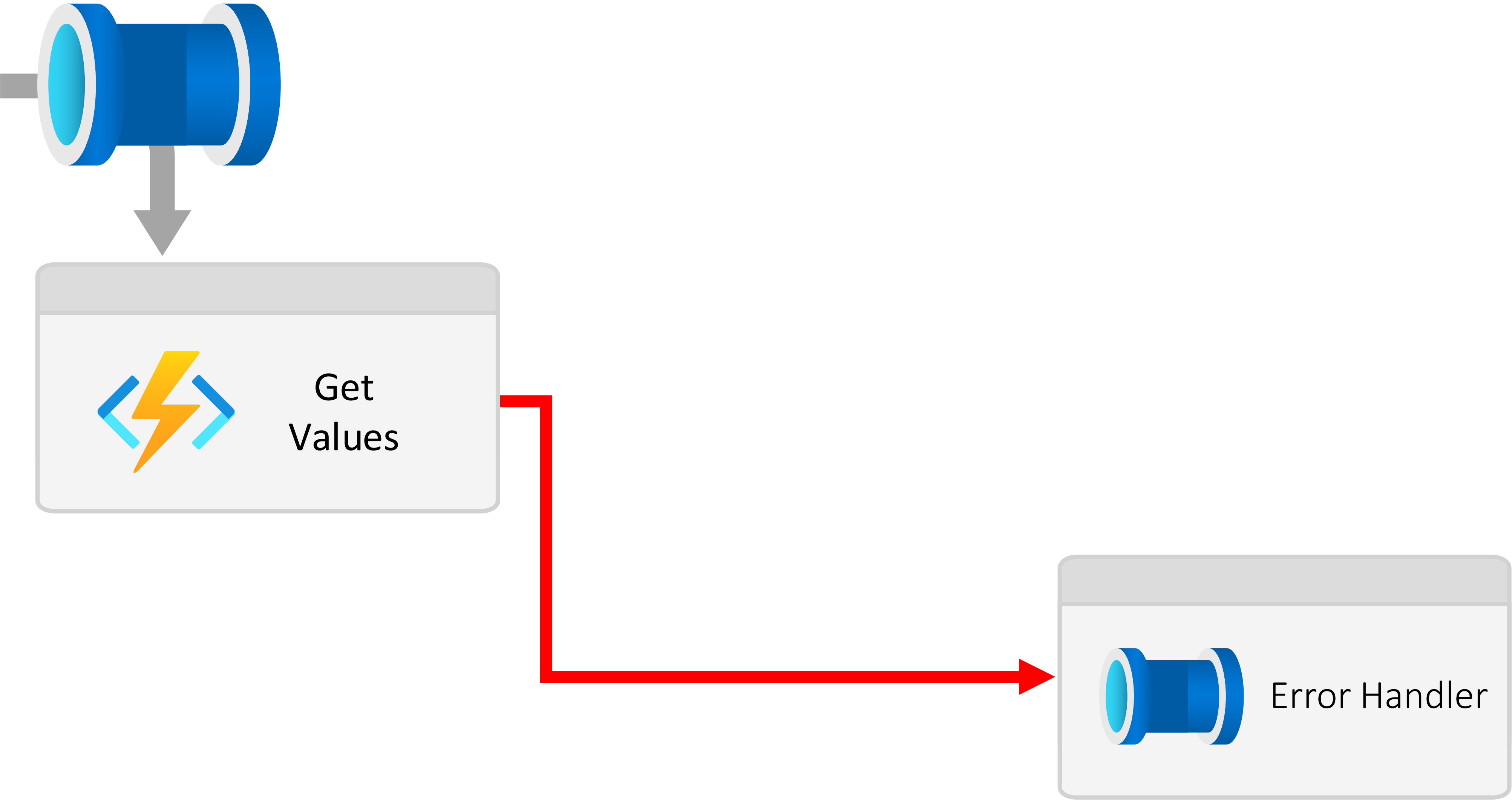
Execution Dependencies



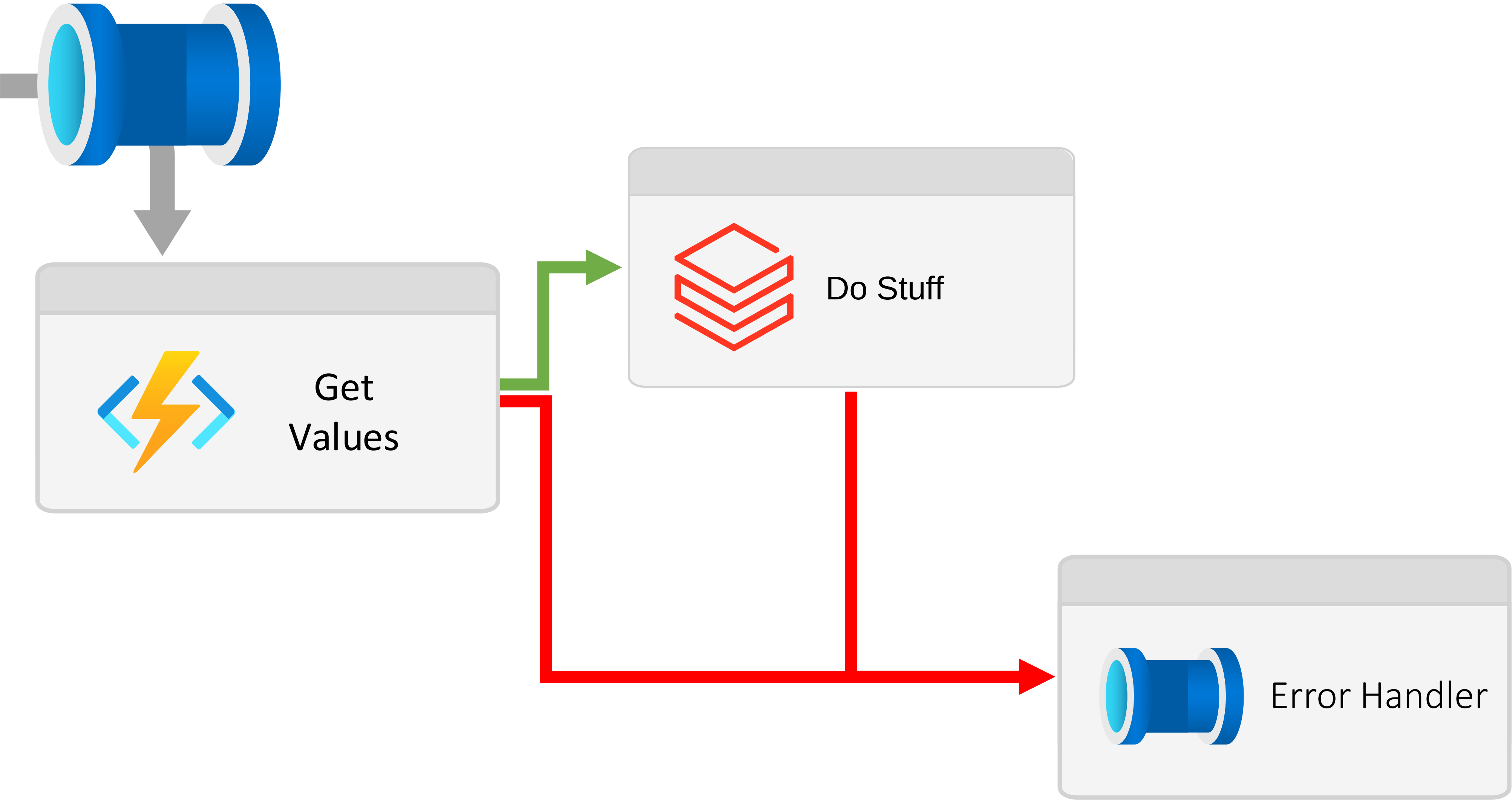
Execution Dependency Options



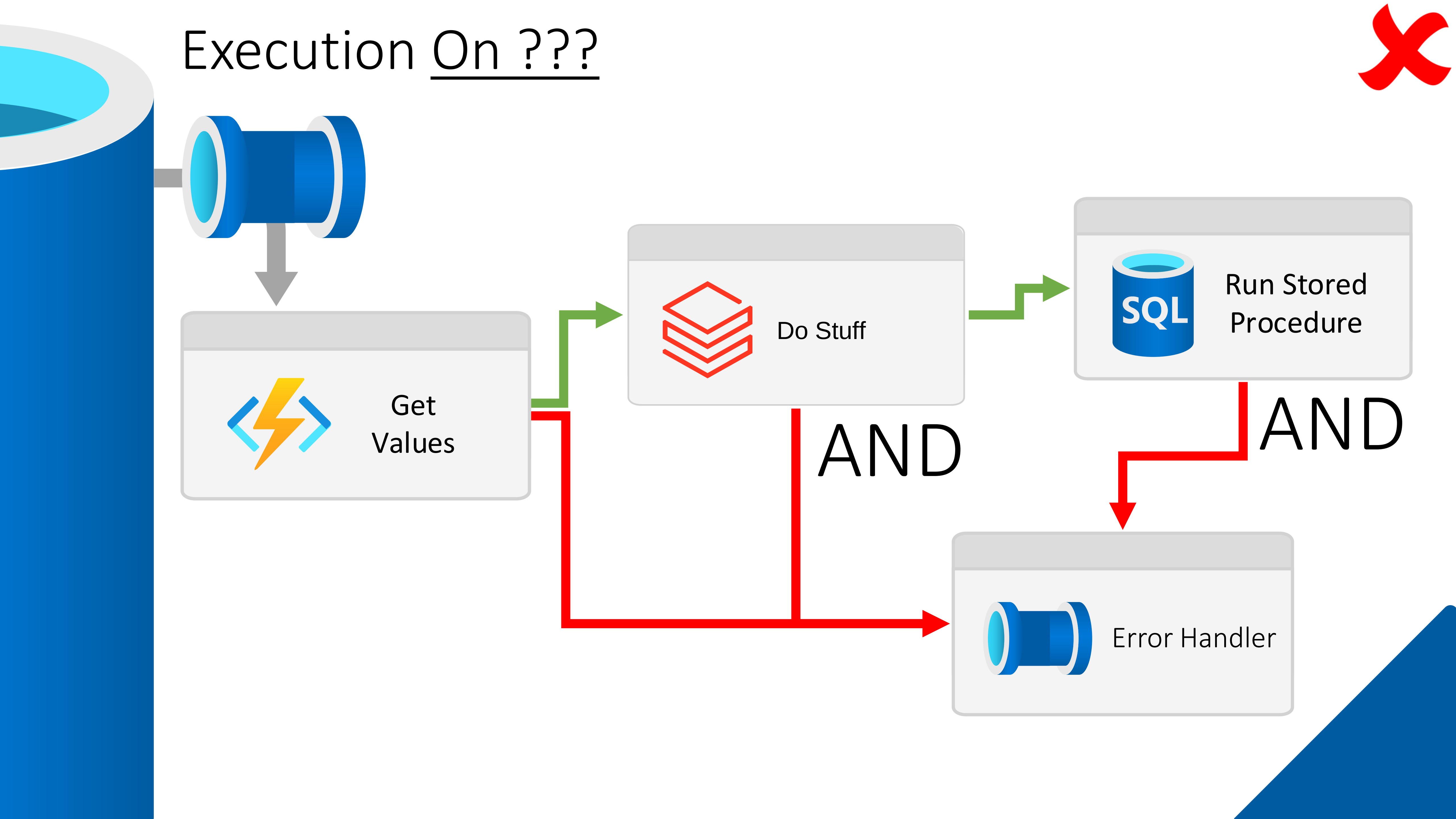
Execution On Failure



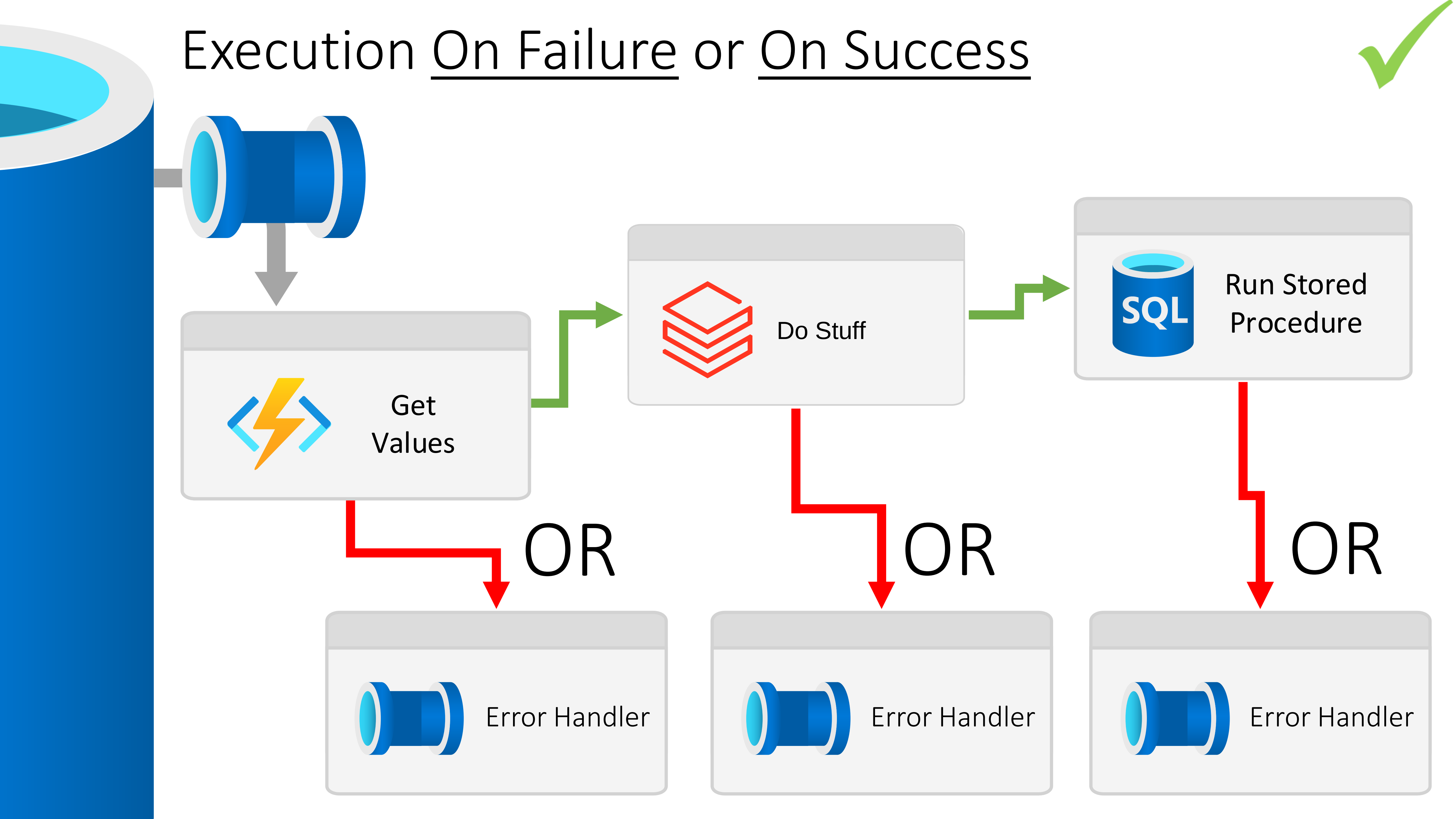
Execution On Failure or On Success



Execution On ???

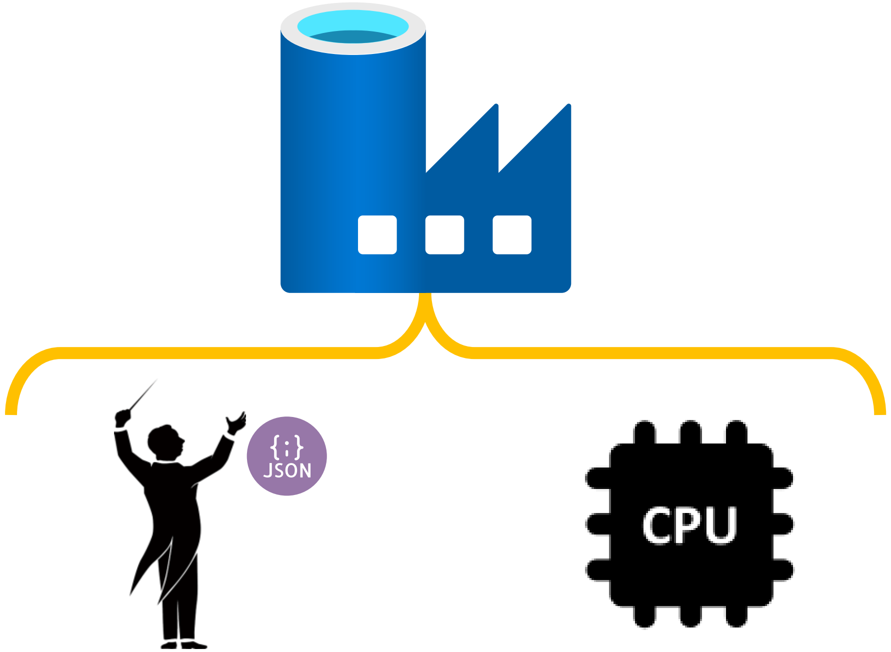


Execution On Failure or On Success



Integration Runtimes

What is an Integration Runtime?

