

Azure Data Integration Pipelines

In Production

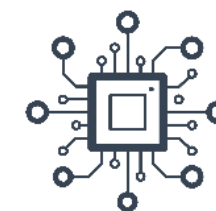
Paul Andrew | Technical Architect in Azure CoE



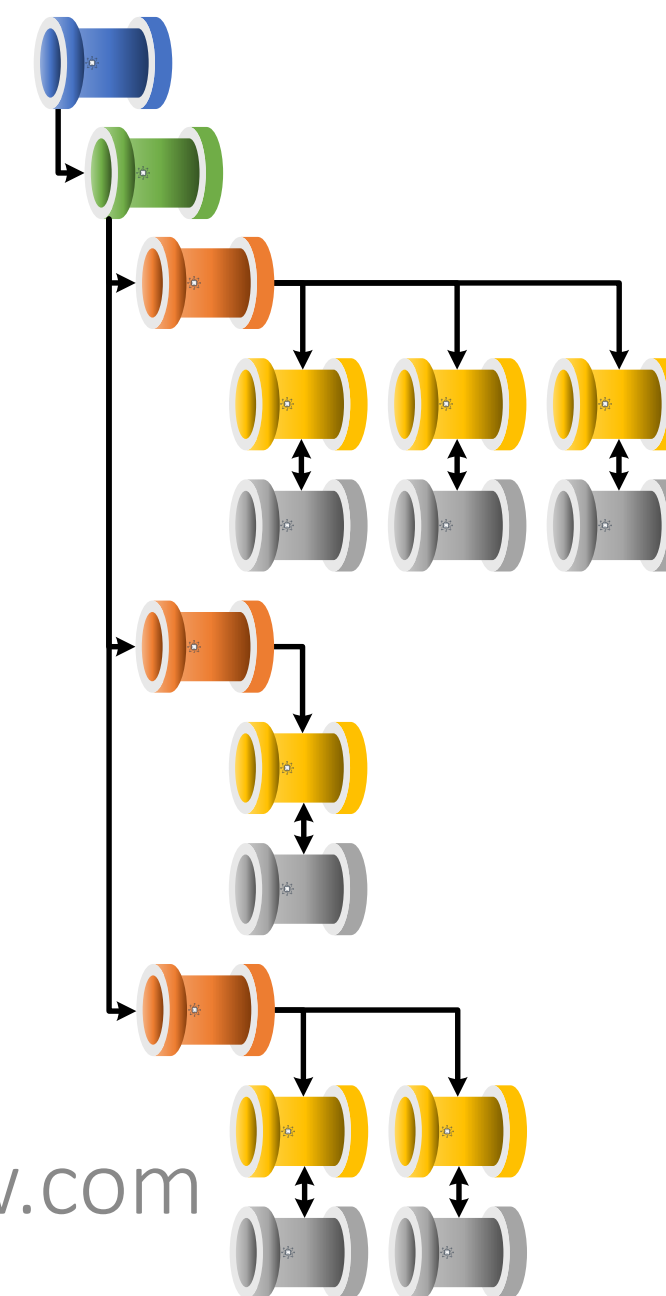
@MrPaulAndrew

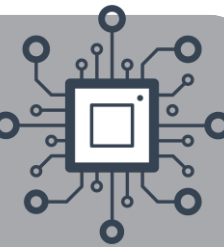


In/MrPaulAndrew



MrPaulAndrew.com





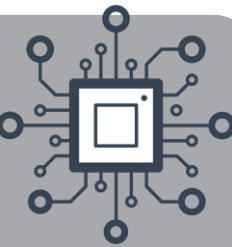
<https://github.com/mrpaulandrew>

CommunityEvents

Demo code, content and slides from various community events.

● C++

[{Event/Location}-{Month}-{Year}](#)



☐☐ Data Integration Pipelines – A Quick Overview

☐☐ Scaled Out Design Patterns

☐☐ Metadata Driven Framework

☐☐ Testing

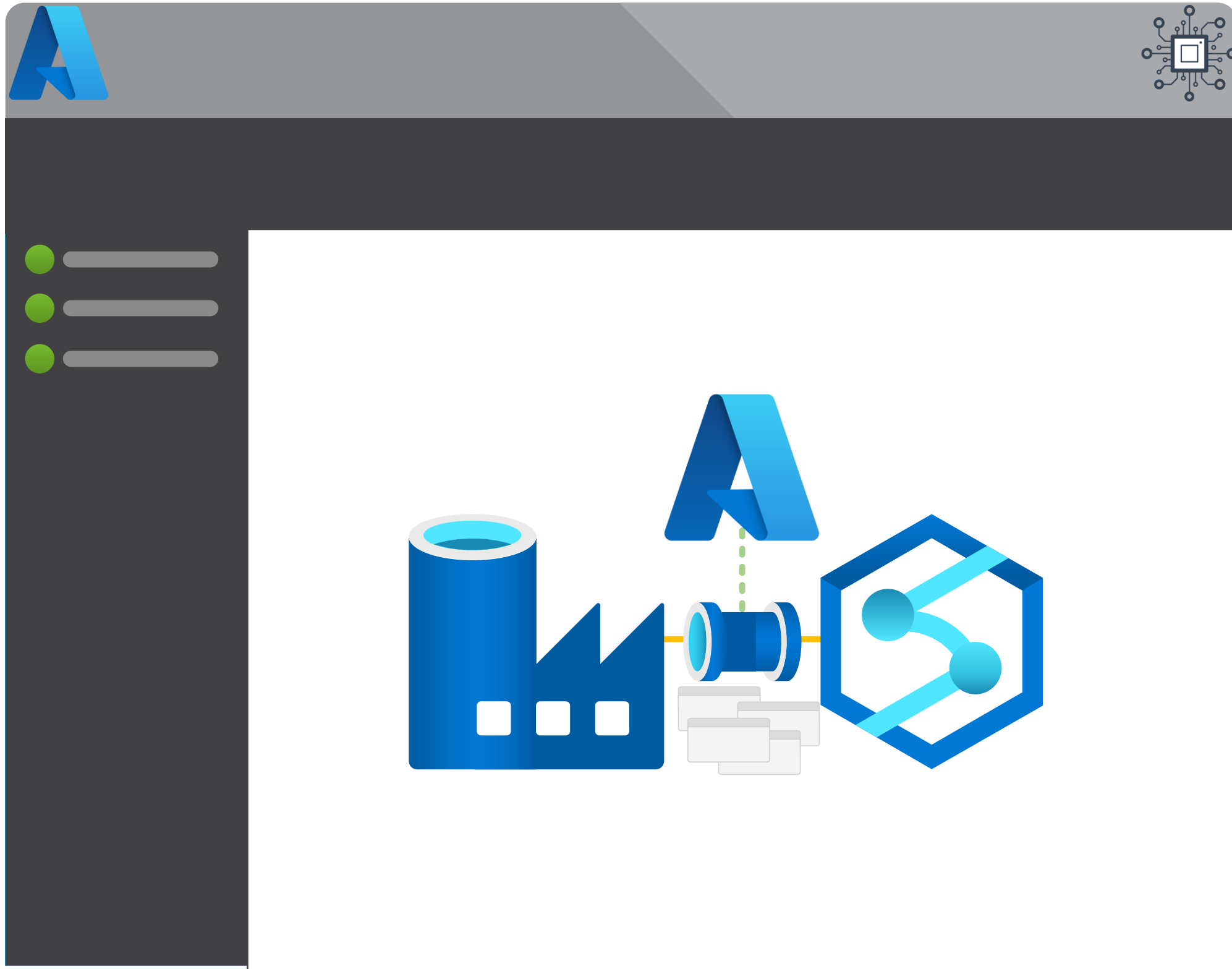
☐☐ Source Code – A Quick Overview

☐☐ Enterprise Deployments

☐☐ VNet Integration

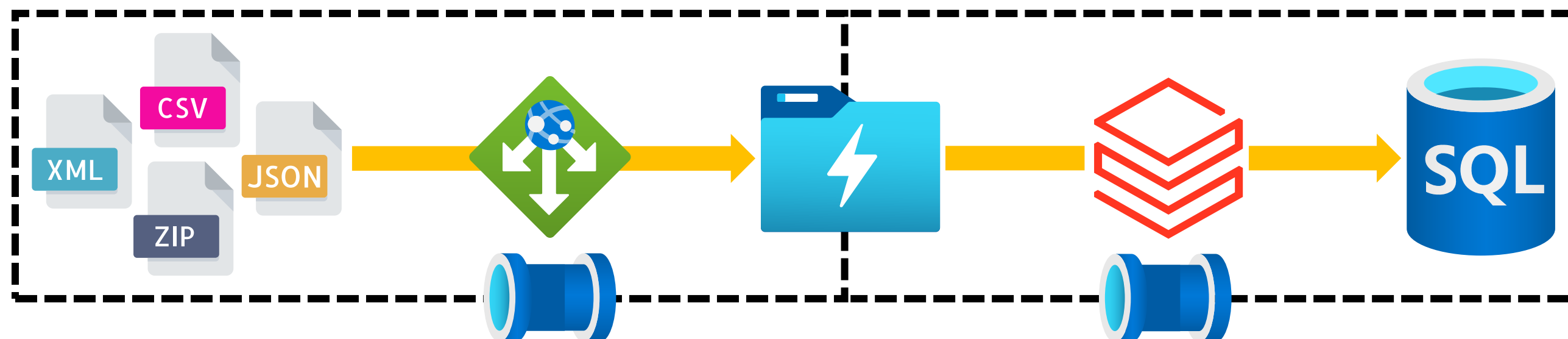
☐☐ Best Practice

Data Integration Pipelines – A Quick Overview

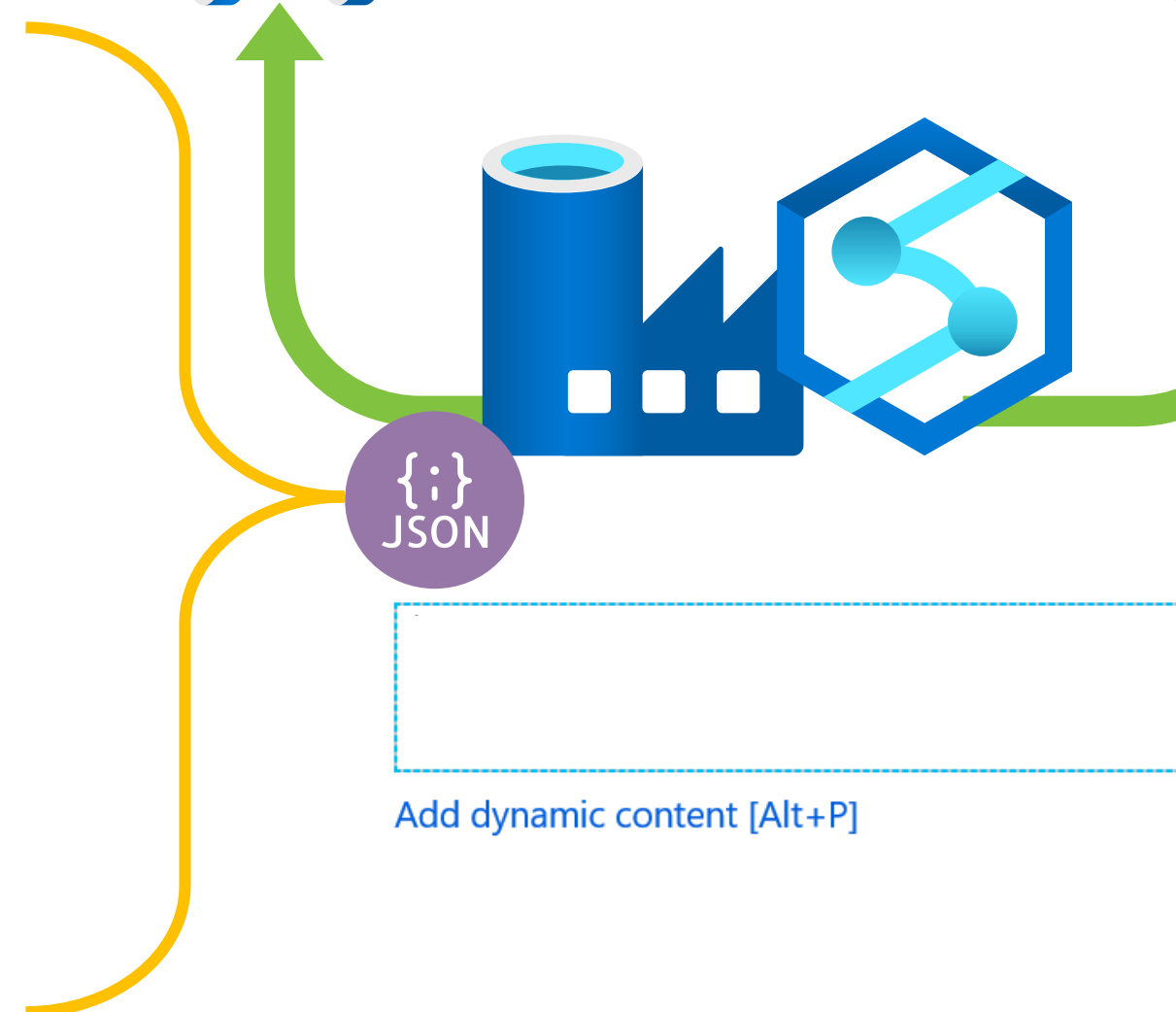




Control Flow



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



Integration Runtimes

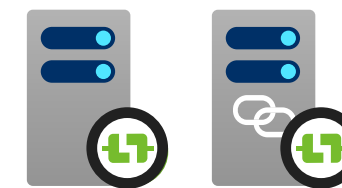
6



Azure IR

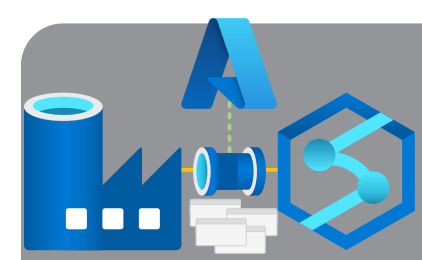


Hosted IR

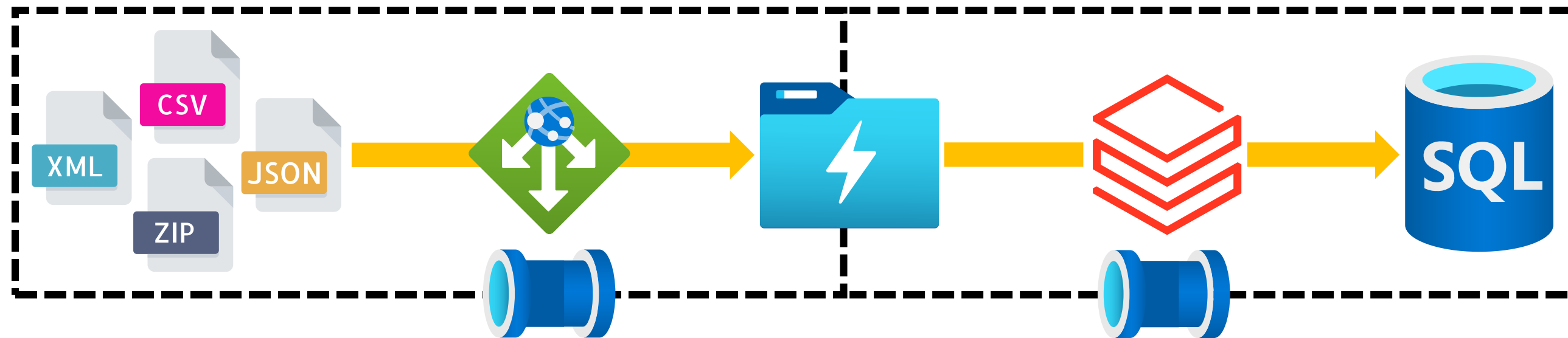
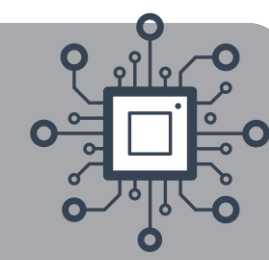


SSIS IR

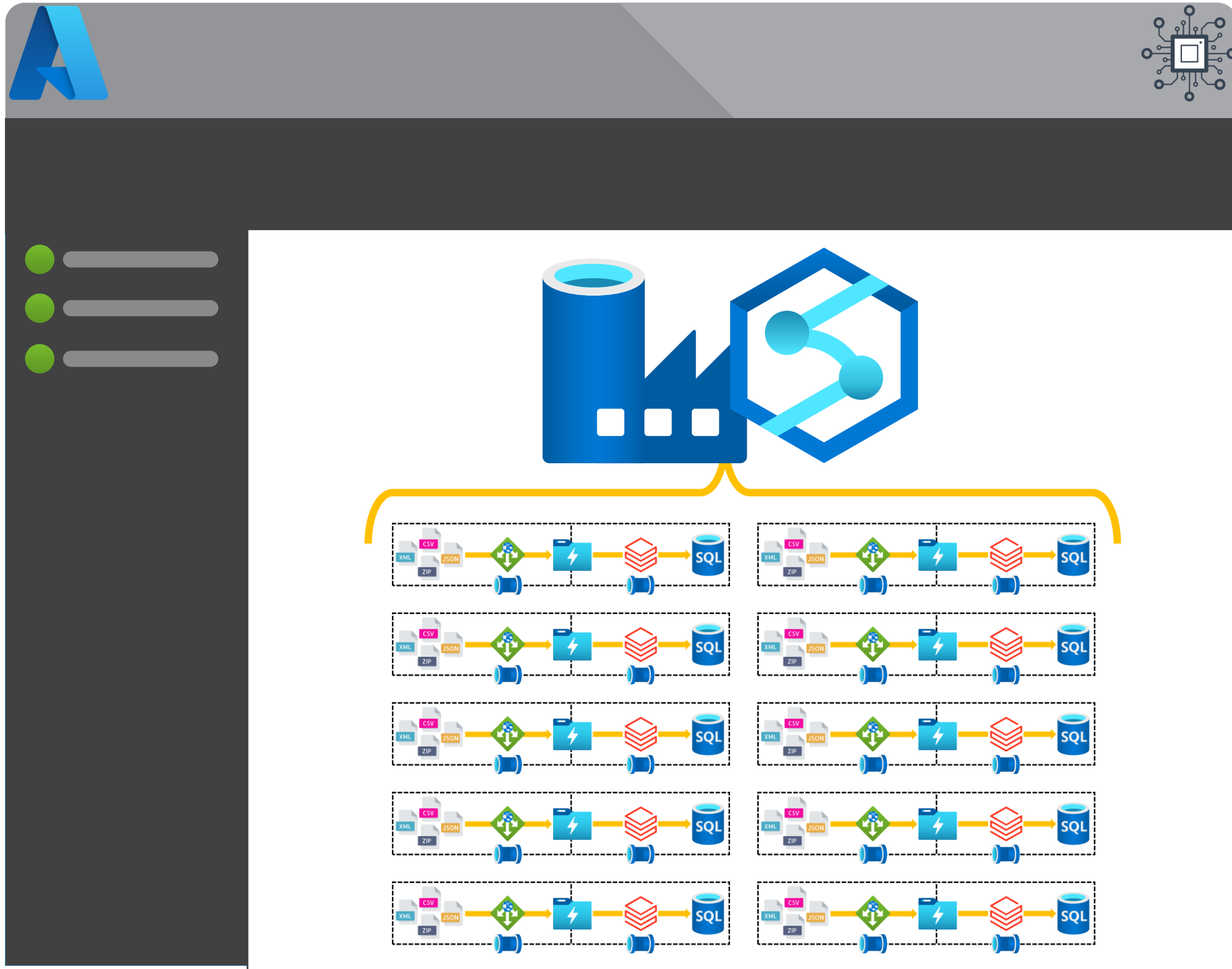


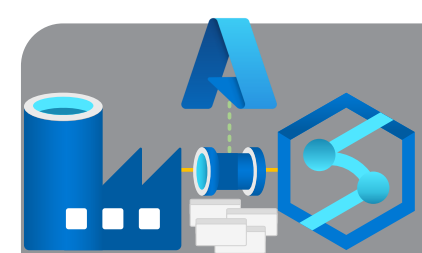


Integration Pipelines as Data Engineers

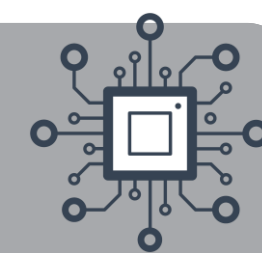


Scaled Out Design Patterns

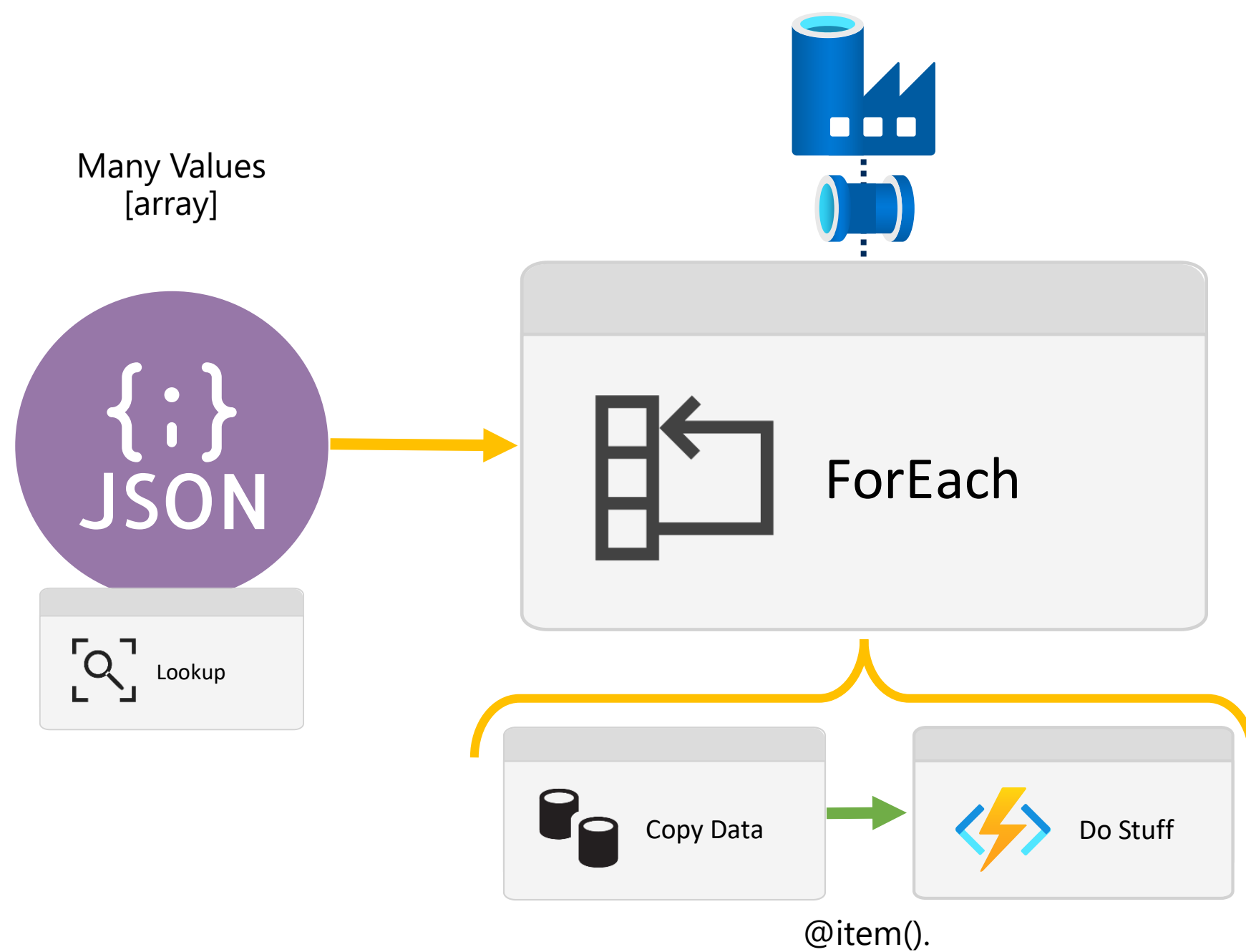




For Each Activity



Scaling Out Control Flow Activities



IsSequential: true



[array]

[0]

[1]

[2]

[3]

[i]



[array]

[0]

[1]

[2]

[3]

[4]

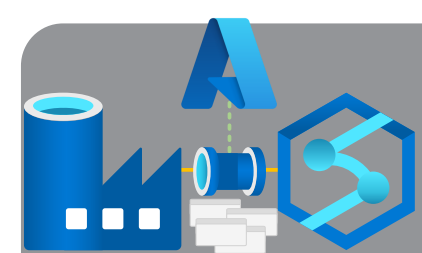
[5]

[6]

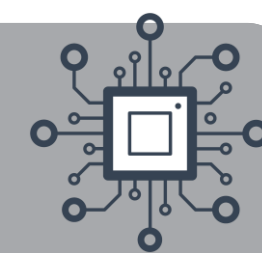
[i]

Batch Count Default: 20

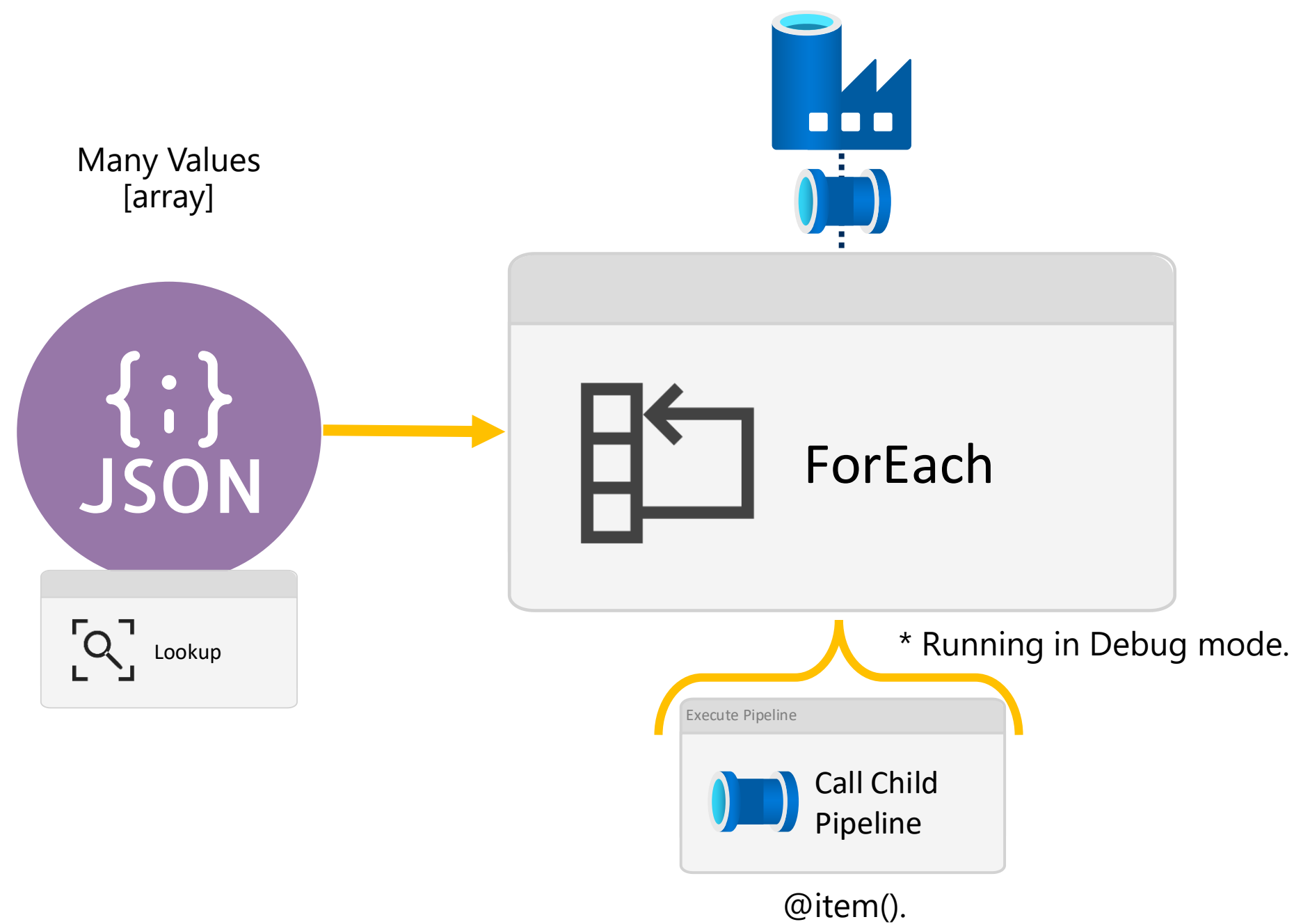
Batch Count Max: 50



For Each Activity



Scaling Out Control Flow Activities



IsSequential: true



[array]

[0]

[1]

[2]

[3]

[i]



[array]

[0]

[1]

[2]

[3]

[4]

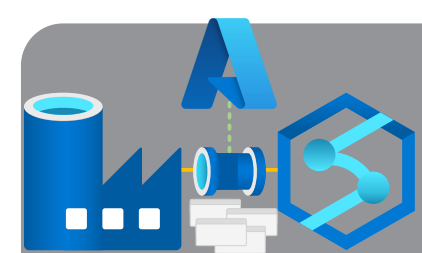
[5]

[6]

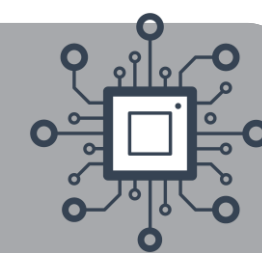
[i]

Batch Count Default: 20

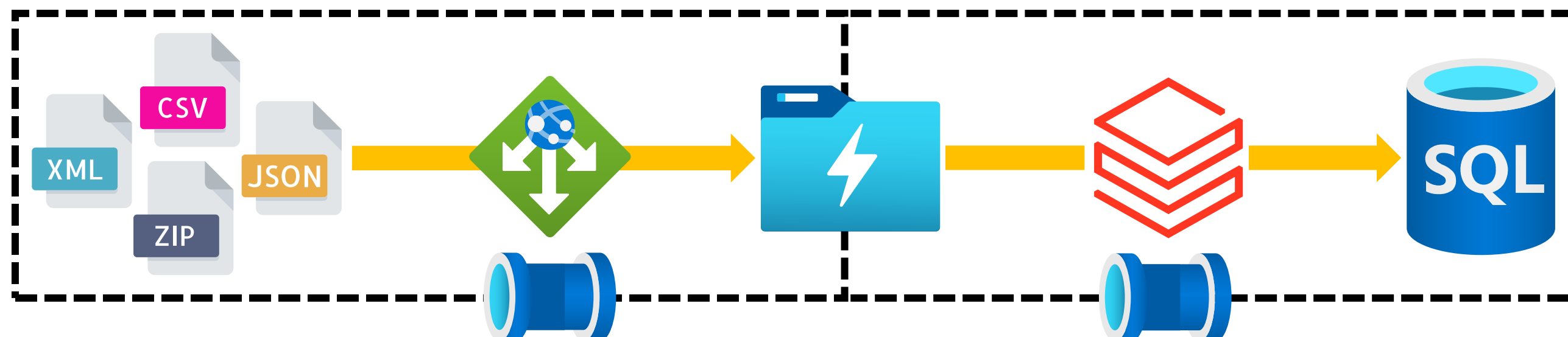
Batch Count Max: 50



Integration Pipelines as Data Engineers



Control Flow



1

Linked Services



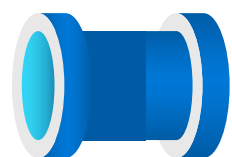
2

Datasets



3

Activities



4

Pipelines



5

Triggers



Add dynamic content [Alt+P]

