STRATEGIC PARTNER

Microsoft

GOLD SPONSOR

elitmind think bright

Future Processing

SILVER SPONSOR

dbWatch DATABASE CONTROL

nationale nederlanden

BRONZE SPONSOR

predica.

Azure Orchestration – Applying Data Factory in Production

Paul Andrew | Principal Consultant & Solution Architect

altius


@MrPaulAndrew


In/MrPaulAndrew

https://github.com/mrpaulandrew

CommunityEvents

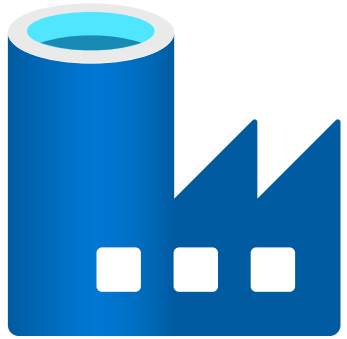
Demo code, content and slides from various community events.

C++

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

AGENDA ??

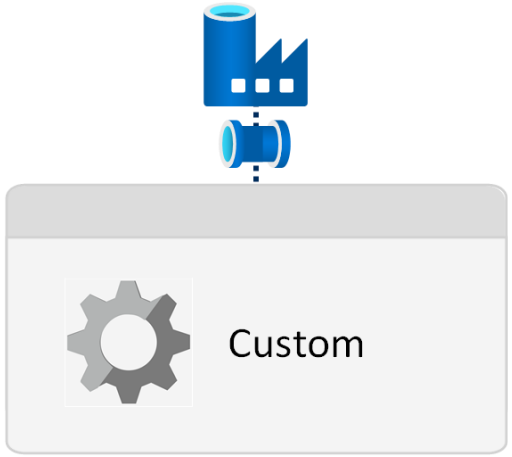
Applying Data Factory in Production



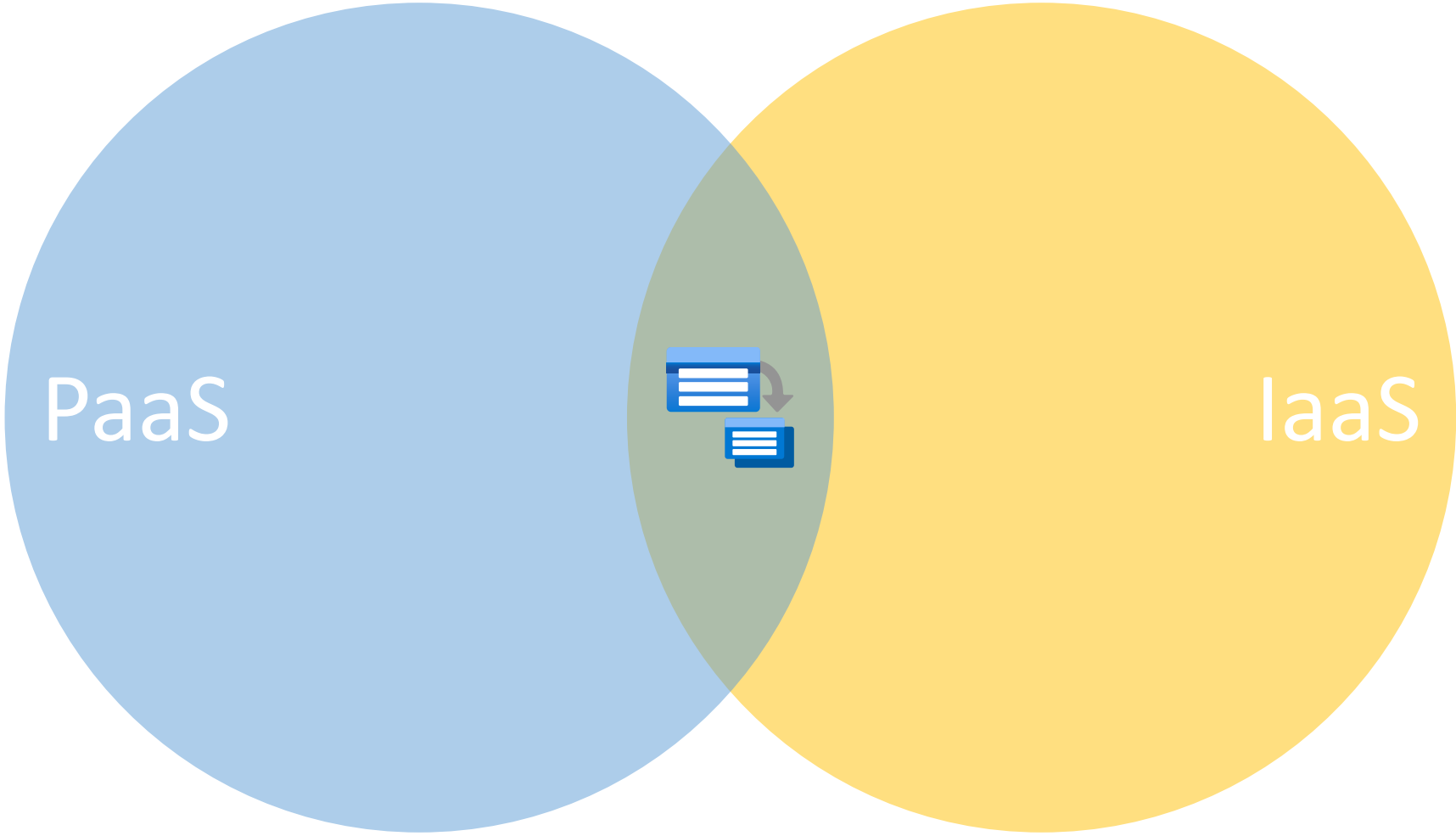
AGENDA – Short Stories

- Custom Activities
- Controlling & Scaling Compute
- Scale Out Execution
- Metadata Driven Pipelines
- Deploying Data Factory
- Best Practices