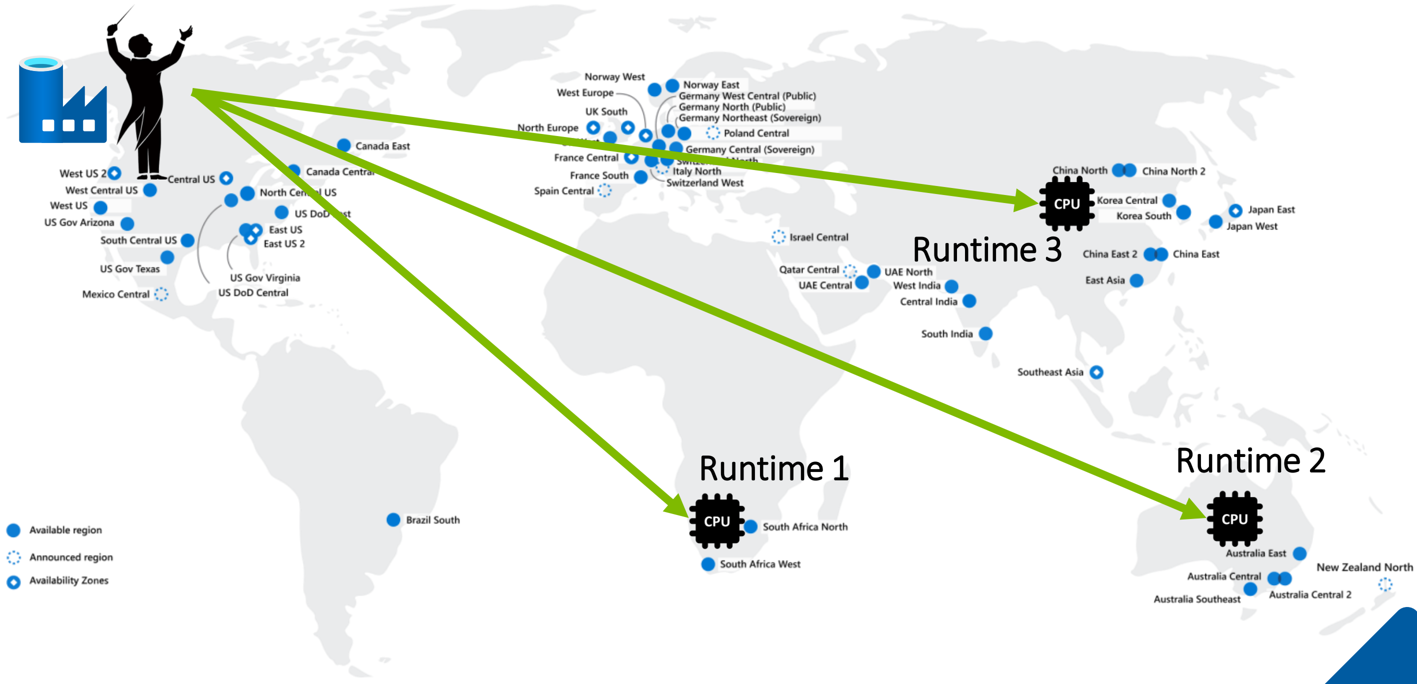


Orchestrator

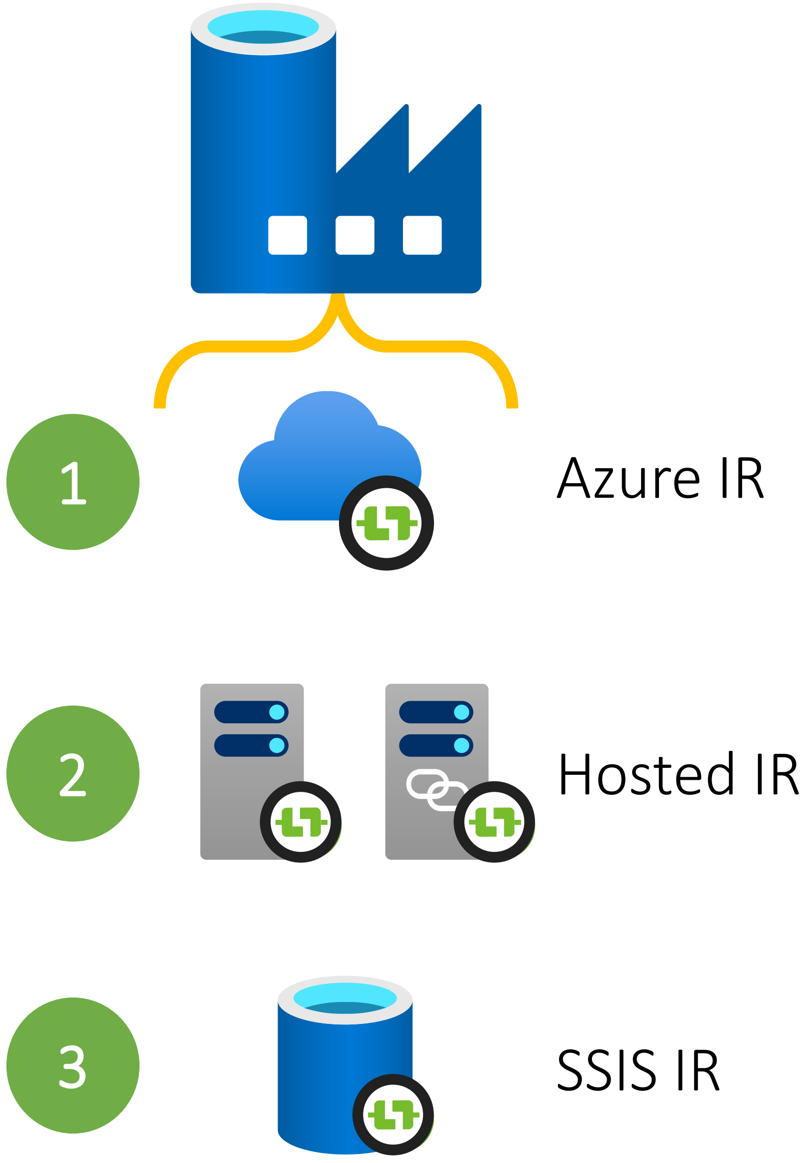
Fixed Region

Runtime

Flexible Location



What can an Integration Runtime do?

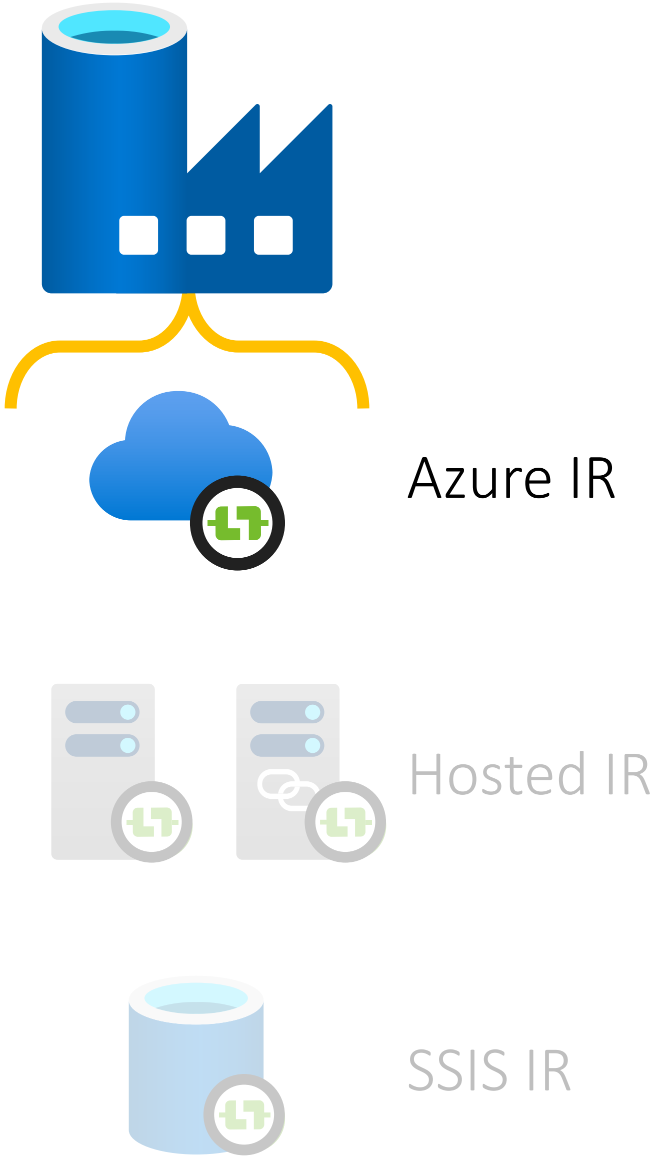


1 Azure IR

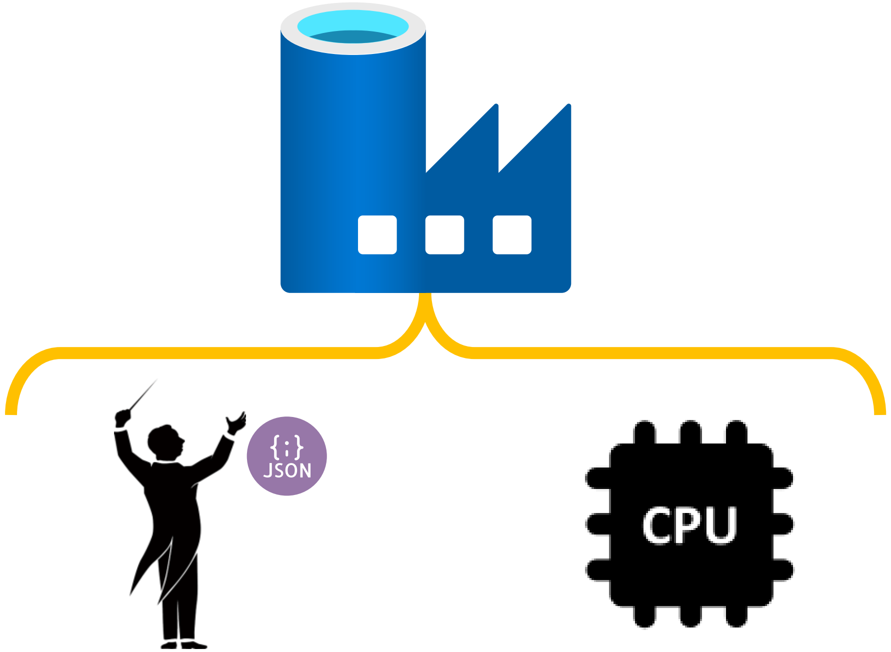
2 Hosted IR

3 SSIS IR

Azure Integration Runtime



Azure Integration Runtime

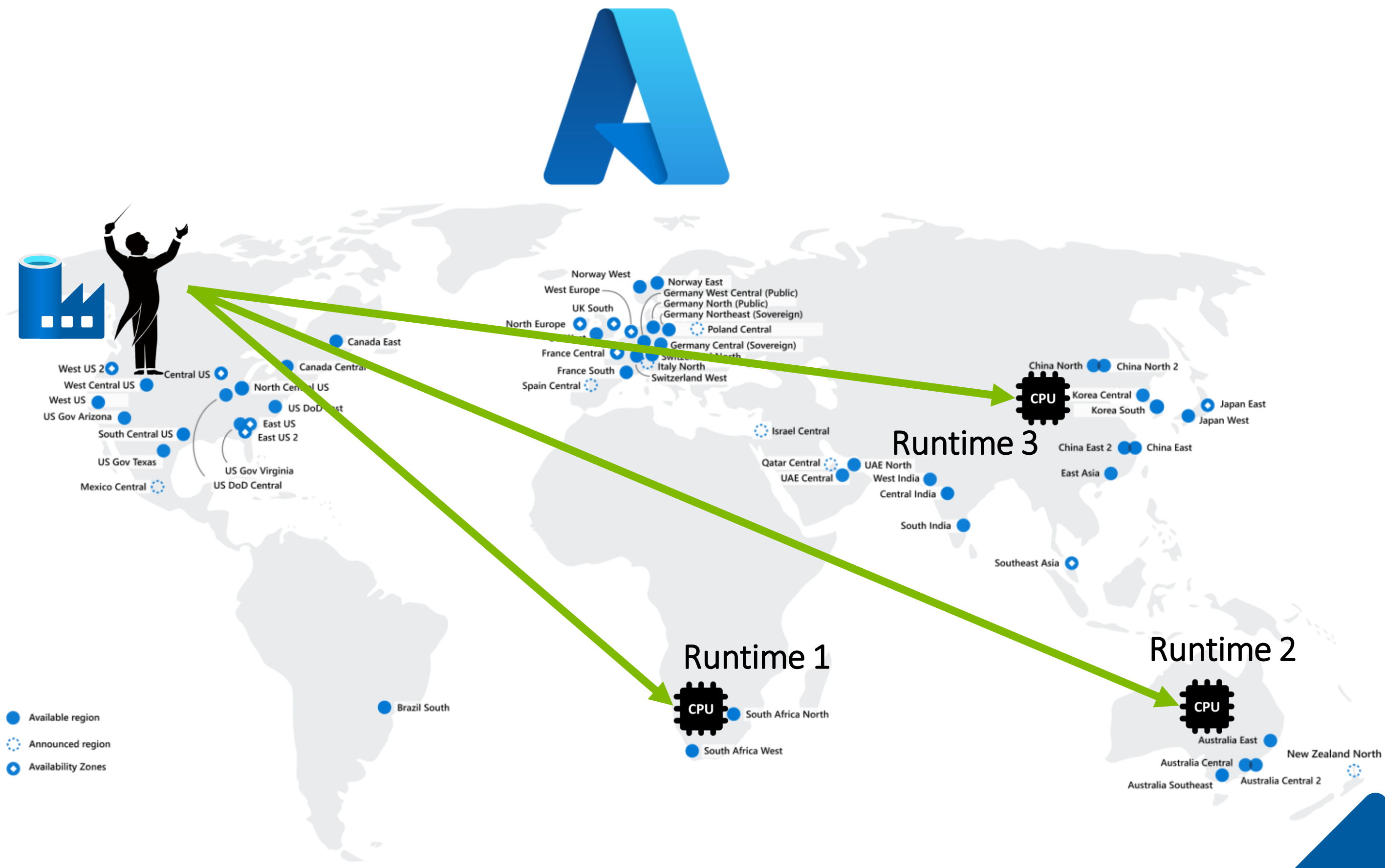


Orchestrator

Fixed Region

Runtime

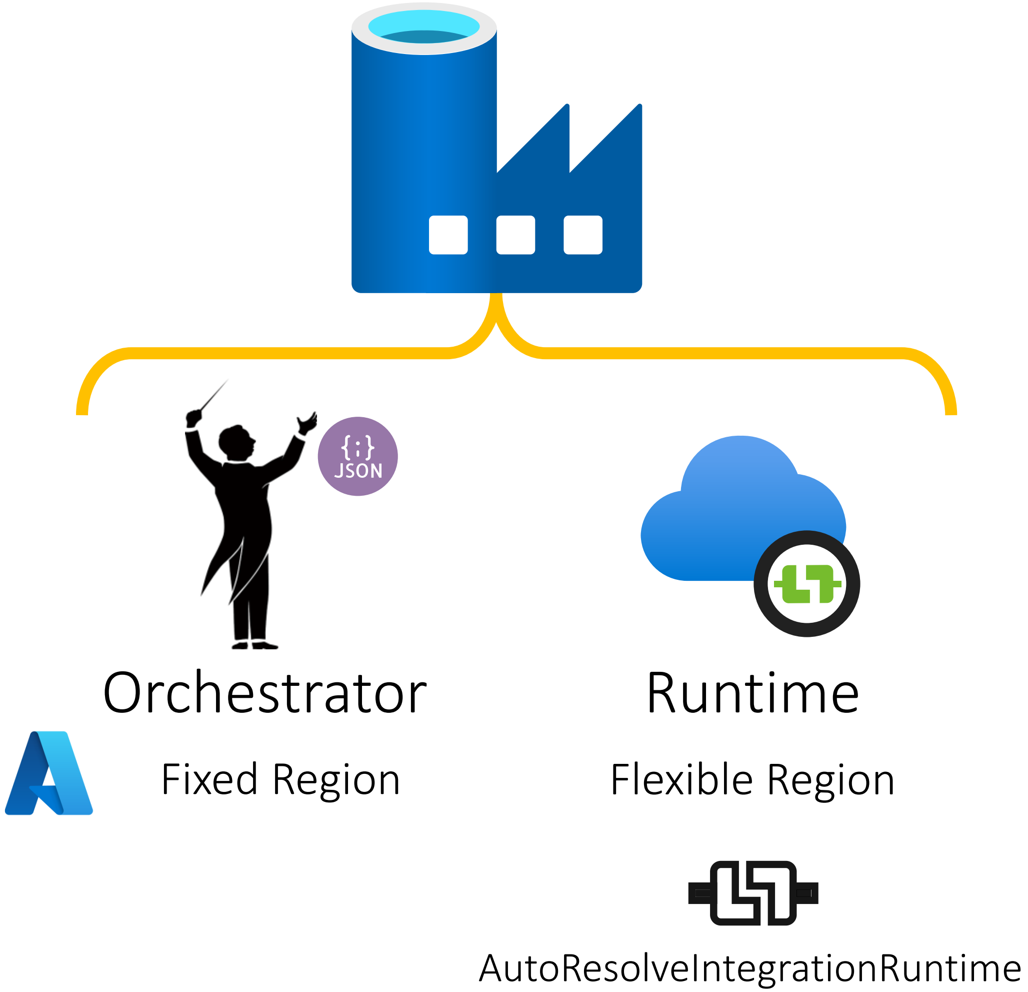
Flexible Location



Azure Integration Runtime

Internal vs External Activities

https://mrpaulandrew.com/2020/12/22/pipelines-understanding-internal-vs-external-activities/