Integration Runtimes

6



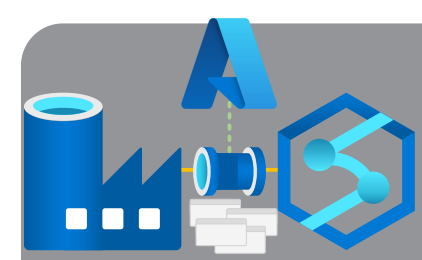
Azure IR



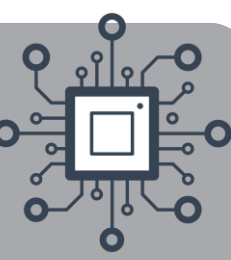
Hosted IR

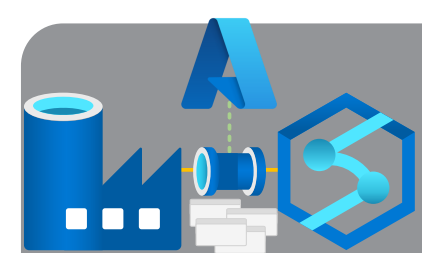


SSIS IR

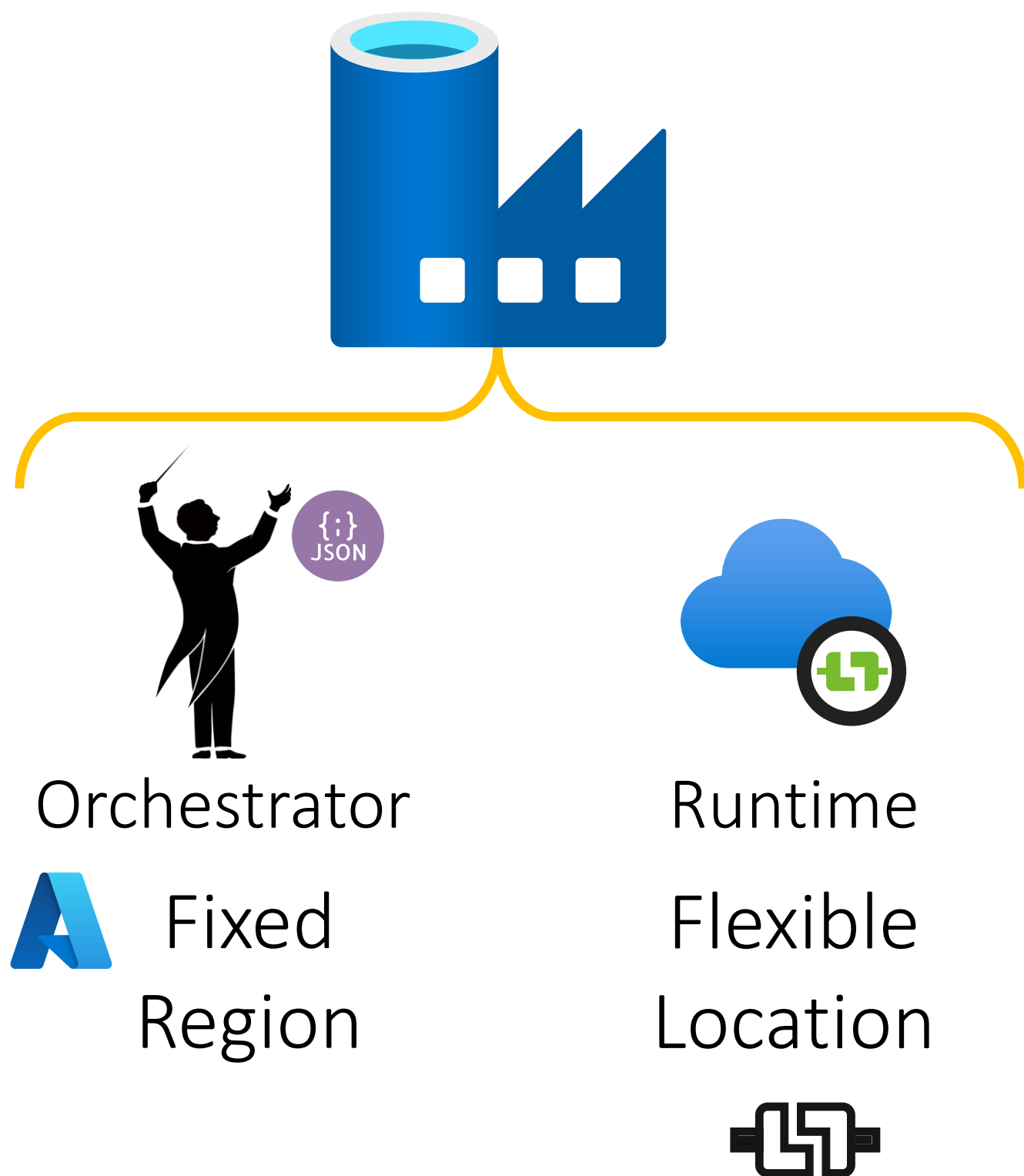
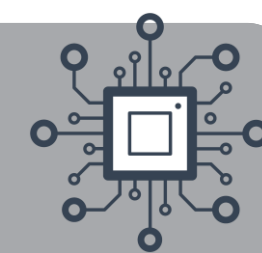


Azure Integration Runtime

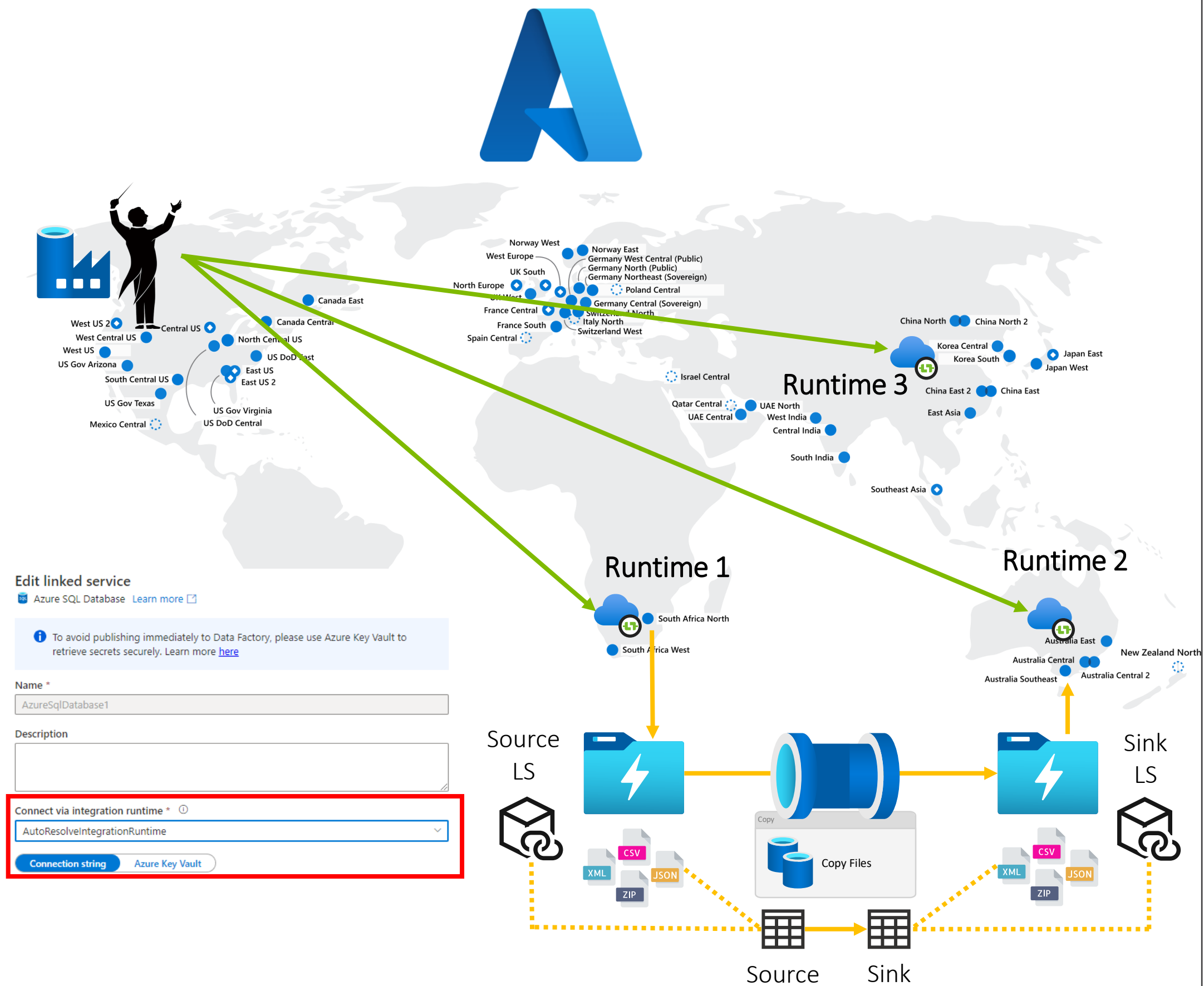


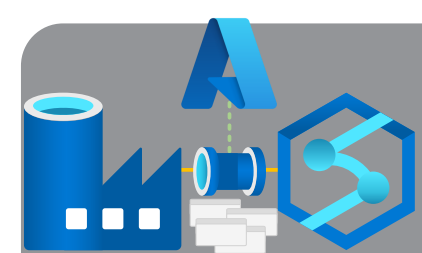


Azure Integration Runtime



AutoResolveIntegrationRuntime

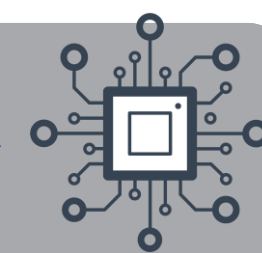


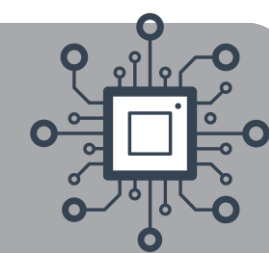
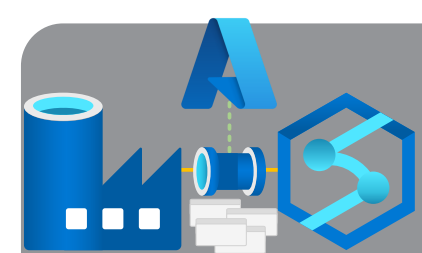


Azure Integration Runtime

Internal vs External Activities

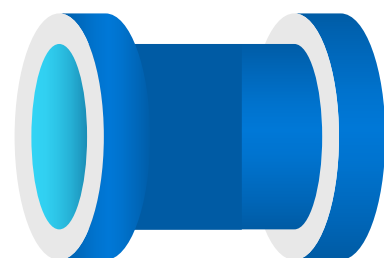
<https://mrpaulandrew.com/2020/12/22/pipelines-understanding-internal-vs-external-activities/>





Concurrency – Pipelines vs Activities

Per Subscription, per IR Region



10,000

Internal

1,000

External

3,000

Example 1



1
IR

1
Pipeline

1
ForEach

+

50
Batches

×

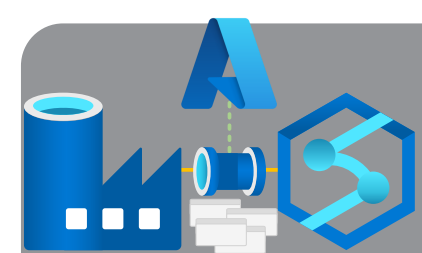
39
Wait Activities

=

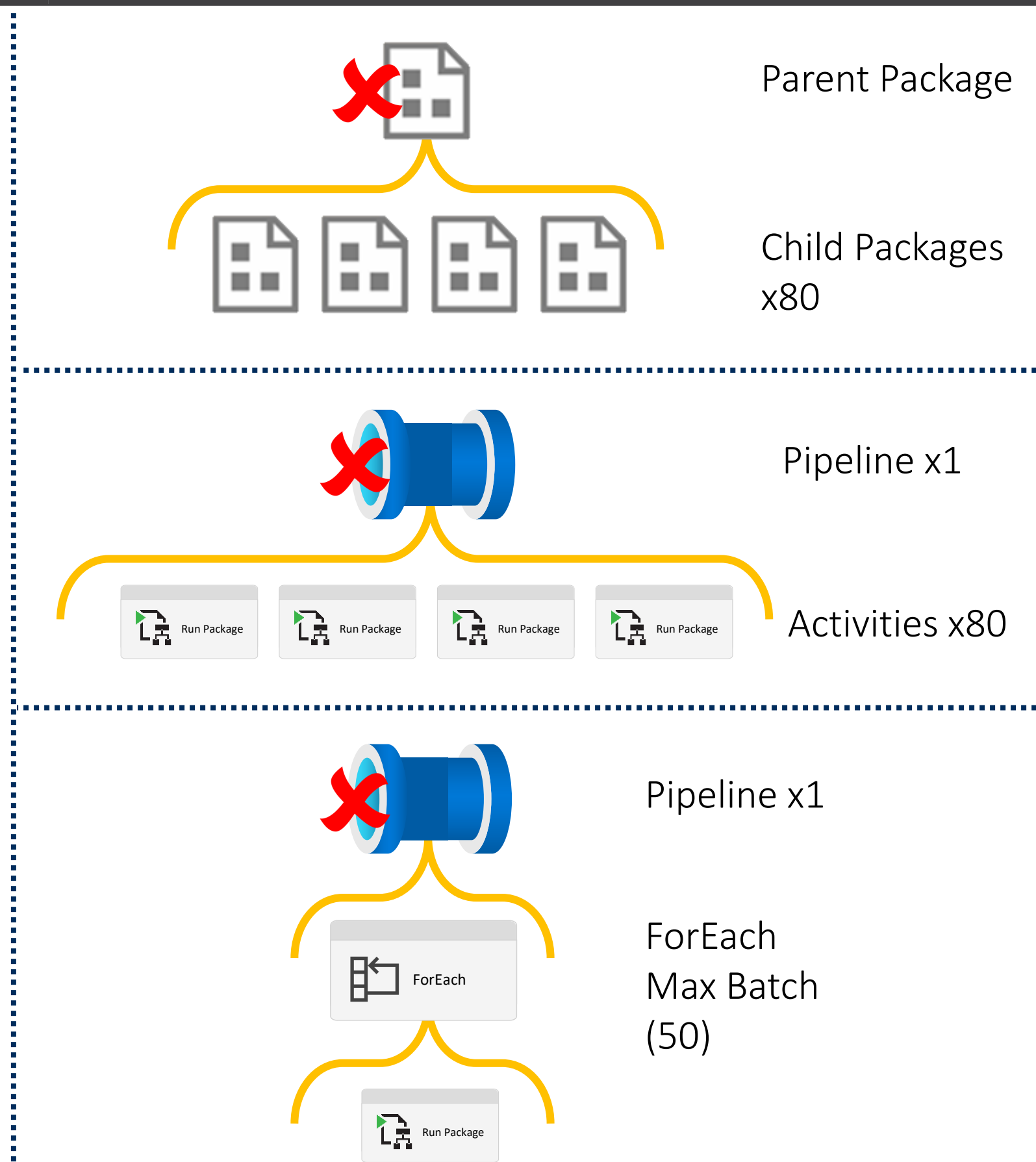
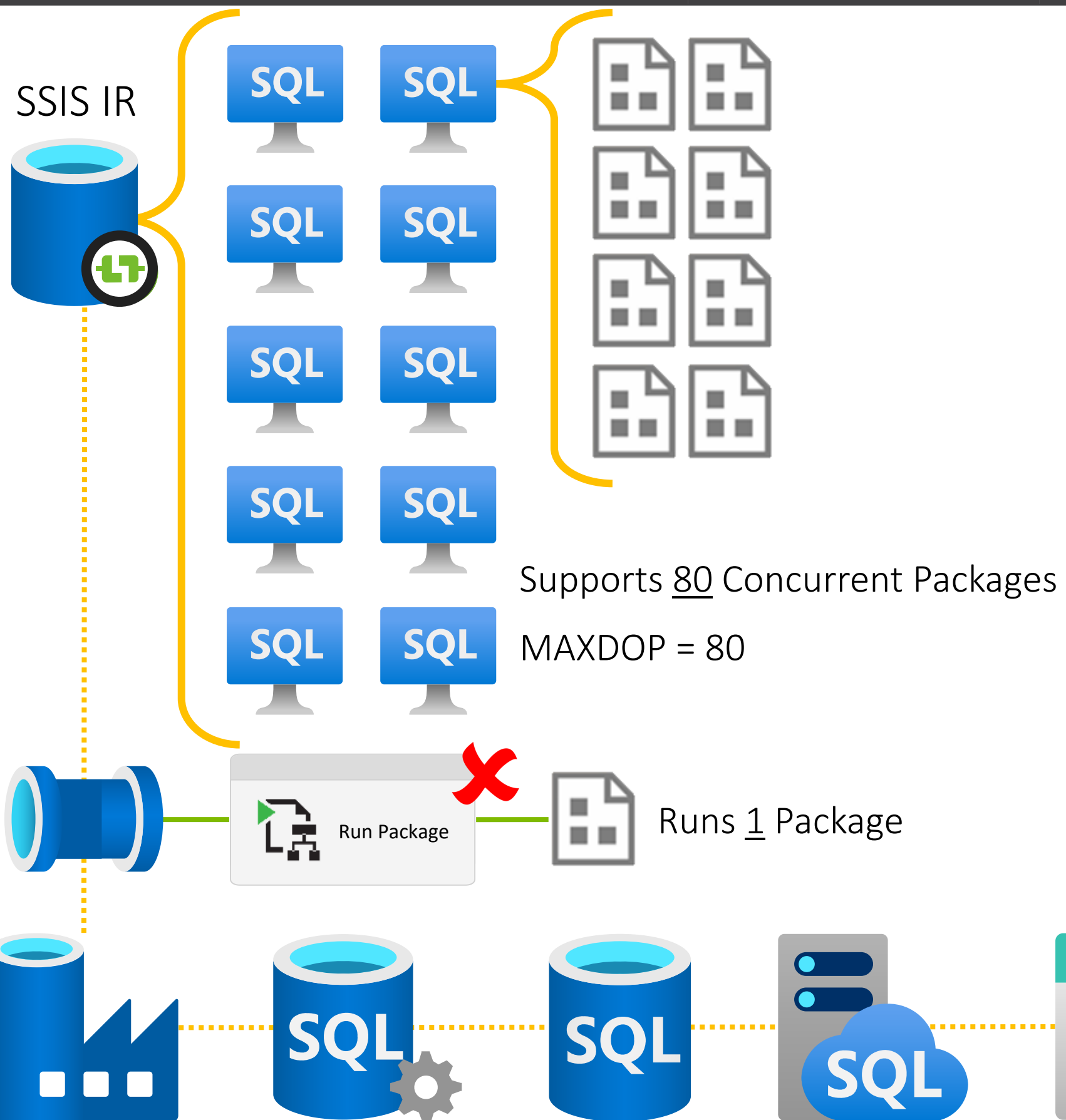
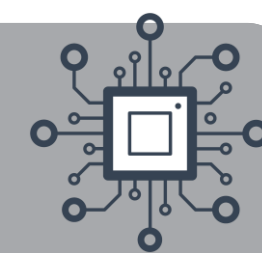
1951
Concurrent Activities

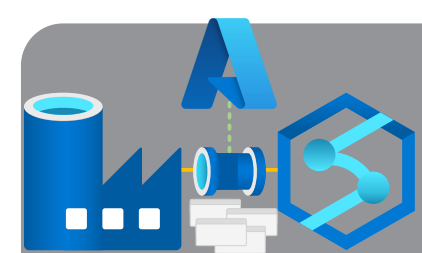


DEMO

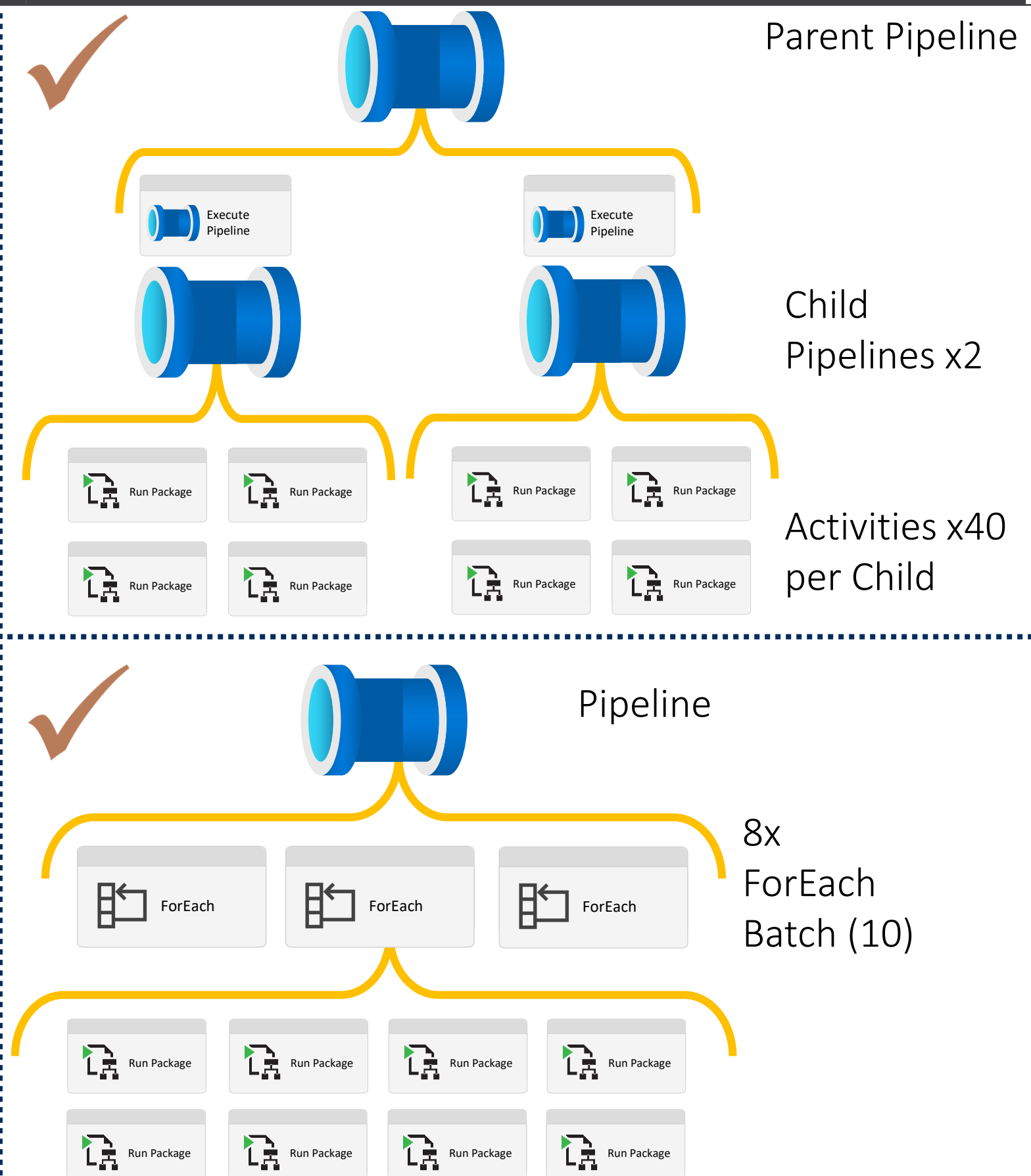
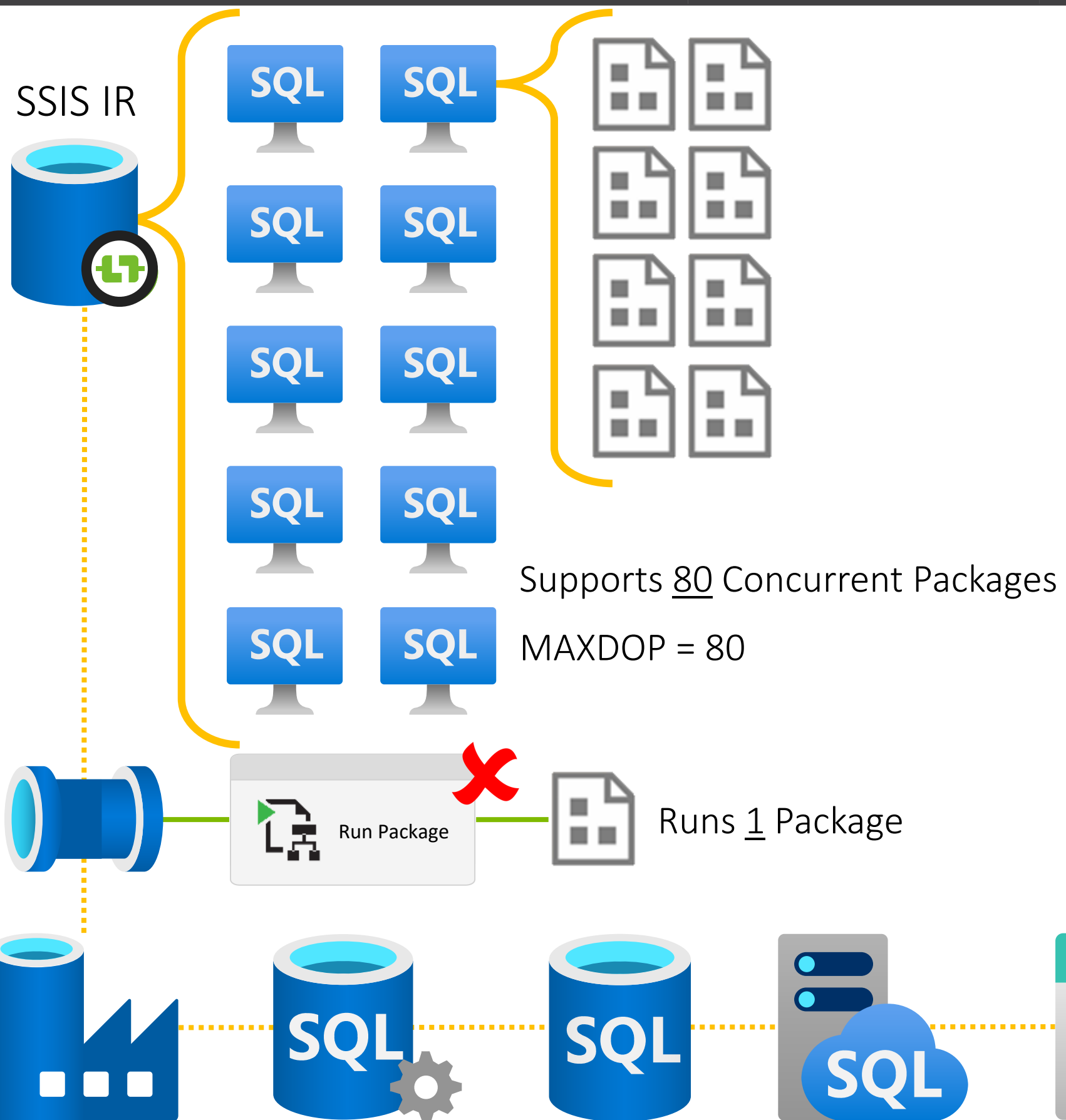
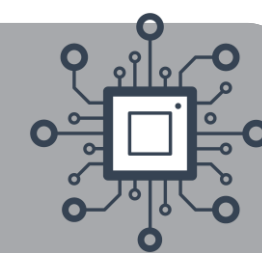


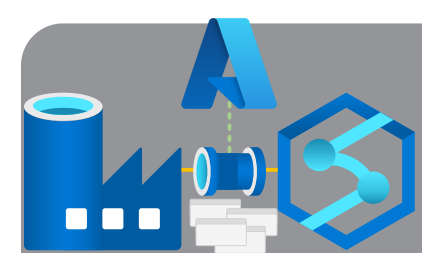
Problem: Using All Of The SSIS IR Compute



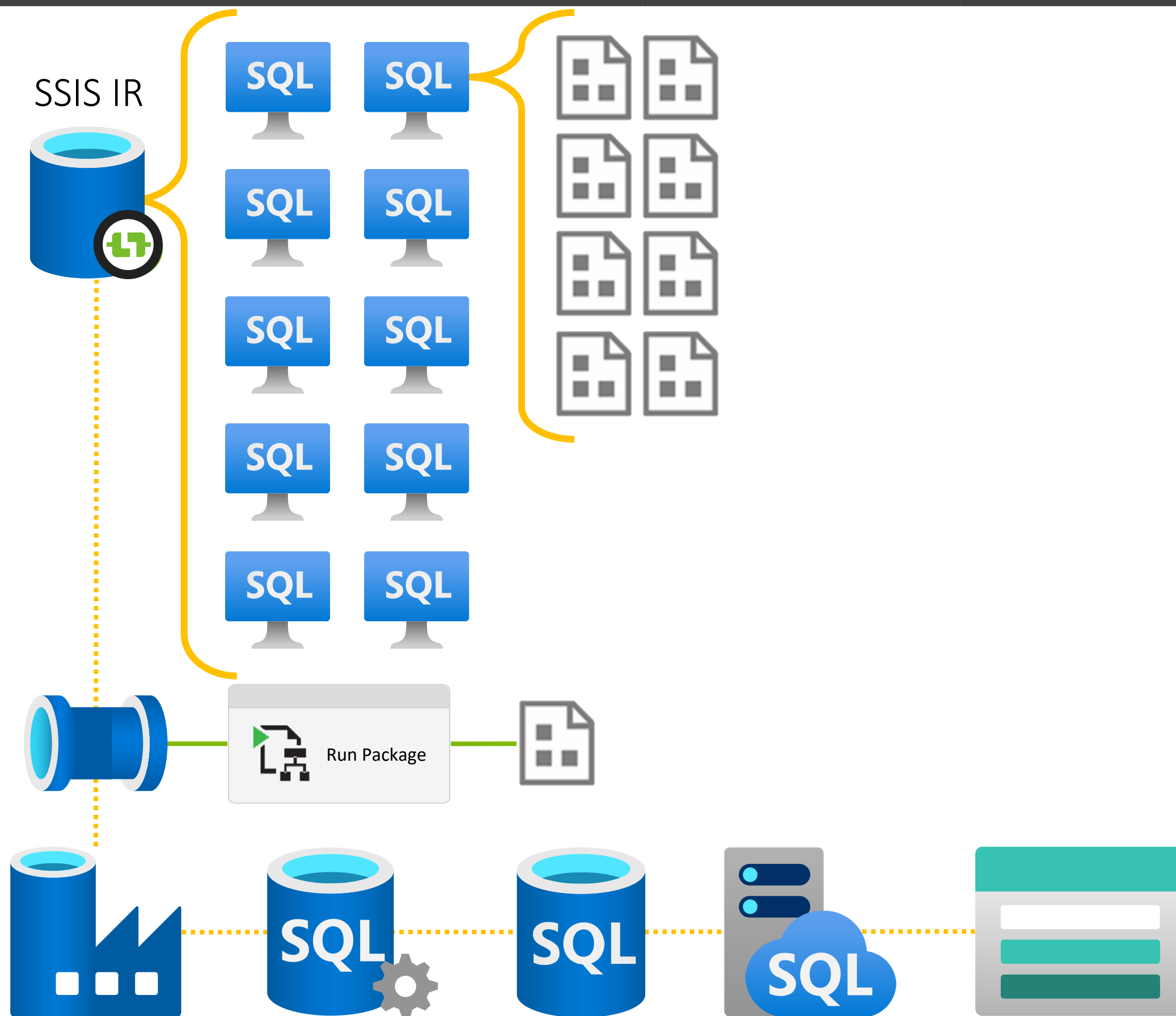
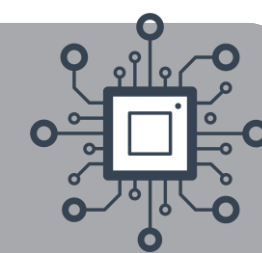


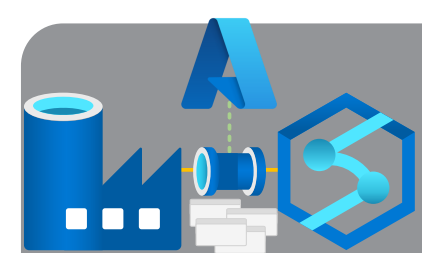
Solution 1 & 2: Static Pipelines



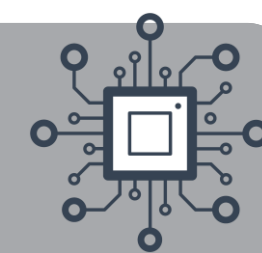


Solution 1 & 2: Static Pipelines

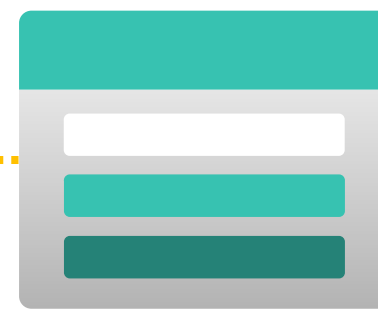
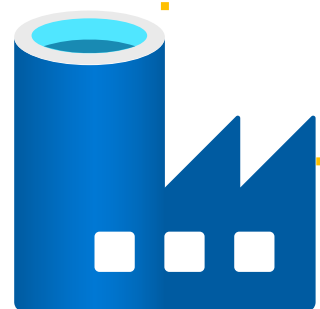
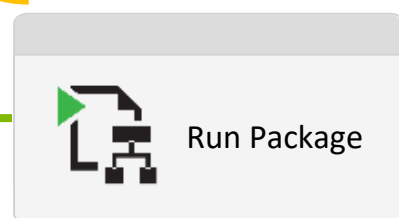
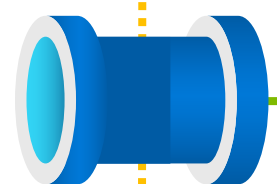
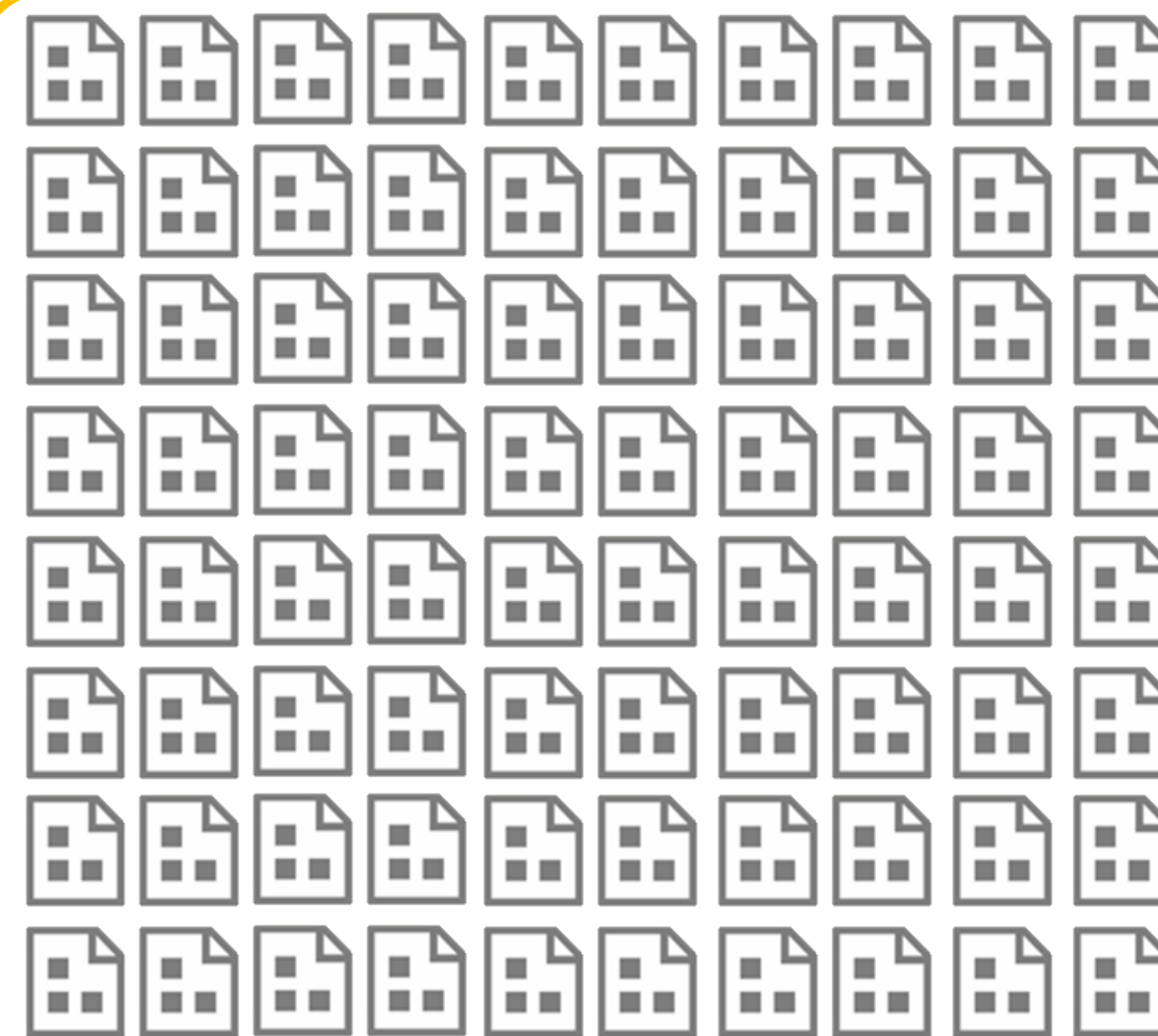
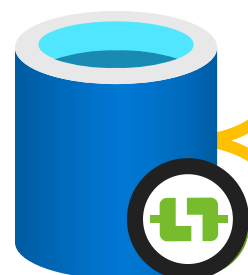




