Azure Data Integration Pipelines

In Production *(The Hard Stuff)*

Paul Andrew | Technical Architect in Azure CoE

Microsoft® MVP Most Valuable Professional

avanade

@MrPaulAndrew

In/MrPaulAndrew

MrPaulAndrew.com

https://github.com/mrpaulandrew

CommunityEvents

Demo code, content and slides from various community events.

C++

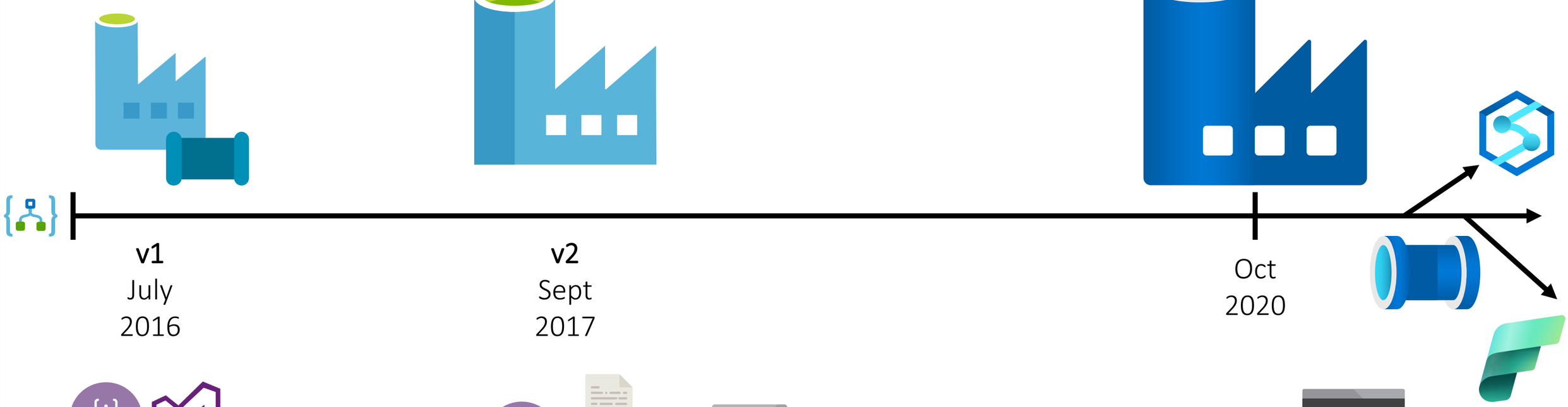
{Event/Location}-{Month}-{Year}

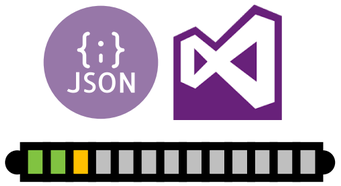
Agenda – Short Stories

- Data Integration Pipelines – A Quick History Lesson
- Scaled Out Design Patterns
- Metadata Driven Framework
- Enterprise Deployments
- Testing
- VNet Integration
- Custom Security Roles & Key Vault Integration

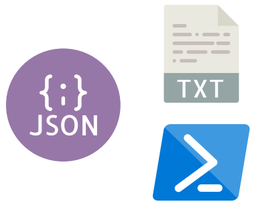
Data Integration Pipelines – A Quick History Lesson

A Quick History Lesson





Time Slices
VS2015 Projects



Notepad
Development

New UI

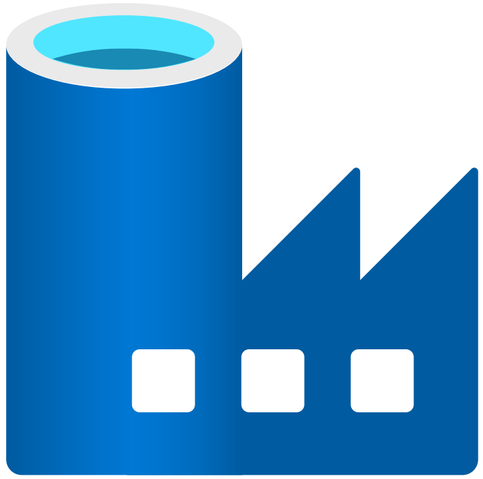
Source
Control

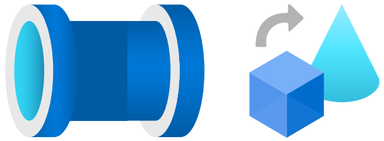
Data
Flows

Private
Endpoints

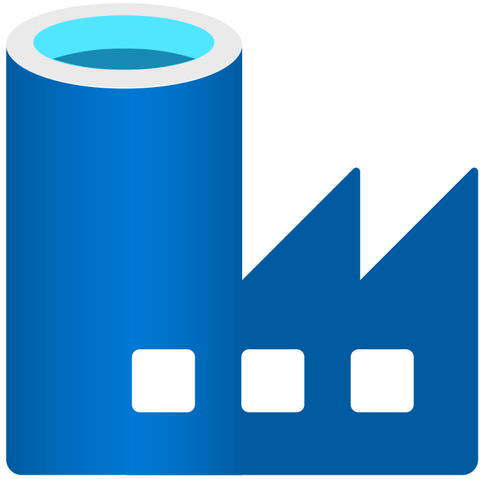
Synapse Analytics vs Data Factory

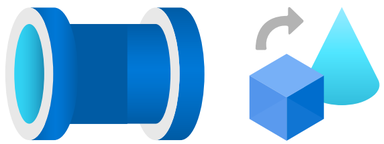
https://docs.microsoft.com/en-us/azure/synapse-analytics/data-integration/concepts-data-factory-differences





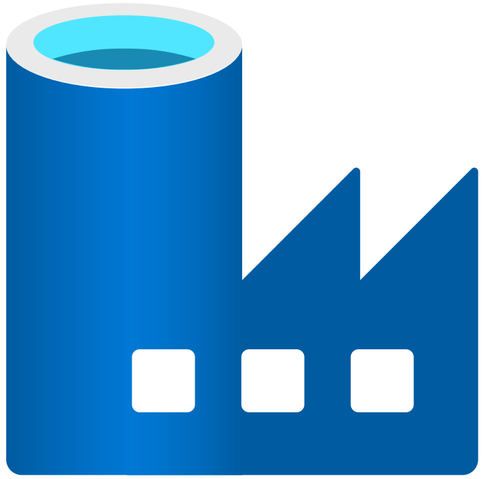
Synapse Analytics vs Data Factory

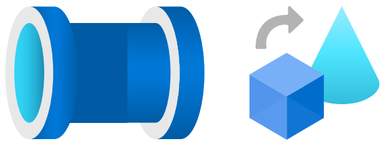




Microsoft Fabric vs Data Factory

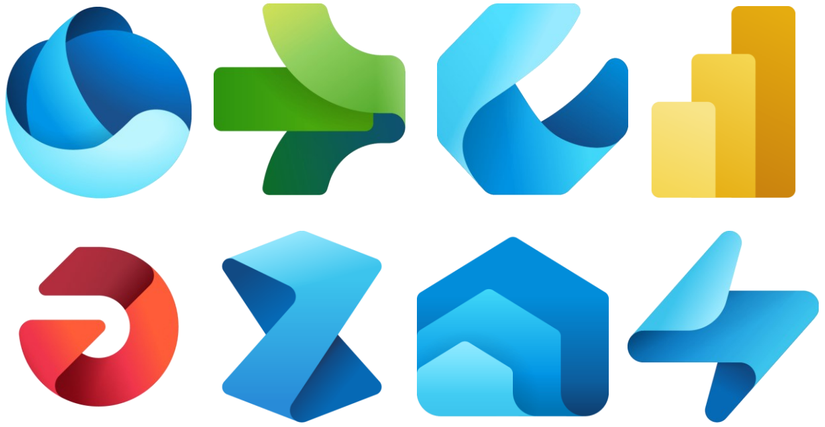
https://mrpaulandrew.com/2023/05/31/what-is-microsoft-fabric-my-point-of-view/