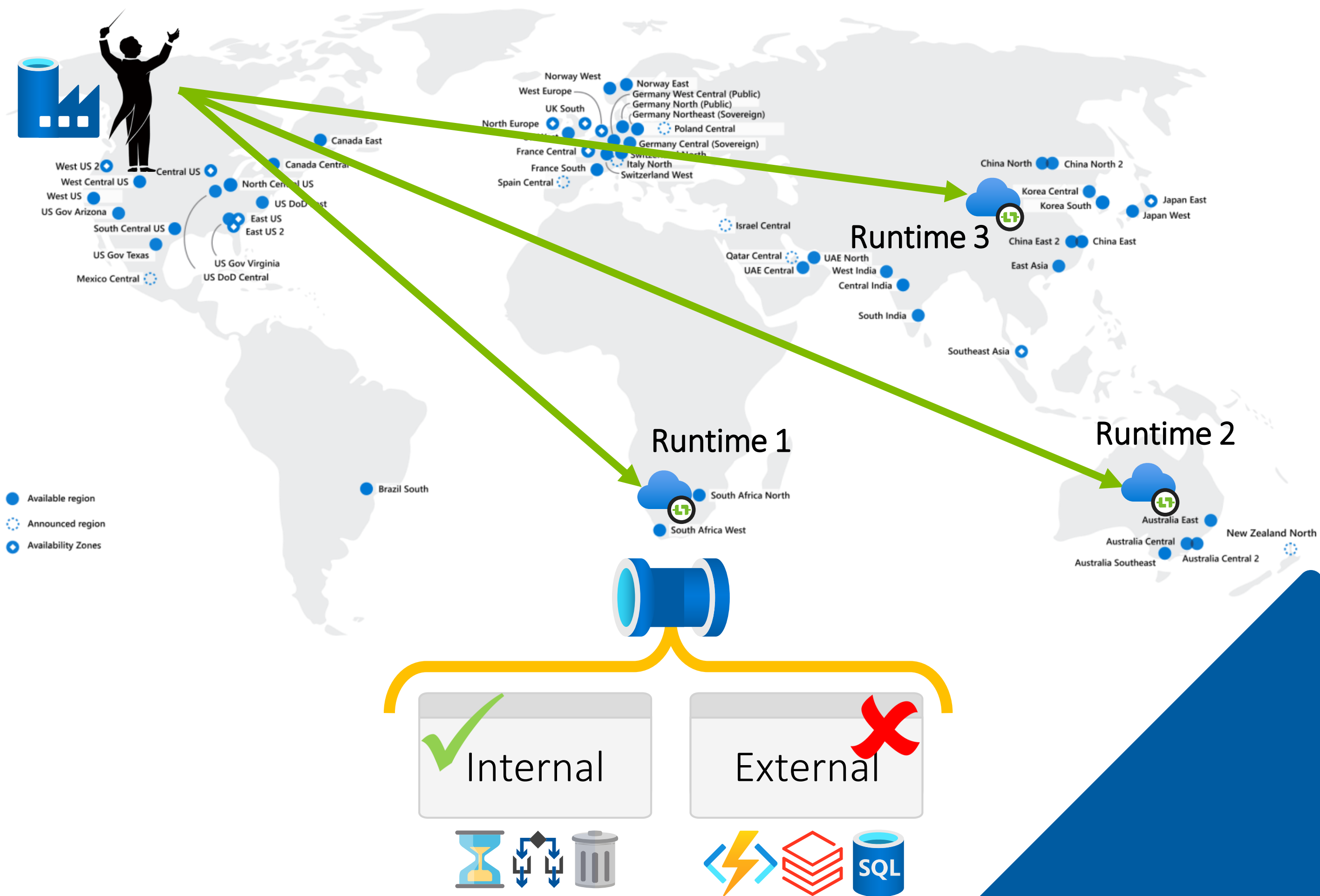
Orchestrator

Fixed Region

Runtime

Flexible Region

AutoResolveIntegrationRuntime



Runtime 3

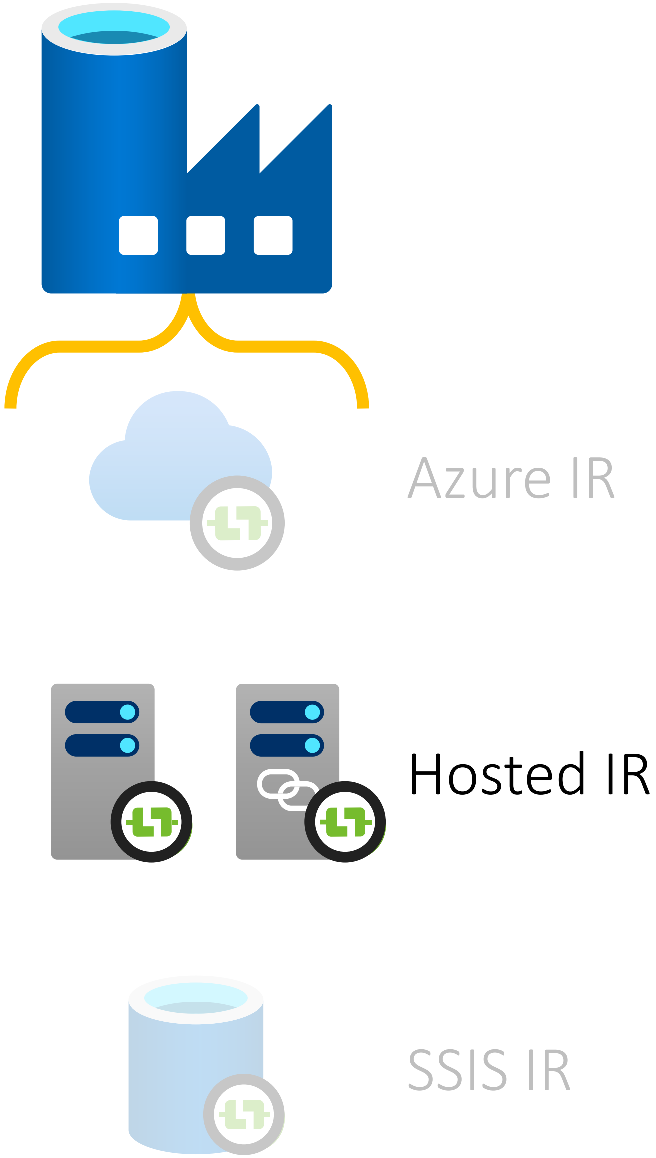
Runtime 1

Runtime 2

Internal

External

Hosted Integration Runtime

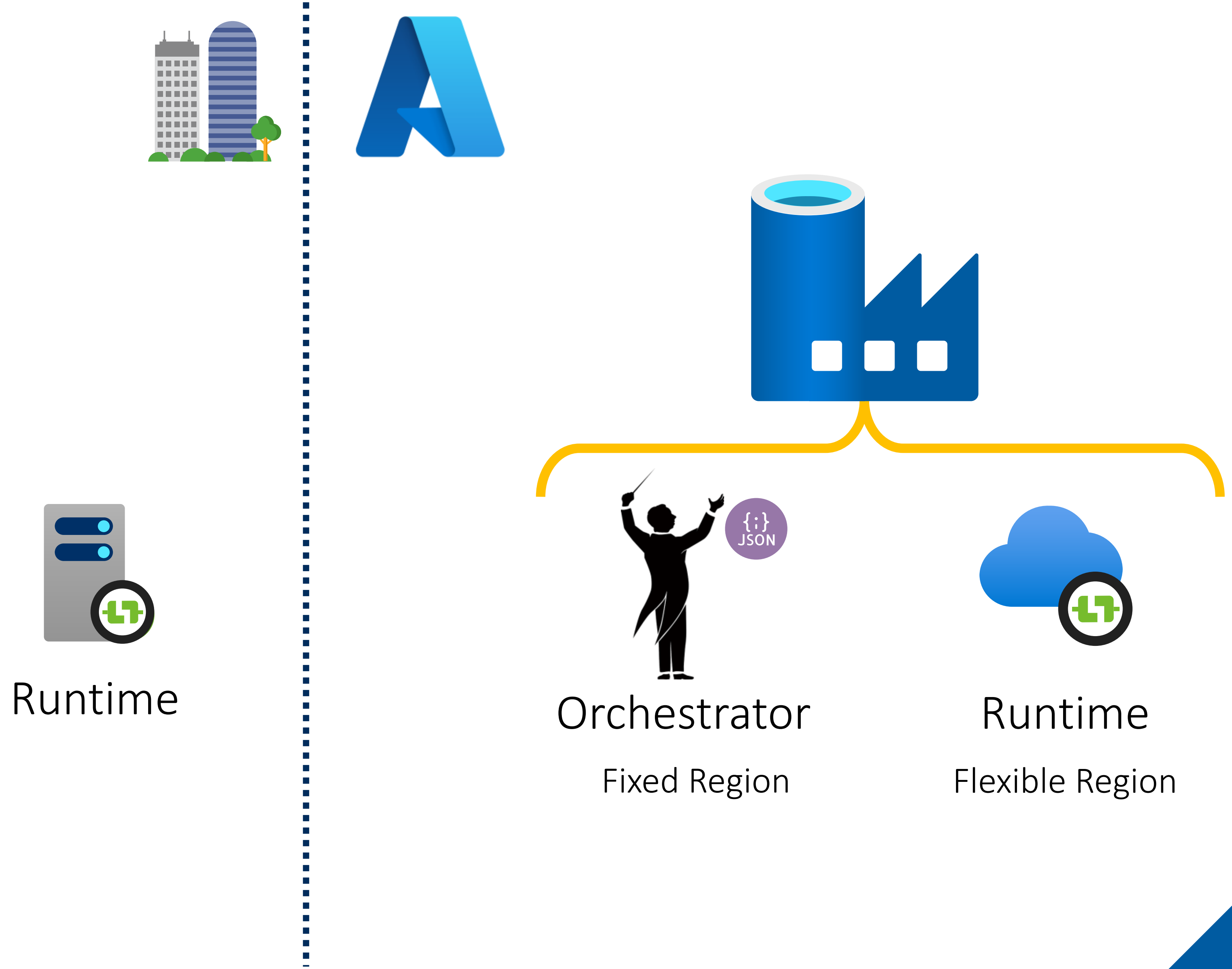


Azure IR

Hosted IR

SSIS IR

Hosted Integration Runtime



Runtime

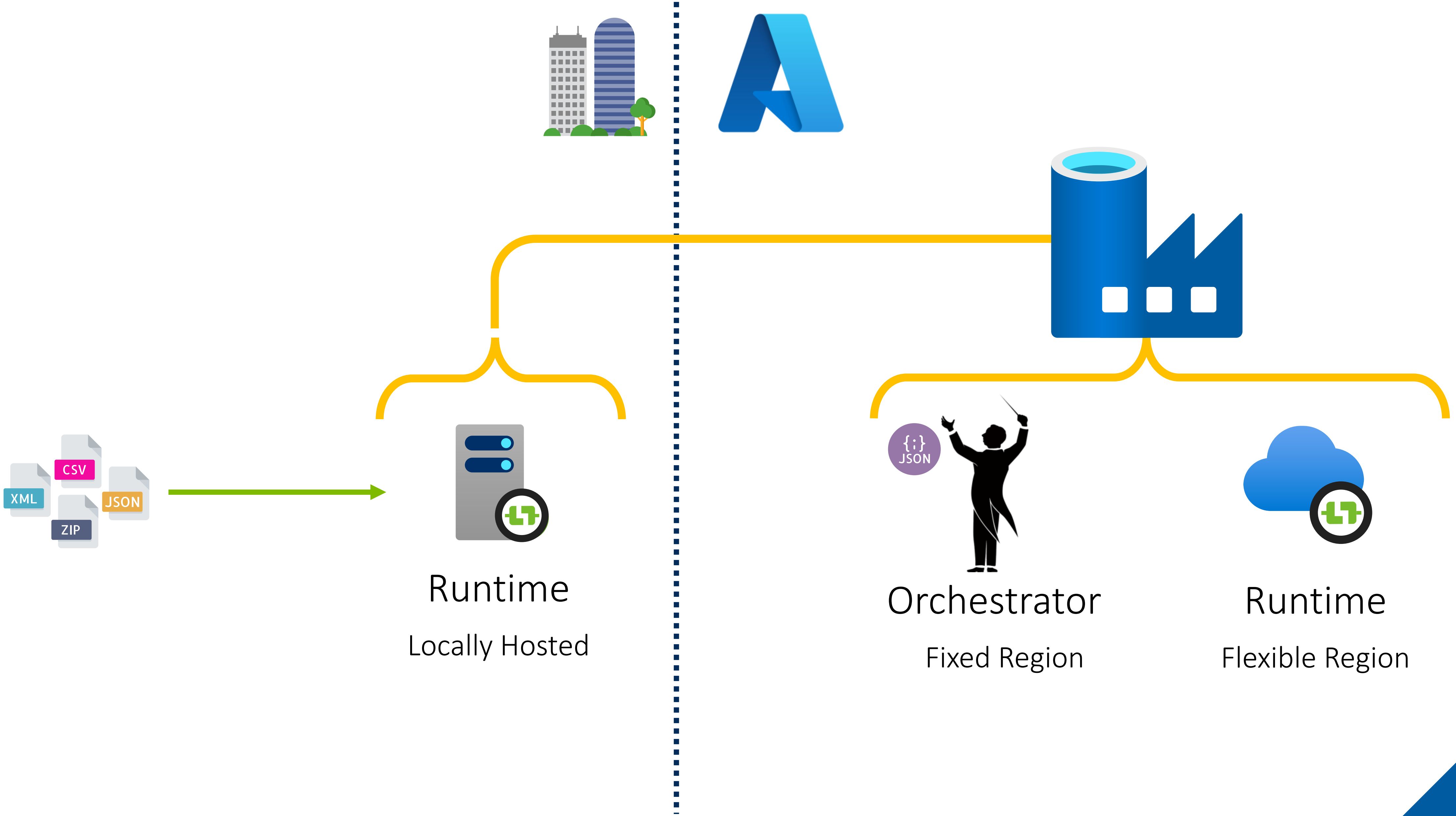
Orchestrator

Fixed Region

Runtime

Flexible Region

Hosted Integration Runtime



Runtime

Locally Hosted

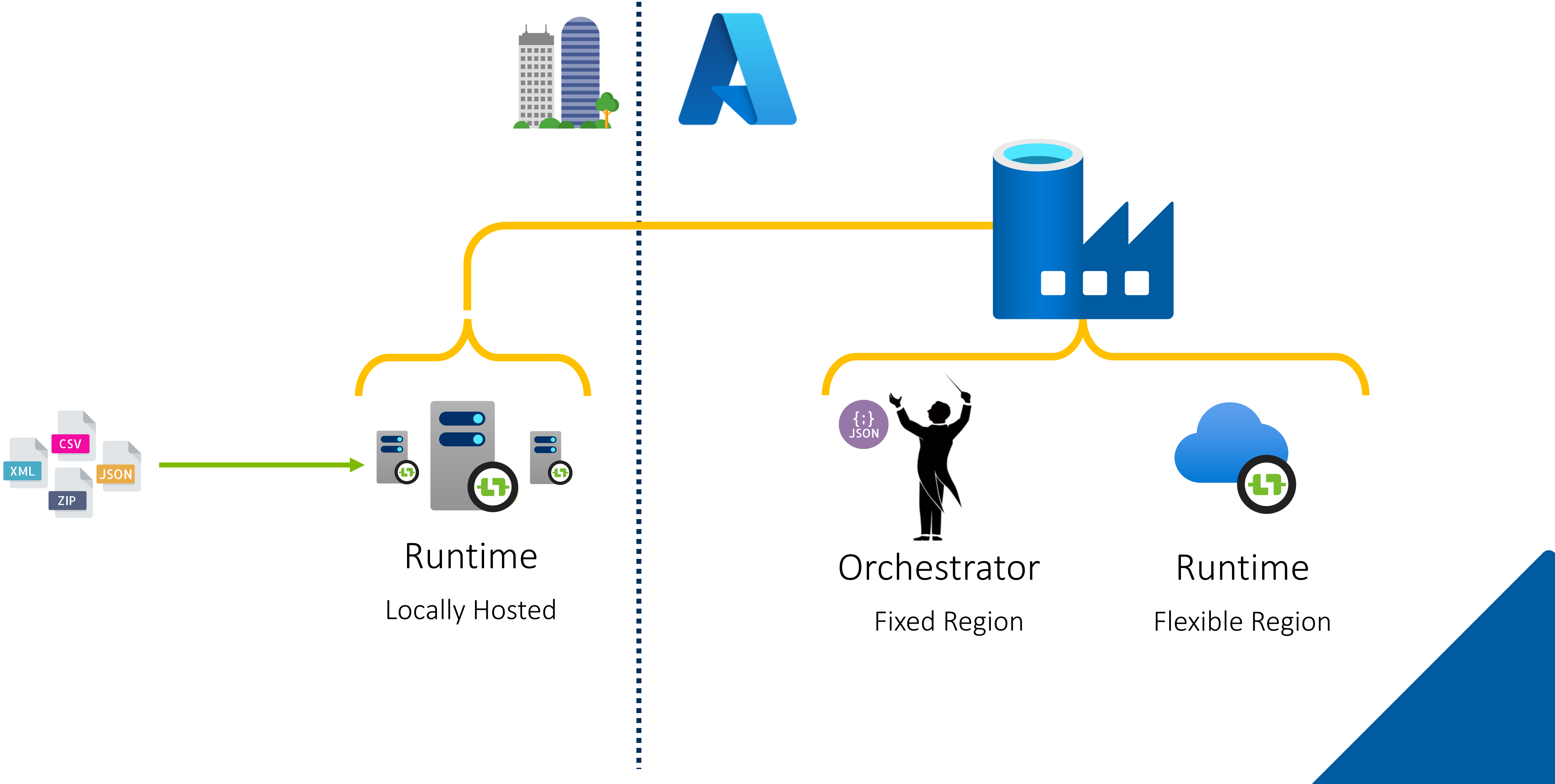
Orchestrator

Fixed Region

Runtime

Flexible Region

Hosted Integration Runtime – Secondary Nodes



Runtime

Locally Hosted

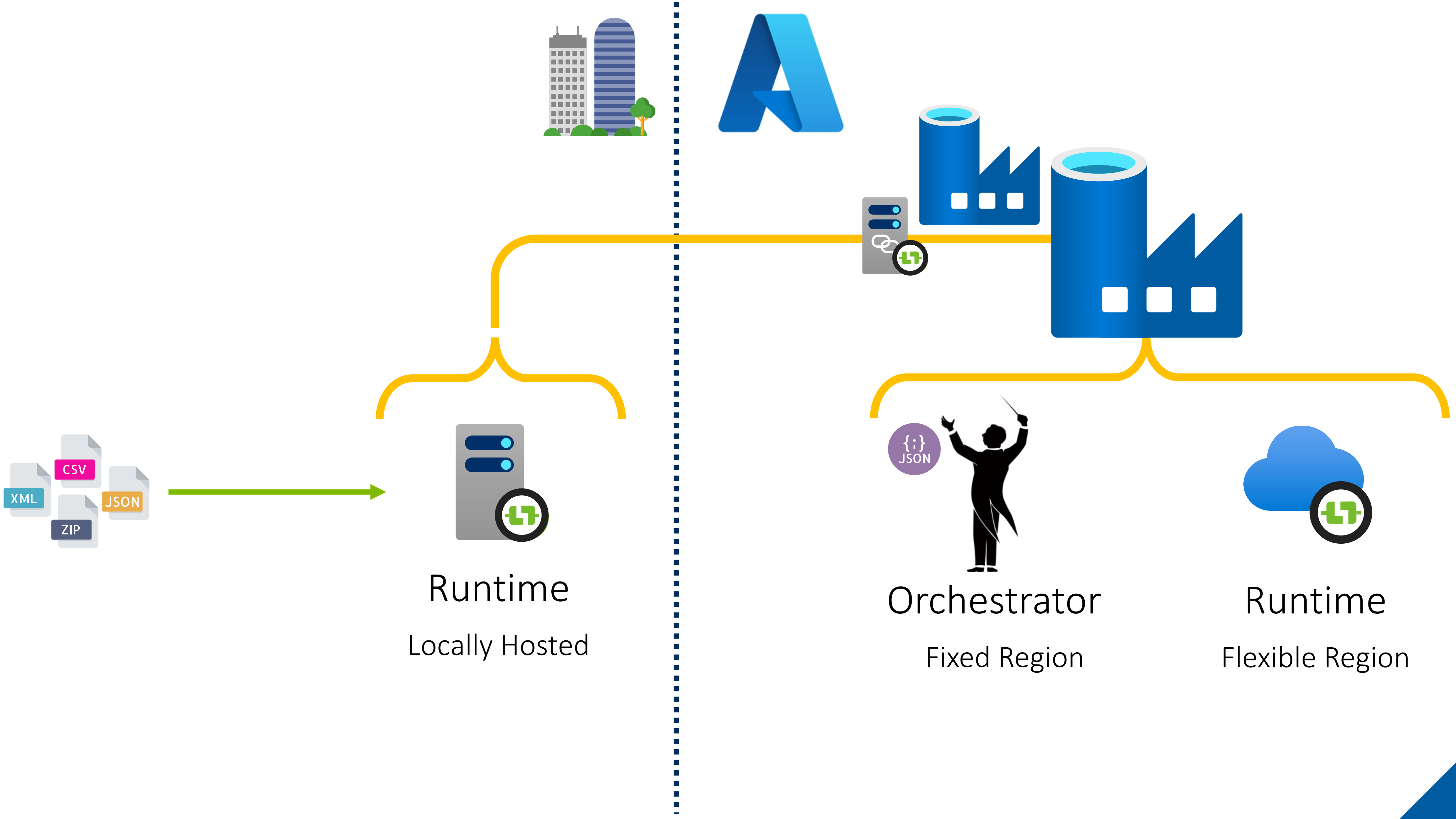
Orchestrator

Fixed Region

Runtime

Flexible Region

Hosted Integration Runtime – Linked



Runtime

Locally Hosted

Orchestrator

