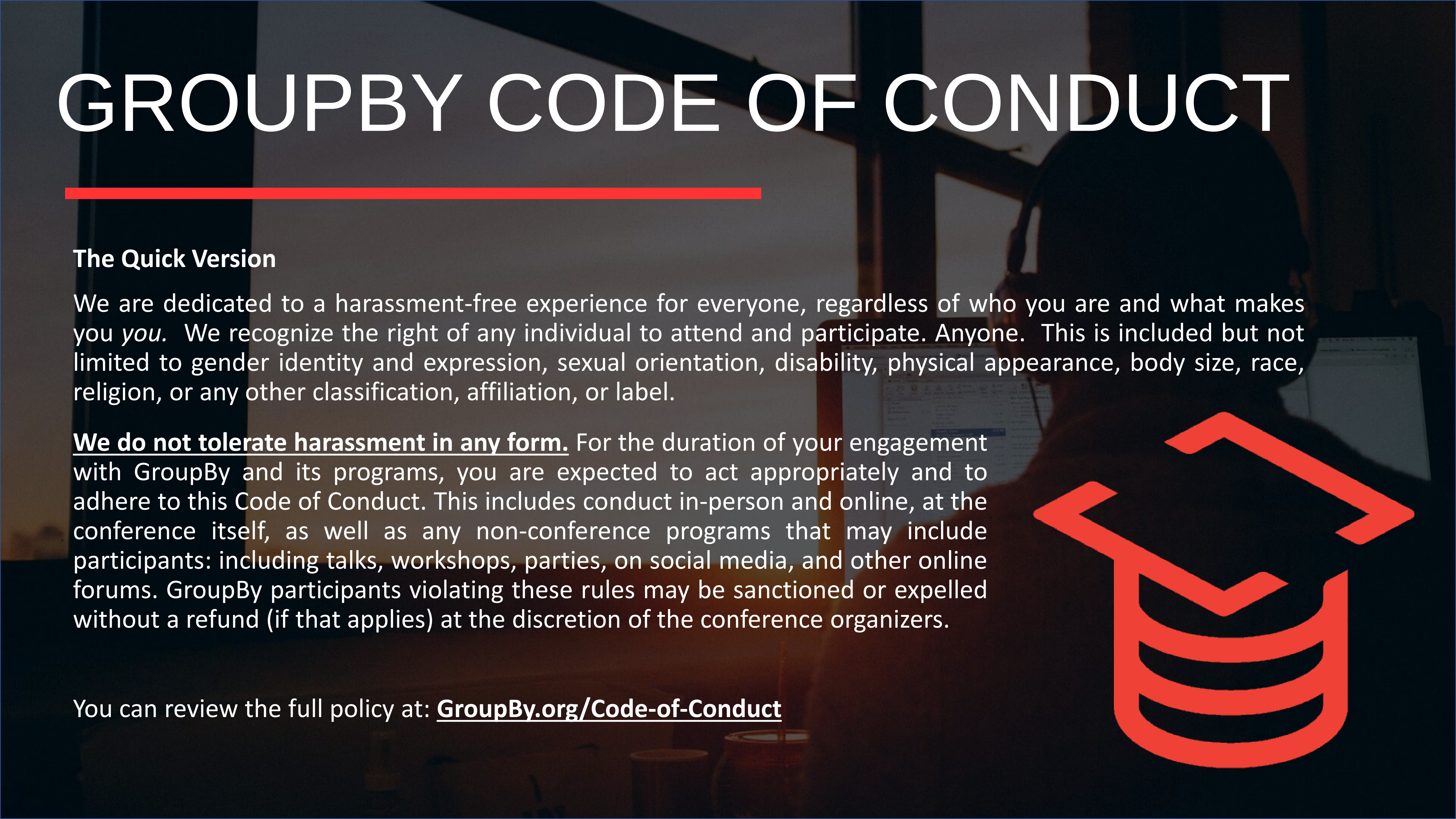
GROUPBY 2020 | OCT 27-28

Free Online Training for Data Professionals.

By the Community, for the Community.



GROUPBY CODE OF CONDUCT

The Quick Version

We are dedicated to a harassment-free experience for everyone, regardless of who you are and what makes you *you*. We recognize the right of any individual to attend and participate. Anyone. This is included but not limited to gender identity and expression, sexual orientation, disability, physical appearance, body size, race, religion, or any other classification, affiliation, or label.

We do not tolerate harassment in any form. For the duration of your engagement with GroupBy and its programs, you are expected to act appropriately and to adhere to this Code of Conduct. This includes conduct in-person and online, at the conference itself, as well as any non-conference programs that may include participants: including talks, workshops, parties, on social media, and other online forums. GroupBy participants violating these rules may be sanctioned or expelled without a refund (if that applies) at the discretion of the conference organizers.

You can review the full policy at: **GroupBy.org/Code-of-Conduct**

A Complete Introduction to Azure Data Factory

Paul Andrew | Principal Consultant & Solution Architect

Microsoft® MVP Most Valuable Professional

altius

@MrPaulAndrew

In/MrPaulAndrew

MrPaulAndrew.com

https://github.com/mrpaulandrew

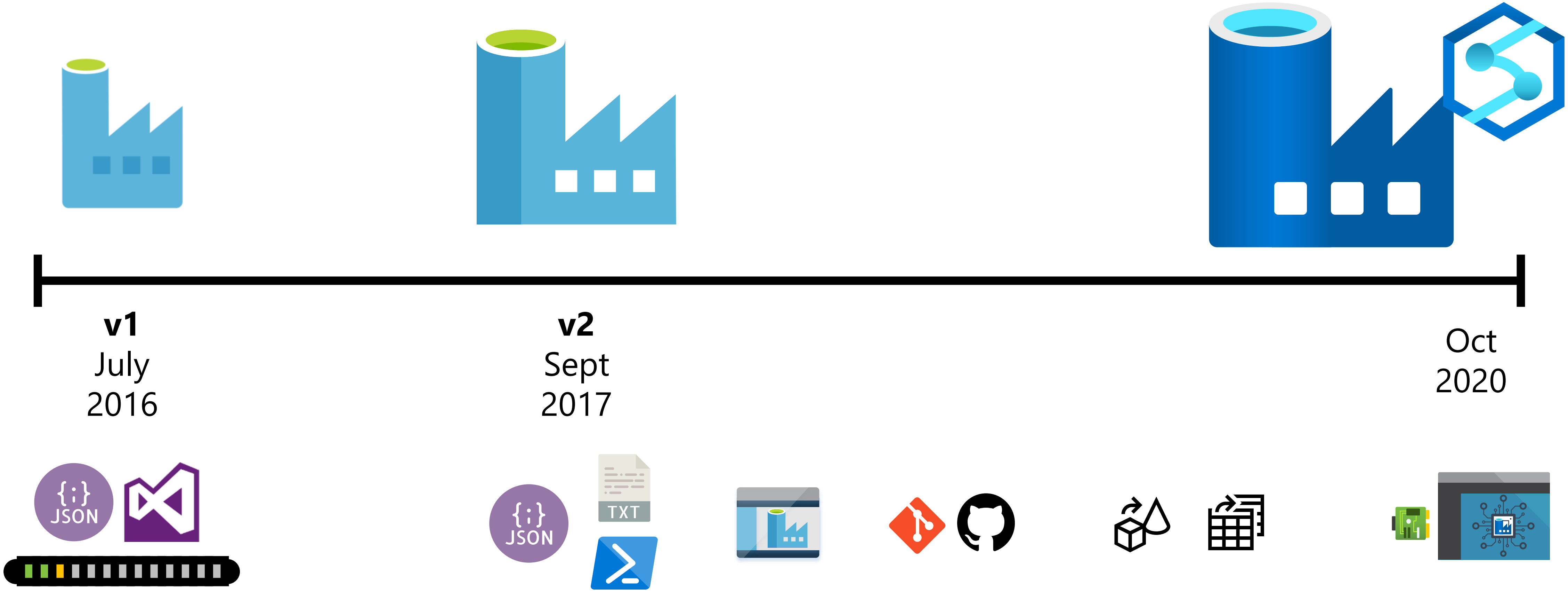
CommunityEvents

Agenda

- What is it and why use it?
- Data Factory Components
- Common Activities
- Execution Dependencies
- Integration Runtimes
- Running SSIS Packages in Azure
- Data Factory Data Flows
- Source Code & ARM 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

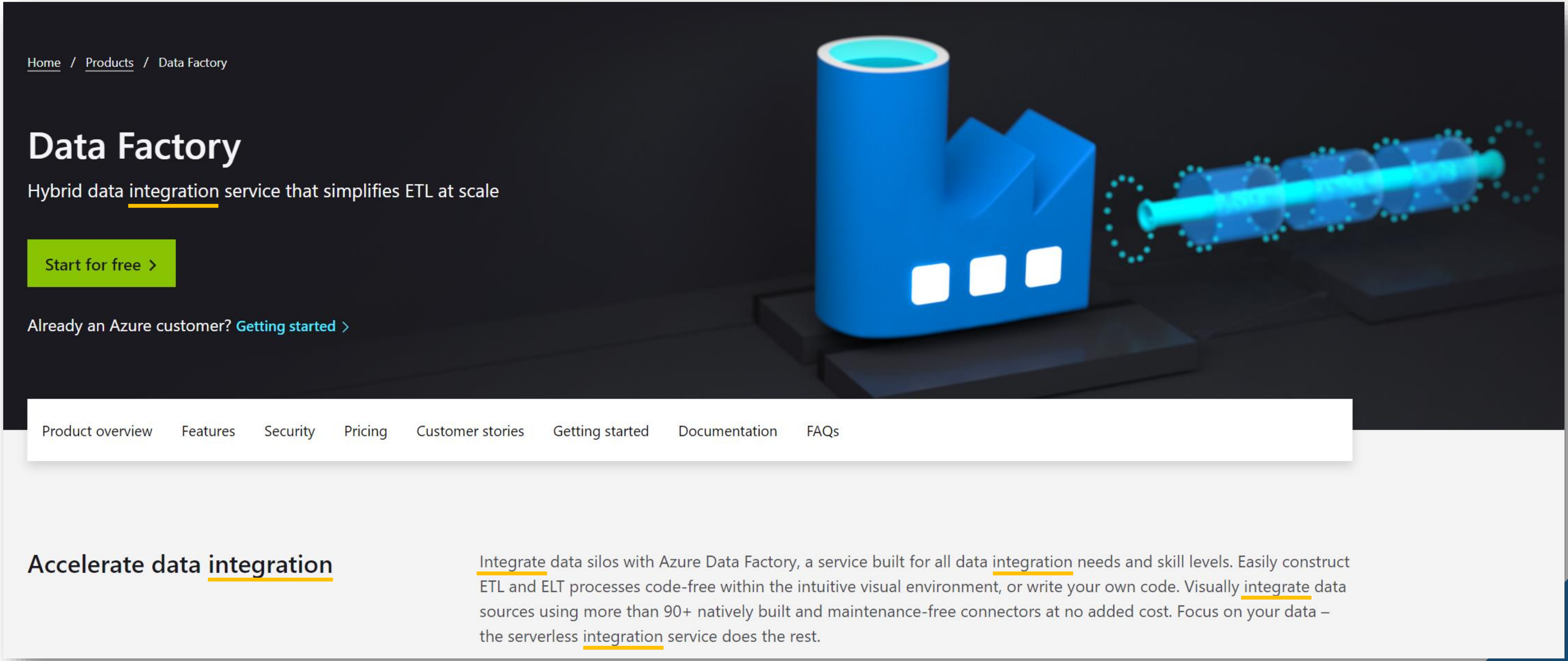
New UI

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 >

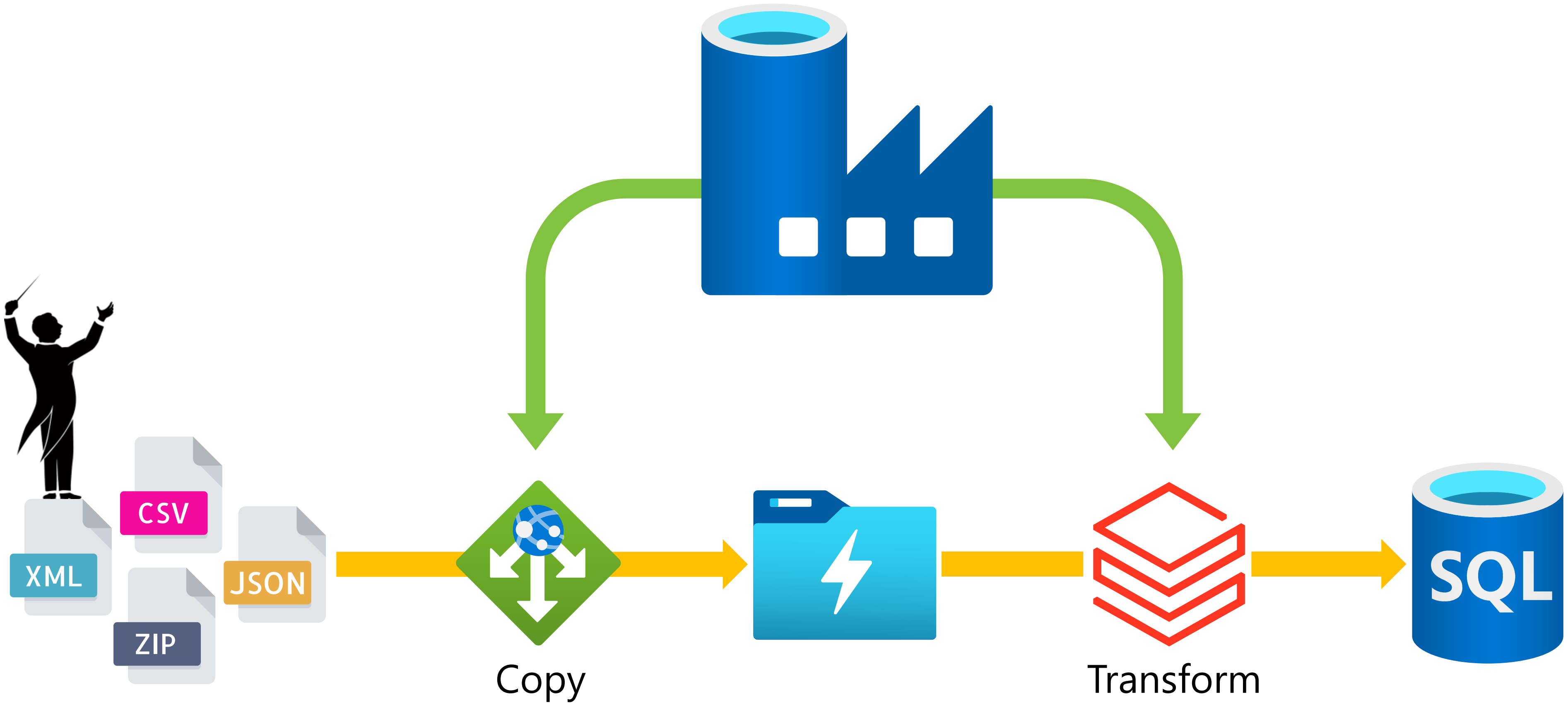
Already an Azure customer? Getting started >

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.

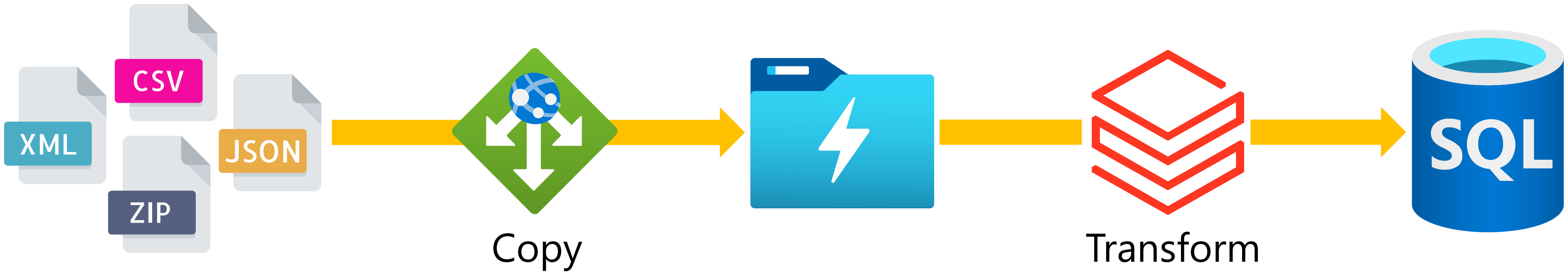
https://azure.microsoft.com/en-gb/services/data-factory

What is Azure Data Factory (ADF)?

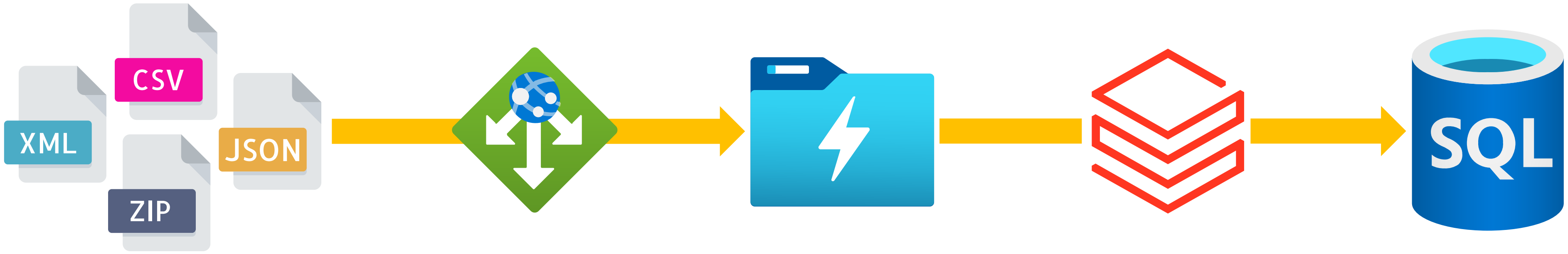


Data Factory Components

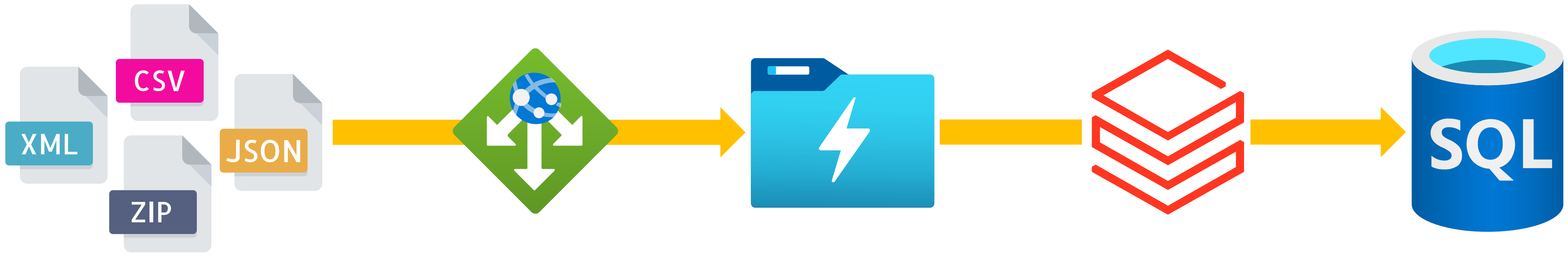
Data Factory Components



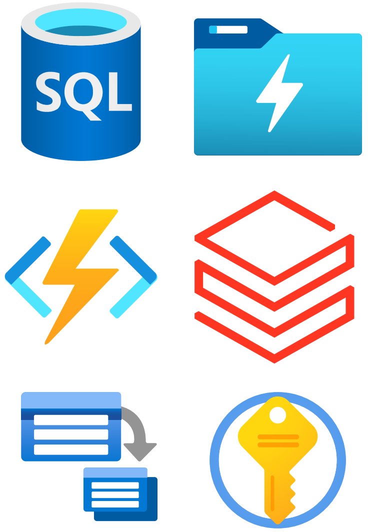
Data Factory Components



Data Factory Components



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

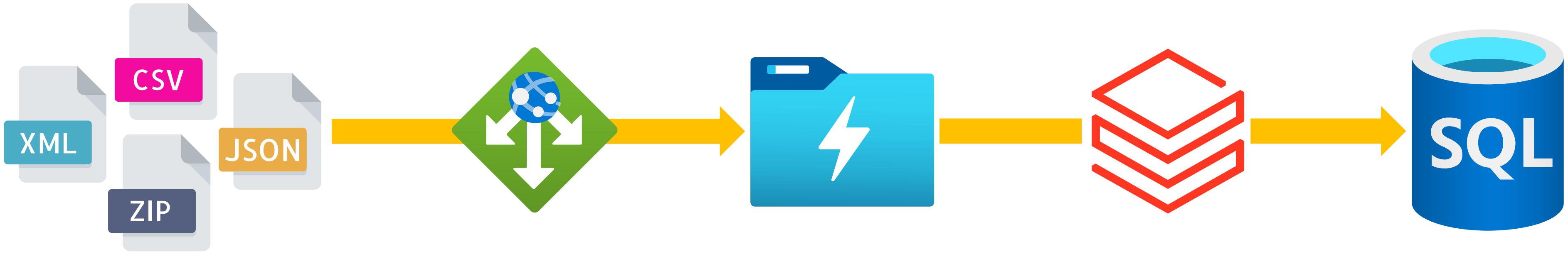


{;} JSON

SQLDBLinkedService

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

Data Factory Components



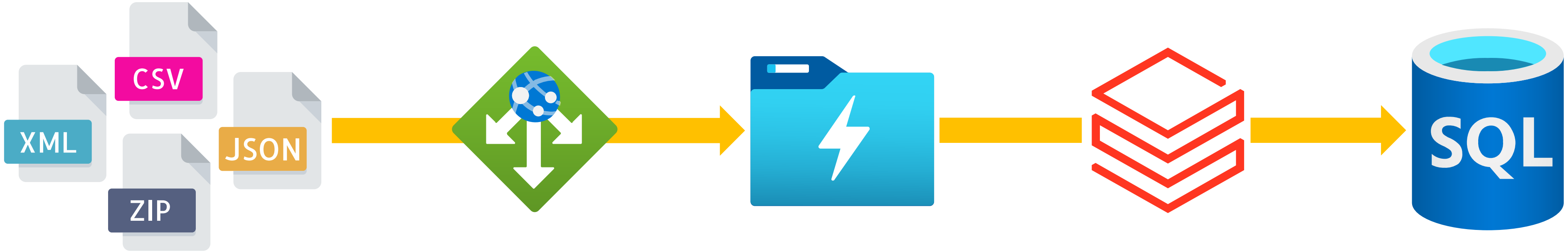
1 **Linked Services**

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

[dbo].[SalesOrders]

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

Data Factory Components



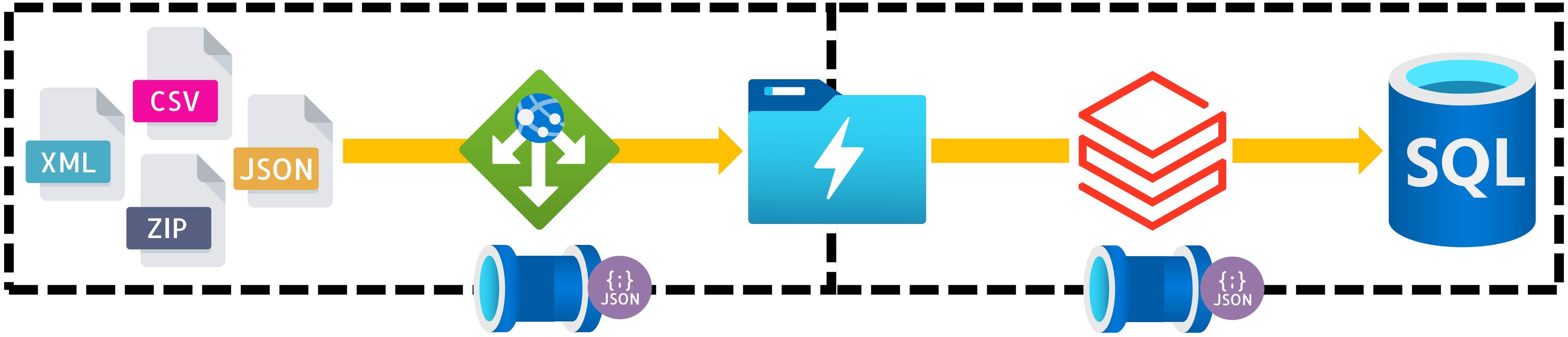
1. **Linked Services**
2. **Datasets**
3. **Activities** – What do we want to happen when we invoke a Linked Service? With what conditions?

{:} JSON

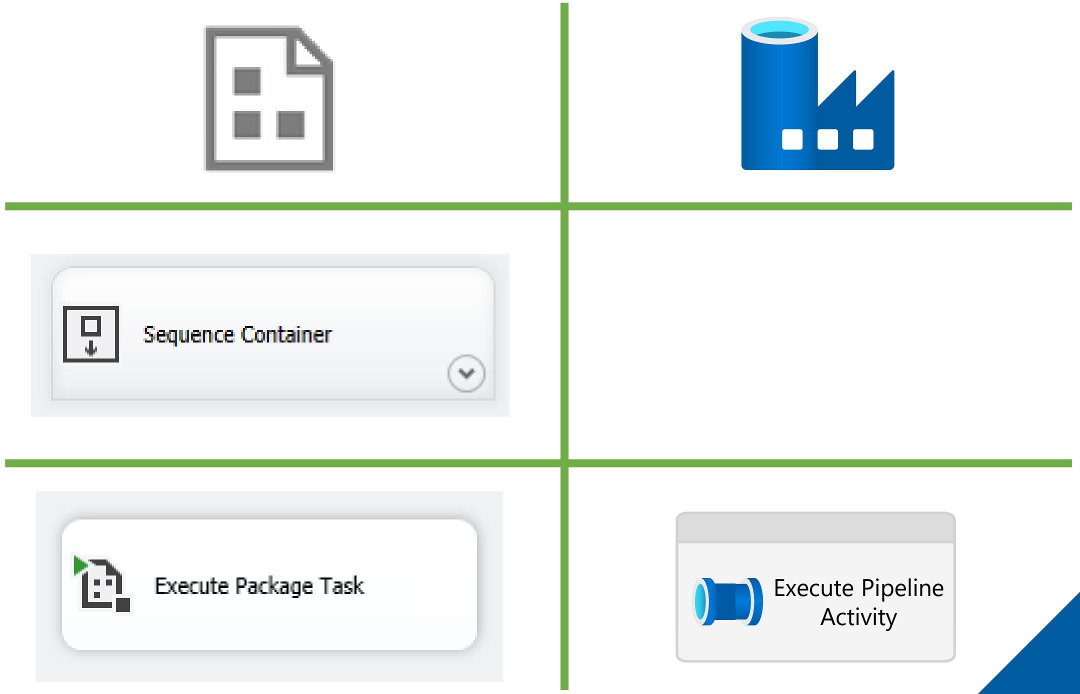
Databricks Notebook Activity

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

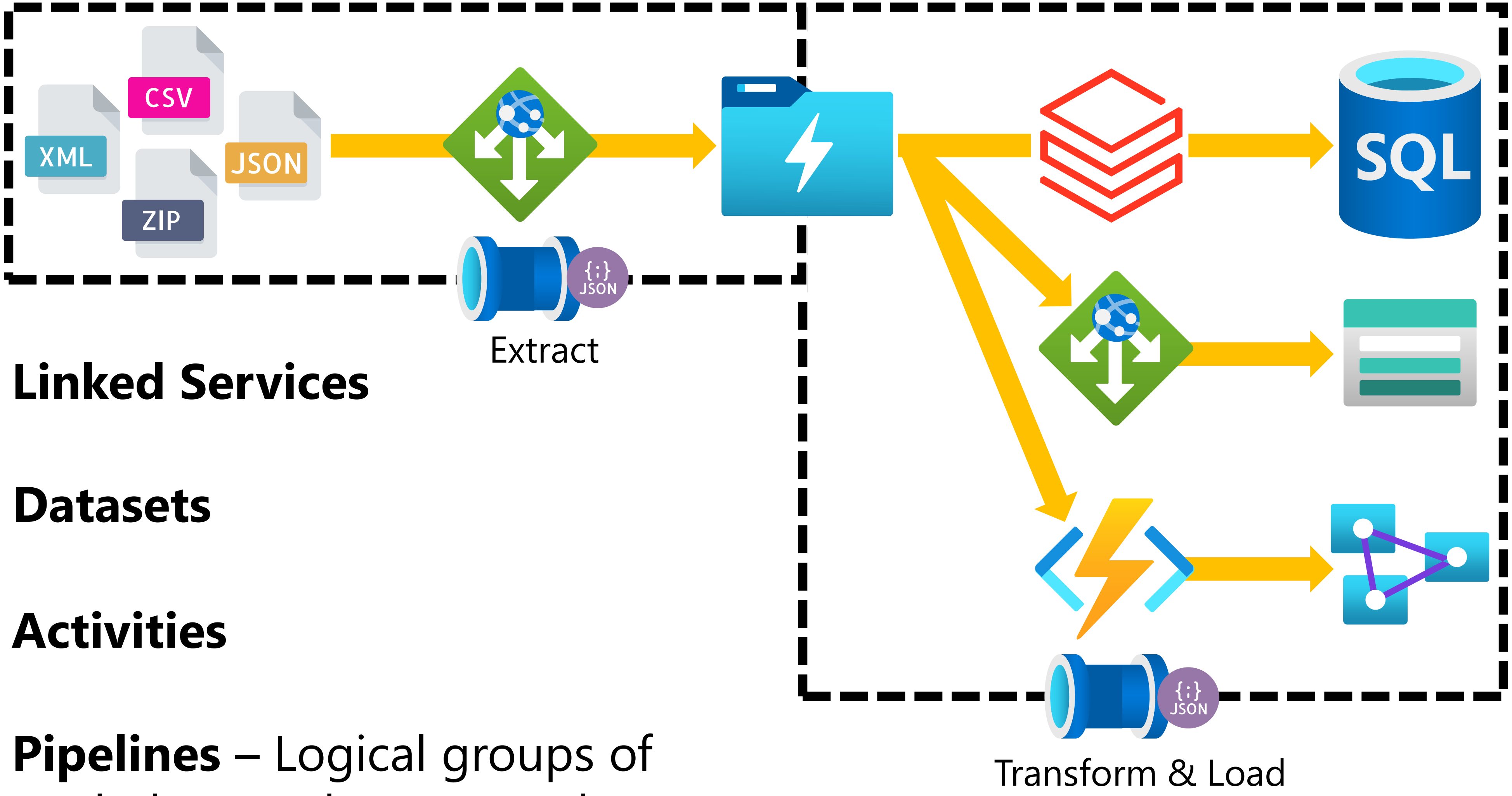
Data Factory Components



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

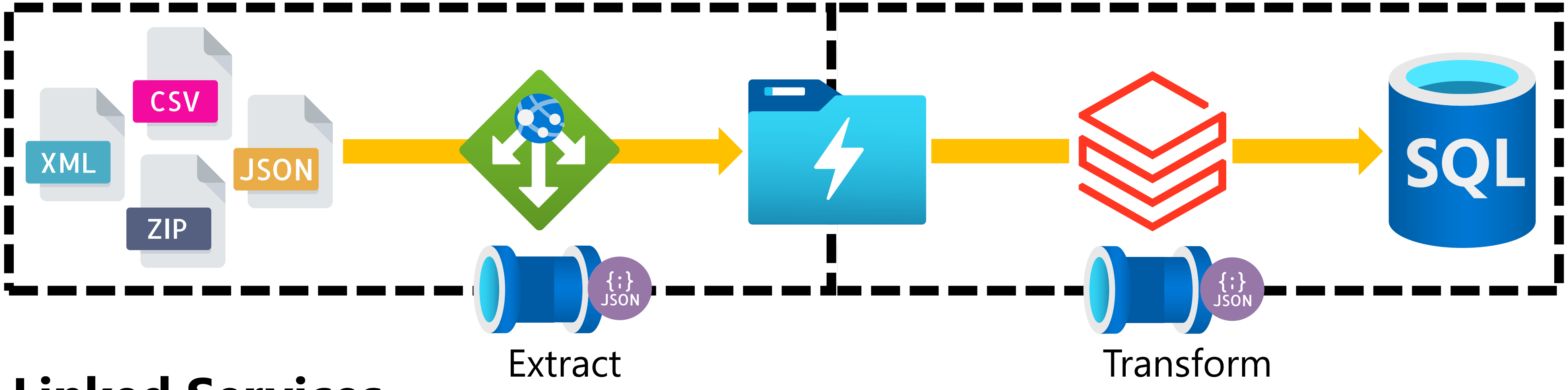


Data Factory Components

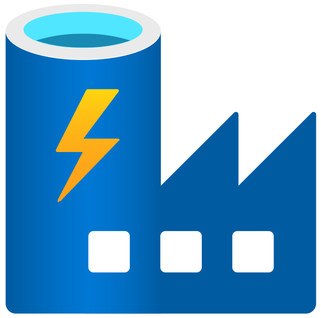


1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines** – Logical groups of work that can be executed.