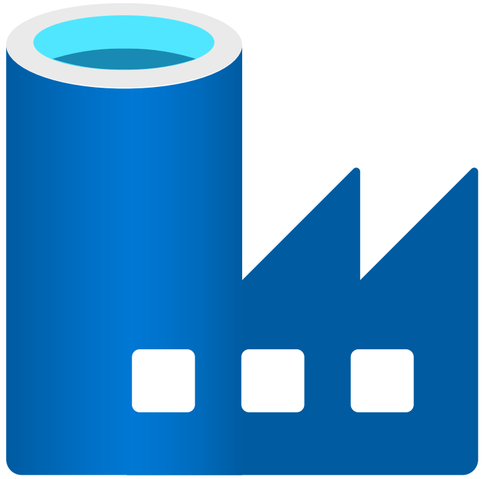


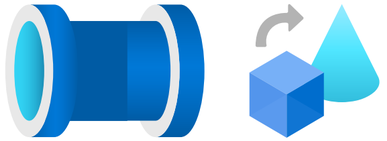


Microsoft Fabric vs Data Factory

https://mrpaulandrew.com/2023/05/31/what-is-microsoft-fabric-my-point-of-view/

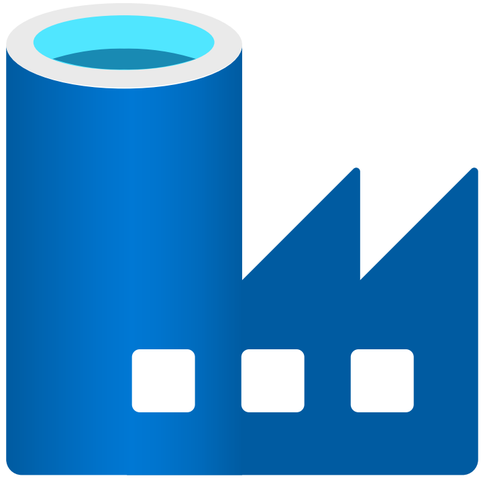


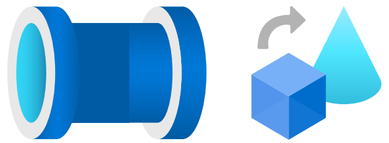




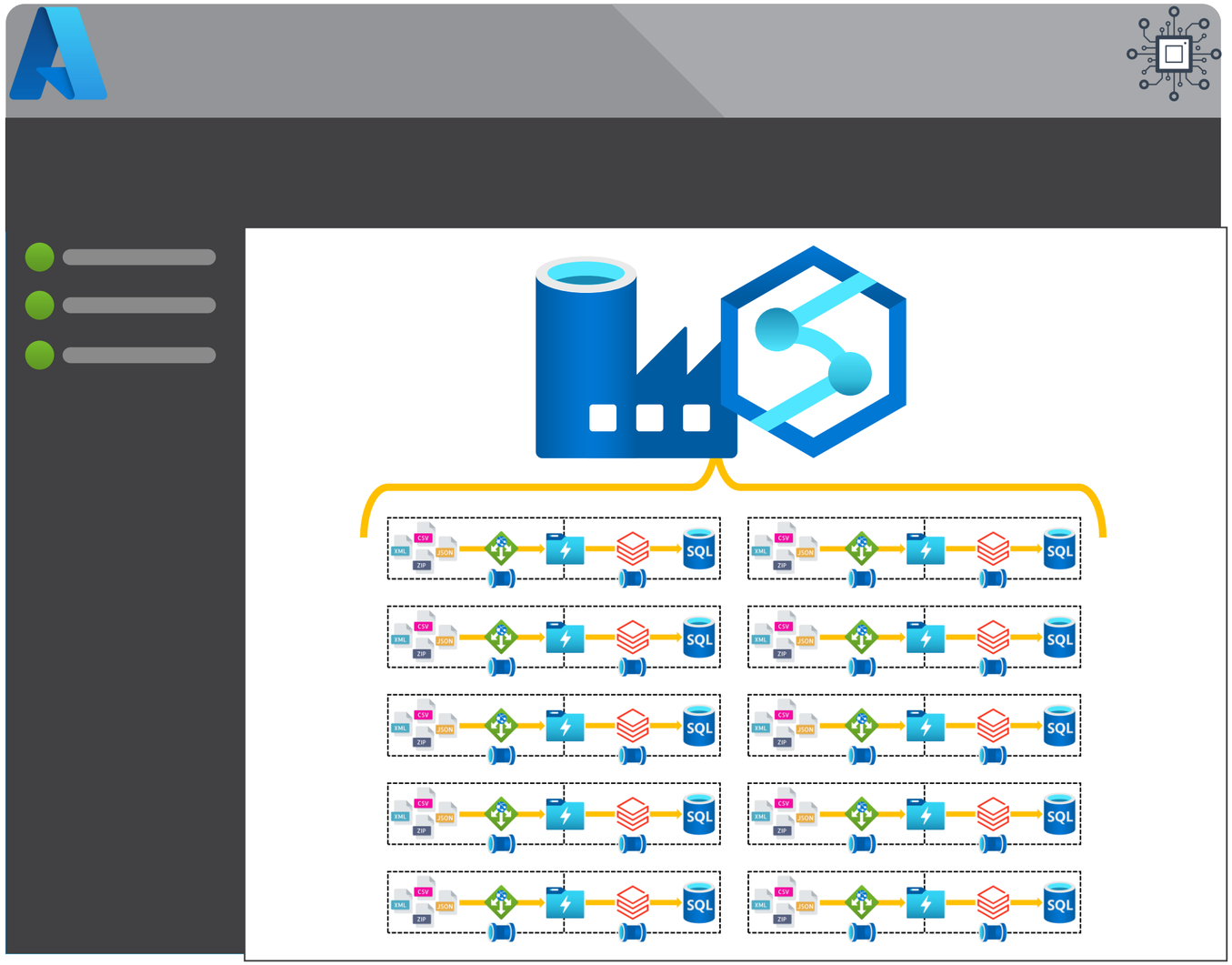
Microsoft Fabric vs Data Factory

https://mrpaulandrew.com/2023/05/31/what-is-microsoft-fabric-my-point-of-view/



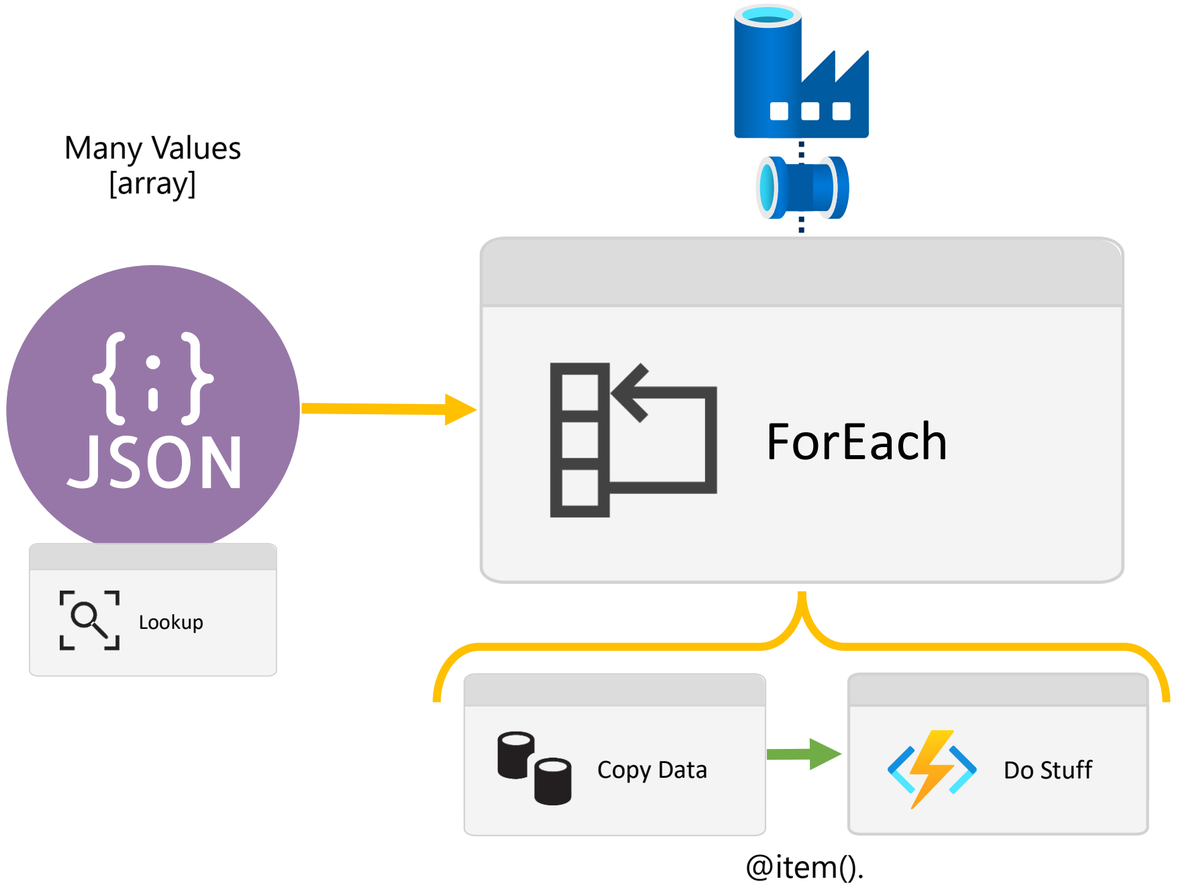


Scaled Out Design Patterns

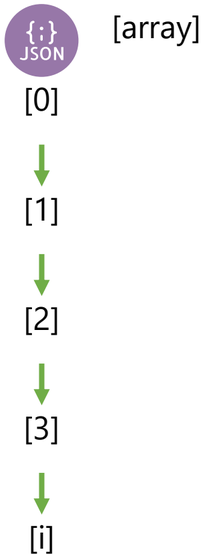


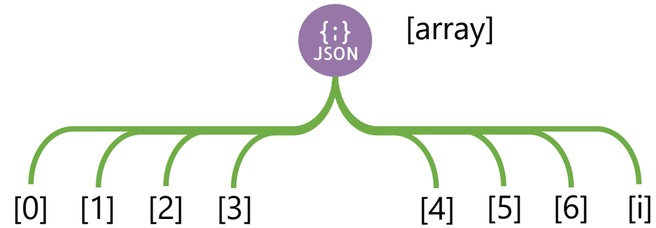
For Each Activity

Scaling Out Control Flow Activities



IsSequential: true



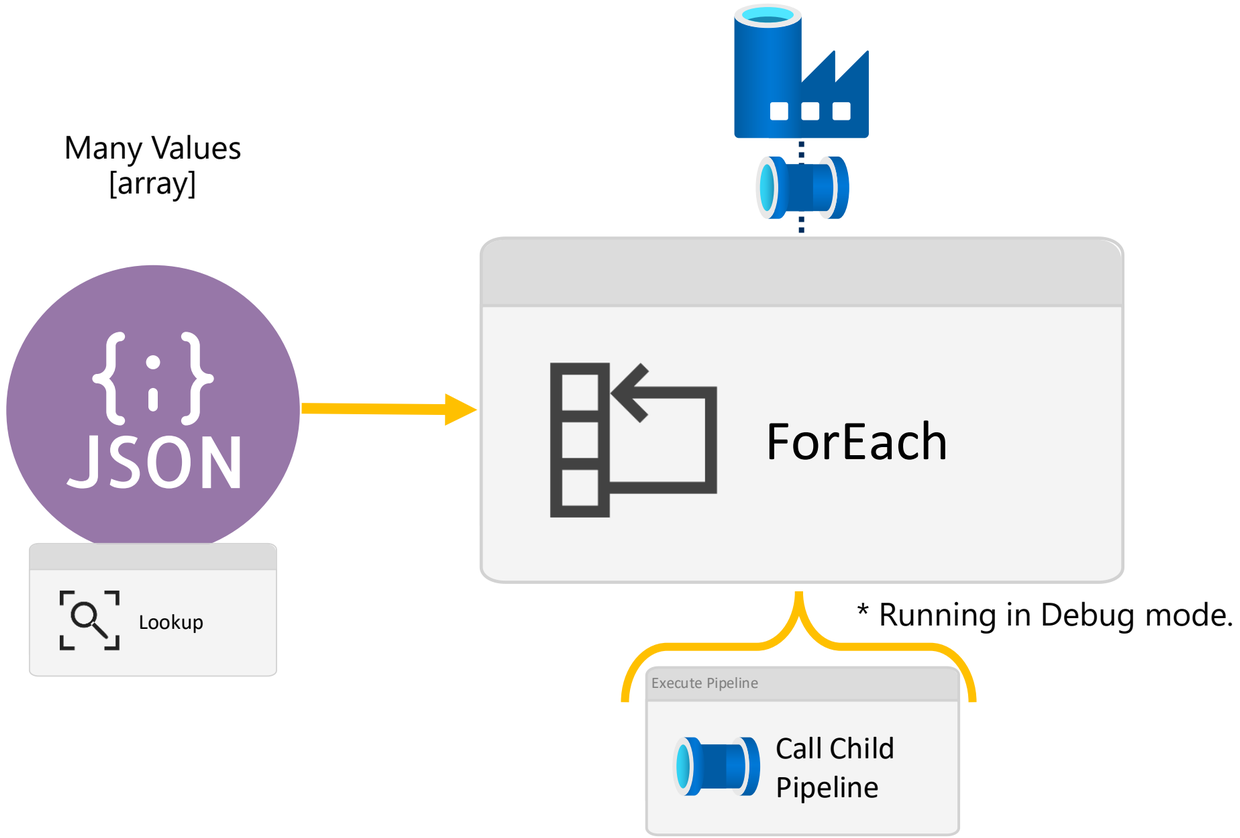


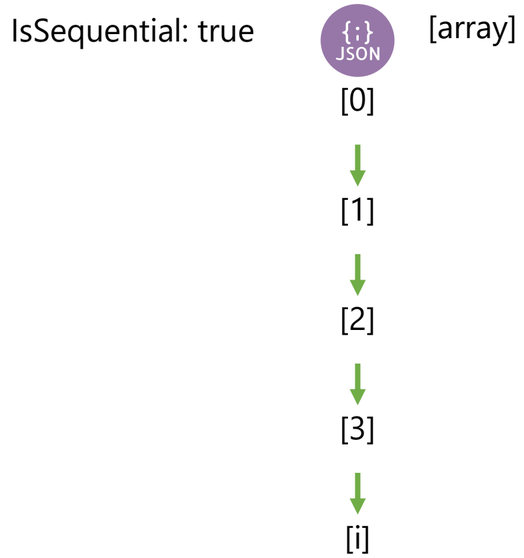
Batch Count Default: 20

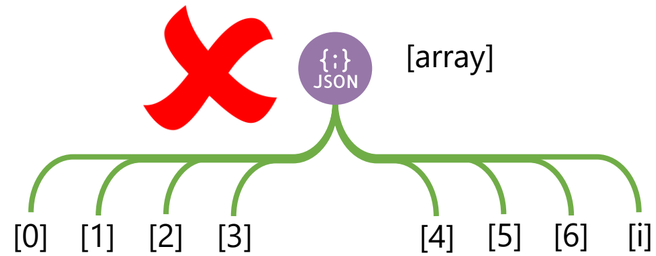
Batch Count Max: 50

For Each Activity

Scaling Out Control Flow Activities



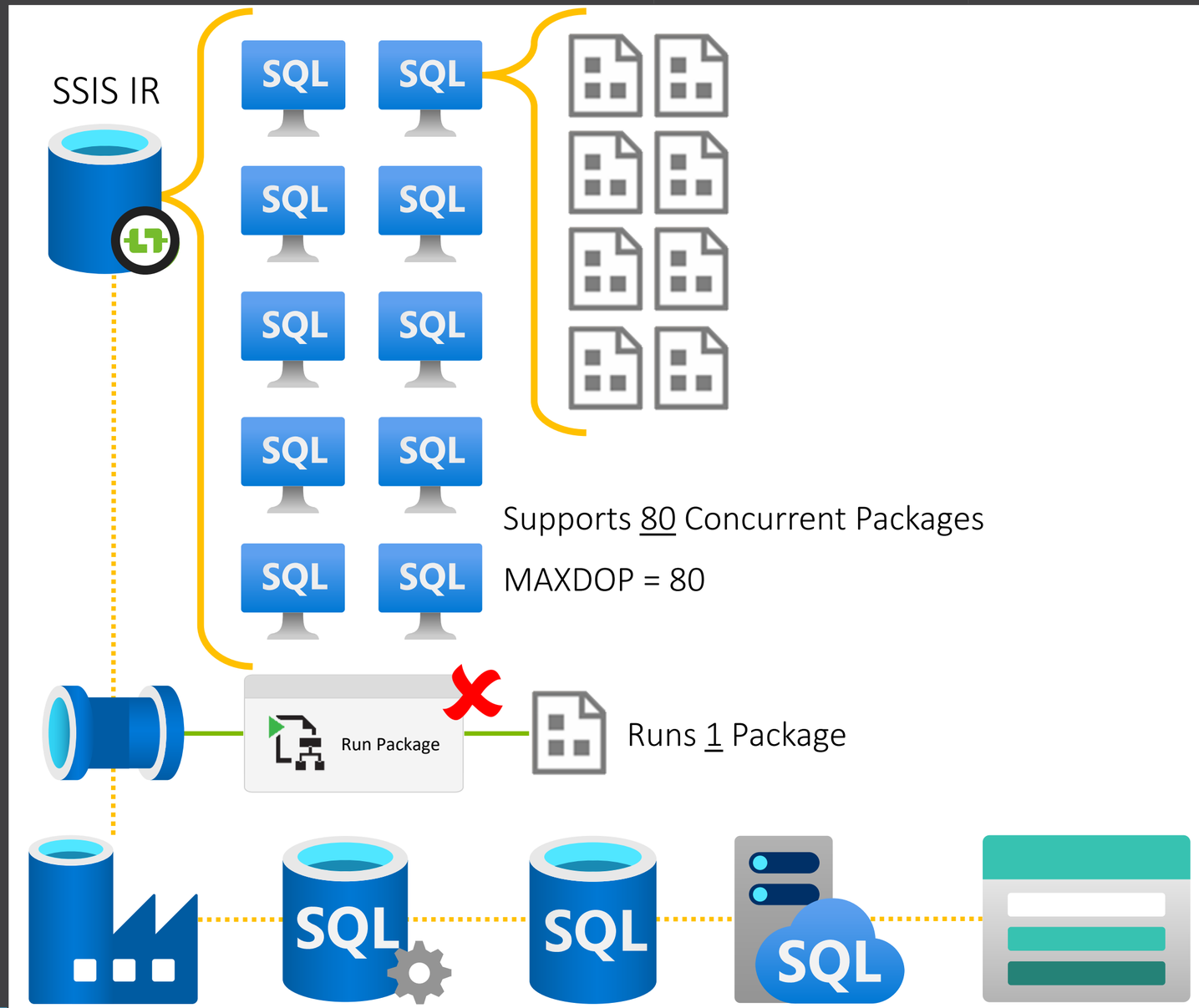




Batch Count Default: 20

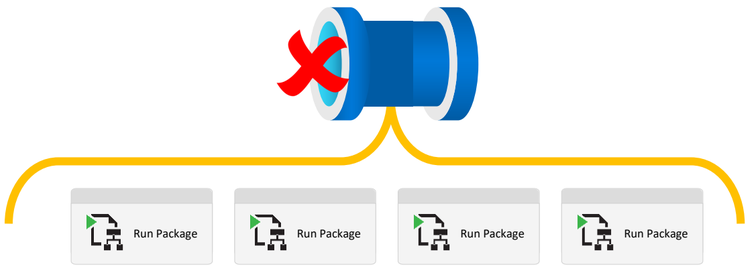
Batch Count Max: 50

Problem: Using All Of The SSIS IR Compute



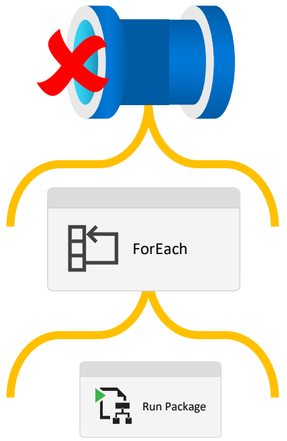
Parent Package

Child Packages x80



Pipeline x1

Activities x80

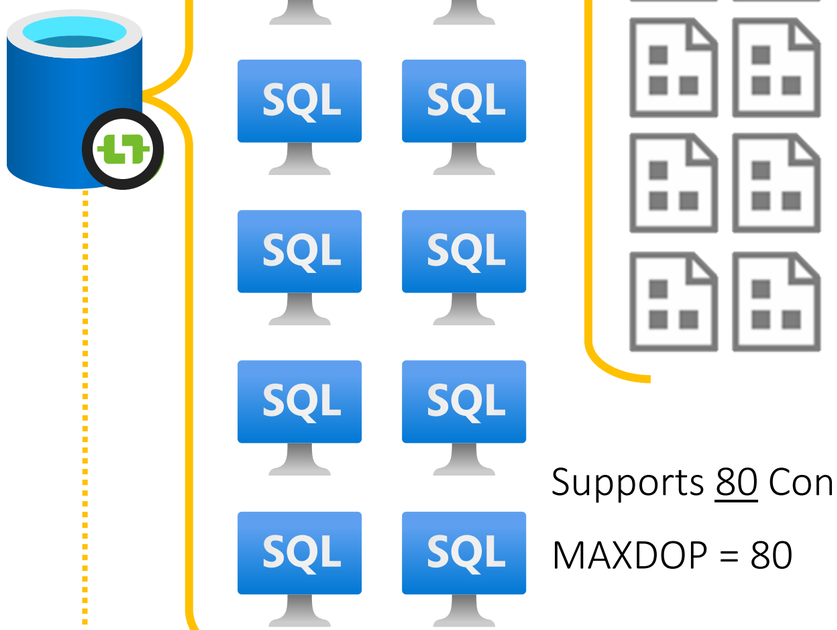


Pipeline x1

ForEach
Max Batch
(50)

Solution: Static Pipelines

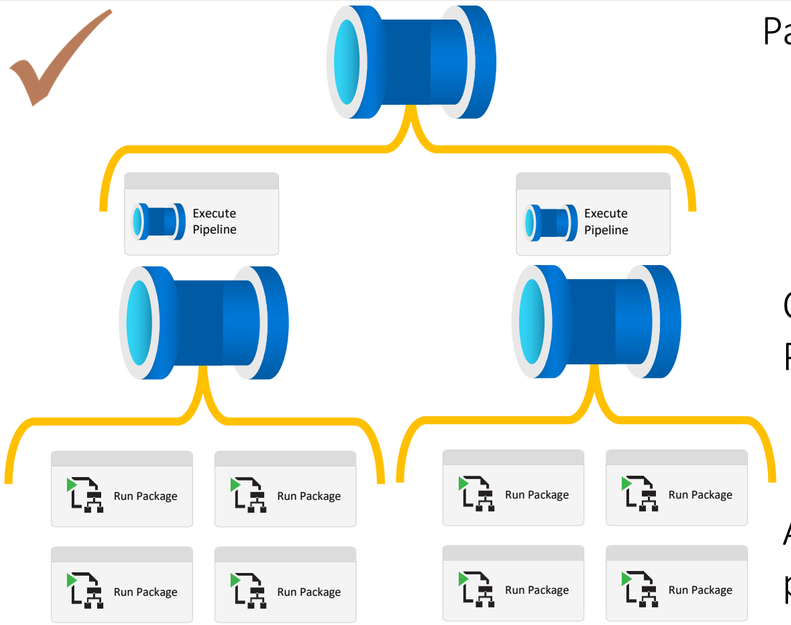
SSIS IR



Supports 80 Concurrent Packages

MAXDOP = 80

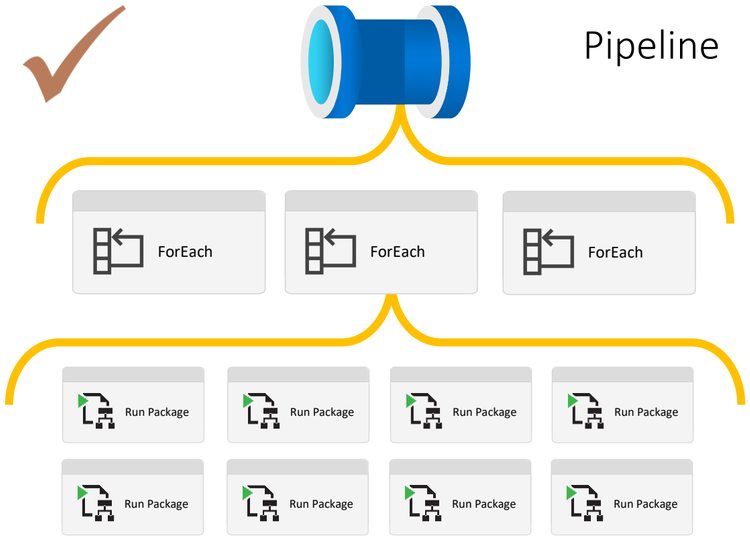
Parent Pipeline



Child Pipelines x2

Activities x40 per Child

Pipeline



8x ForEach Batch (10)

Solution: Nested ForEach Activities & Bucket Metadata

SSIS IR

```
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

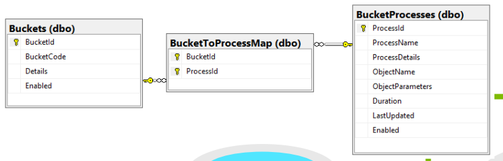
FE L2
MAXDOP 50

@item().BucketId

Array[BucketContents]

Run Package