Fixed Region

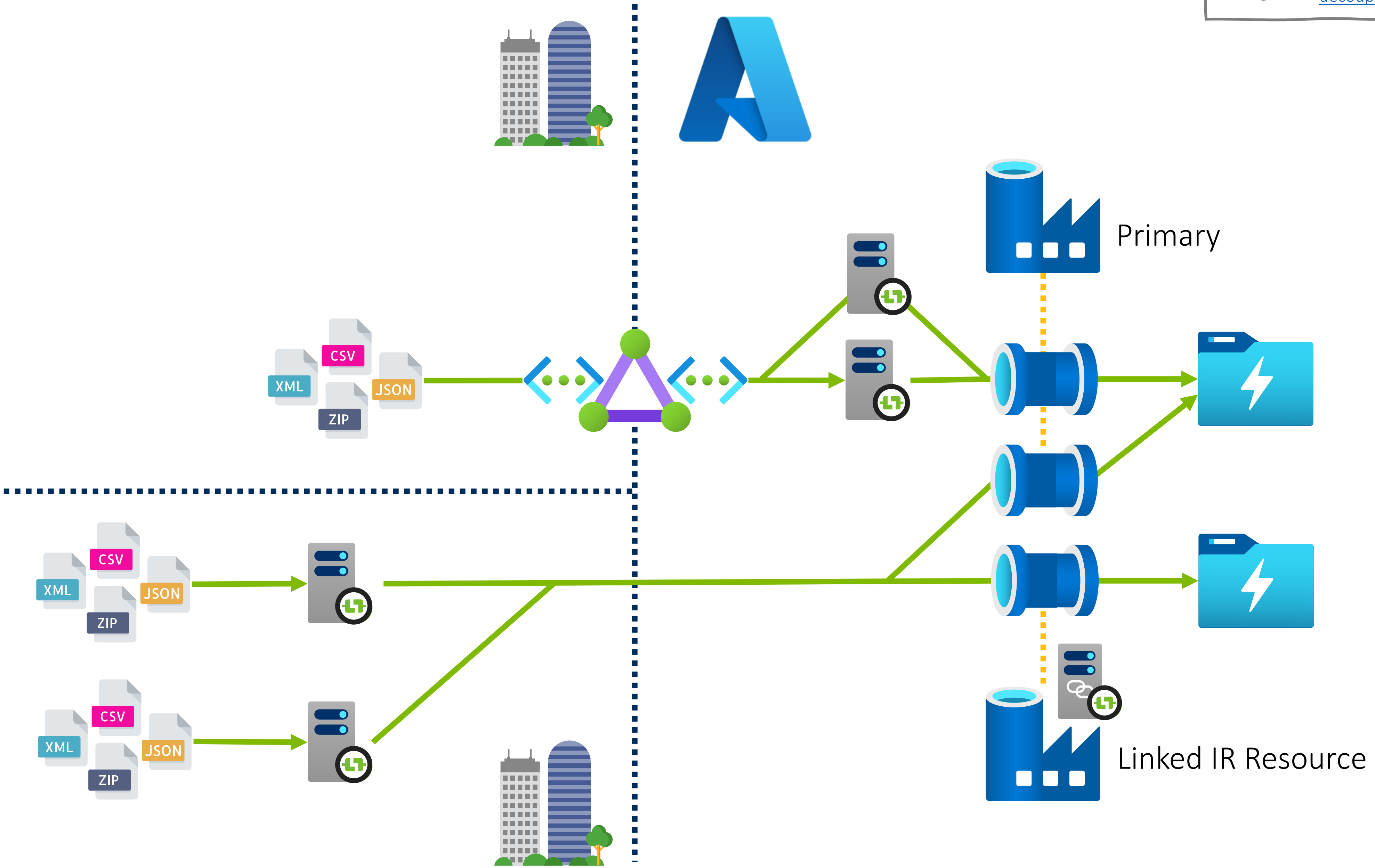
Runtime

Flexible Region

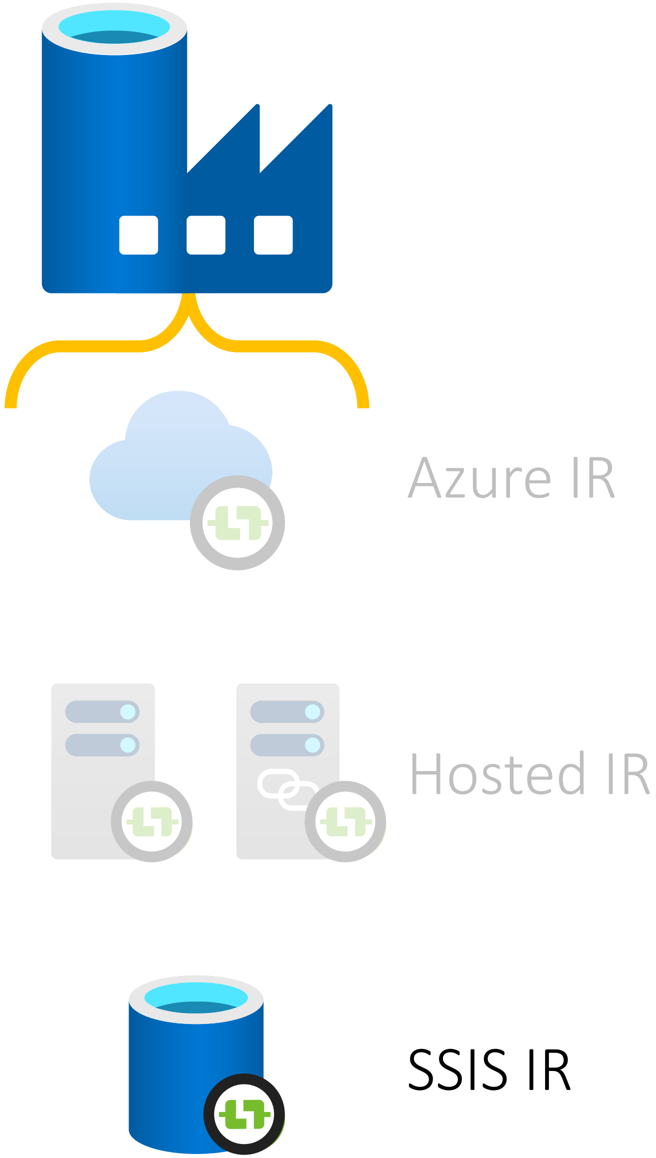
Hosted IR Advanced Patterns

Scaling Azure Data Integration Pipelines

https://mrpaulandrew.com/2021/08/10/scaling-azure-data-integration-pipelines-decoupling-data-extract-and-transform/



SSIS Integration Runtime



Azure IR

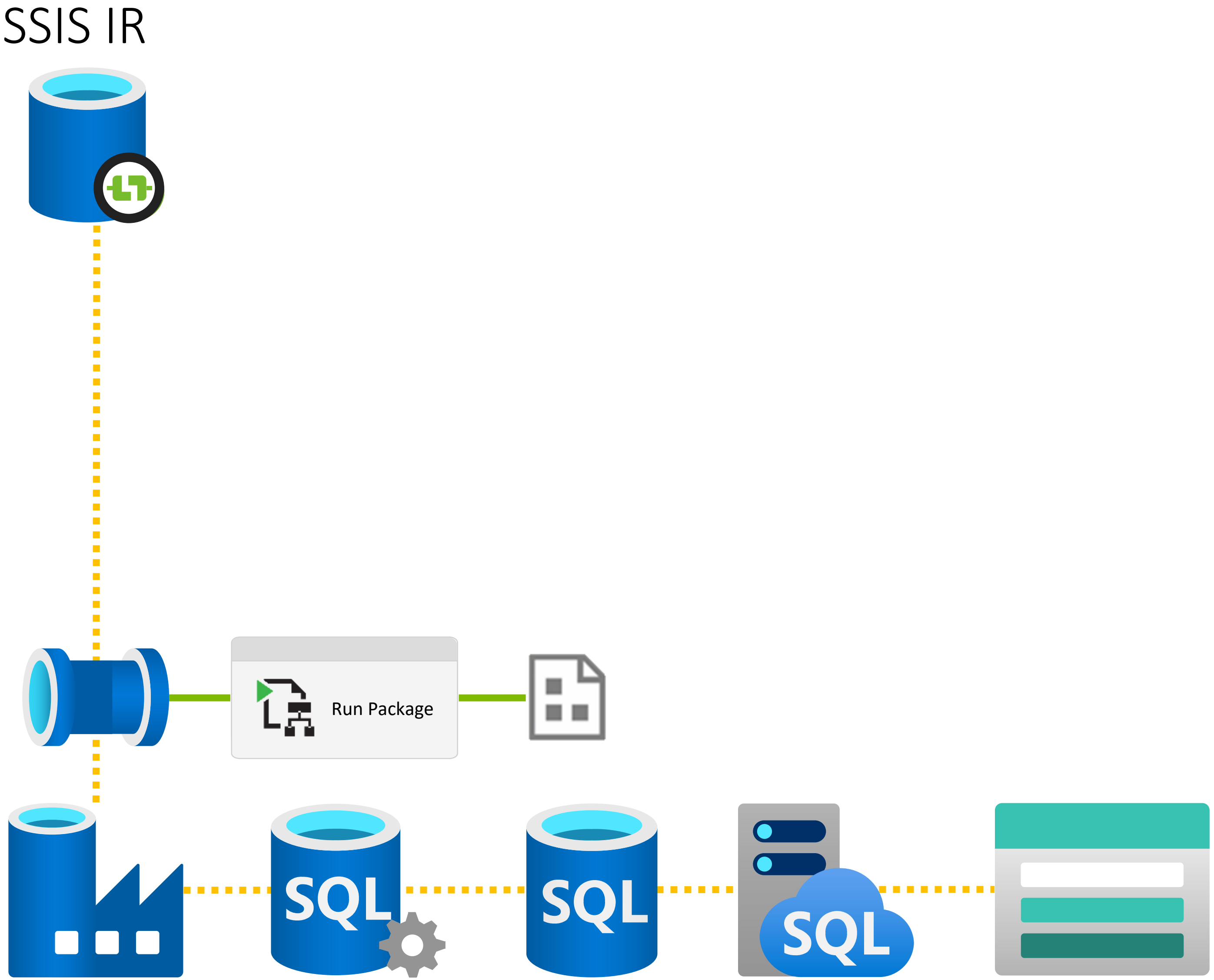
Hosted IR

SSIS IR

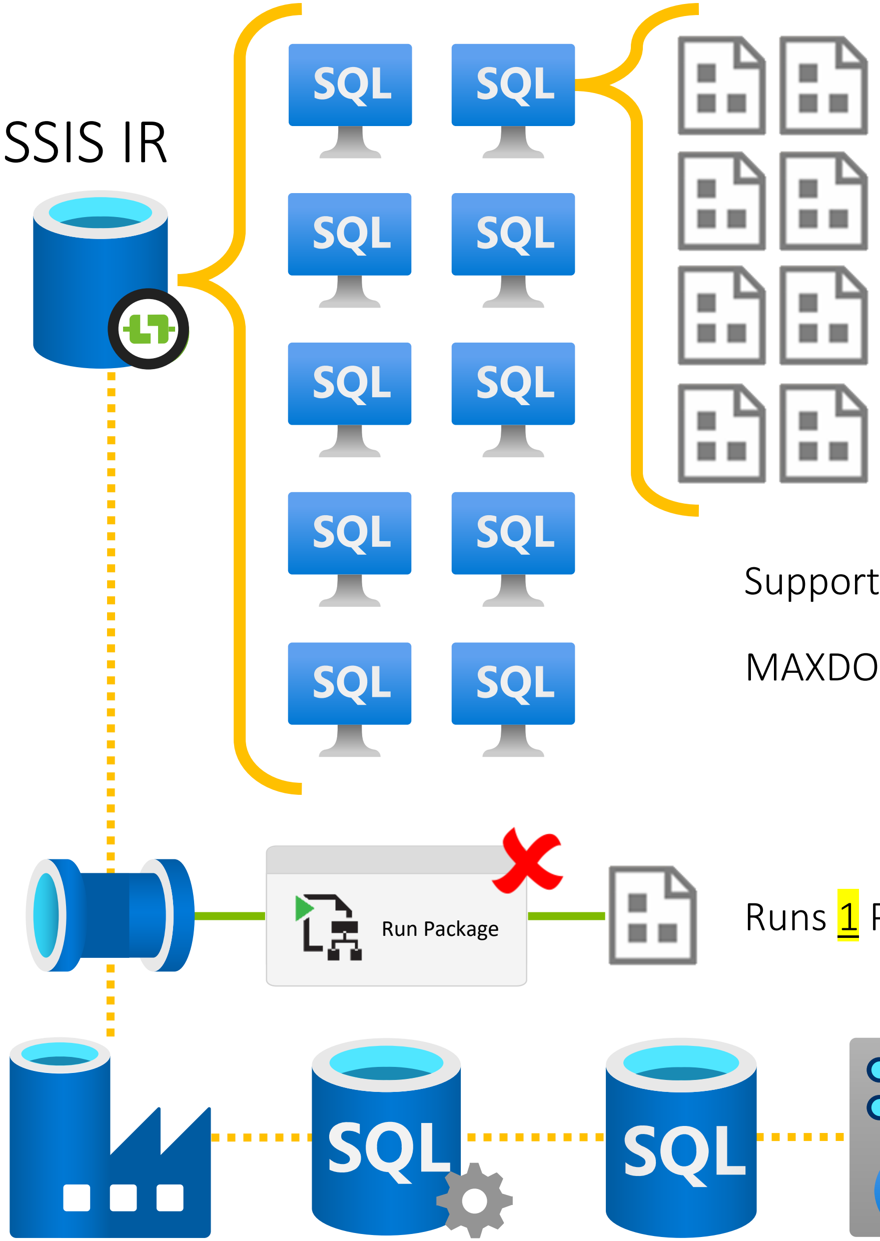
Running an SSIS Package in Azure

SSIS IR

Running an SSIS Package in Azure



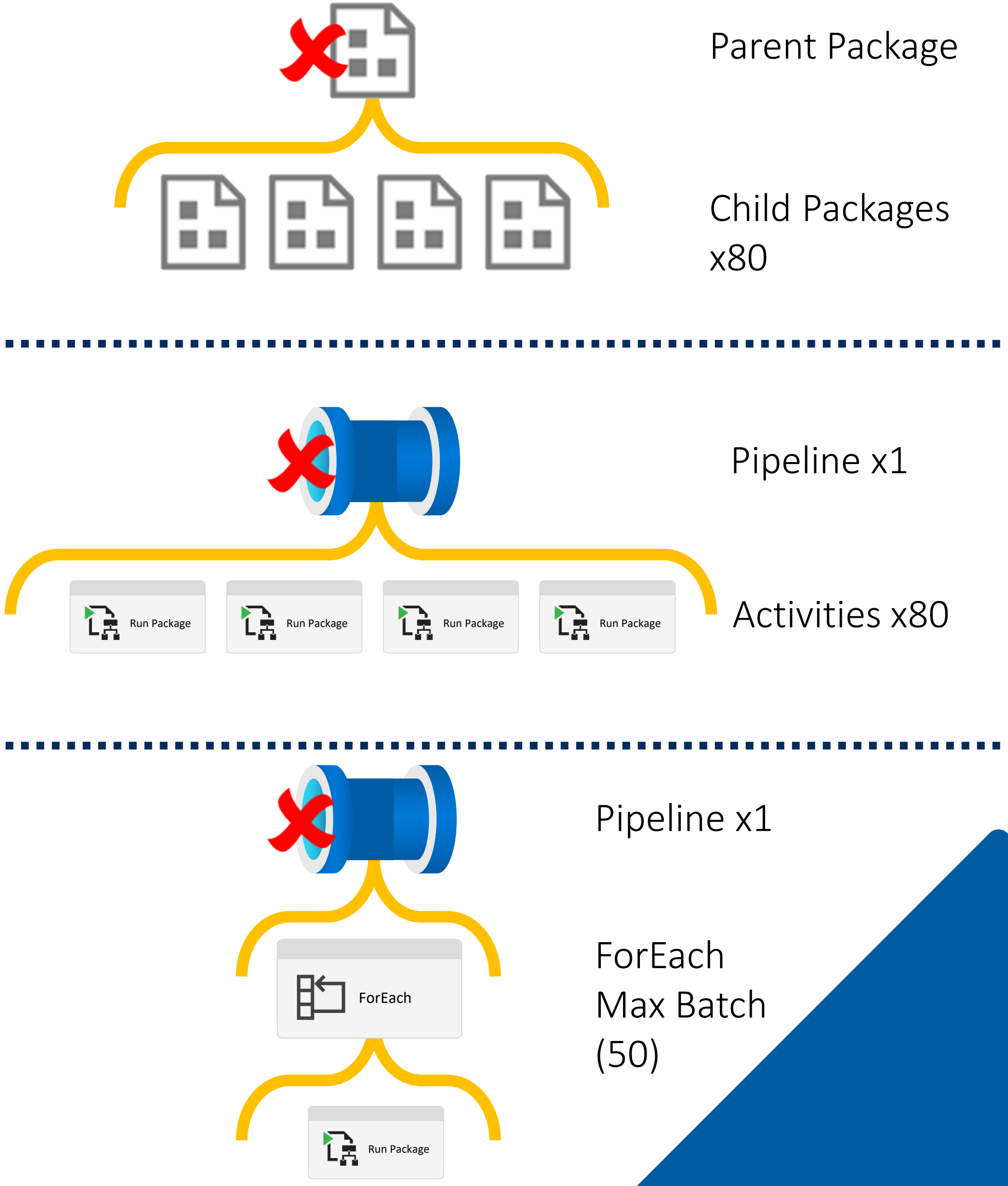
Problem: Using All Of The SSIS IR Compute



Supports 80 Concurrent Packages

MAXDOP = 80

Runs 1 Package



Parent Package

Child Packages x80

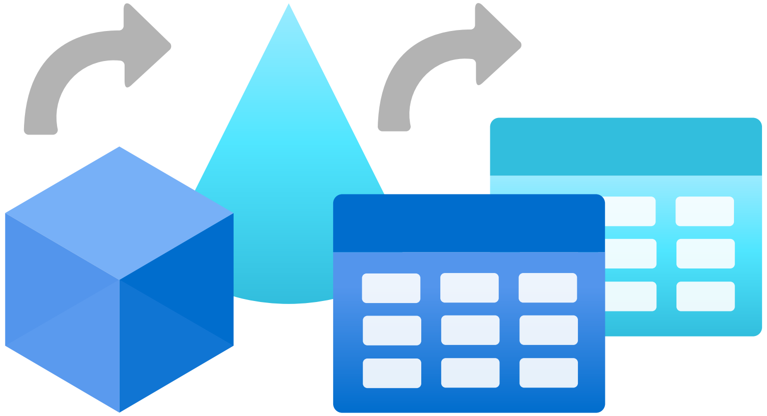
Pipeline x1

Activities x80

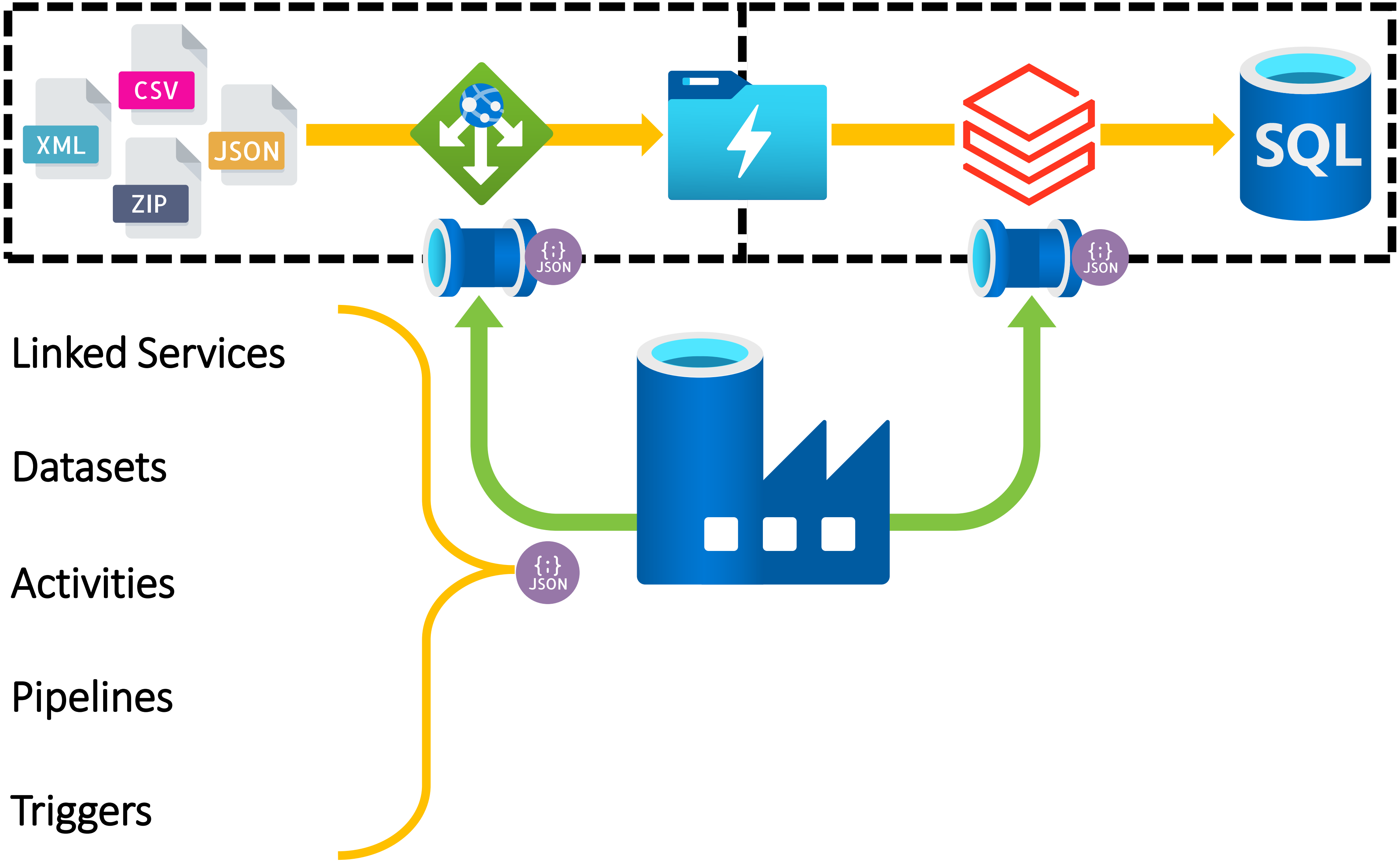
Pipeline x1

ForEach Max Batch (50)