Data Factory Components

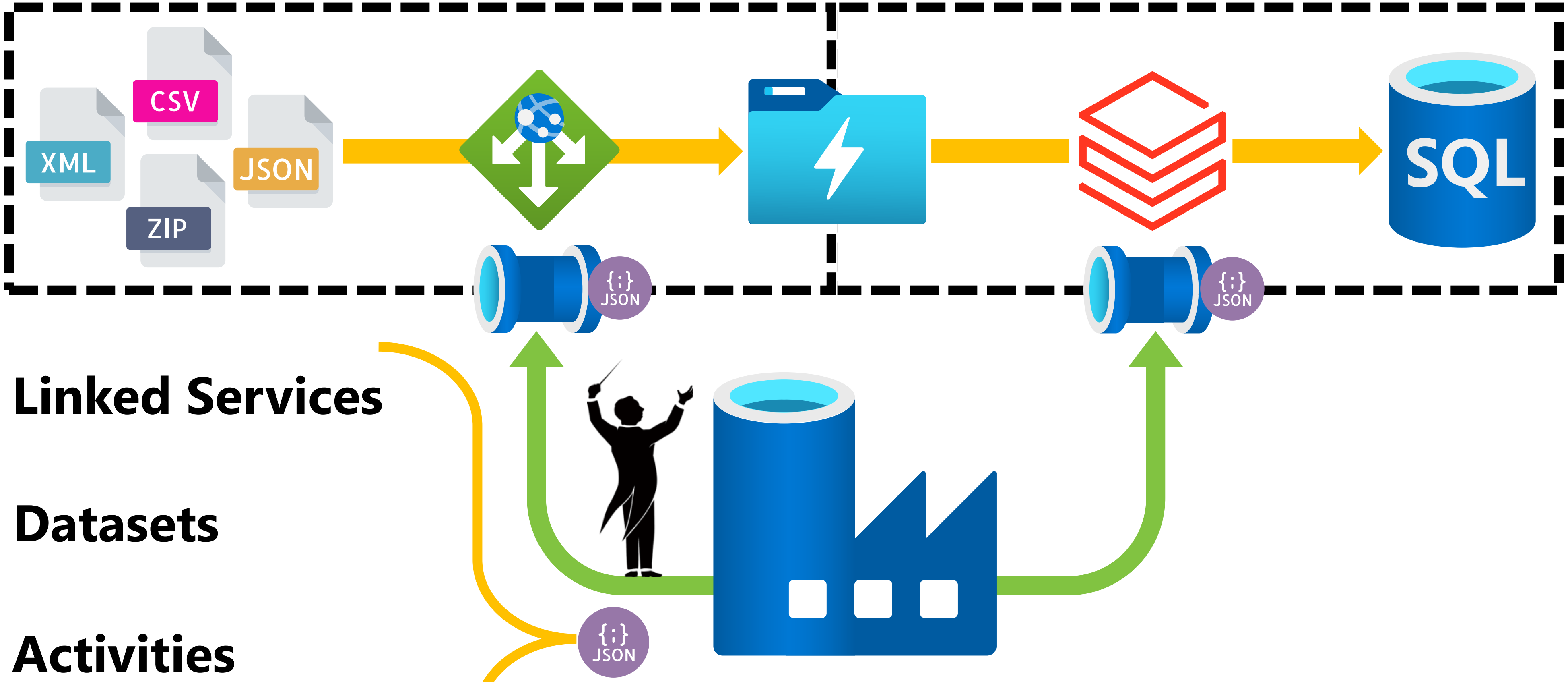


1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines**
5. **Triggers** – Telling our when pipelines to run.



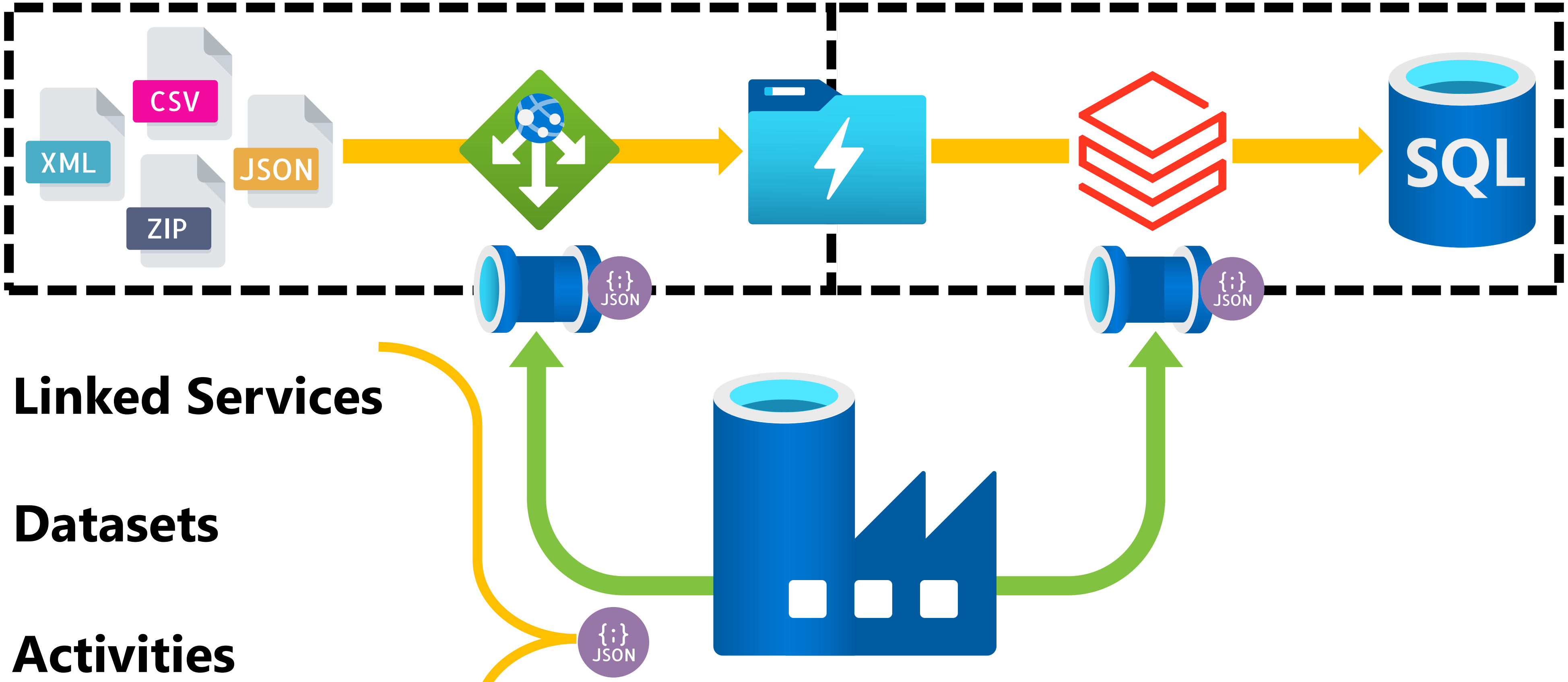
- Manually
- Programmatically
- Schedule
- Tumbling Windows
- File Event

Data Factory Components



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

Data Factory Control Flow Components

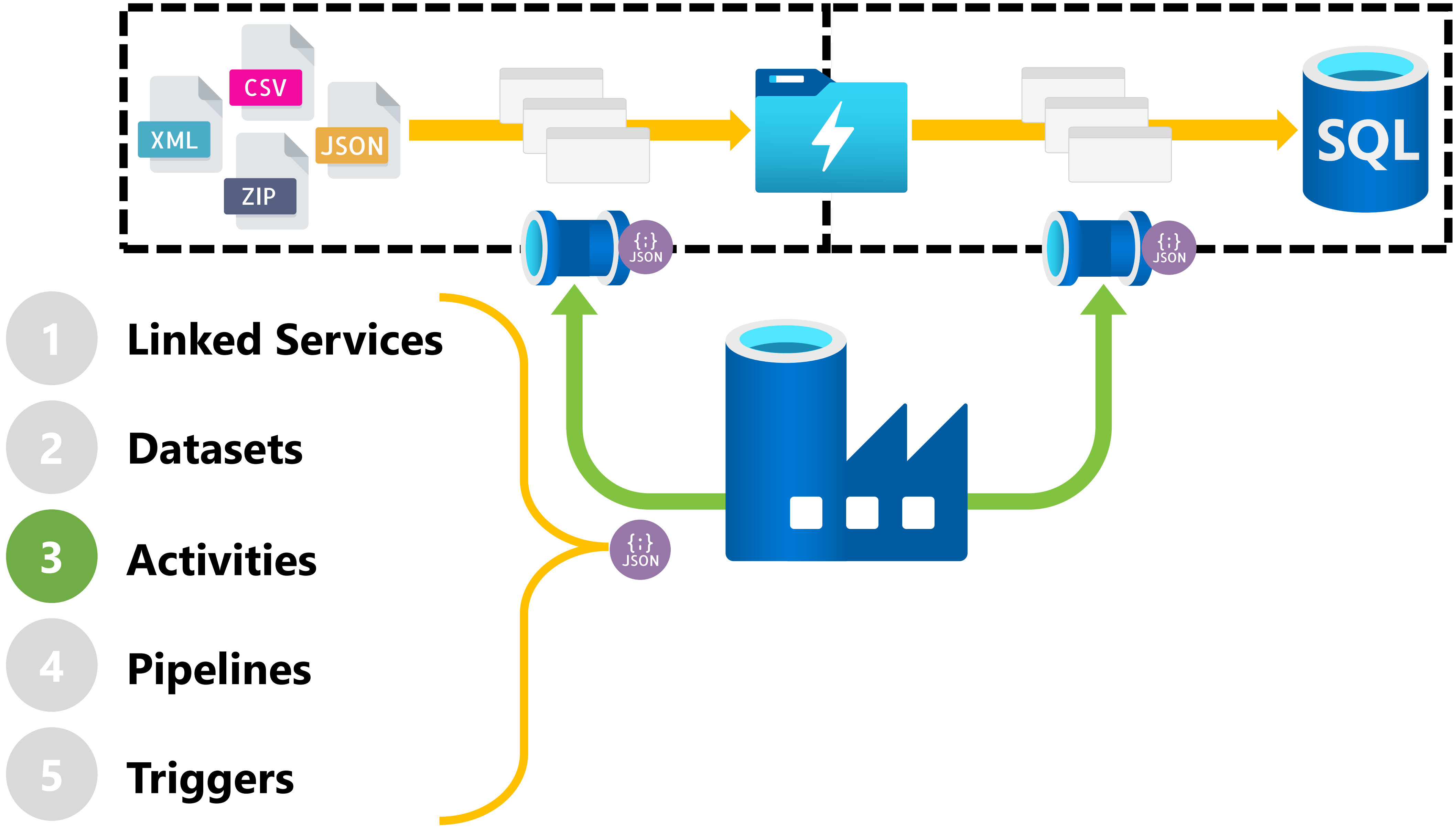


1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

Common Activities

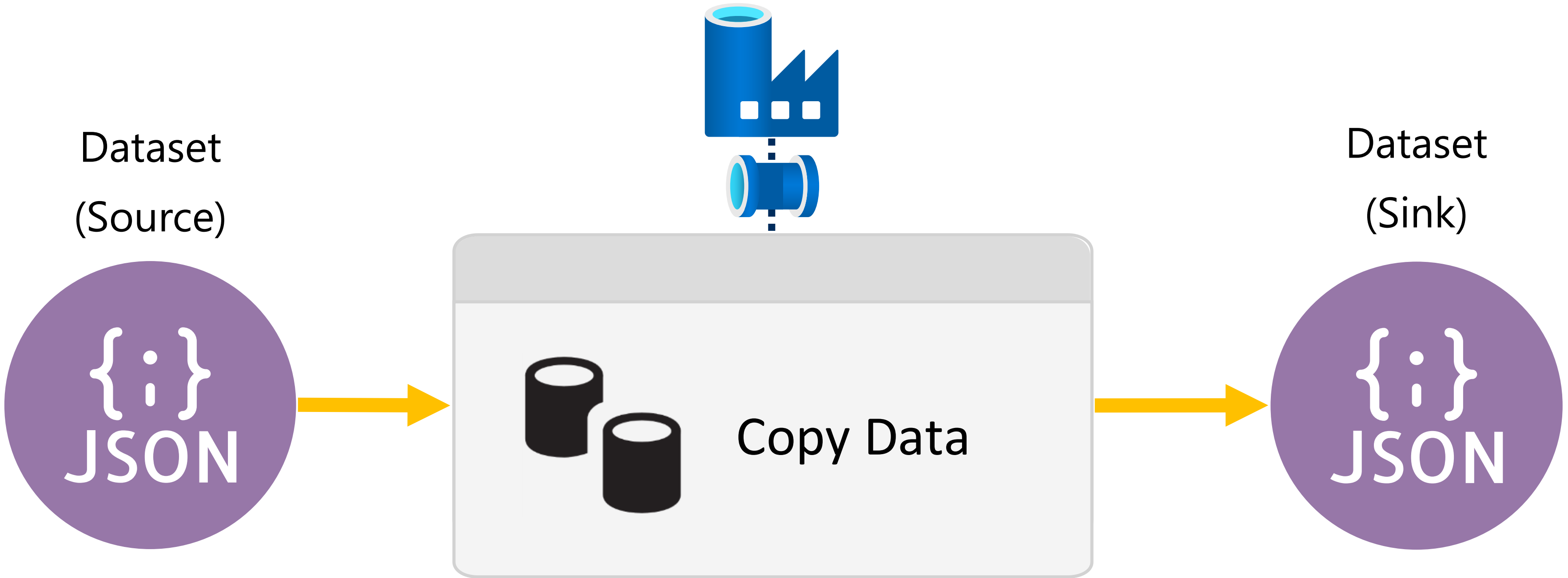
```
SELECT TOP 5
    [ActivityName],
    [Inputs],
    [Outputs],
    [Details]
FROM
    [metadata].[AdfActivities]
WHERE
    [Notes] = 'Pauls Favourites';
```

Data Factory Common Activities



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

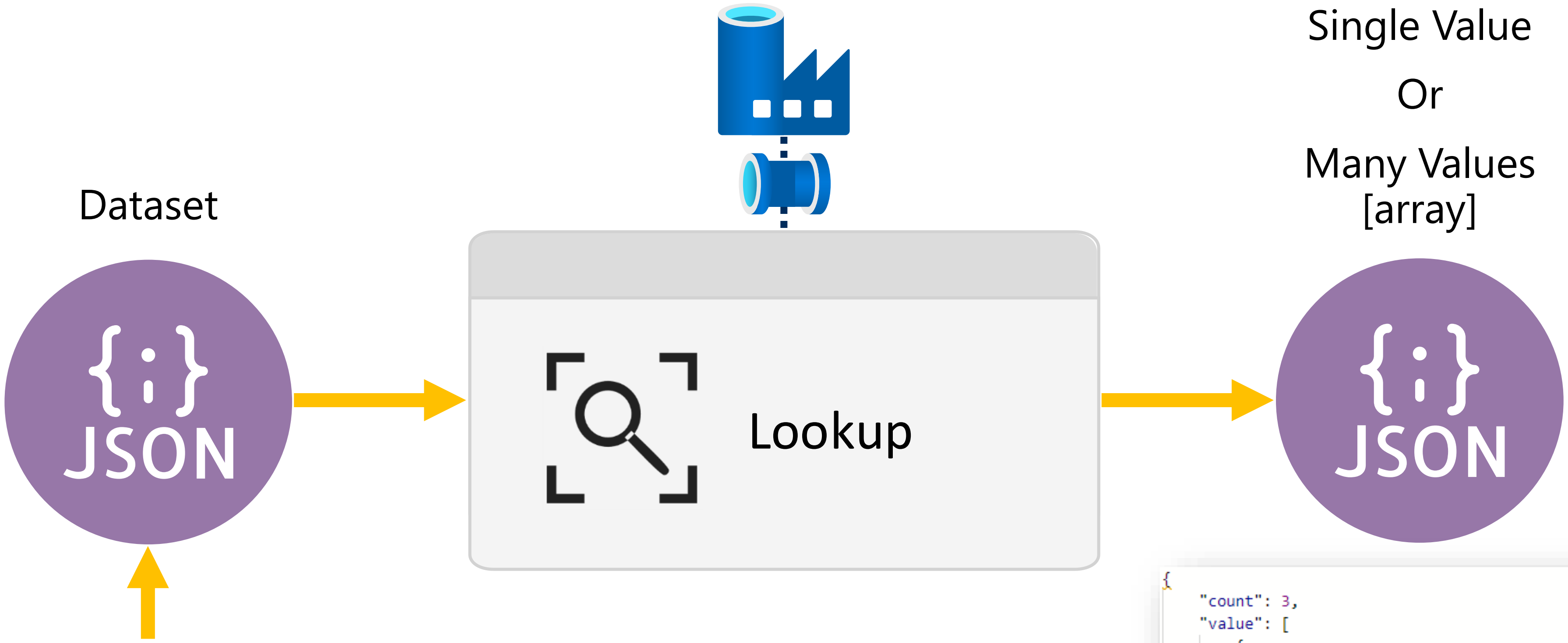
Copy

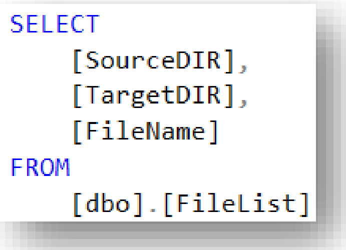


- Auto Scaling
- Transactional Restarts
- Handle Zip Compression
- Attribute Mapping and Schema Drift
- Handle Failed Rows
- Add Custom Attributes
- Parse Excel & JSON Files