```
;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
```



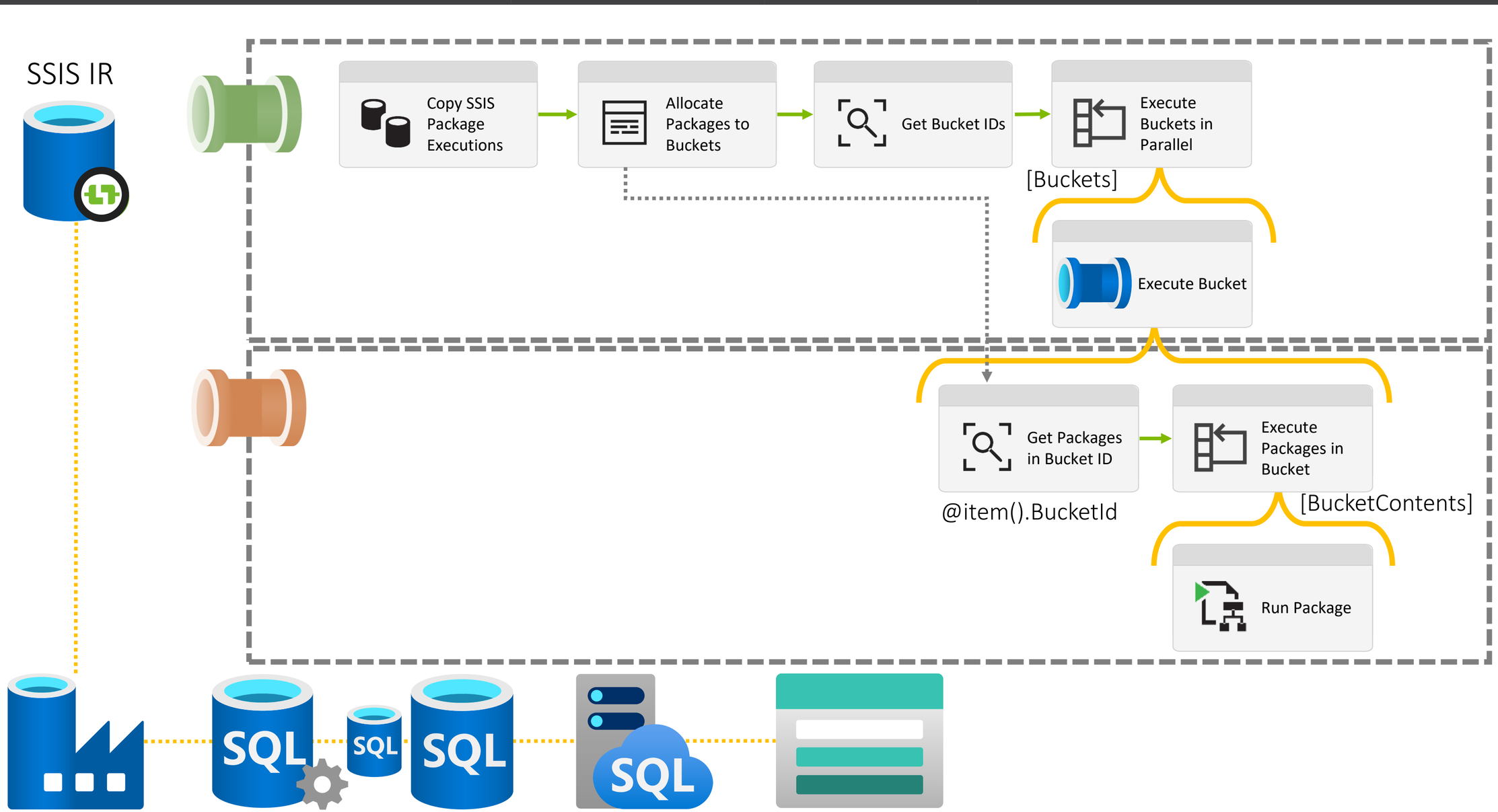
SSISDB

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

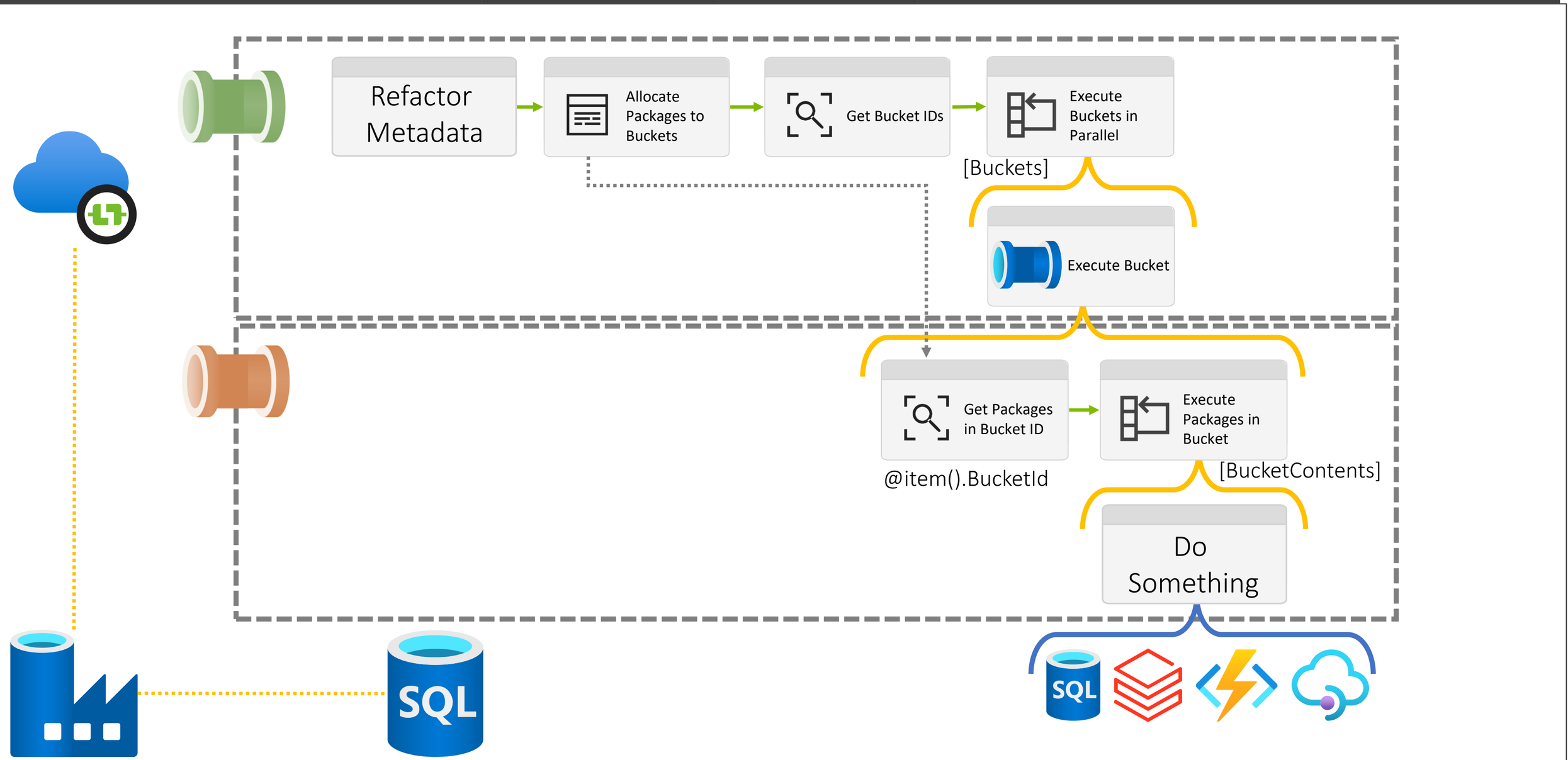
(FE L1) x (FE L2) = NEW MAXDOP

50 x 50 = 2500

Solution: Nested ForEach Activities & Bucket Metadata

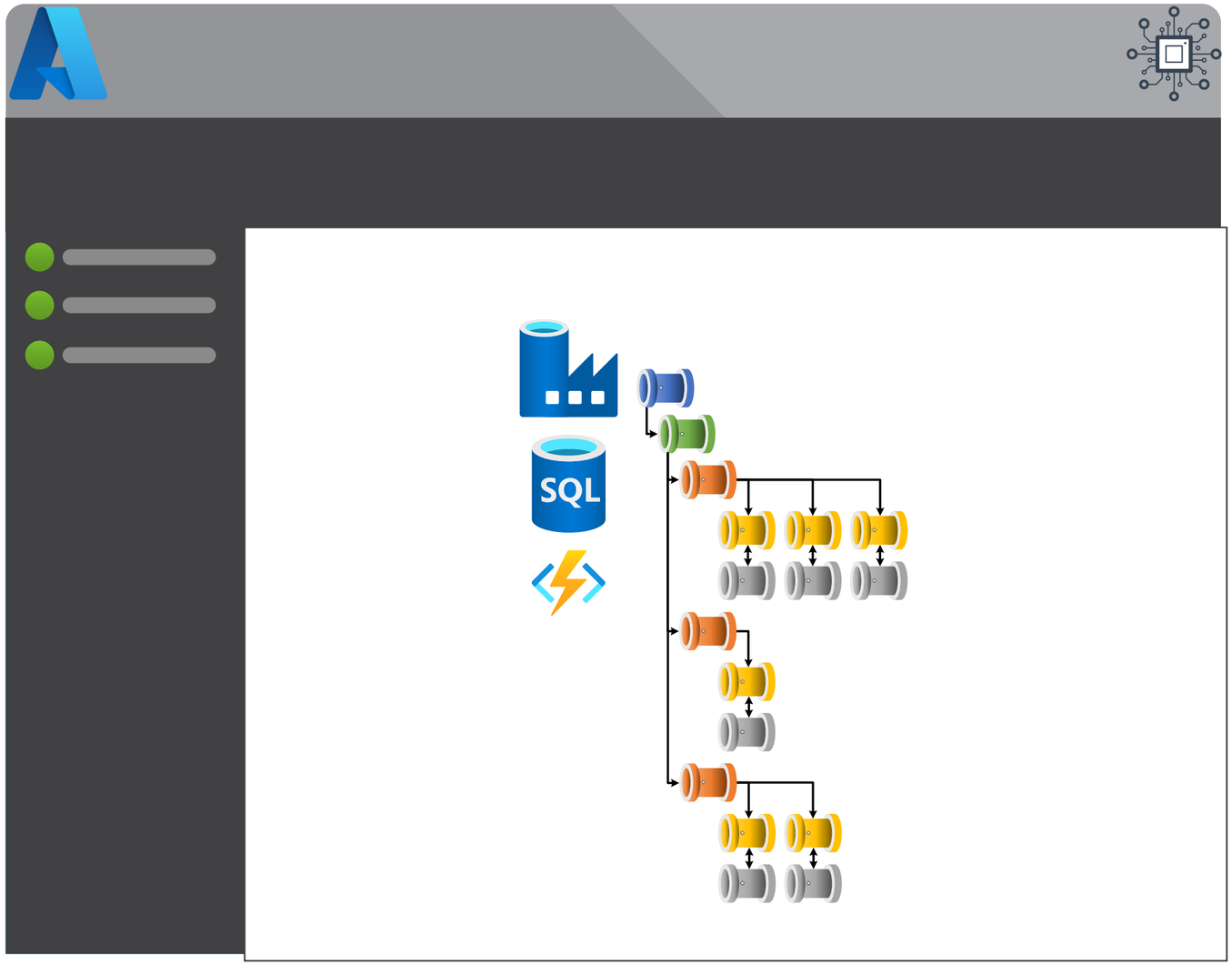


Solution: Nested ForEach Activities & Bucket Metadata



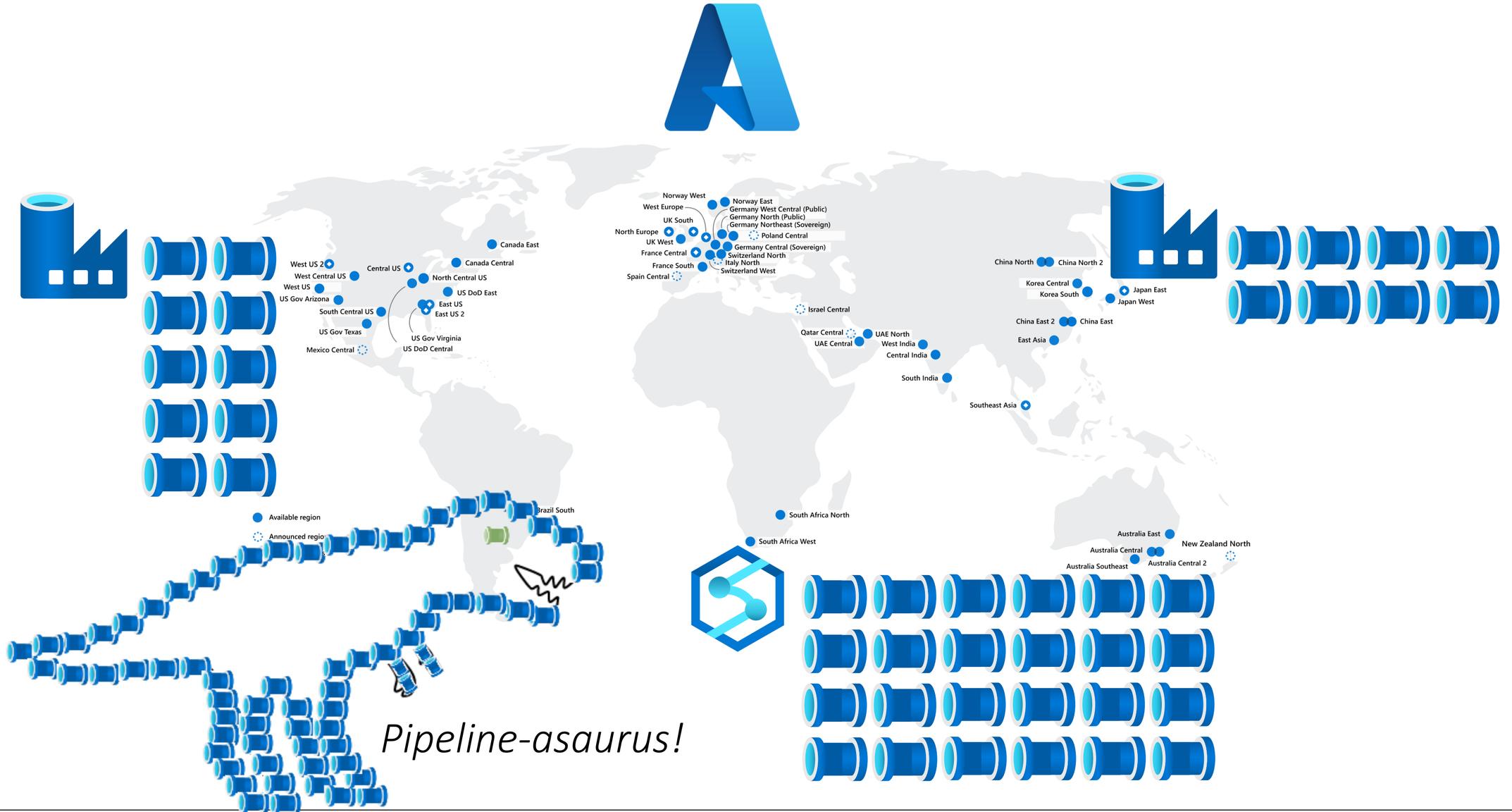
Demo

Metadata Driven Orchestration

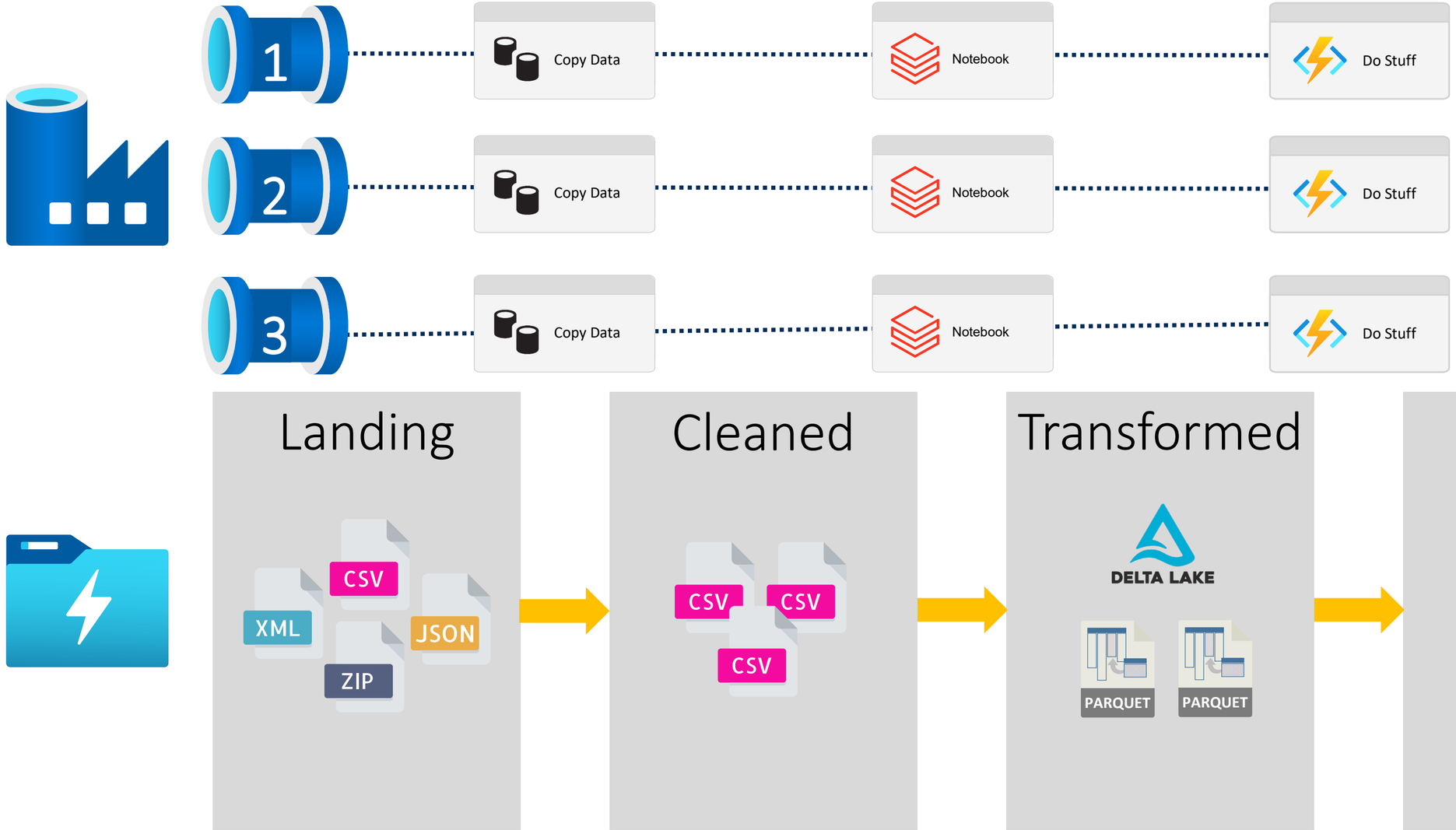


Problem

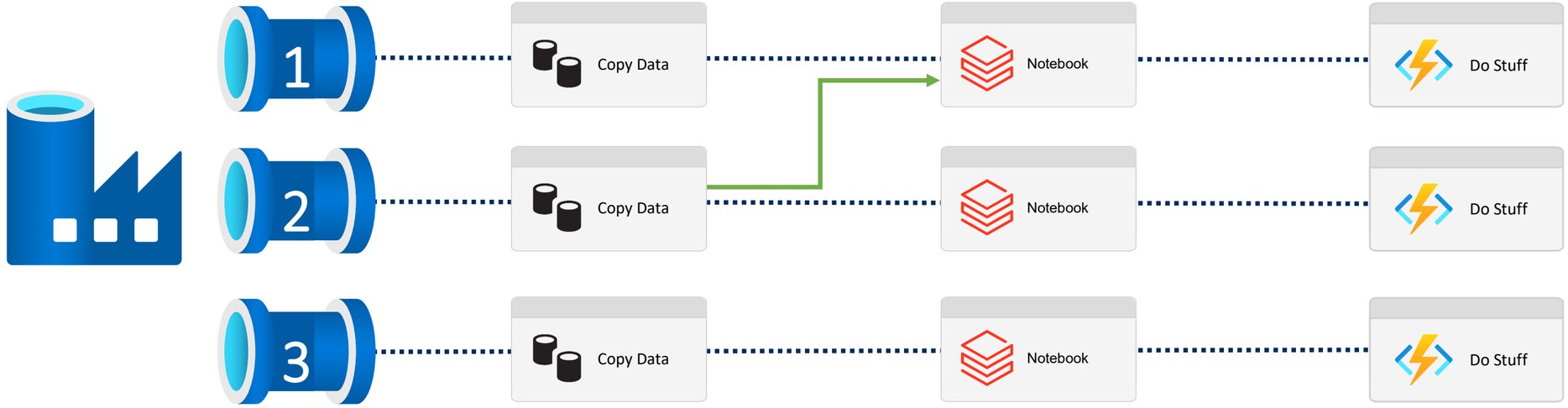
How should we structure and trigger our Integration Pipelines?

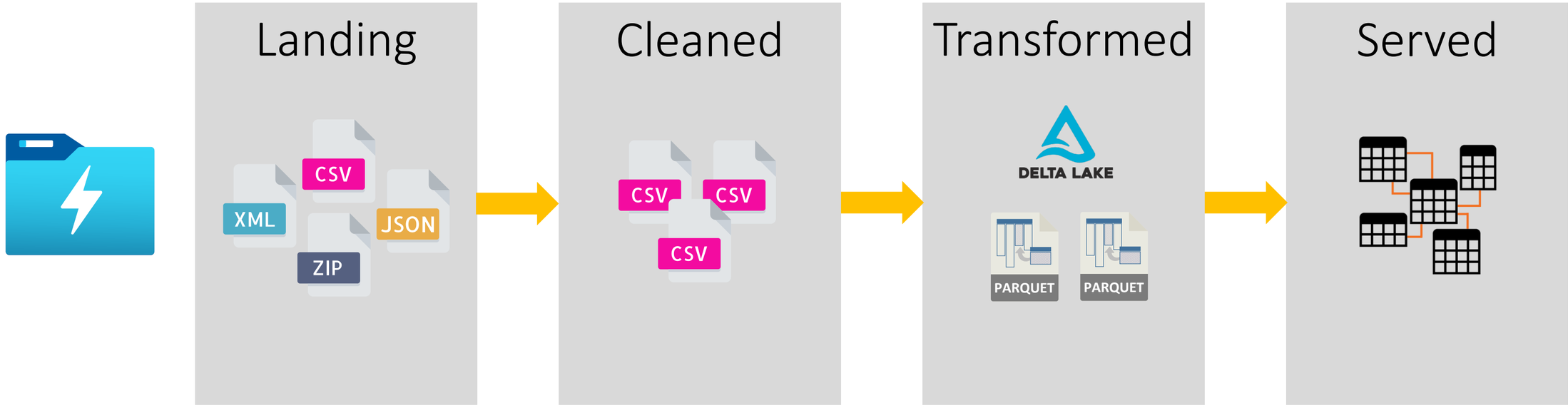


Problem

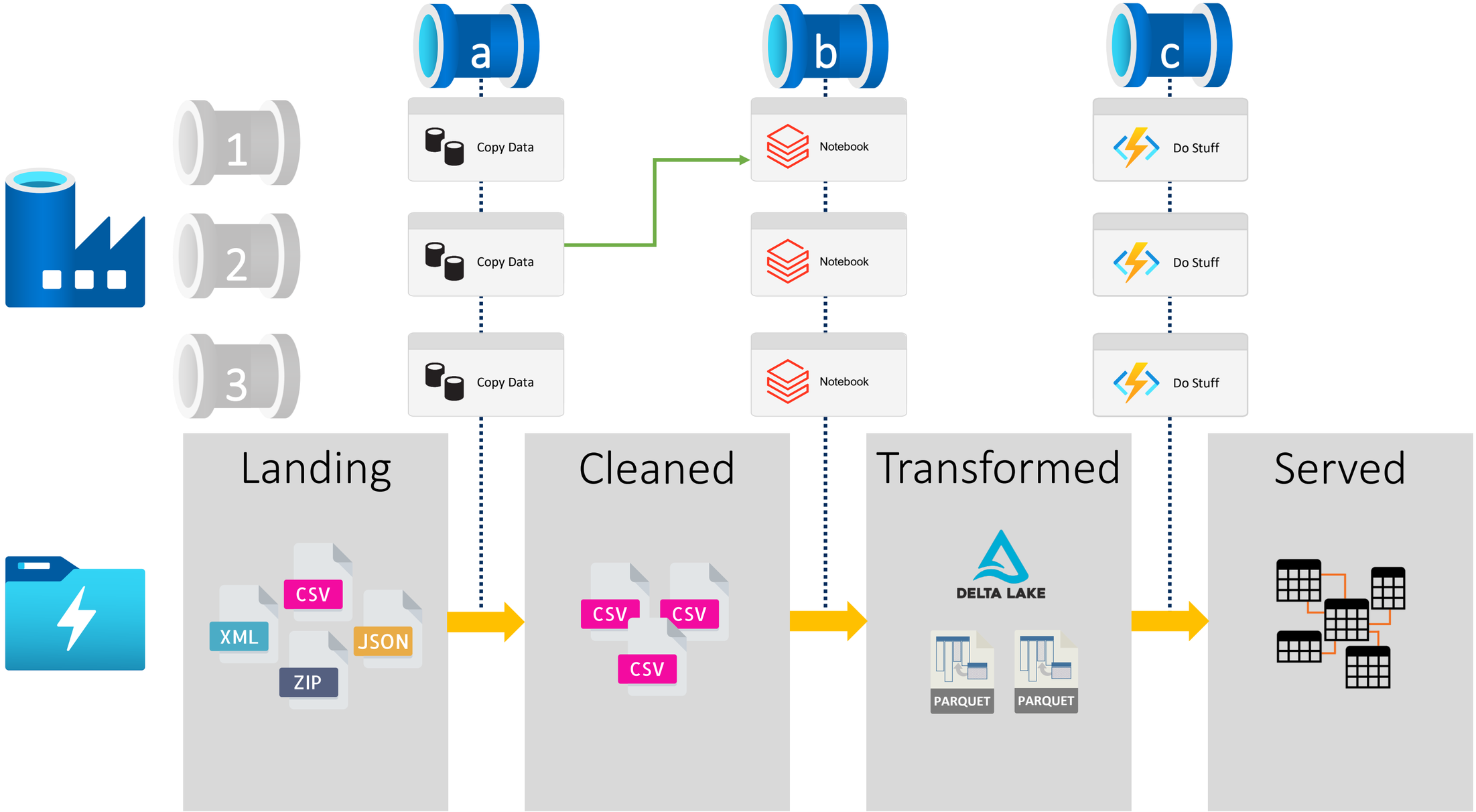

1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff
Landing
Cleaned
Transformed
Served
CSV
XML
JSON
ZIP
CSV
CSV
CSV
DELTA LAKE
PARQUET
PARQUET

Problem

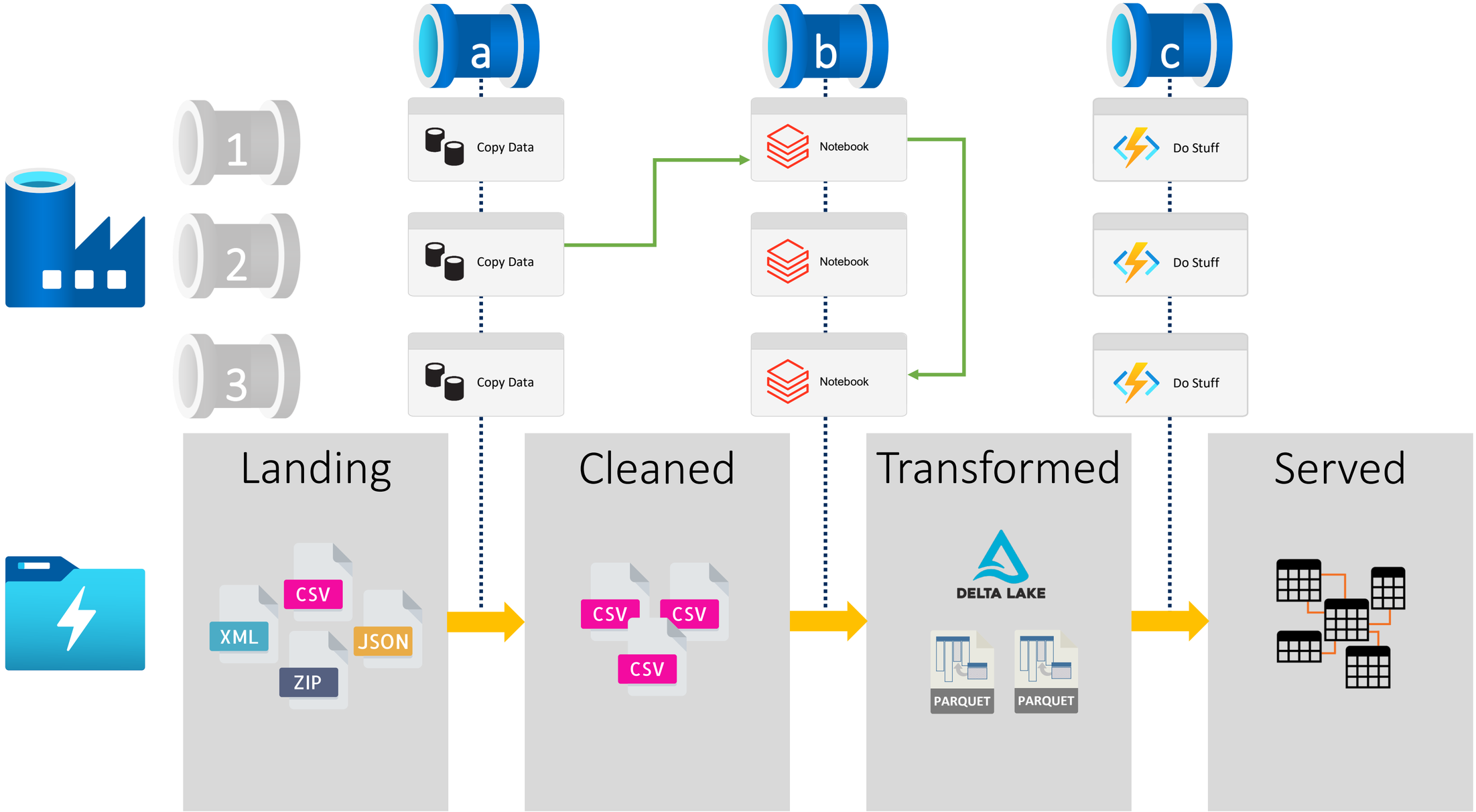

1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff


Landing
CSV
XML
JSON
ZIP
Cleaned
CSV
CSV
CSV
Transformed
DELTA LAKE
PARQUET
PARQUET
Served

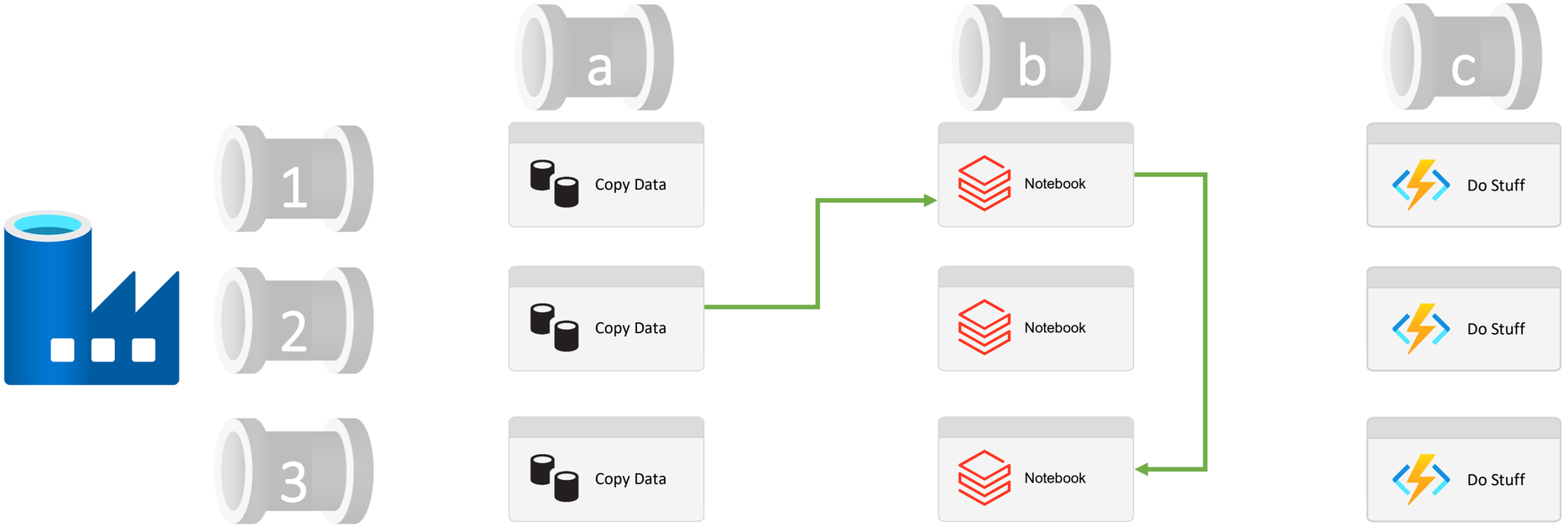
Problem

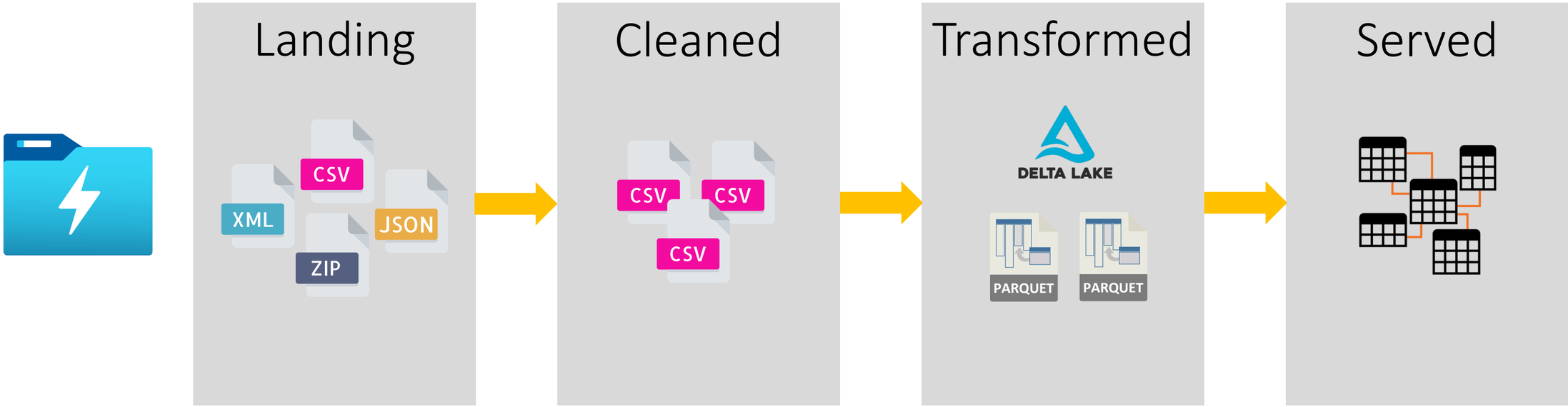

a
b
c
1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff
Landing
Cleaned
Transformed
Served
CSV
XML
JSON
ZIP
CSV
CSV
CSV
DELTA LAKE
PARQUET
PARQUET