Data Flows

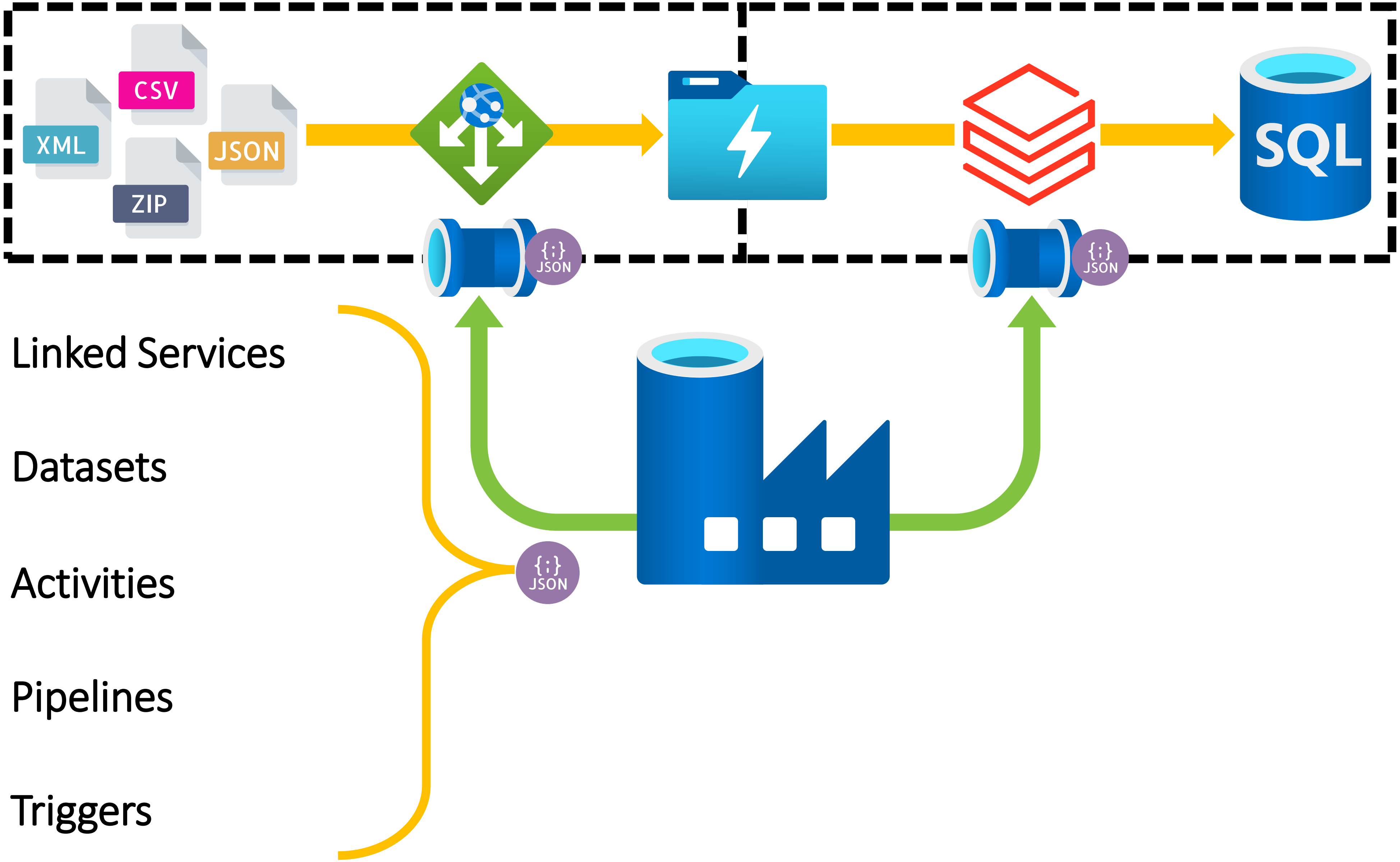


Integration Components

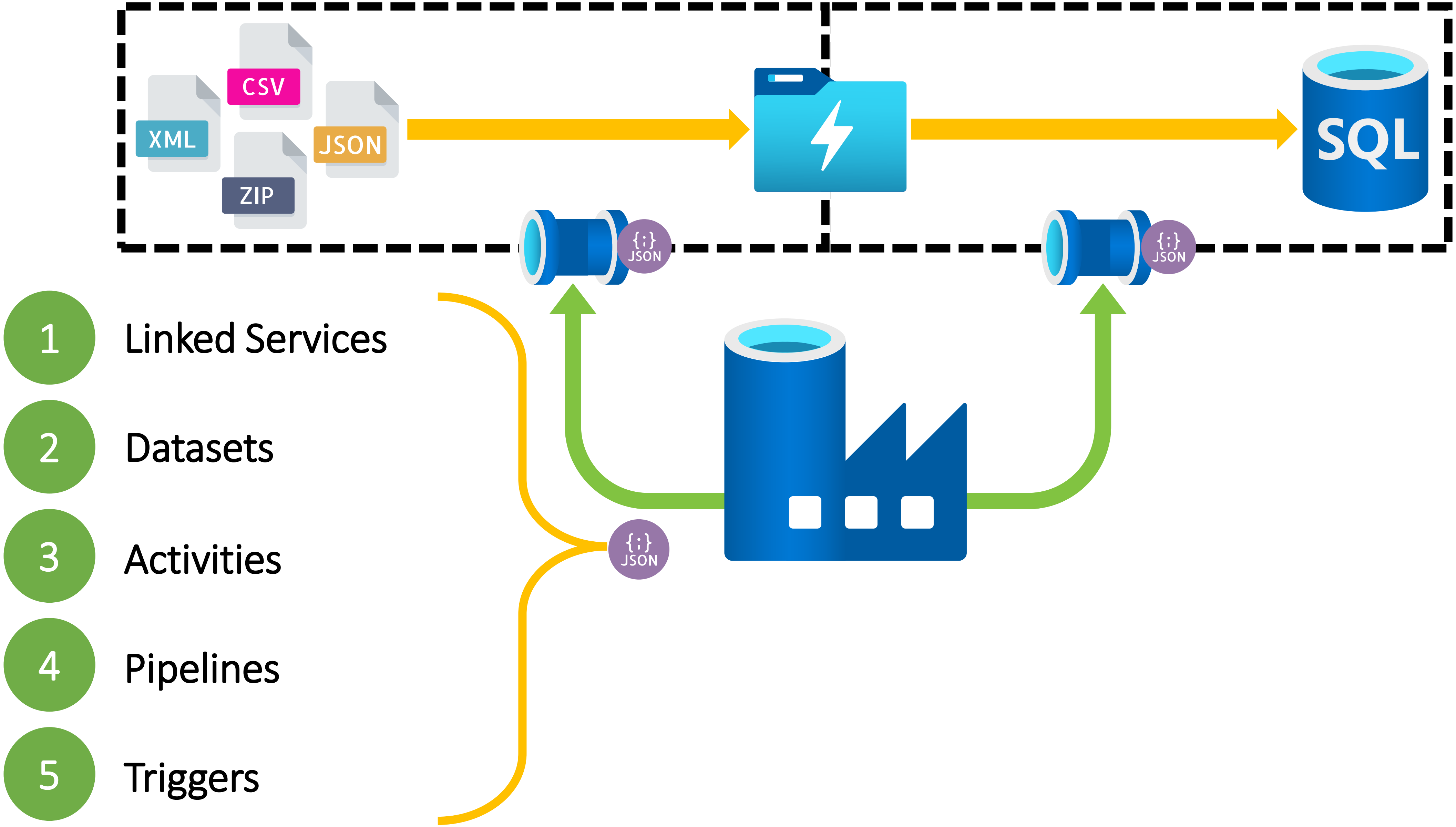


1. Linked Services
2. Datasets
3. Activities
4. Pipelines
5. Triggers

Integration Control Flow Components



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



1 Linked Services

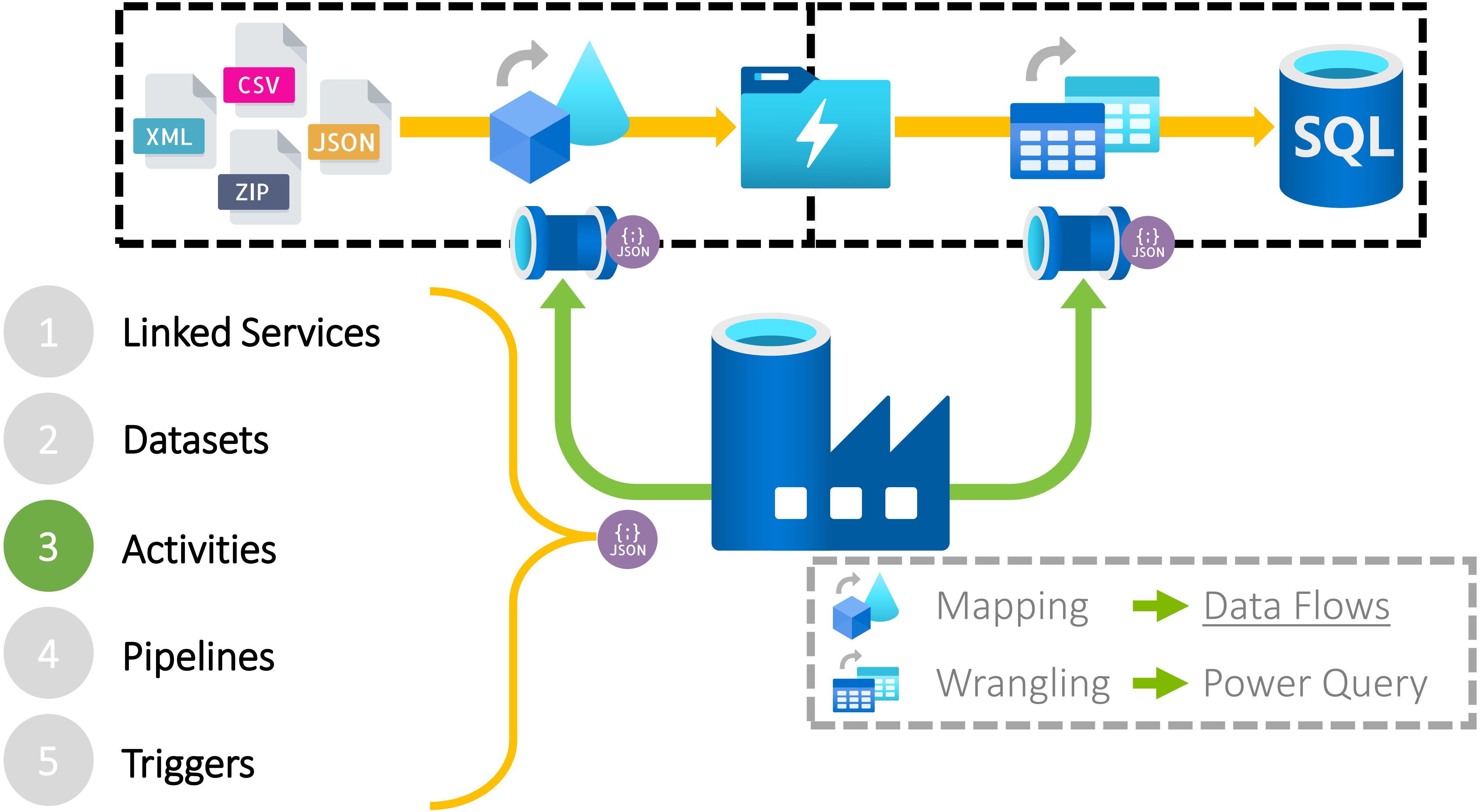
2 Datasets

3 Activities

4 Pipelines

5 Triggers

Integration Data Flow (Transformation) Activities



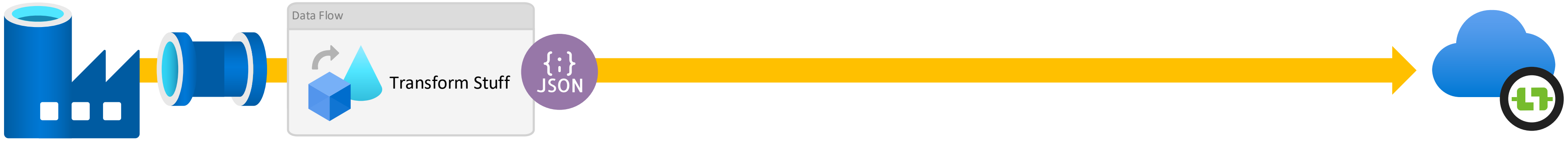
1. Linked Services
2. Datasets
3. Activities
4. Pipelines
5. Triggers

Mapping → Data Flows

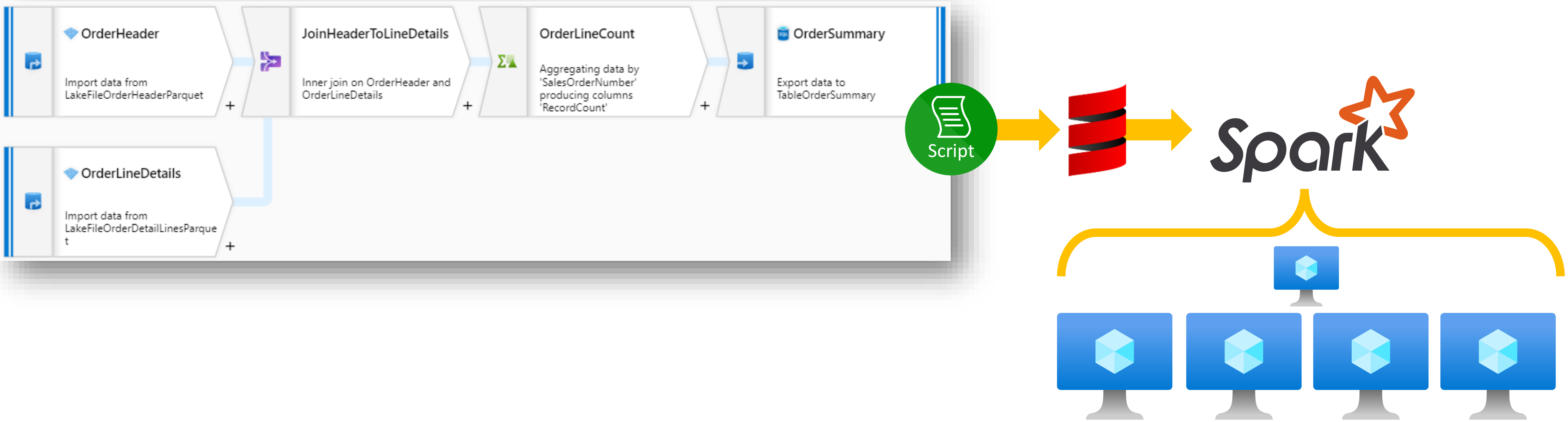
Wrangling → Power Query

What is a ~~Mapping~~ Data Flow?

Control Flow



Data Flow



OrderHeader

Import data from LakeFileOrderHeaderParquet

JoinHeaderToLineDetails

Inner join on OrderHeader and OrderLineDetails

OrderLineCount

Aggregating data by 'SalesOrderNumber' producing columns 'RecordCount'

OrderSummary

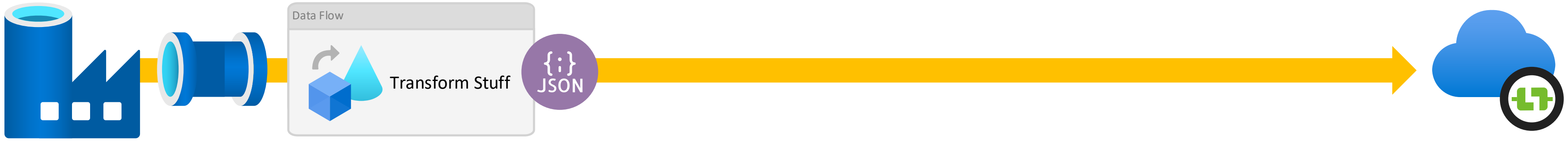
Export data to TableOrderSummary

OrderLineDetails

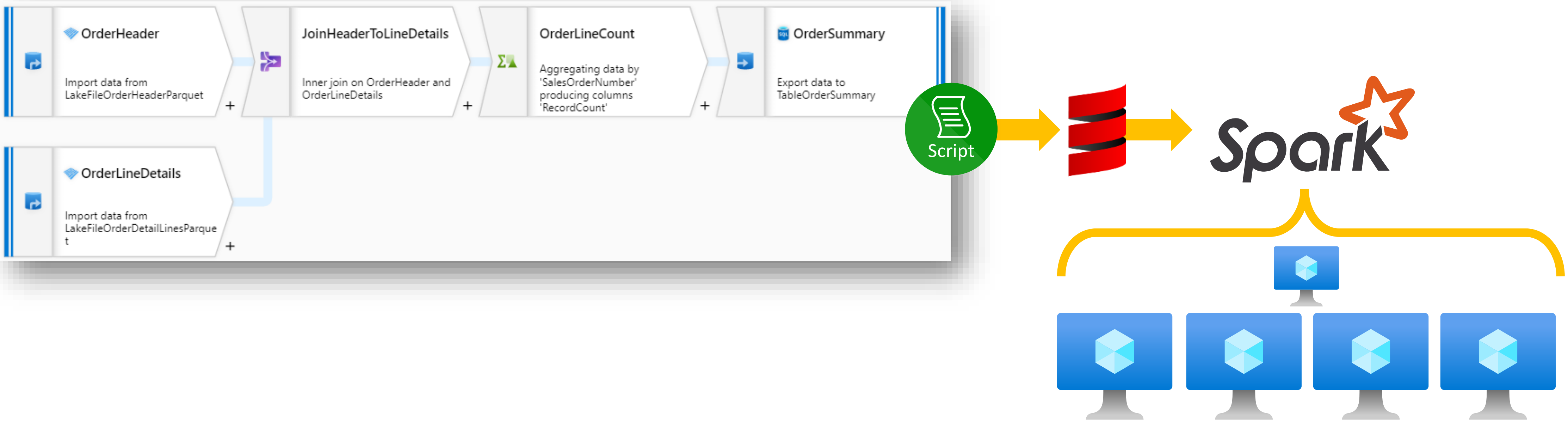
Import data from LakeFileOrderDetailLinesParque t

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

Control Flow



Data Flow



OrderHeader

Import data from LakeFileOrderHeaderParquet

JoinHeaderToLineDetails

Inner join on OrderHeader and OrderLineDetails

OrderLineCount

Aggregating data by 'SalesOrderNumber' producing columns 'RecordCount'