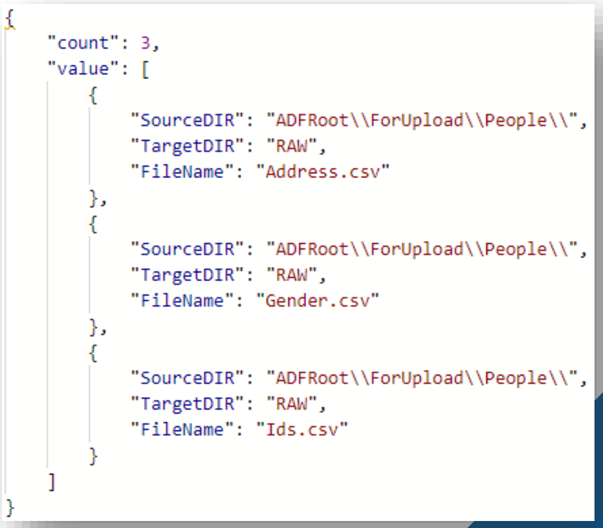
Lookup

Get value to support other control flow activities





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



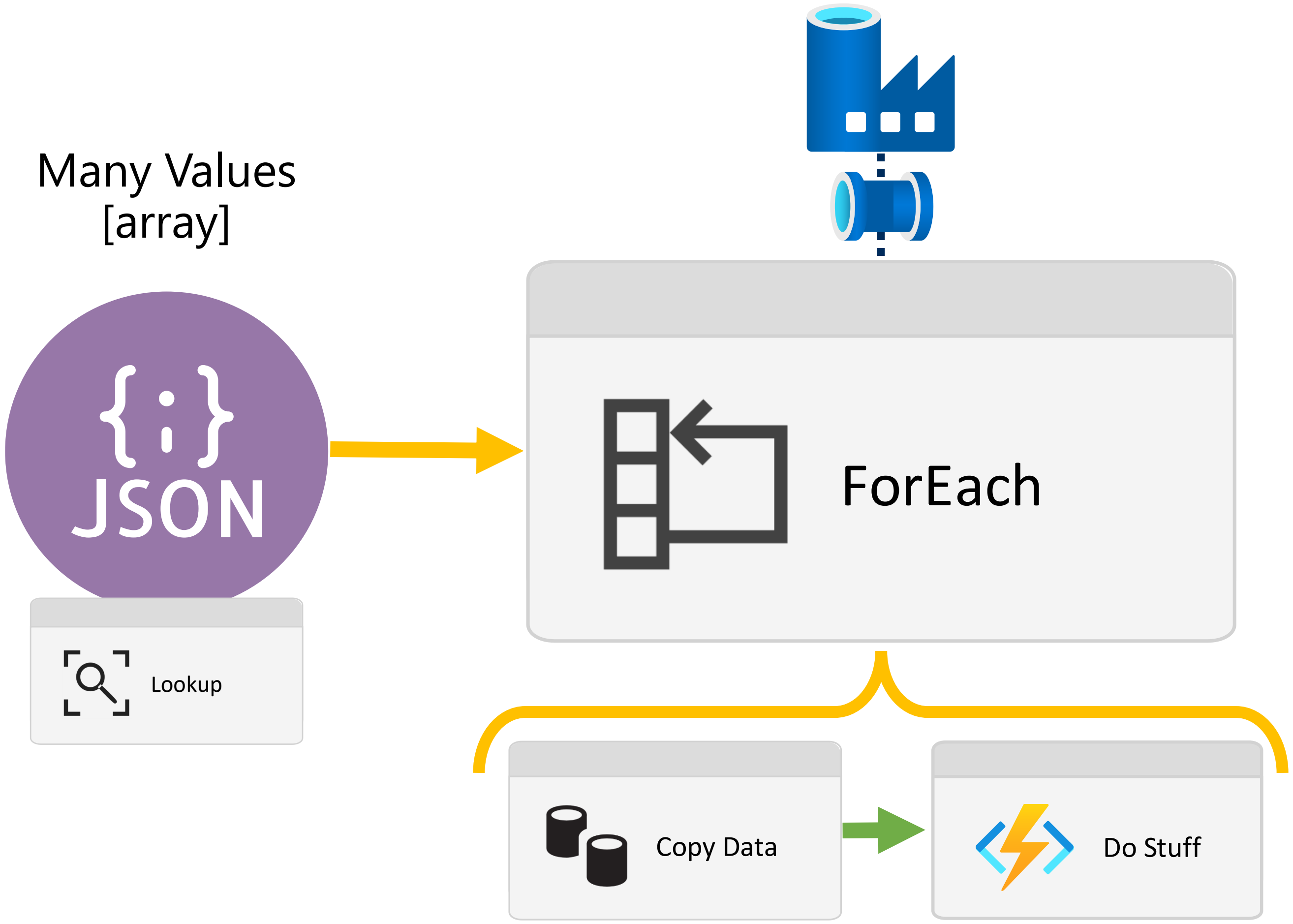
```
{
    "count": 3,
    "value": [
        {
            "SourceDIR": "ADFRoot\\ForUpload\\People\\",
            "TargetDIR": "RAW",
            "FileName": "Address.csv"
        },
        {
            "SourceDIR": "ADFRoot\\ForUpload\\People\\",
            "TargetDIR": "RAW",
            "FileName":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

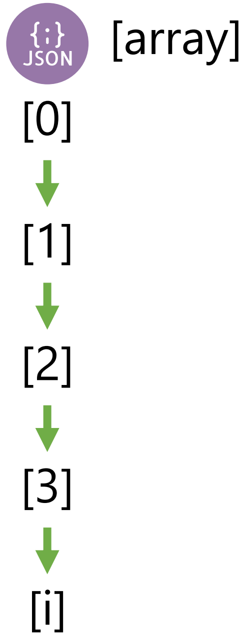
Scaling Out Control Flow Activities

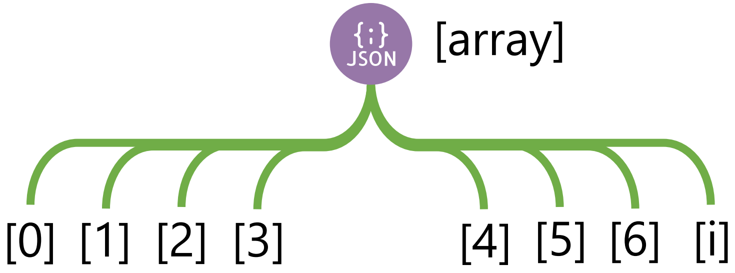
Many Values
[array]



@item().

IsSequential:
true





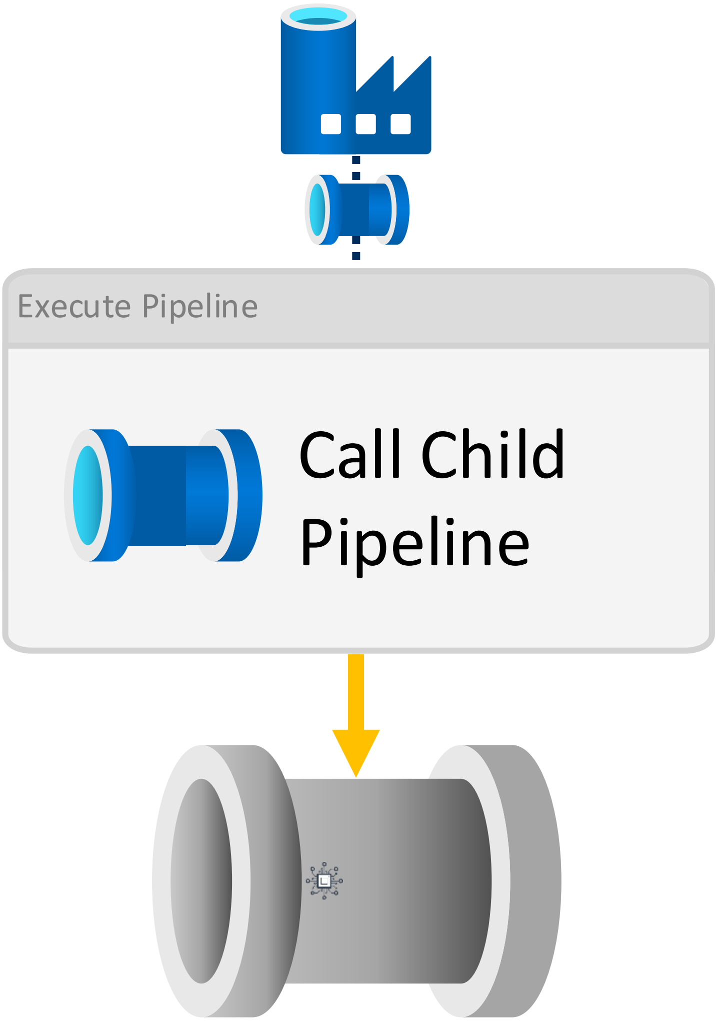
Batch Count Default: 20

Batch Count Max: 50

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

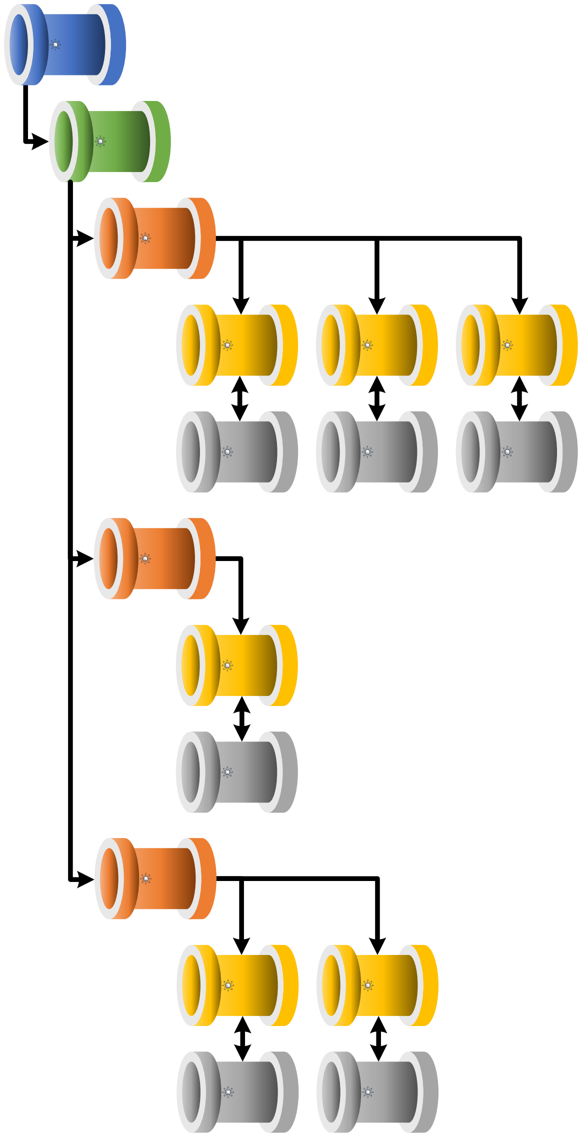
Execute Pipeline

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



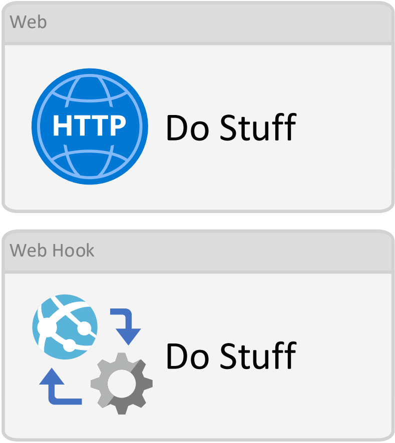
Execute Pipeline

Call Child Pipeline



Azure Function

Extend Data Factory with Rest Calls

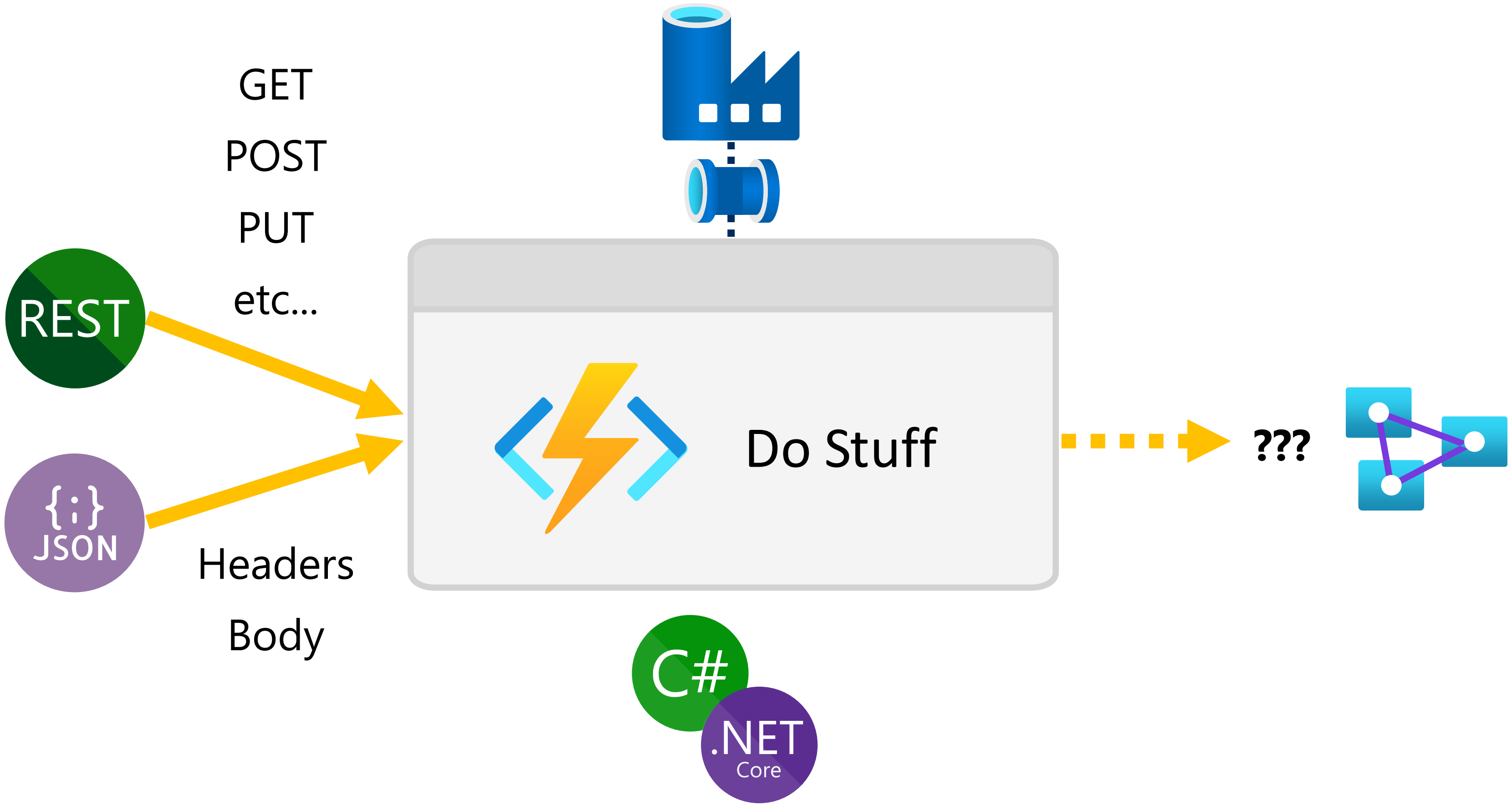


Web

Do Stuff

Web Hook

Do Stuff



GET

POST

PUT

etc…

Headers

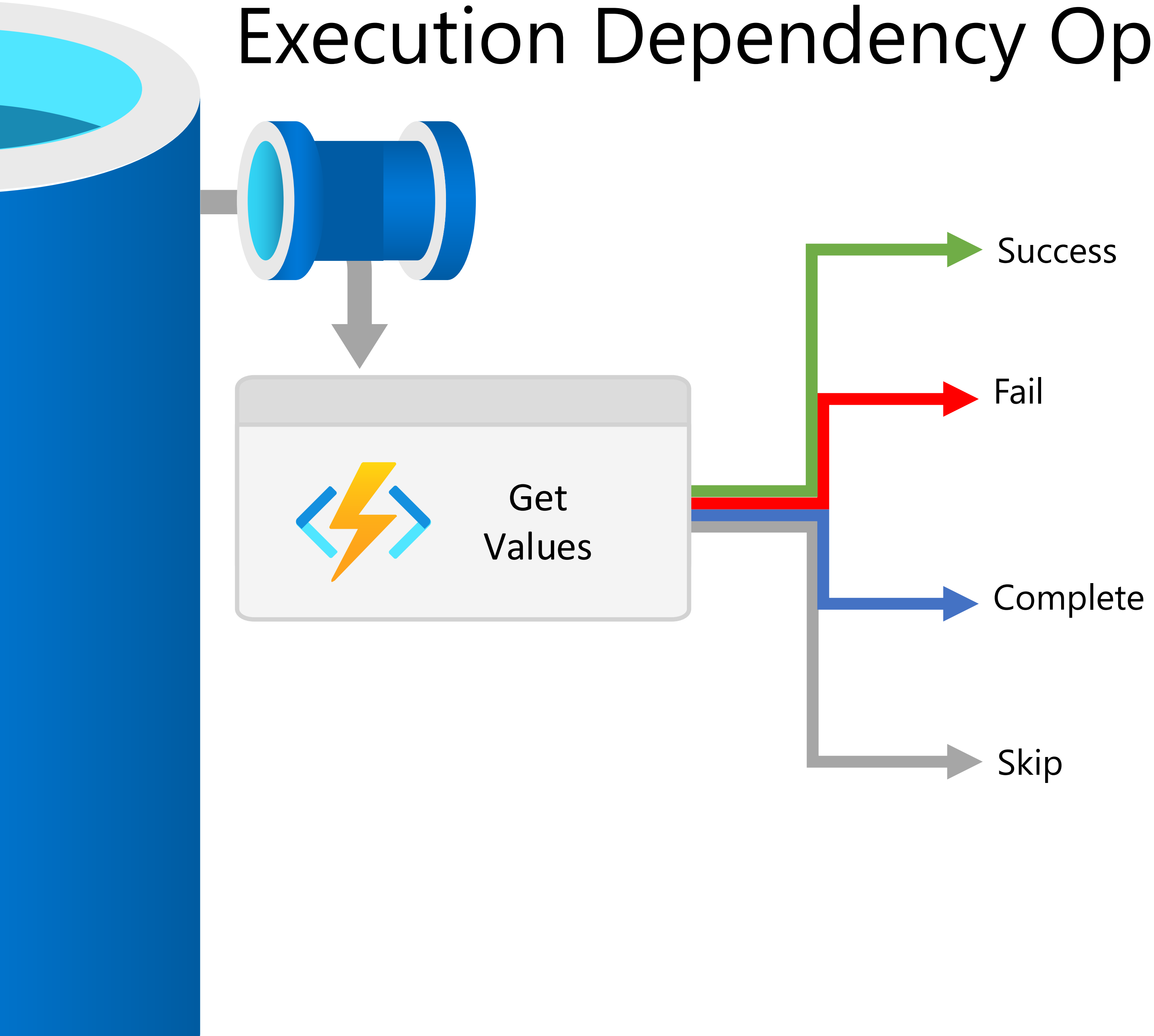
Body

Do Stuff

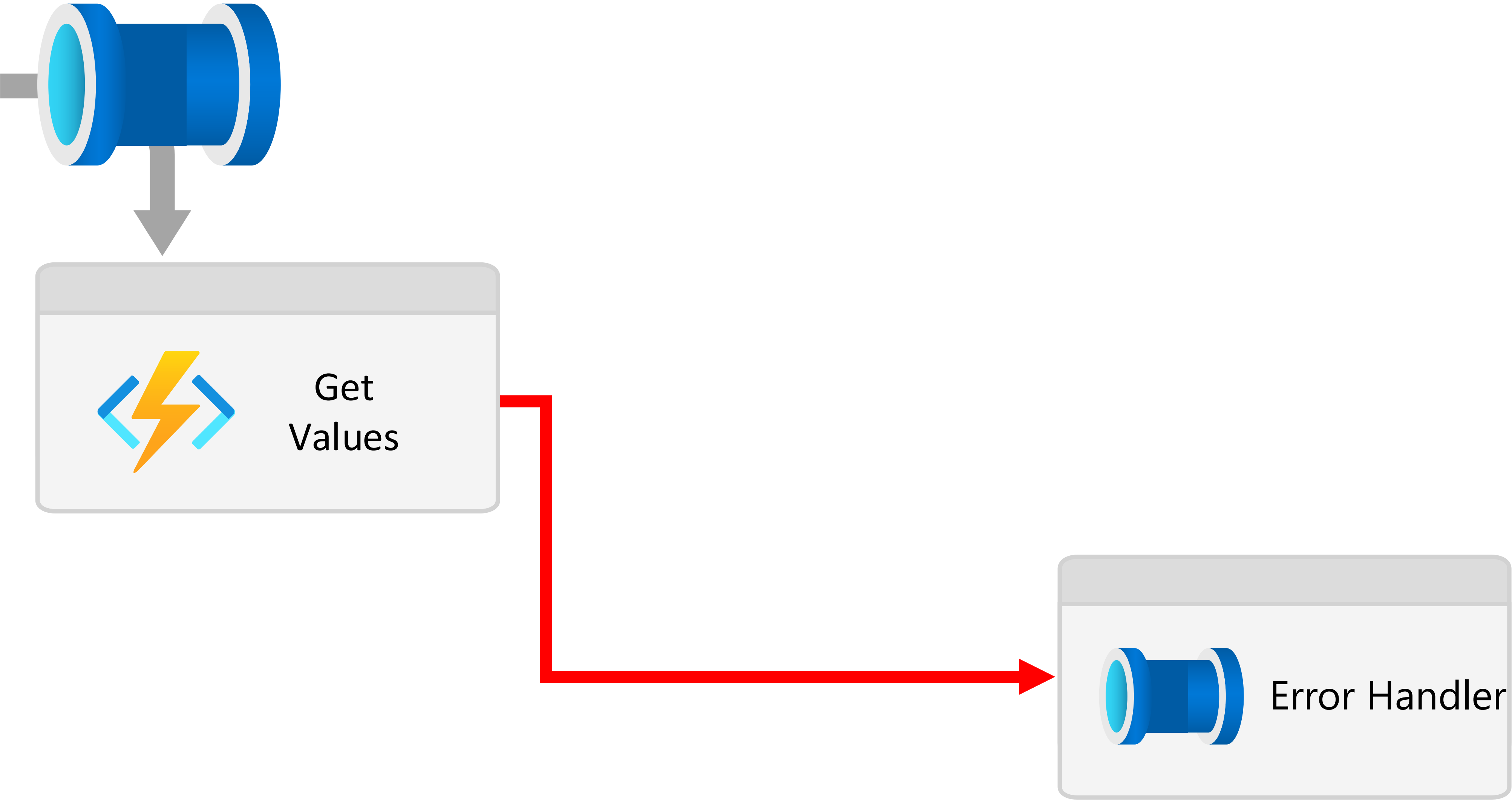
???

Execution Dependencies

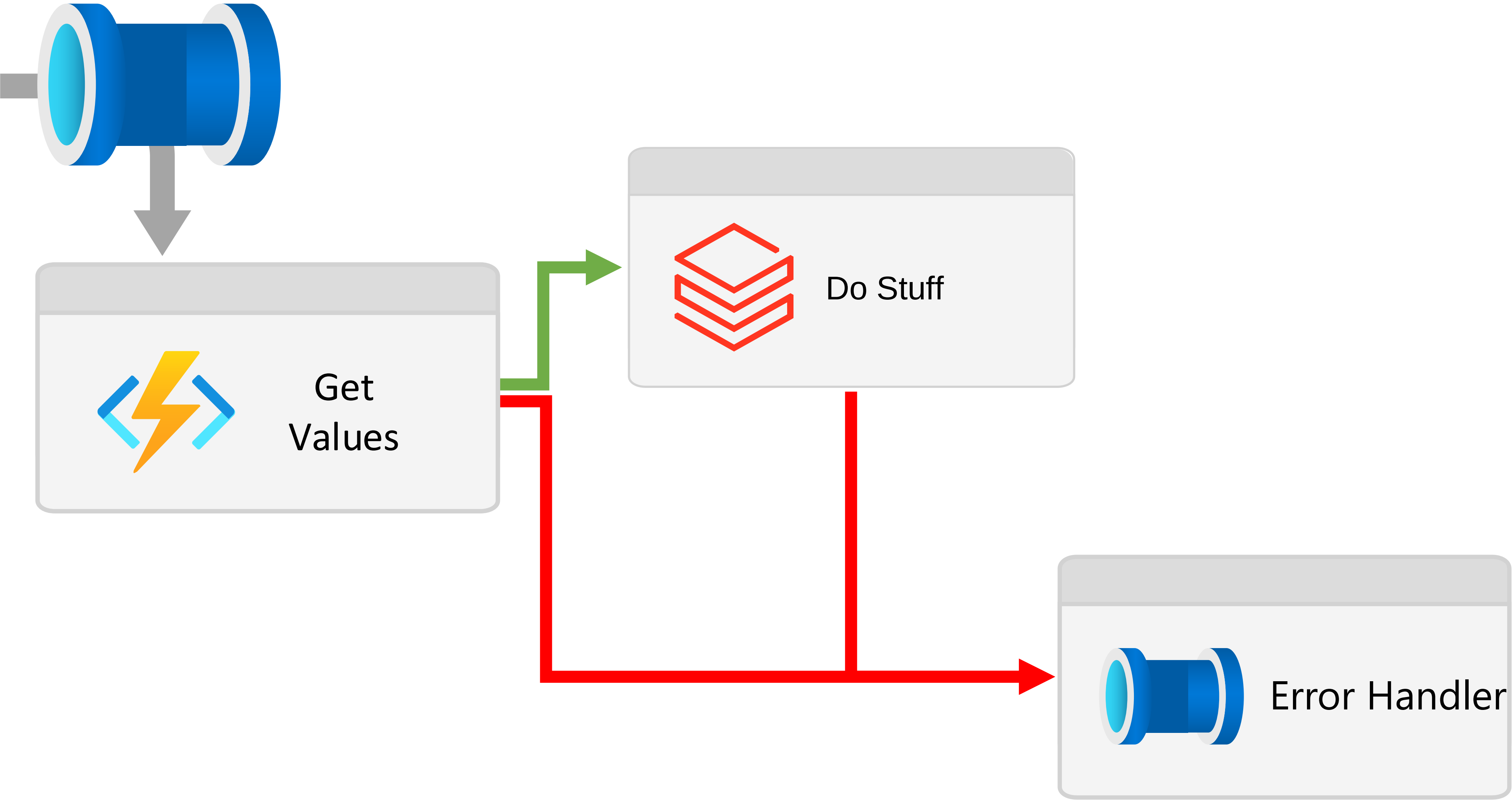
Execution Dependency Options



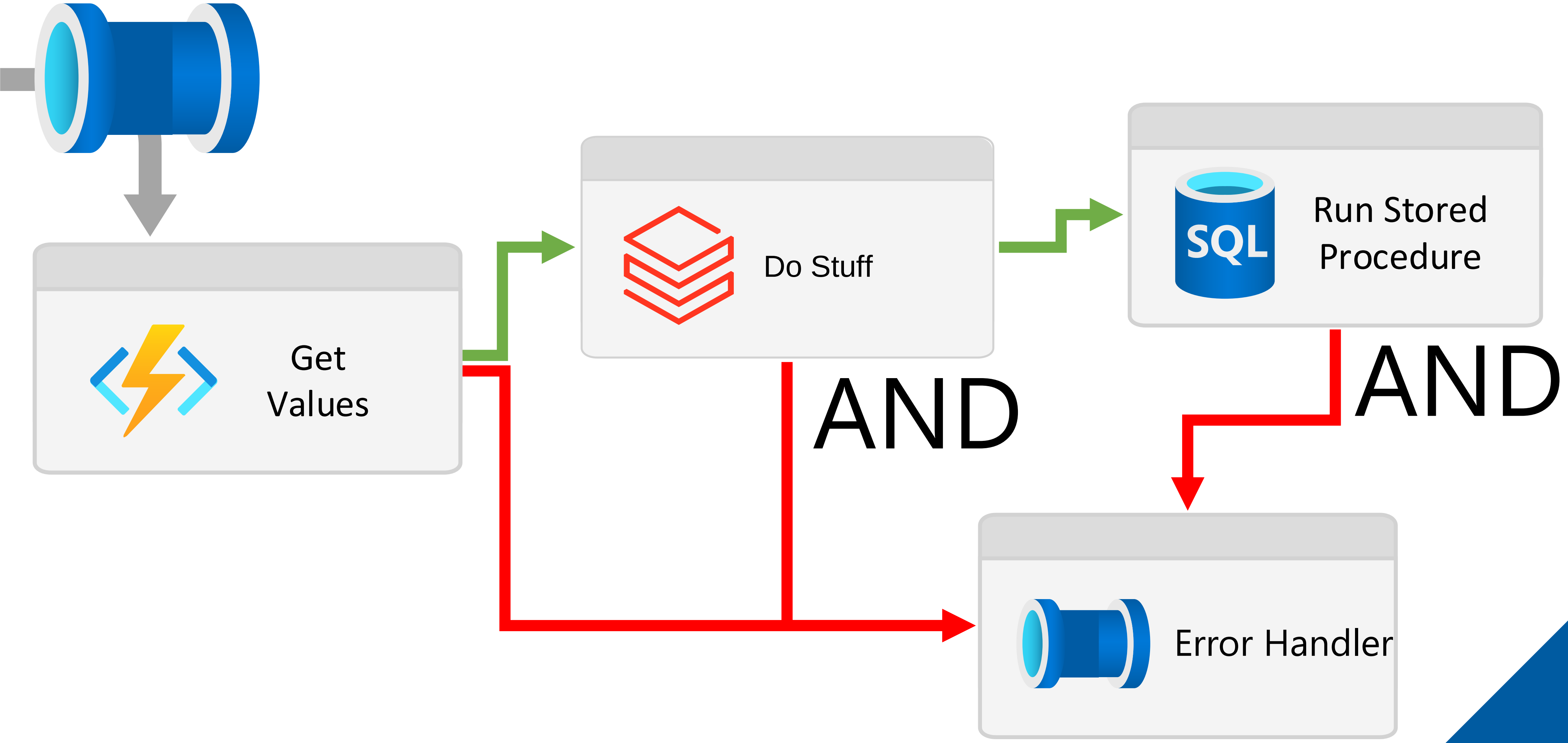
Execution On Failure



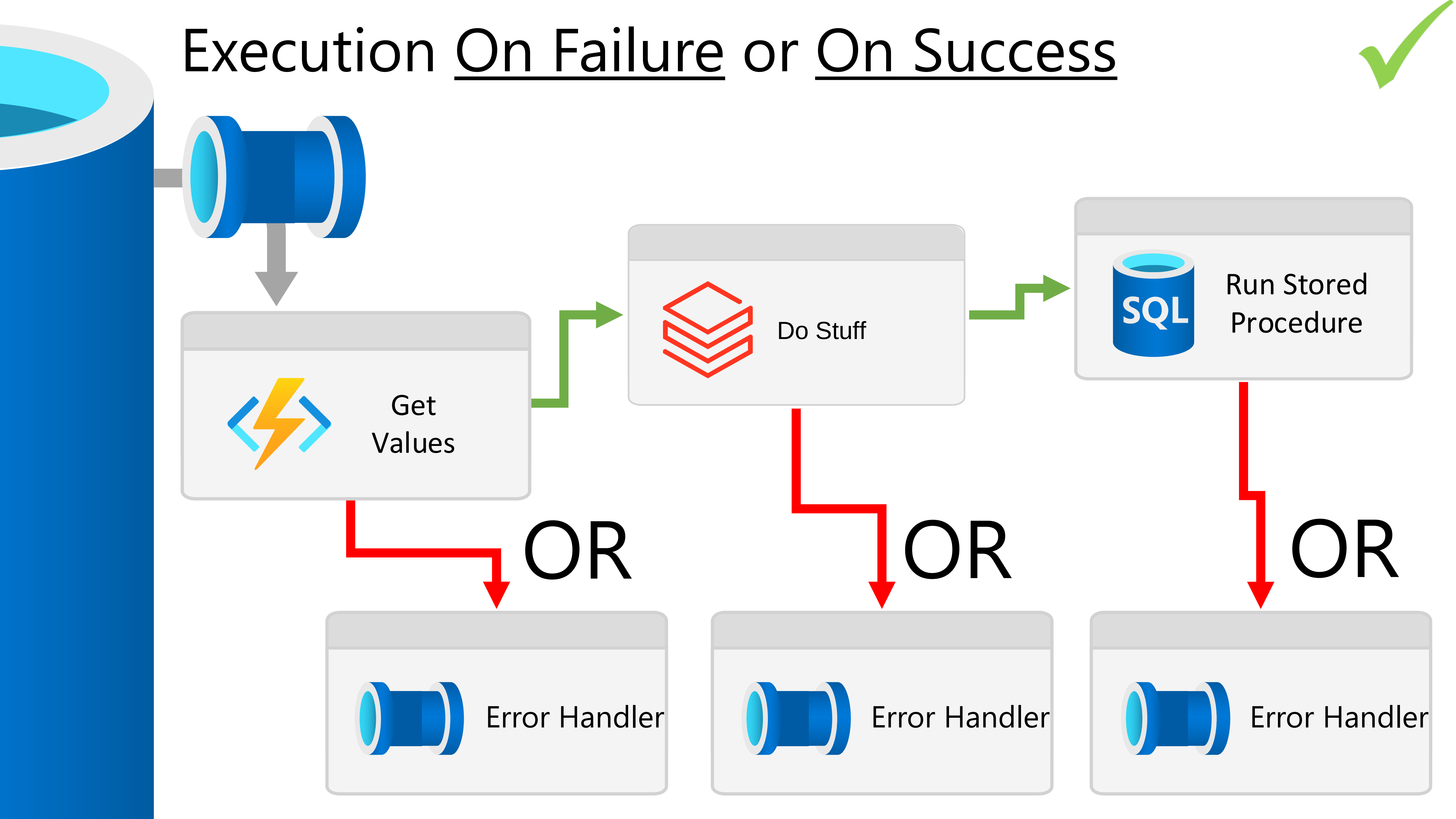
Execution On Failure or On Success



Execution On ???