Problem


a
b
c
1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff
Landing
Cleaned
Transformed
Served
CSV
XML
JSON
ZIP
CSV
CSV
CSV
DELTA LAKE
PARQUET
PARQUET

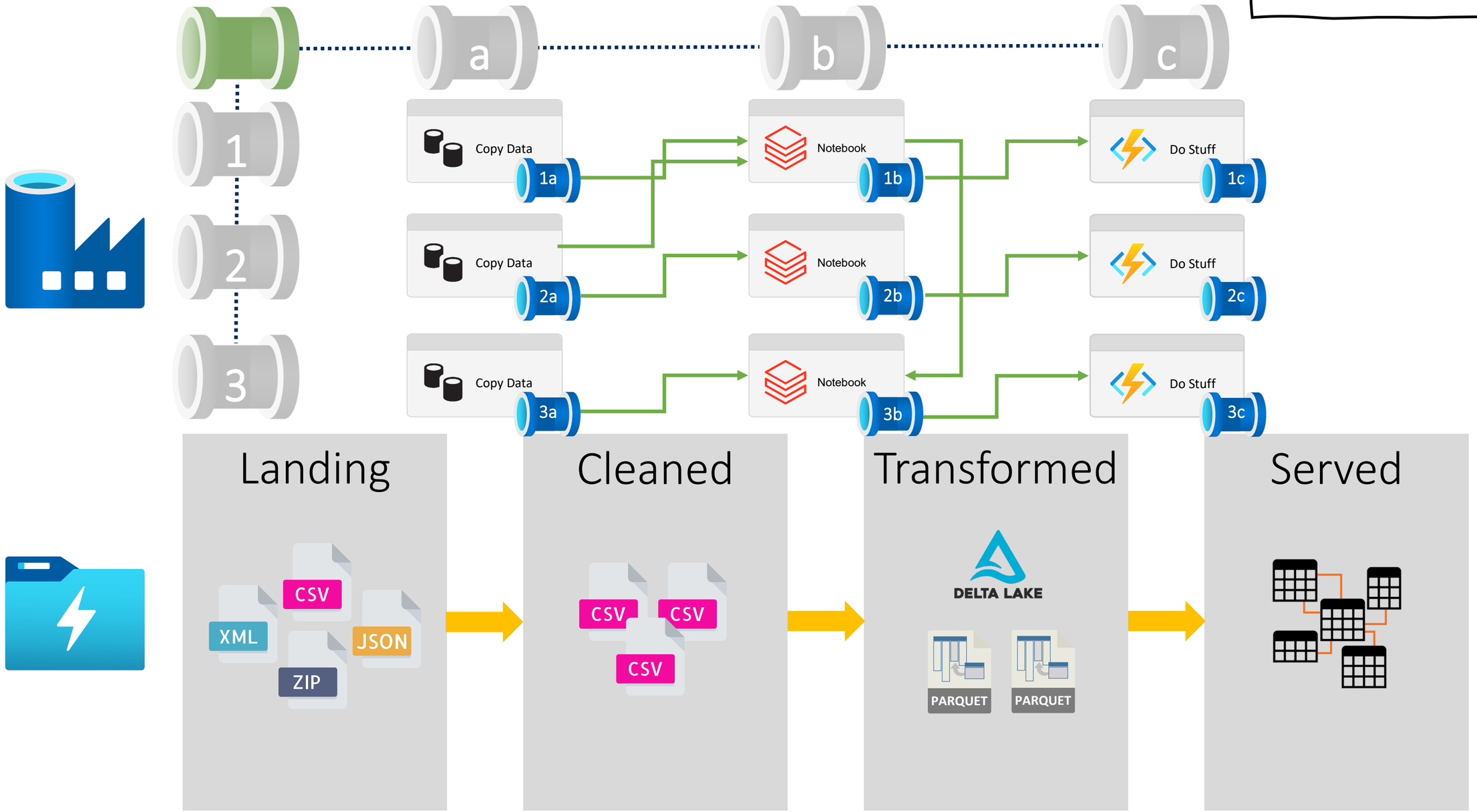
Problem


a
b
c
1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff

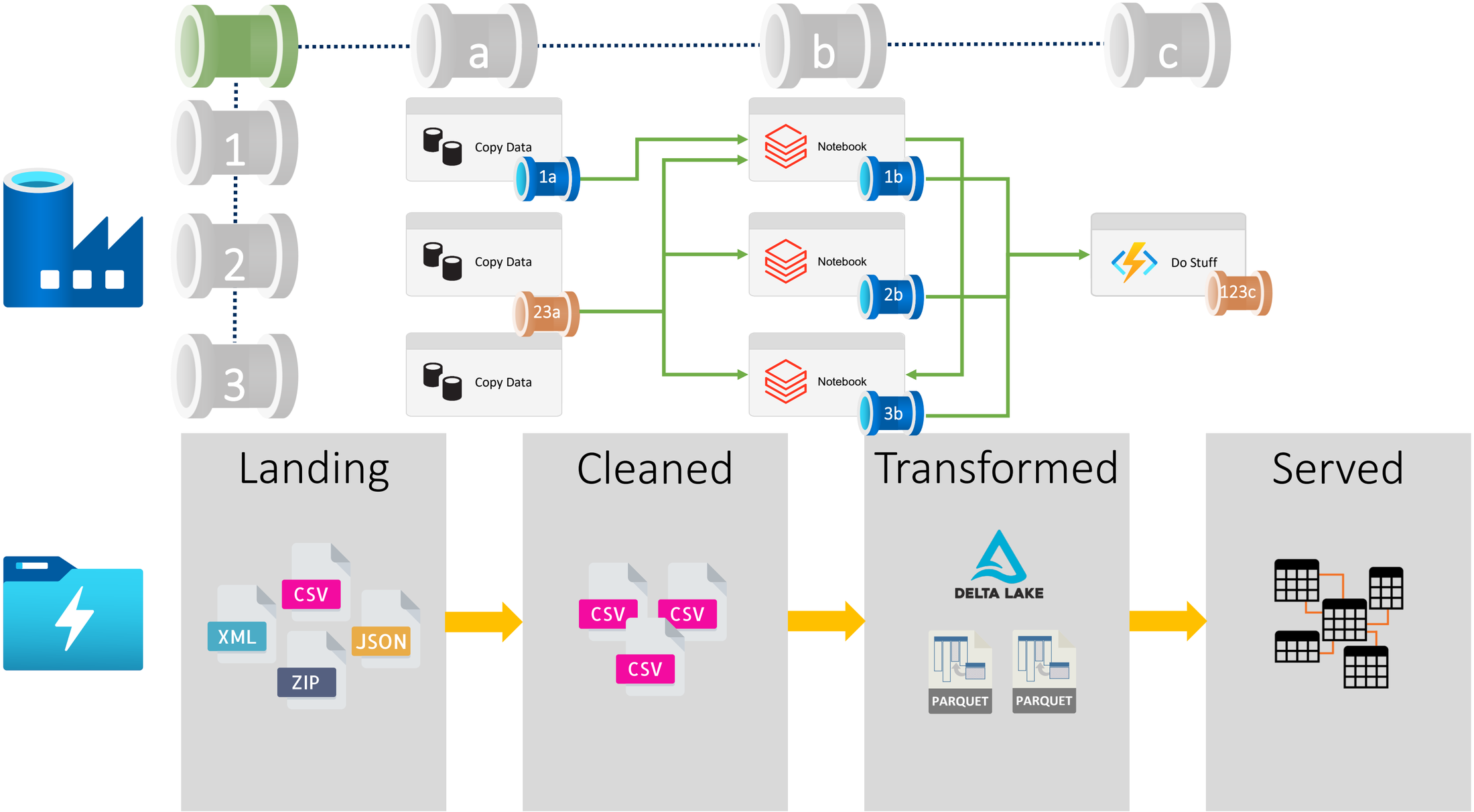

Landing
CSV
XML
JSON
ZIP
Cleaned
CSV
CSV
CSV
Transformed
DELTA LAKE
PARQUET
PARQUET
Served

Problem

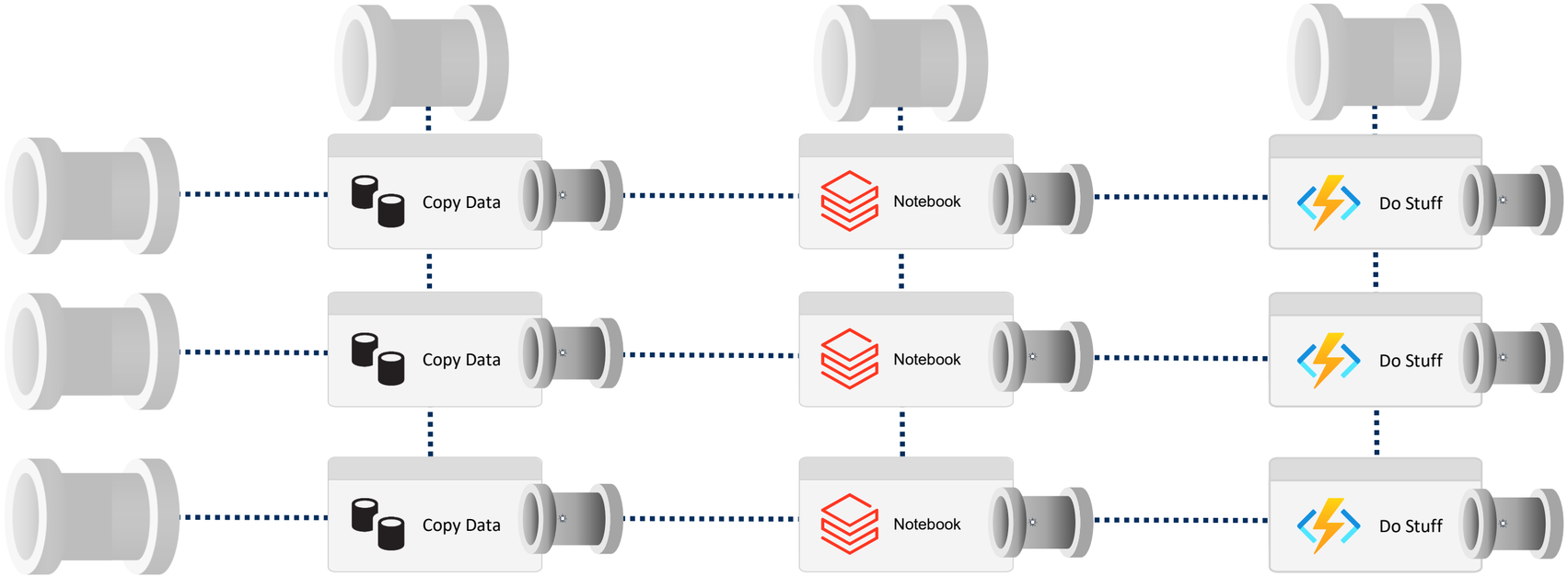
Only 40 Activities per Pipeline.



Problem


a
b
c
1
2
3
Copy Data
1a
Copy Data
23a
Copy Data
Notebook
1b
Notebook
2b
Notebook
3b
Do Stuff
123c
Landing
CSV
XML
JSON
ZIP
Cleaned
CSV
CSV
CSV
Transformed
DELTA LAKE
PARQUET
PARQUET
Served

Solution



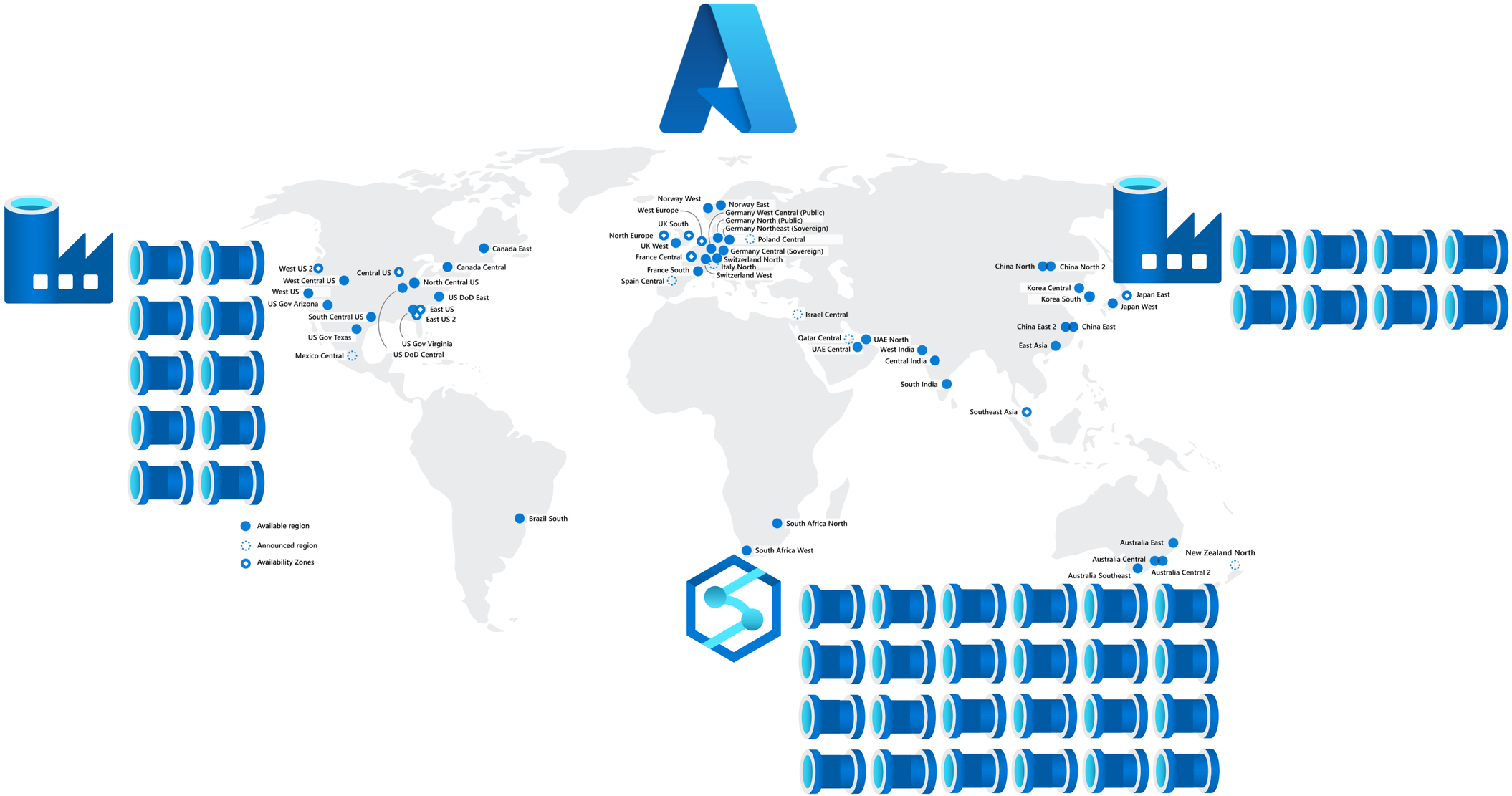
Stages
1
2
3

Pipelines
a
b
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

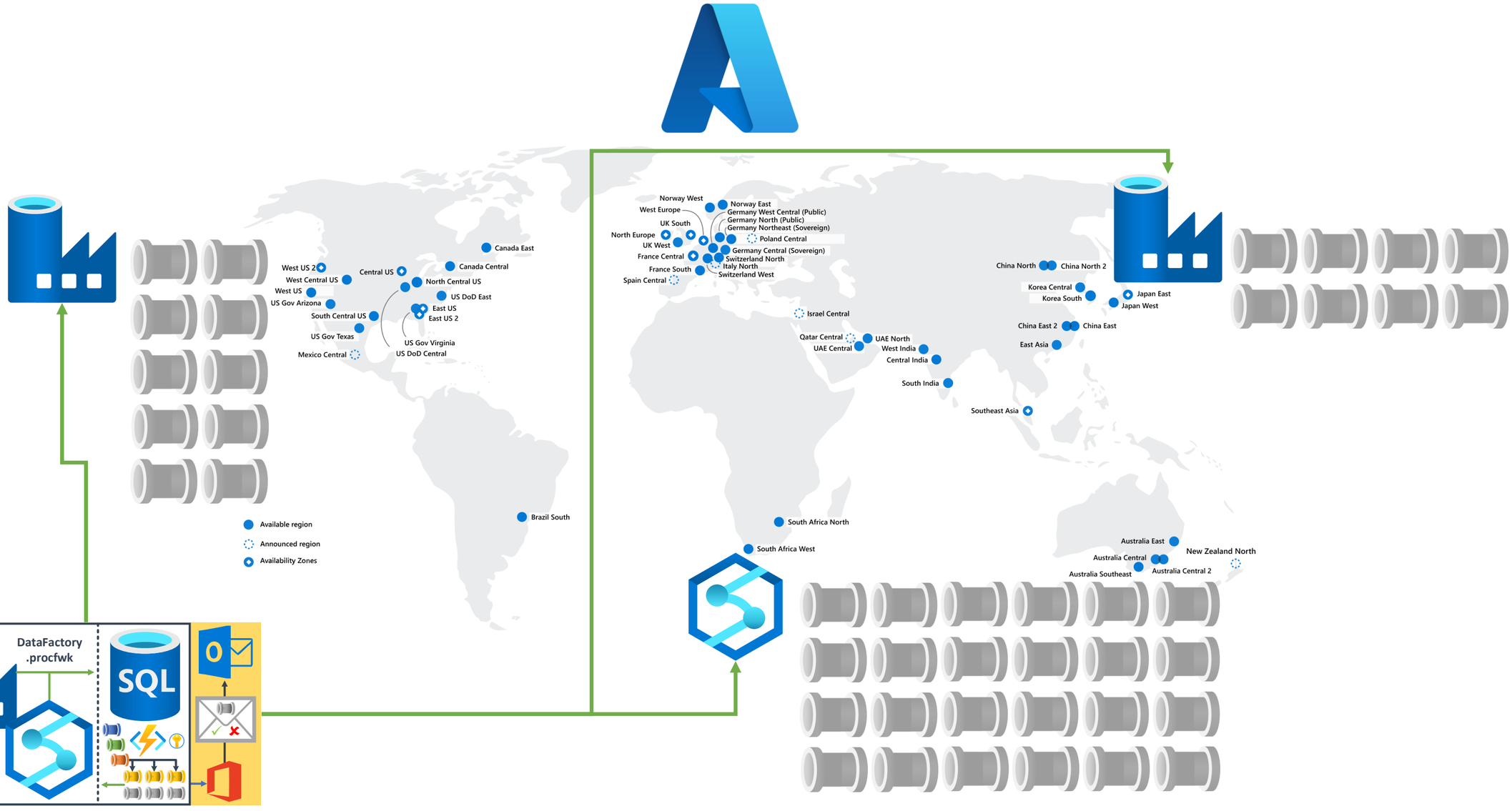
Problem

How should we structure and trigger our Integration Pipelines?



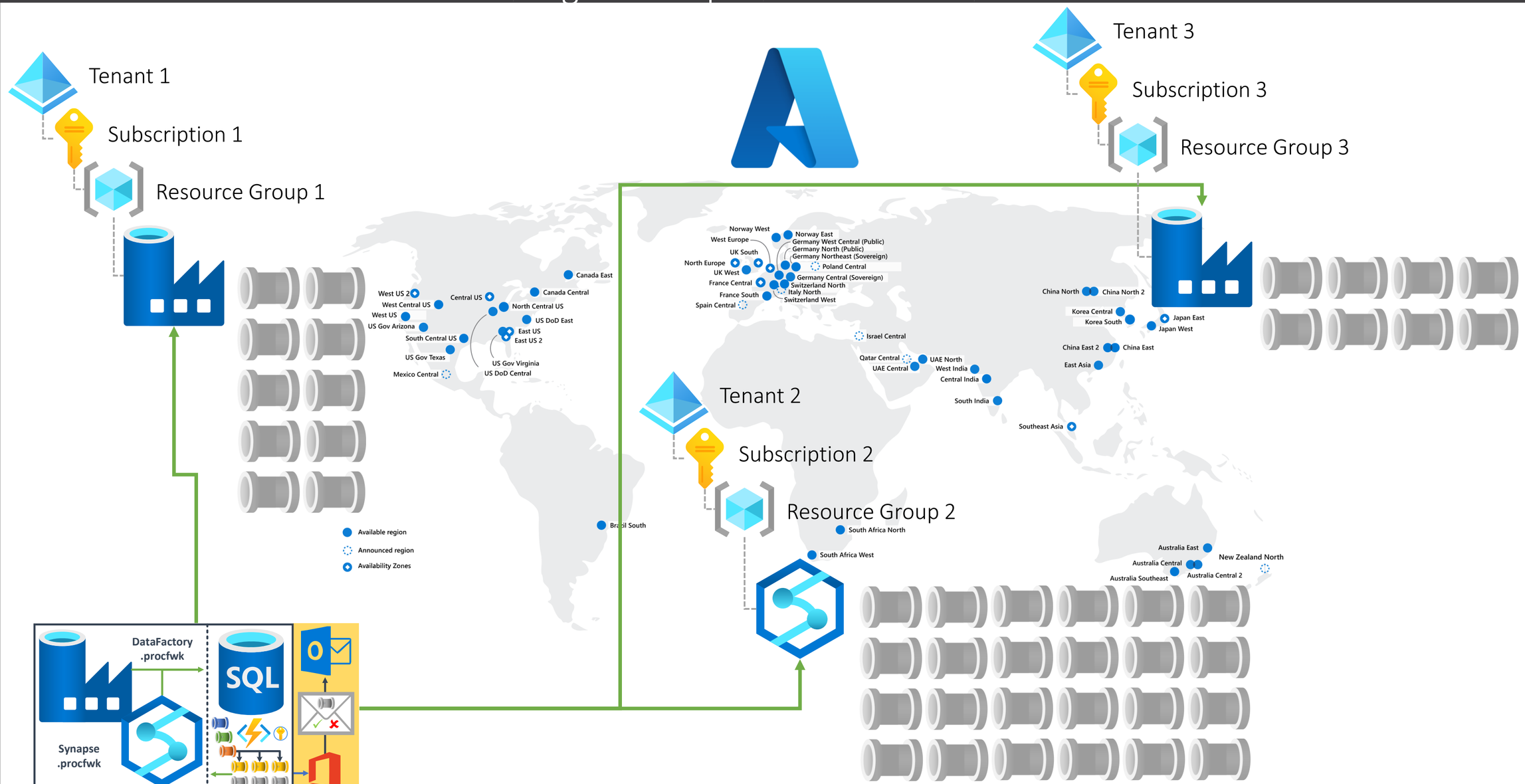
Solution

Use Metadata to Drive Integration Pipeline execution



Solution

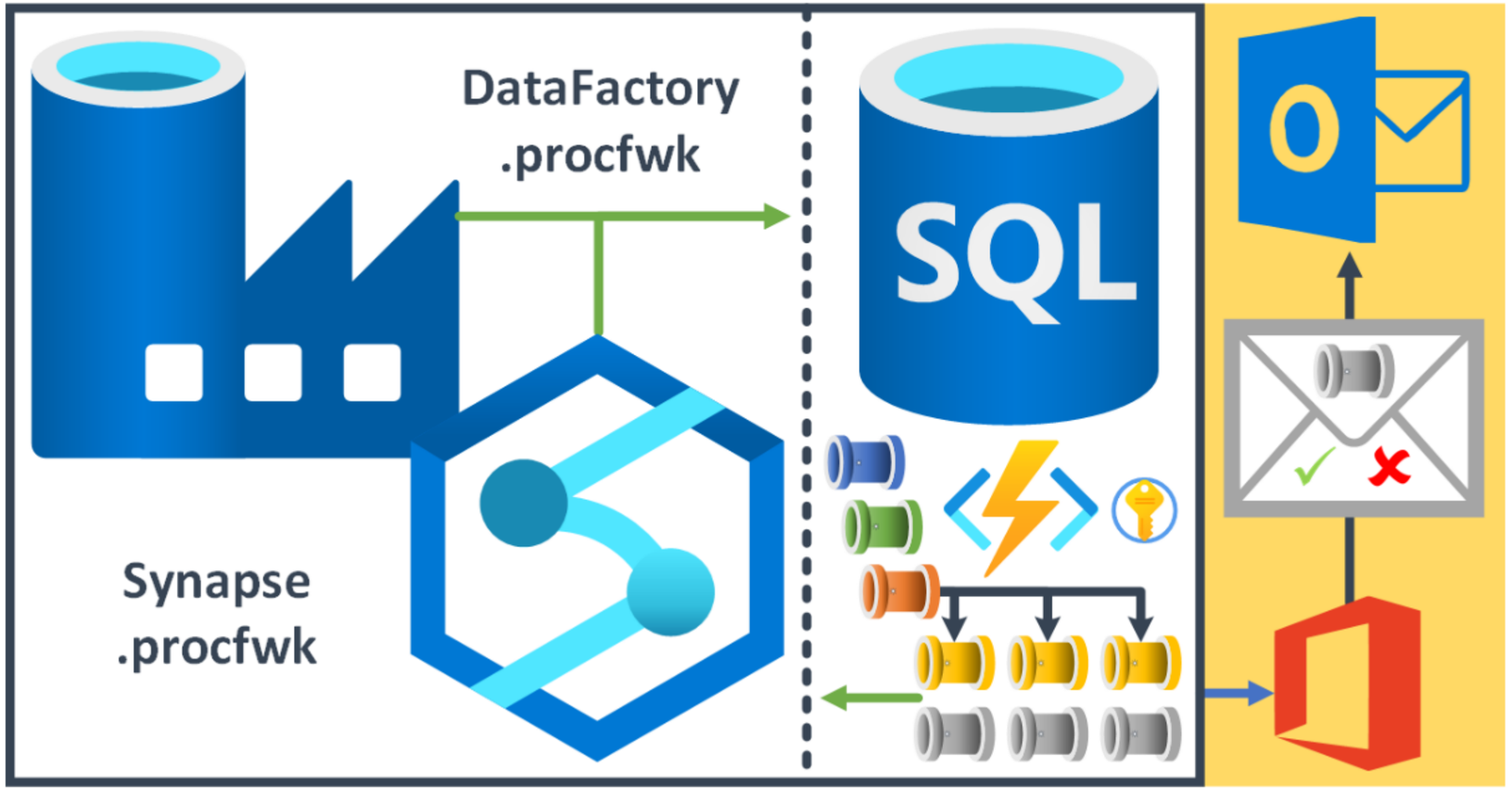
Use Metadata to Drive Integration Pipeline execution



Introducing ProcFwk.com

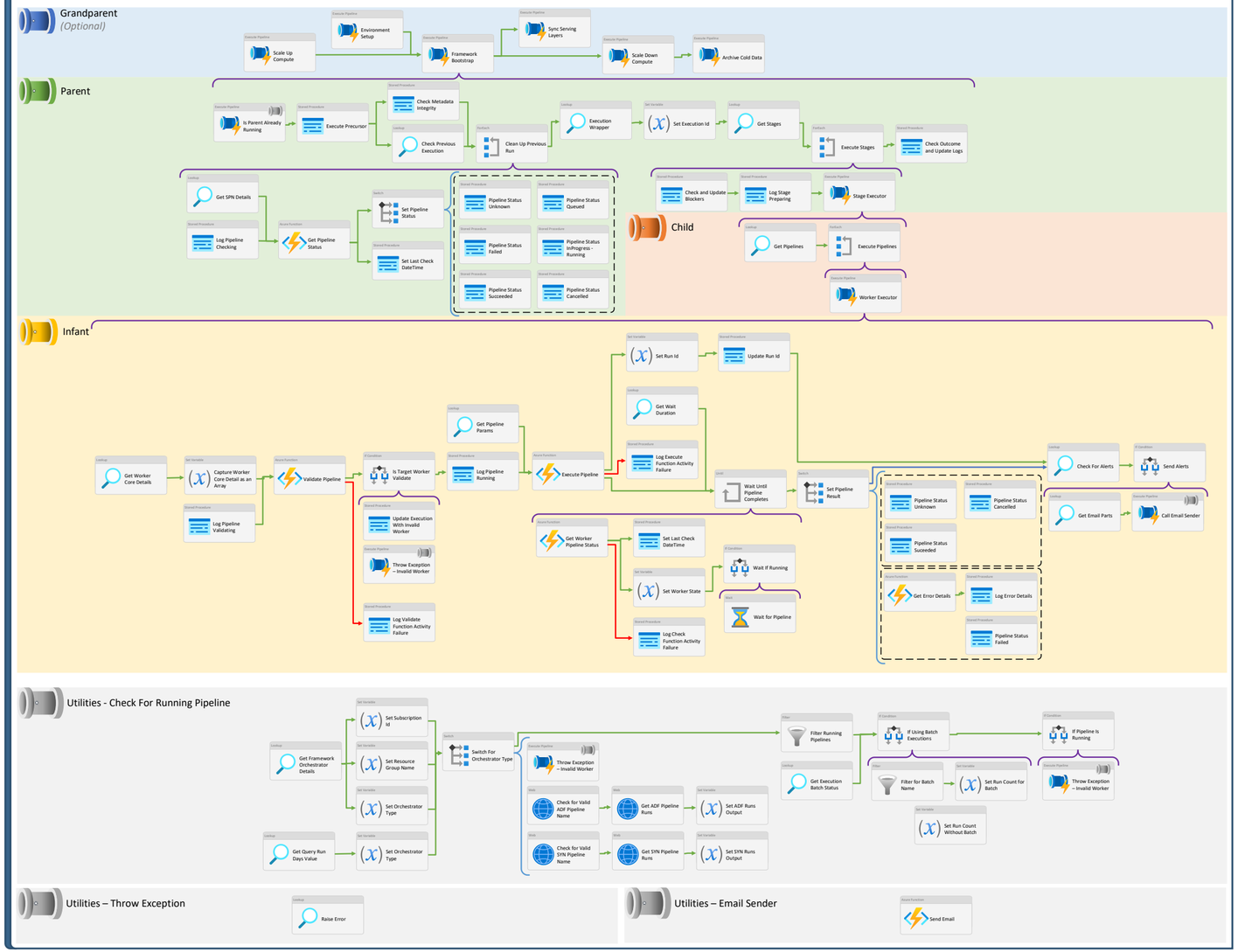
stackoverflow

procfwk



ProcFwk.com Activity Chain

Orchestration Framework

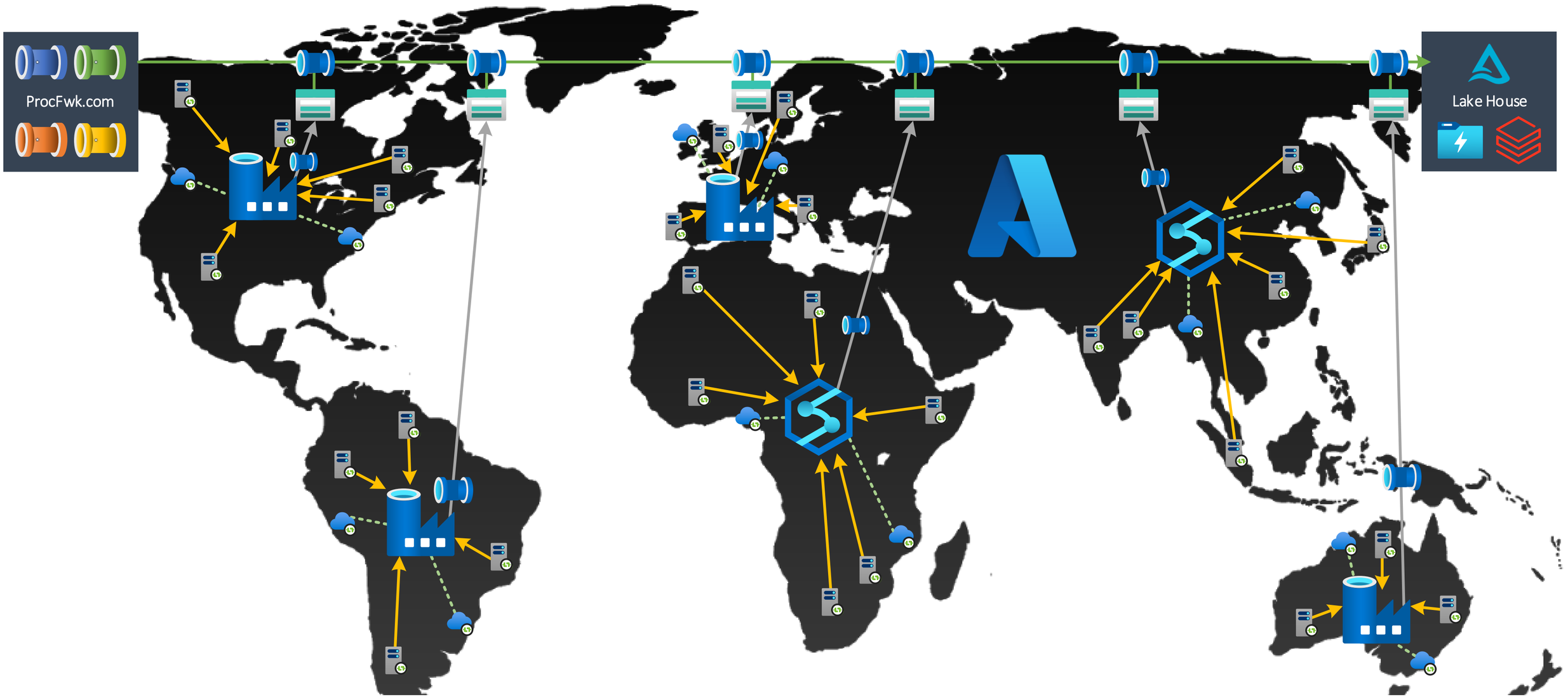