OrderSummary

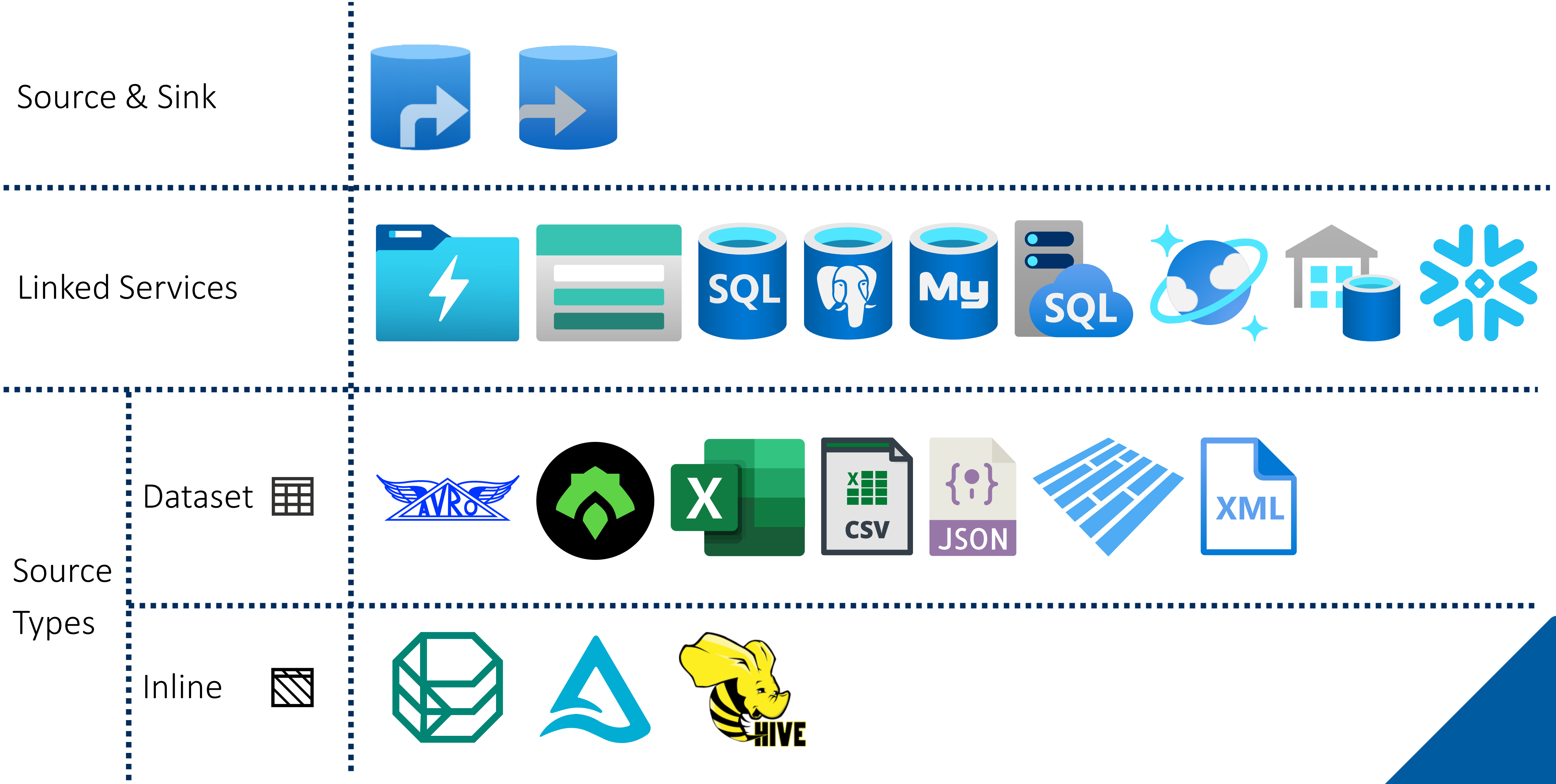
Export data to TableOrderSummary

OrderLineDetails

Import data from LakeFileOrderDetailLinesParque t

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

- Flatten
- Parse

- Filter
- Sort
- Alter Row

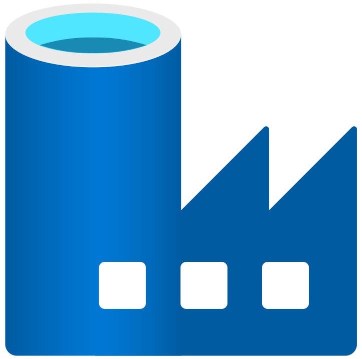
Key

Input & Output Modifiers

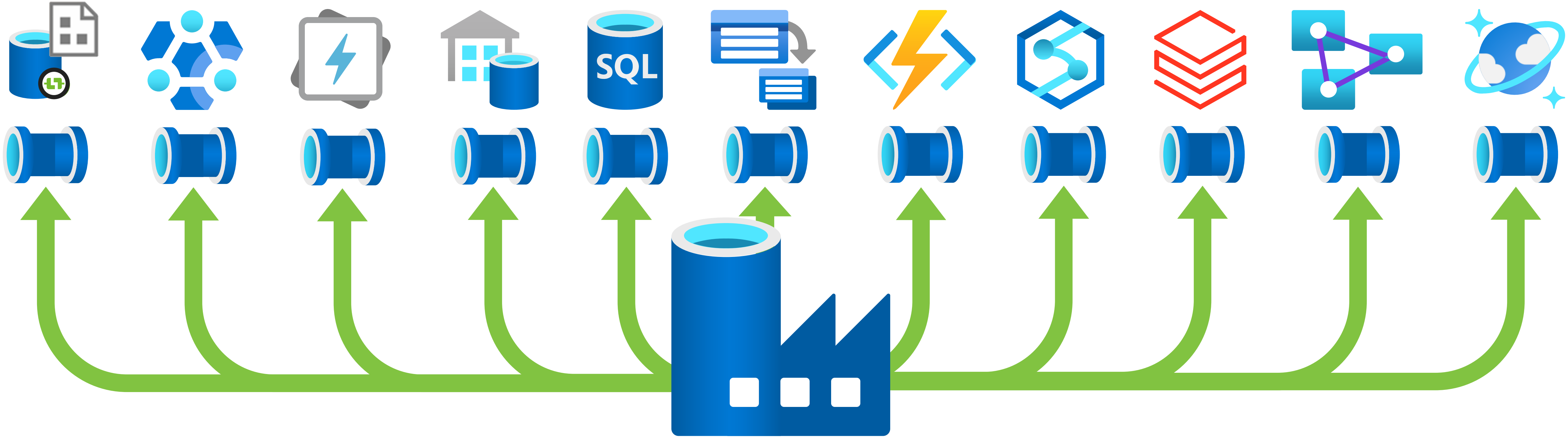
Schema Modifiers

Formatters

Row Modifiers



Other Data Transformation Services in Azure



SSIS Packages

HD Insight

Data Lake Analytics

Synapse SQLDW

SQL Database

Batch Service

Durable Functions

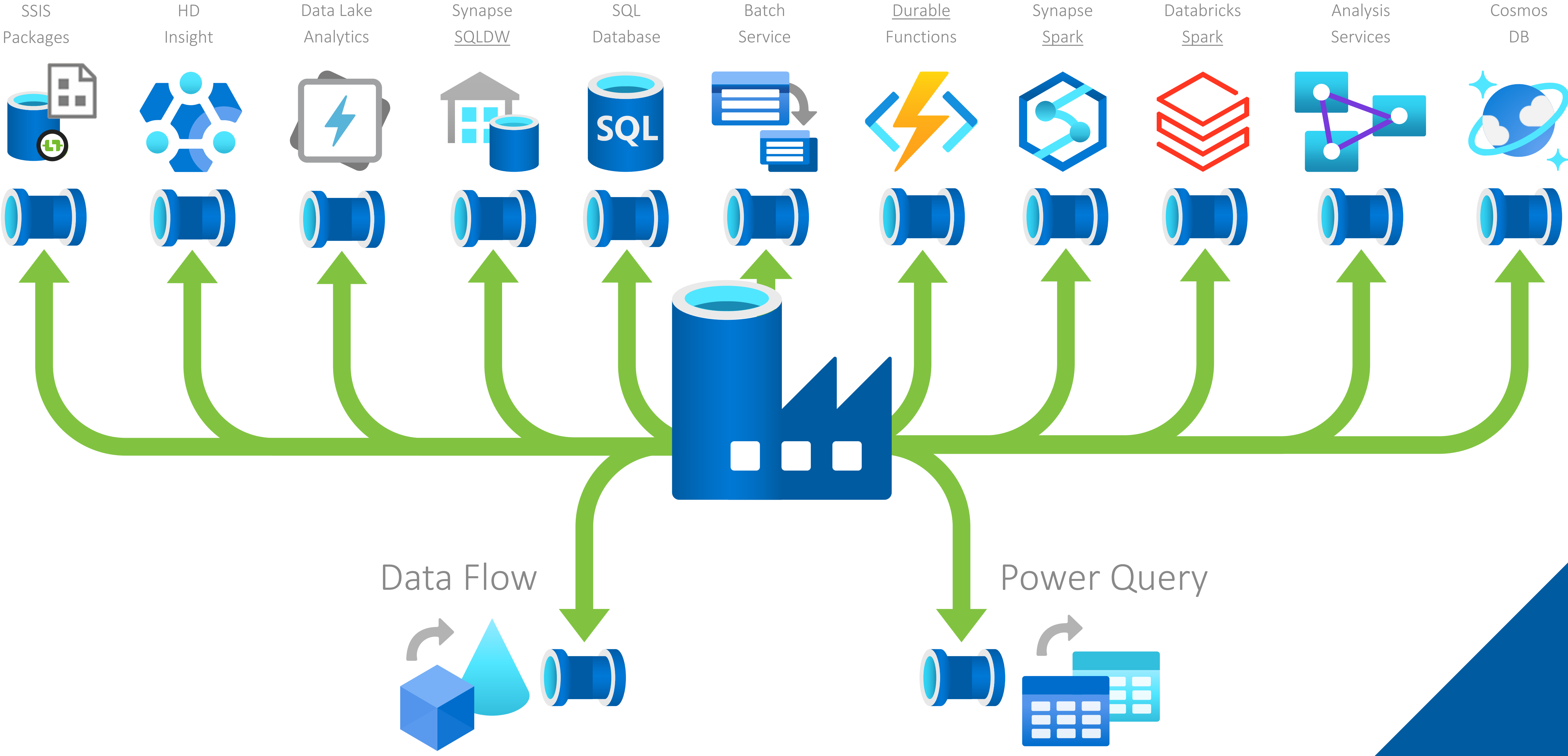
Synapse Spark

Databricks Spark

Analysis Services

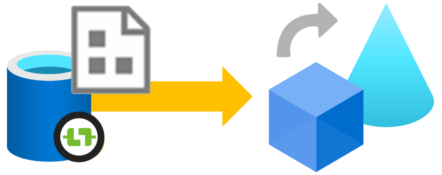
Cosmos DB

When Should We Use These Integration Pipeline Transformation Activities?

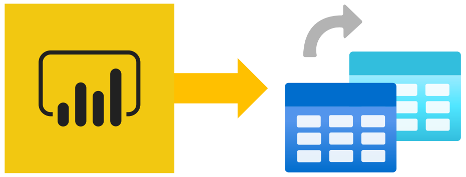


Use Cases

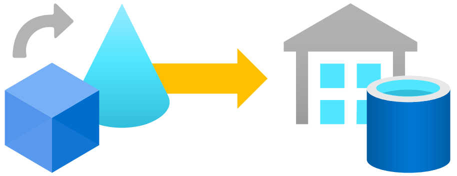
SSIS developers who are transferring existing skills to cloud native technologies have a very low barrier to entry and don’t need to worry about distributed compute to get started.



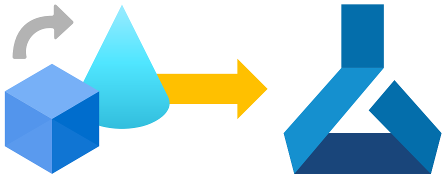
Data engineering made easy for the power users who has grown out of Power BI following a series of Data Lake exploration sessions.



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



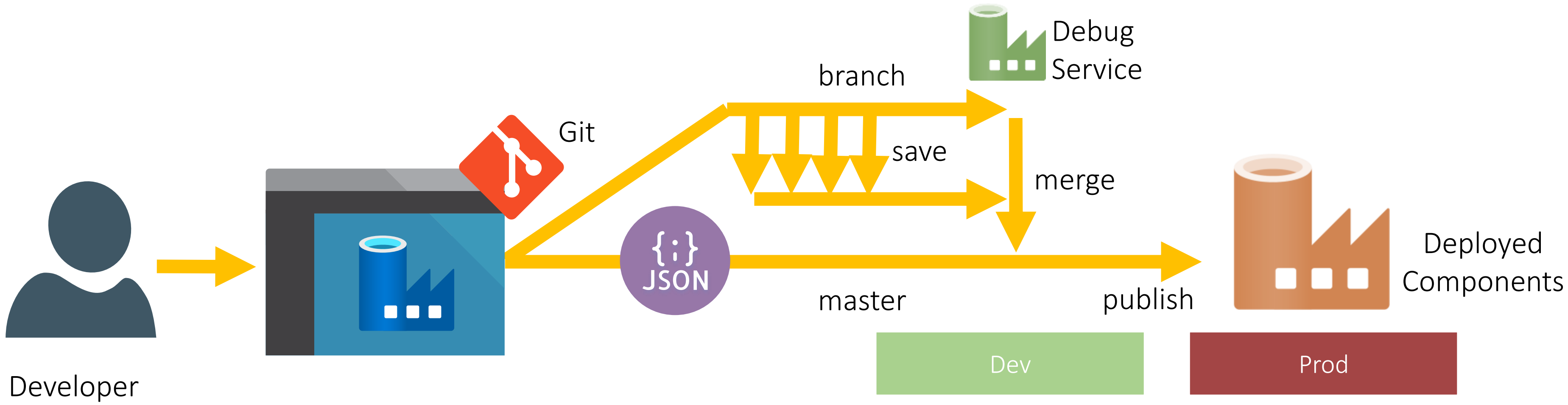
Simpler and quicker data wrangling for data scientists that want to quickly prepare multiple raw datasets ready for model training and testing, also with the ability to use large amounts of compute.



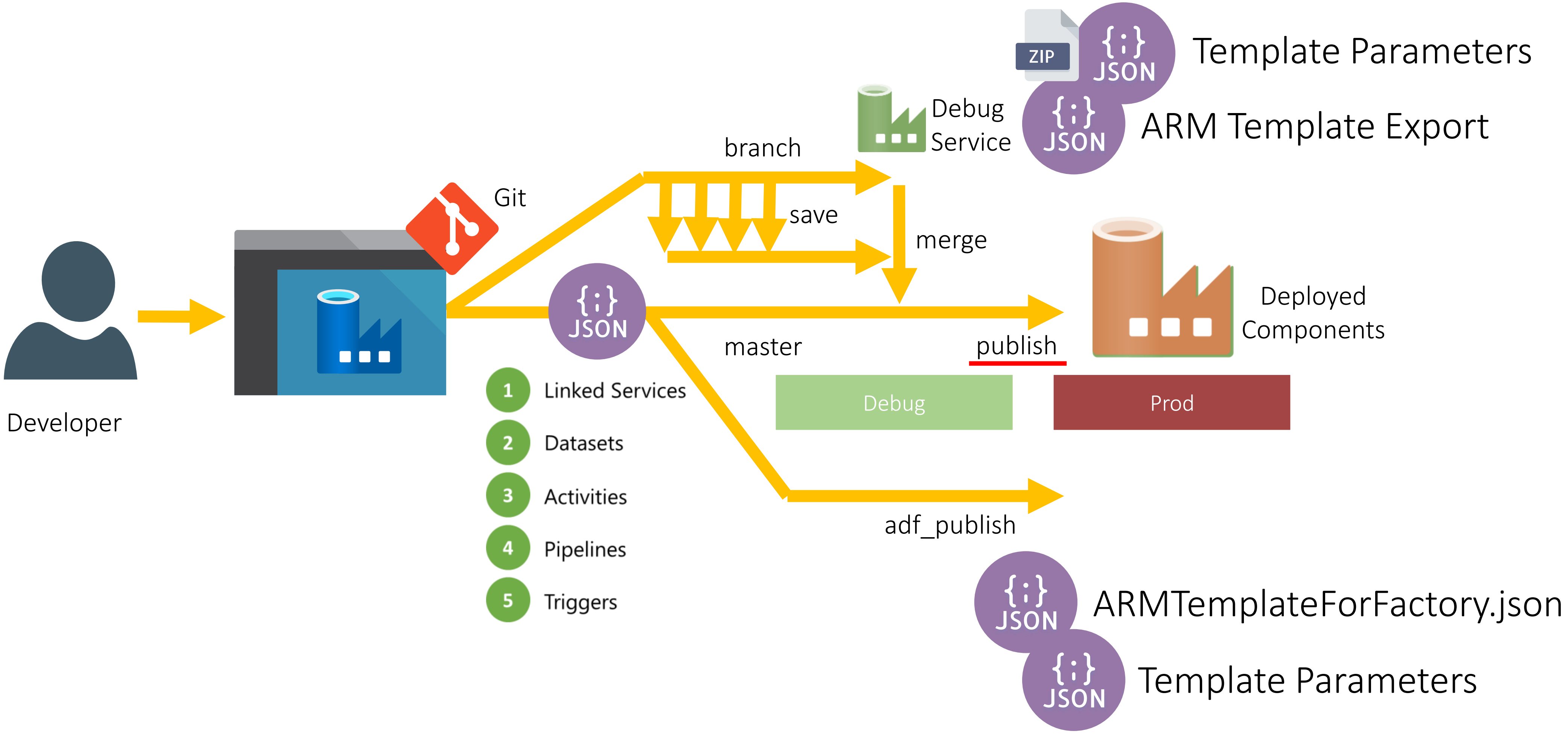
Data Flows used to deliver all data transformation workloads as part of a end to end cloud based data analytics/warehouse solution.

Data Flows script dynamically generated from external metadata and injected into like we once did with BIML for SSIS packages.

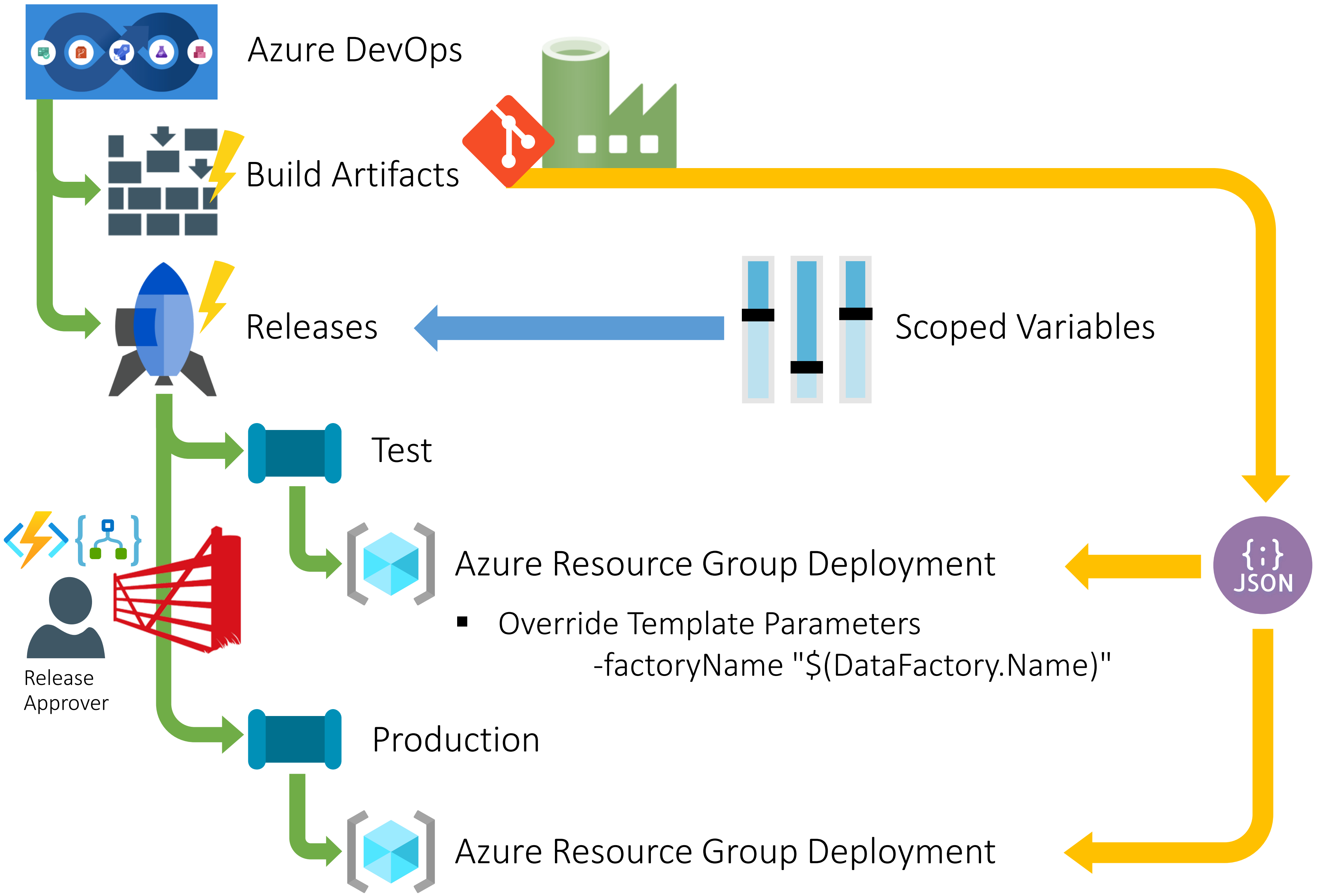
Source Control & Deployments


Debug
Service
branch
Git
save
merge
Deployed
Components
master
publish
Dev
Prod
Developer