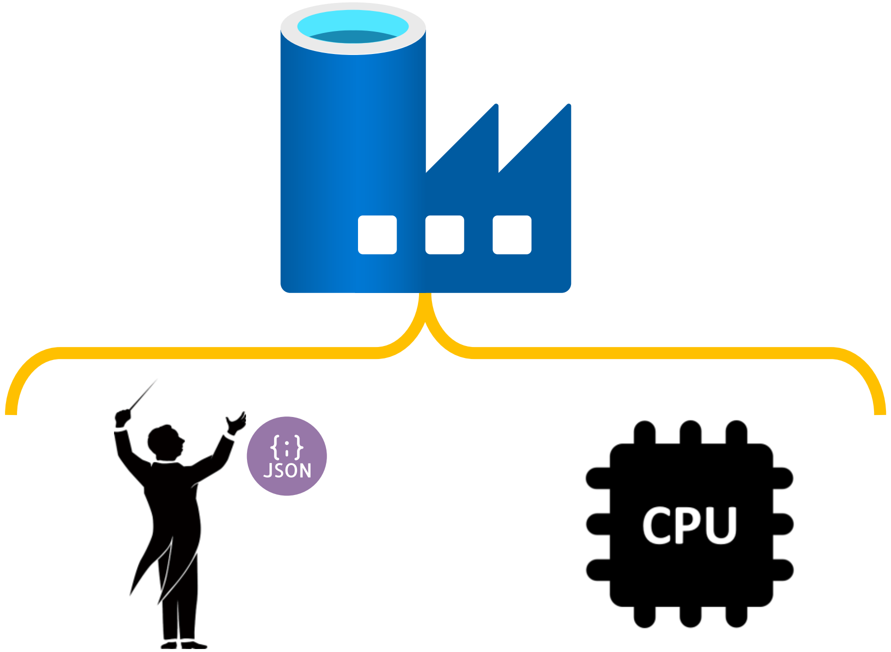


Execution On Failure or On Success ✓



Integration Runtimes

What is an Integration Runtime?

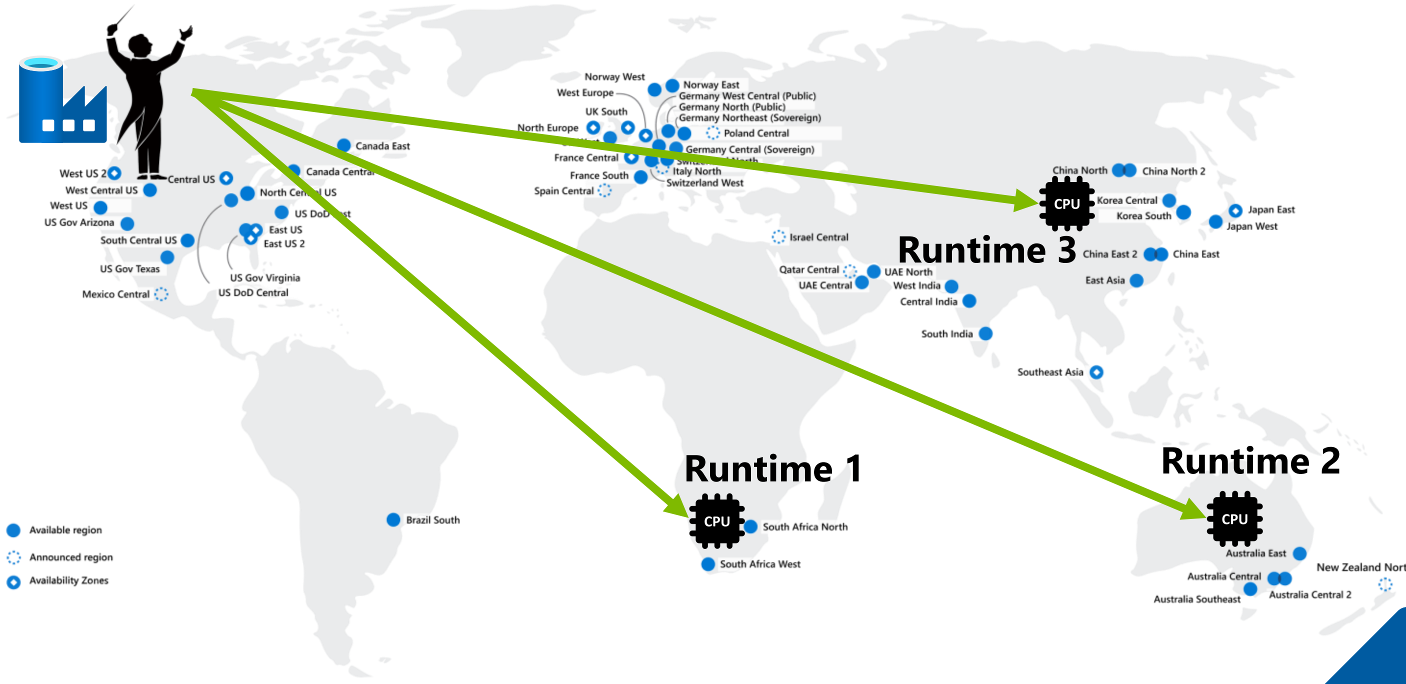


Orchestrator

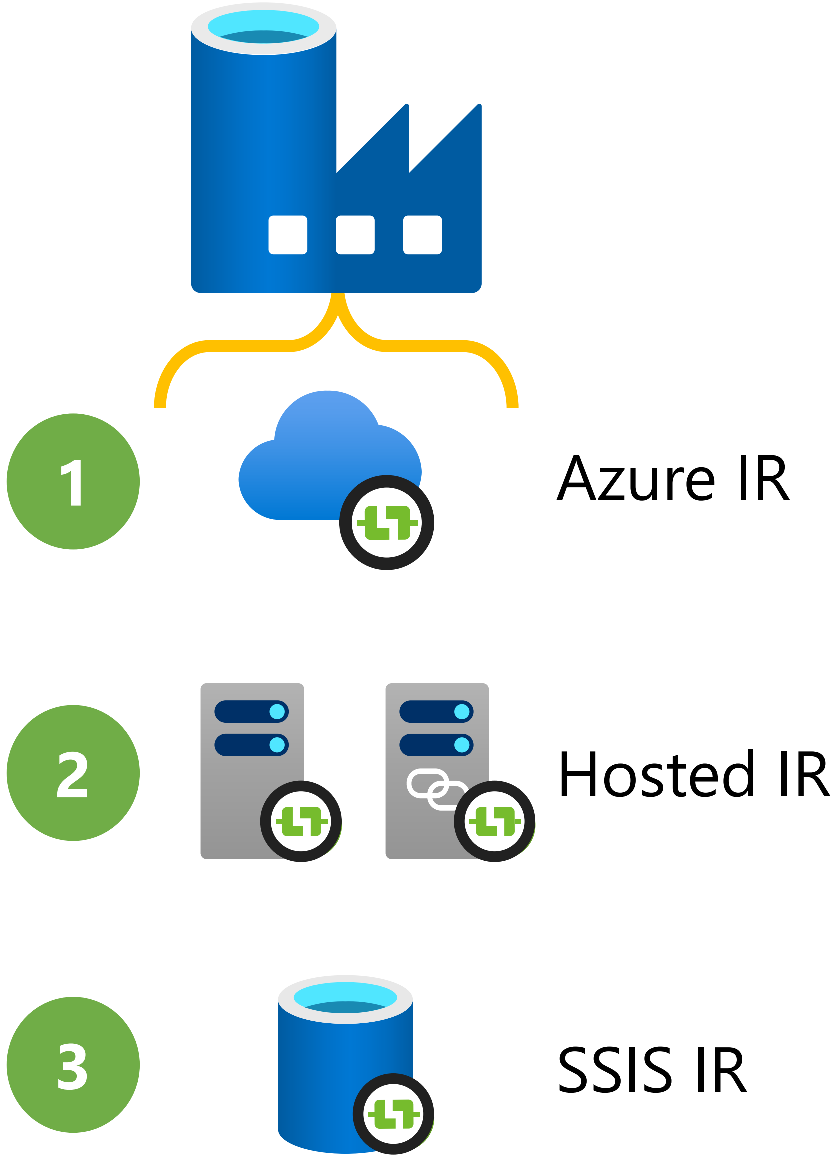
Fixed Region

Runtime

Flexible Location



What can an Integration Runtime do?

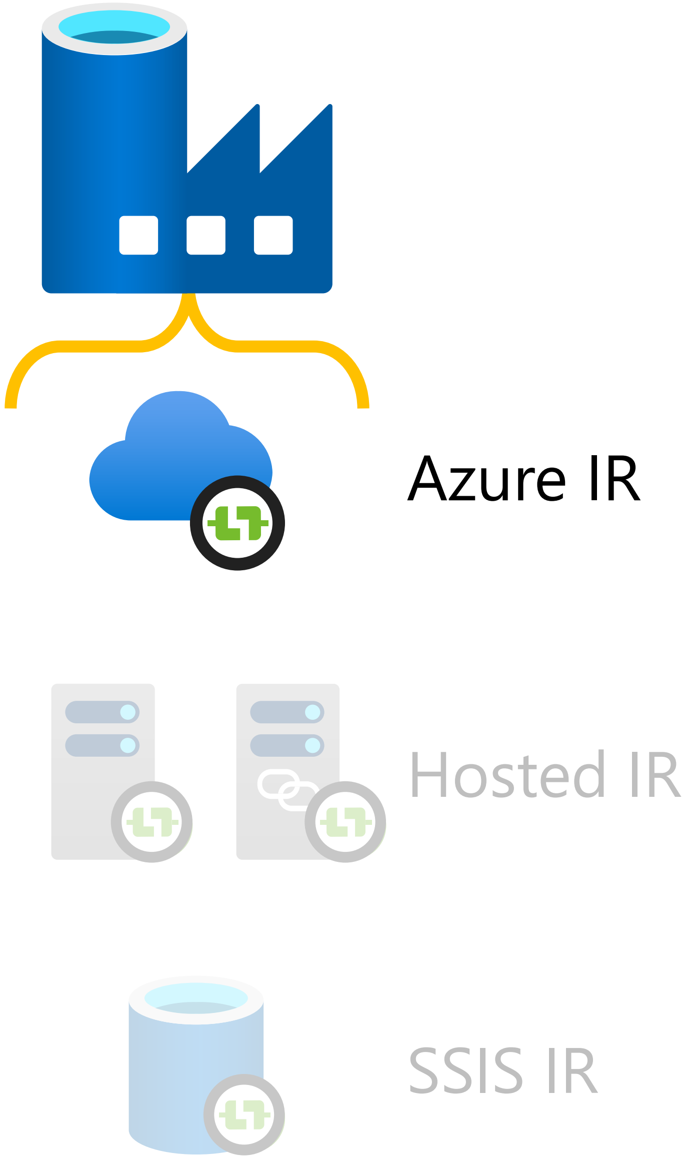


1 Azure IR

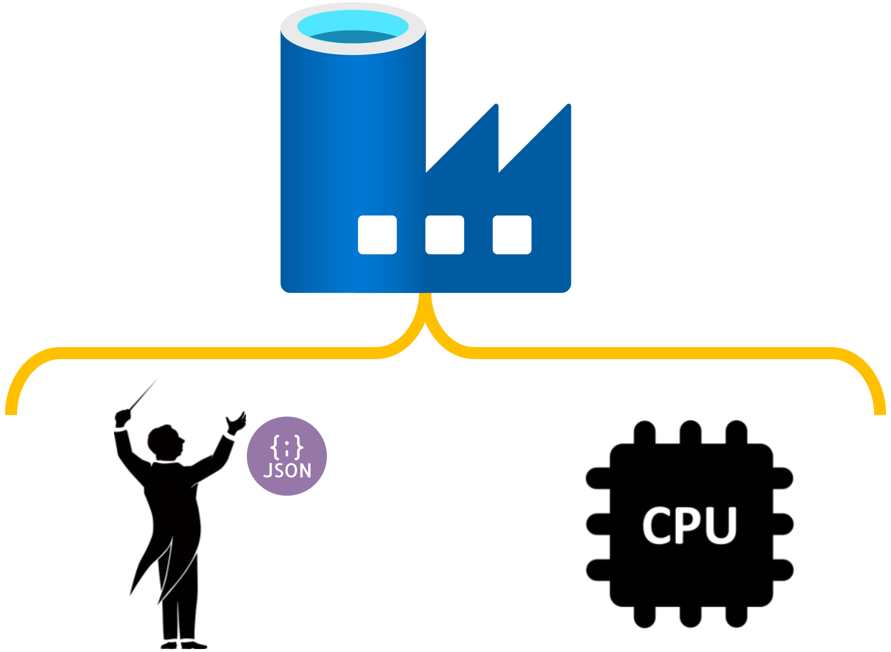
2 Hosted IR

3 SSIS IR

Azure Integration Runtime



Azure Integration Runtime

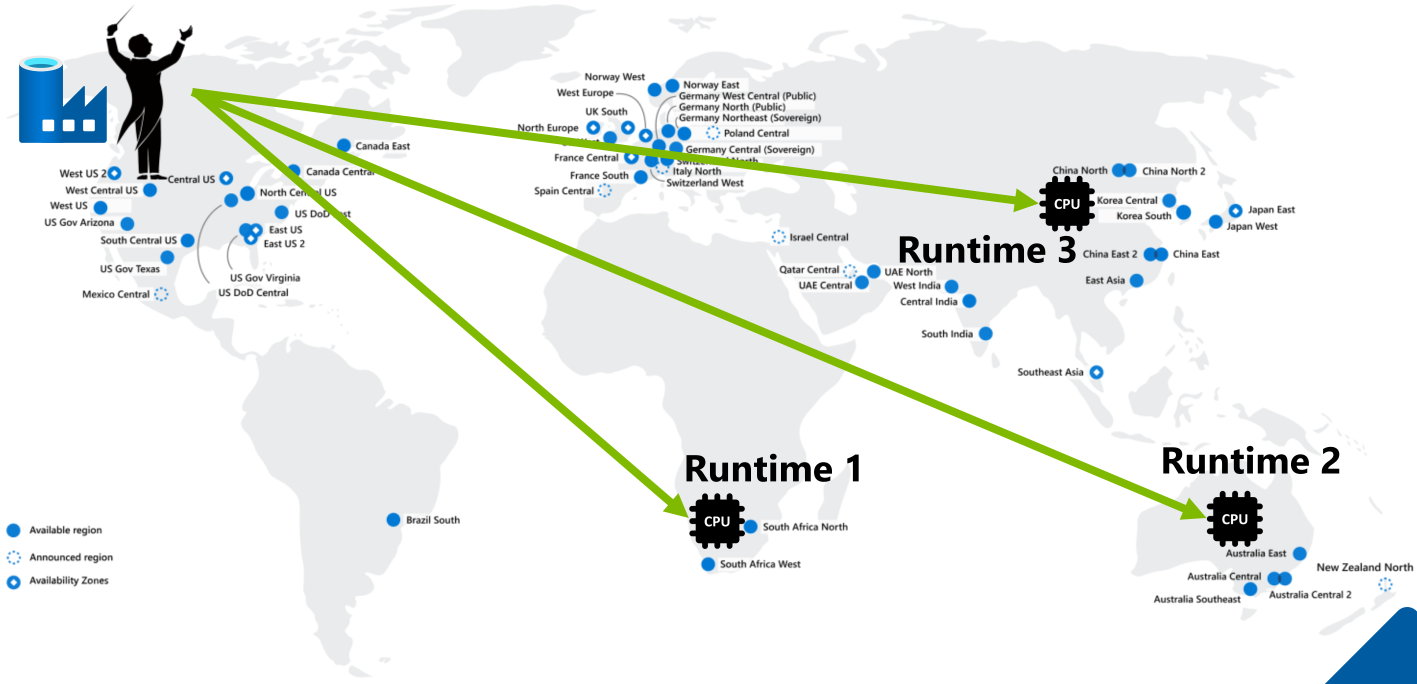


Orchestrator

Fixed Region

Runtime

Flexible Location

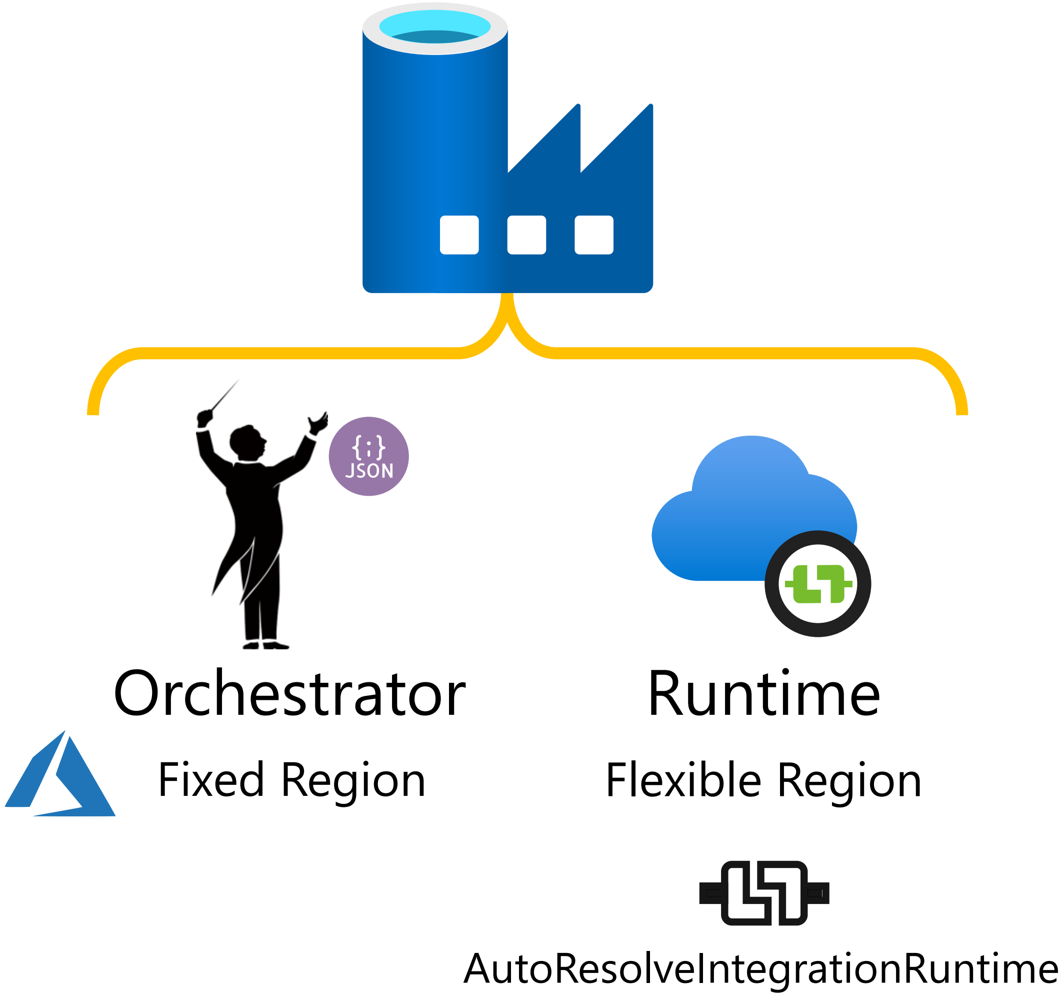


Runtime 3

Runtime 1

Runtime 2

Azure Integration Runtime



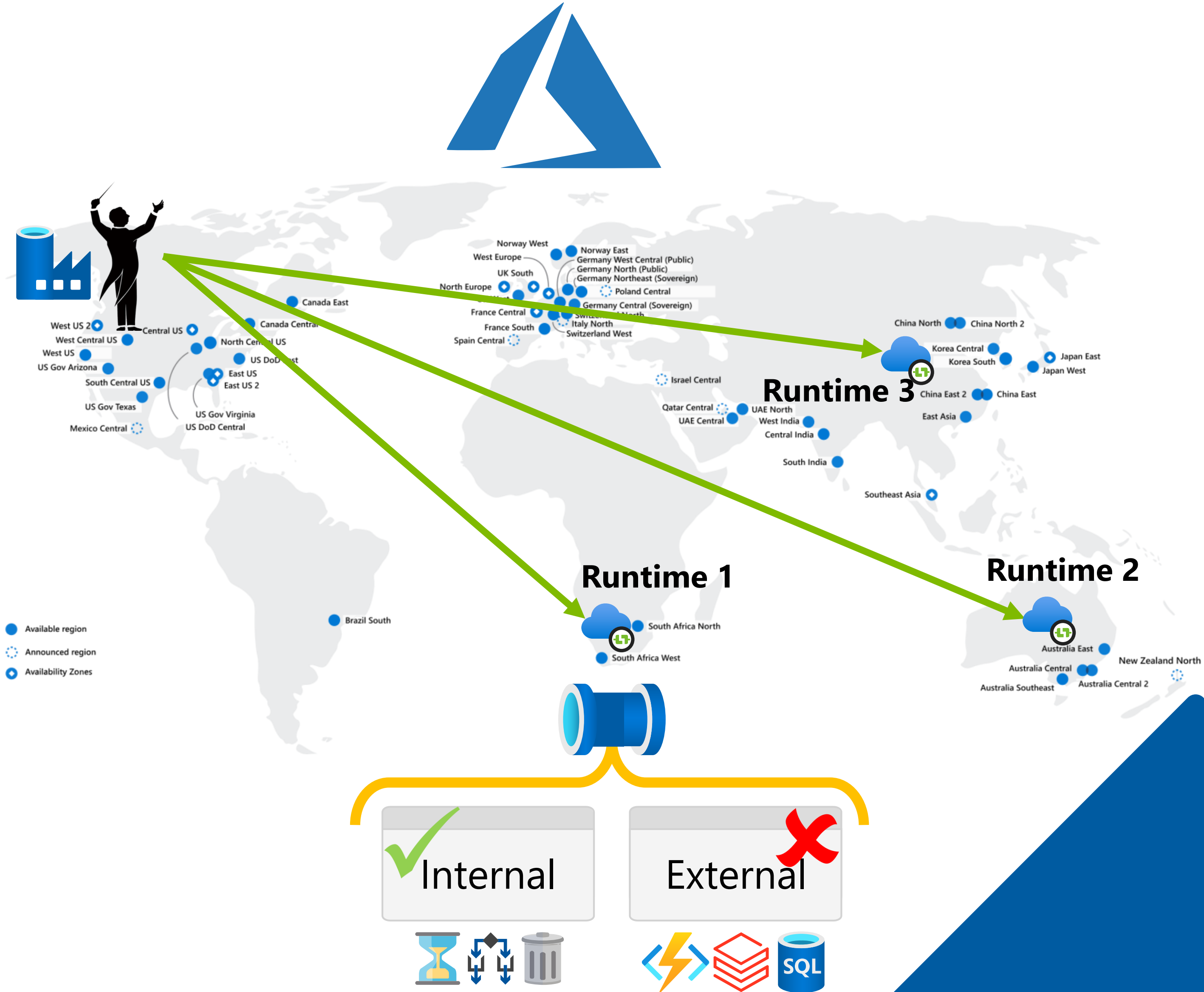
Orchestrator

Fixed Region

Runtime

Flexible Region

AutoResolveIntegrationRuntime



Runtime 3

Runtime 1

Runtime 2

✓ Internal

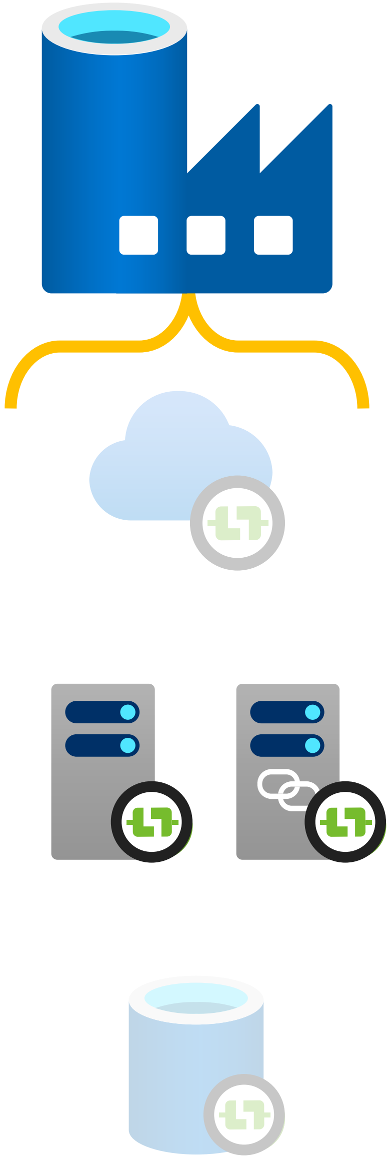
External ✗

● Available region

◌ Announced region

◈ Availability Zones

Hosted Integration Runtime

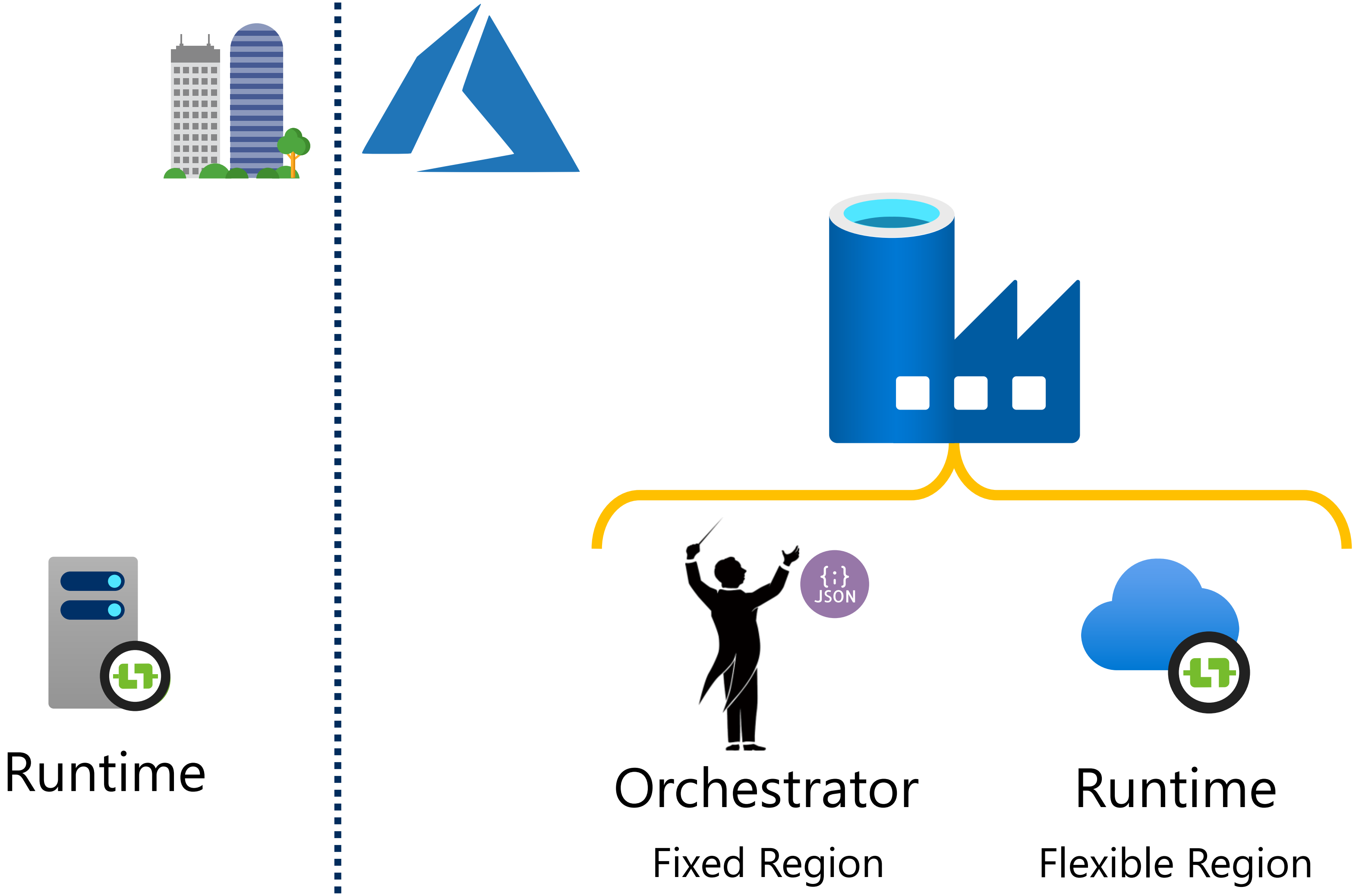


Azure IR

Hosted IR

SSIS IR

Hosted Integration Runtime



Runtime

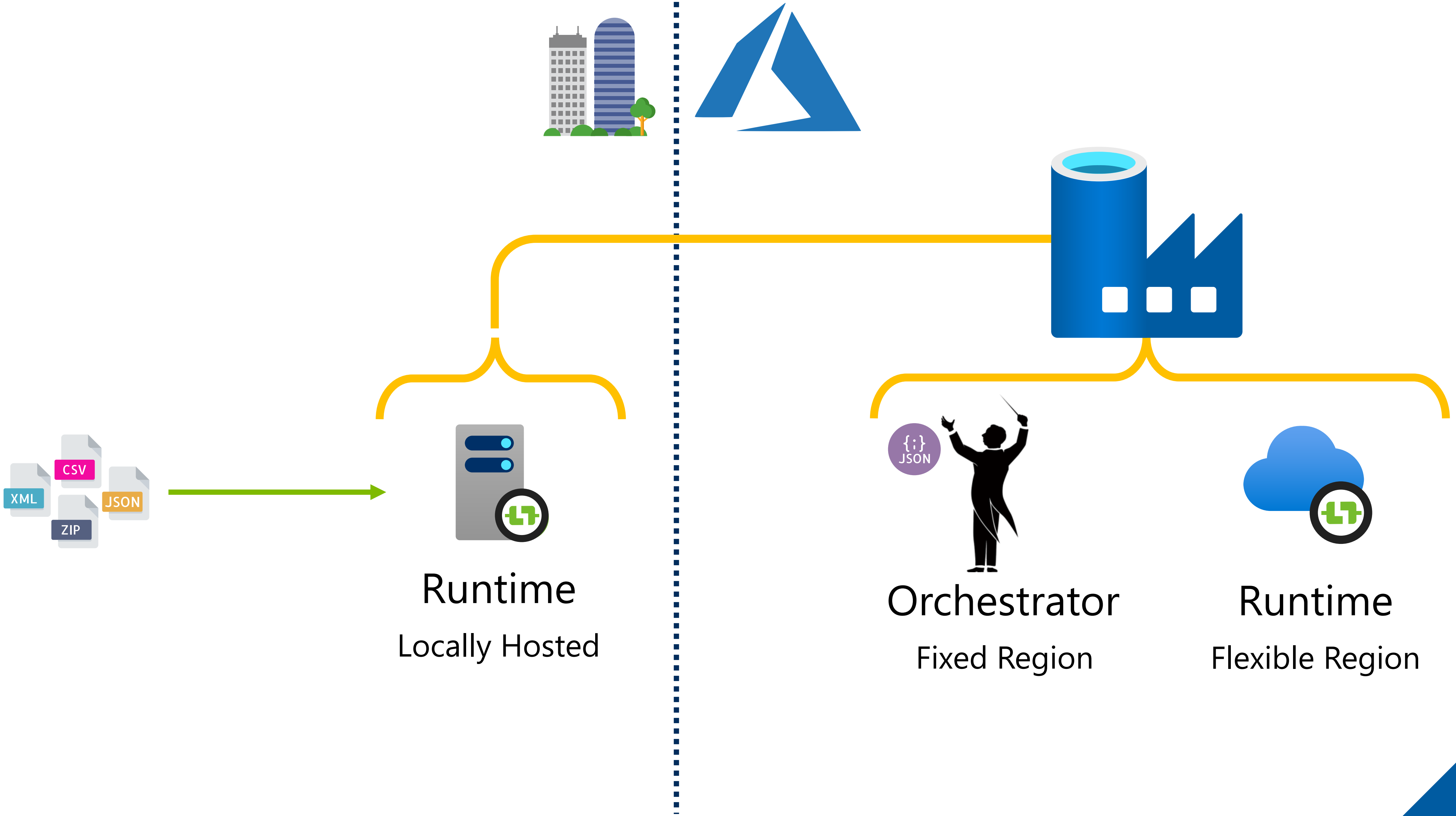
Orchestrator

Fixed Region

Runtime

Flexible Region

Hosted Integration Runtime



Runtime

Locally Hosted

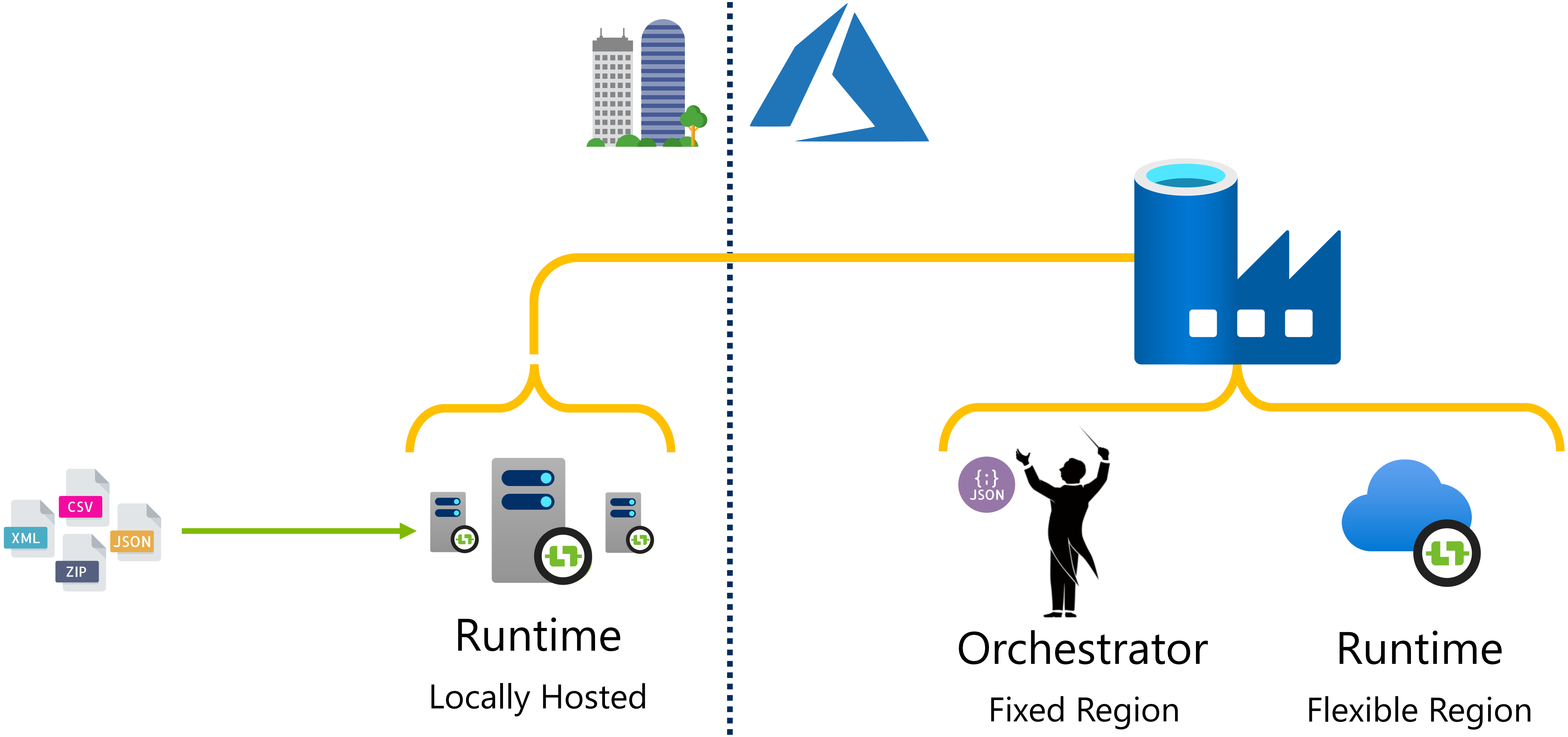
Orchestrator

Fixed Region

Runtime

Flexible Region

Hosted Integration Runtime – Secondary Nodes



Runtime

Locally Hosted

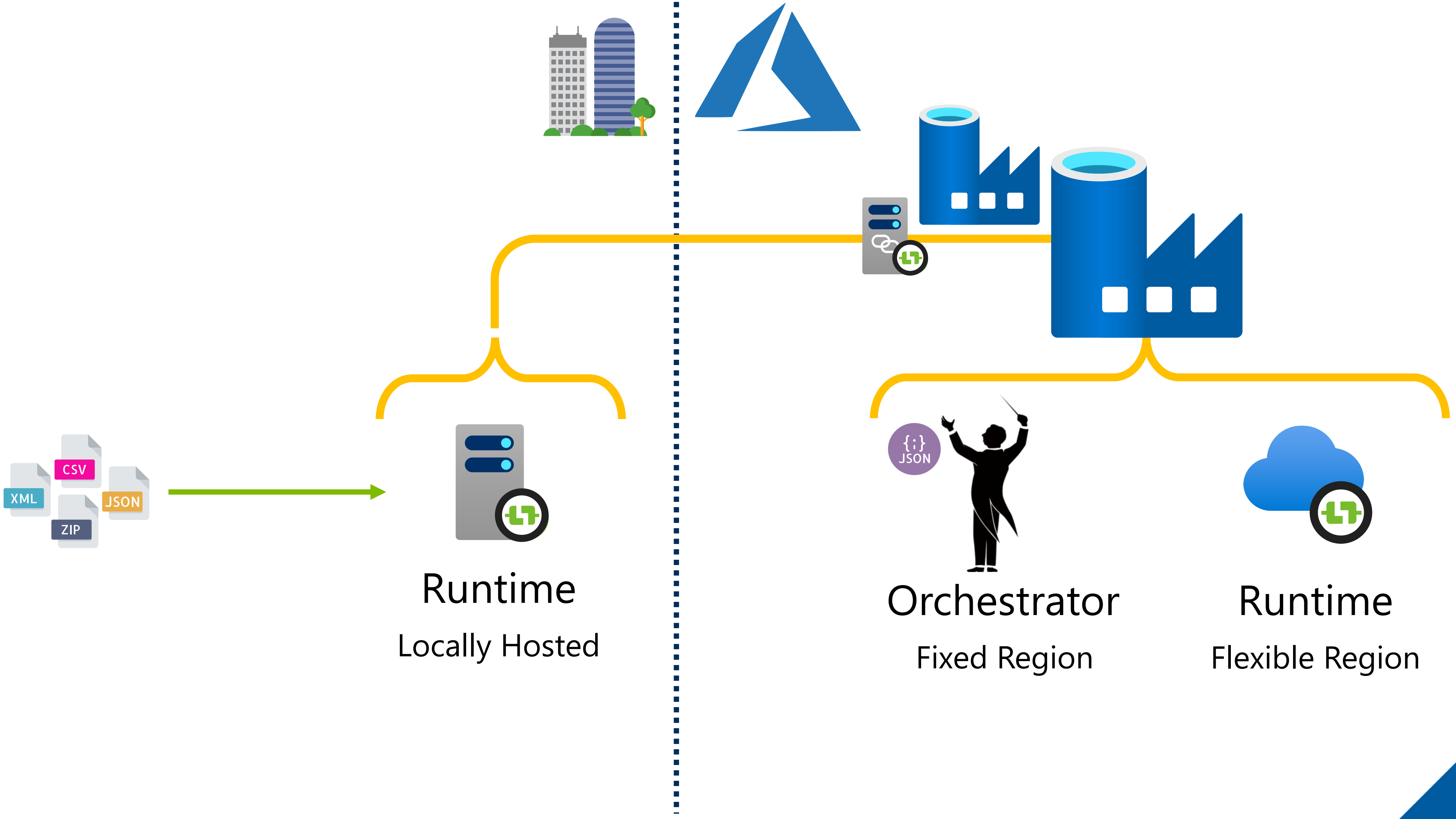
Orchestrator

Fixed Region

Runtime

Flexible Region

Hosted Integration Runtime – Linked



Runtime

Locally Hosted

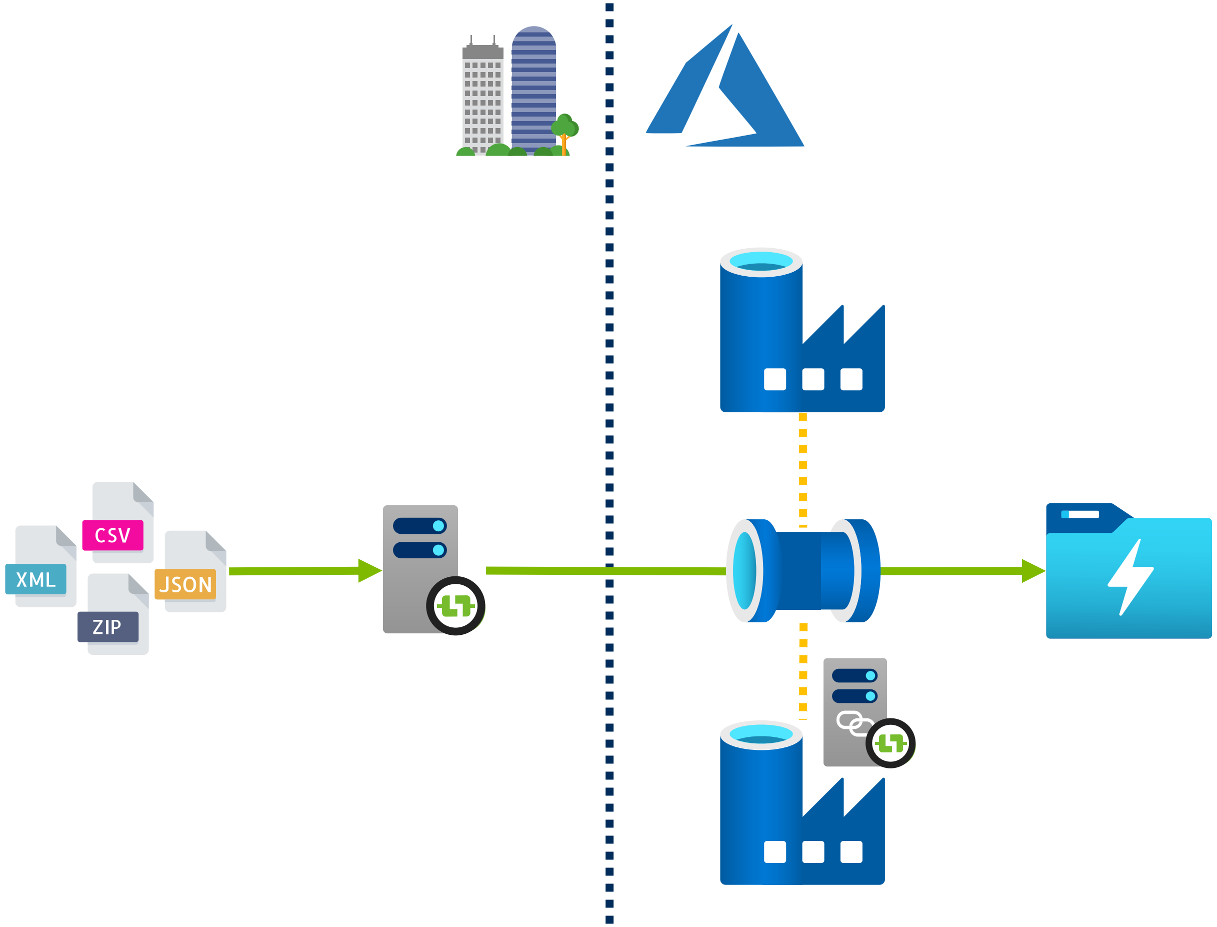