Solution 3: Packages Refactored on a Single Node IR



SSIS IR

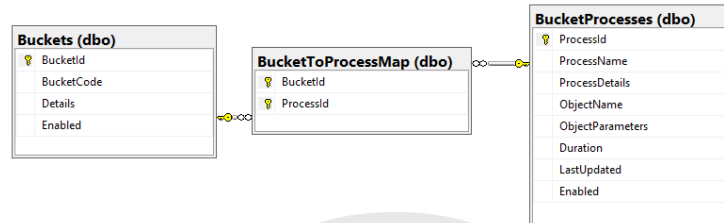


Solution 4: Nested ForEach Activities & Bucket Metadata

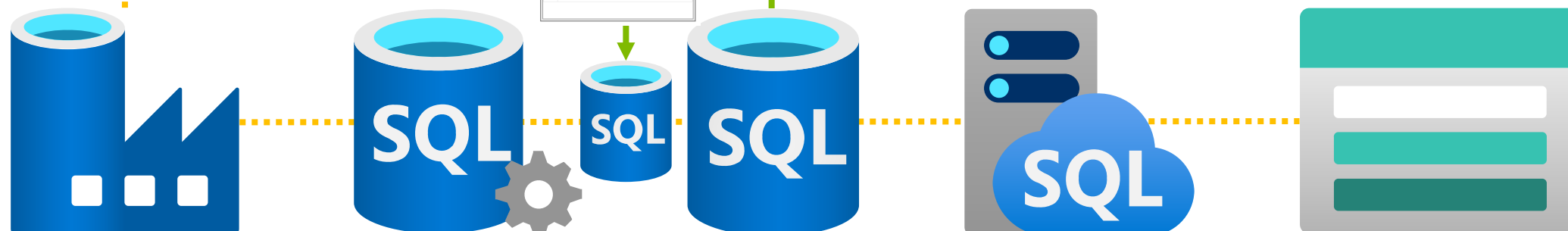
SSIS IR



```
;WITH maxBuckets AS
(
    SELECT MAX([BucketId]) AS 'MaxBucket' FROM [dbo].[Buckets]
)
INSERT INTO [dbo].[BucketToProcessMap]
SELECT
    CASE
        WHEN (ROW_NUMBER() OVER (ORDER BY p.[Duration] DESC) * 1)
        % maxBuckets.[MaxBucket] = 0 THEN maxBuckets.[MaxBucket]
        ELSE (ROW_NUMBER() OVER (ORDER BY p.[Duration] DESC) * 1)
        % maxBuckets.[MaxBucket]
    END AS 'NewBucketId',
    p.[ProcessId]
FROM
    [dbo].[BucketProcesses] p
CROSS JOIN maxBuckets
```



SSISDB



```
SELECT [BucketId] FROM [dbo].[Buckets] WHERE [Enabled] = 1
```

Copy SSIS
Package
Executions

Allocate
Packages to
Buckets

Get Bucket IDs

Execute
Buckets in
Parallel

FE L1
MAXDOP 50

Array[Buckets]

Execute Bucket

Get Packages
in Bucket ID

Execute
Packages in
Bucket

FE L2
MAXDOP 50

@item().BucketId

Array[BucketContents]

Run Package

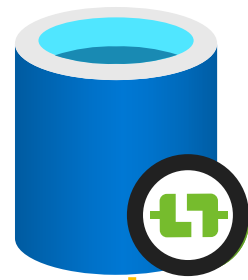
```
SELECT p.* FROM
[dbo].[BucketToProcessMap] m
INNER JOIN [dbo].[BucketProcesses] p
ON m.[ProcessId] = p.[ProcessId]
WHERE m.[BucketId] = @BucketId
```

$$(FE L1) \times (FE L2) = NEW MAXDOP$$
$$50 \times 50 = 2500$$



Solution 4: Nested ForEach Activities & Bucket Metadata

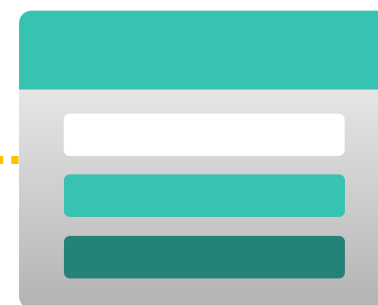
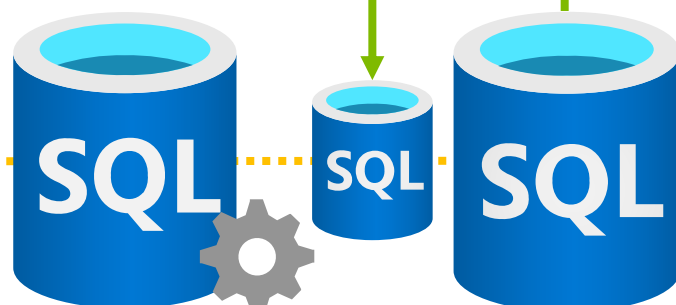
SSIS IR



Compute size	S4	S6	S7	S9	S12
Max DTUs	200	400	800	1600	3000
Included storage (GB) ¹	250	250	250	250	250
Max storage (GB)	1024	1024	1024	1024	1024
Max in-memory OLTP storage (GB)	N/A	N/A	N/A	N/A	N/A
Max concurrent workers (requests)	400	800	1600	3200	6000
Max concurrent sessions	4800	9600	19200	30000	30000

<https://docs.microsoft.com/en-us/azure/azure-sql/database/resource-limits-dtu-single-databases>

SSISDB



Copy SSIS Package Executions

Allocate Packages to Buckets

Get Bucket IDs

Execute Buckets in Parallel

FE L1
MAXDOP 50

Array[Buckets]

Execute Bucket

Get Packages in Bucket ID

Execute Packages in Bucket

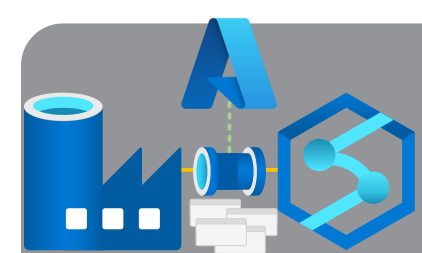
FE L2
MAXDOP 50

@item().BucketId

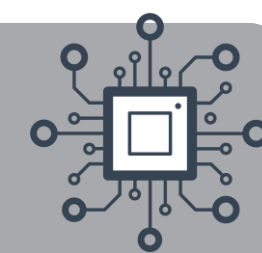
Array[BucketContents]

Run Package

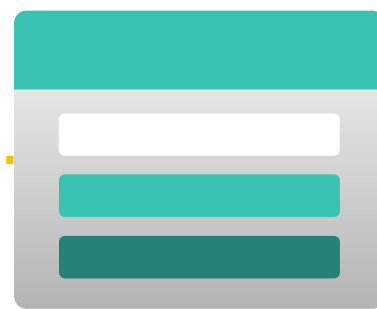
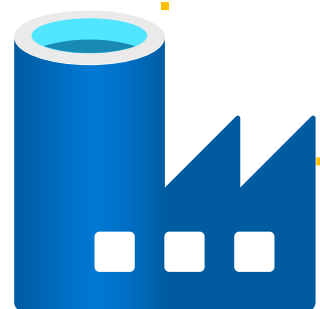
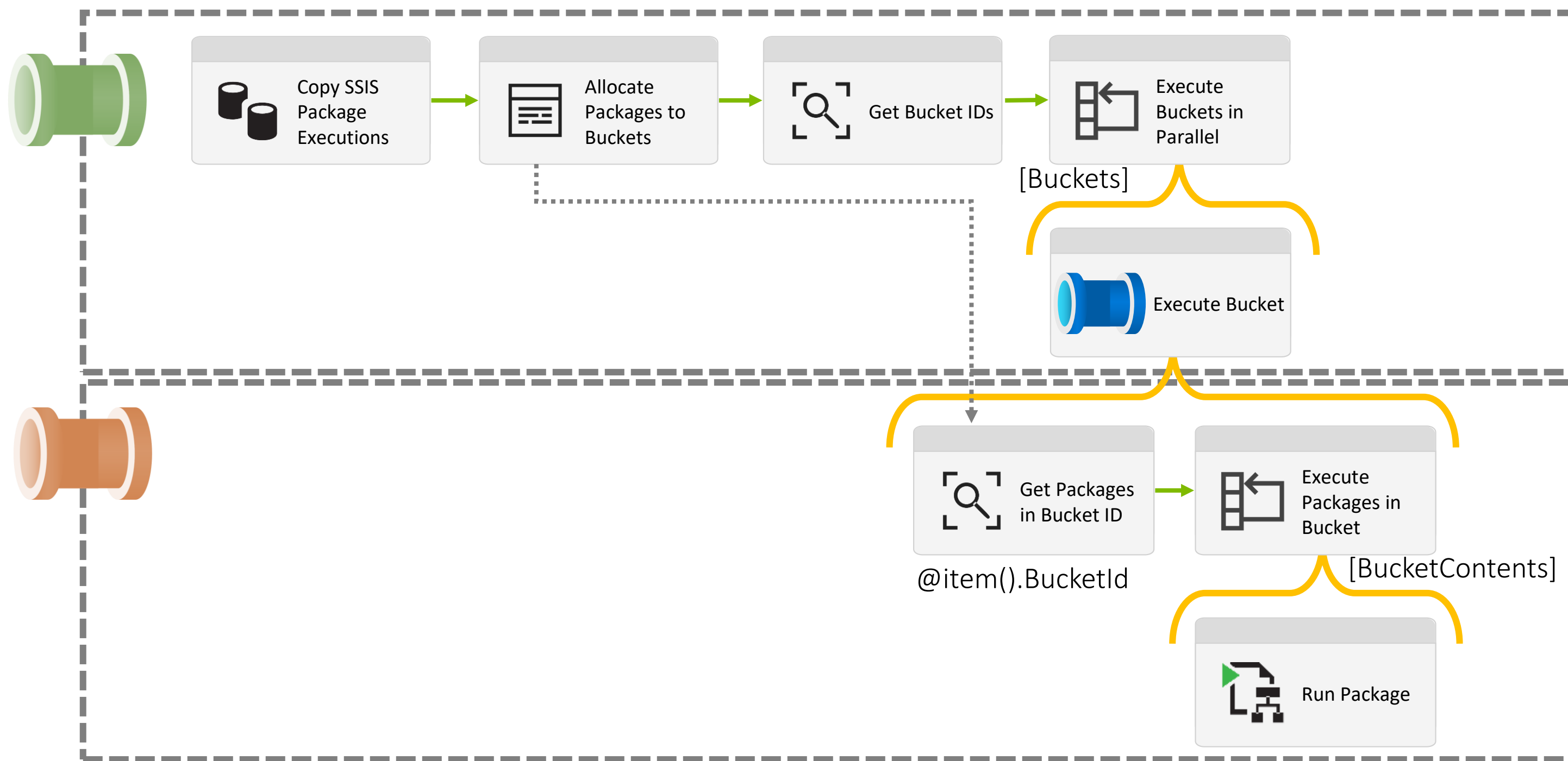
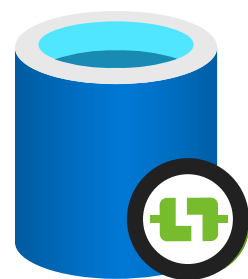
$$(FE L1) \times (FE L2) = NEW MAXDOP$$
$$50 \times 50 = 2500$$



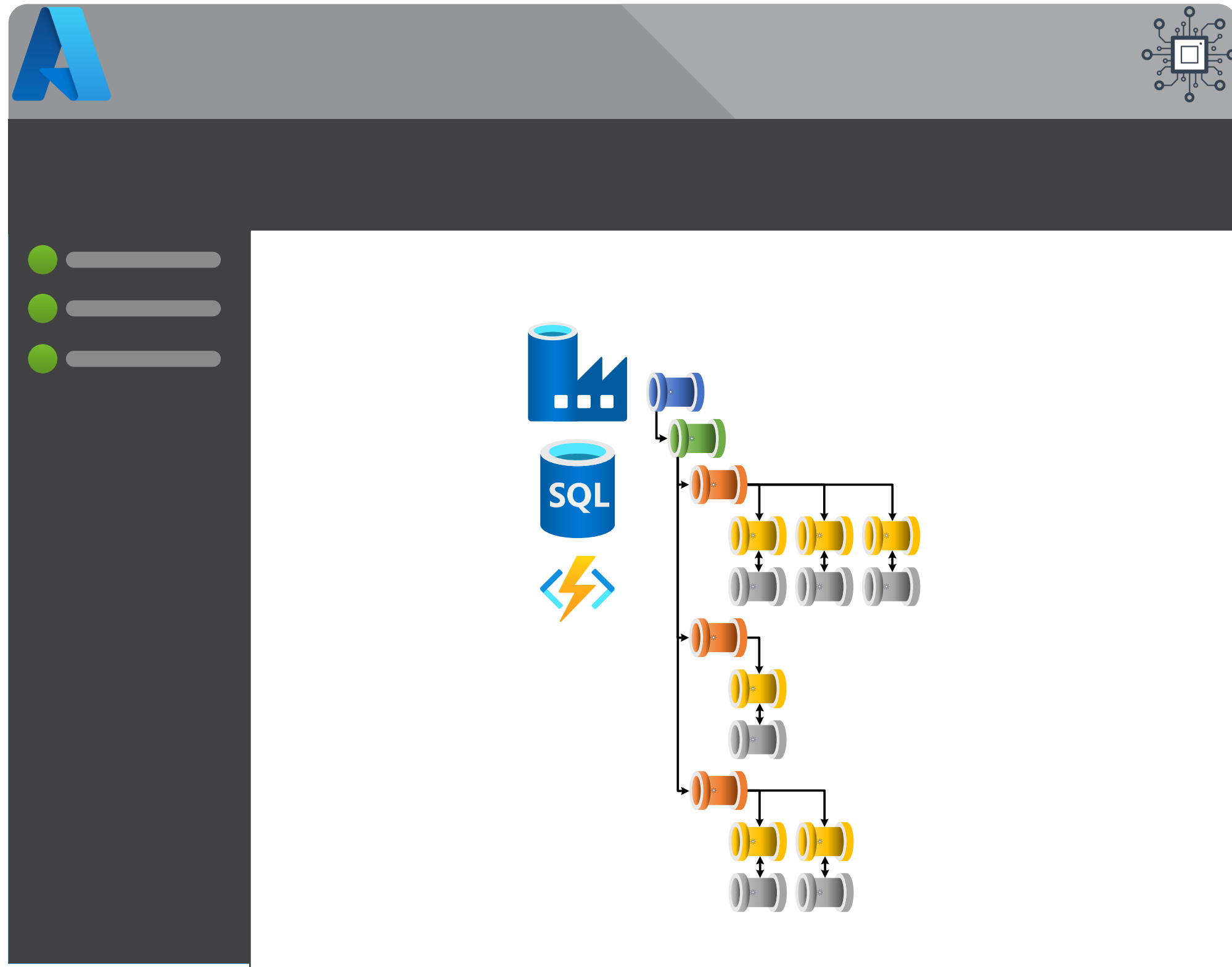
Solution 4: Nested ForEach Activities & Bucket Metadata

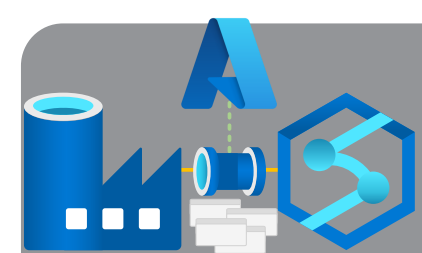


SSIS IR

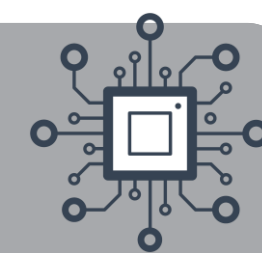


Metadata Driven Framework

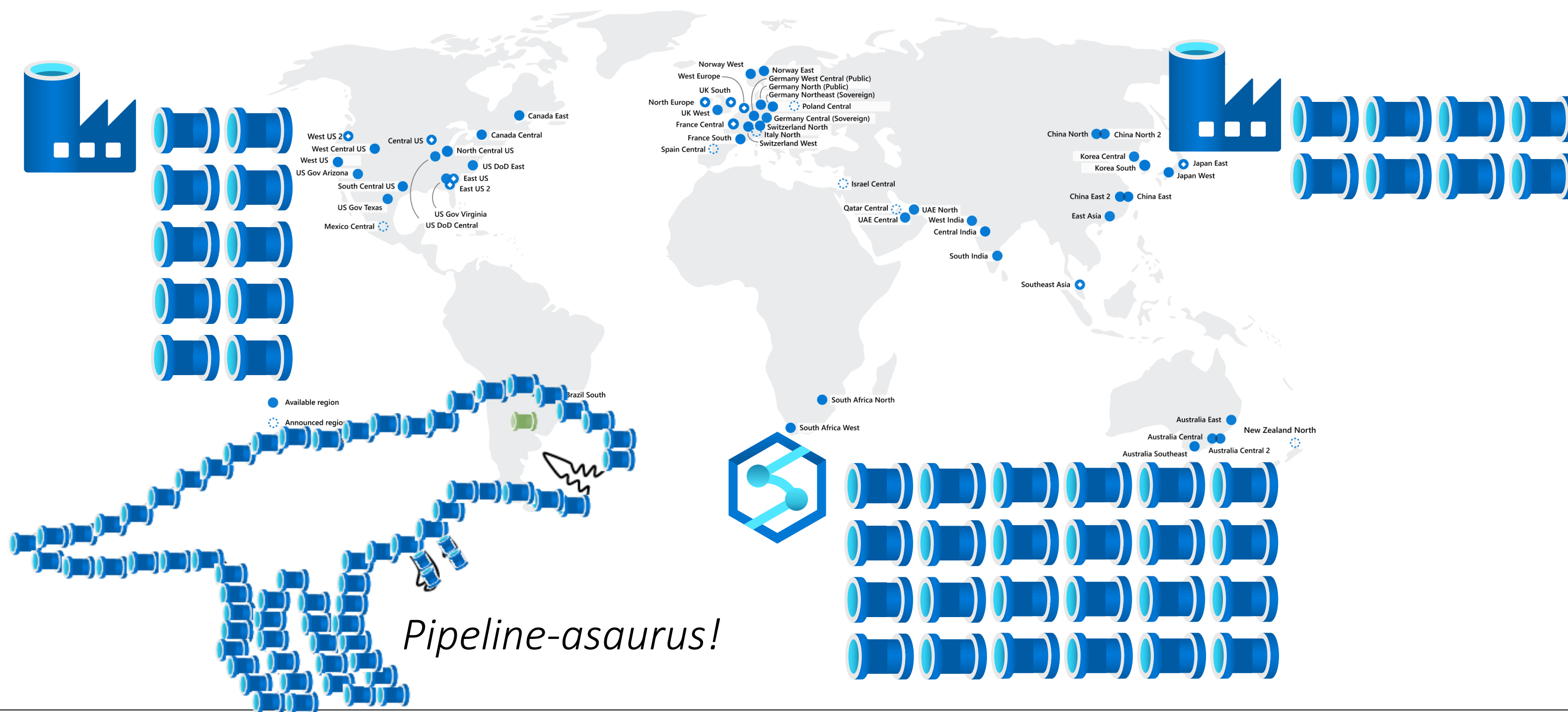




Problem

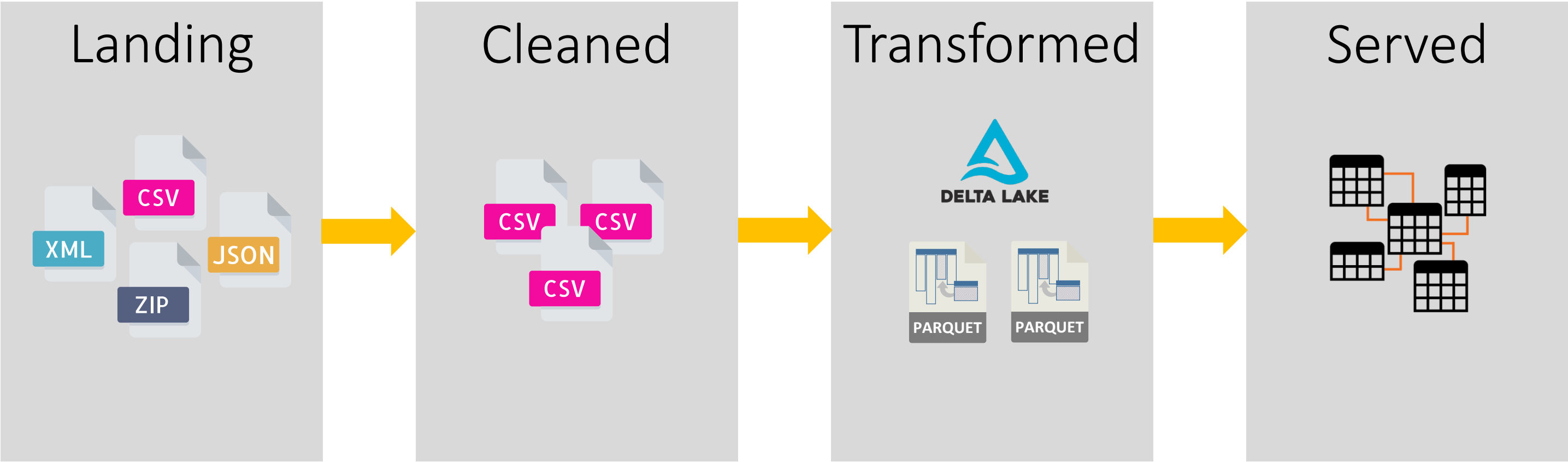
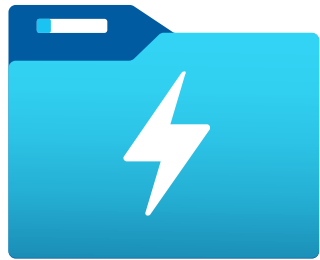
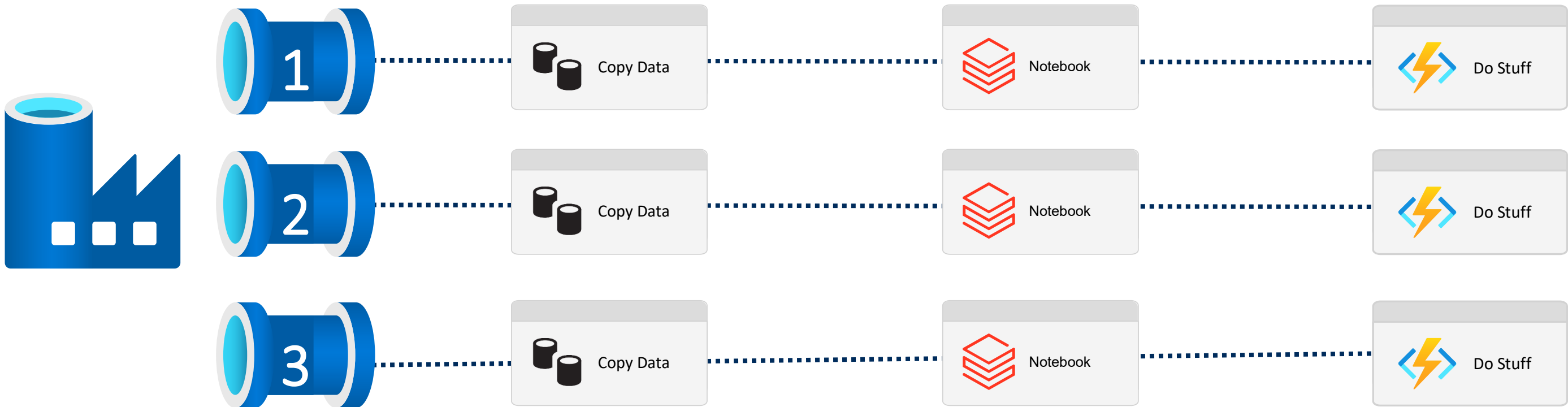
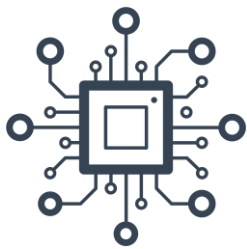


How should we structure and trigger our Integration Pipelines?

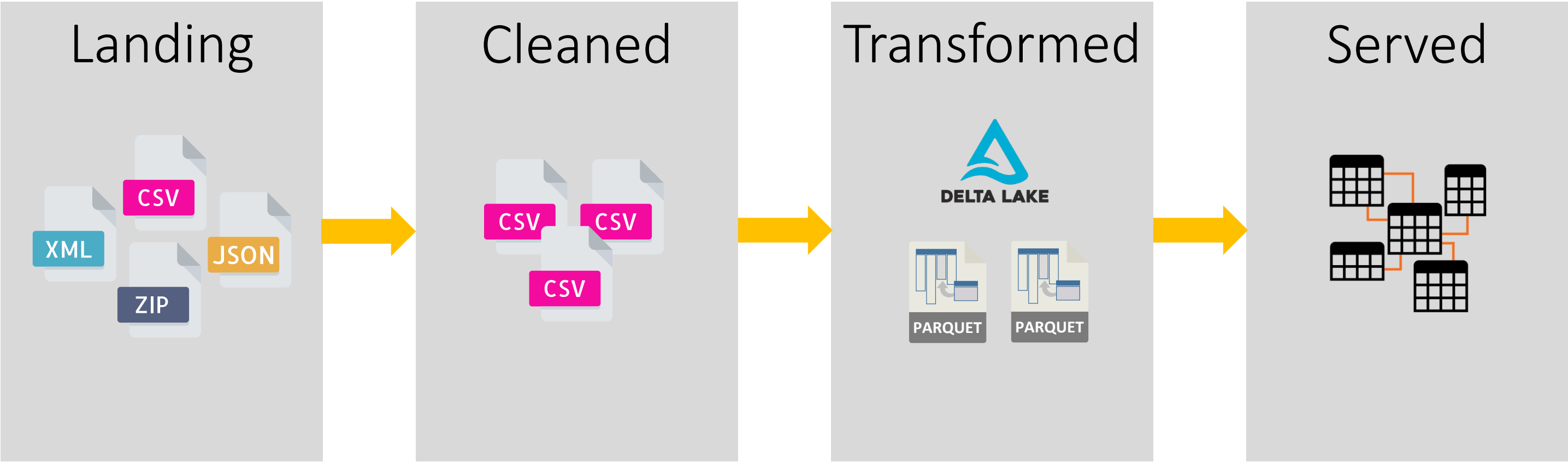
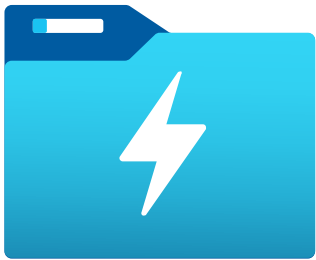
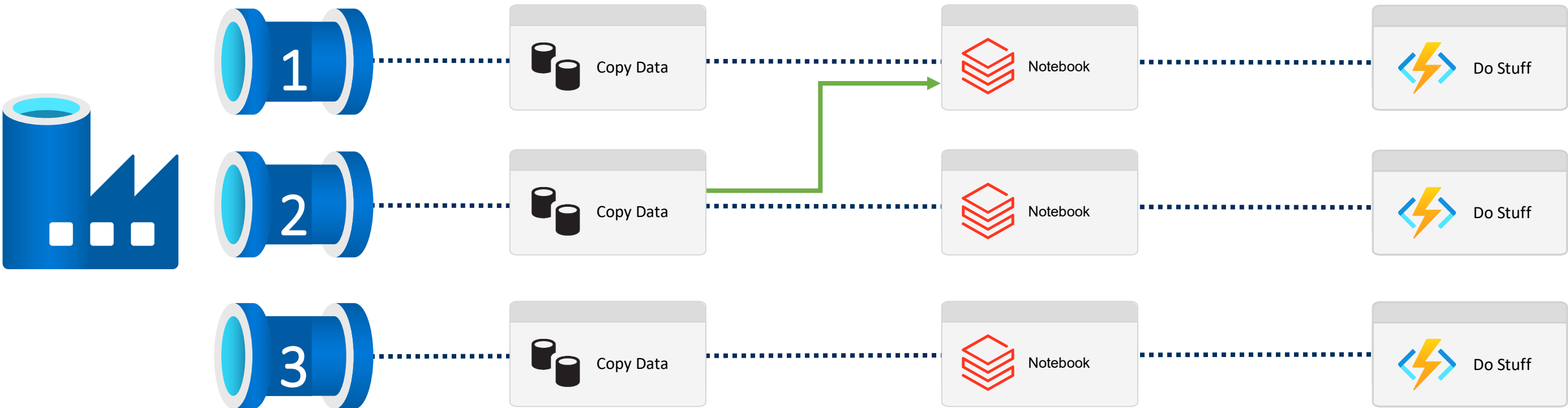
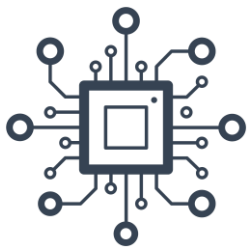


Pipeline-asaurus!