Getting Our ADF Source Code



Data Factory Continuous Delivery



Azure DevOps

Build Artifacts

Releases

Scoped Variables

Test

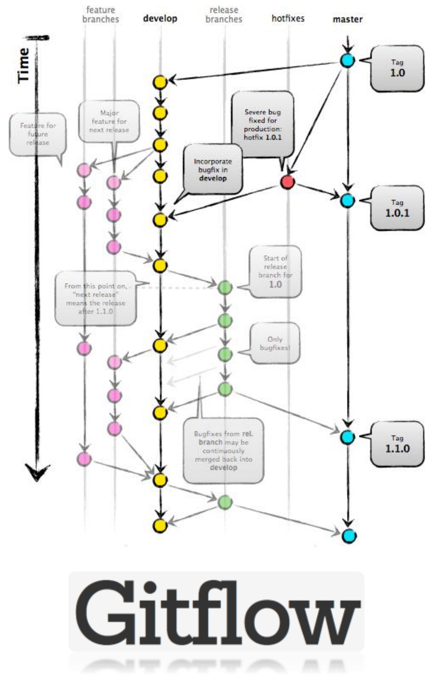
Azure Resource Group Deployment

- Override Template Parameters
  -factoryName "$(DataFactory.Name)"

Production

Azure Resource Group Deployment

ARMTemplateForFactory.json



Gitflow

Data Factory Continuous Delivery - Simple

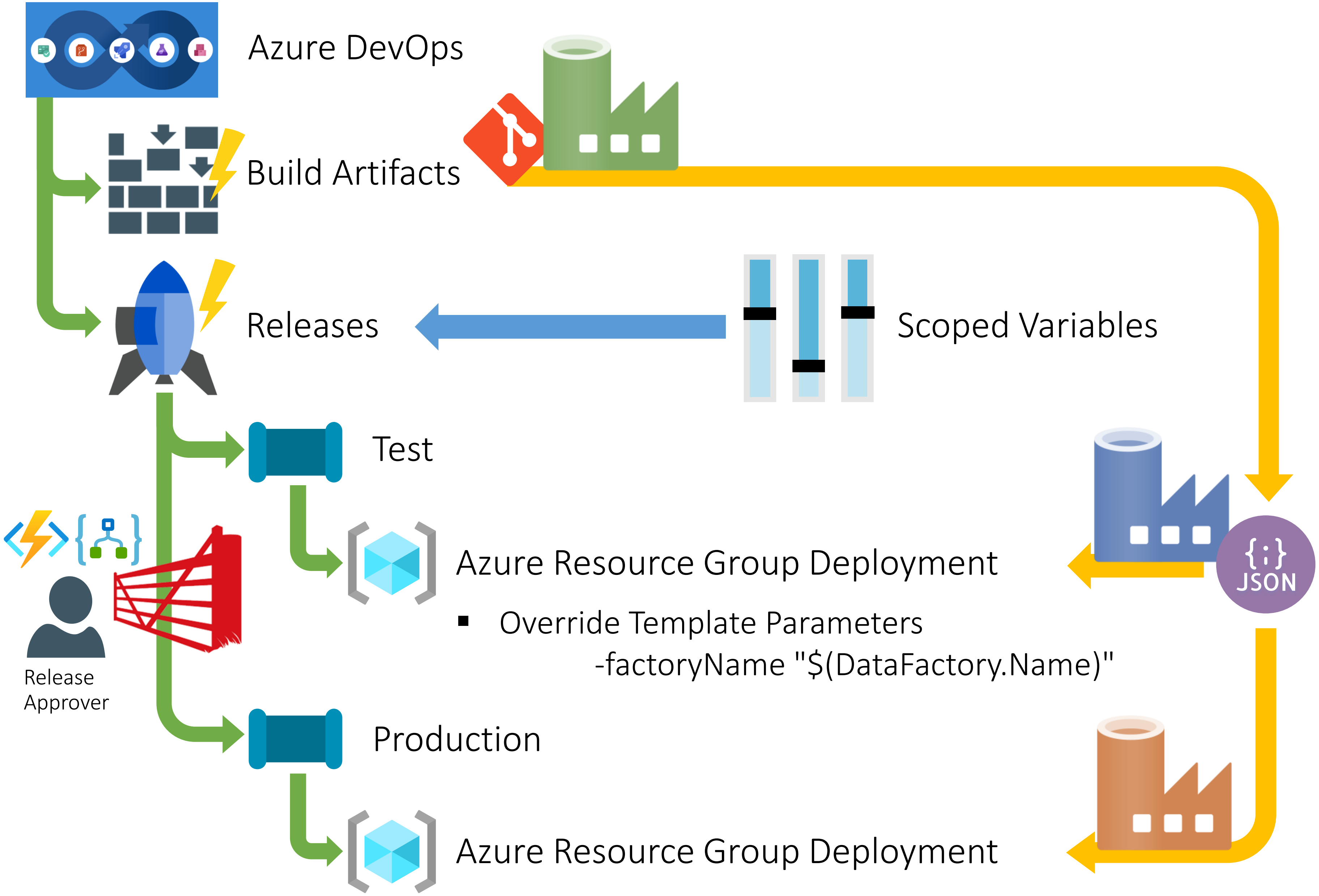
1 Linked Services

2 Datasets

3 Activities

4 Pipelines

5 Triggers



Azure DevOps

Build Artifacts

Releases

Scoped Variables

Test

Azure Resource Group Deployment

- Override Template Parameters

  -factoryName "$(DataFactory.Name)"

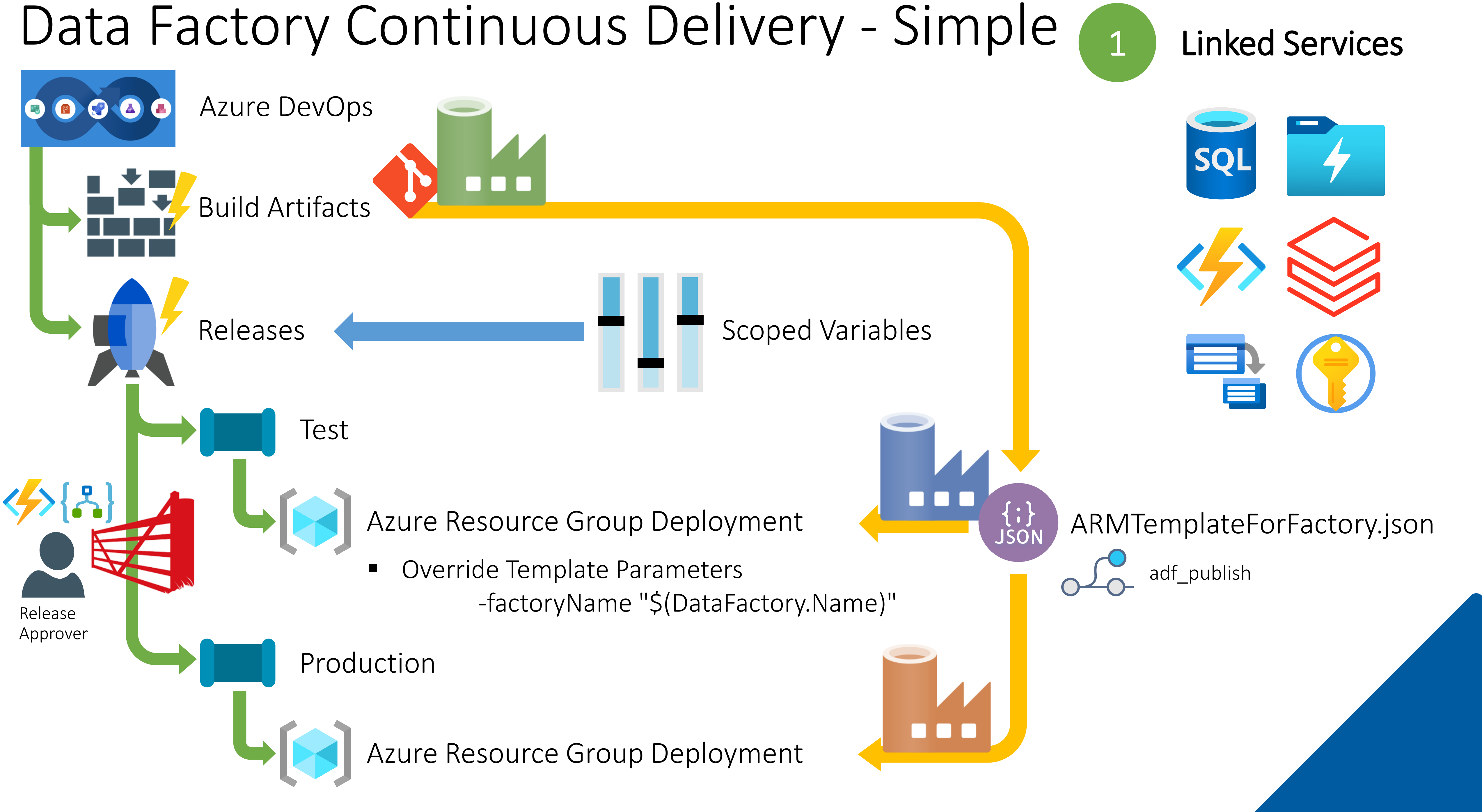
Release Approver

Production

Azure Resource Group Deployment

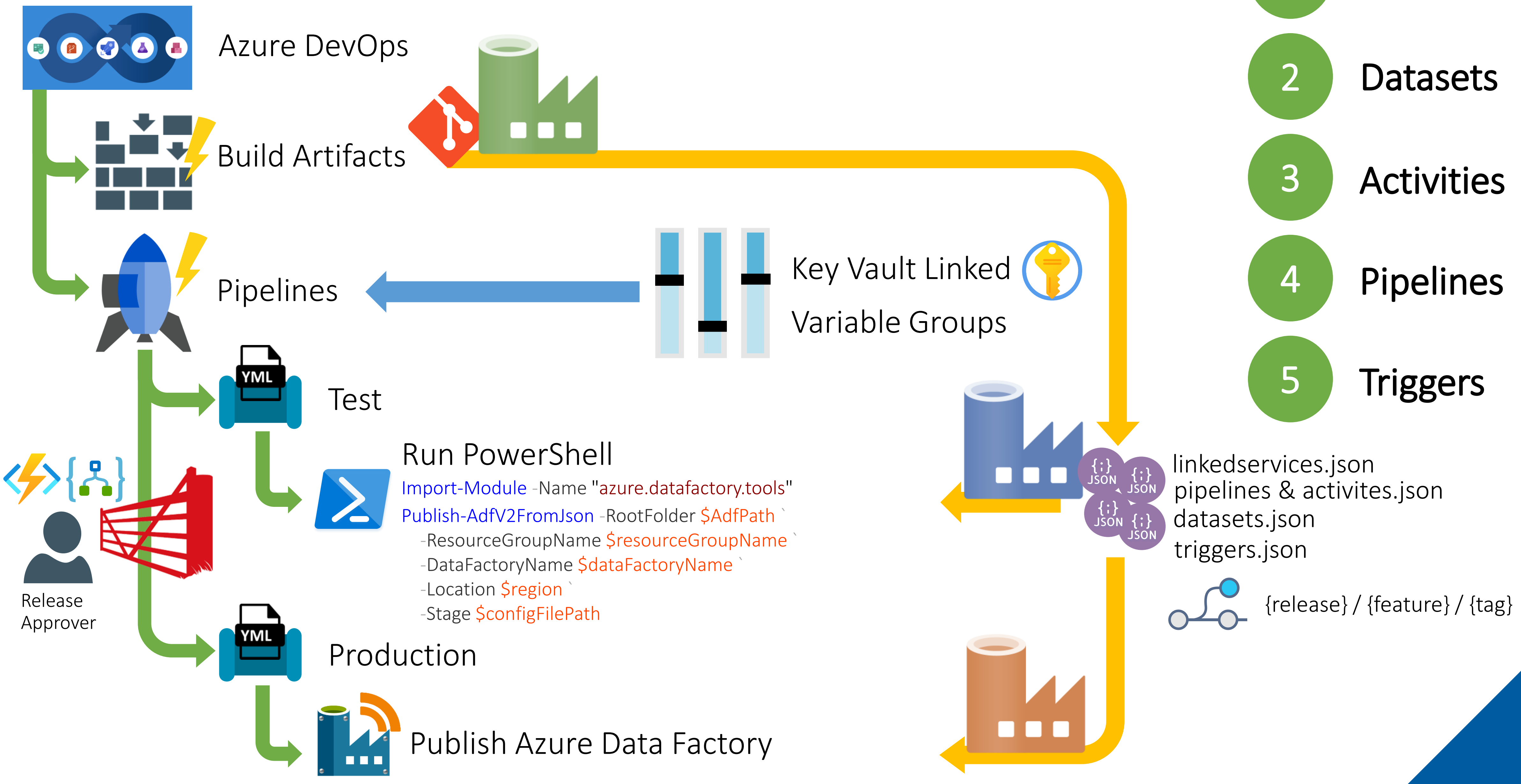
ARMTemplateForFactory.json

Data Factory Continuous Delivery - Simple



- Override Template Parameters
  -factoryName "$(DataFactory.Name)"

Data Factory Continuous Delivery - Complex



```
Import-Module -Name "azure.datafactory.tools"
Publish-AdfV2FromJson -RootFolder $AdfPath `
    -ResourceGroupName $resourceGroupName `
    -DataFactoryName $dataFactoryName `
    -Location $region `
    -Stage $configFilePath
```

1 Linked Services

2 Datasets

3 Activities

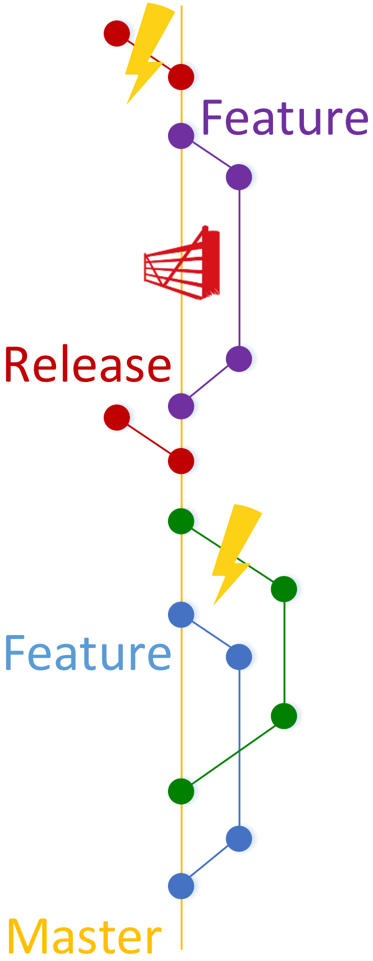
4 Pipelines

5 Triggers

https://www.powershellgallery.com/packages/azure.datafactory.tools

Data Factory DevOps Story Summary

What is your code branching strategy?



Which source control tool to use?

How many environments do we want?

What deployment method do we want to use?

What artifacts are we going to use?...

OR

How much control do you want?