Orchestrator

Fixed Region

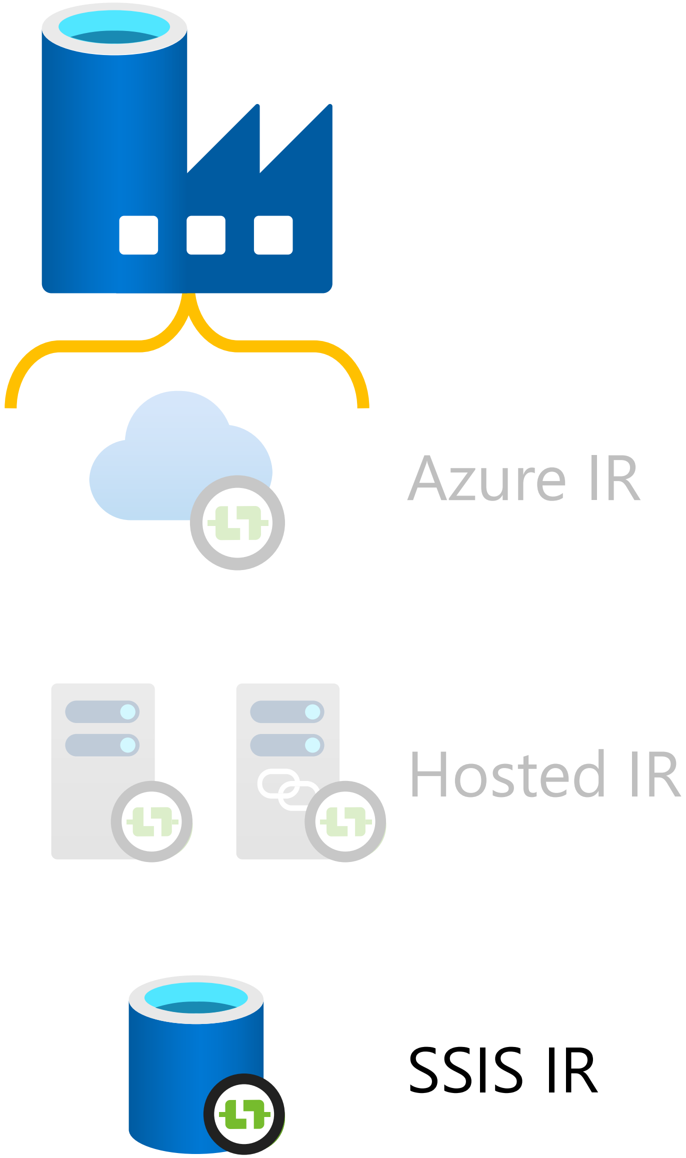
Runtime

Flexible Region

DEMO


CSV
XML
JSON
ZIP

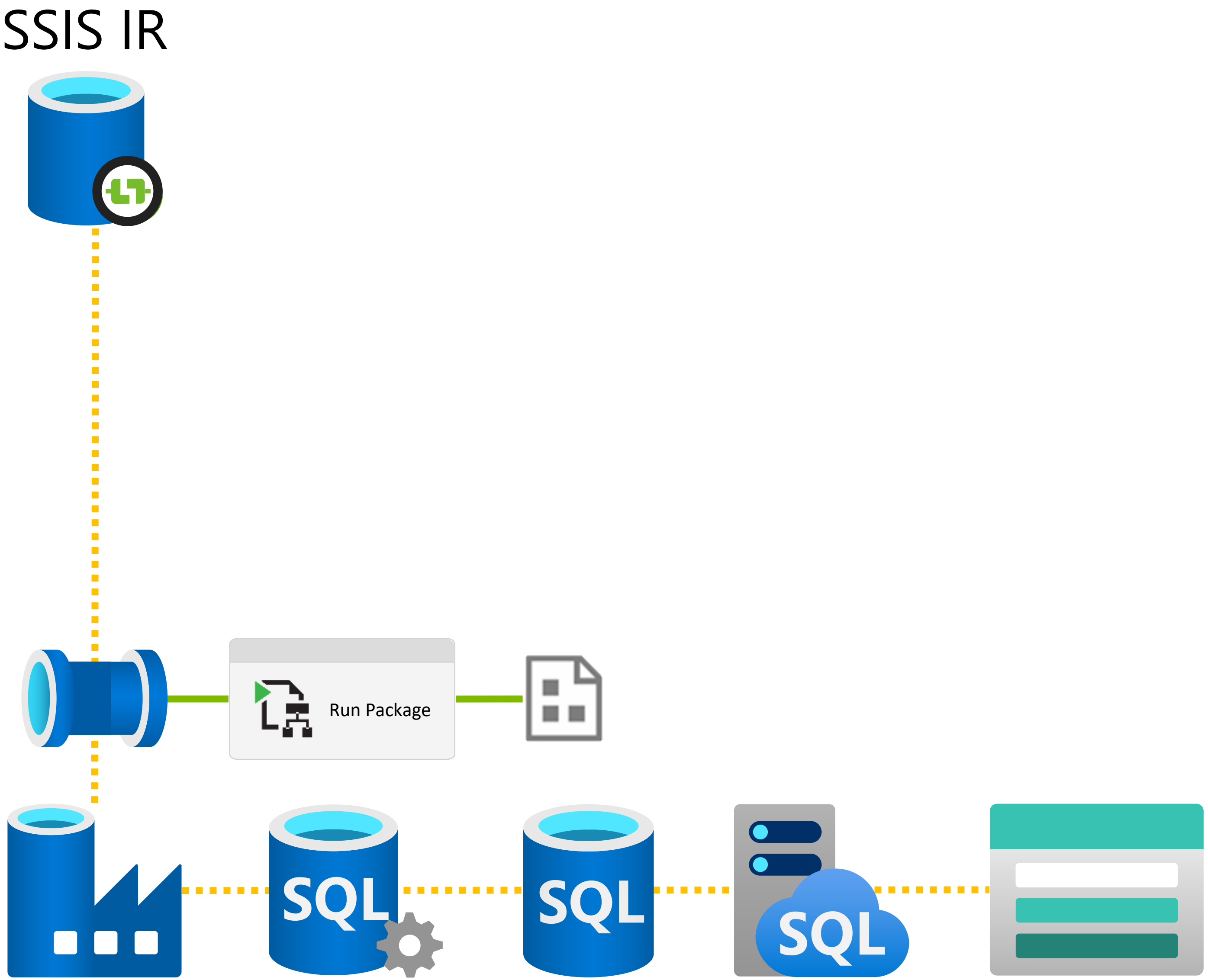
SSIS Integration Runtime



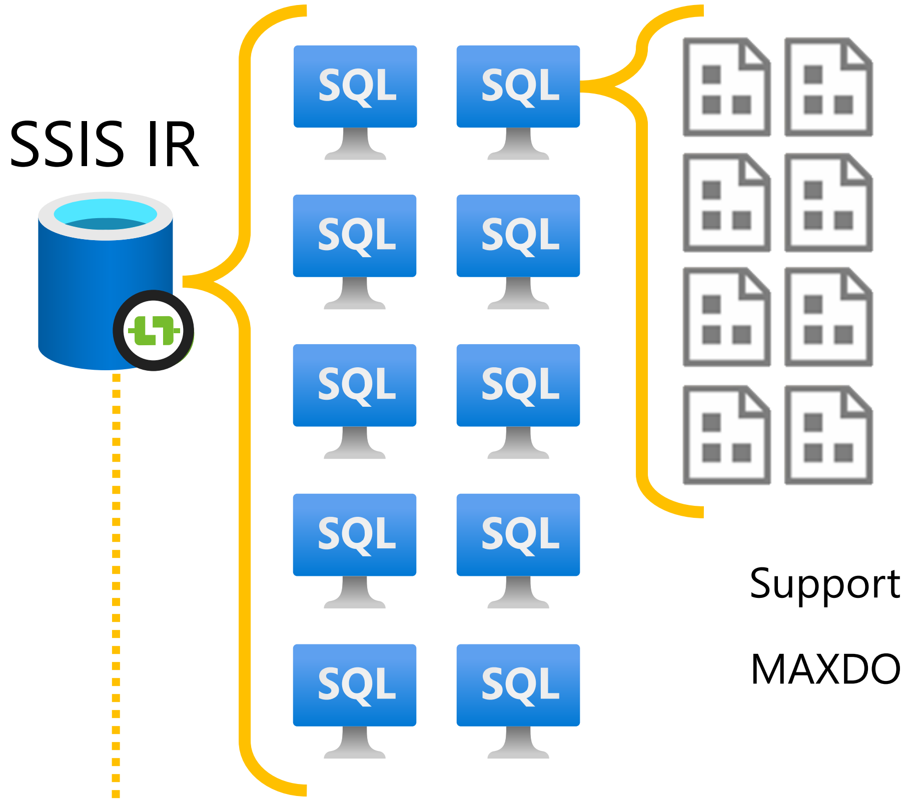
Running an SSIS Package in Azure

SSIS IR

Running an SSIS Package in Azure

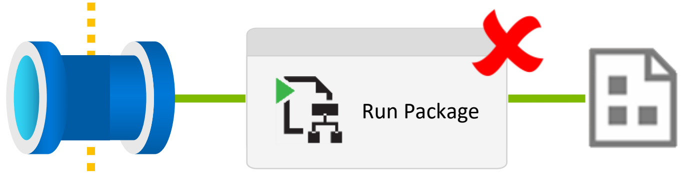


Problem: Using All Of The SSIS IR Compute

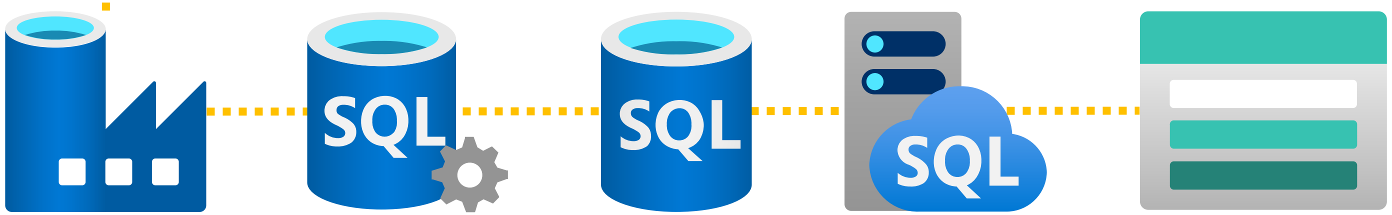


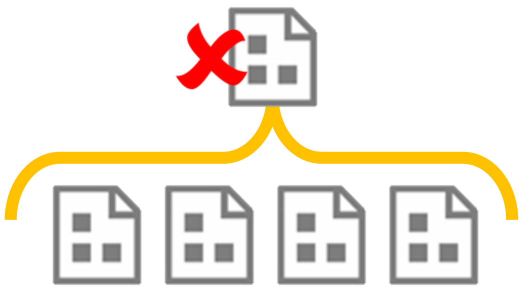
Supports 80 Concurrent Packages

MAXDOP = 80



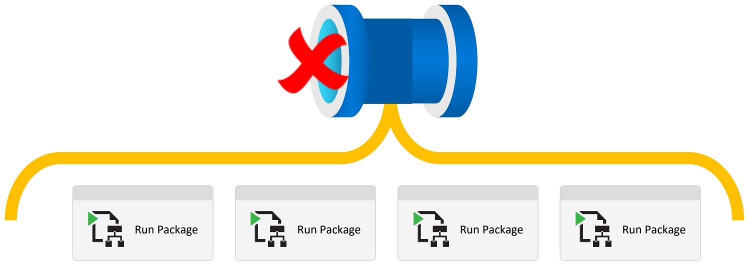
Runs 1 Package





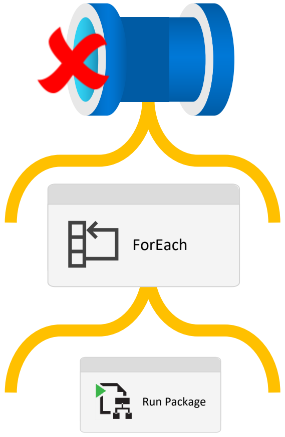
Parent Package

Child Packages x80



Pipeline x1

Activities x80

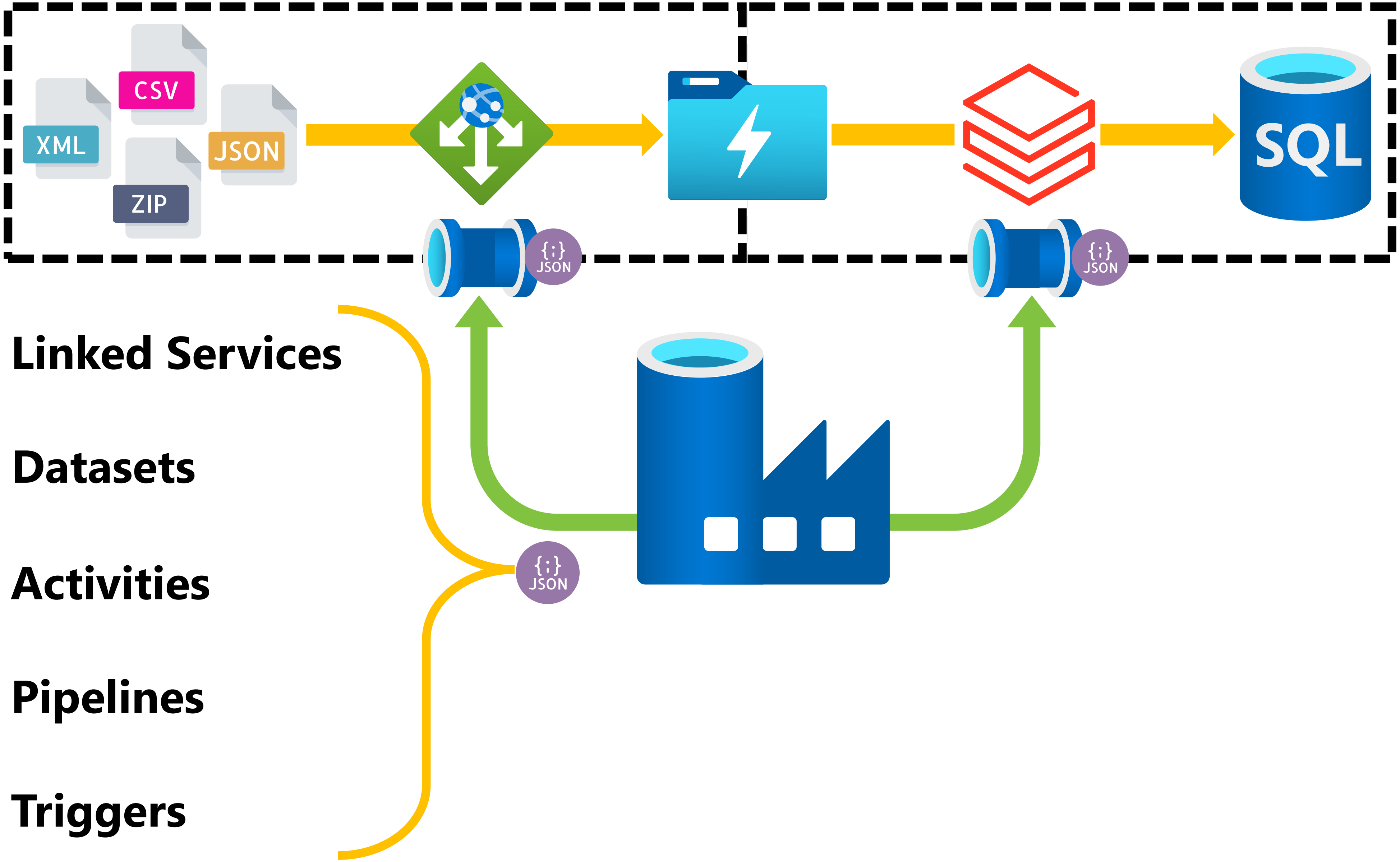


Pipeline x1

ForEach
Max Batch
(50)

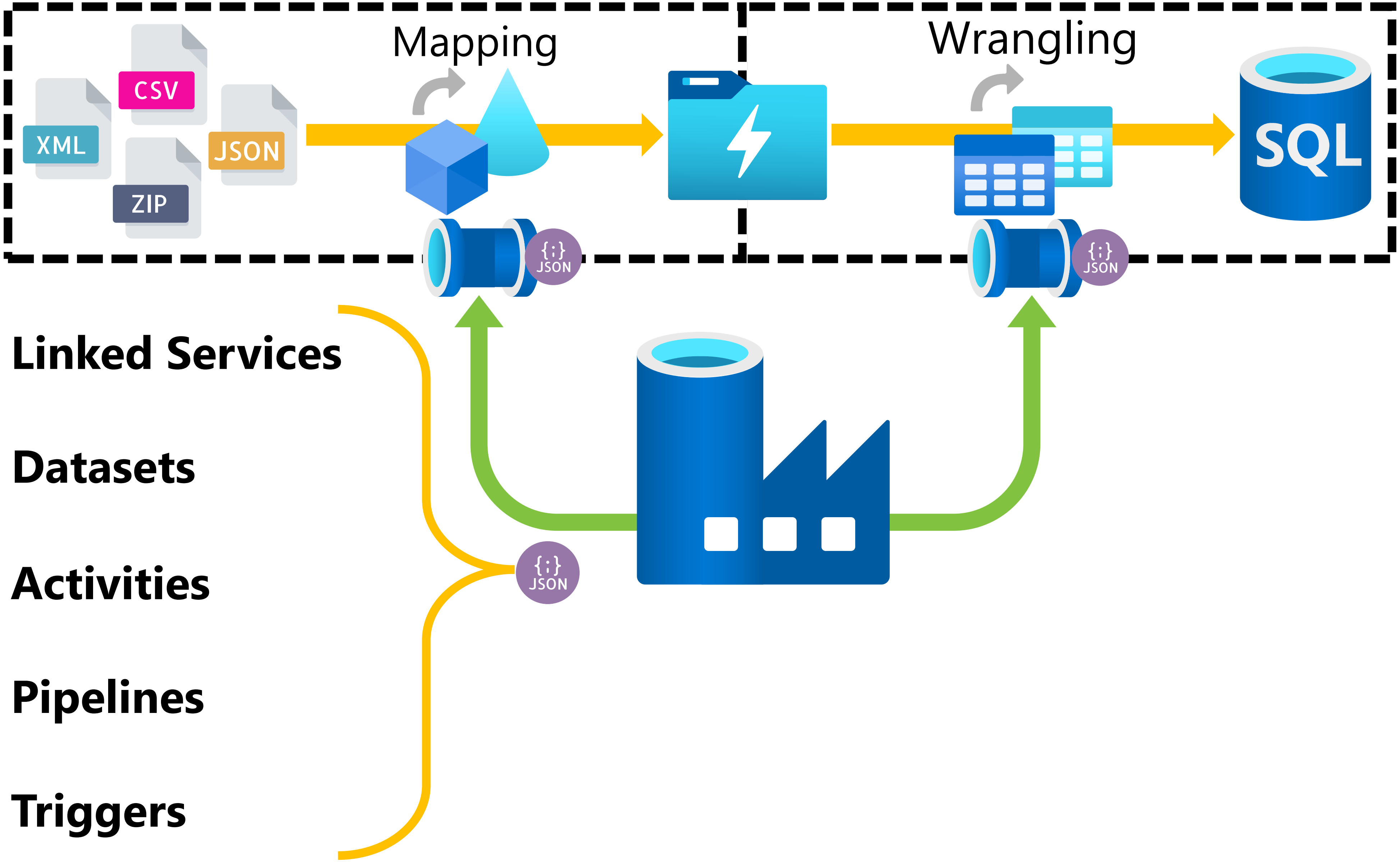
Data Factory Data Flows

Data Factory Control Flow Components



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

Data Factory Data Flow Components



1 **Linked Services**

2 **Datasets**

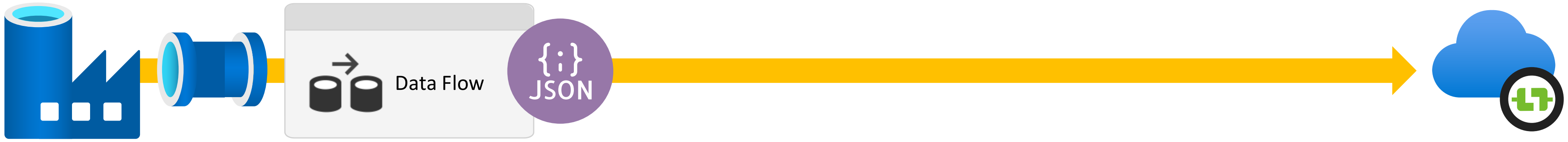
3 **Activities**

4 **Pipelines**

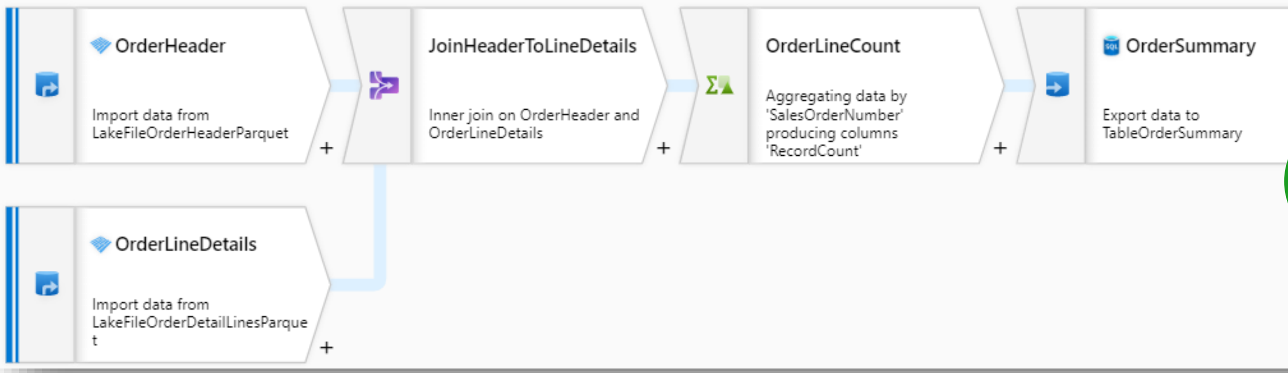
5 **Triggers**

What is a Mapping Data Flow?

Control Flow



Data Flow



OrderHeader

Import data from LakeFileOrderHeaderParquet

JoinHeaderToLineDetails

Inner join on OrderHeader and OrderLineDetails

OrderLineCount

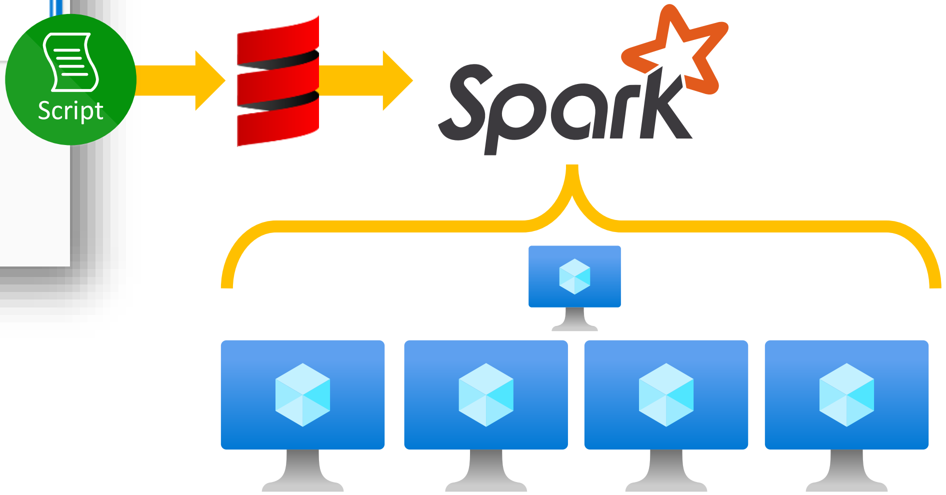
Aggregating data by 'SalesOrderNumber' producing columns 'RecordCount'

OrderSummary

Export data to TableOrderSummary

OrderLineDetails

Import data from LakeFileOrderDetailLinesParquet



A: Graphic data transformation tool that sits on top of Apache Spark.