Worker Pipelines

- Worker 1 - Extract
- Worker 2 - Clean
- Worker 3 - Transform
- Worker 4 - Load
- Worker 5 - Serve
- Worker n - .....

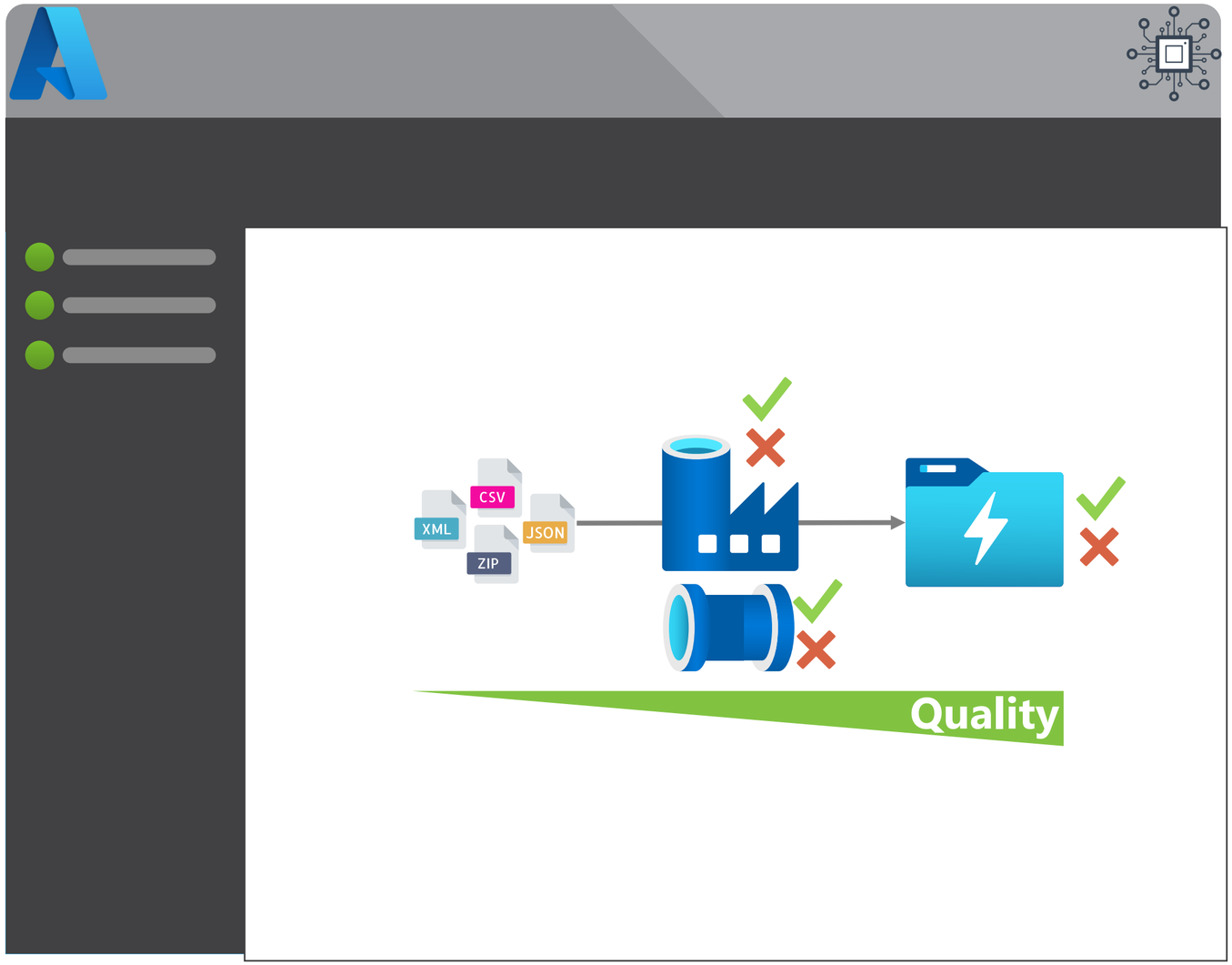
Go to Visio file in GitHub:
github.com/mrpaulandrew/procfwk/blob/master/Images

Hub & Spoke Integration Architecture



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

Testing



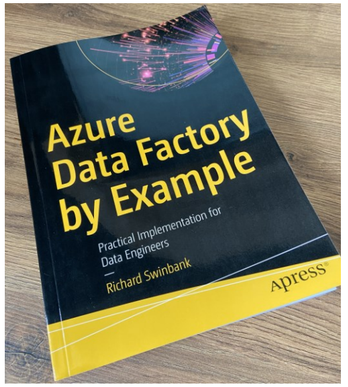
Types of Testing

- **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

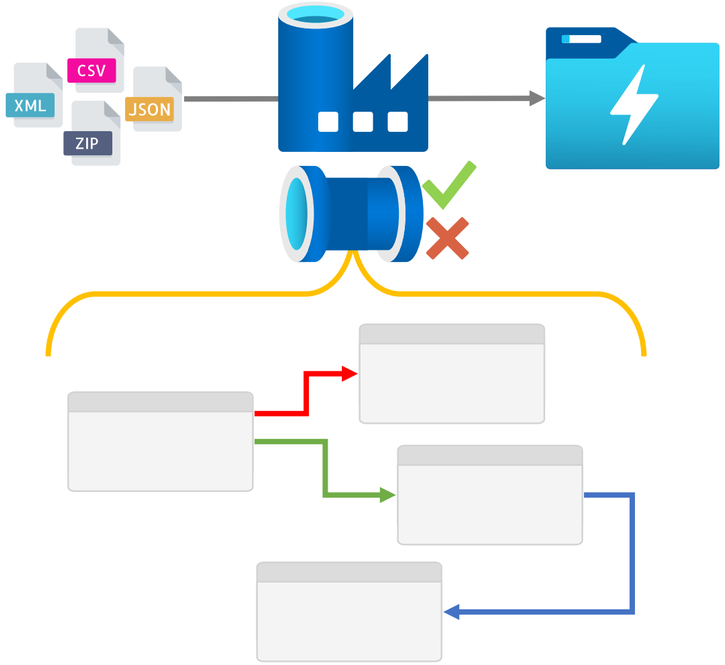
https://mrpaulandrew.com/2021/06/29/azure-data-factory-by-exampe-a-review-of-my-technical-review/

Types of Testing

What do they mean in our pipelines?

Integration Tests

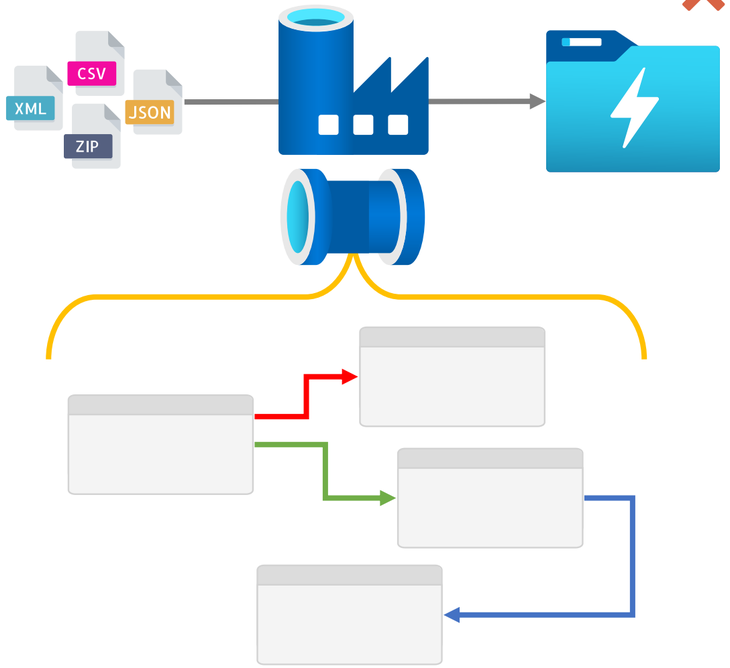
Simple Run Outcome



Functional Tests

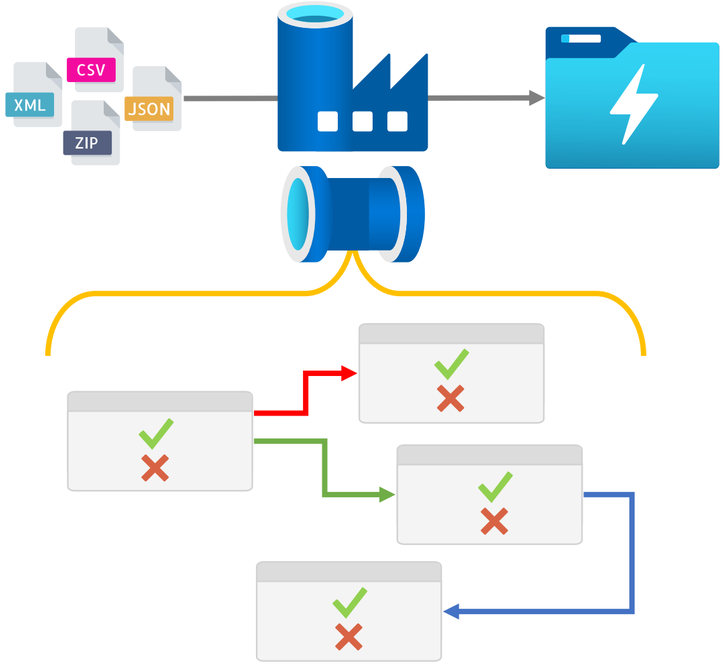
Sample Data

Specific Results



Unit Tests

Sample Data

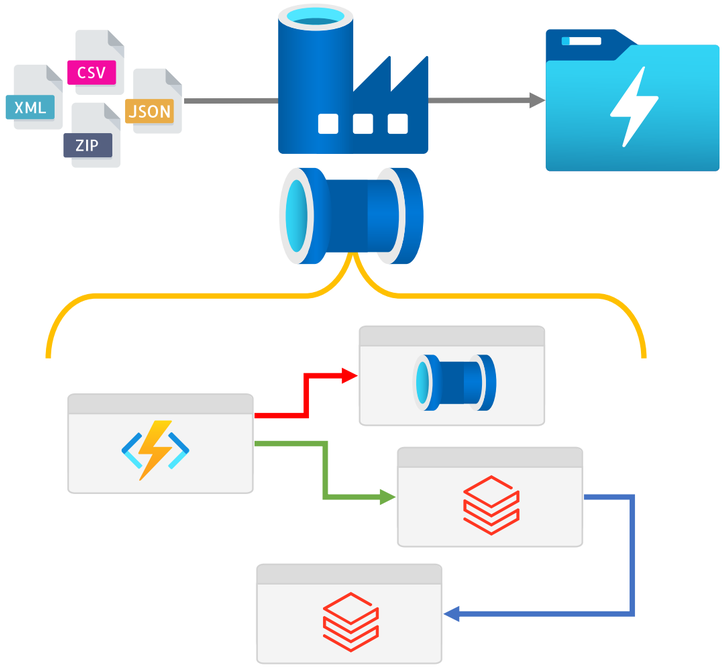


Specific Activity Inputs & Outputs

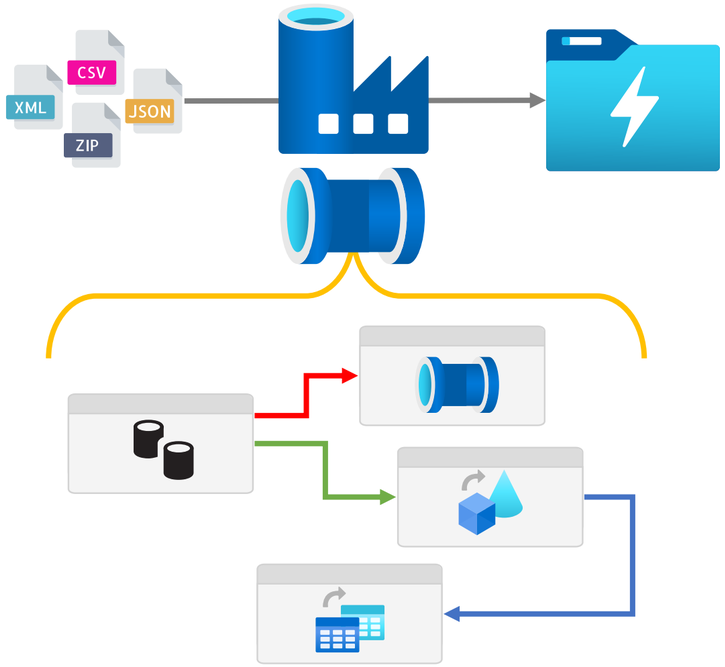
Types of Testing

Does it depend on external resources?

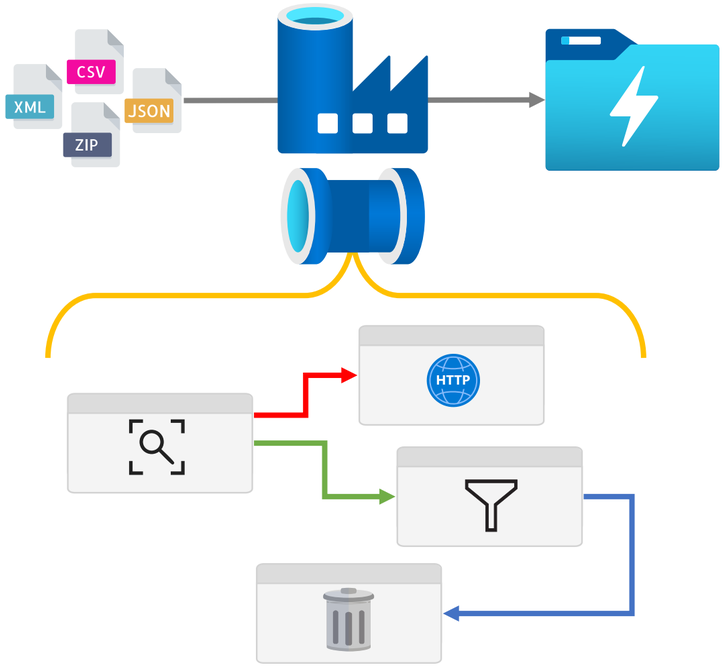
Integration Tests



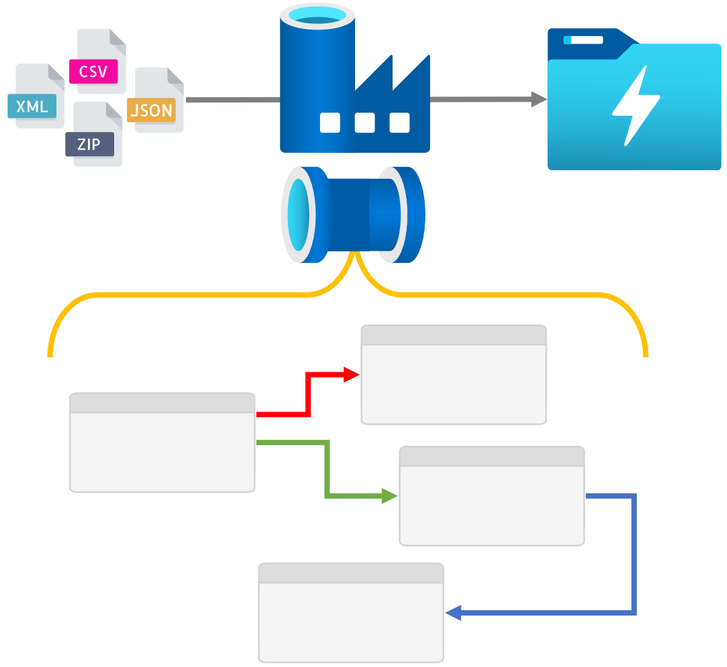
Functional Tests

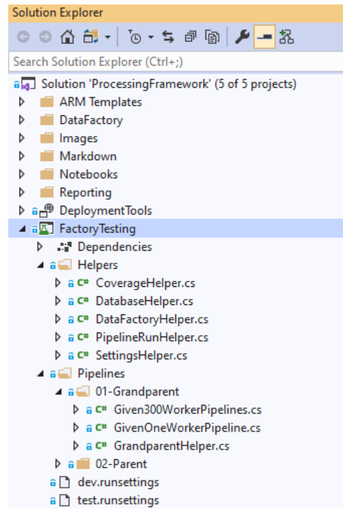


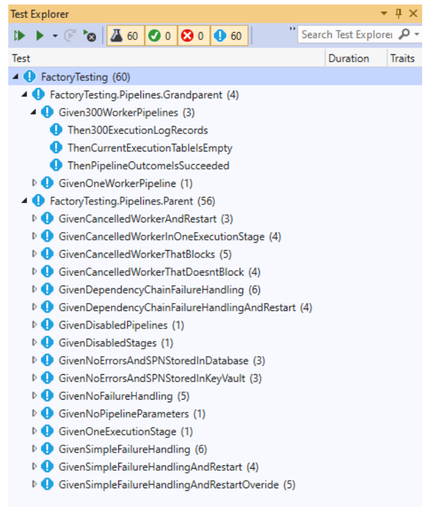
Unit Tests



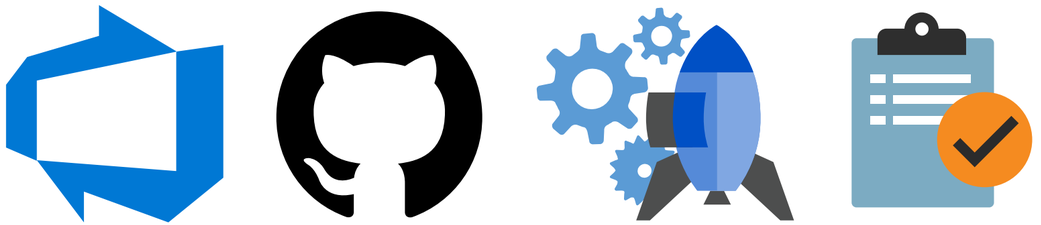
How To Run & Automate Tests