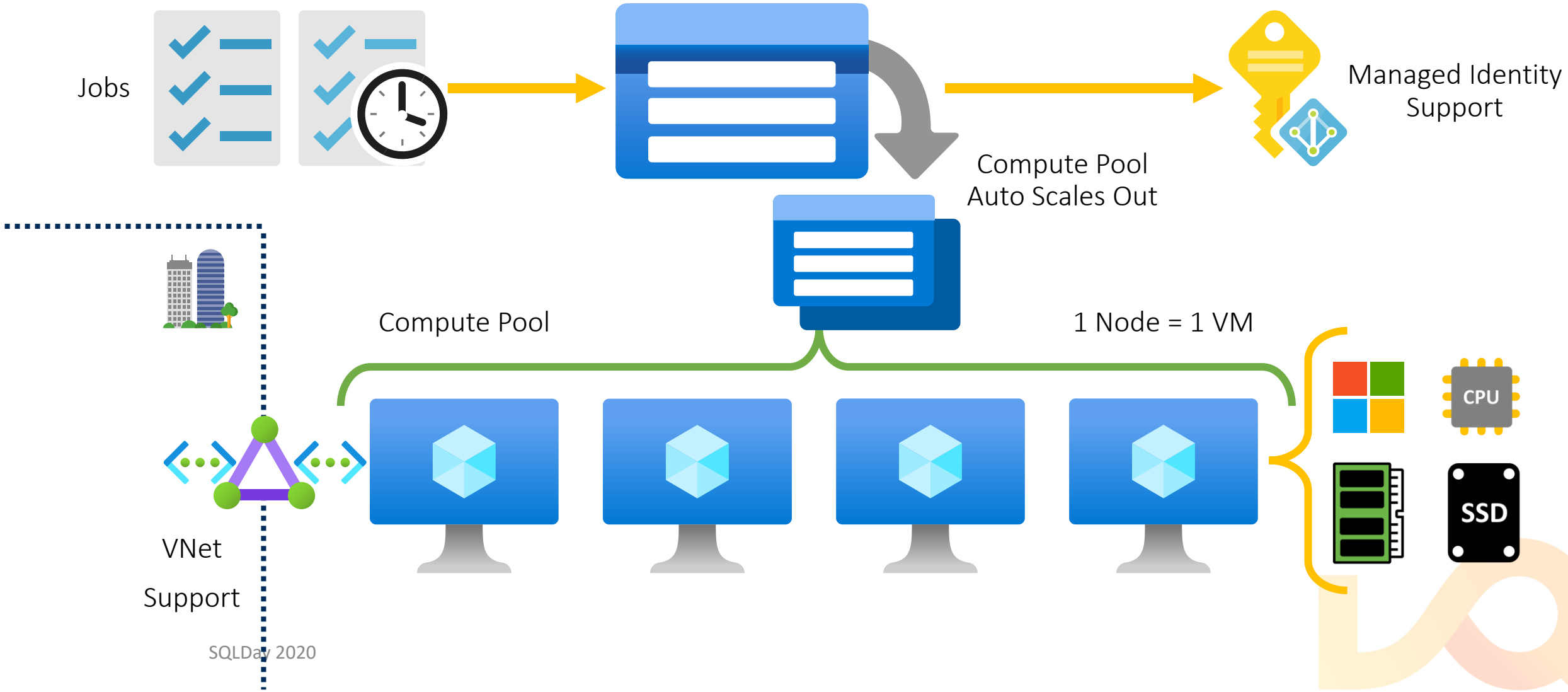
Custom Activities



Azure Batch Service



Azure Batch Service



Custom Activity

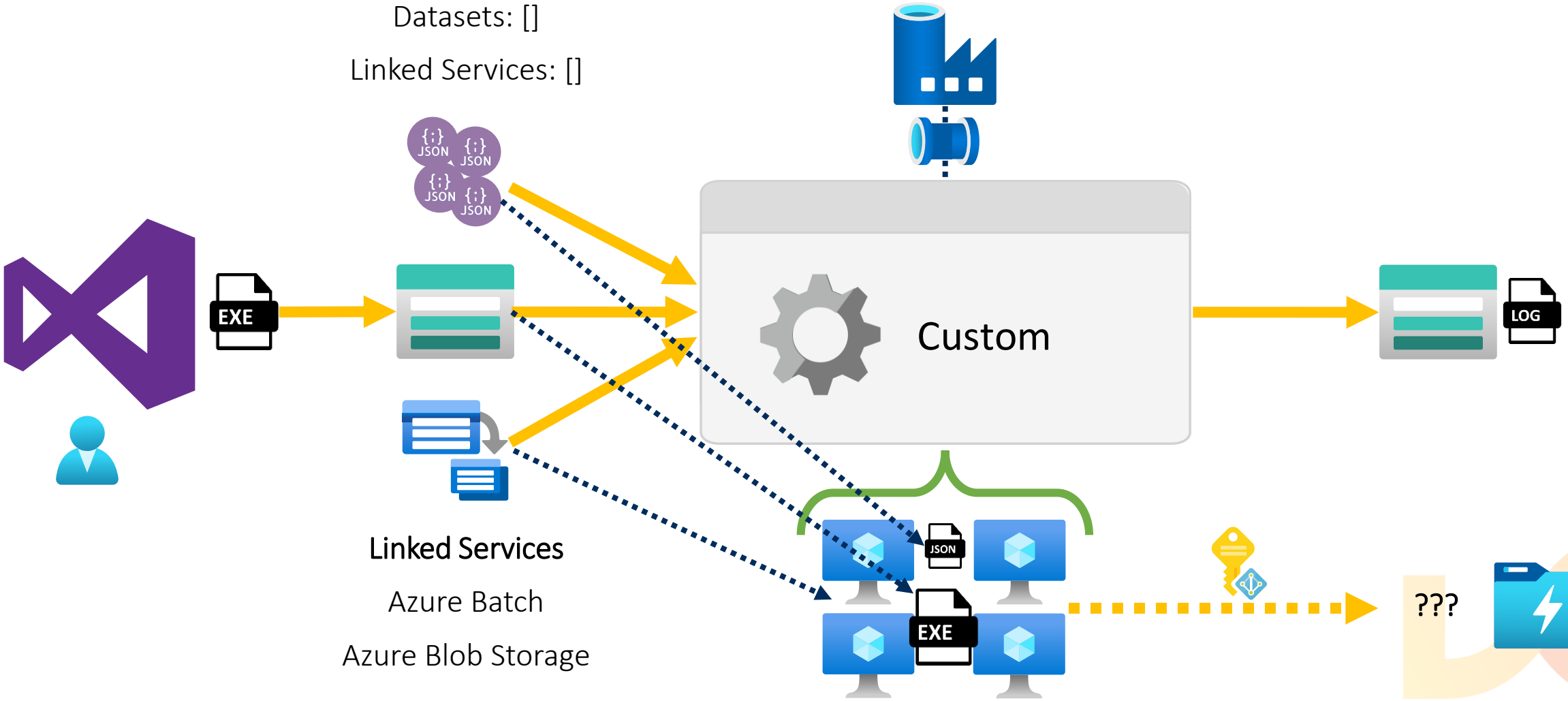
https://mrpaulandrew.com/2018/11/12/creating-an-azure-data-factory-v2-custom-activity/

Extend Data Factory with Custom Code

References Objects

Datasets: []

Linked Services: []