What can a Mapping Data Flow do? - Transformations

- New Branch
- Join
- Conditional Split
- Exists
- Union
- Lookup

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

- Filter
- Sort
- Alter Row

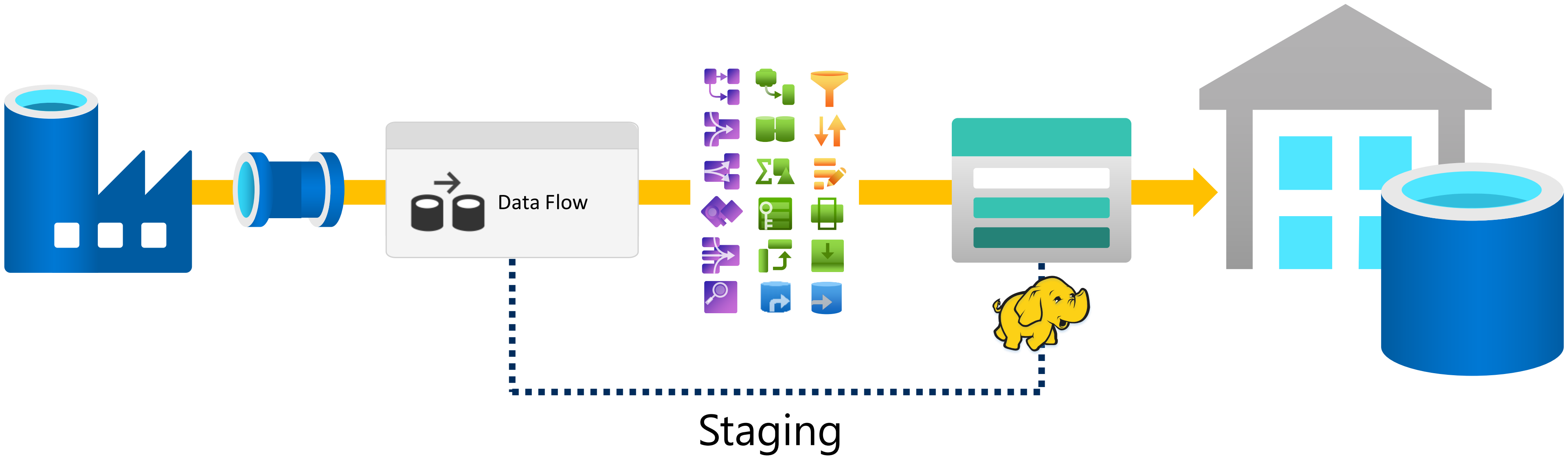
Key

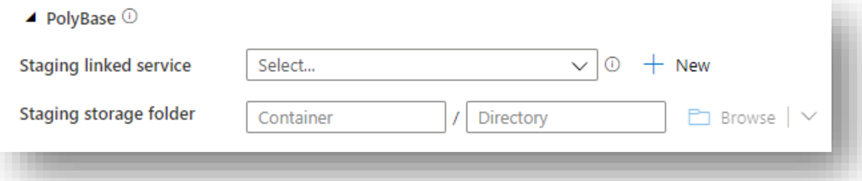
Input & Output Modifiers

Schema Modifiers

Row Modifiers

What can a Mapping Data Flow do? - PolyBase





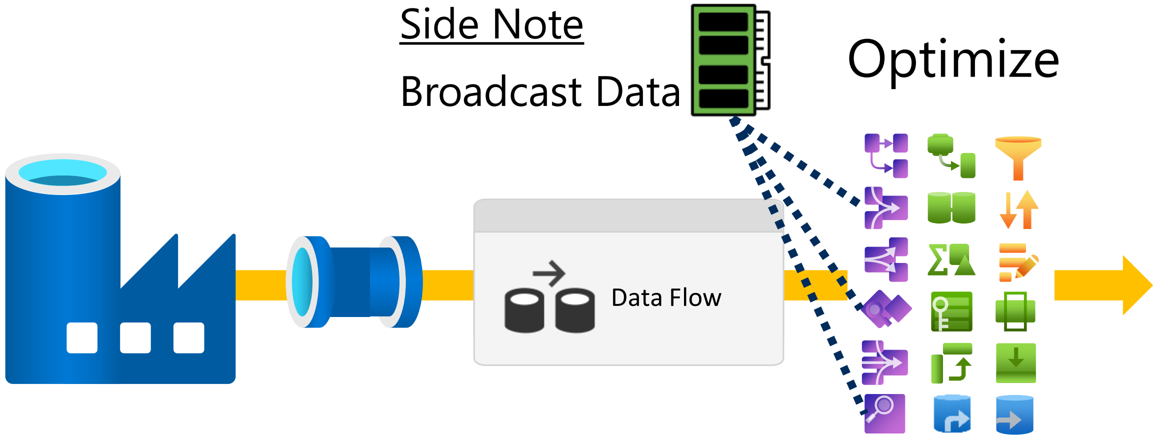
◢ PolyBase ⓘ

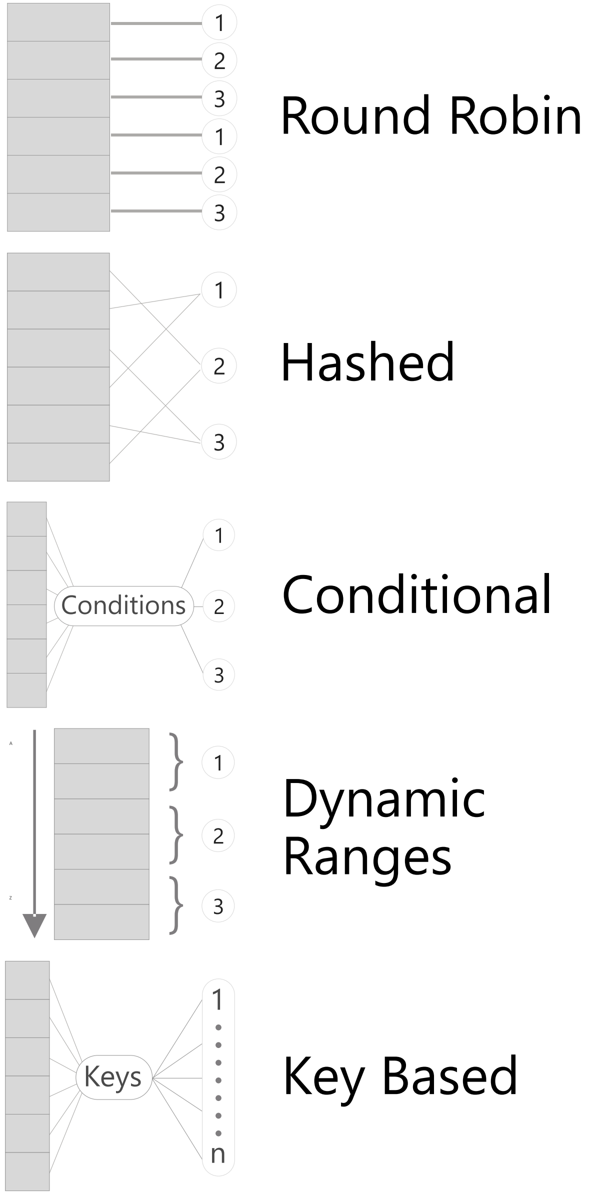
Staging linked service: Select... ⓘ + New

Staging storage folder: Container / Directory Browse

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

What can a Mapping Data Flow do? - Partition Handling

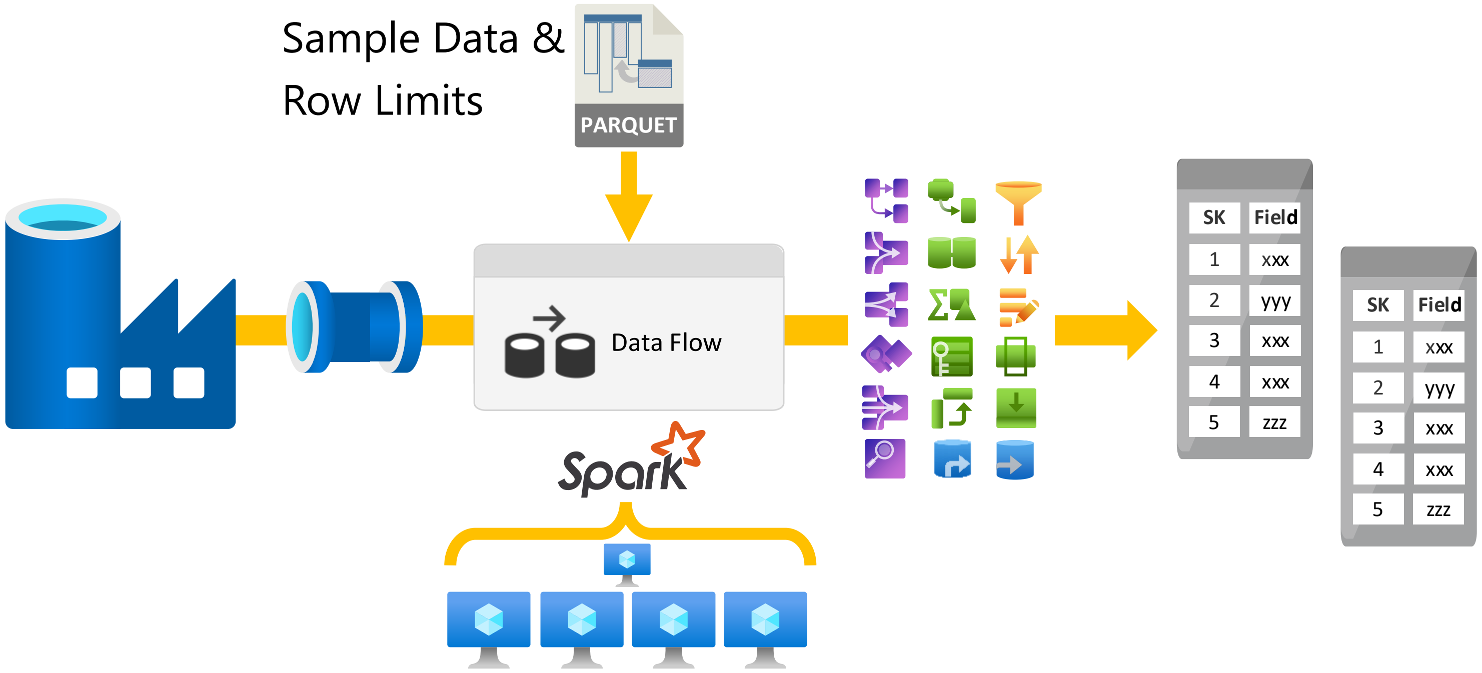




What can a Mapping Data Flow do? - Debugging

Enable Data Flow Debug Mode

Data Preview



Sample Data & Row Limits

PARQUET

Data Flow

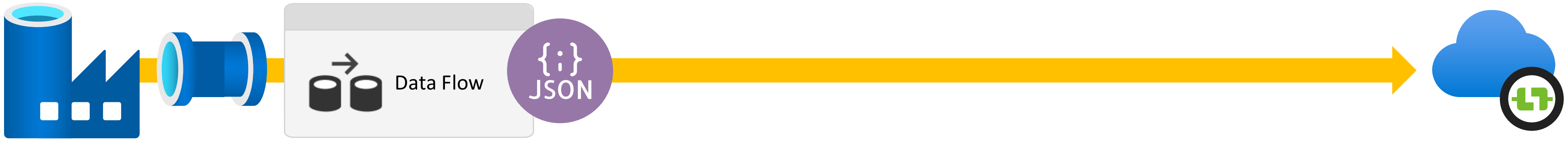
Spark

SK	Field
1	xxx
2	yyy
3	xxx
4	xxx
5	zzz

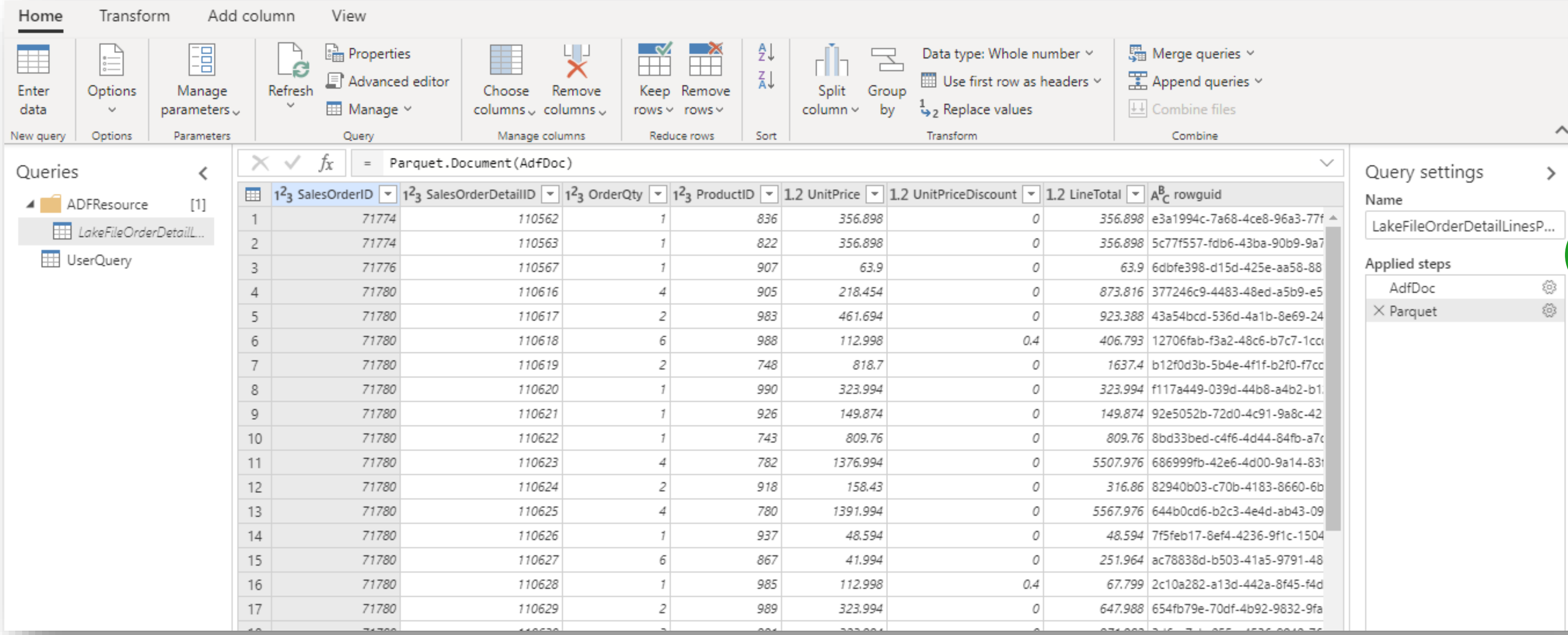
SK	Field
1	xxx
2	yyy
3	xxx
4	xxx
5	zzz

What is a Wrangling Data Flow?

Control Flow



Data Flow



Queries

- ADFResource [1]
  - LakeFileOrderDetailL...
- UserQuery

= Parquet.Document(AdfDoc)

	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-a7
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

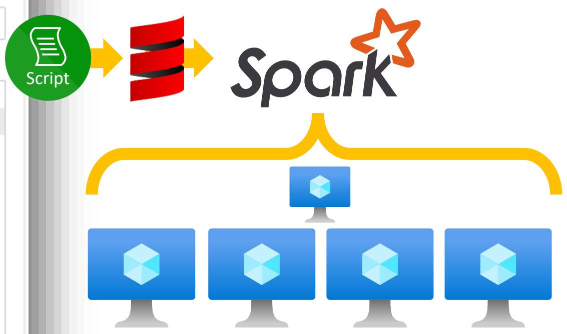
Query settings

Name

LakeFileOrderDetailLinesP...

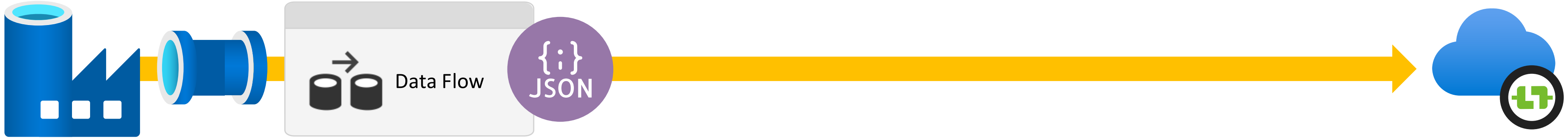
Applied steps

- AdfDoc
- Parquet

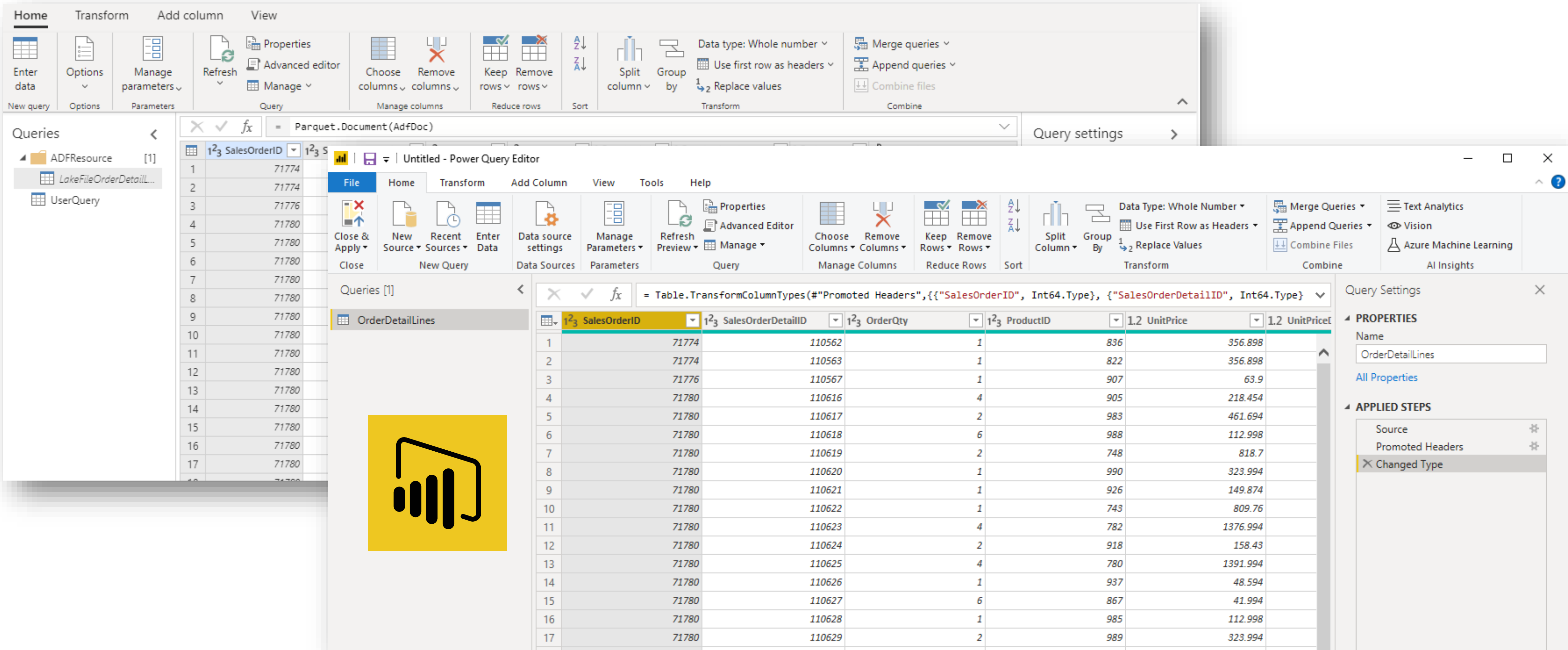


What can a Wrangling Data Flow do? - Home

Control Flow



Data Flow