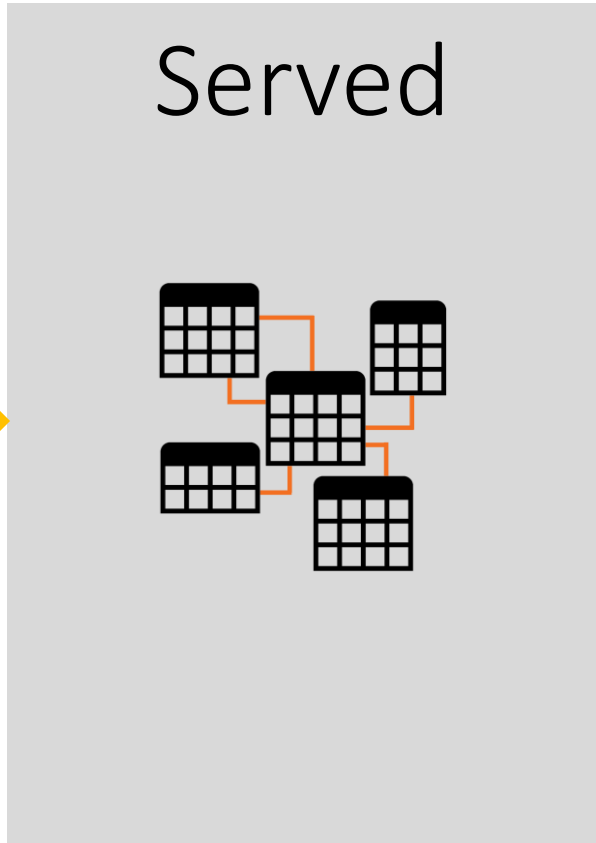
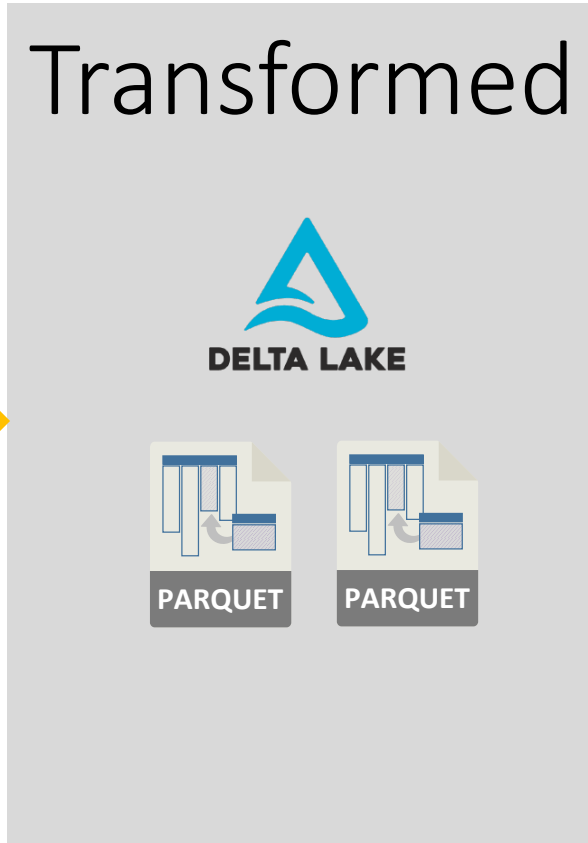
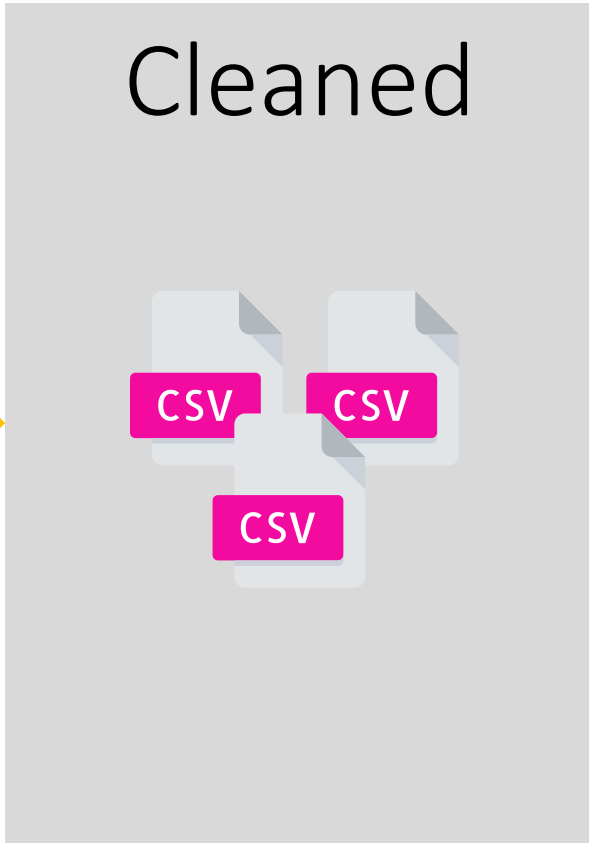
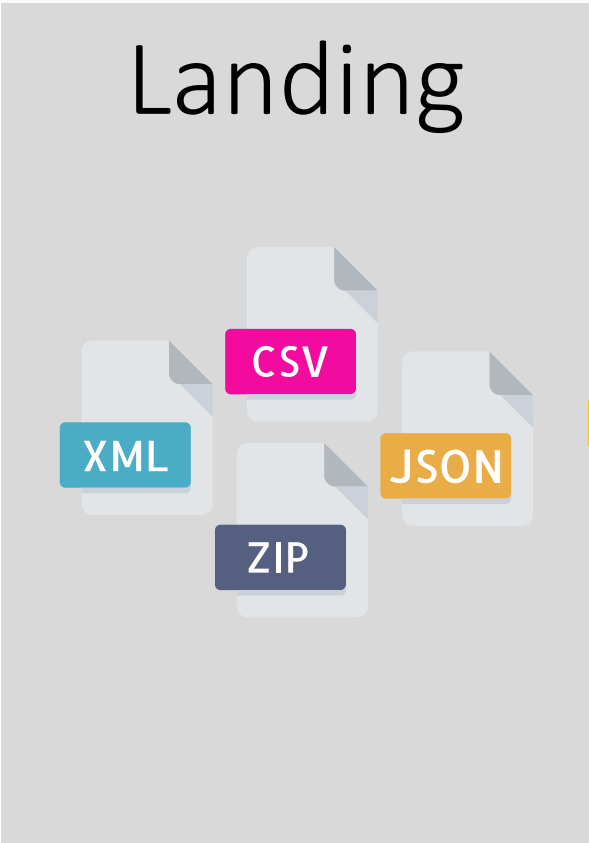
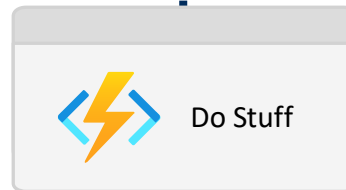
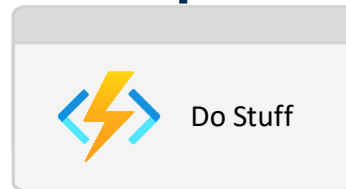
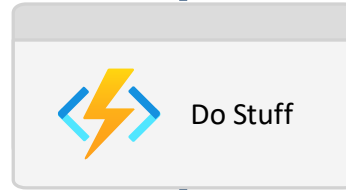
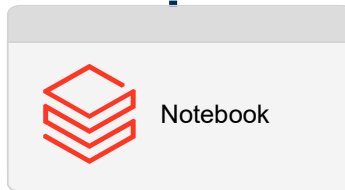
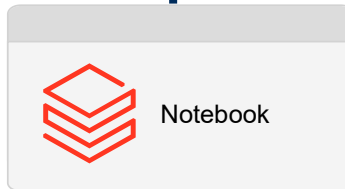
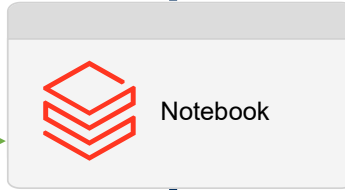
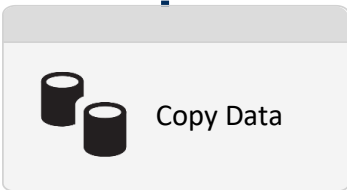
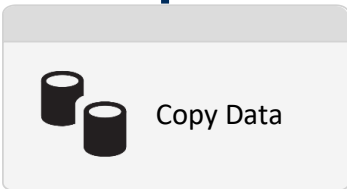
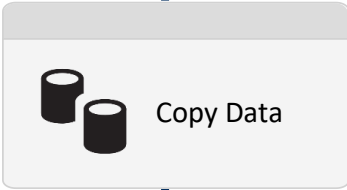
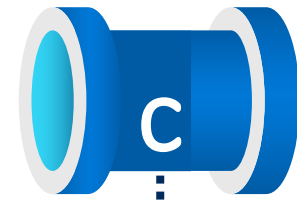
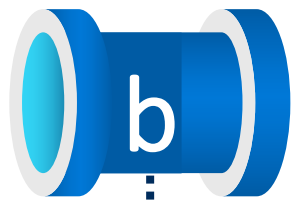
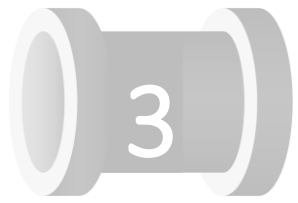
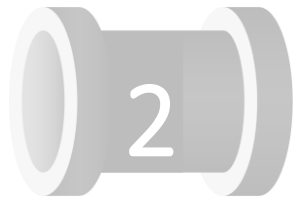
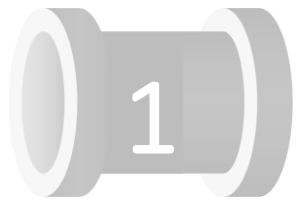
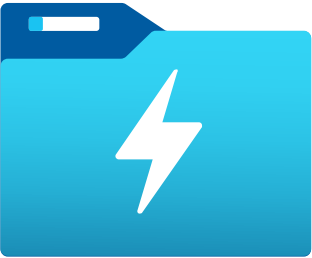
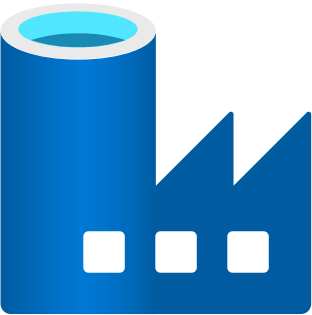
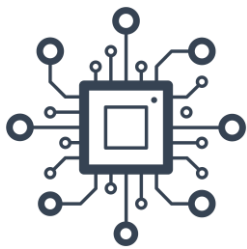
Problem



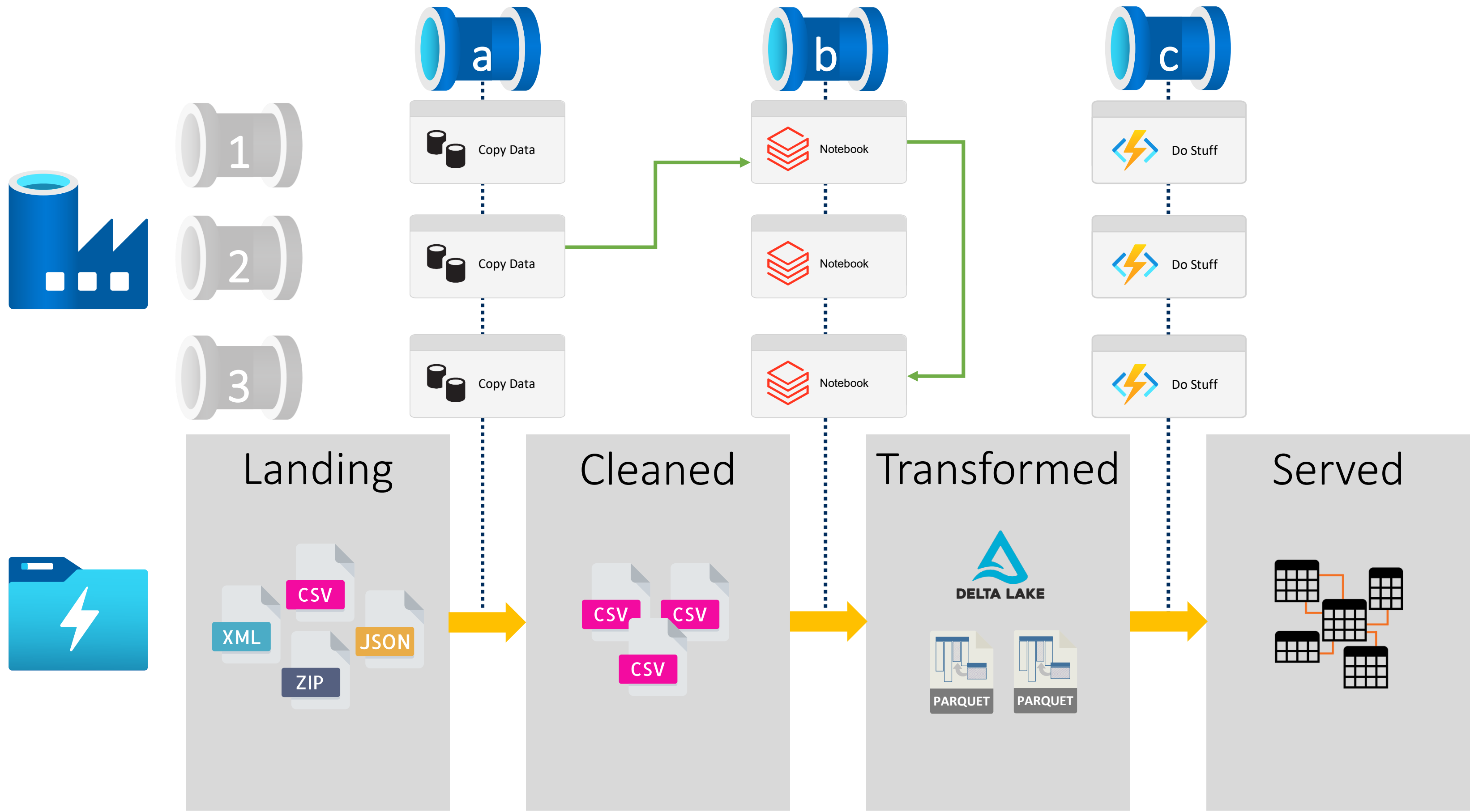
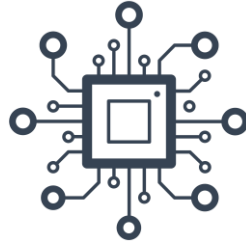
Problem



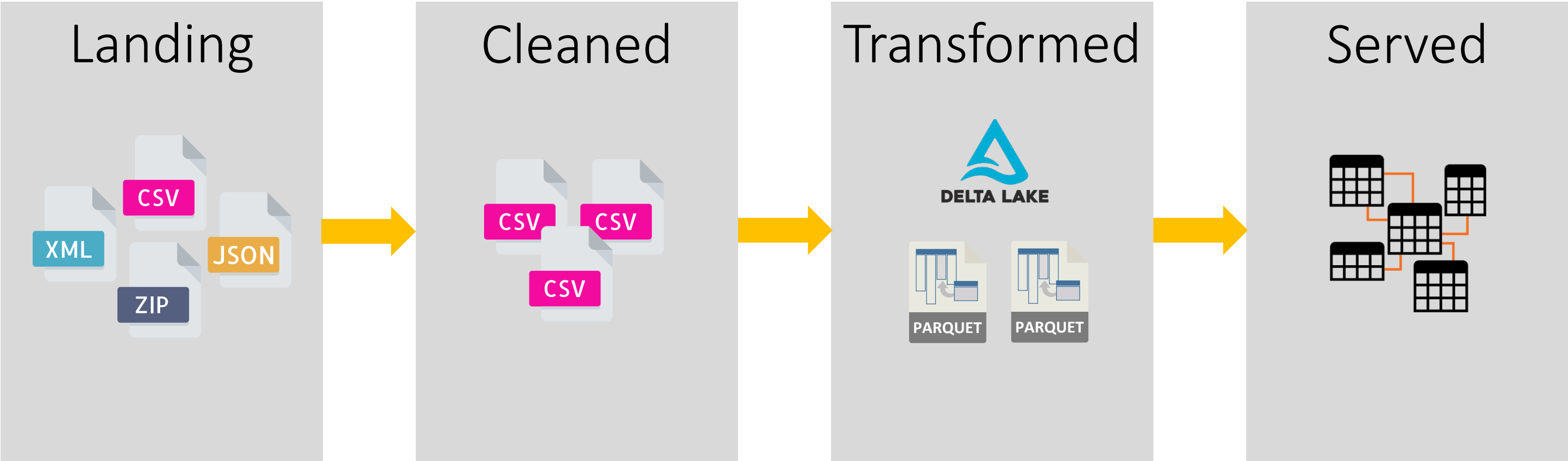
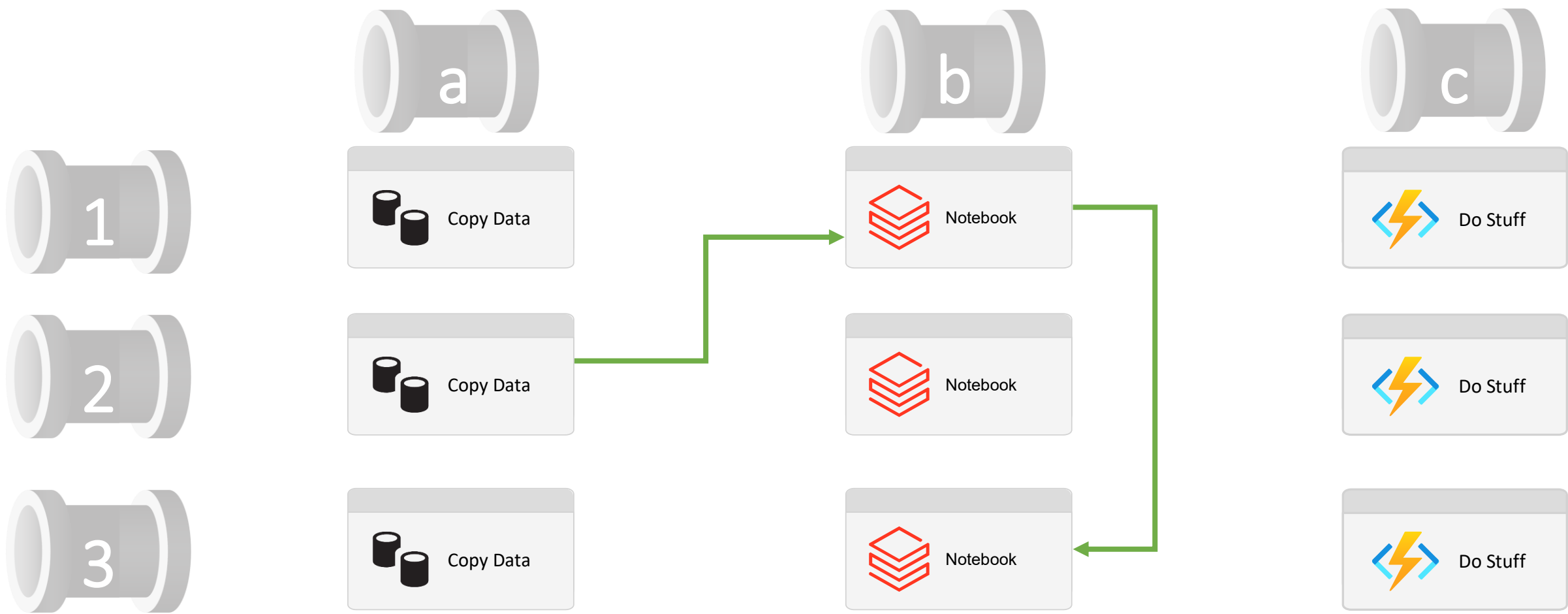
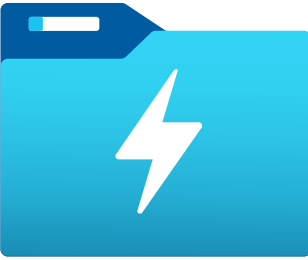
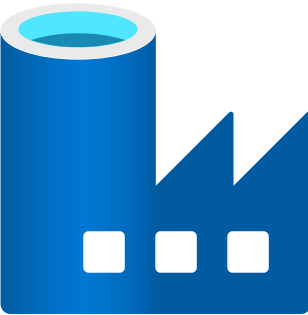
Problem




Problem

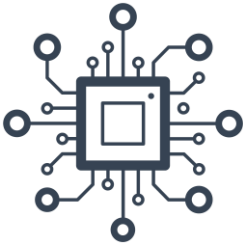
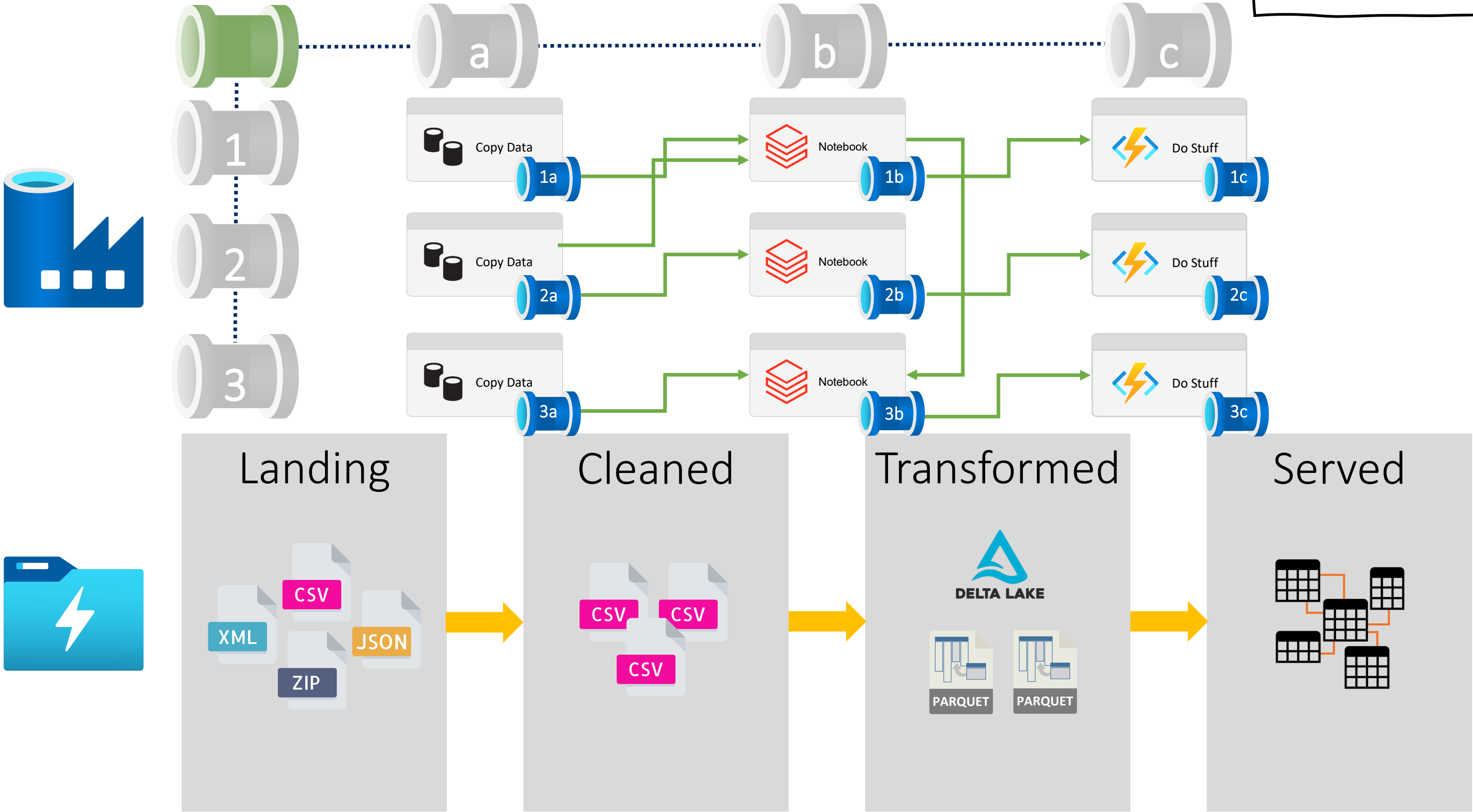


Problem

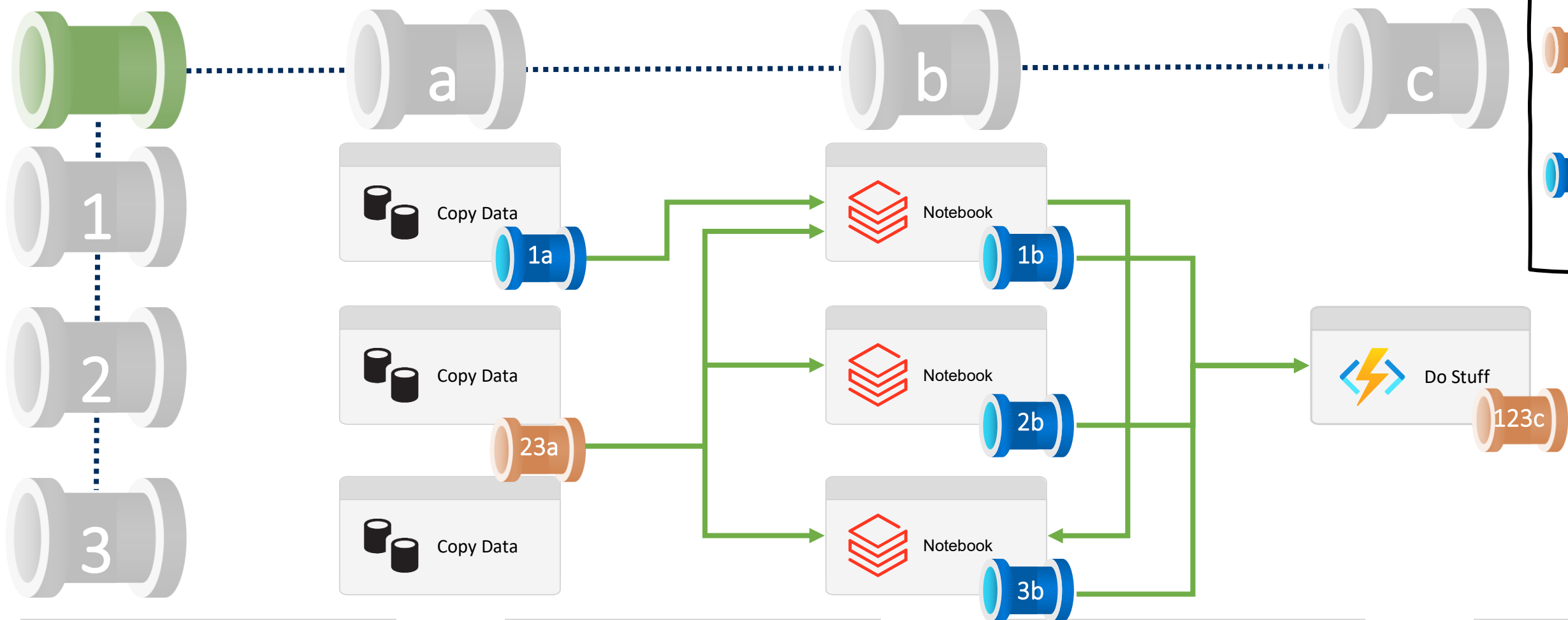
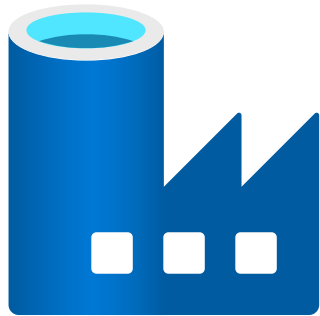





Problem

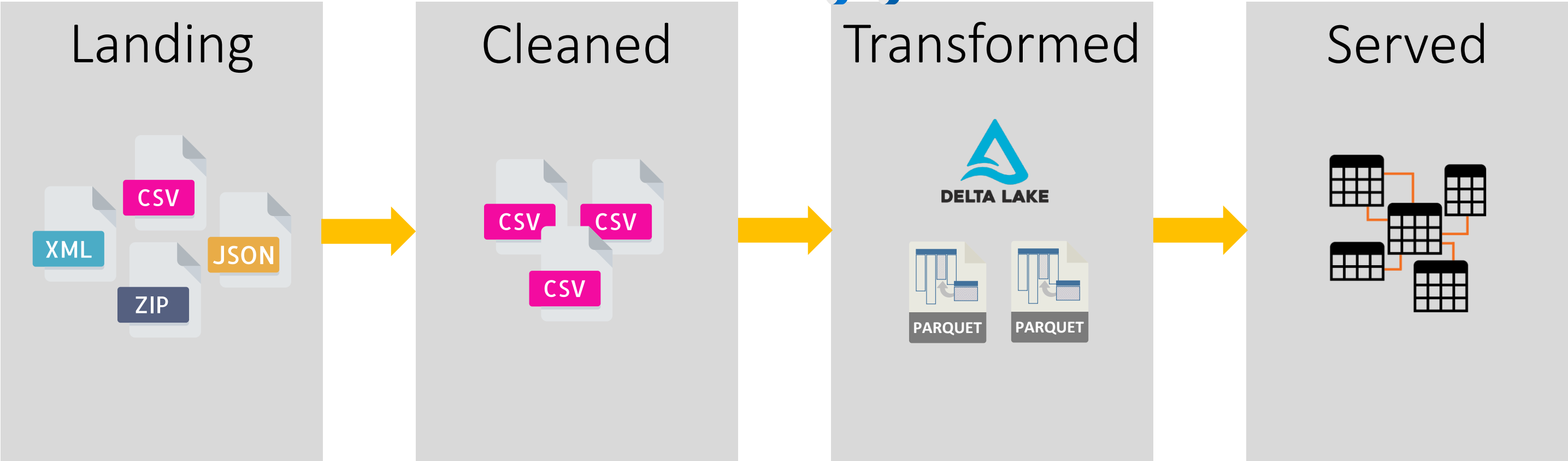
 Only 40 Activities per Pipeline.

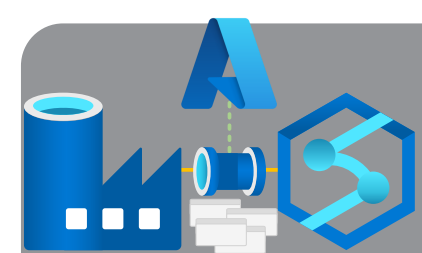


Problem

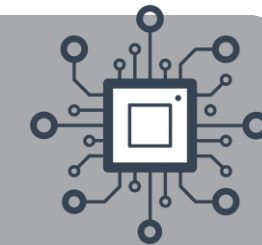


-  Grandparent pipeline for all processing.
-  Parent pipeline to consolidate work.
-  Child pipelines for low level dependencies.

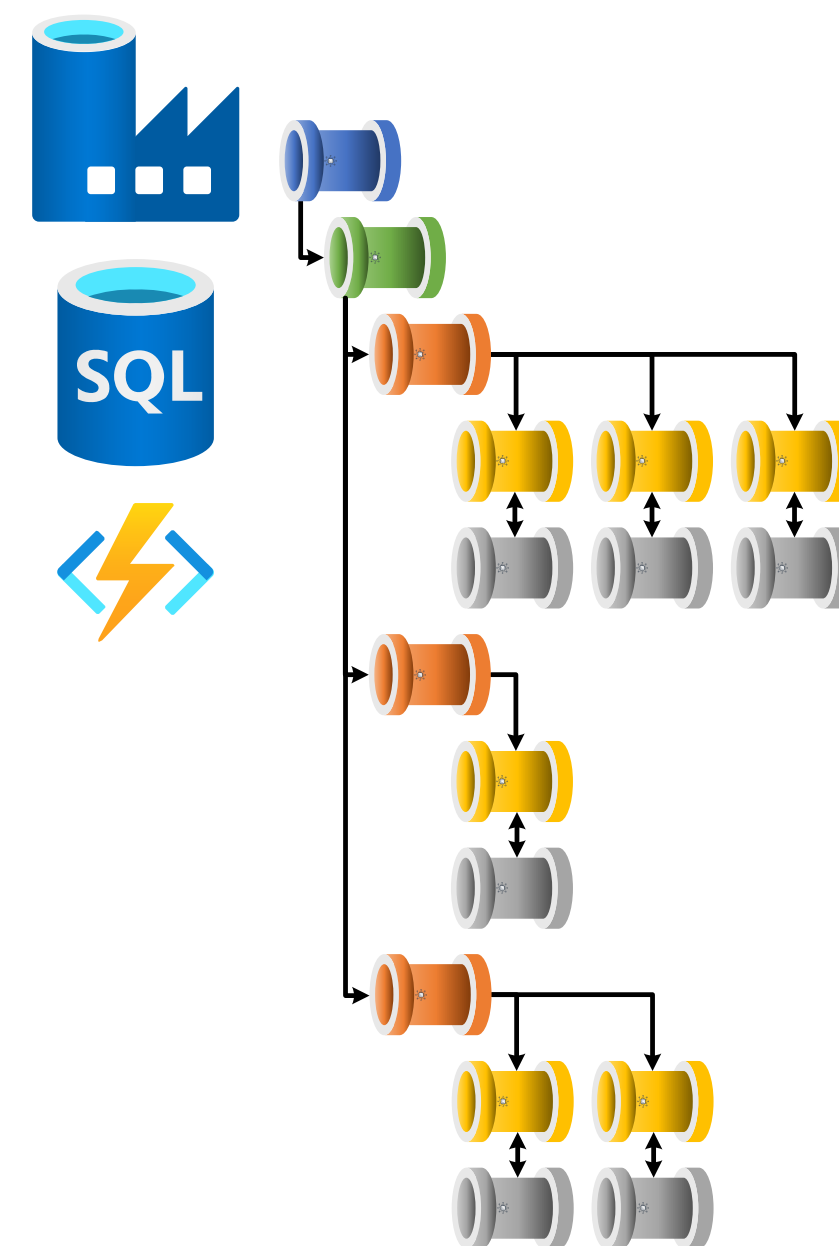
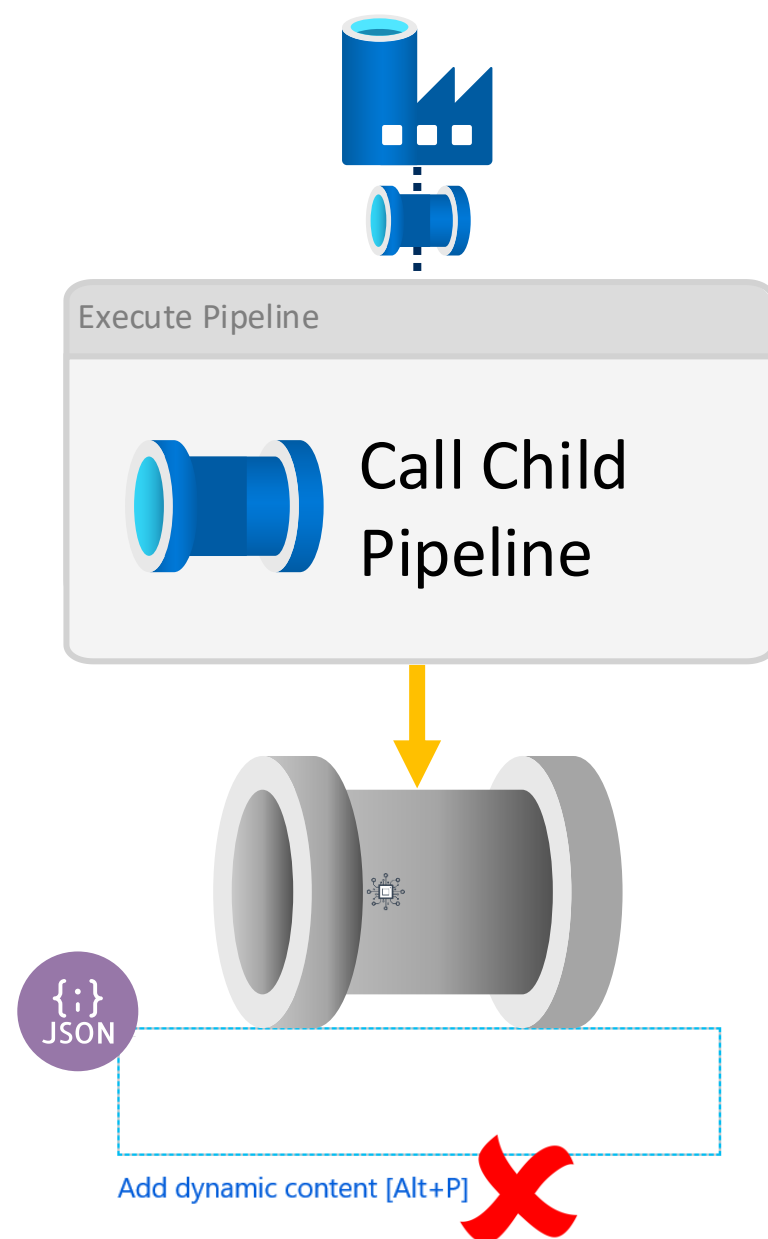