ARMTemplate ForFactory.json

linkedservices.json

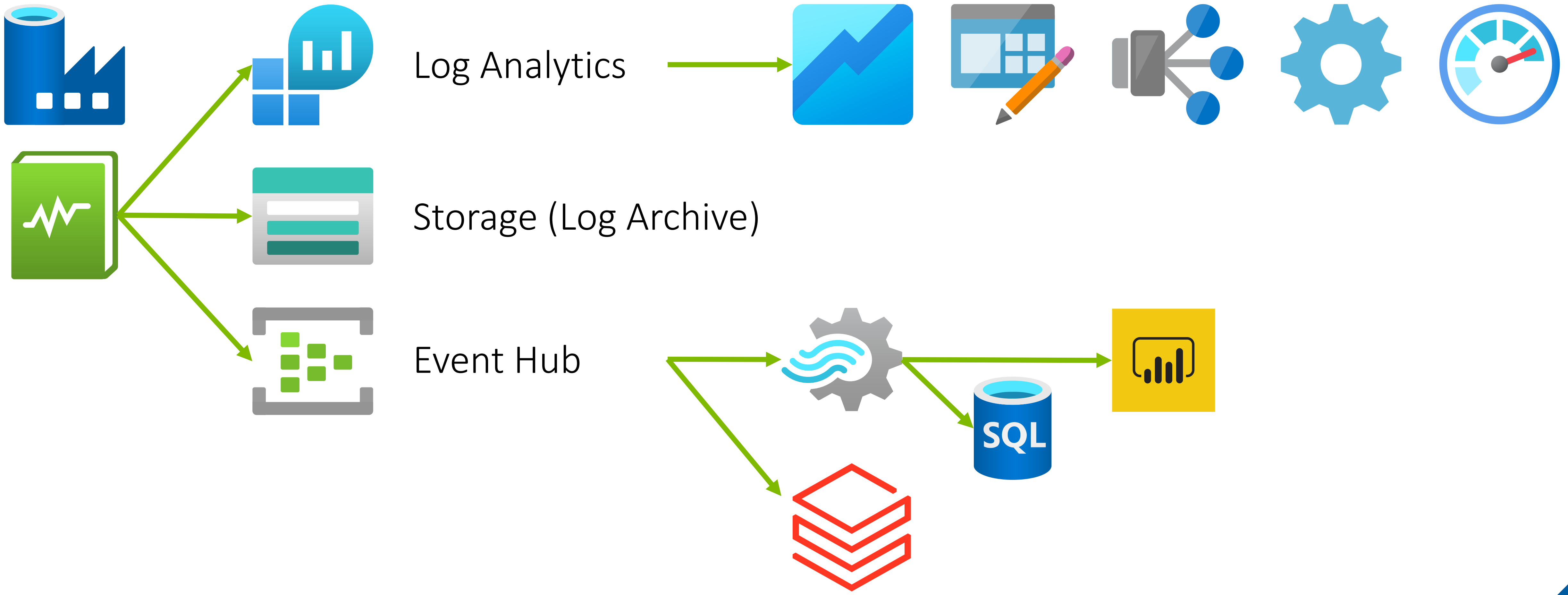
pipelines & activites.json

datasets.json

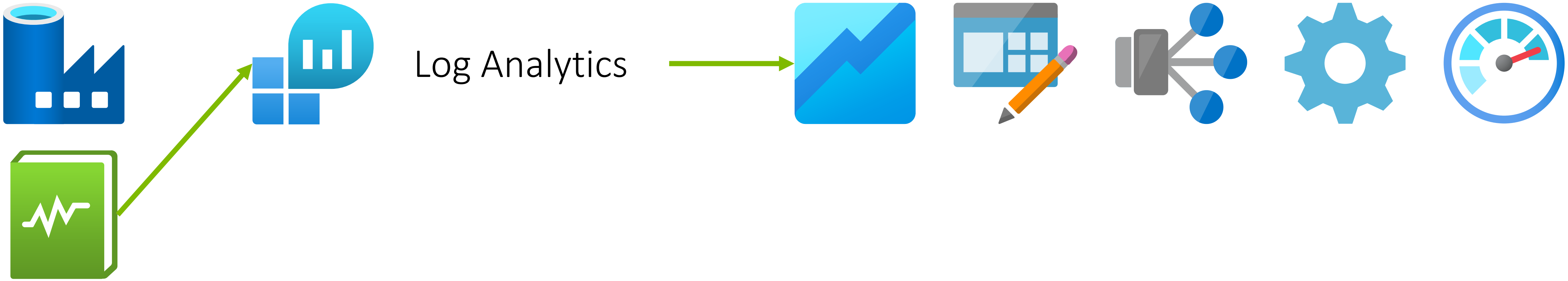
triggers.json

Monitoring & Logging

Diagnostic Settings

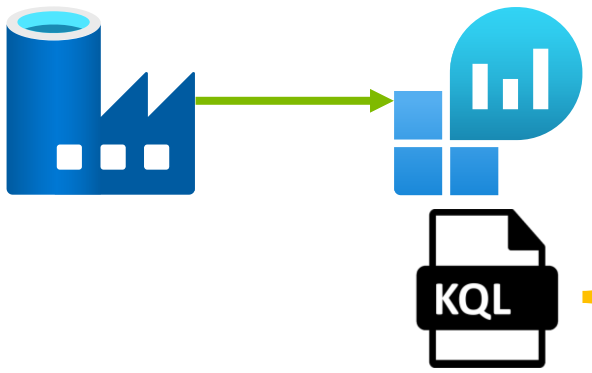


Diagnostic Settings

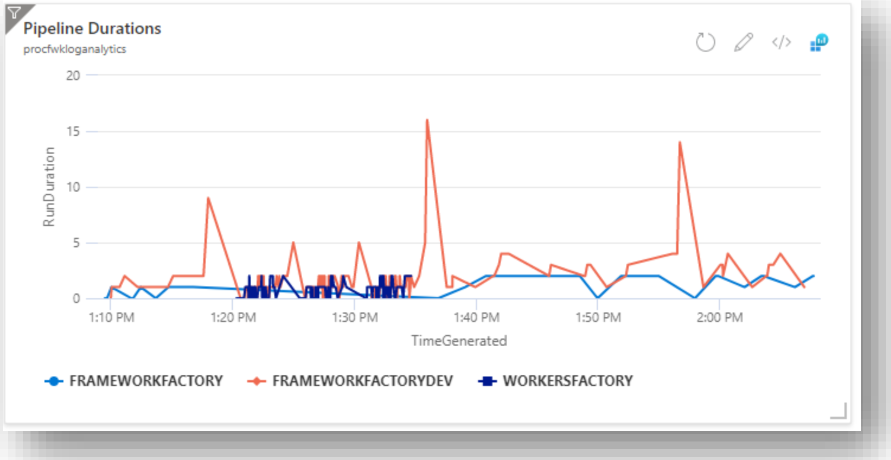


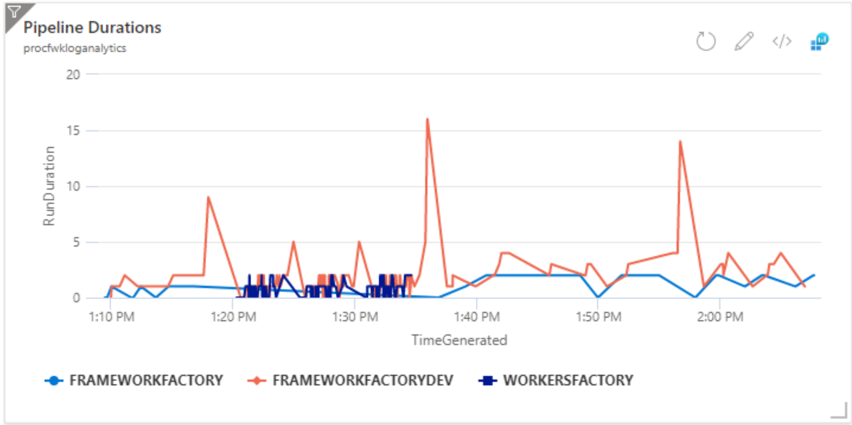
Log Analytics

Using Log Analytics



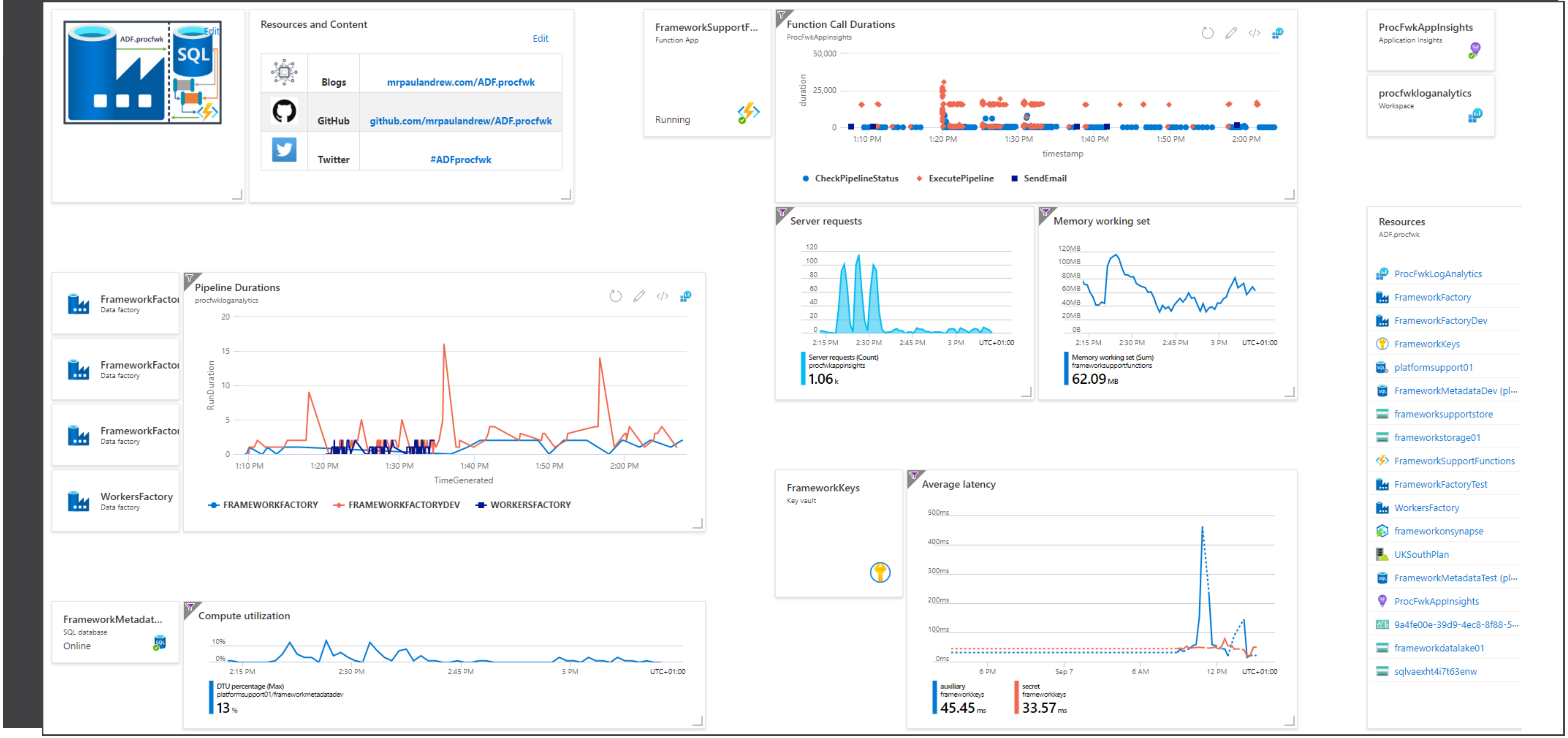
```
ADFPipelineRunDurations
 | project
         TimeGenerated,
         Start,
         End,
         ['DataFactory'] = substring(ResourceId, 121, 100),
         Status,
         PipelineName,
         Parameters,
         ["RunDuration"] = datetime_diff('Minute', End, Start)
 | where
         TimeGenerated > ago(1h)
         and Status !in ('InProgress','Queued','Cancelling')
```





Pipeline Durations

procfwkloganalytics



Resources and Content

Edit

	Blogs	mrpaulandrew.com/ADF.procfwk
	GitHub	github.com/mrpaulandrew/ADF.procfwk
	Twitter	#ADFprocfwk

FrameworkSupportF...
Function App
Running

Function Call Durations
ProcFwkAppInsights

ProcFwkAppInsights
Application Insights

procfwkloganalytics
Workspace

Server requests

Server requests (Count)
procfwkappinsights
1.06 k

Memory working set

Memory working set (Sum)
frameworksupportfunctions
62.09 MB

Resources
ADF.procfwk

- ProcFwkLogAnalytics
- FrameworkFactory
- FrameworkFactoryDev
- FrameworkKeys
- platformsupport01
- FrameworkMetadataDev (pl...
- frameworksupportstore
- frameworkstorage01
- FrameworkSupportFunctions
- FrameworkFactoryTest
- WorkersFactory
- frameworkonsynapse
- UKSouthPlan
- FrameworkMetadataTest (pl...
- ProcFwkAppInsights
- 9a4fe00e-39d9-4ec8-8f88-5...
- frameworkdatalake01
- sqlvaexht4i7t63enw

FrameworkFactor
Data factory

FrameworkFactor
Data factory

FrameworkFactor
Data factory

WorkersFactory
Data factory

Pipeline Durations
procfwkloganalytics

FrameworkKeys
Key vault

Average latency

auxiliary
frameworkkeys
45.45 ms

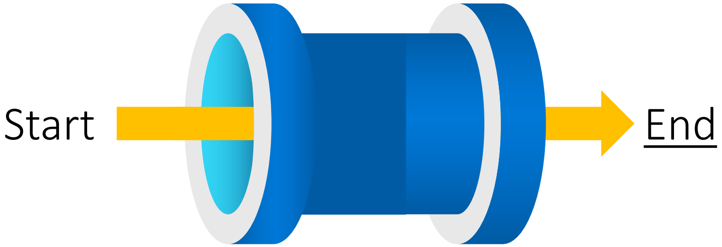
secret
frameworkkeys
33.57 ms

FrameworkMetadat...
SQL database
Online

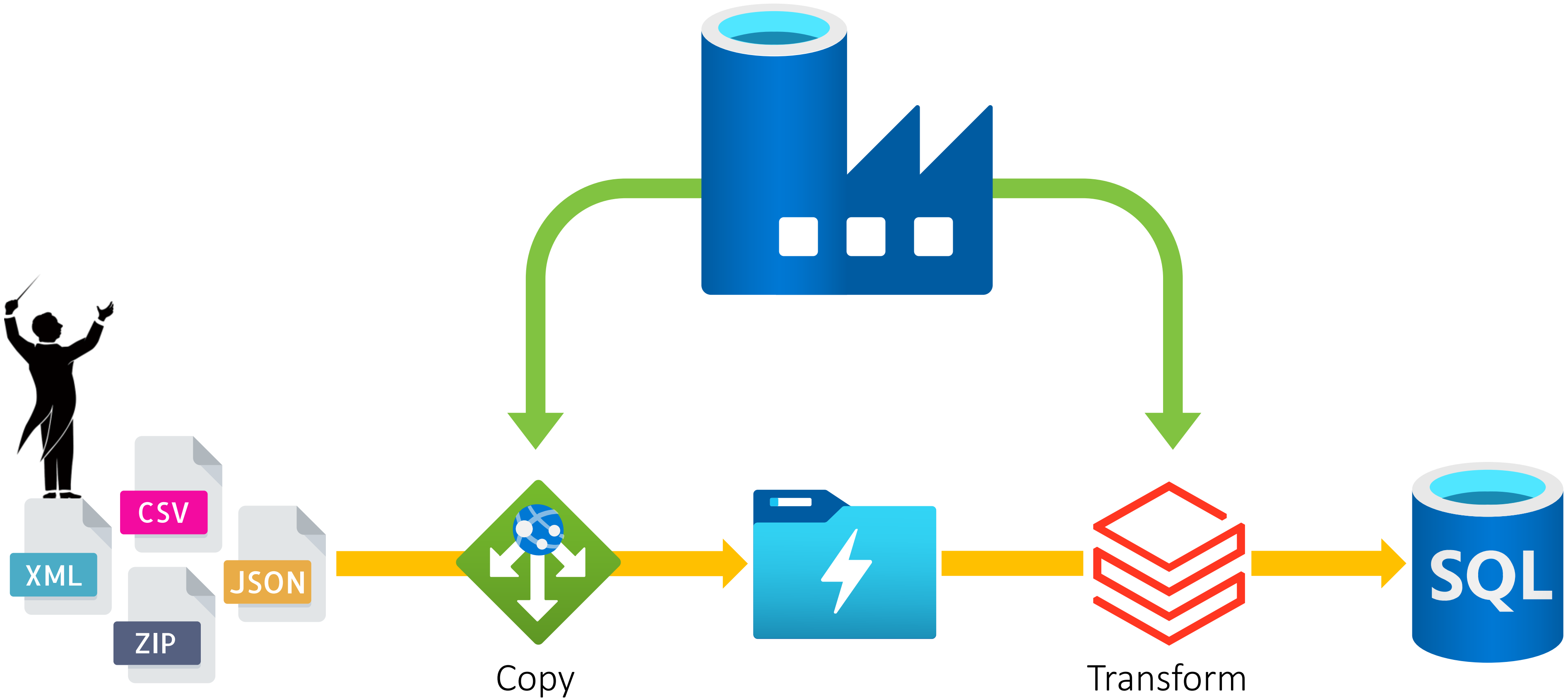
Compute utilization

DTU percentage (Max)
platformsupport01/frameworkmetadatadev
13 %

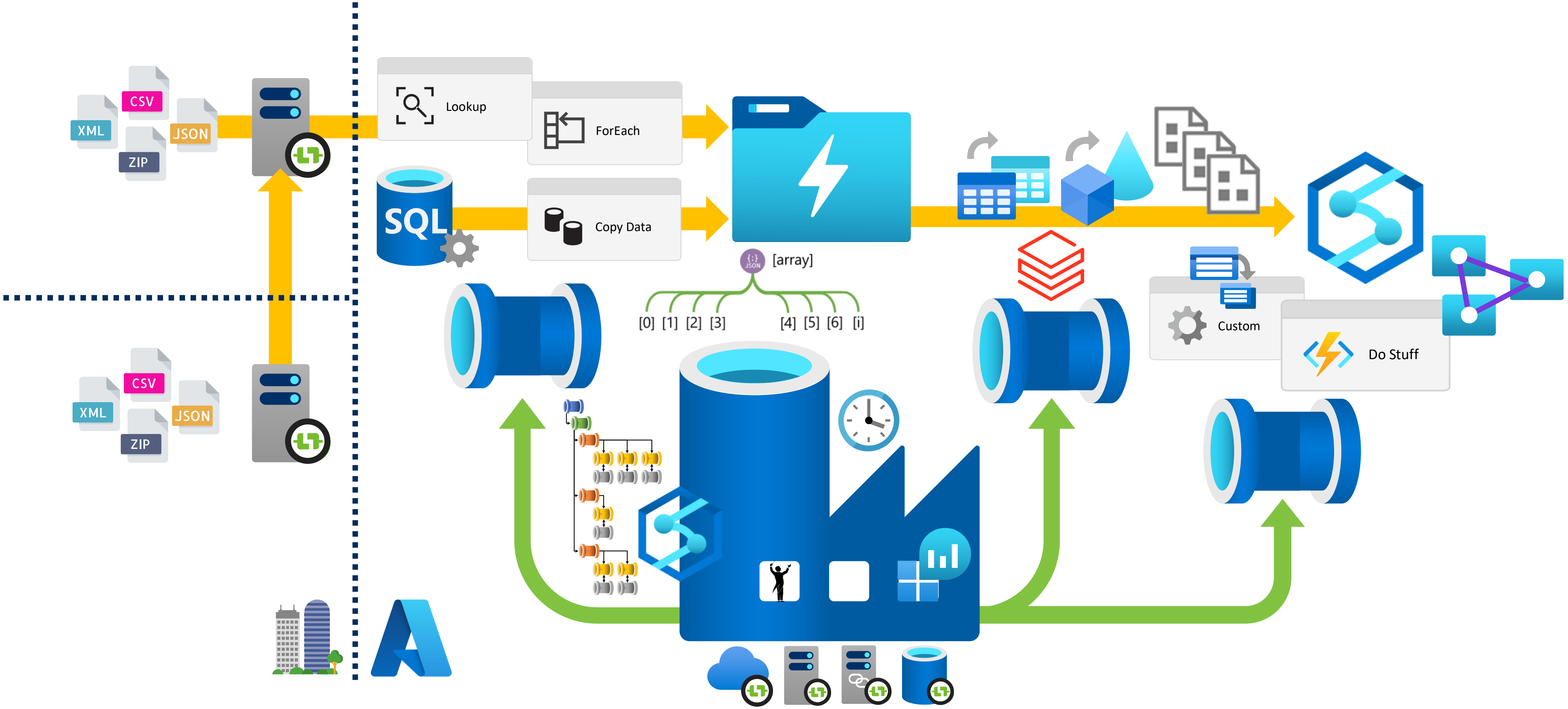
Conclusions



What is Azure Data Factory (ADF)?



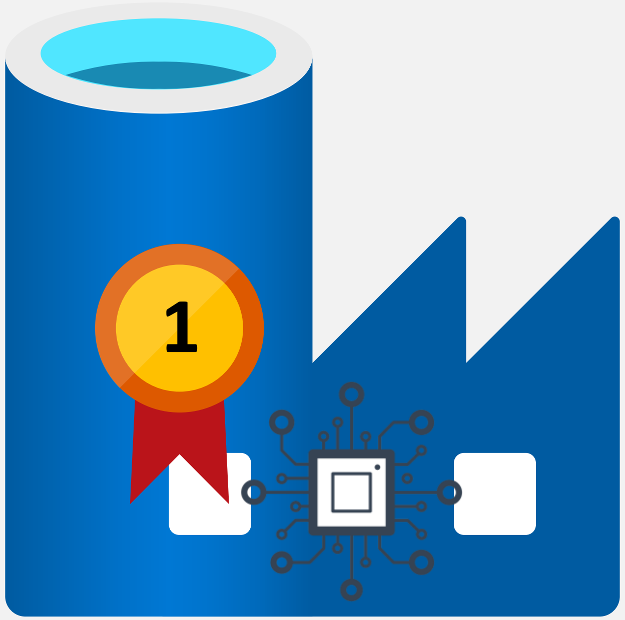
What are Azure ~~Data Factory~~ Integration Pipelines?



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

What Next?

Best Practices for Implementing Azure Data Factory



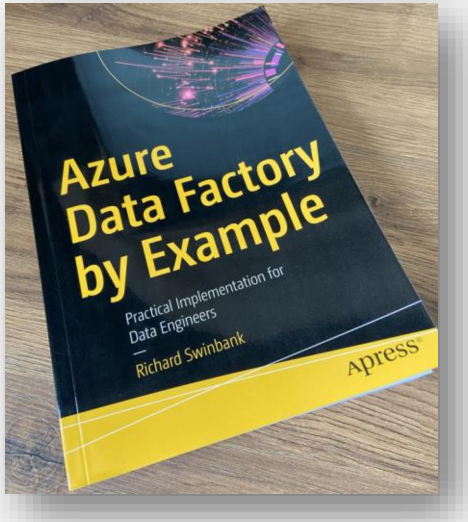
- Environment Setup
- Multiple Data Factory Instance’s
- Deployments
- Automated Testing
- Naming Conventions
- Pipeline Hierarchies
- Pipeline & Activity Descriptions
- Annotations
- Factory Component Folders
- Linked Service Security via Azure Key Vault
- Security Custom Roles
- Dynamic Linked Services
- Generic Datasets
- Metadata Driven Processing
- Parallel Execution
- Hosted Integration Runtimes
- Azure Integration Runtimes
- Wider Platform Orchestration
- Custom Error Handler Paths
- Monitoring via Log Analytics
- Timeouts & Retry
- Service Limitations
- Using Templates
- Documentation

Best Practices for Implementing ADF

https://mrpaulandrew.com/2019/12/18/best-practices-for-implementing-azure-data-factory/

What Next?

Azure Data Factory by Example



Author: Richard Swinbank @RichardSwinbank

Technical Reviewer: Paul Andrew

ISBN-13978-1484270288

Thank you for listening...

Paul Andrew

MVP Microsoft® Most Valuable Professional

avanade

Blog:	mrpaulandrew.com	
YouTube:	c/mrpaulandrew	
Email:	paul@mrpaulandrew.com	
Twitter:	@mrpaulandrew	
LinkedIn:	In/mrpaulandrew	
GitHub:	github.com/mrpaulandrew	/CommunityEvents /ContentCollateral

Robert Walters Technology

ROBERT WALTERS

‘Lets Talk-Data Engineering’

Thanks for joining us for the session!

ROBERT WALTERS

Robert Walters Technology

‘Lets Talk-Data Engineering’

Daniel Bone

Recruitment Consultant with 3 years experience across IT / BI / Data

Founder of the ‘Lets Talk – Data Engineering’ group

Email: Daniel.Bone@robertwalters.com
Phone number: 07766850780
LinkedIn: https://www.linkedin.com/in/daniel-bone-01a3b4199/