```
= Parquet.Document(AdfDoc)
```

	SalesOrderID
1	71774
2	71774
3	71776
4	71780
5	71780
6	71780
7	71780
8	71780
9	71780
10	71780
11	71780
12	71780
13	71780
14	71780
15	71780
16	71780
17	71780

```
= Table.TransformColumnTypes(#"Promoted Headers",{{"SalesOrderID", Int64.Type}, {"SalesOrderDetailID", Int64.Type}
```

	SalesOrderID	SalesOrderDetailID	OrderQty	ProductID	UnitPrice	UnitPriceI
1	71774	110562	1	836	356.898	
2	71774	110563	1	822	356.898	
3	71776	110567	1	907	63.9	
4	71780	110616	4	905	218.454	
5	71780	110617	2	983	461.694	
6	71780	110618	6	988	112.998	
7	71780	110619	2	748	818.7	
8	71780	110620	1	990	323.994	
9	71780	110621	1	926	149.874	
10	71780	110622	1	743	809.76	
11	71780	110623	4	782	1376.994	
12	71780	110624	2	918	158.43	
13	71780	110625	4	780	1391.994	
14	71780	110626	1	937	48.594	
15	71780	110627	6	867	41.994	
16	71780	110628	1	985	112.998	
17	71780	110629	2	989	323.994	

PROPERTIES

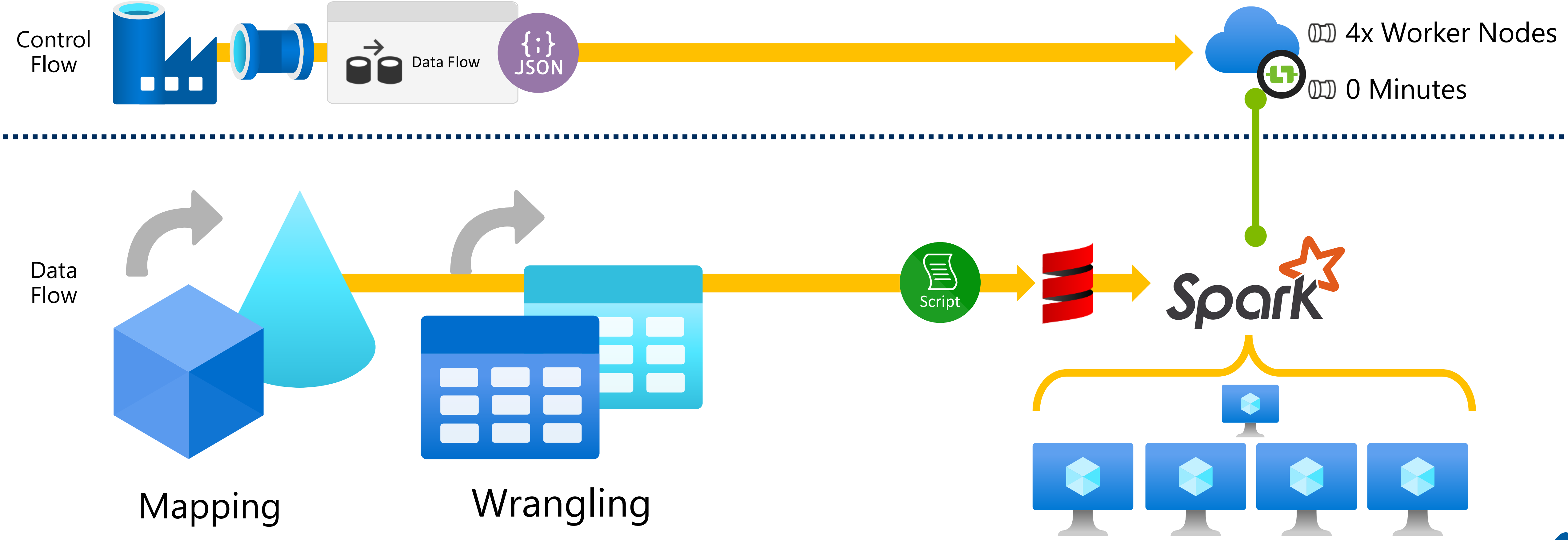
Name: OrderDetailLines

All Properties

APPLIED STEPS

- Source
- Promoted Headers
- Changed Type

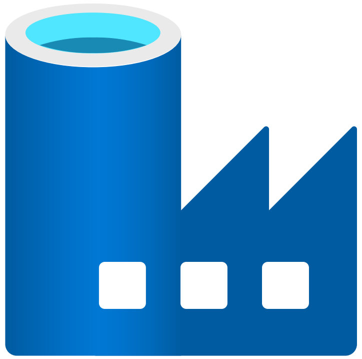
Data Flow Cluster Configuration



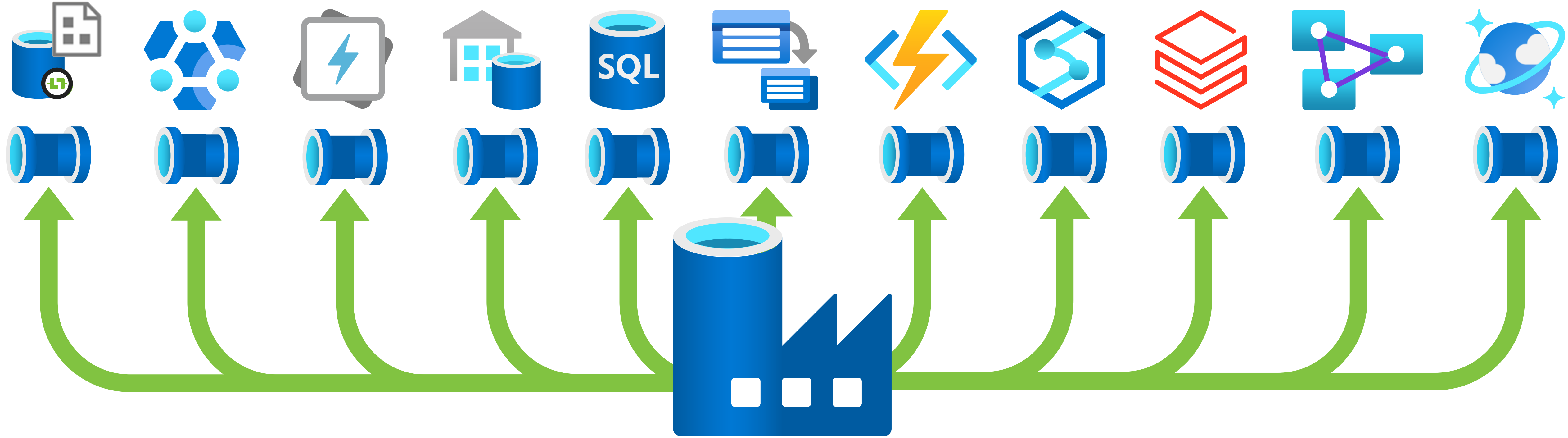
Default Azure IR

- General Purpose
- 4x Worker Nodes
- 0 Minutes

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



Other Data Transformation Services in Azure



SSIS Packages

HD Insight

Data Lake Analytics

Synapse SQLDW

SQL Database

Batch Service

Durable Functions

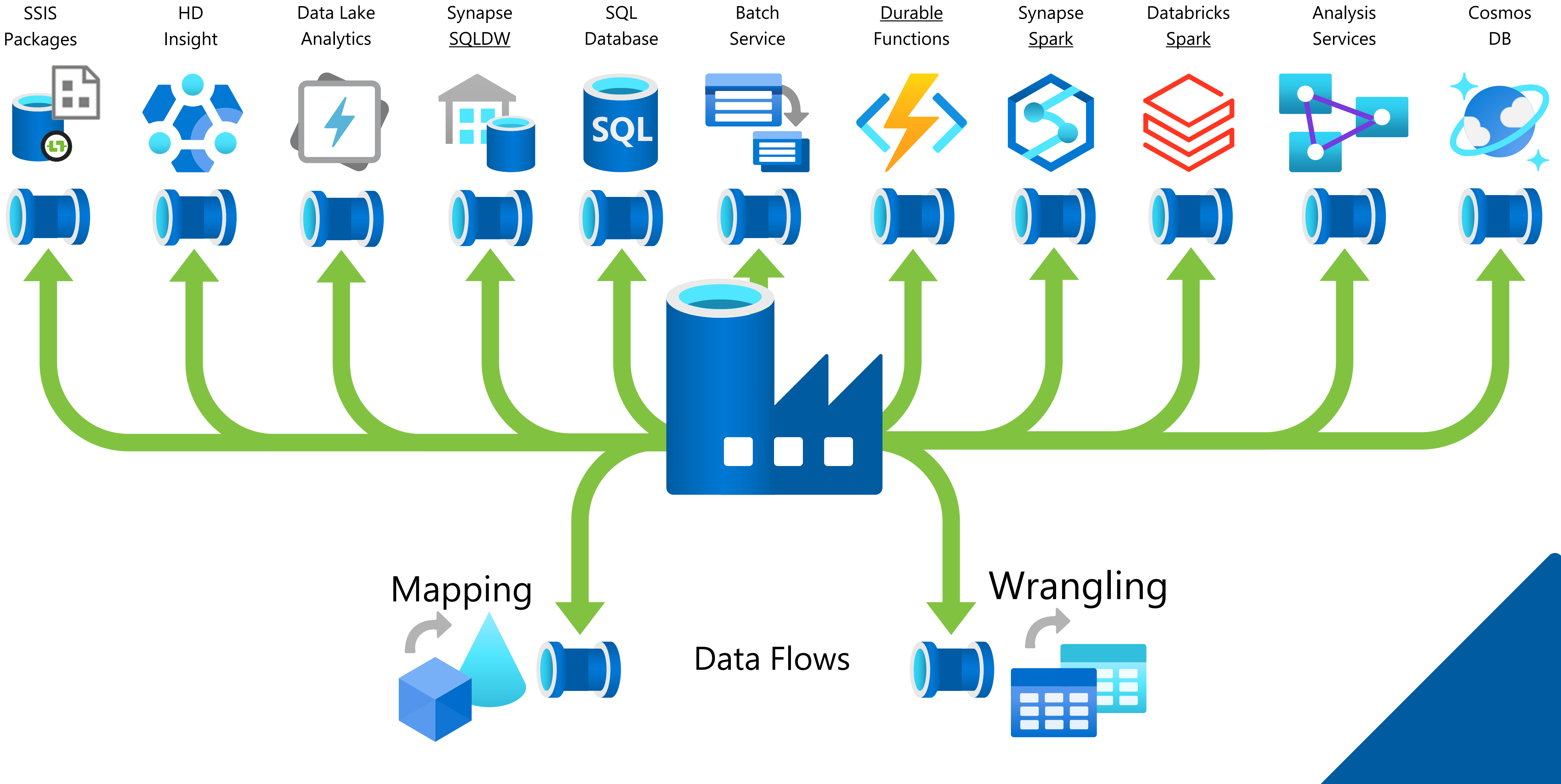
Synapse Spark

Databricks Spark

Analysis Services

Cosmos DB

When Should We Use Data Flows?

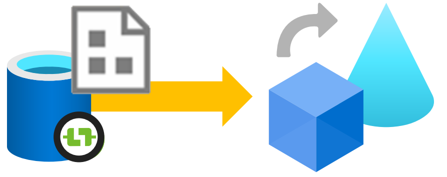


Data Transformations in Azure Comparisons

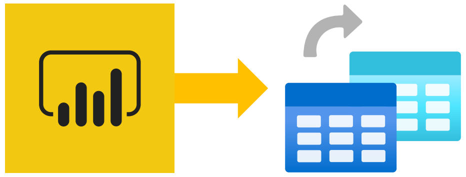
Transformation Method	Graphical UI	Scales Out	Scales Up	Cloud Native Tech
T-SQL (SQLDB)	✗	✗	✓	✗
SSIS	✓	✗	✓	✗
Scala (Databricks)	✗	✓	✓	✓
Data Factory Data Flows	✓	✓	✓	✓

Use Cases

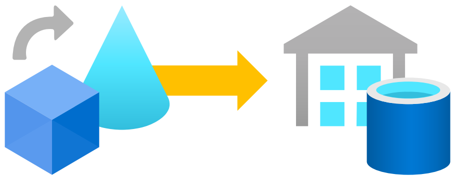
SSIS developers who are transferring existing skills to cloud native technologies have a very low barrier to entry and don’t need to worry about distributed compute to get started.



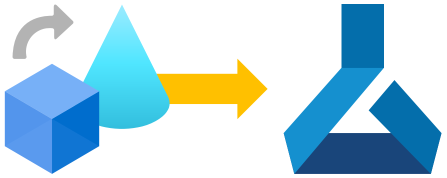
Data engineering made easy for the power users who has grown out of Power BI following a series of Data Lake exploration sessions.



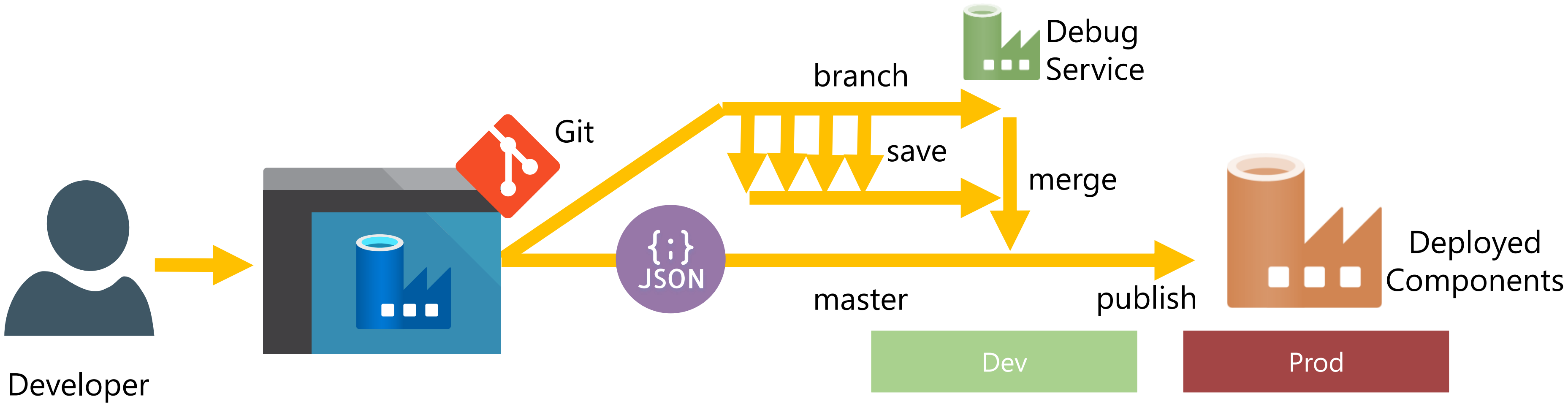
Data insight teams needing to do rapid prototyping and data warehouse loading within a single Azure Resource making deployments simple and release cycles short.



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



Source Code & ARM Deployments



Developer

Git

{;}
JSON

branch

save

merge

Debug
Service

master

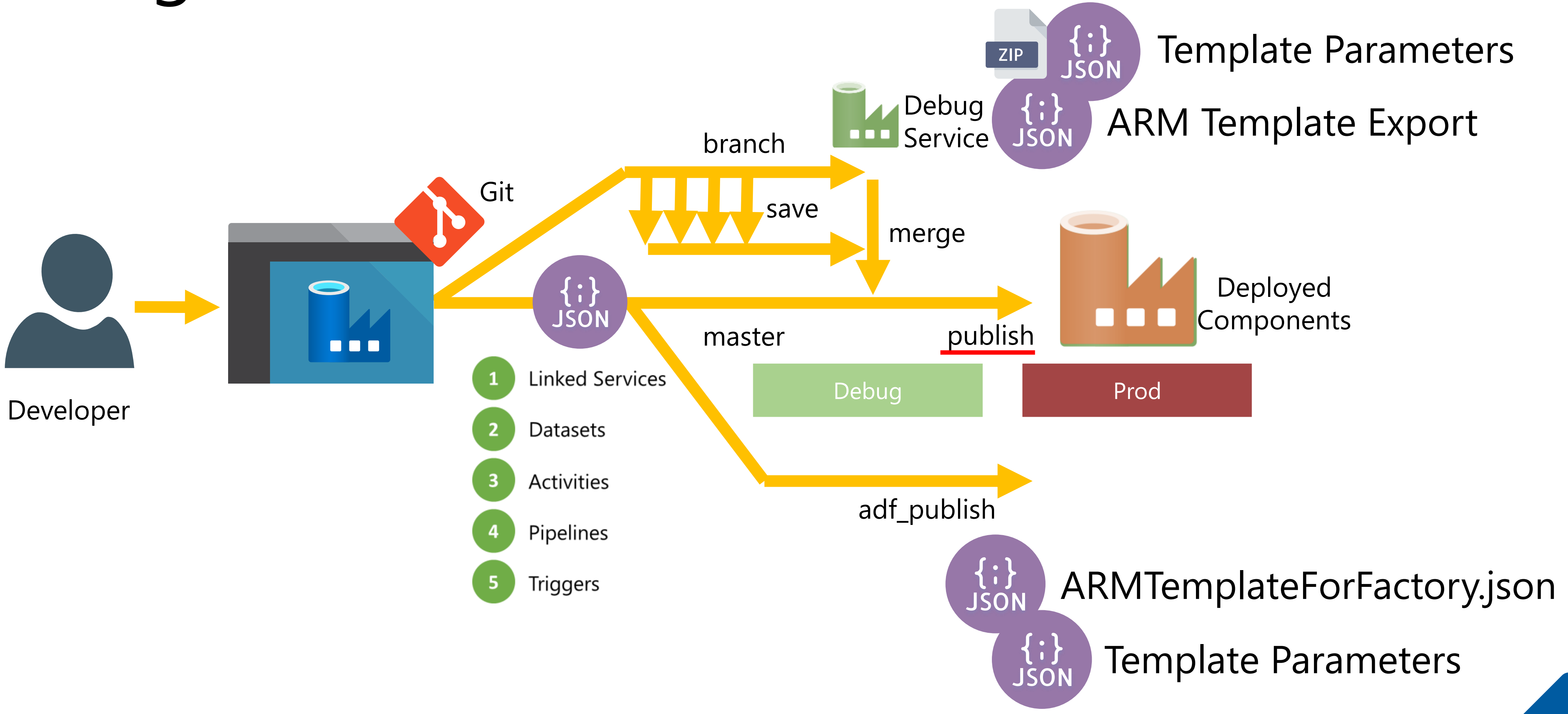
publish

Deployed
Components

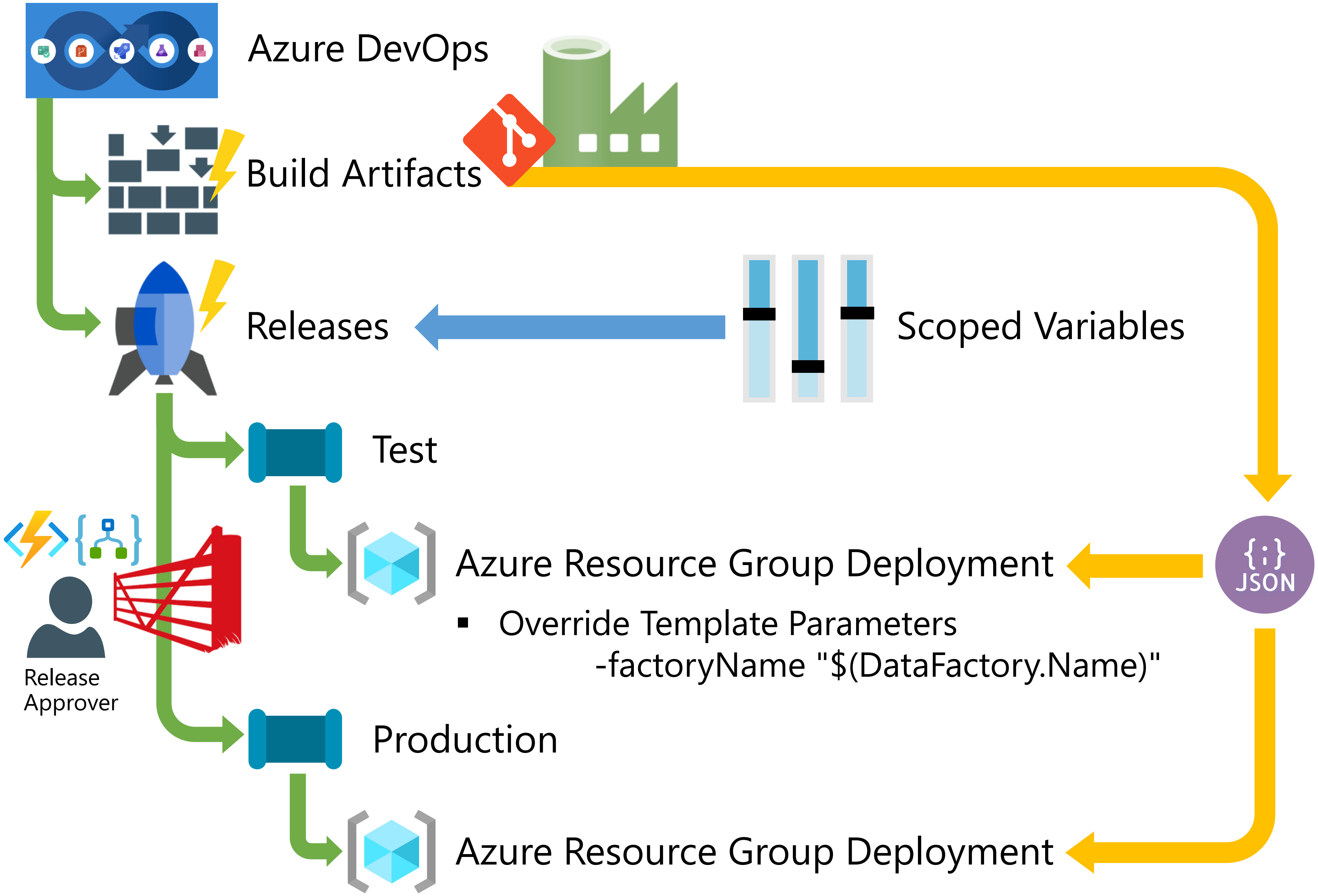
Dev

Prod

Getting Our ADF Source Code



Data Factory Continuous Delivery



Azure DevOps

Build Artifacts

Releases

Scoped Variables

Test

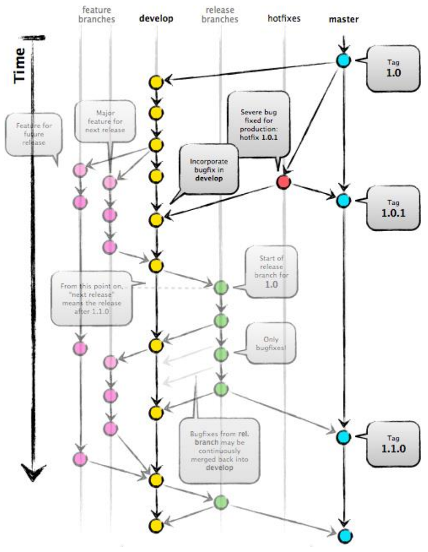
Azure Resource Group Deployment

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

ARMTemplateForFactory.json

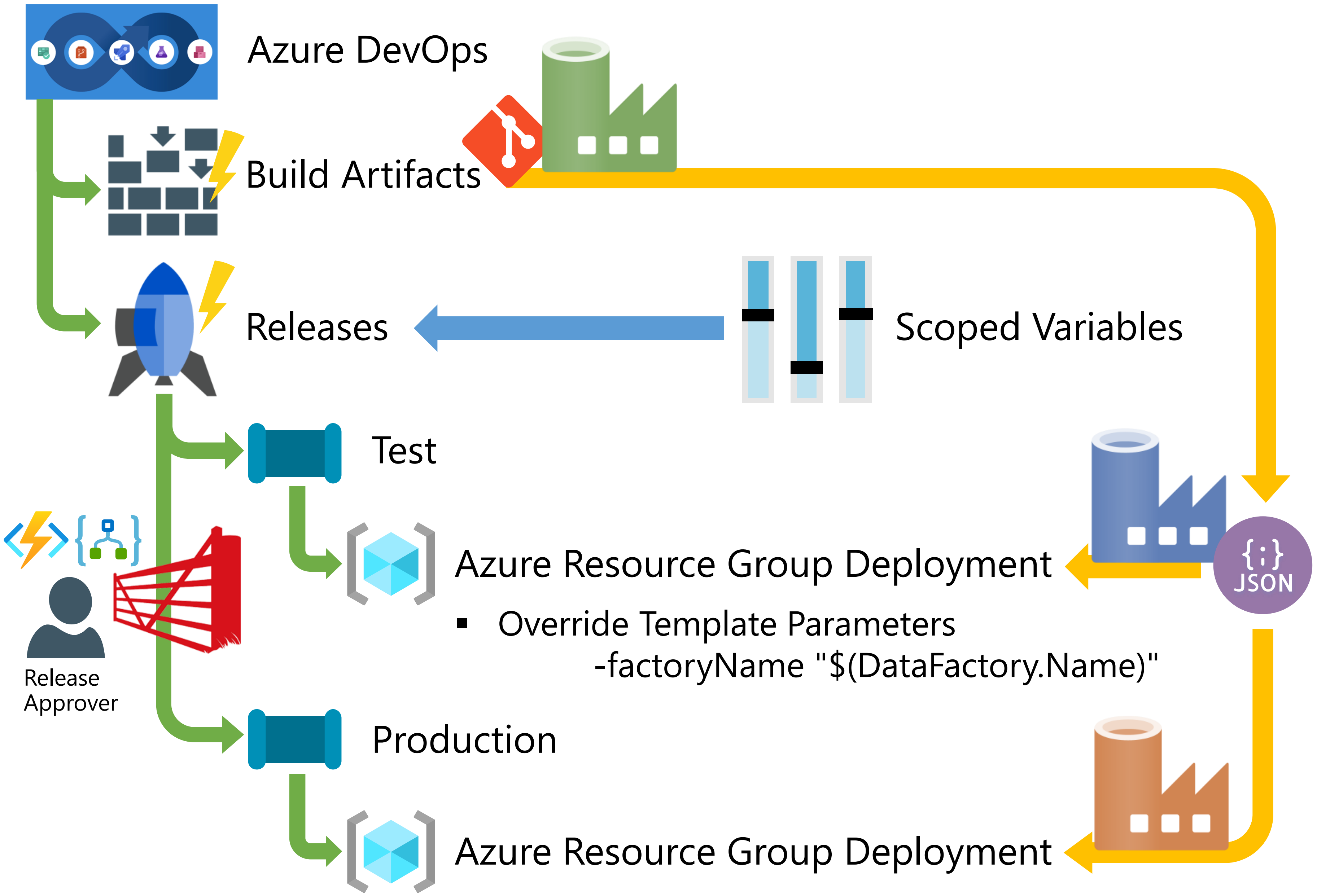
Production

Azure Resource Group Deployment


feature branches
develop
release branches
hotfixes
master
Time
Tag 1.0
Severe bug fixed for production: hotfix 1.0.1
Feature for future release
Major feature for next release
Incorporate bugfix in develop
Tag 1.0.1
Start of release branch for 1.0
From this point on, "next release" means the release after 1.1.0
Only bugfixes!
Bugfixes from rel. branch may be continuously merged back into develop
Tag 1.1.0

Gitflow

Data Factory Continuous Delivery



Azure DevOps

Build Artifacts

Releases

Scoped Variables

Test

Azure Resource Group Deployment

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

ARMTemplateForFactory.json

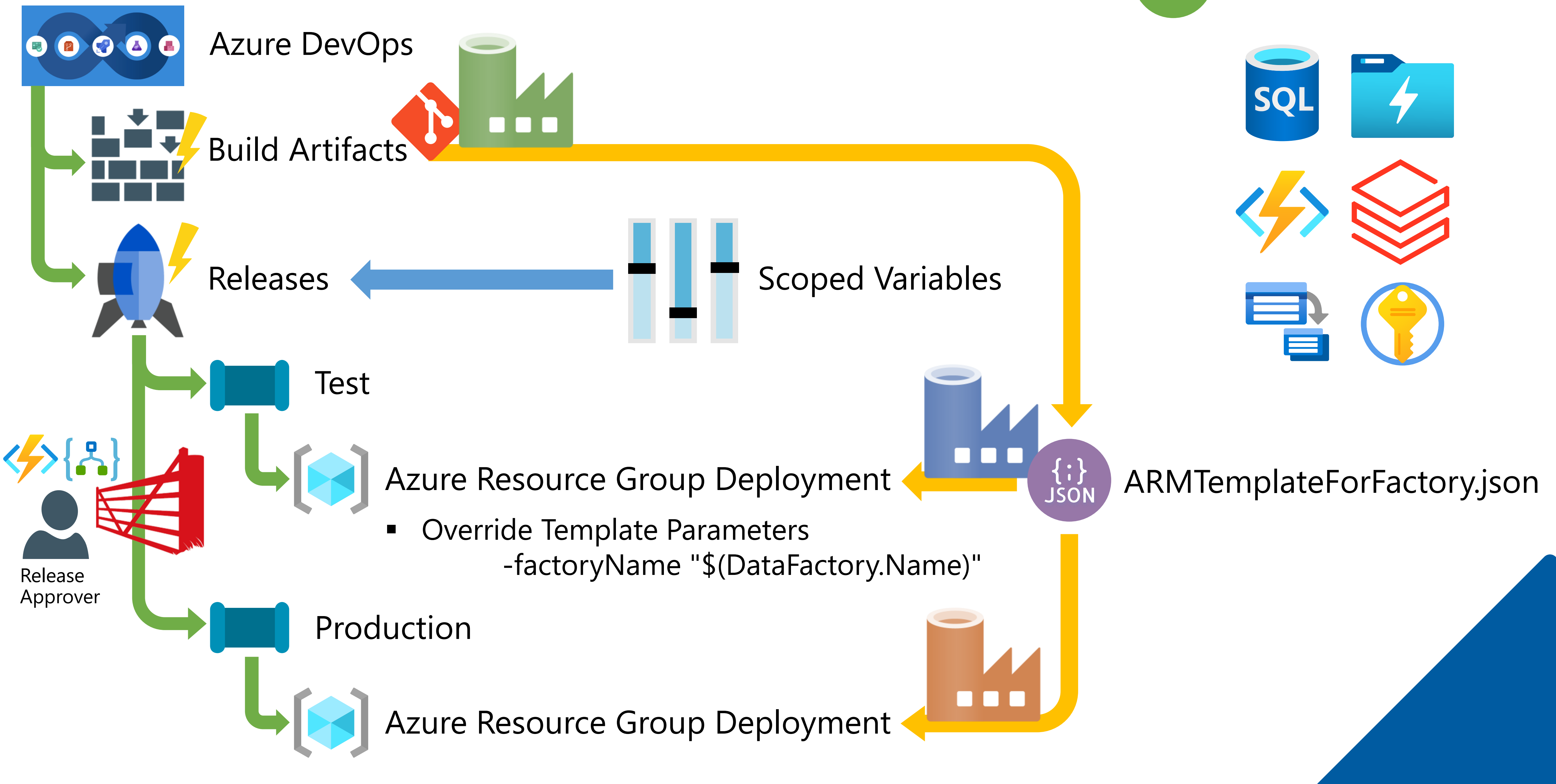
Release Approver

Production

Azure Resource Group Deployment

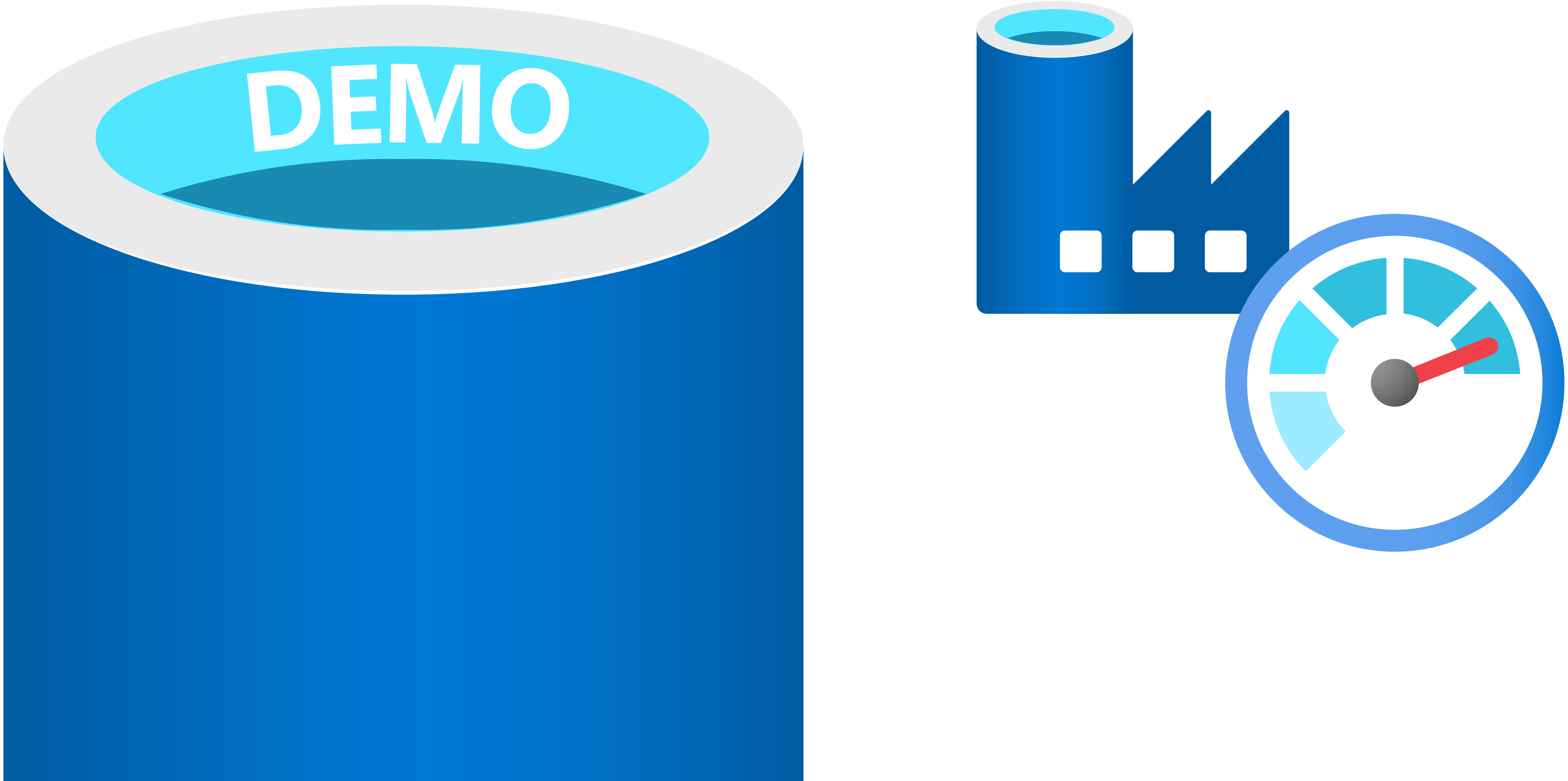
1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

Data Factory Continuous Delivery



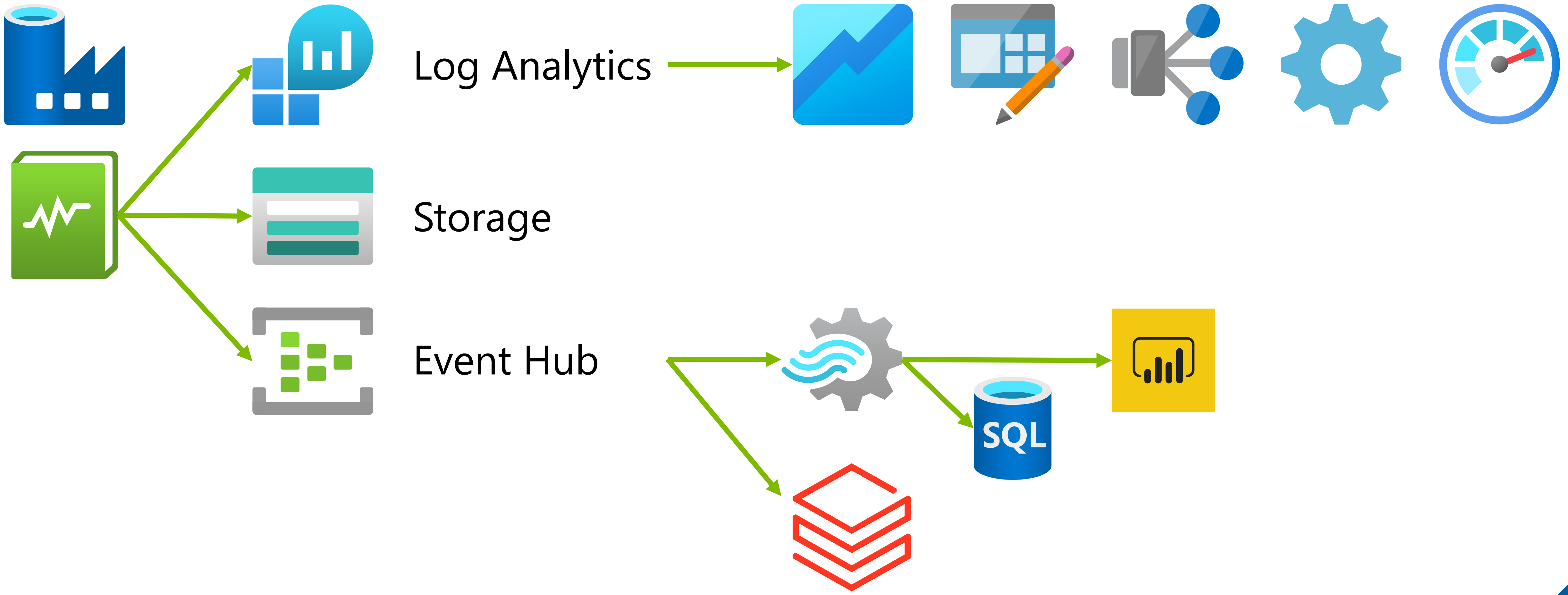
- Override Template Parameters
  -factoryName "$(DataFactory.Name)"

Monitoring & Logging



DEMO

Diagnostic Settings

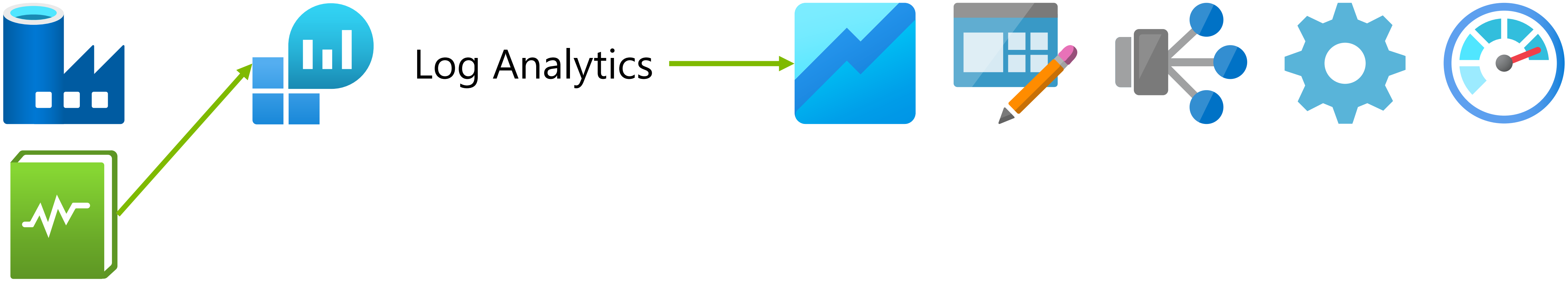


Log Analytics

Storage

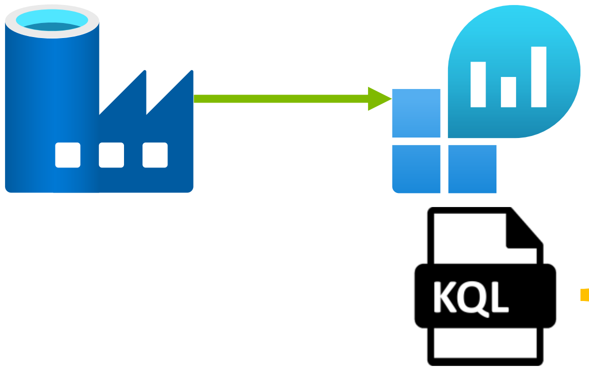
Event Hub

Diagnostic Settings

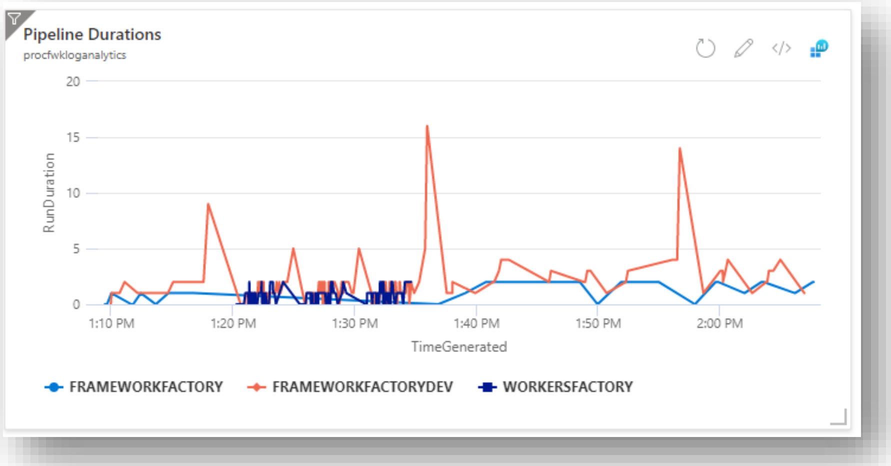


Log Analytics

Using Log Analytics

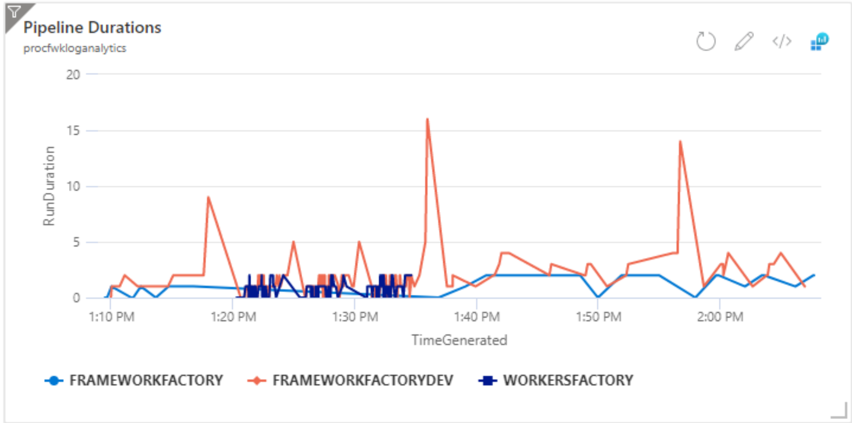


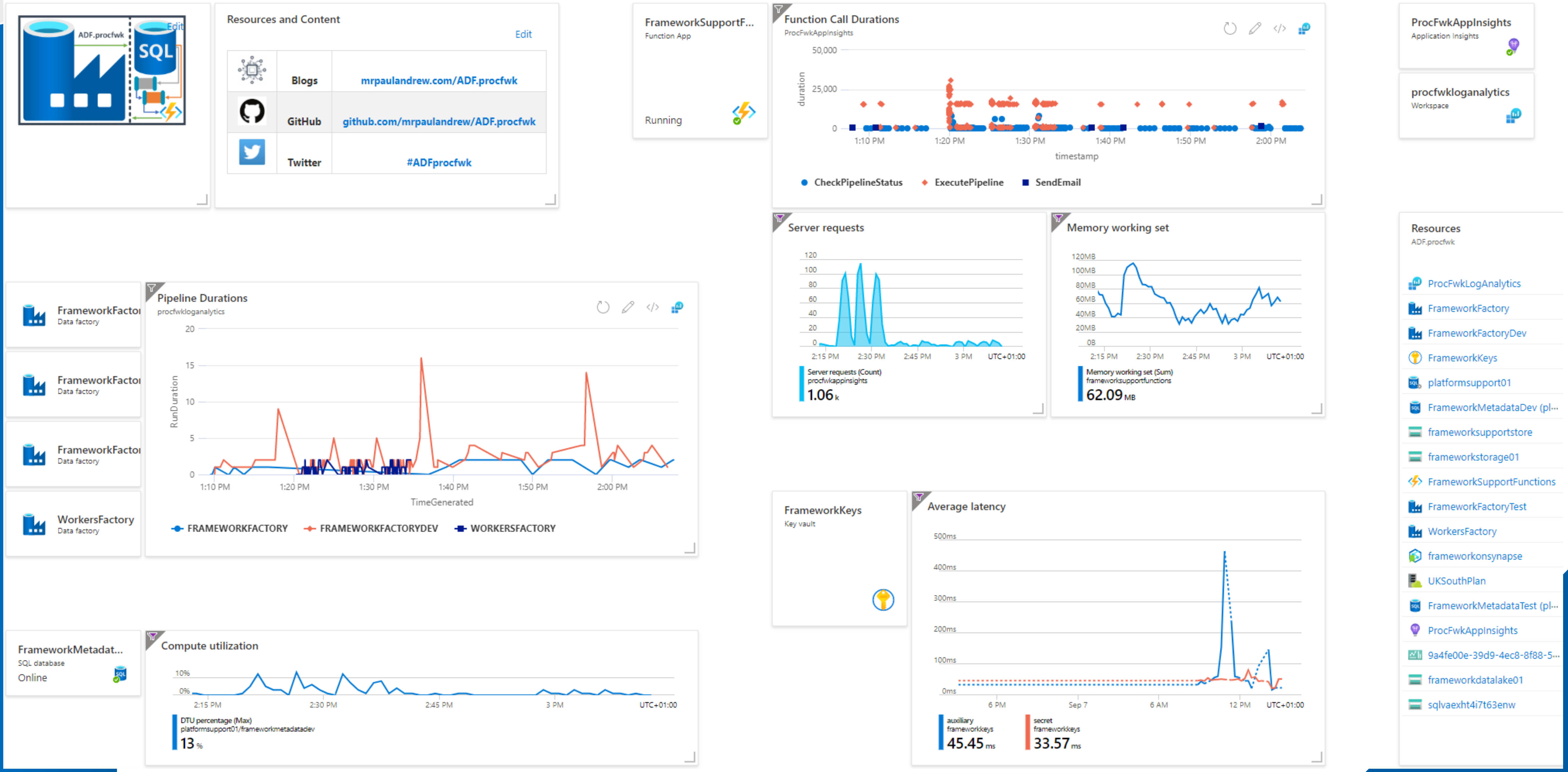
```
ADFPipelineRunDurations
 | project
         TimeGenerated,
         Start,
         End,
         ['DataFactory'] = substring(ResourceId, 121, 100),
         Status,
         PipelineName,
         Parameters,
         ["RunDuration"] = datetime_diff('Minute', End, Start)
 | where
         TimeGenerated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