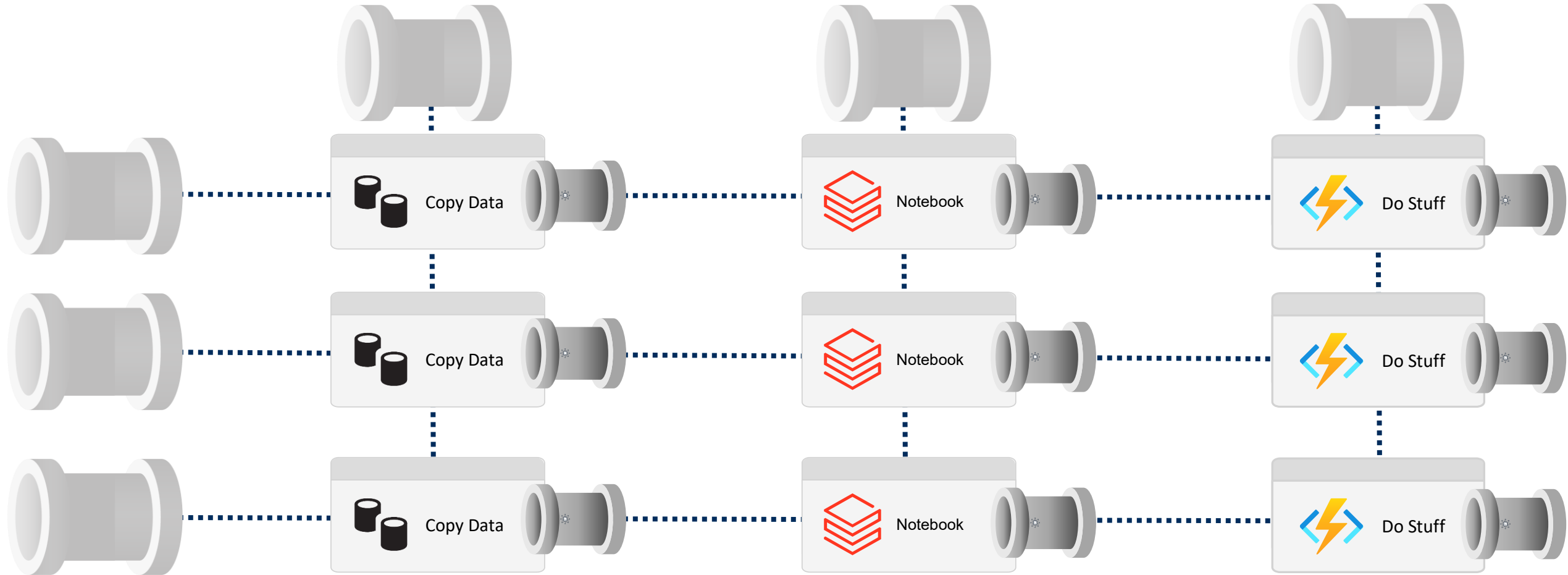
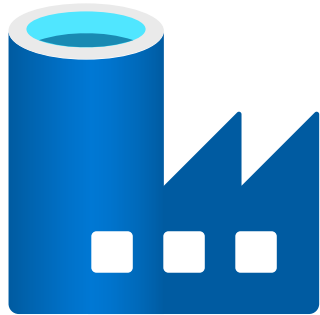
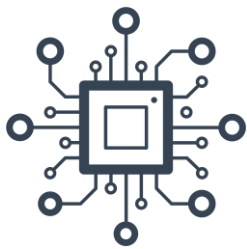
Solution



Use Metadata to Drive Integration Pipeline execution

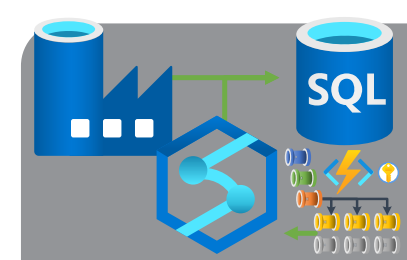


Solution

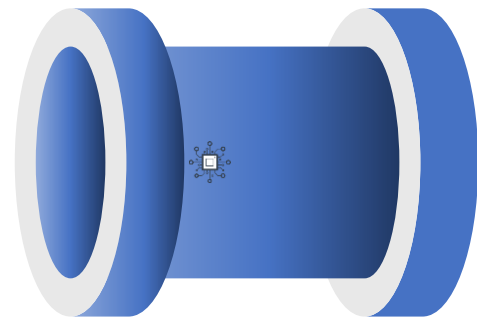
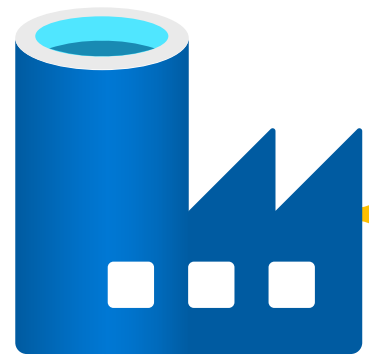
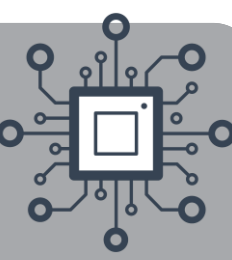


Stages	Pipelines
1	a
2	b
3	c
	d
	e
	f
	g
	h
	i

Stage	Pipeline
1	a
1	b
1	c
2	d
2	e
3	f
3	g
3	h
3	i

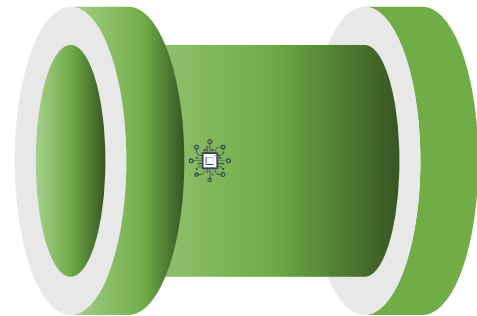


Framework Pipeline Hierarchy



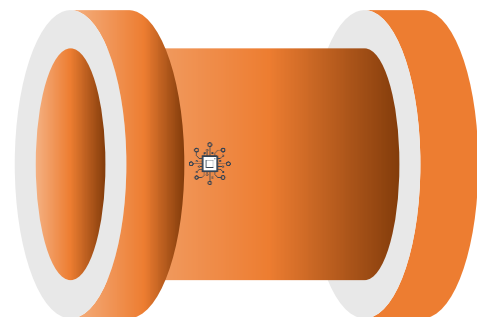
- Grandparent

Role: Optional level platform setup, for example, scale up/out compute services ready for the framework to run.



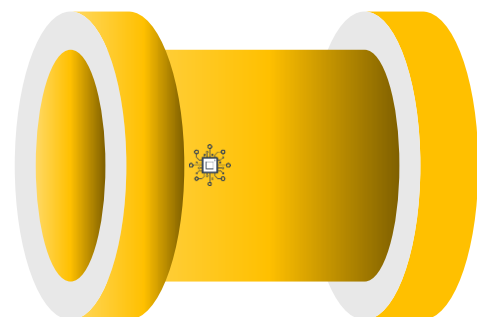
- Parent

Role: Execution run wrapper for batches and execution stage iterator.



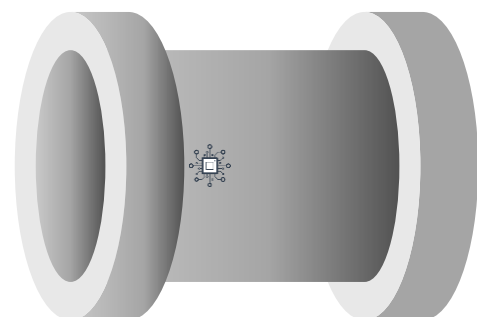
- Child

Role: Scale out triggering of worker pipelines within the execution stage(s).



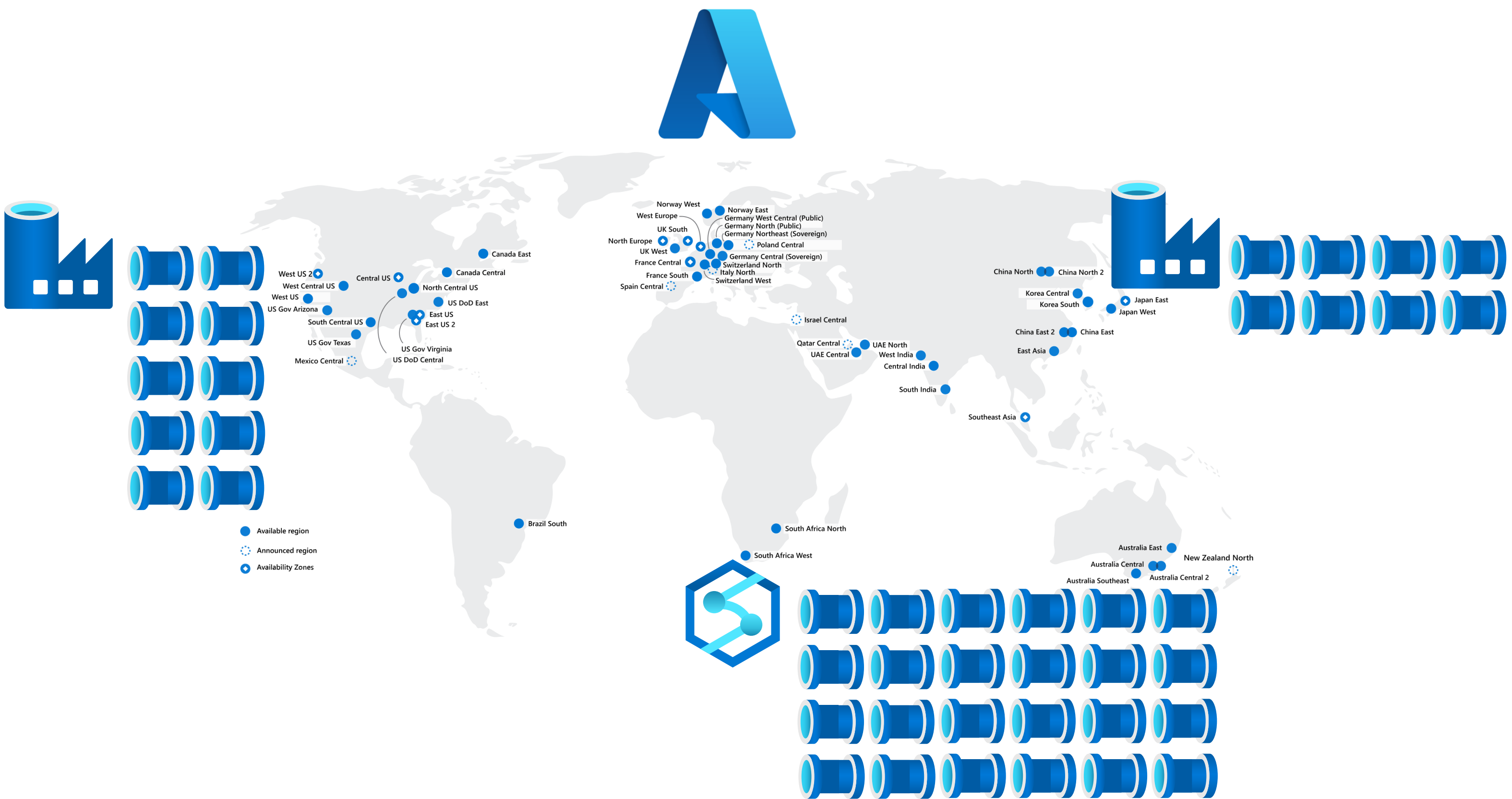
- Infant

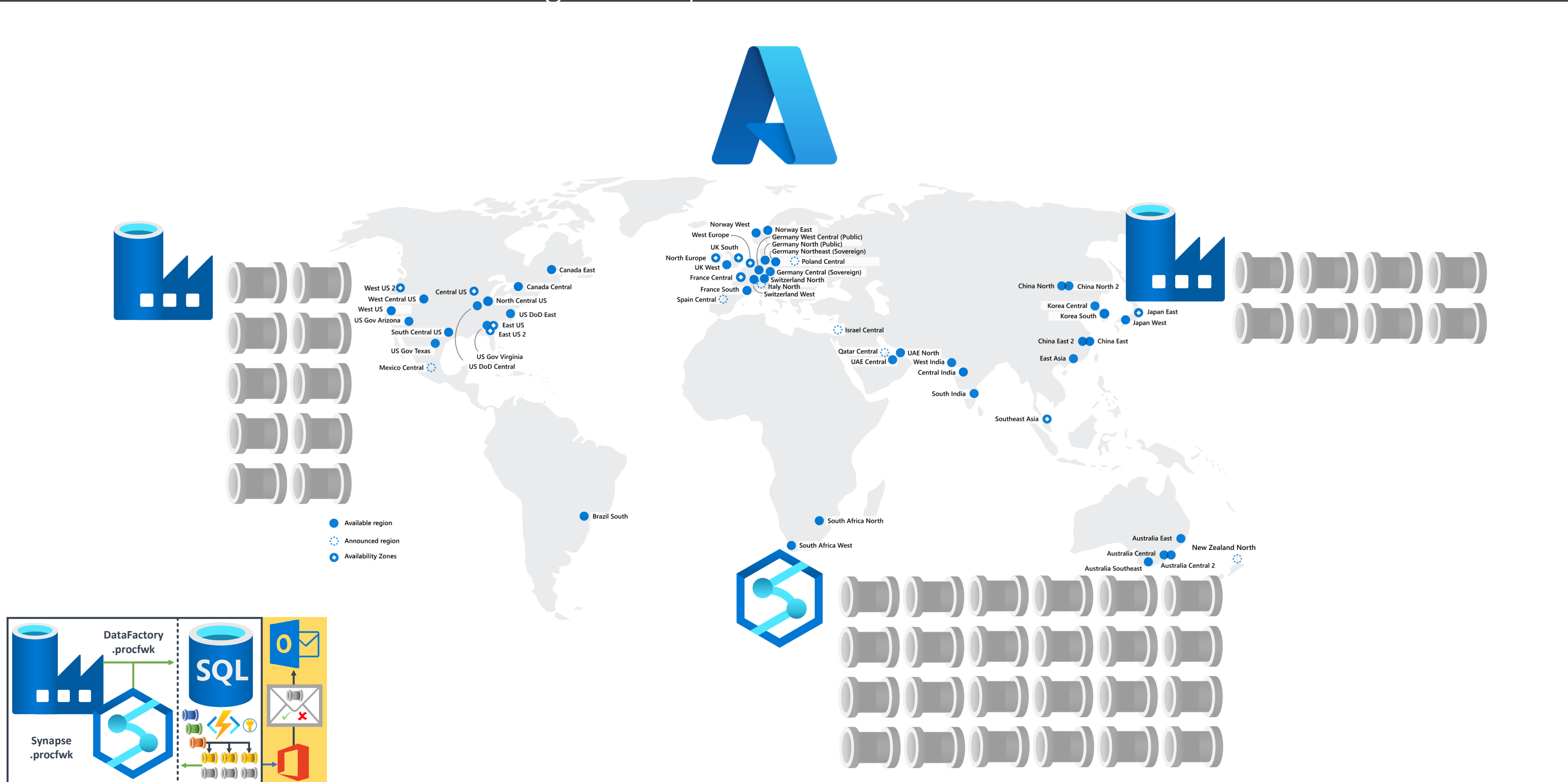
Role: Worker validator, executor, monitor and reporting of the outcome for the single worker pipeline.

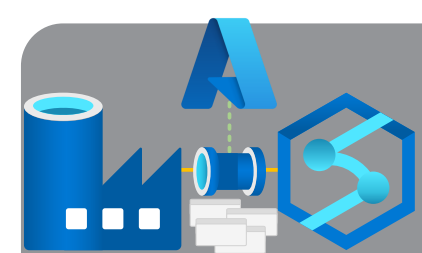


- Worker

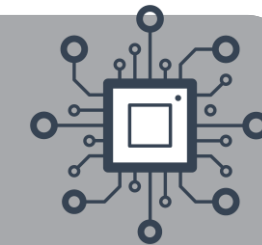
Role: Anything specific to the process needing to be performed.





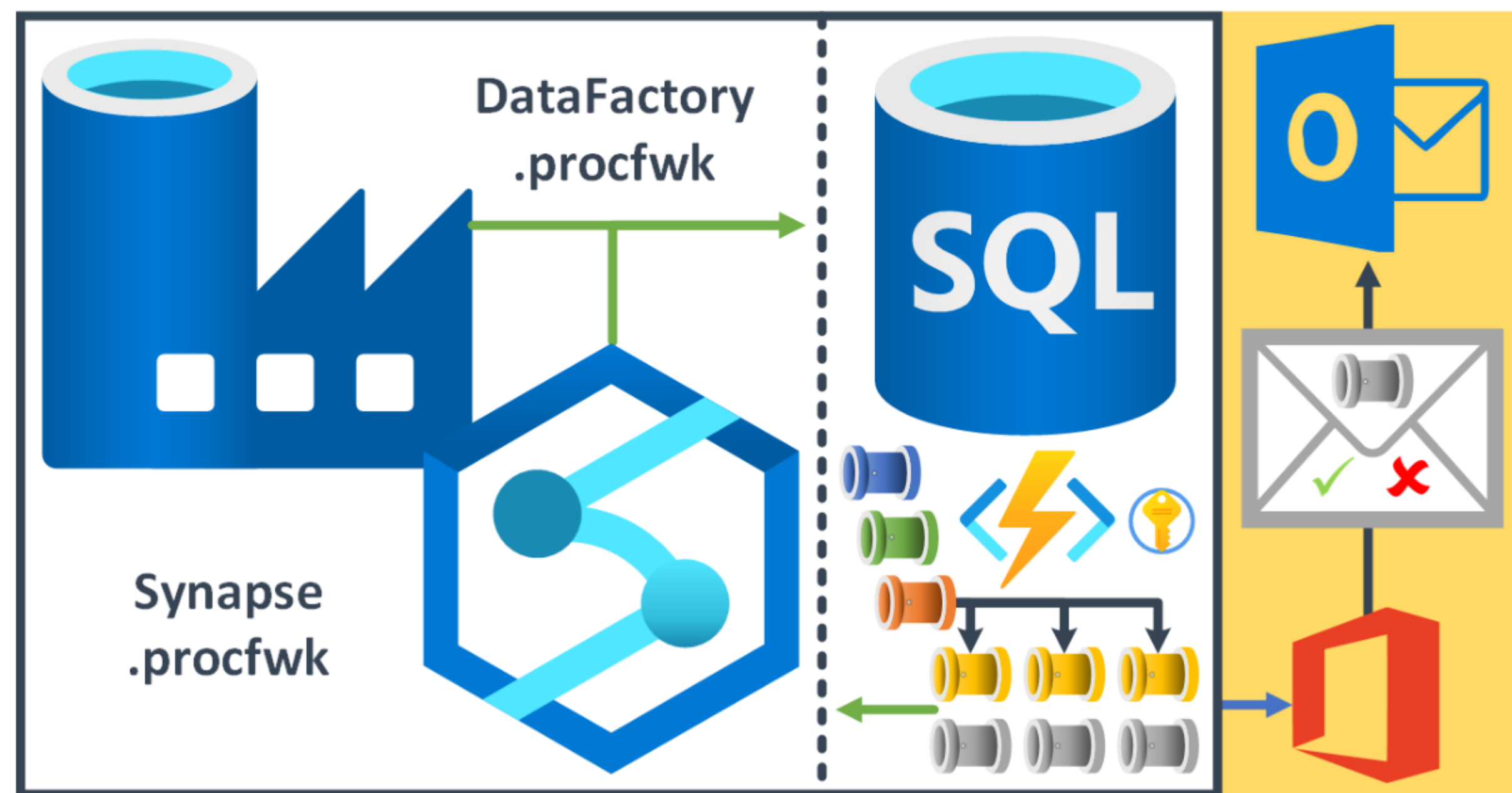


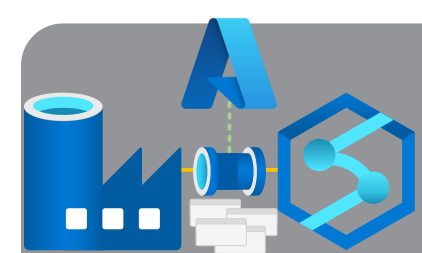
Introducing [ProcFwk.com](https://procfwk.com)



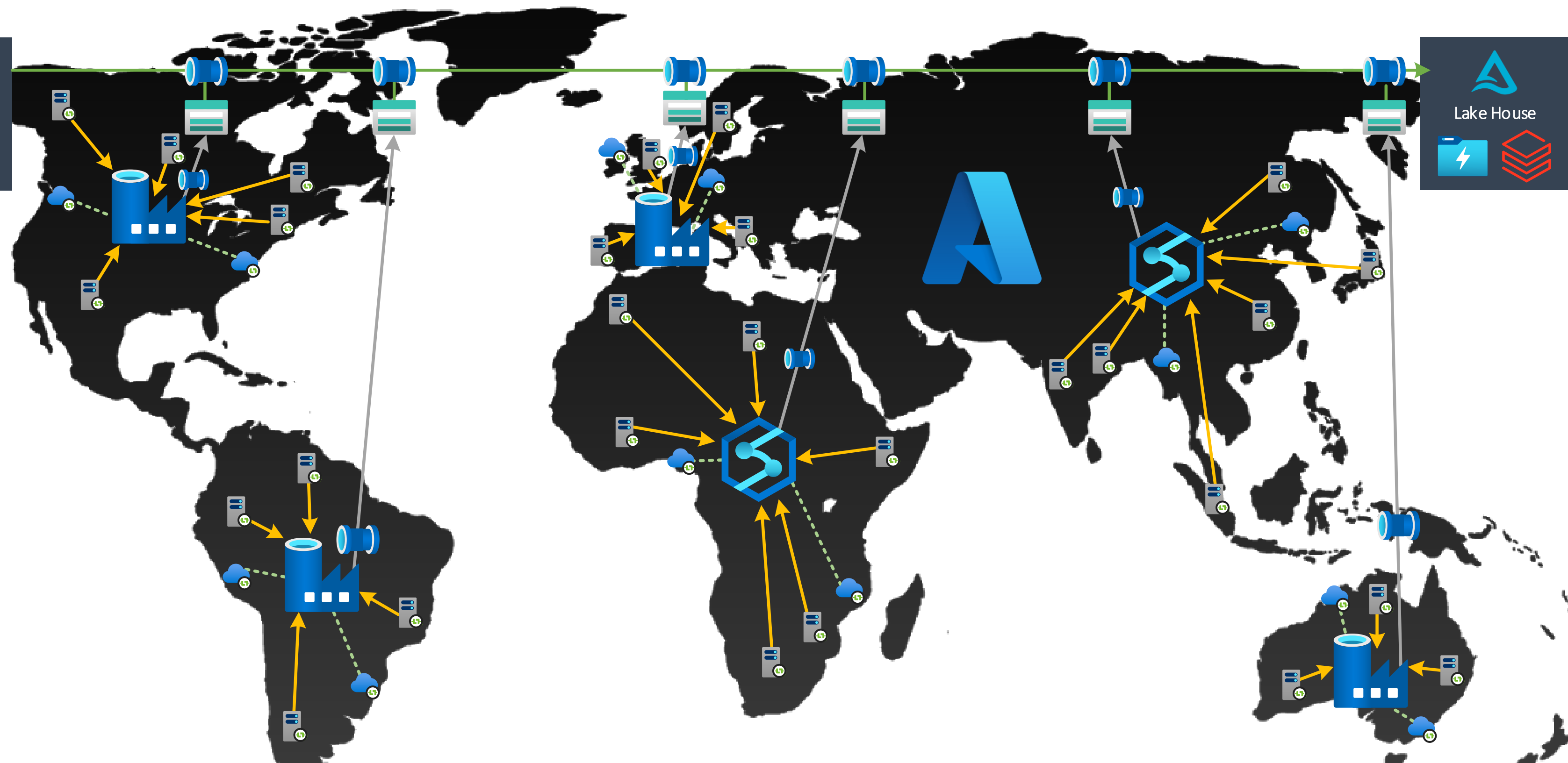
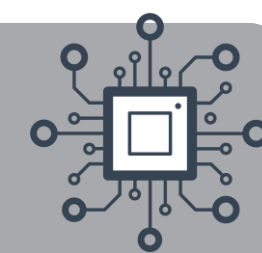
stackoverflow

procfwk



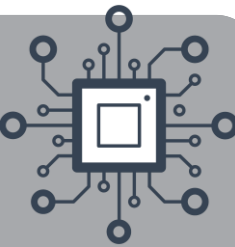


Hub & Spoke Integration Architecture



Testing





🔗 **Integration Tests** - A test of a pipeline as-is, without eliminating any effects of external dependencies.

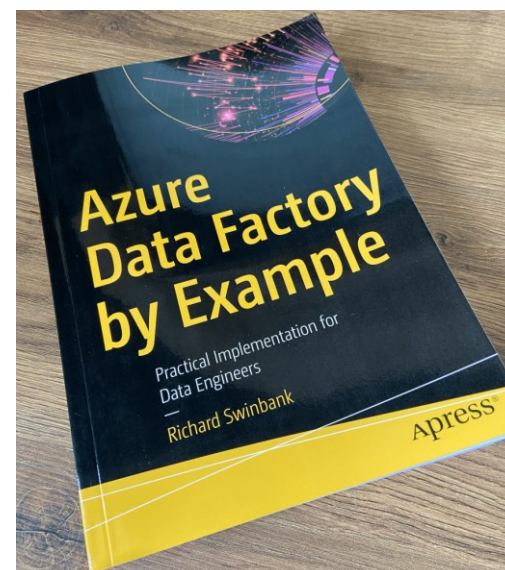
🔗 **Functional Tests** - An isolated test of whether the pipeline is doing things right – is the pipeline producing the desired result?

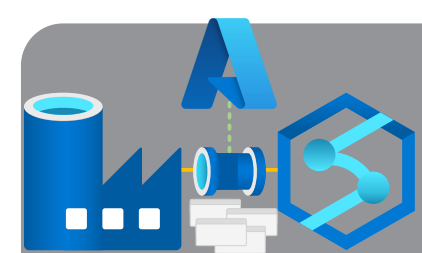
🔗 **Unit Tests** - An isolated test of whether the pipeline is doing the right things – do the pipeline's activities get executed in the way you expect?

Source: Richard Swinbank richardswinbank.net

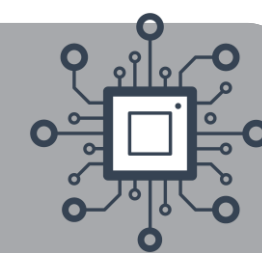
ISBN-13978-1484270288

<https://mrpaulandrew.com/2021/06/29/azure-data-factory-by-exampe-a-review-of-my-technical-review/>





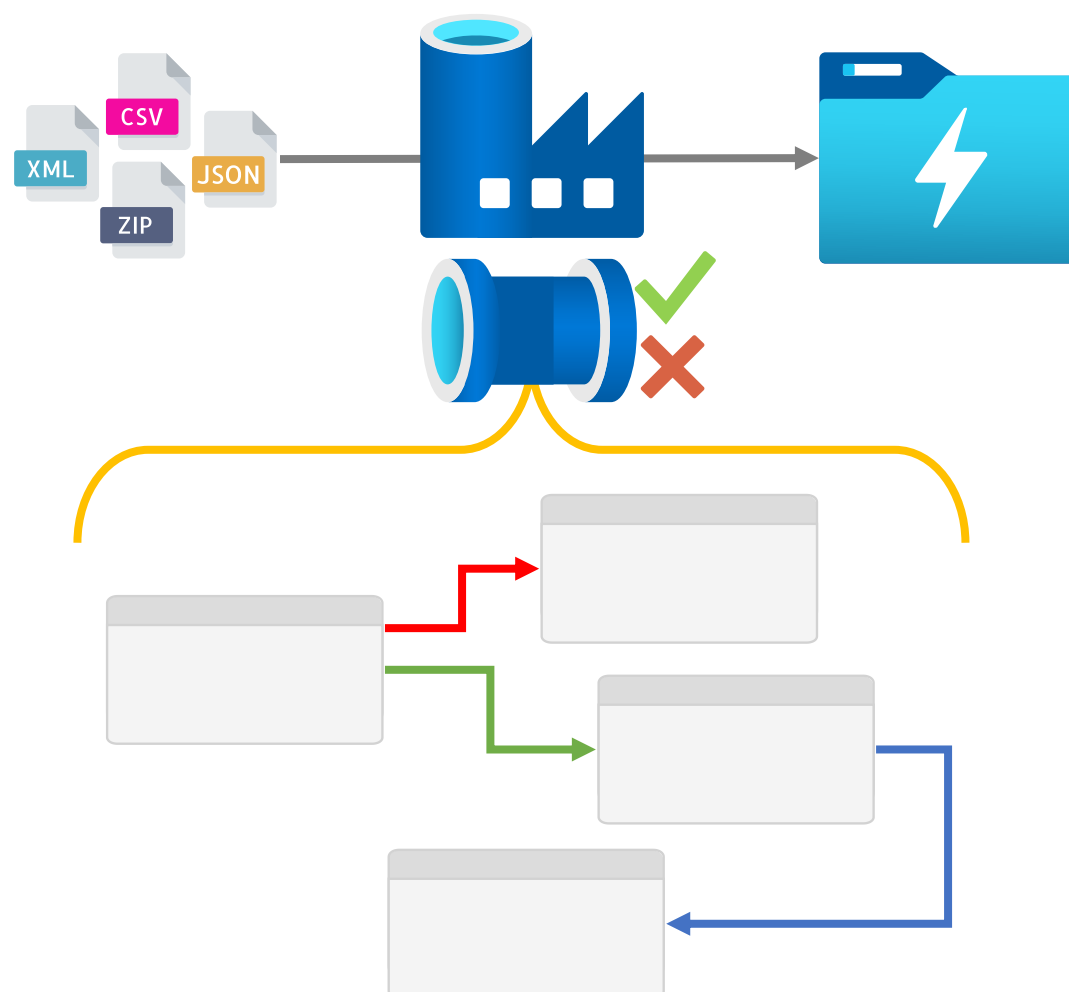
Types of Testing



What do they mean in our pipelines?