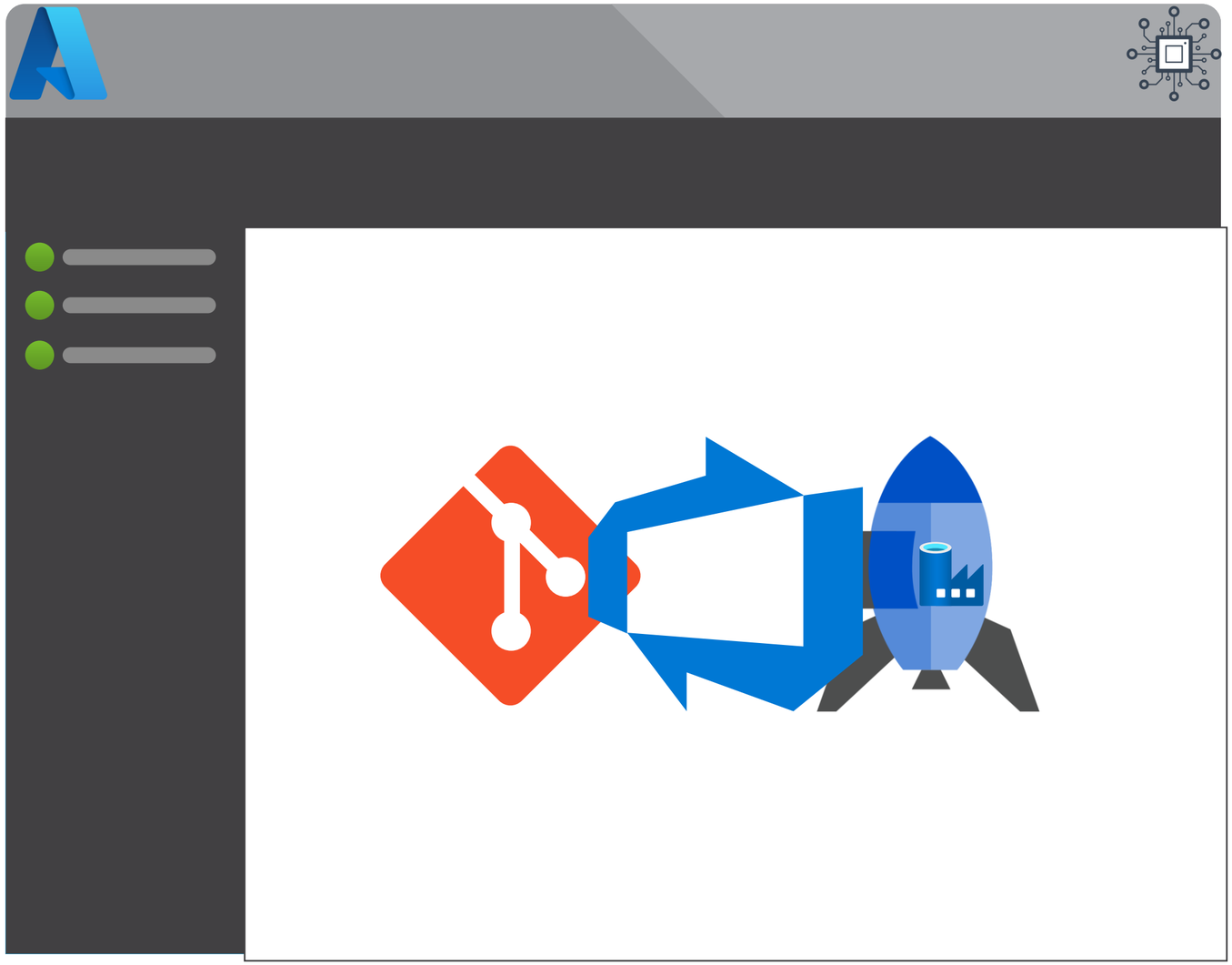




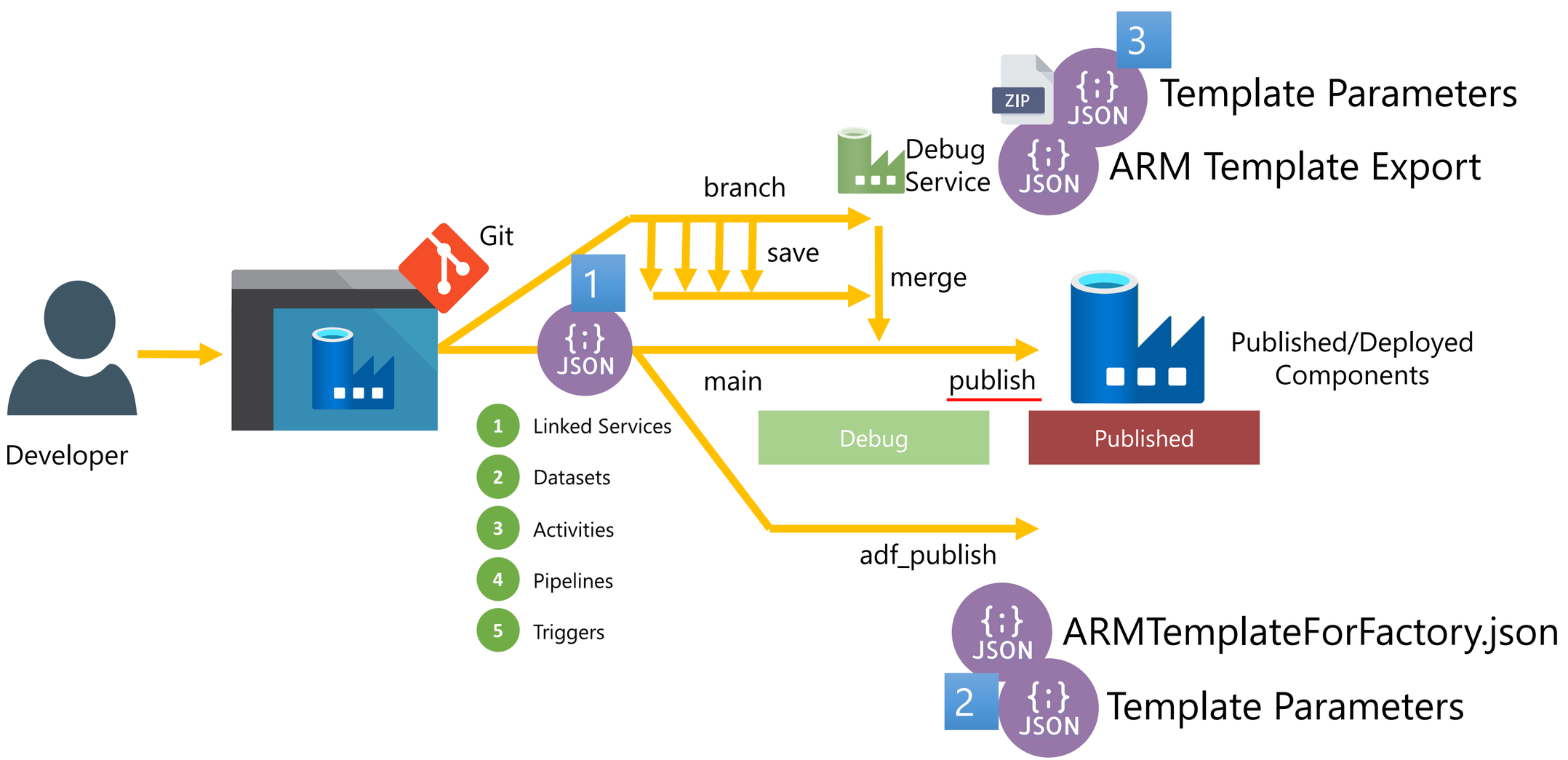
github.com/richardswinbank/community/tree/main/adf-testing-series



CI/CD

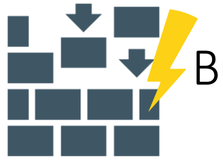


Getting Our ADF Source Code



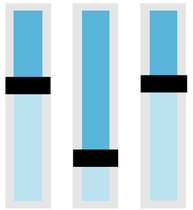
Data Factory Continuous Delivery

Azure DevOps



Build Artifacts

Pipelines



Key Vault Linked
Variable Groups

YML

Test

Run PowerShell

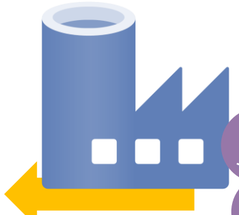
```
Import-Module -Name "azure.datafactory.tools"
Publish-AdfV2FromJson -RootFolder $AdfPath `
  -ResourceGroupName $resourceGroupName `
  -DataFactoryName $dataFactoryName `
  -Location $region `
  -Stage $configFilePath
```

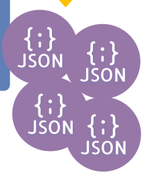
Release
Approver

YML

Production

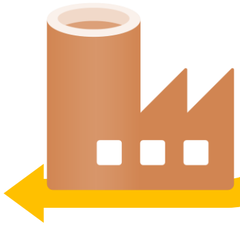
Publish Azure Data Factory





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

{release} / {feature} / {tag}



1 Linked Services

2 Datasets

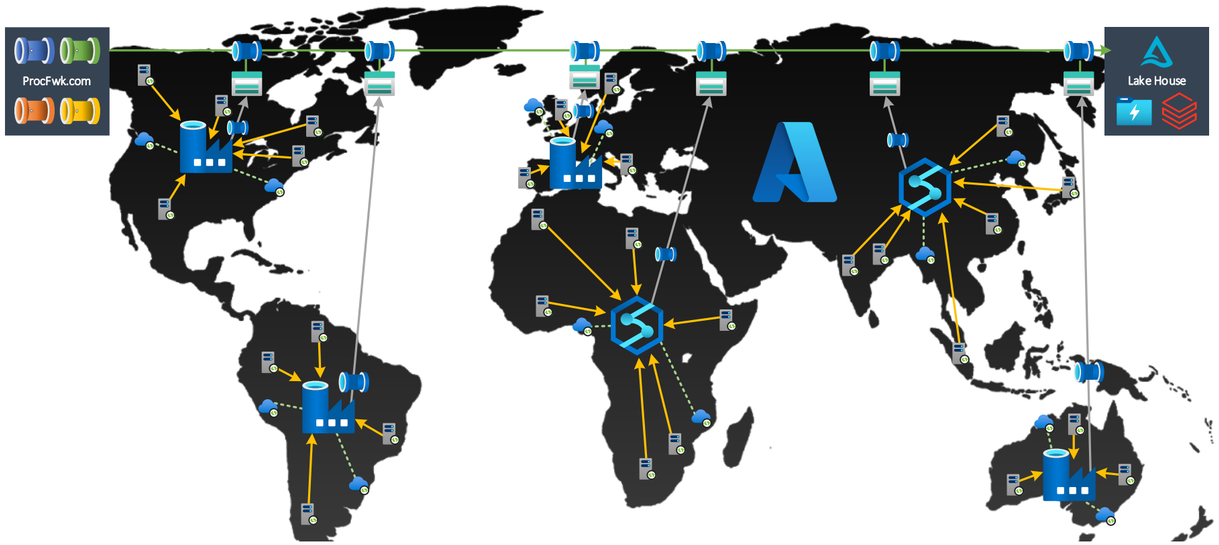
3 Activities

4 Pipelines

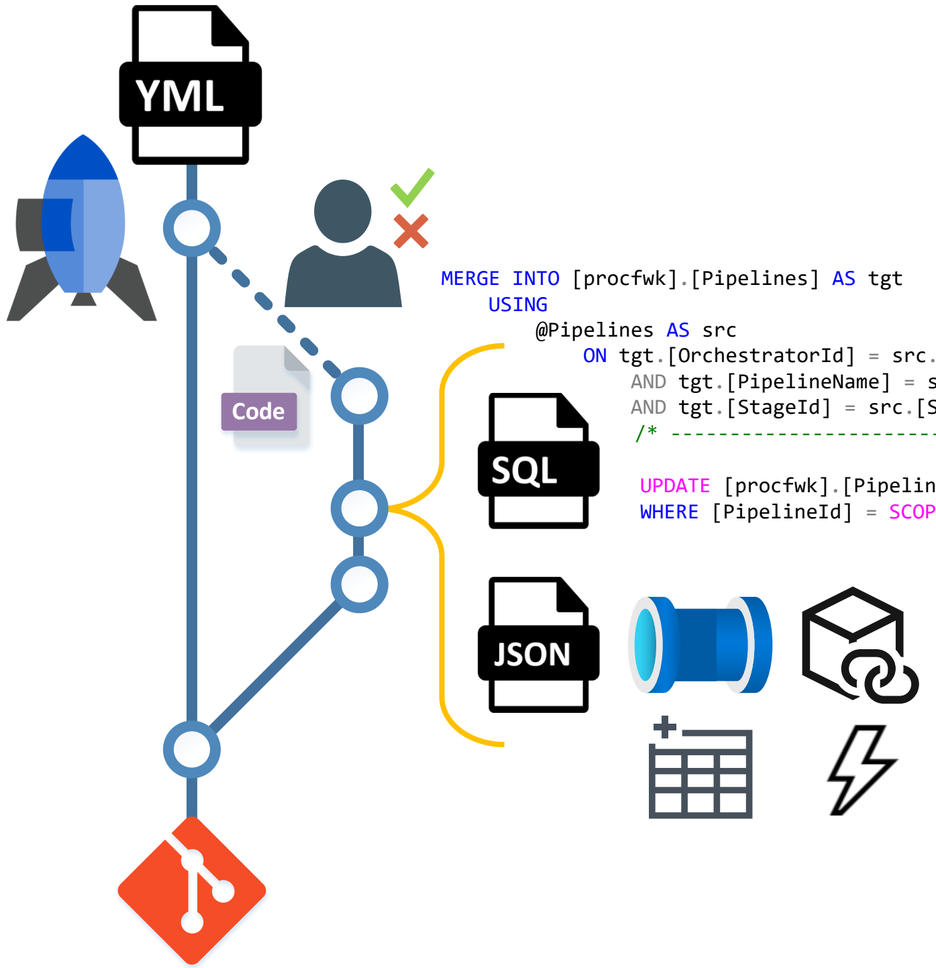
5 Triggers

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

How Small Can Deployments Be?



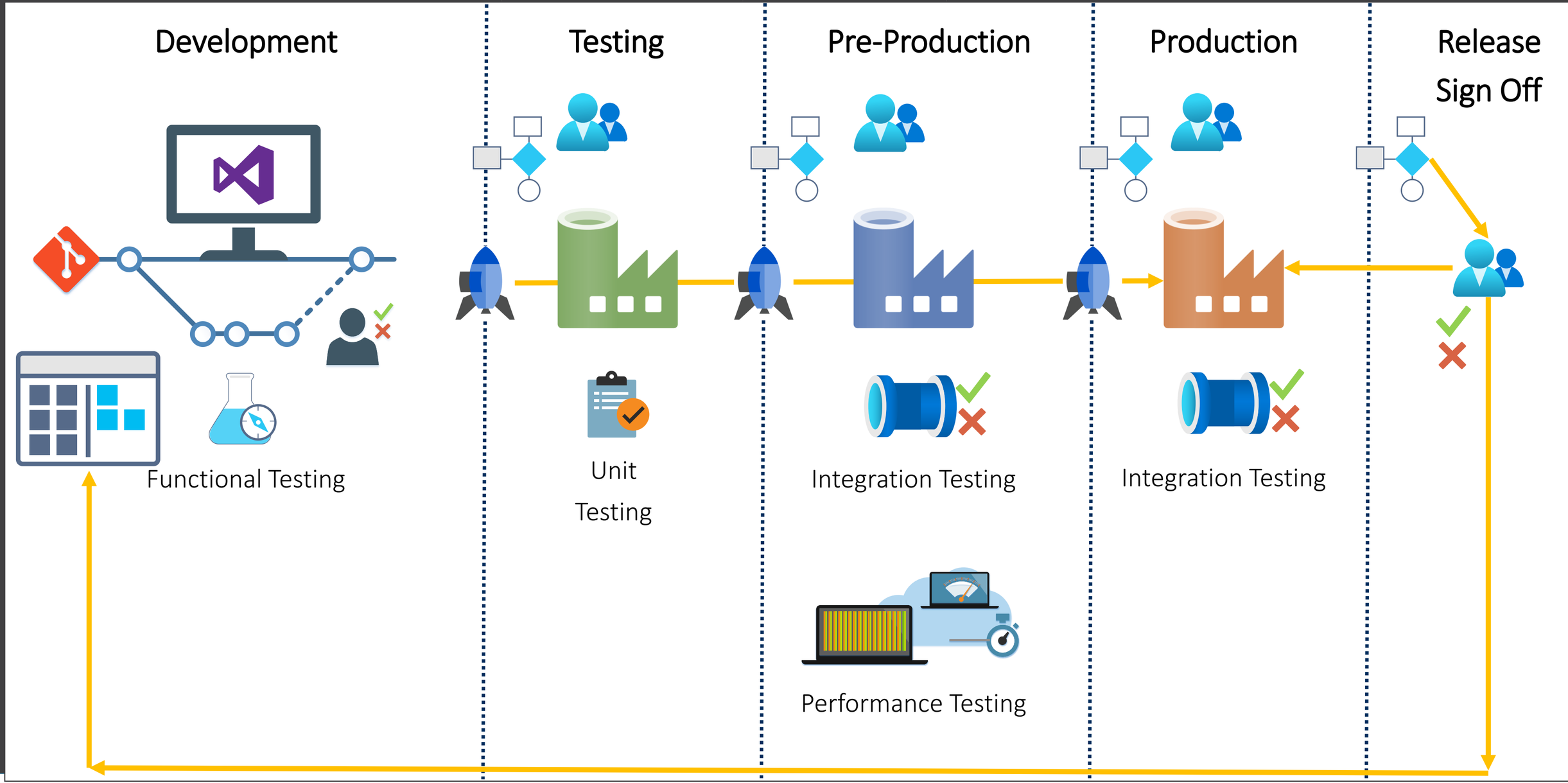
```
UPDATE [procfwk].[Pipelines] SET [Enabled] = 1
WHERE [PipelineId] = SCOPE_IDENTITY();
```



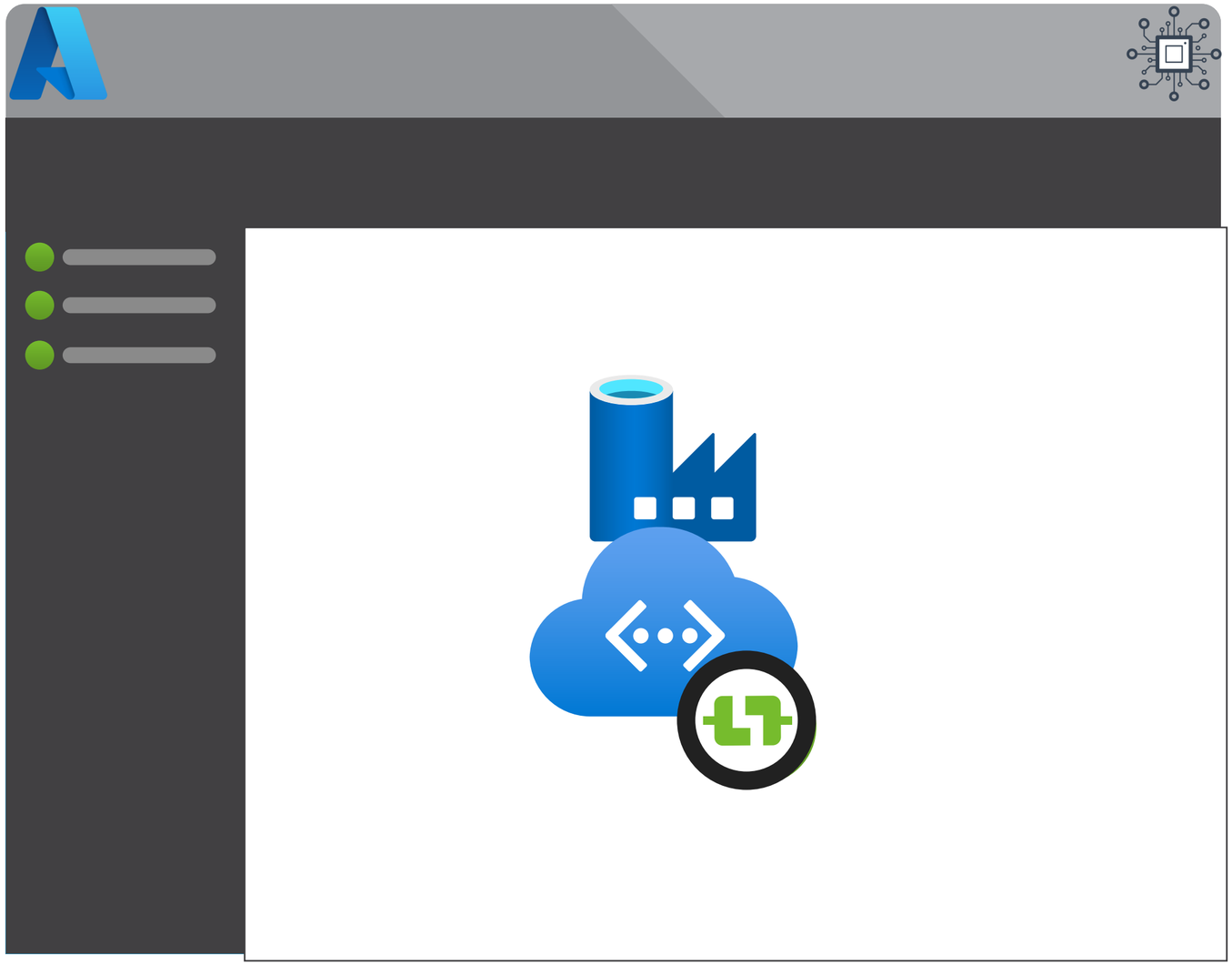