Pipeline Durations

procfwkloganalytics


Pipeline Durations
procfwkloganalytics
20
15
10
5
0
RunDuration
1:10 PM
1:20 PM
1:30 PM
1:40 PM
1:50 PM
2:00 PM
TimeGenerated
FRAMEWORKFACTORY
FRAMEWORKFACTORYDEV
WORKERSFACTORY



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

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

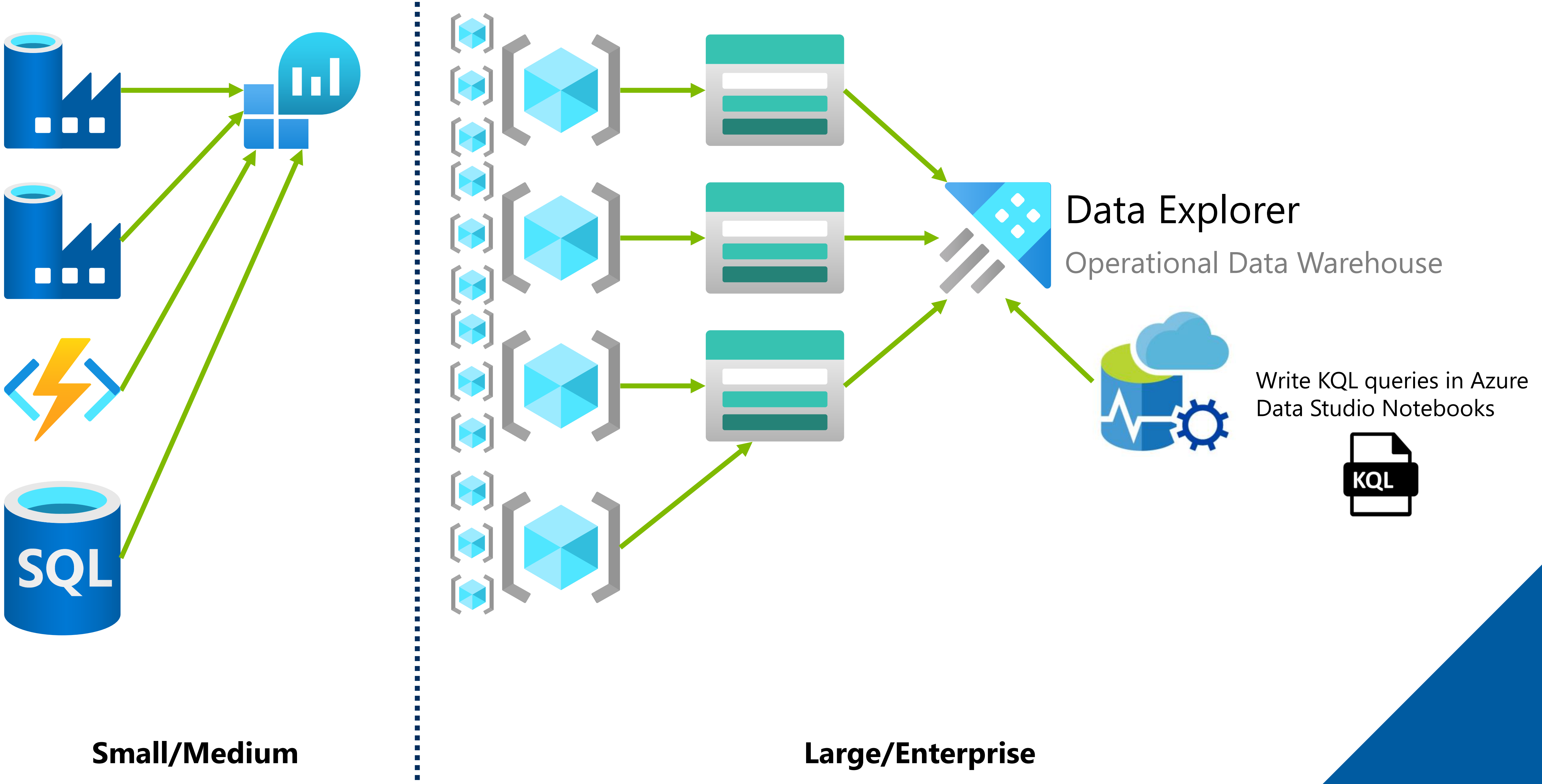
Compute utilization

DTU percentage (Max)
platformsupport01/frameworkmetadatadev
13 %

Resources
ADF.procfwk

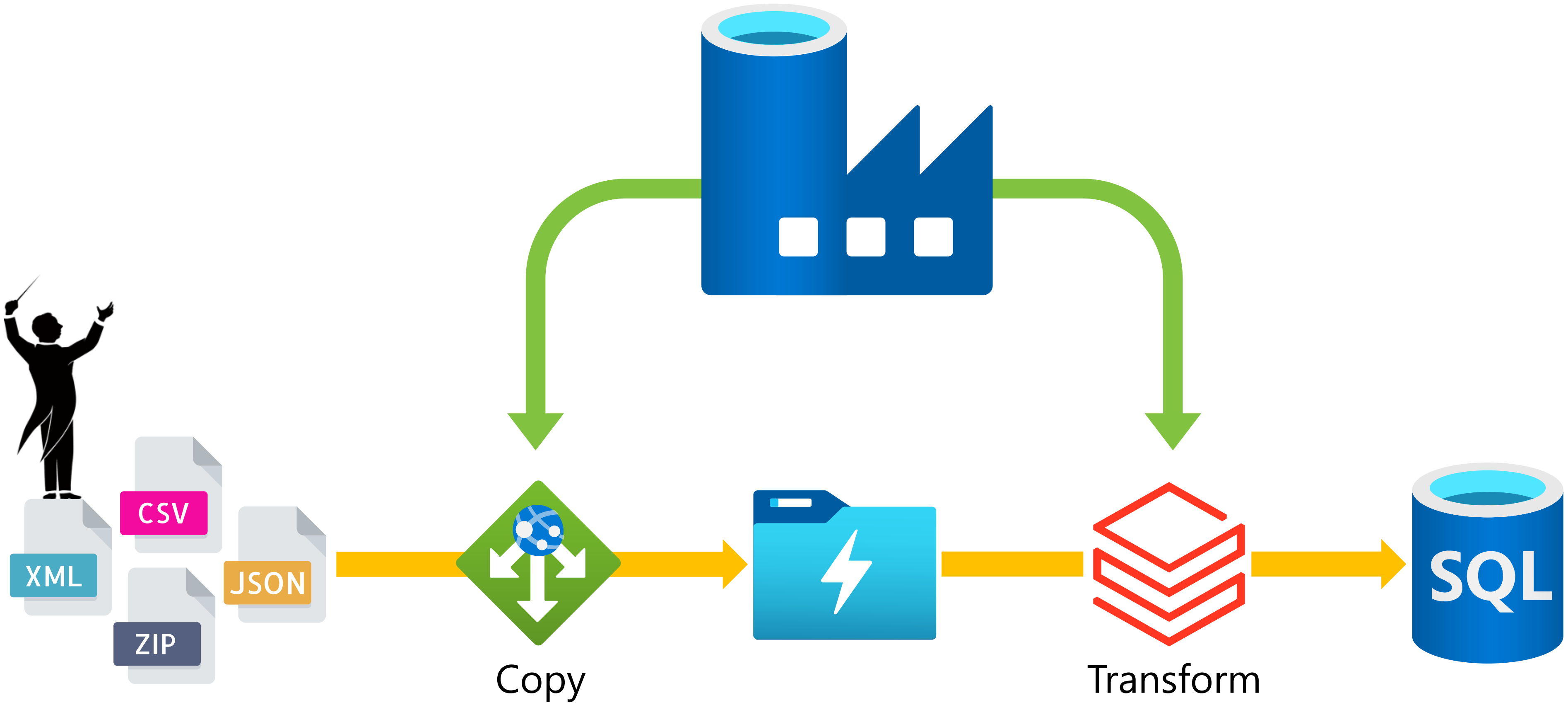
- ProcFwkLogAnalytics
- FrameworkFactory
- FrameworkFactoryDev
- FrameworkKeys
- platformsupport01
- FrameworkMetadataDev (pl...
- frameworksupportstore
- frameworkstorage01
- FrameworkSupportFunctions
- FrameworkFactoryTest
- WorkersFactory
- frameworkonsynapse
- UKSouthPlan
- FrameworkMetadataTest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

Using Data Explorer

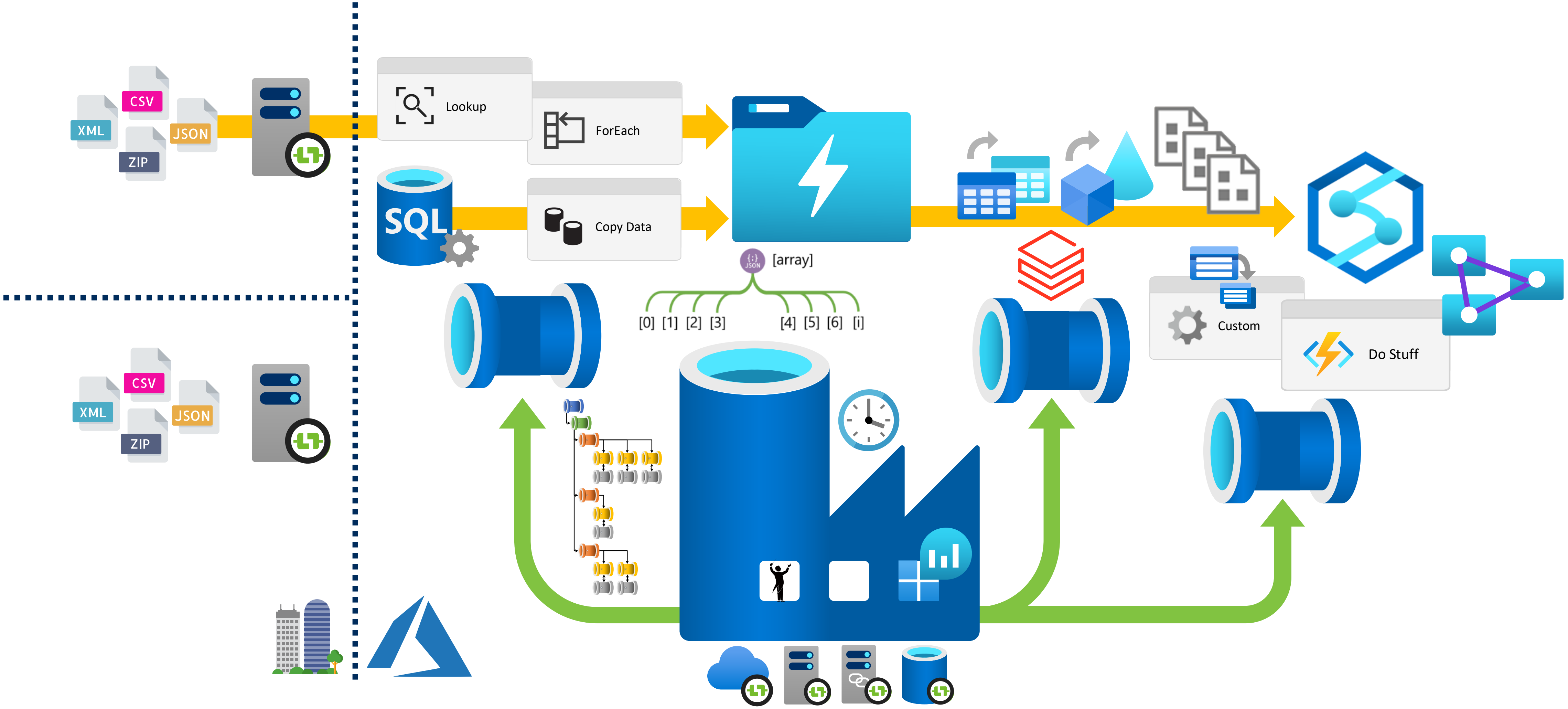

SQL
Data Explorer
Operational Data Warehouse
Write KQL queries in Azure Data Studio Notebooks
KQL
Small/Medium
Large/Enterprise

Conclusions

What is Azure Data Factory (ADF)?



What is Azure Data Factory?



1. A complete Microsoft Azure integration tool.
2. Orchestrator of our <u>Control Flow</u> operations – with scale out Activities.
3. Orchestrator of our <u>Data Flow</u> 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

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

Thank you for listening…

Paul Andrew MVP Microsoft® Most Valuable Professional altius

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