Integration Tests

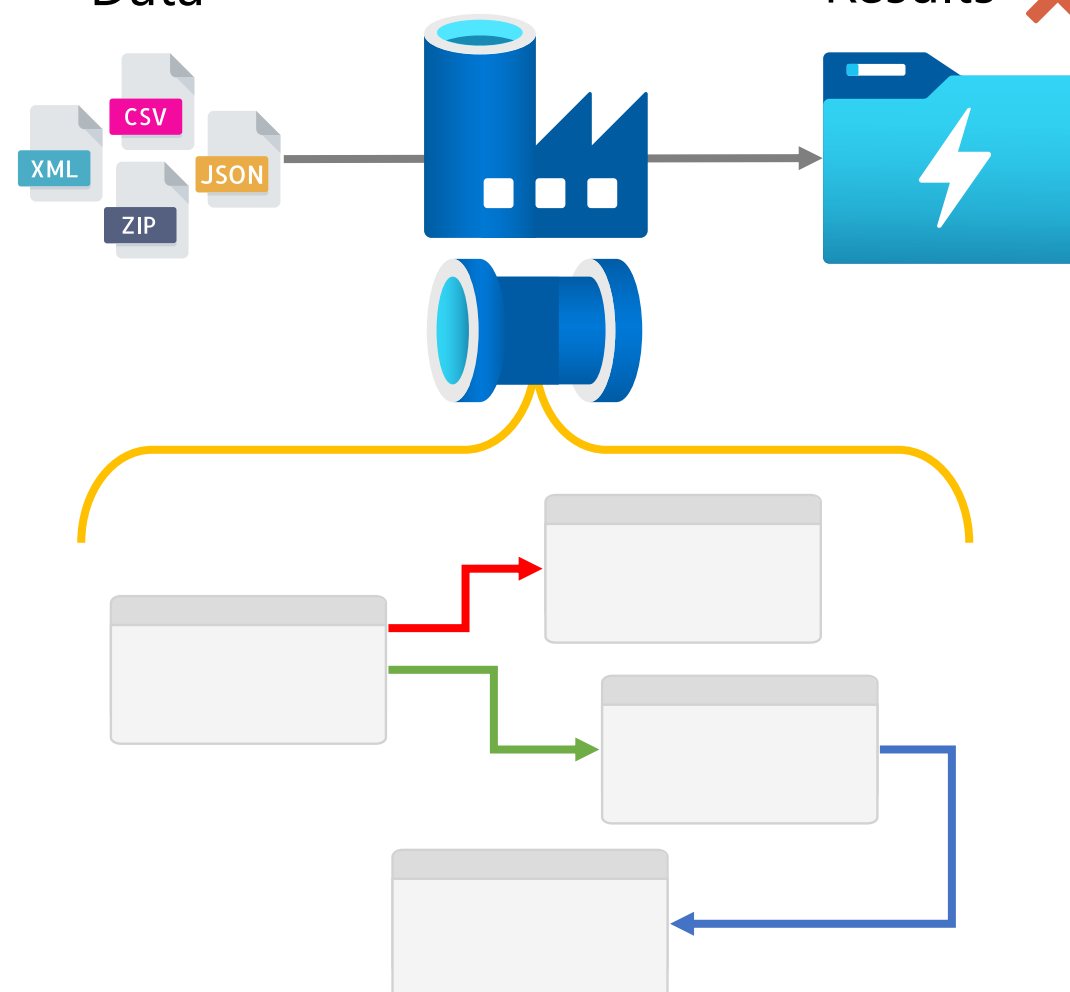
Simple Run Outcome



Functional Tests

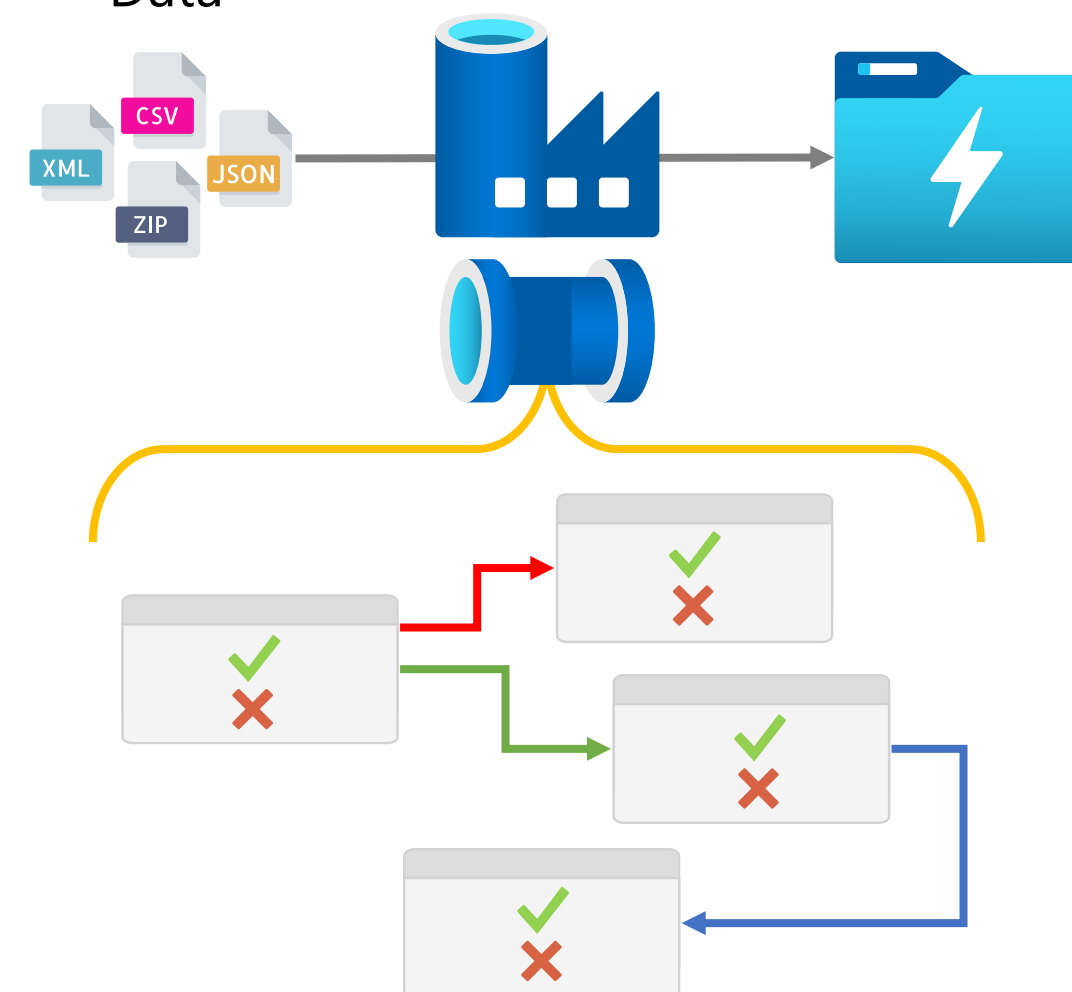
Sample Data

Specific Results ✓
✗

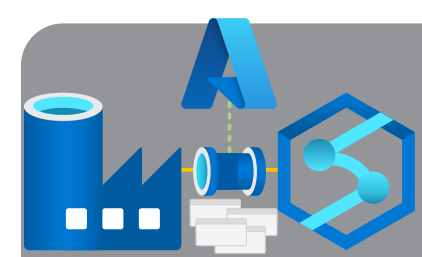


Unit Tests

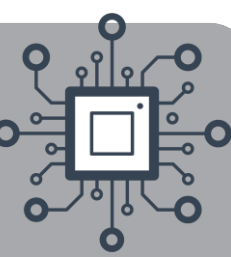
Sample Data



Specific Activity Inputs & Outputs

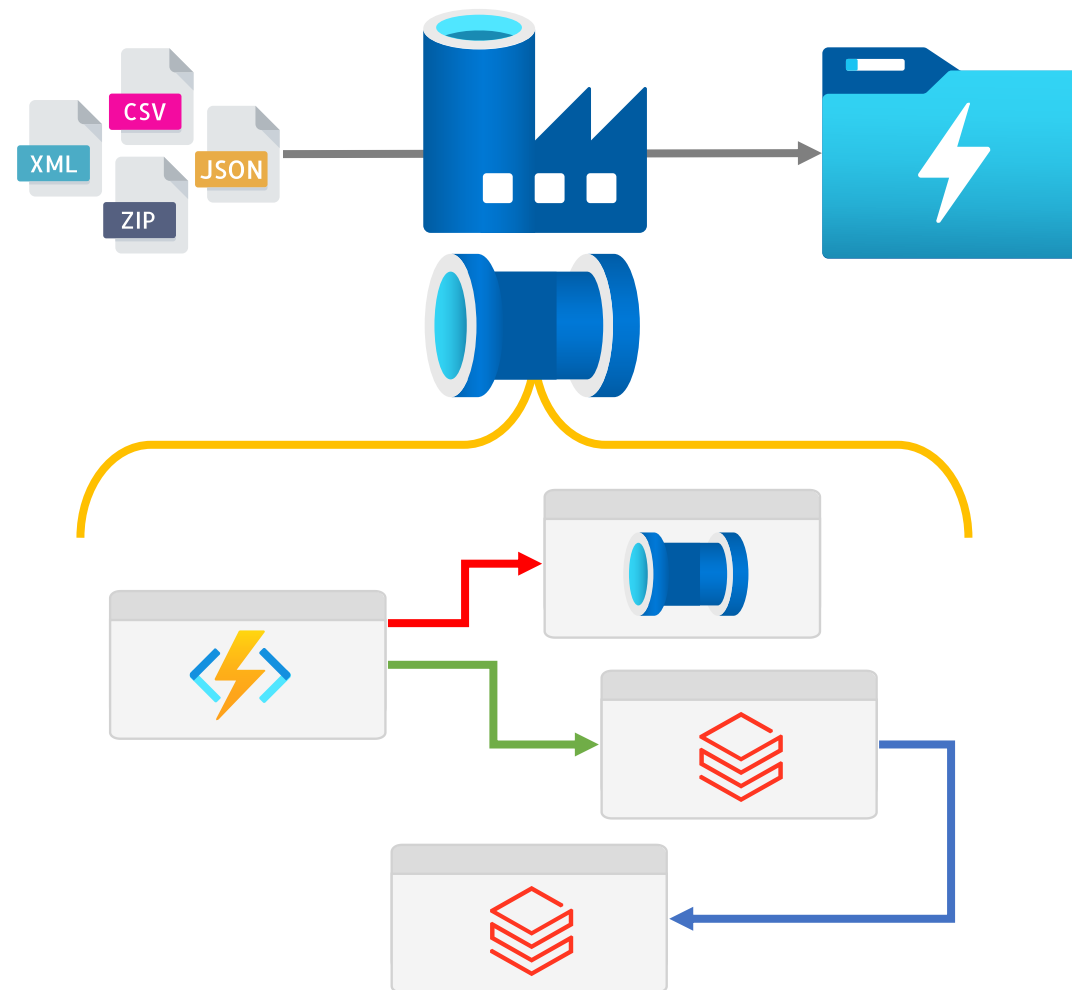


Types of Testing

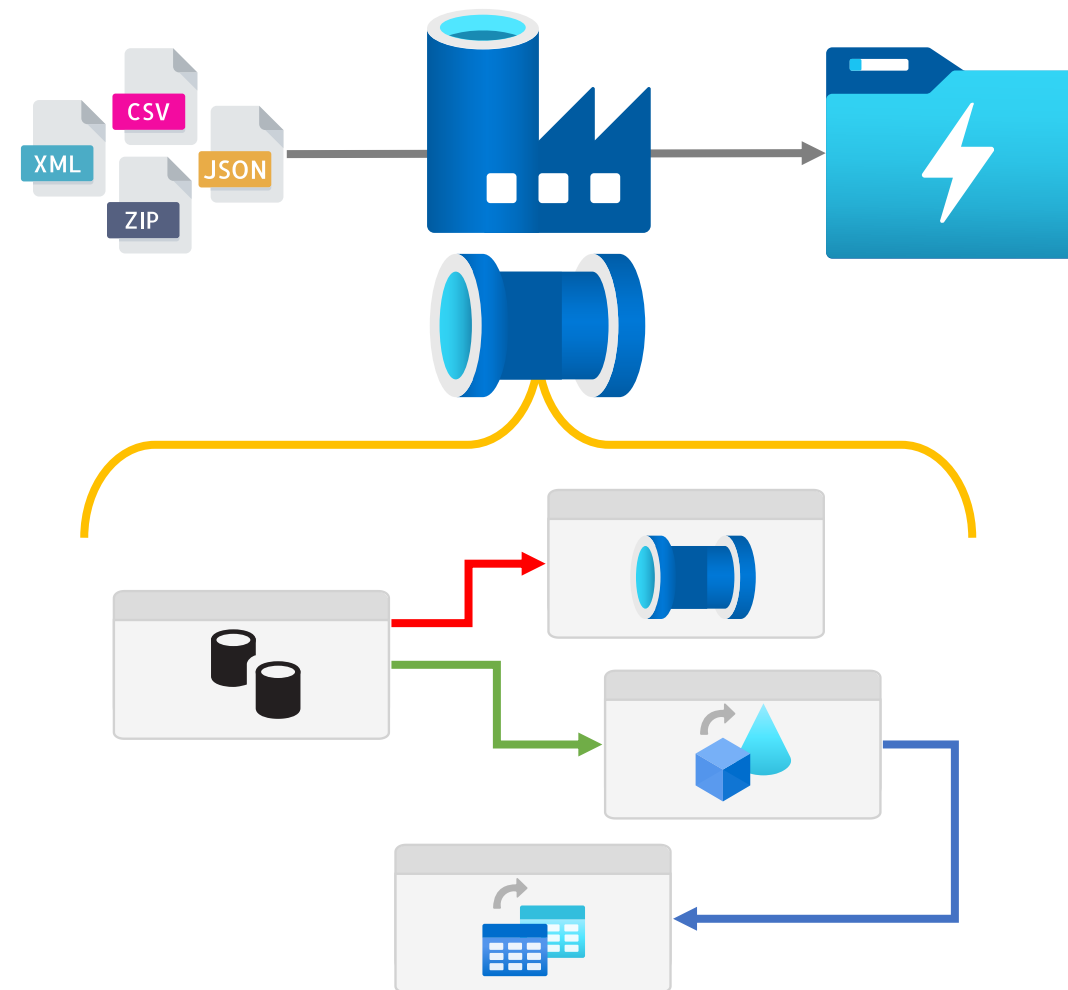


Does it depend on external resources?

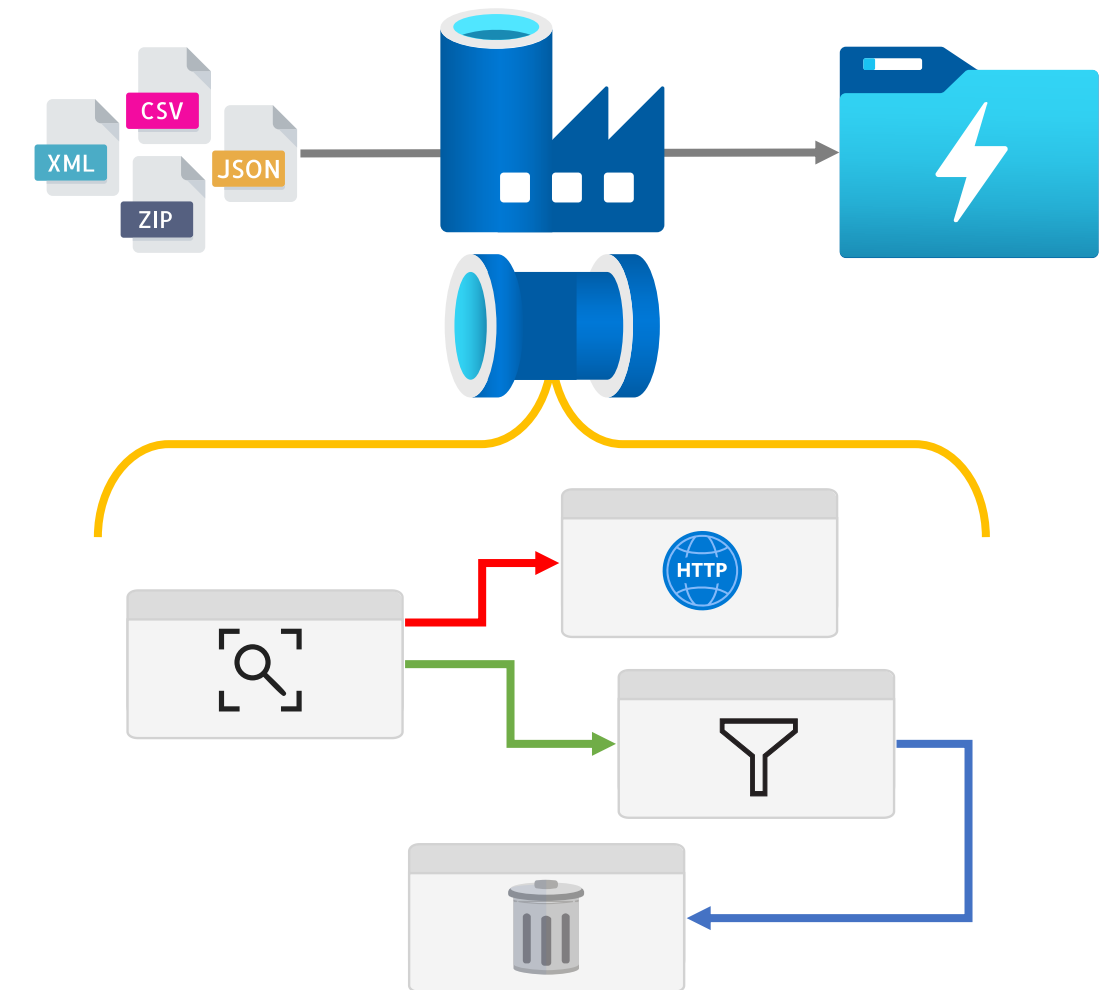
Integration Tests

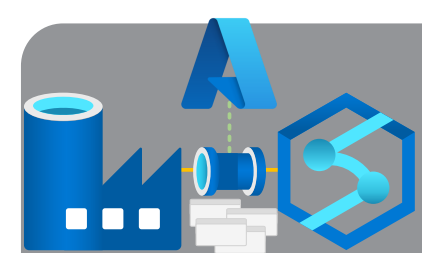


Functional Tests

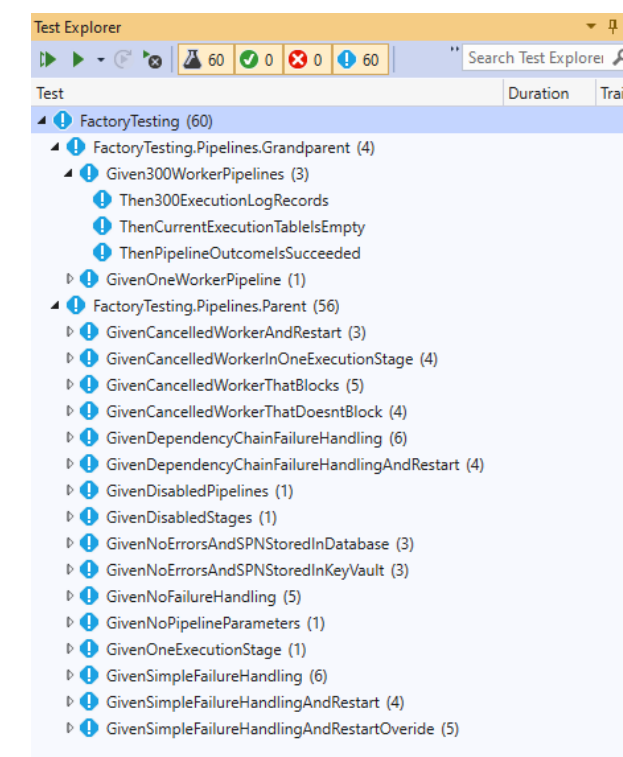
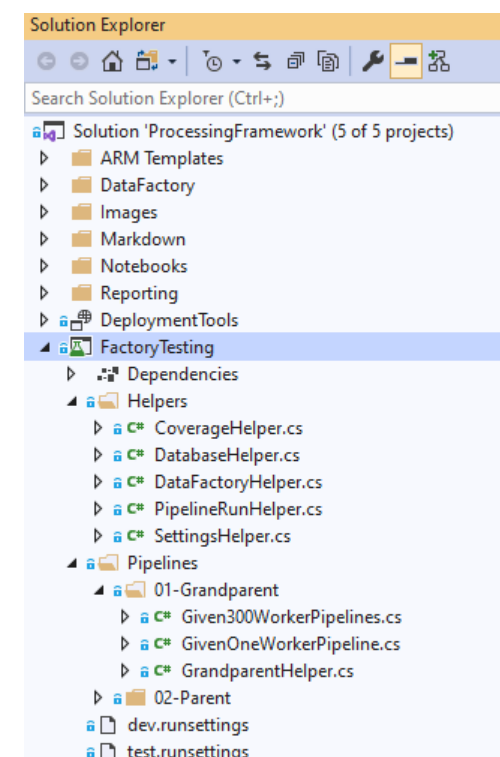
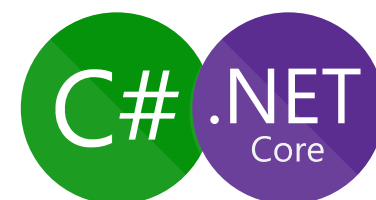
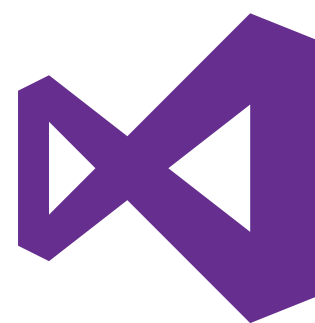
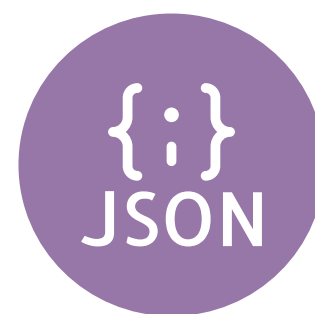
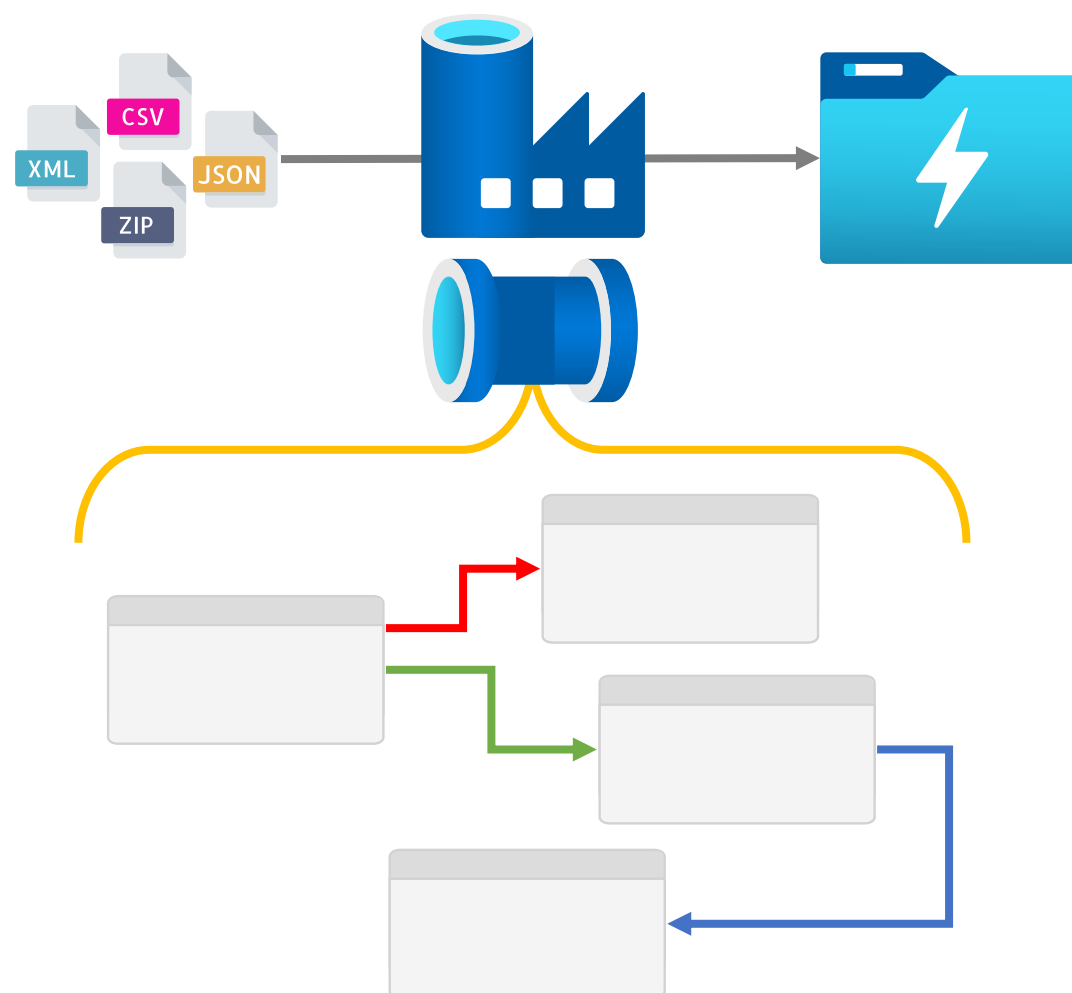
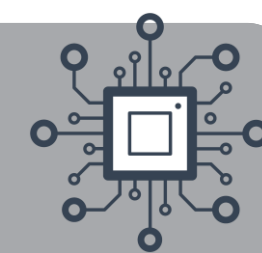


Unit Tests

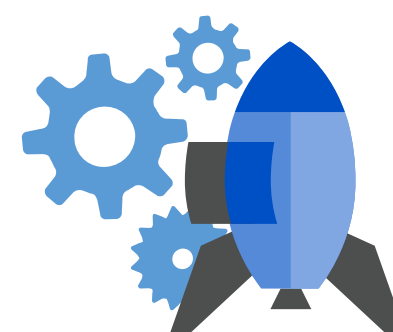
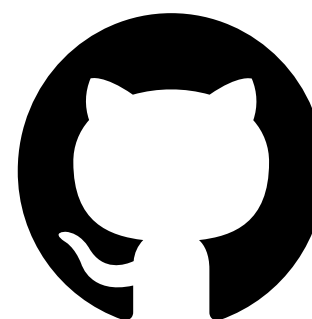
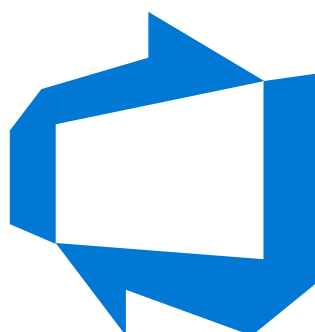


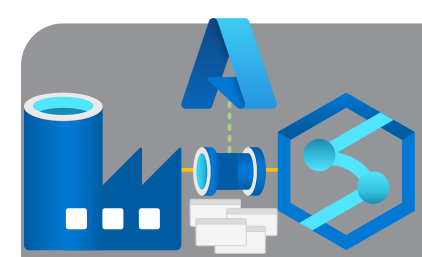


How To Run & Automate Tests

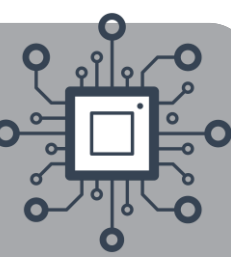


github.com/richardswinbank/community/tree/main/adf-testing-series

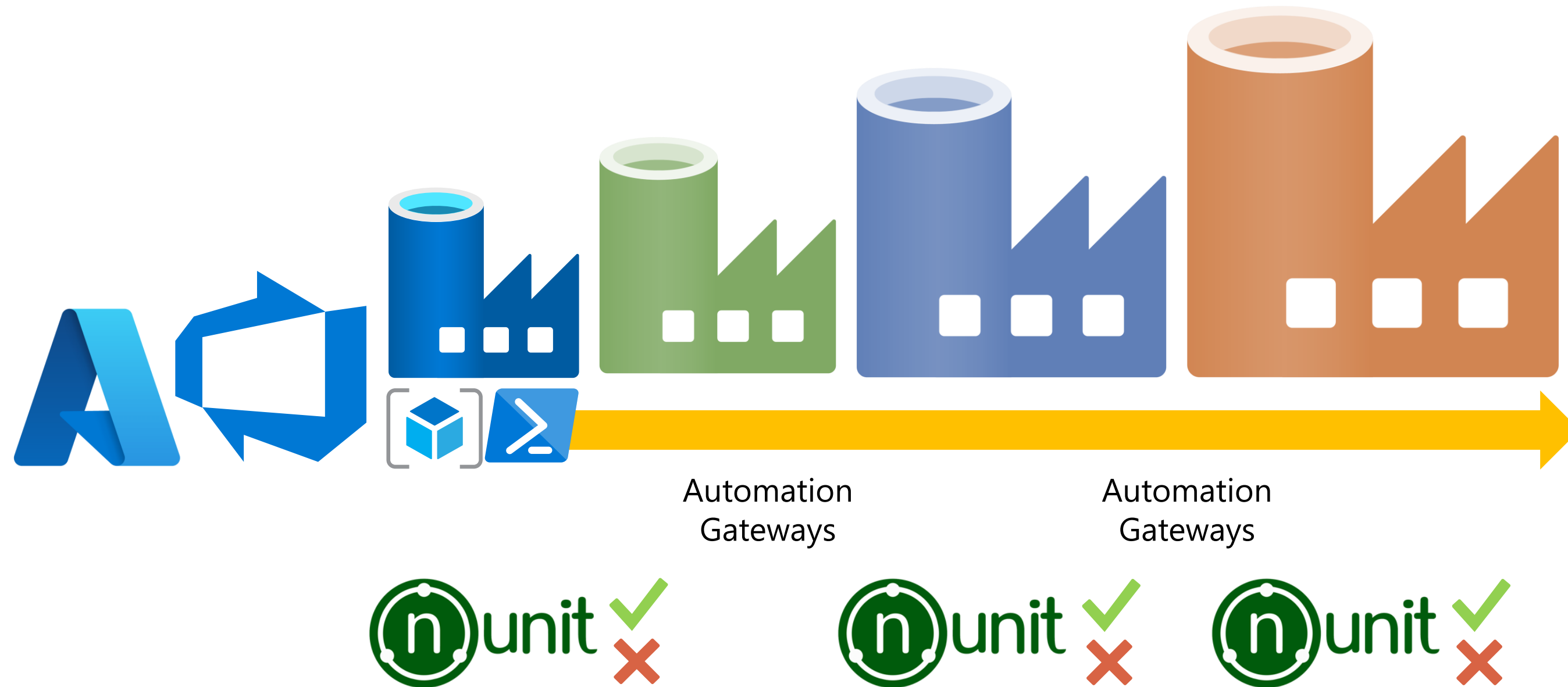




Testing Code vs Actual Code



Data Factory – Chicken vs Egg

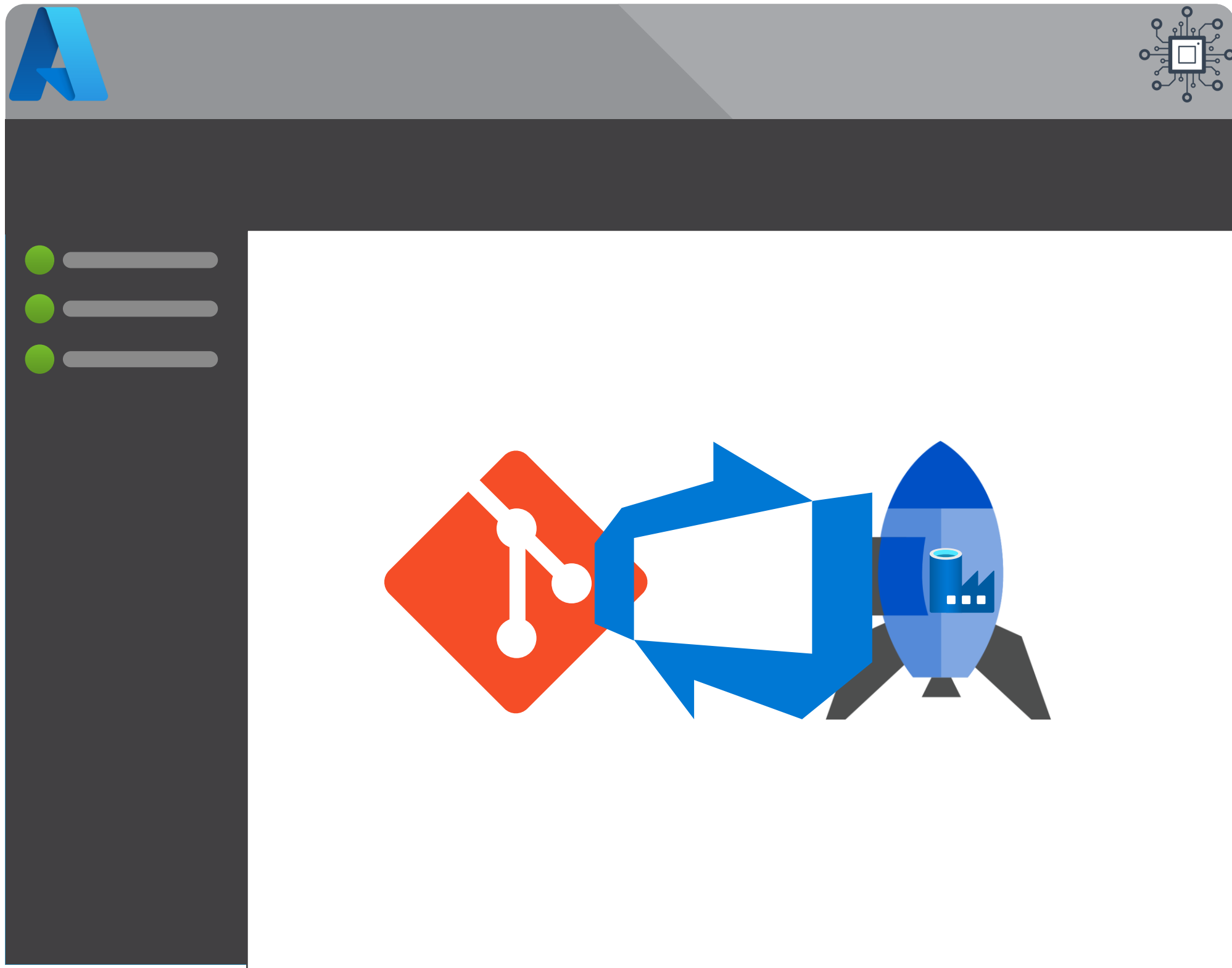


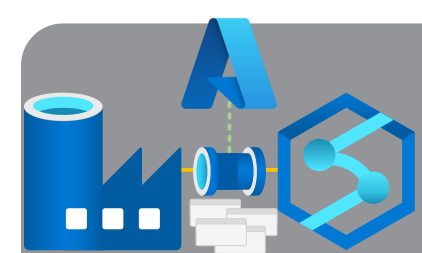
Should automate this?

When should testing scripts run?