```
MERGE INTO [procfwk].[Pipelines] AS tgt
    USING
        @Pipelines AS src
            ON tgt.[OrchestratorId] = src.[OrchestratorId]
            AND tgt.[PipelineName] = src.[PipelineName]
            AND tgt.[StageId] = src.[StageId]
            /* ------------------------------------ */

            UPDATE [procfwk].[Pipelines] SET [Enabled] = 0
            WHERE [PipelineId] = SCOPE_IDENTITY();
```

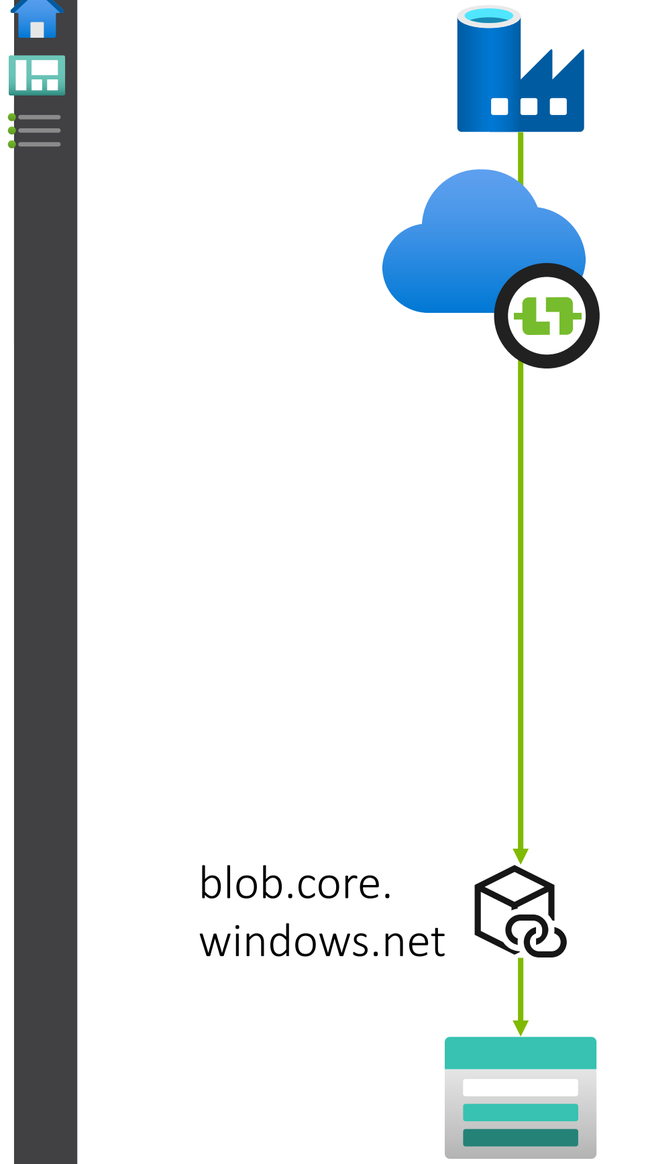
Deployment Life Cycle & Gateway

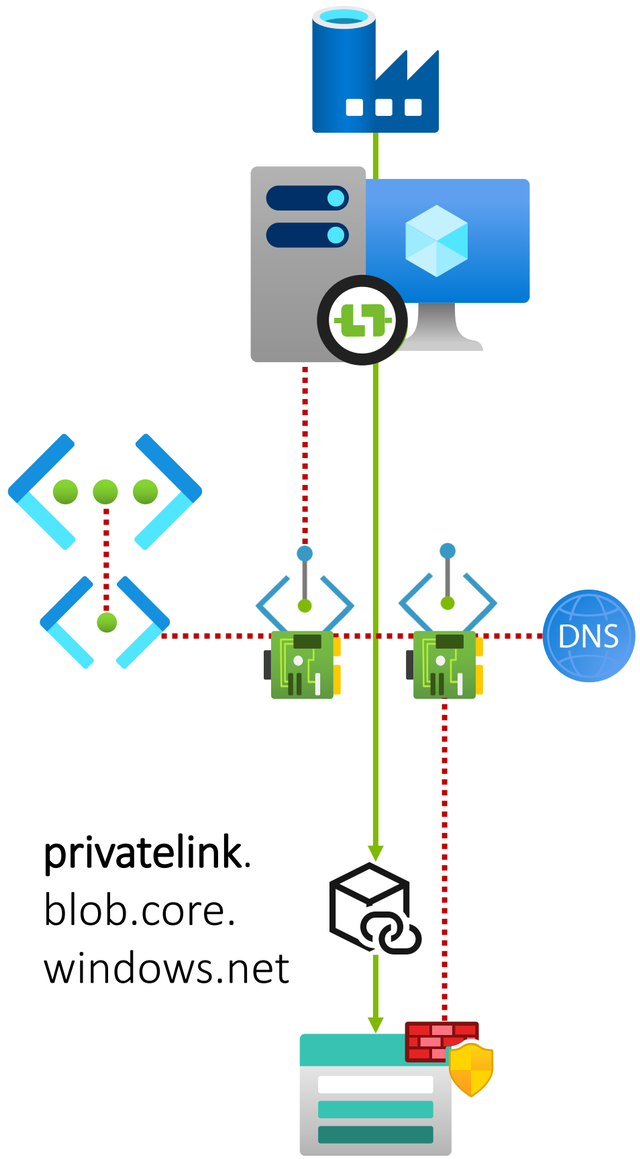


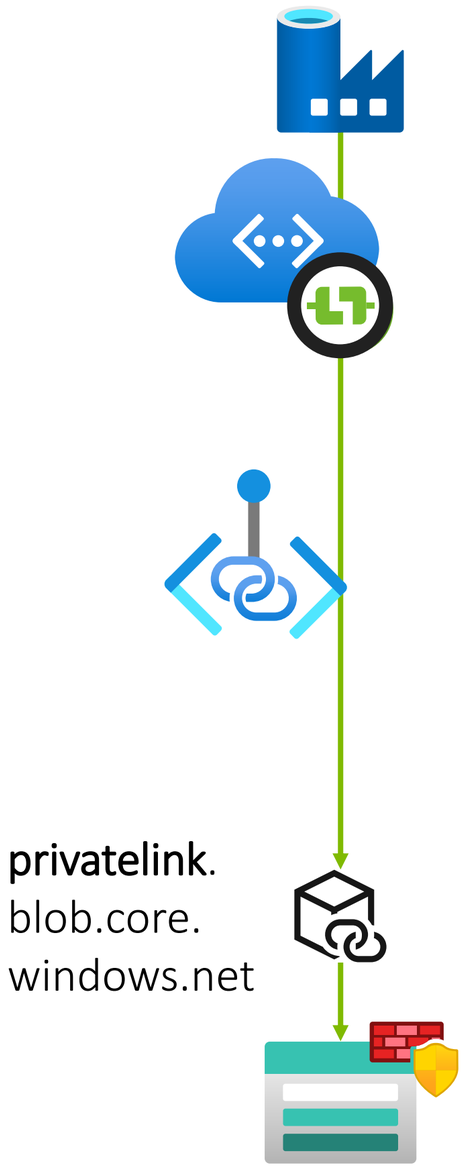
VNet Integration



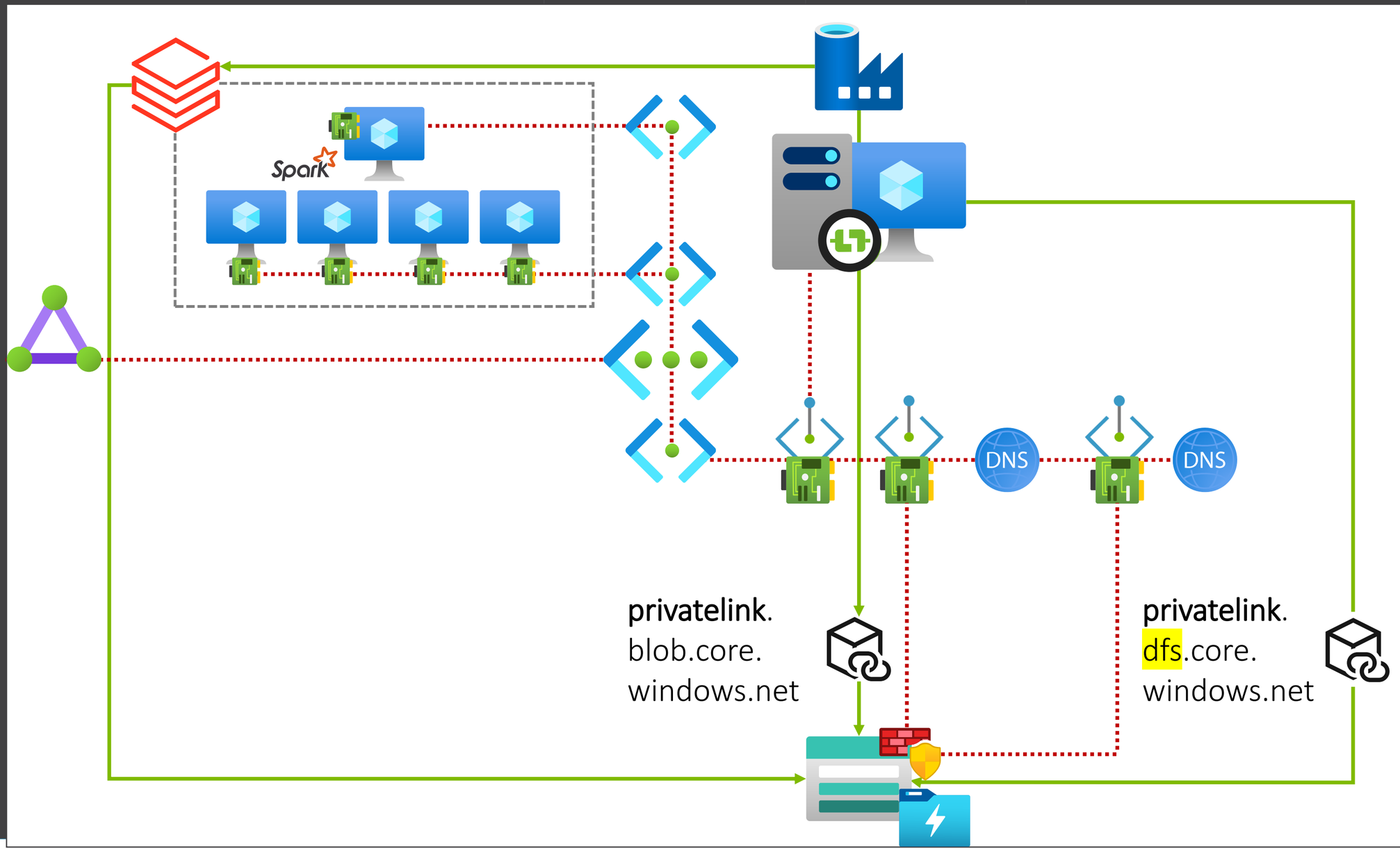
Connections Options





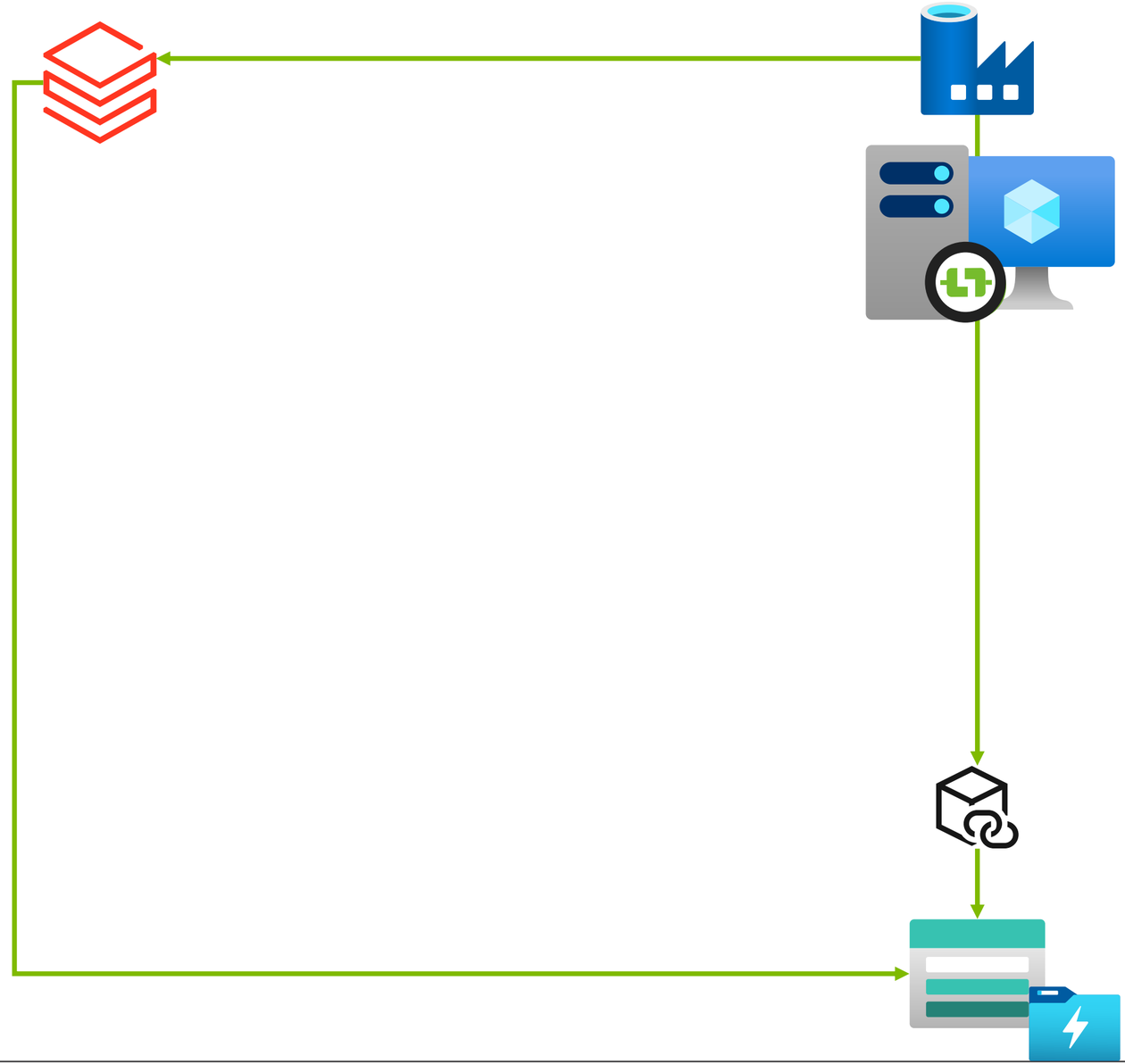


Further VNet Connections

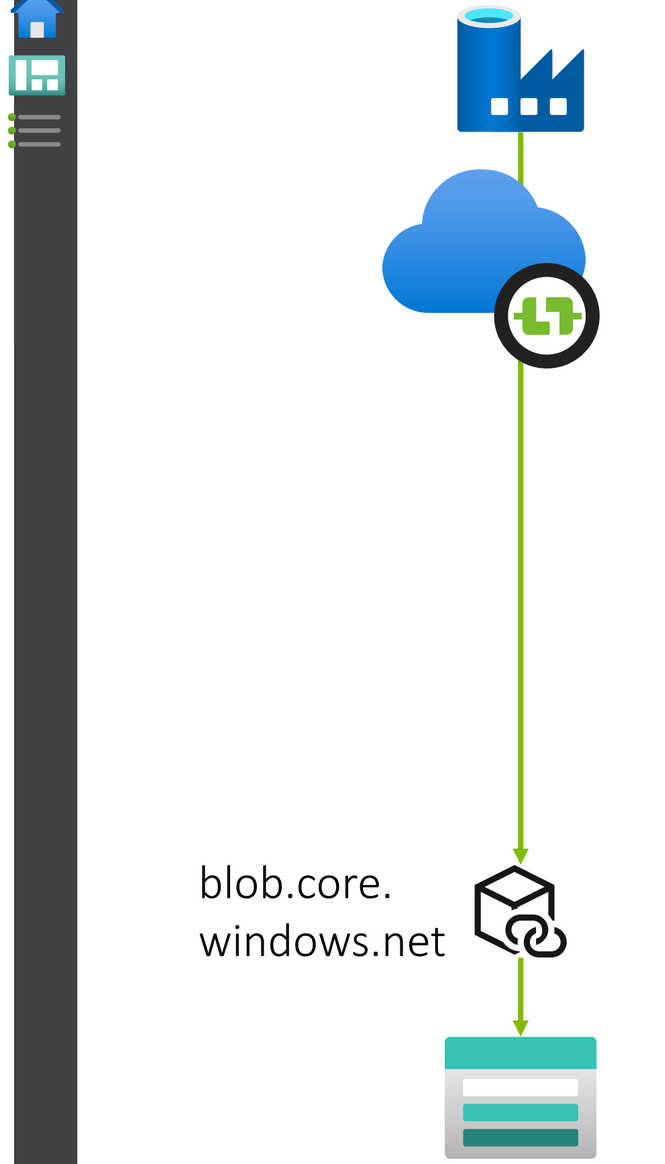


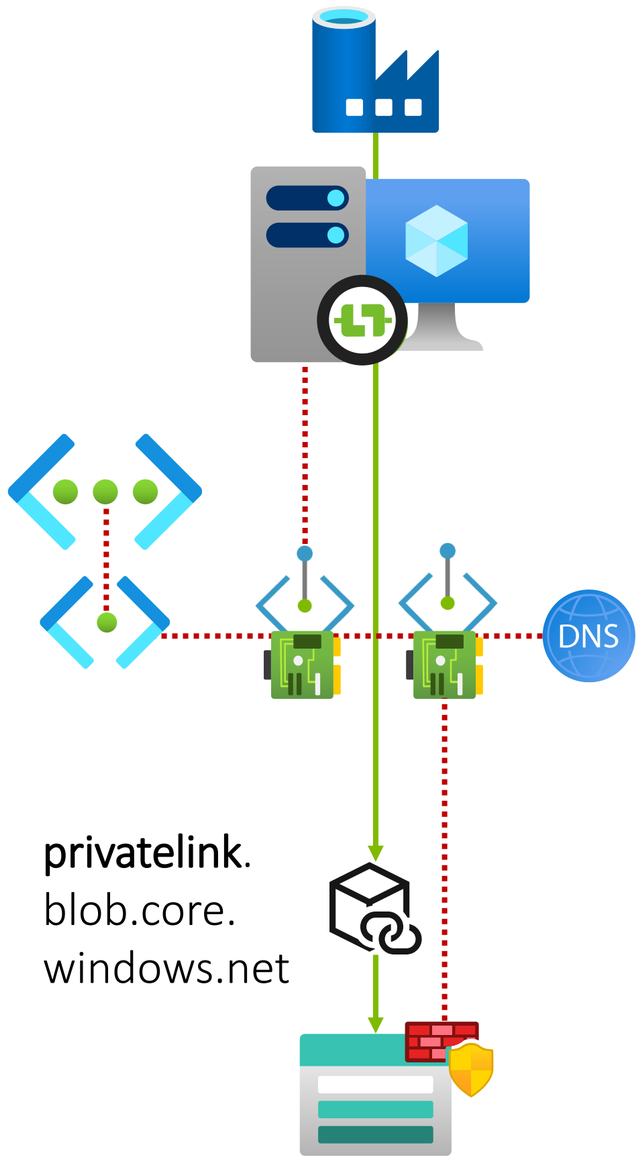
Further ~~VNet~~ Connections

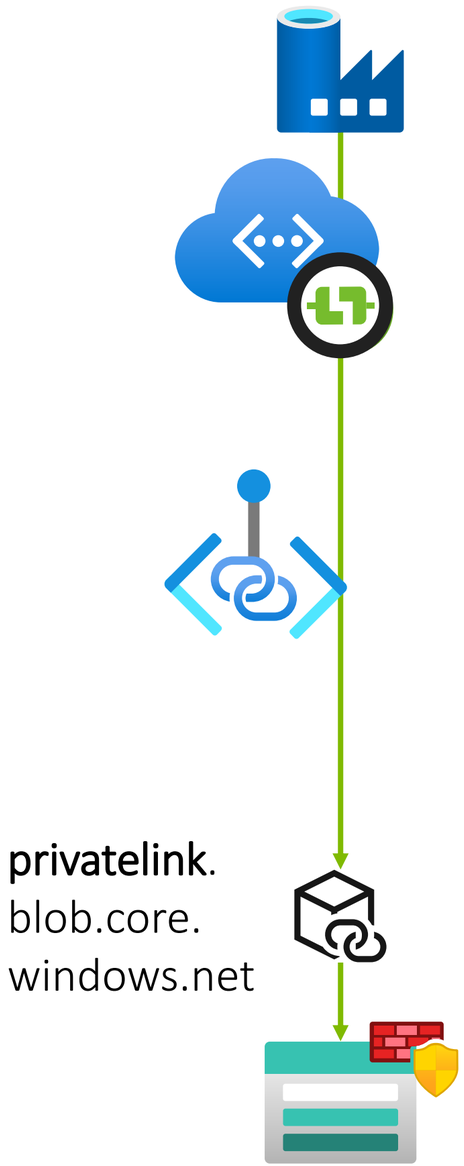
For a Data Engineer



Connection Options



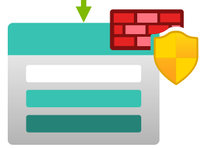




Connection Options

Managed Identity

blob.core.
windows.net

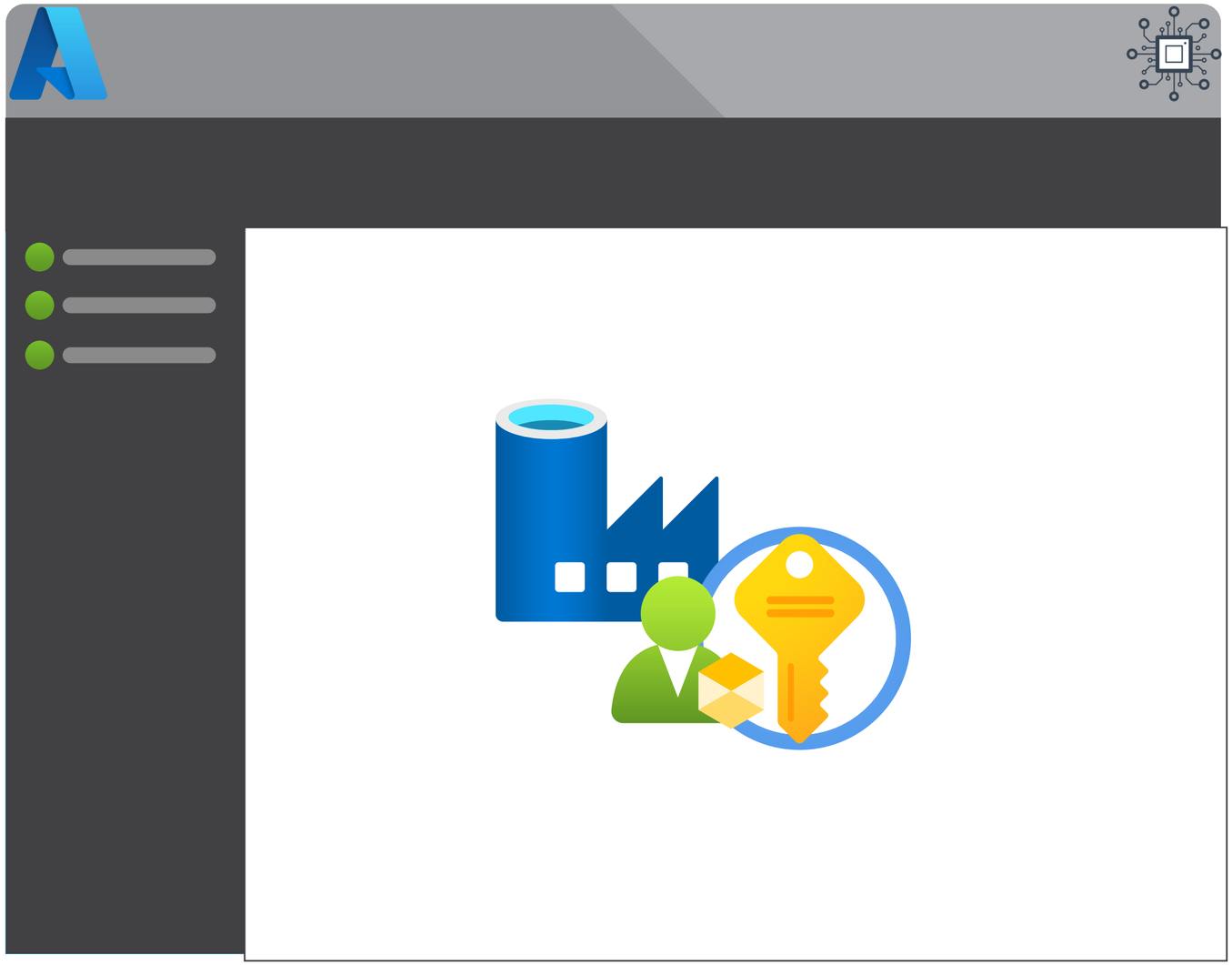


** Trusted access based on system-assigned managed identity.*

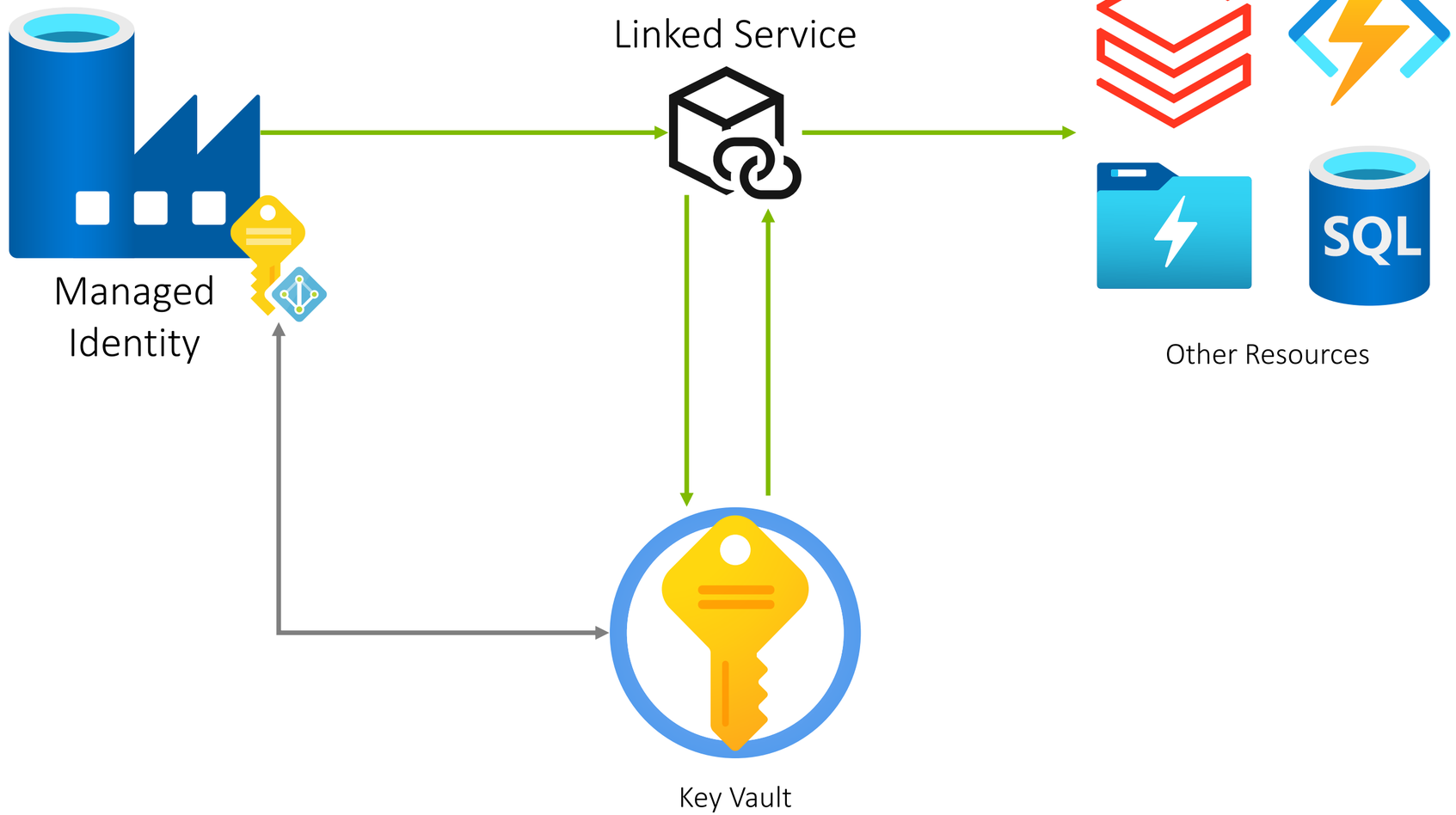
Microsoft.DataFactory/factories

https://docs.microsoft.com/en-us/azure/storage/common/storage-network-security?tabs=azure-portal

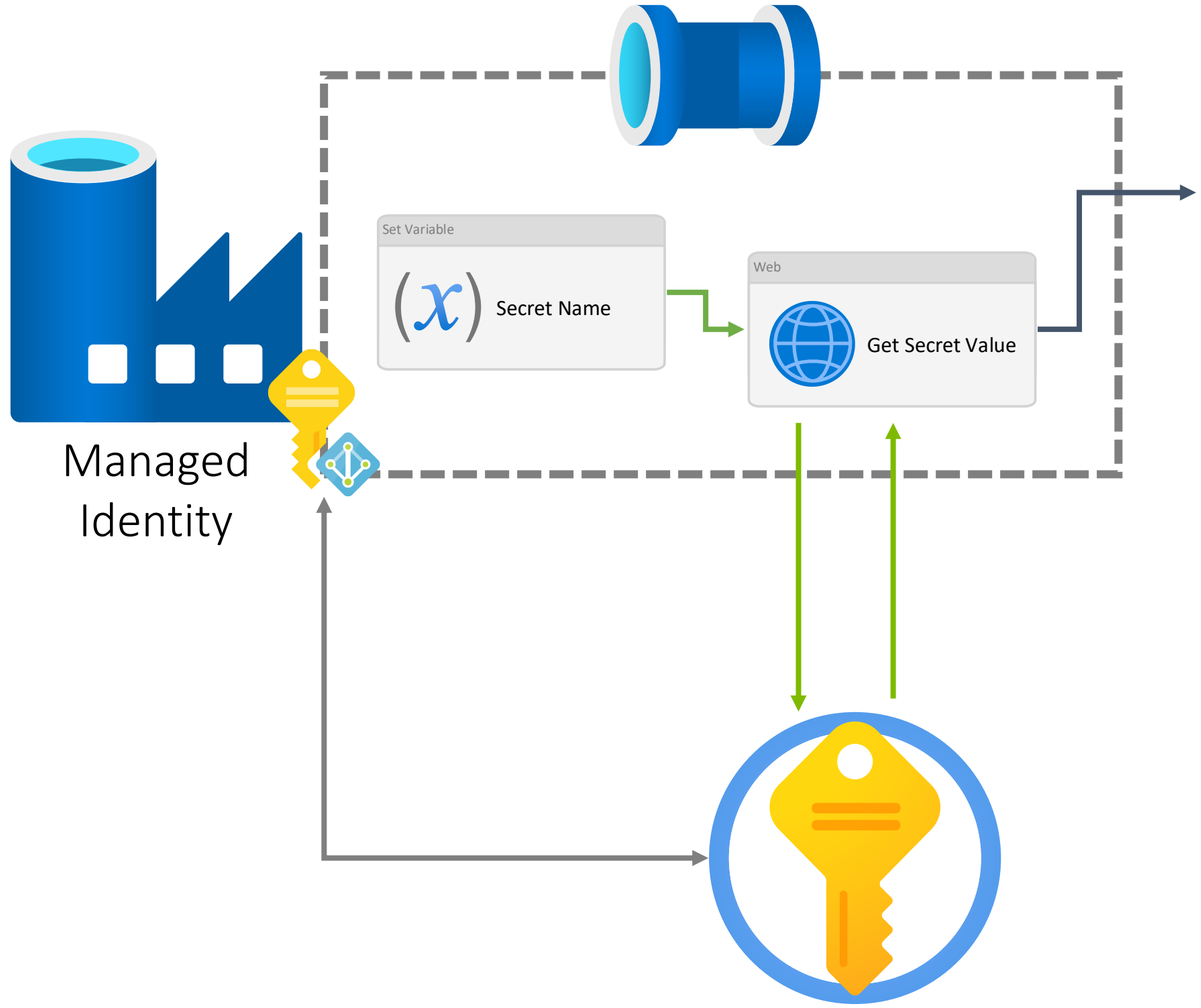
Custom Security Roles & Key Vault Integration



Key Vault Integration – Option 1



Key Vault Integration – Option 2



Output