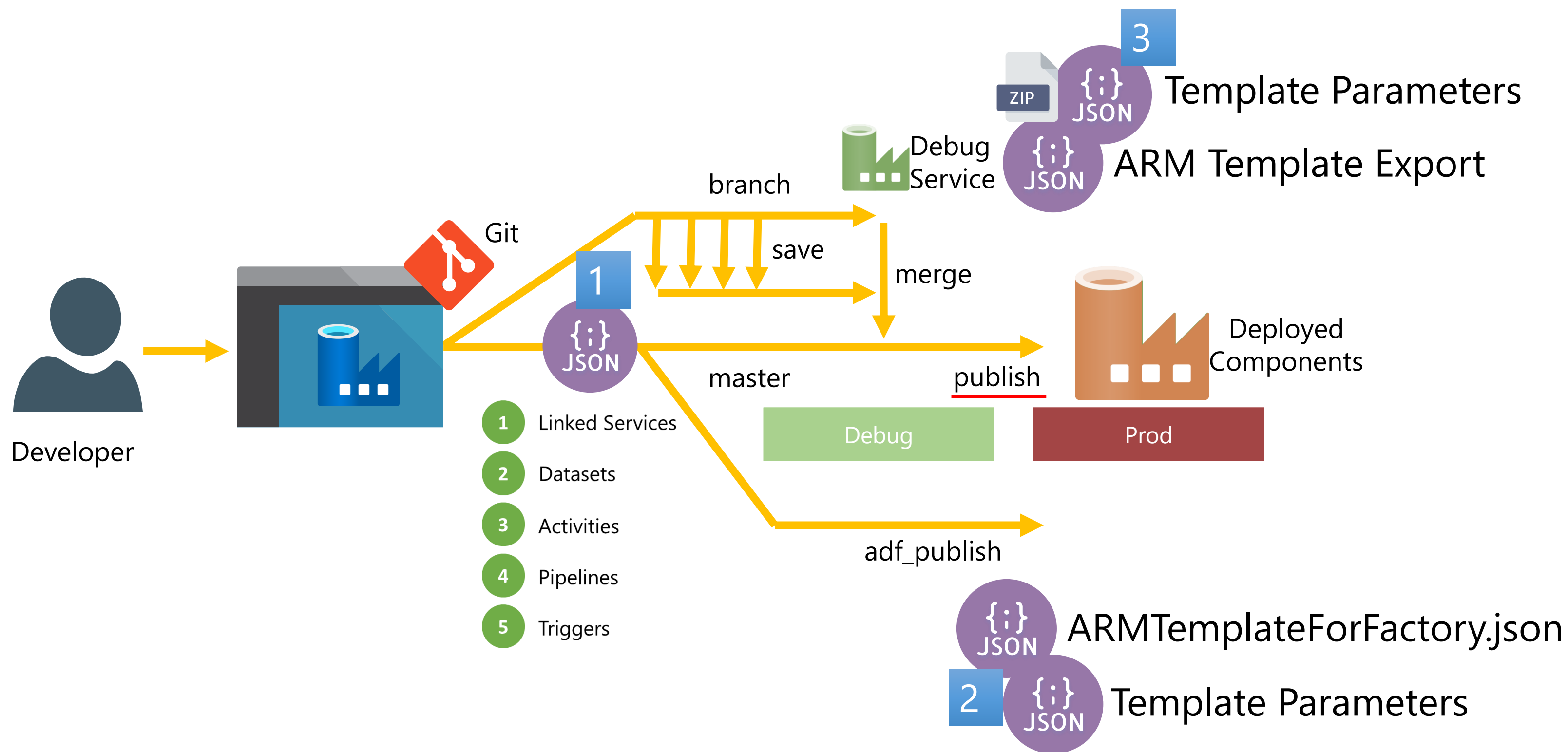
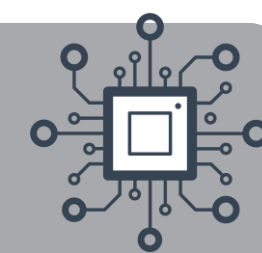
What if a test case fails?

CI/CD



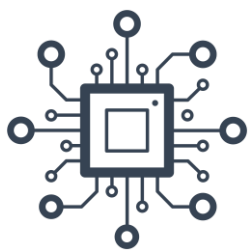
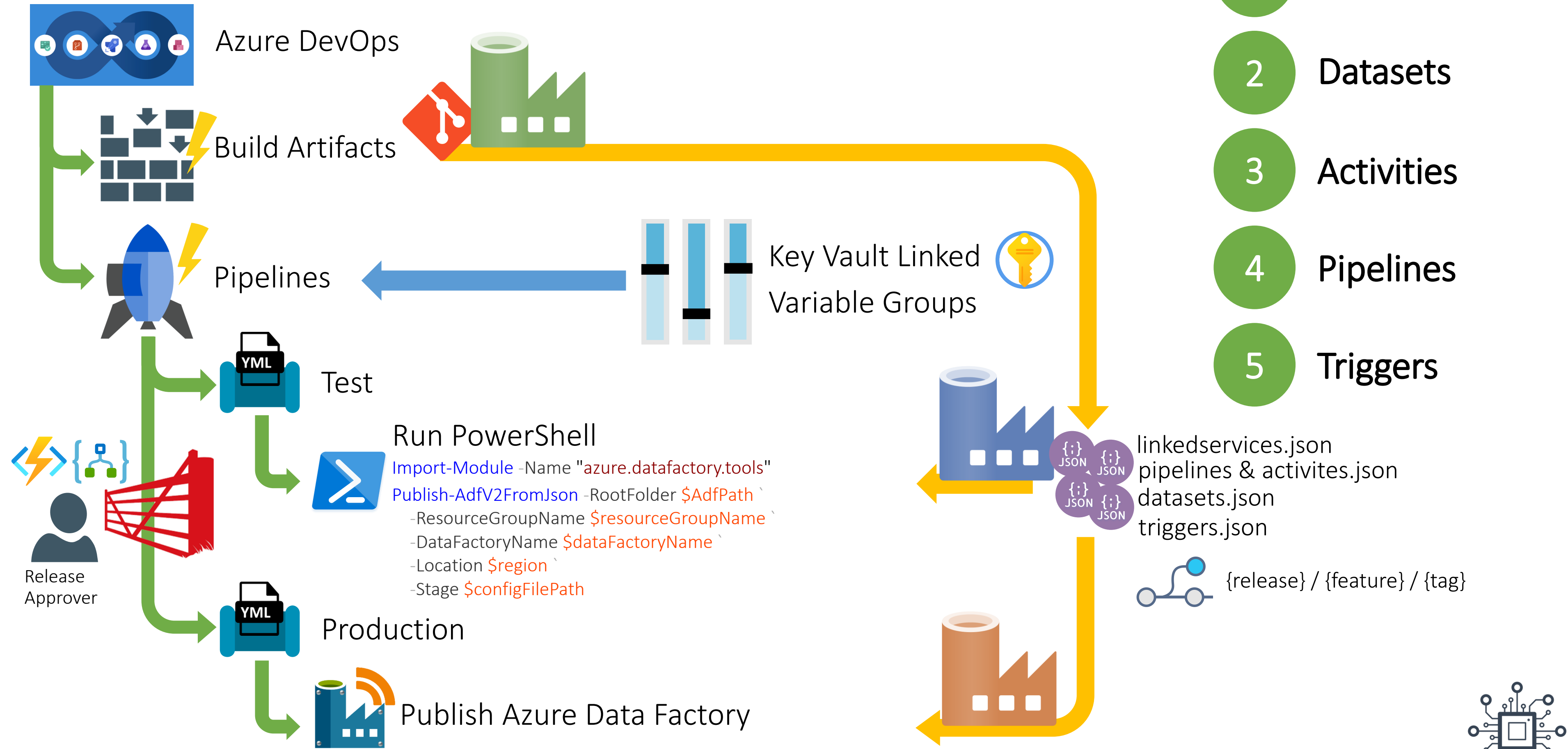


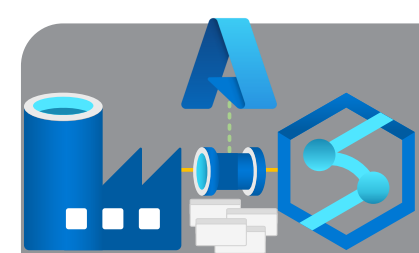
Getting Our ADF Source Code



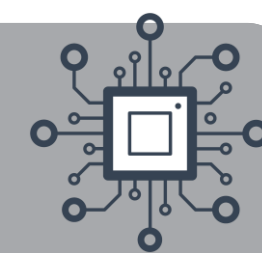
Data Factory Continuous Delivery - Complex

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

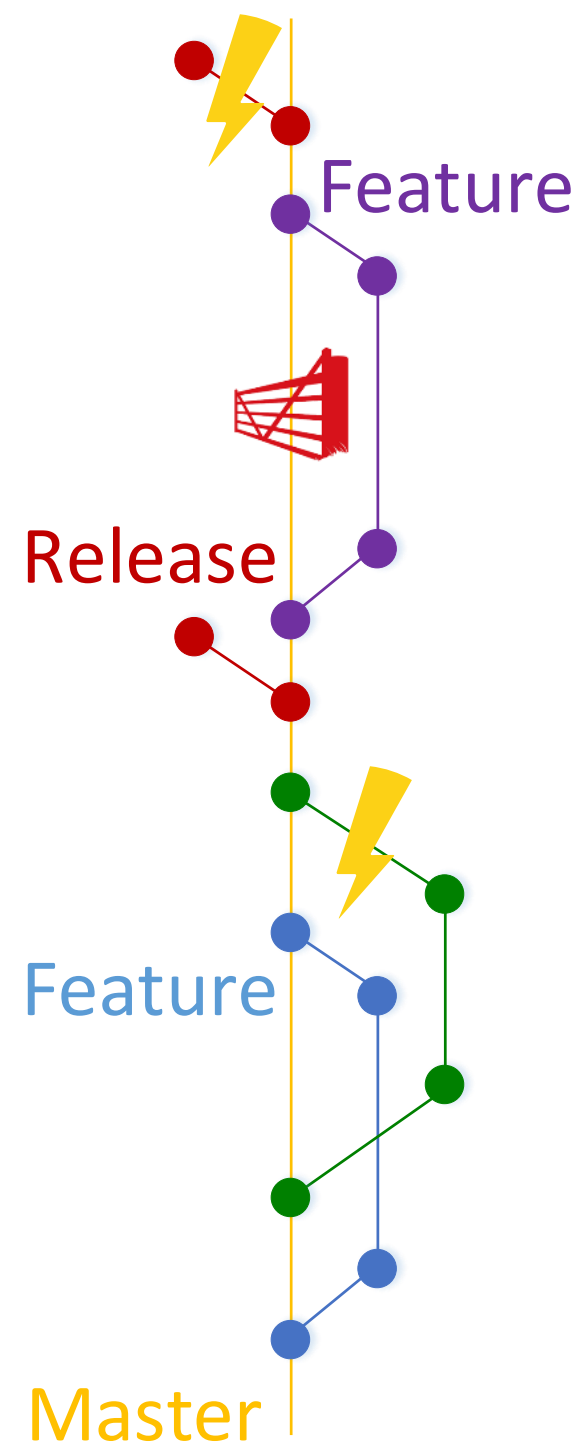




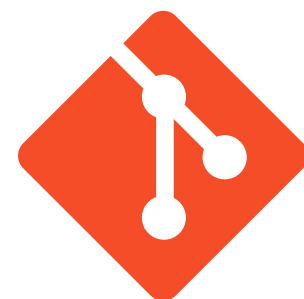
Data Factory DevOps Story Summary



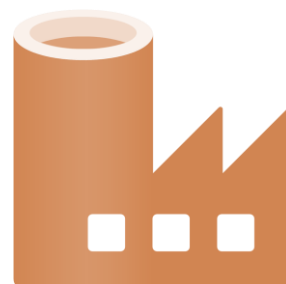
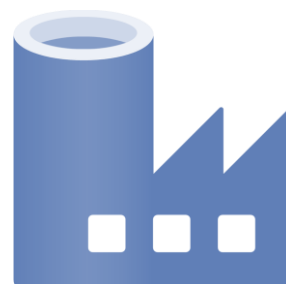
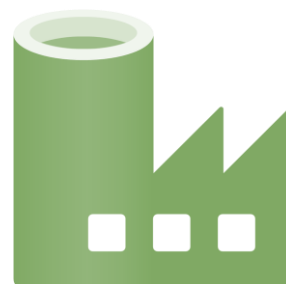
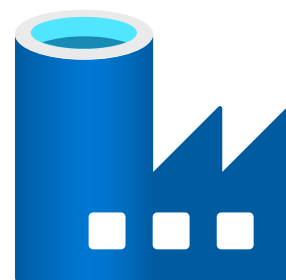
What is your code branching strategy?



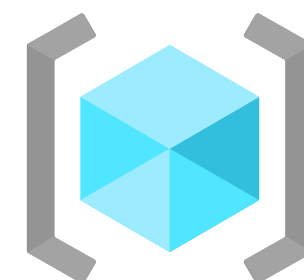
Which source control tool to use?



How many environments do we want?



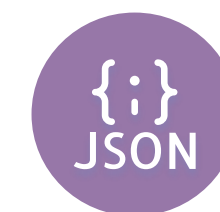
What deployment method do we want to use?



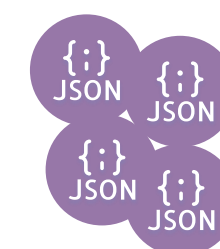
What artifacts are we going to use?...

OR

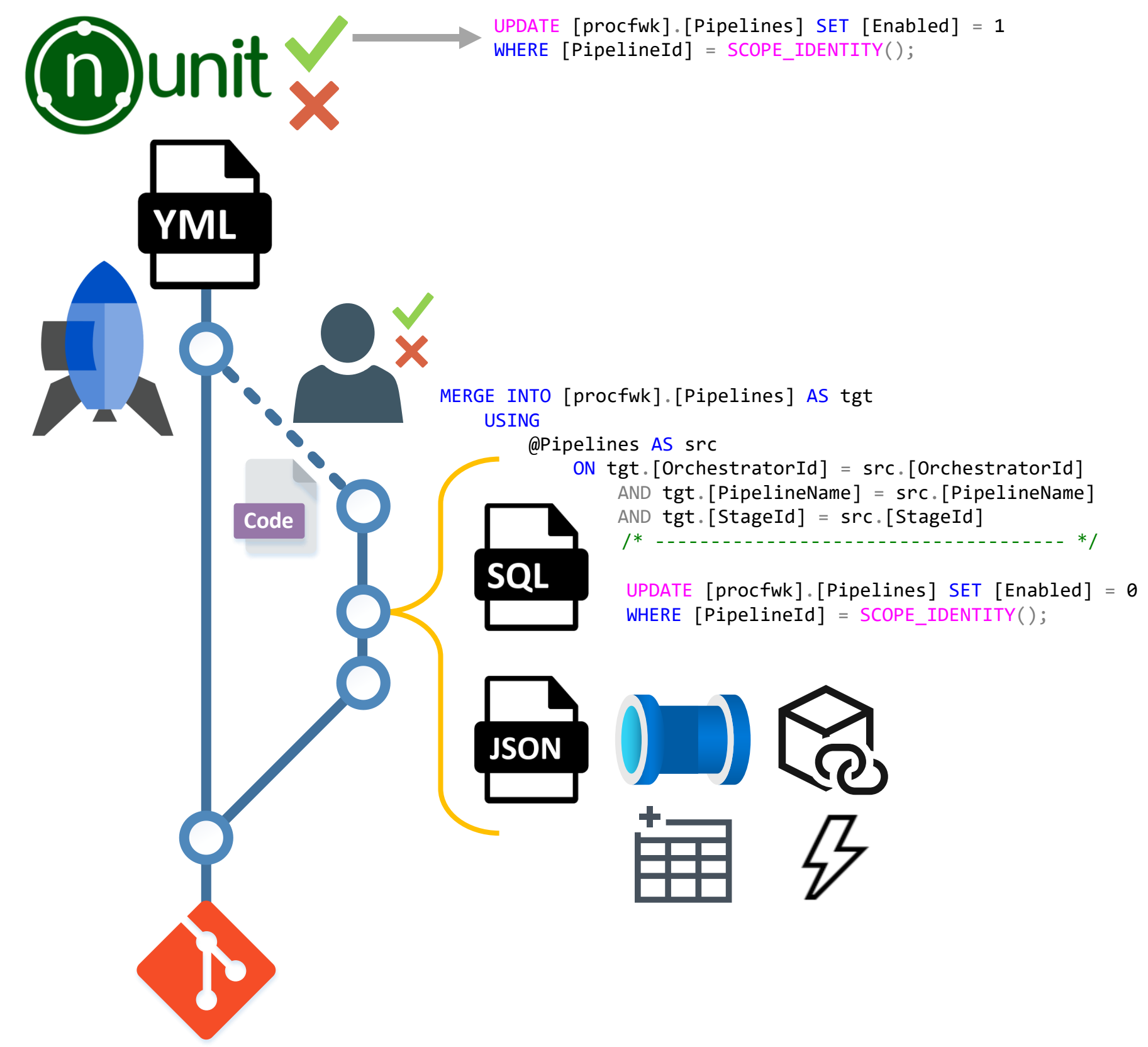
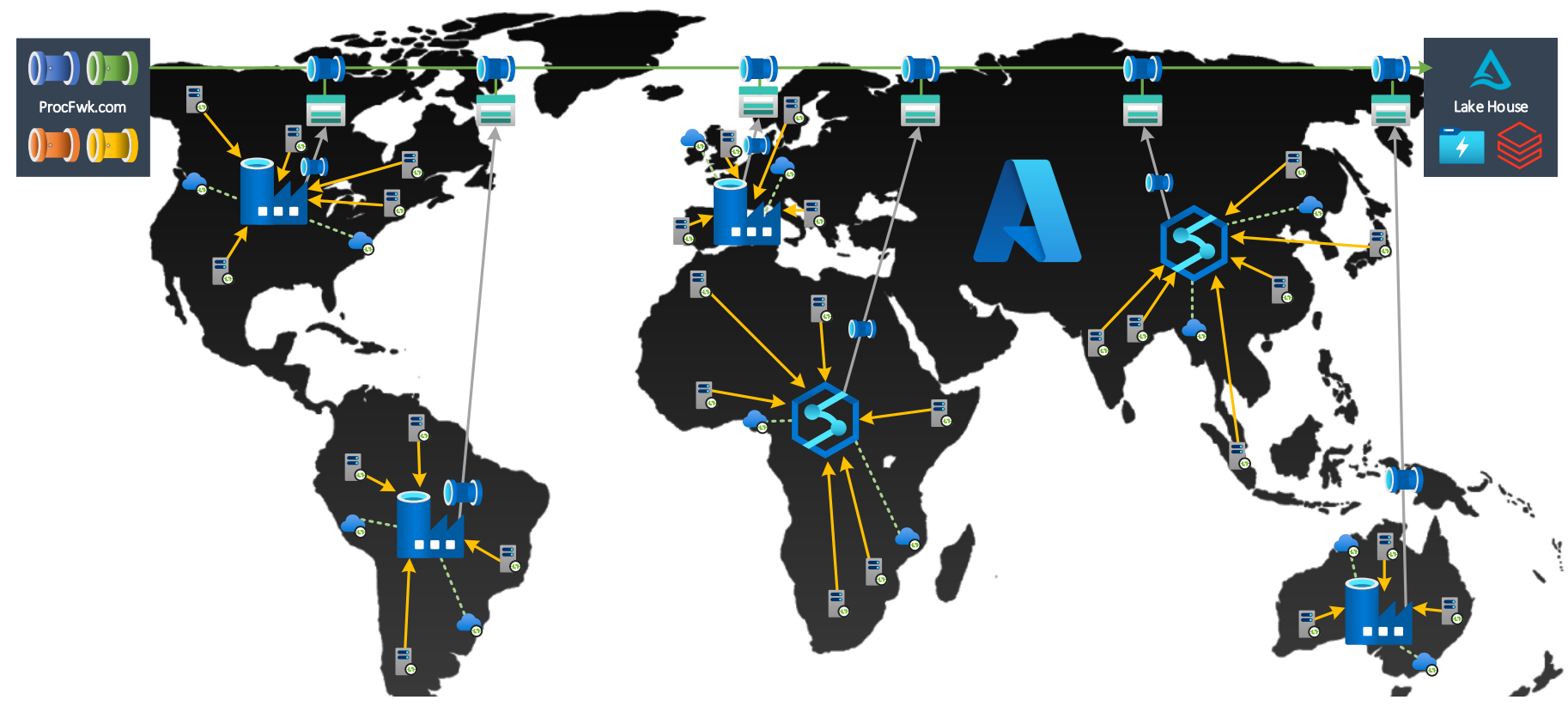
How much control do you want?

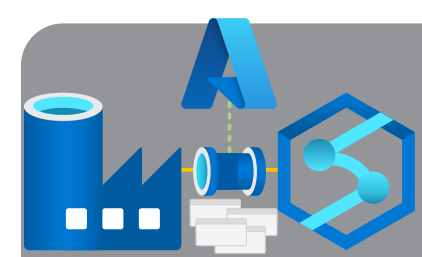


ARMTemplate
ForFactory.json

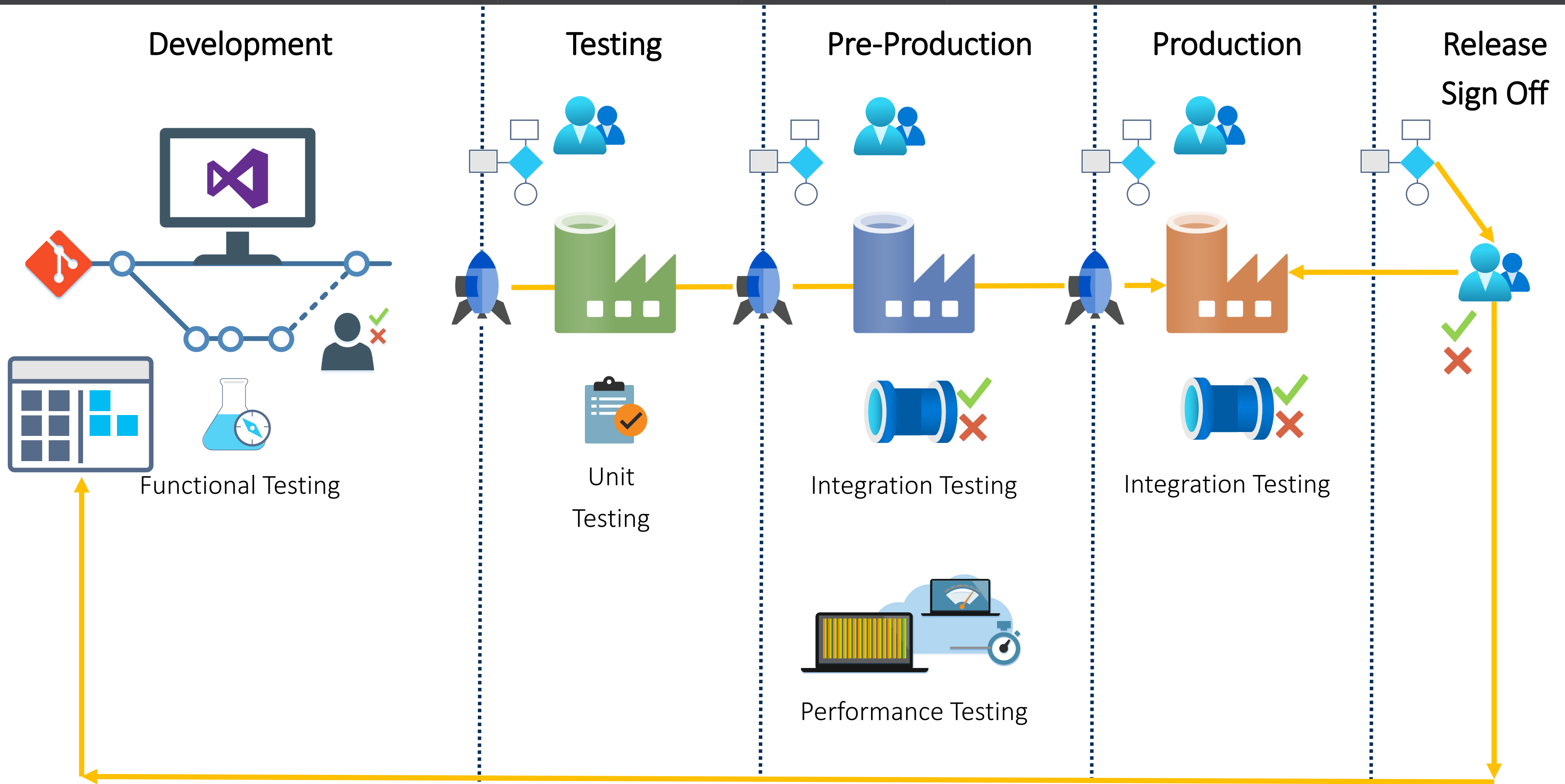
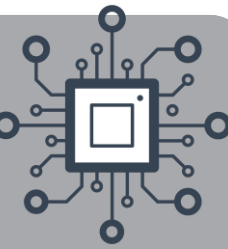


linkedservices.json
pipelines &
activites.json
datasets.json
triggers.json



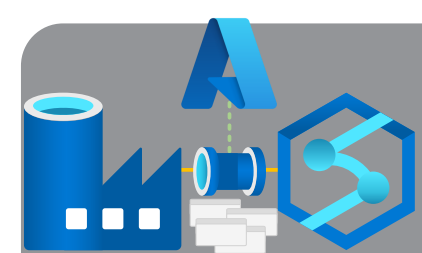


Deployment Life Cycle & Gateway

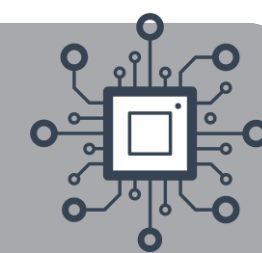


VNet Integration

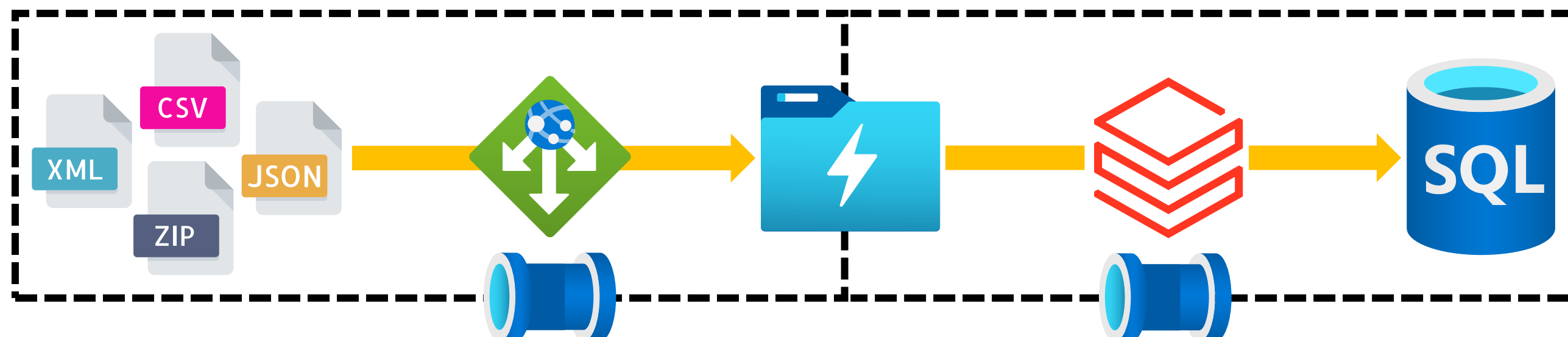




Integration Pipelines as Data Engineers



Control Flow



1

Linked Services



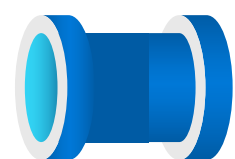
2

Datasets



3

Activities



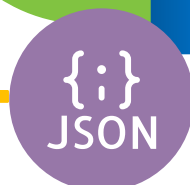
4

Pipelines



5

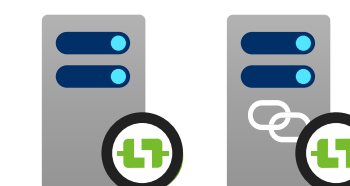
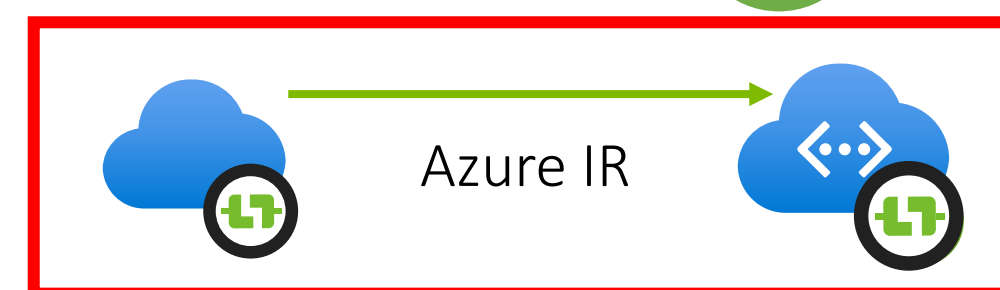
Triggers



Add dynamic content [Alt+P]

Integration Runtimes

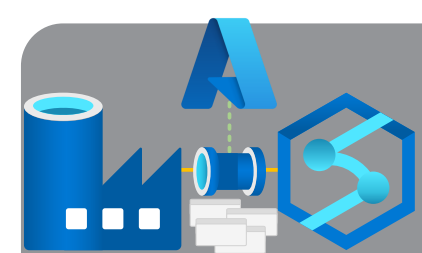
6



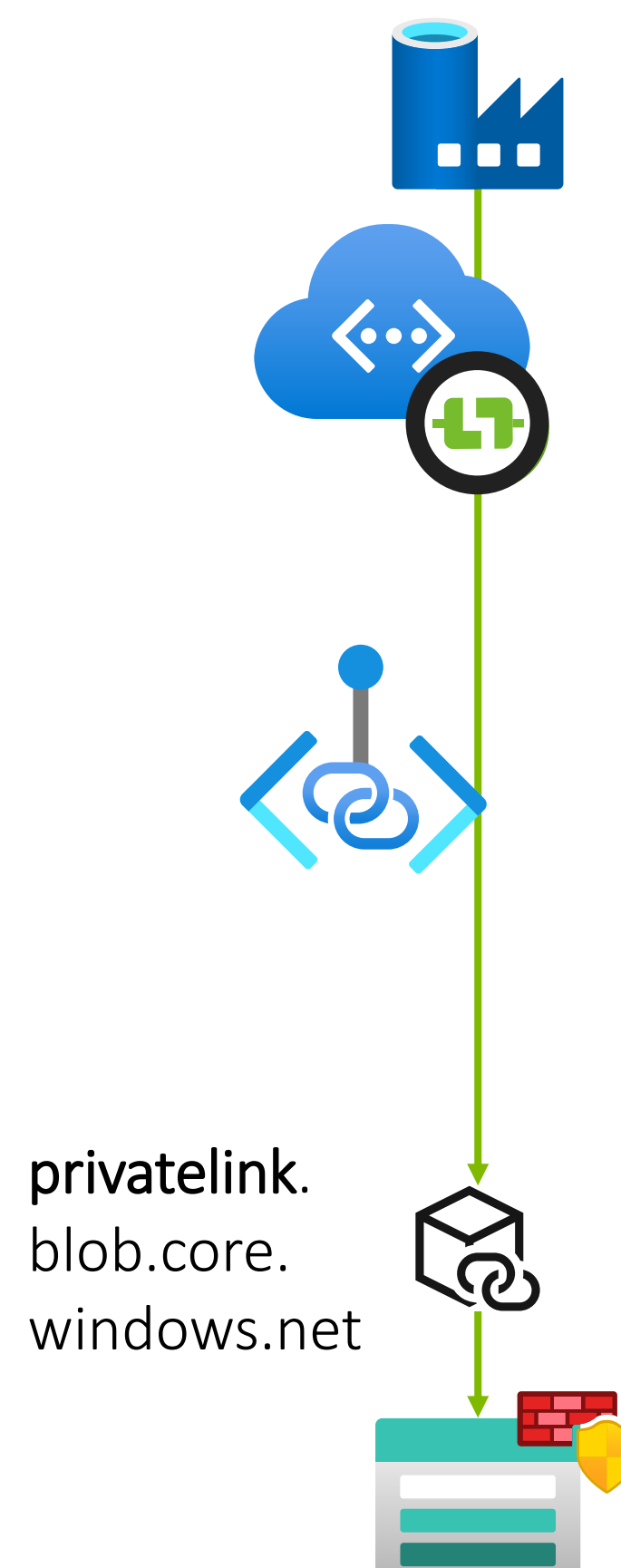
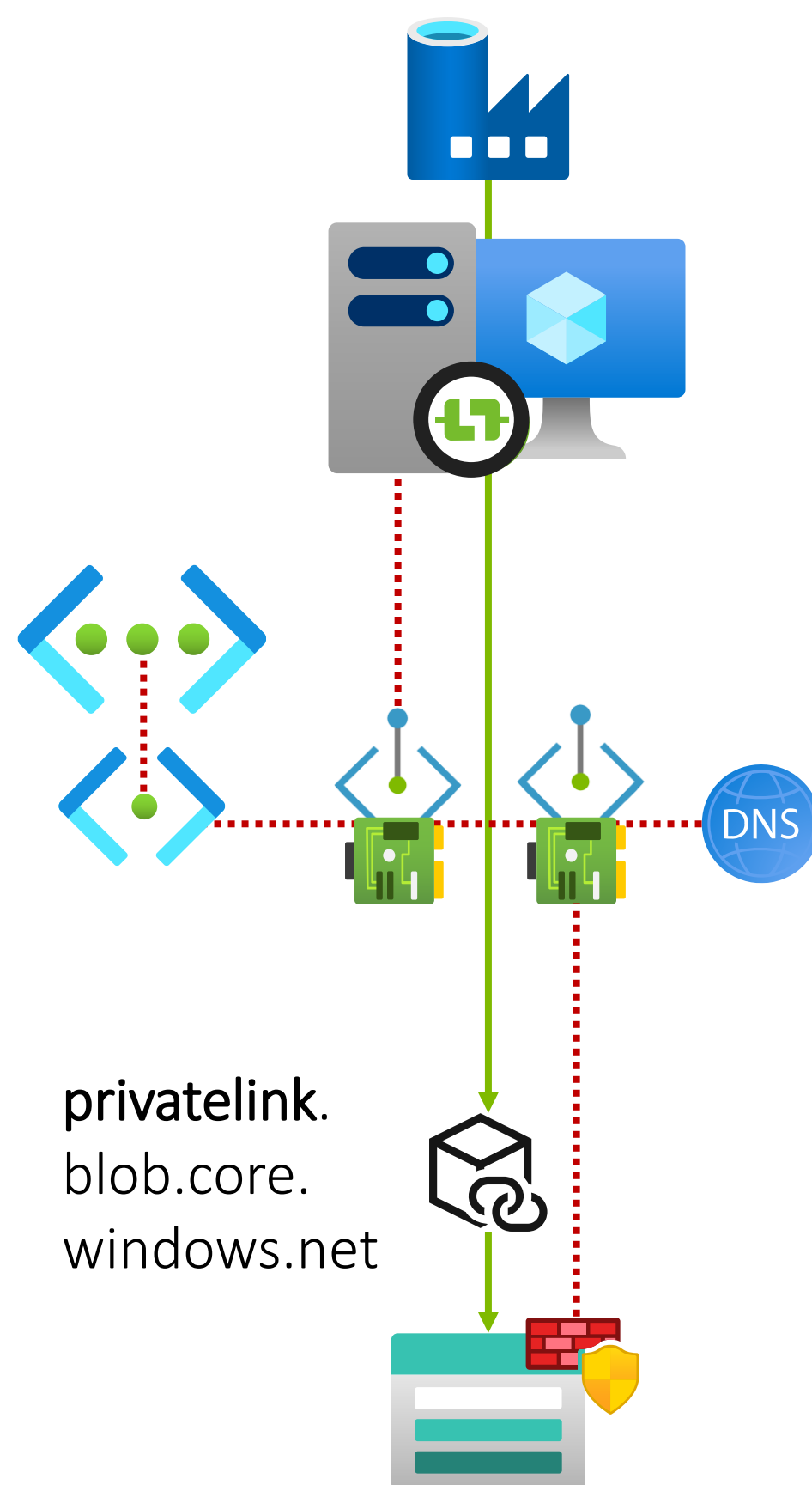
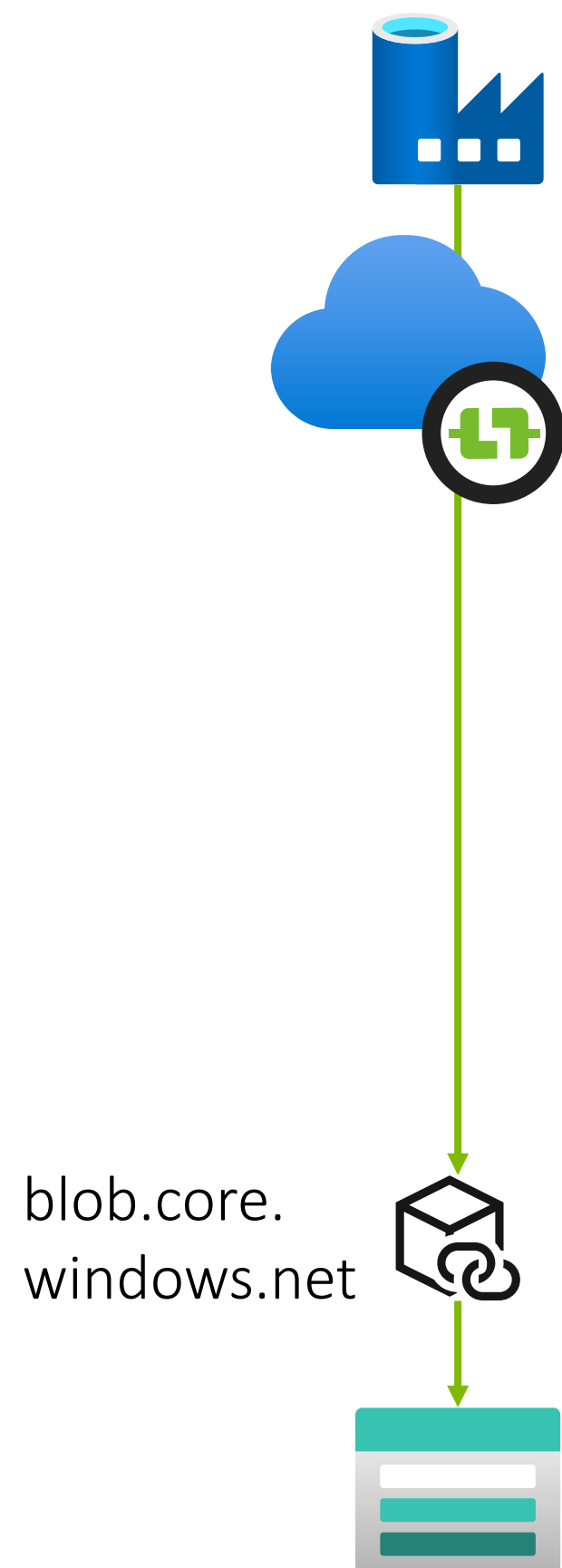
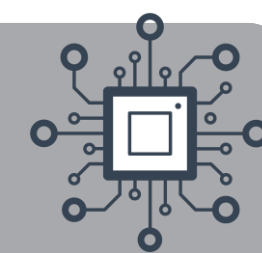
Hosted IR