```
{
  "value": "HelloWorld!",
  "id": "https://trainingkeys01.vault.azure.net/secrets/DemoKeyGetWithWebActivity/0b8ccf8e52b241eaac58ba33c7a4d8c6",
  "attributes": {
    "enabled": true,
    "created": 1645623501,
    "updated": 1645623501,
    "recoveryLevel": "Recoverable+Purgeable"
  },
  "tags": {},
  "ADFWebActivityResponseHeaders": {
    "Pragma": "no-cache",
    "x-ms-keyvault-region": "uksouth",
    "x-ms-request-id": "a17107f2-89e3-45b4-81d1-637d92d575d0",
    "x-ms-keyvault-service-version": "1.9.291.1",
    "x-ms-keyvault-network-info": "conn_type=Ipv4;addr=51.104.25.10;act_addr_fam=InterNetwork;",
    "Strict-Transport-Security": "max-age=31536000;includeSubDomains",
    "X-Content-Type-Options": "nosniff",
    "Cache-Control": "no-cache",
    "Date": "Wed, 23 Feb 2022 13:42:03 GMT",
    "X-Powered-By": "ASP.NET",
    "Content-Length": "258",
    "Content-Type": "application/json; charset=utf-8",
    "Expires": "-1"
  },
  "effectiveIntegrationRuntime": "AutoResolveIntegrationRuntime (UK South)",
  "executionDuration": 0,
  "durationInQueue": {
    "integrationRuntimeQueue": 1
  },
  "billingReference": {
    "activityType": "ExternalActivity",
    "billableDuration": [
      {
        "meterType": "AzureIR",
        "duration": 0.016666666666666666,
        "unit": "Hours"
      }
    ]
  }
}
```


https://docs.microsoft.com/en-us/azure/data-factory/how-to-use-azure-key-vault-secrets-pipeline-activities

Accessing Everything via Data Factory

Azure Tenant

Azure Data Factory

Connection Administrator

Pipeline Executor

Dataset Reader

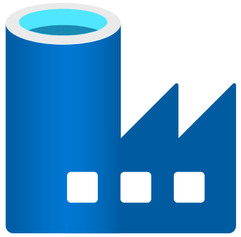
Azure Active Directory

User Account

- Owner
- Contributor
- Reader
- Data Factory Contributor
- Log Analytics Contributor
- Log Analytics Reader
- Managed Application Contributor Role
- Managed Application Operator Role
- Managed Applications Reader
- Monitoring Contributor
- Monitoring Metrics Publisher
- Monitoring Reader
- Resource Policy Contributor
- User Access Administrator

Accessing Data Factory – Custom Roles

Azure Tenant



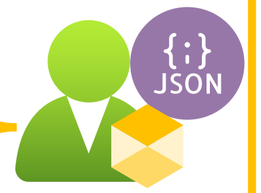
Azure Data Factory

Azure Active Directory

User Account

- Owner
- Contributor
- Reader
- Data Factory Contributor
- Log Analytics Contributor
- Log Analytics Reader
- Managed Application Contributor Role
- Managed Application Operator Role
- Managed Applications Reader
- Monitoring Contributor
- Monitoring Metrics Publisher
- Monitoring Reader
- Resource Policy Contributor
- User Access Administrator

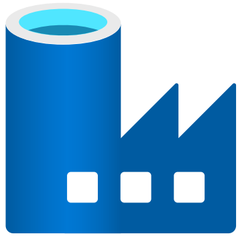
```
"Actions": [
    "Microsoft.DataFactory/operations/read",
    "Microsoft.DataFactory/factories/pipelines/read",
    "Microsoft.DataFactory/factories/linkedServices/read",
    "Microsoft.DataFactory/factories/datasets/read",
    "Microsoft.DataFactory/factories/dataflows/read",
    "Microsoft.DataFactory/datafactories/read"
],
```



ADF Pipeline Executor

ADF Reader

Accessing Data Factory – Custom Roles



Thank you for listening…

Paul Andrew

Microsoft® Most Valuable Professional

avanade

Blog: mrpaulandrew.com
YouTube: c/mrpaulandrew
Email: paul@mrpaulandrew.com

Twitter: @mrpaulandrew
LinkedIn: In/mrpaulandrew

GitHub: github.com/mrpaulandrew