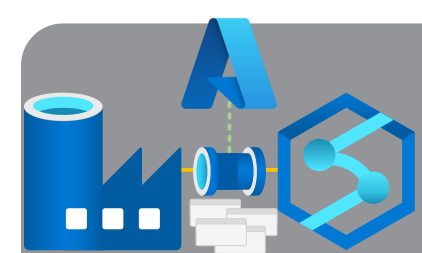


SSIS IR

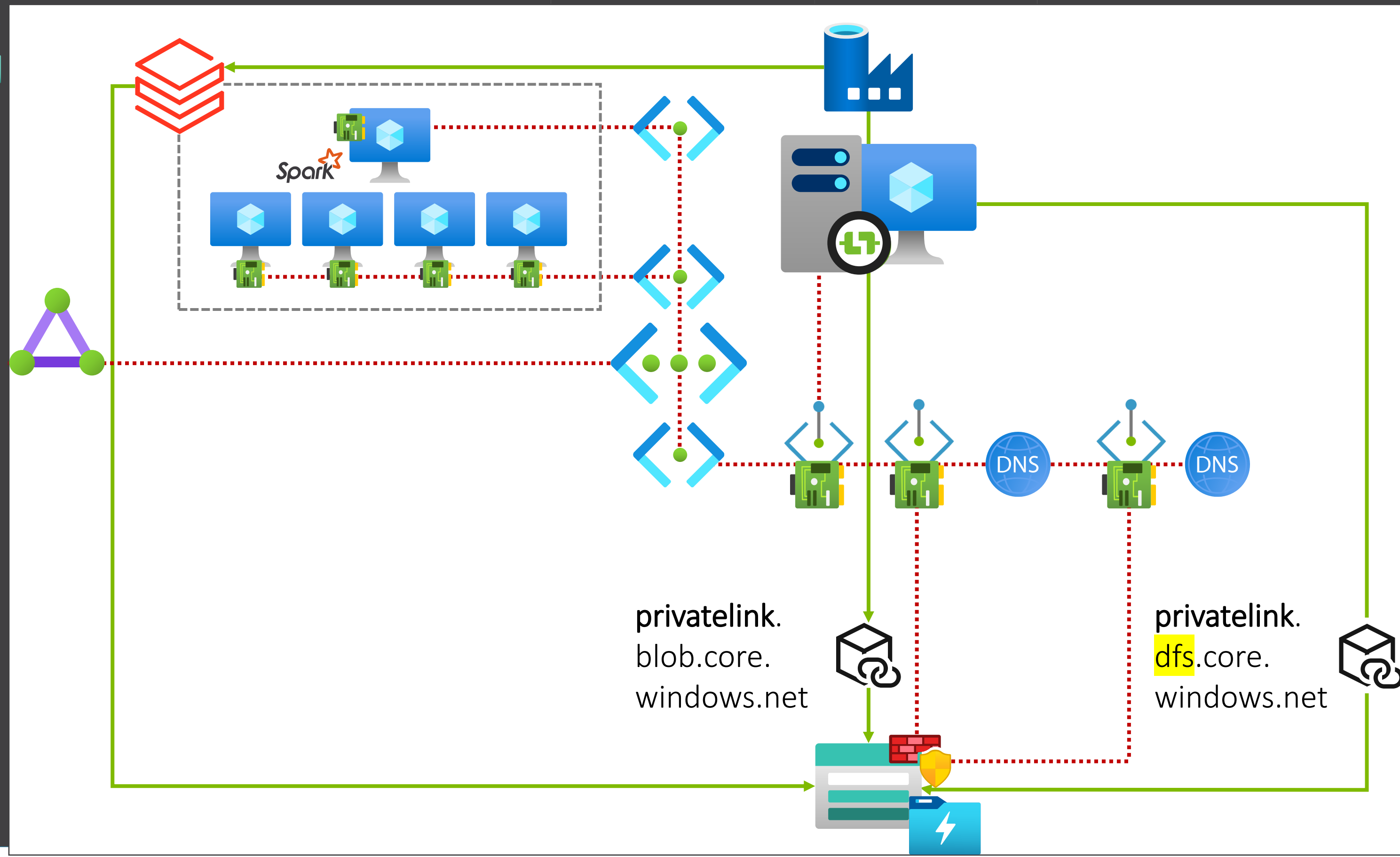
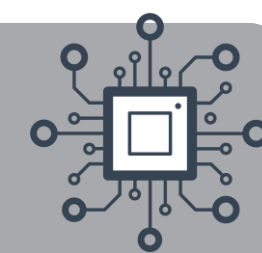


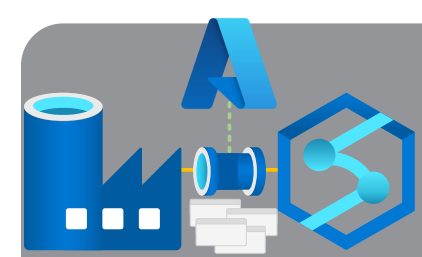
Managed vs Unmanaged Connections



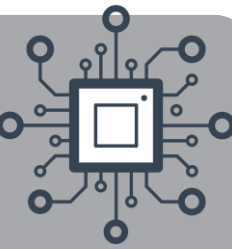


Further VNet Connections

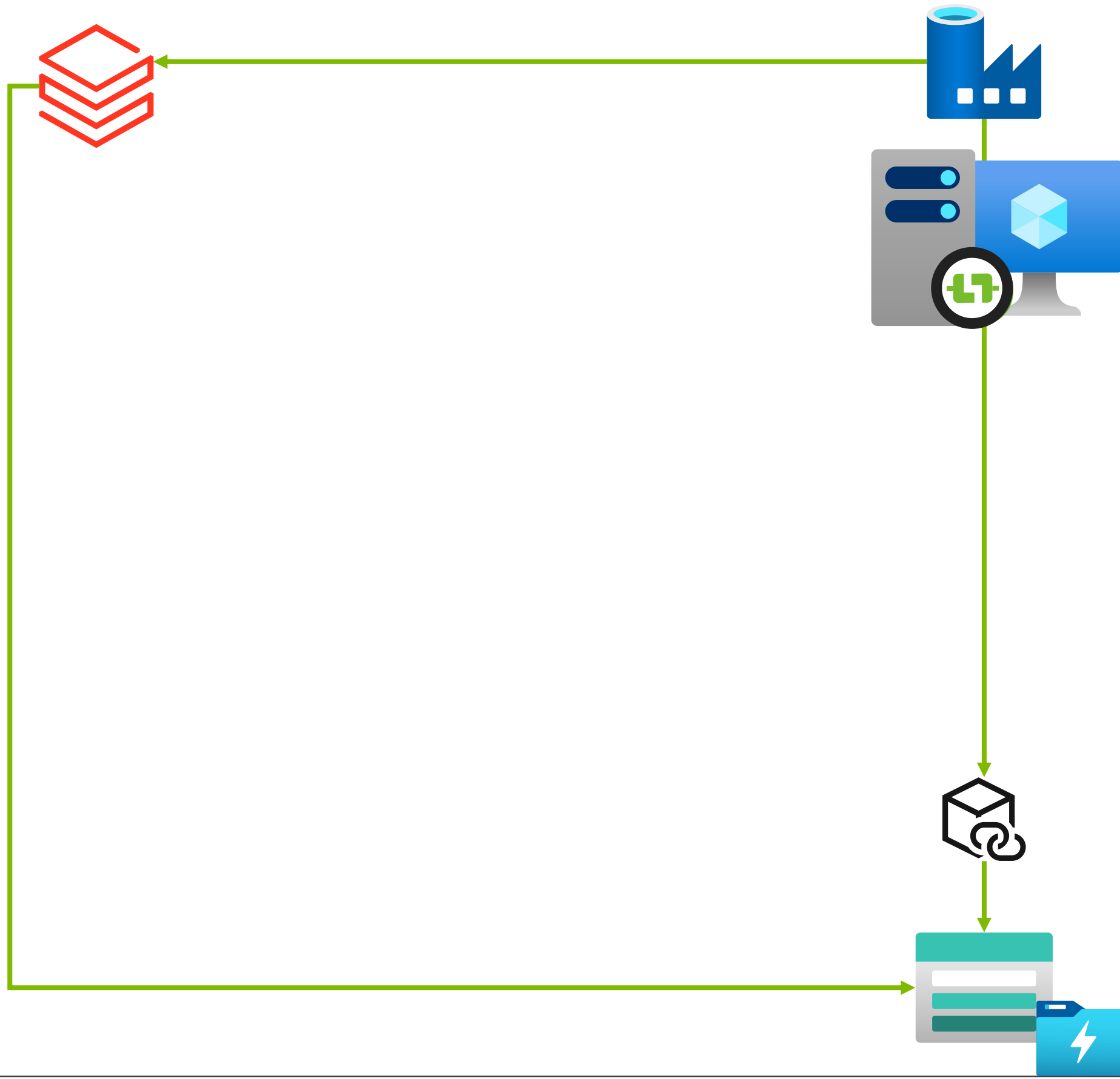


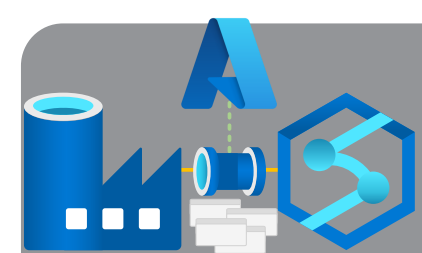


Further ~~V~~Net Connections

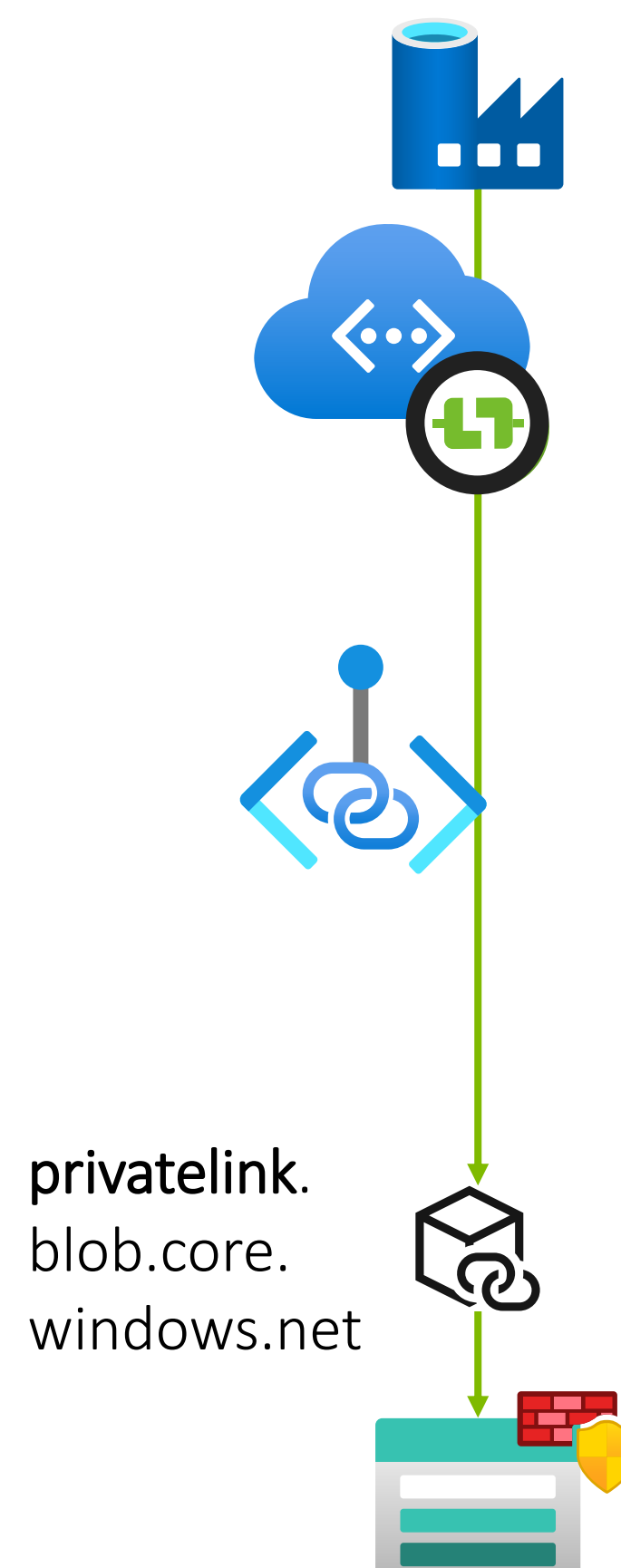
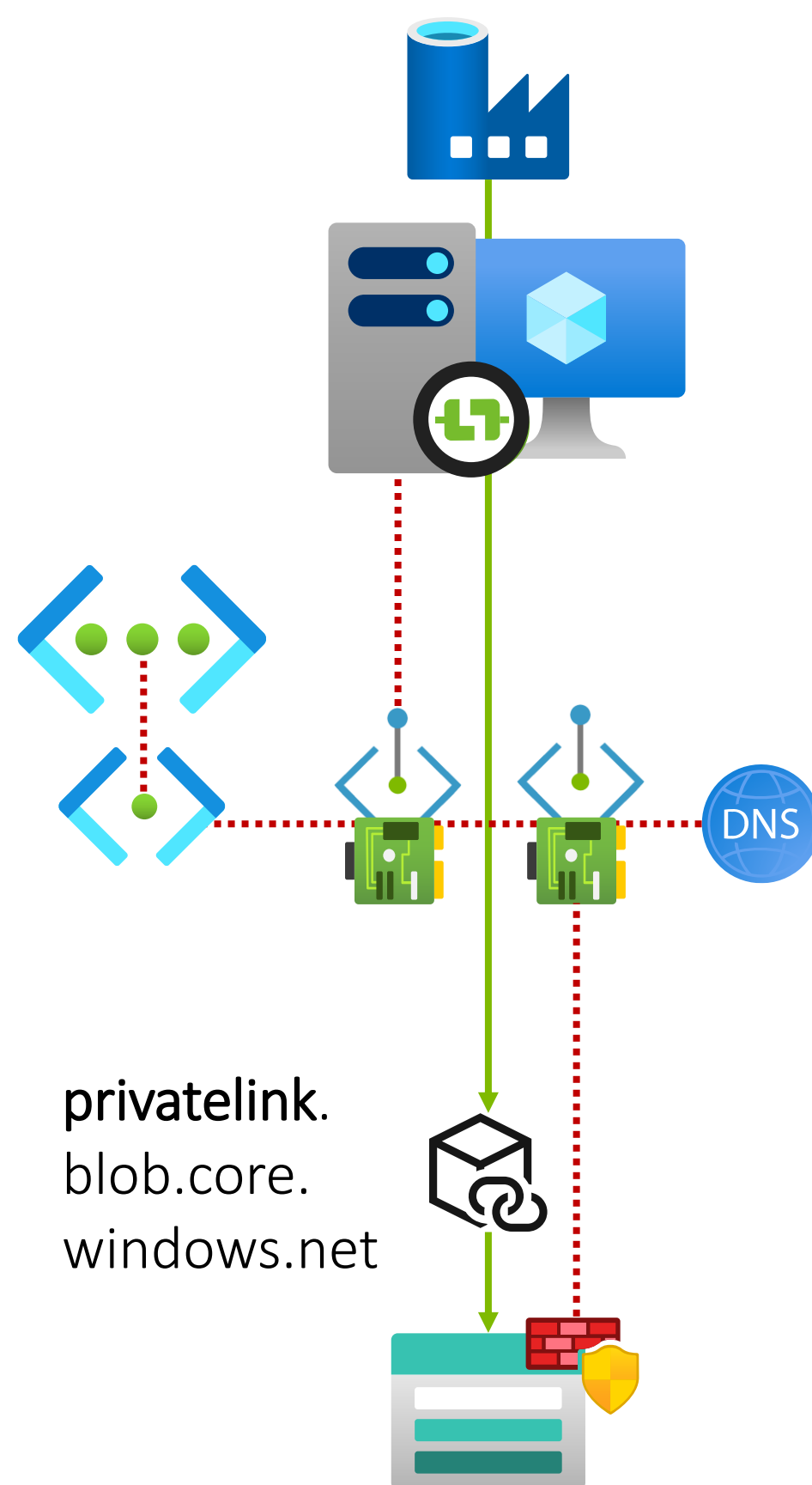
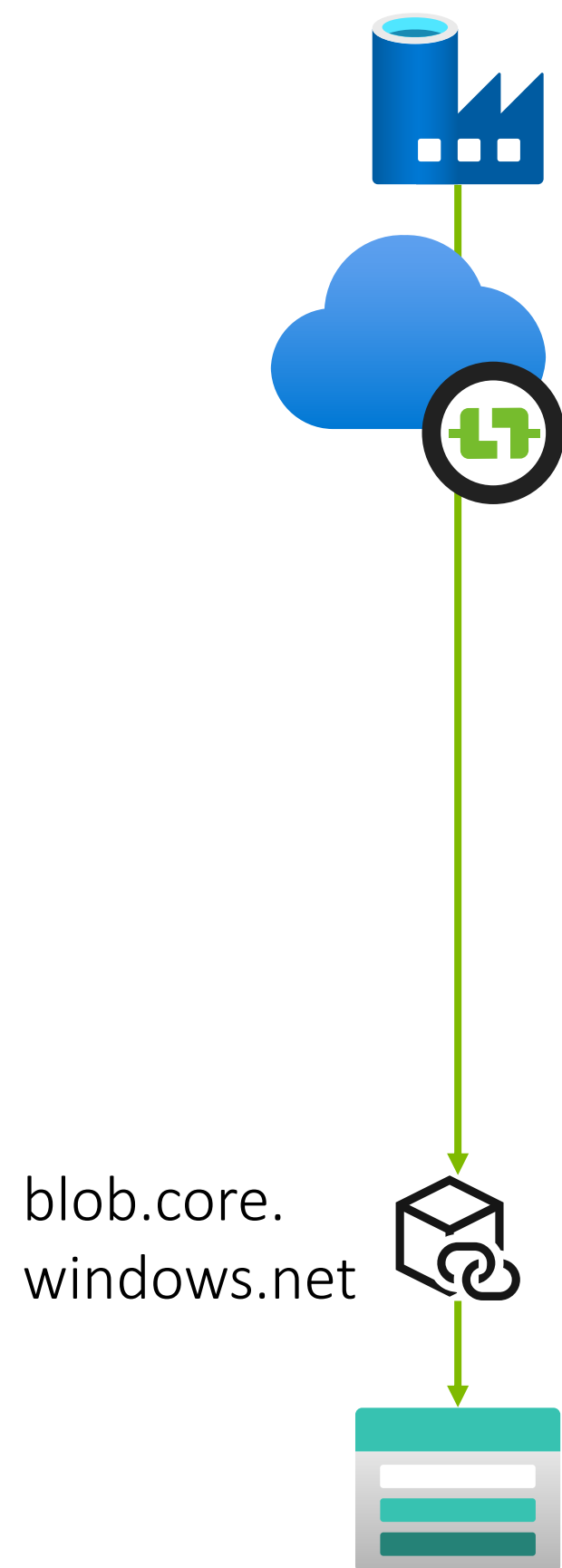
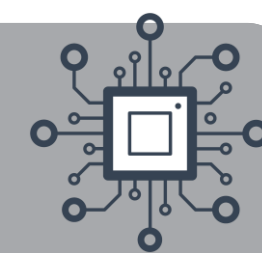


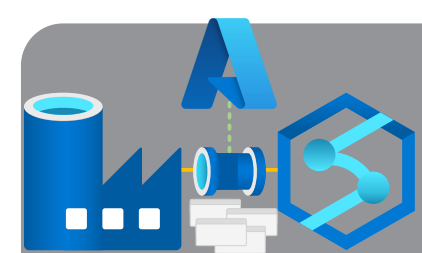
For a Data Engineer



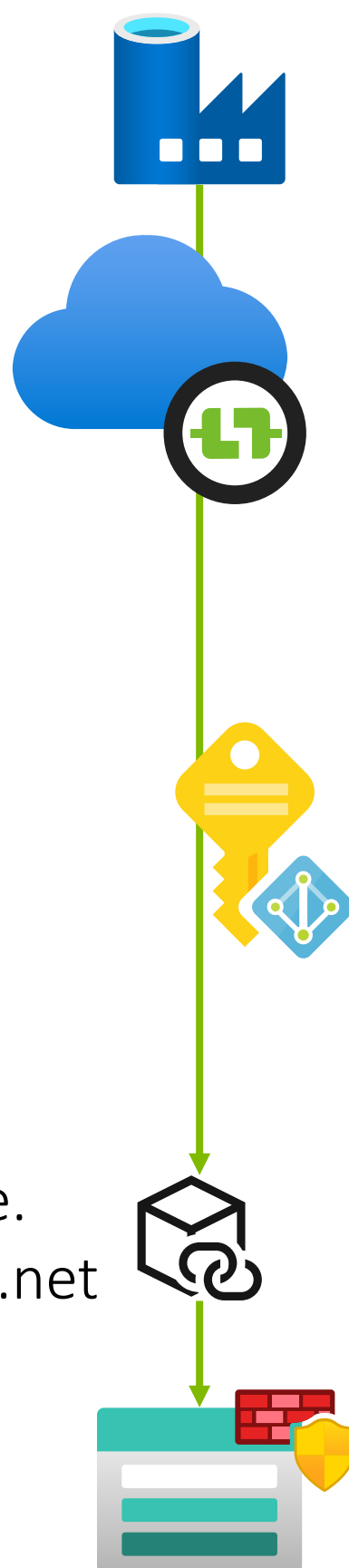
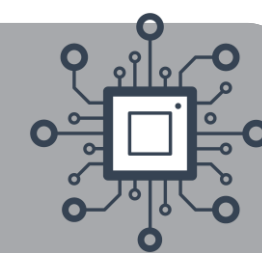


Managed vs Unmanaged Connections





Public Connections



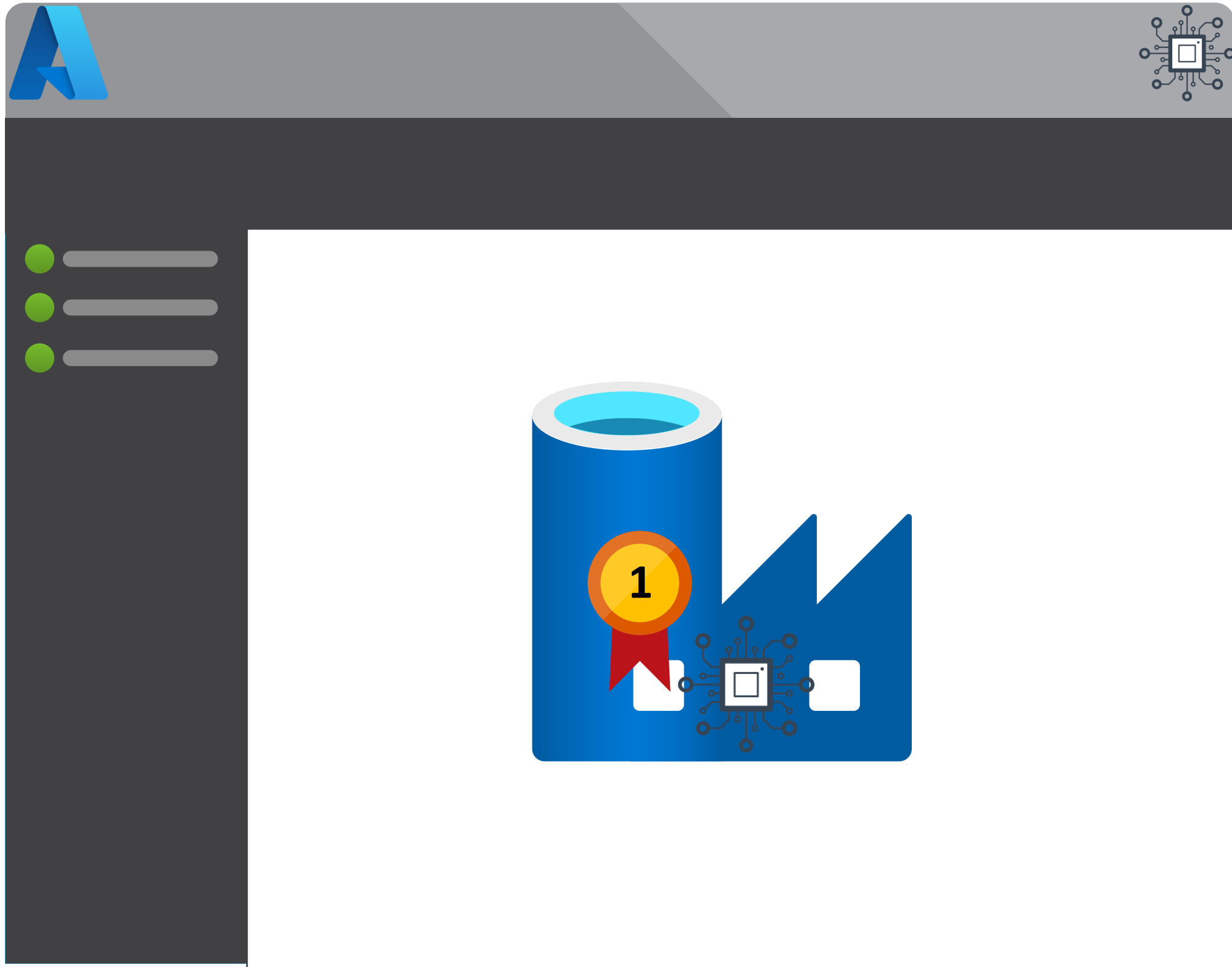
Managed
Identity

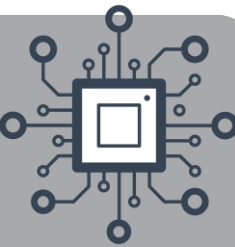
** Trusted access based on system-assigned managed identity.*

Microsoft.DataFactory/factories

blob.core.
windows.net

Best Practice





Environment Setup & Developer Debugging

Deployments

Automated Testing

Naming Conventions

Pipeline Hierarchies

Pipeline & Activity Descriptions

Factory Component Folders

Linked Service Security via Azure Key Vault

Dynamic Linked Services

Generic Datasets

Metadata Driven Processing

Parallel Execution

Hosted Integration Runtimes

Azure Integration Runtimes

Wider Platform Orchestration

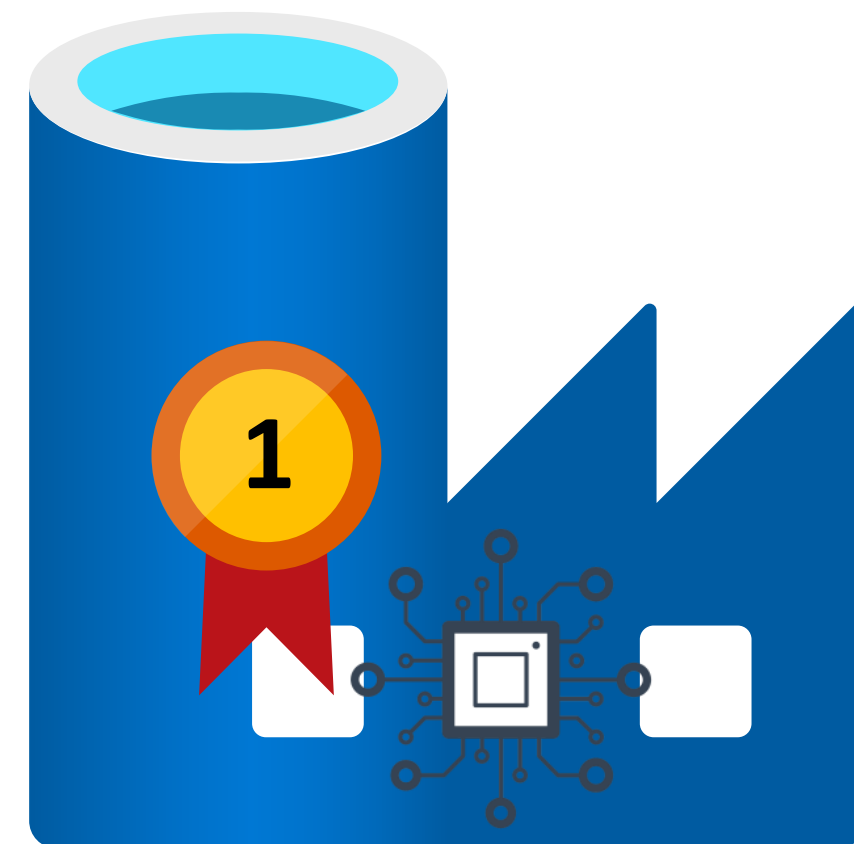
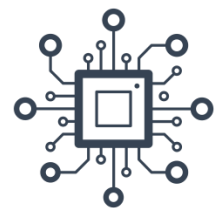
Custom Error Handler Paths

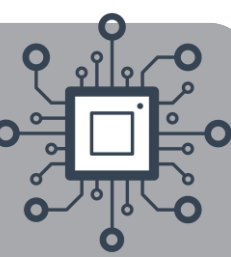
Monitoring via Log Analytics

Service Limitations

Using Pipeline Templates

Documentation





Thank you for listening...

Paul Andrew



Blog: mrpaulandrew.com
YouTube: [c/mrpaulandrew](https://www.youtube.com/c/mrpaulandrew)
Email: paul@mrpaulandrew.com

Twitter: [@mrpaulandrew](https://twitter.com/mrpaulandrew)
LinkedIn: [In/mrpaulandrew](https://www.linkedin.com/company/mrpaulandrew)

GitHub: github.com/mrpaulandrew