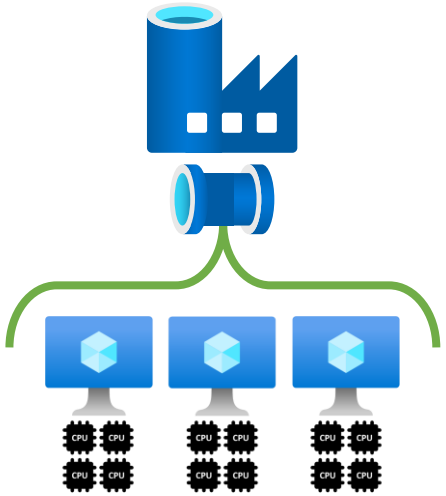


Linked Services

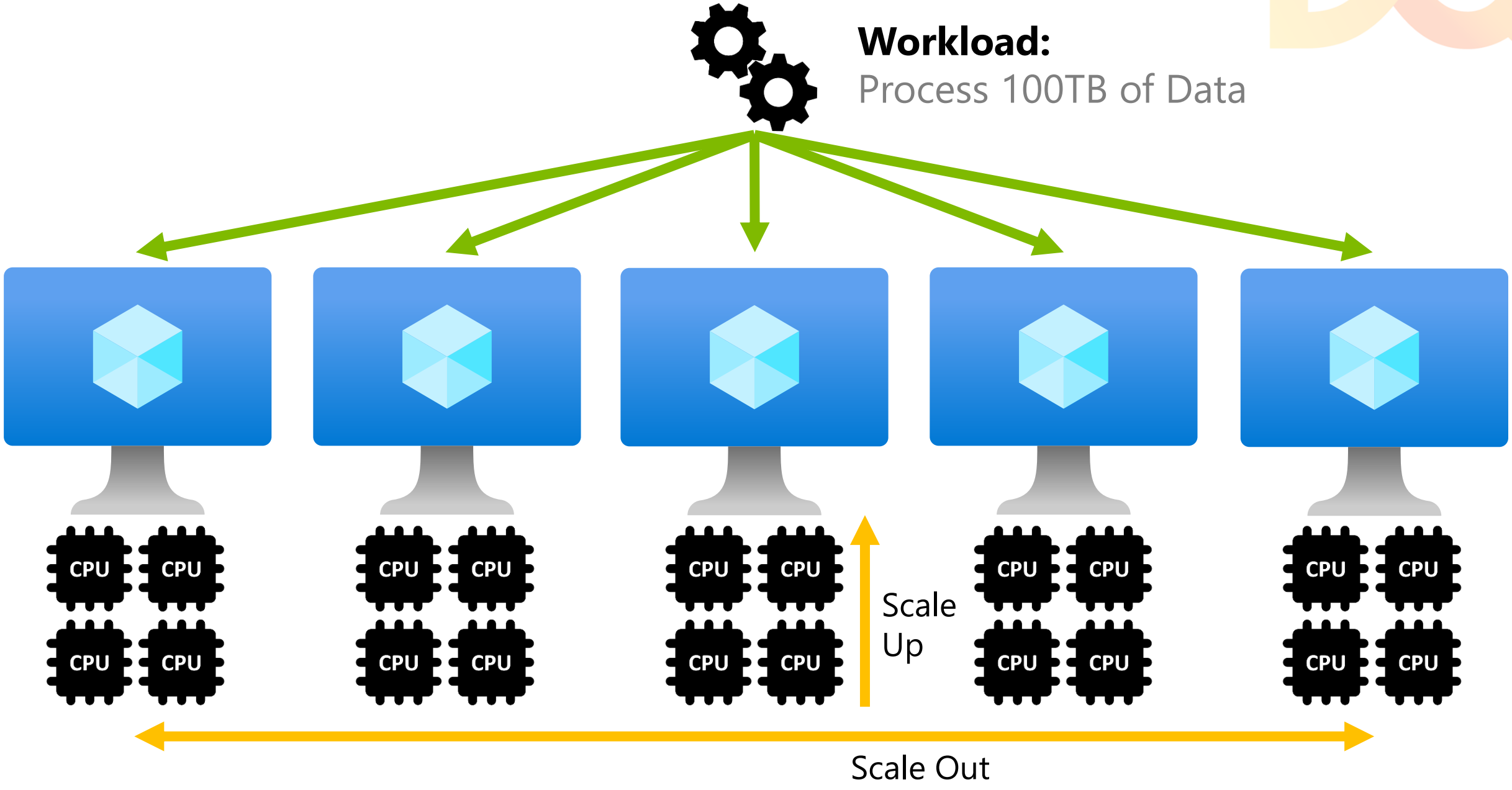
Azure Batch

Azure Blob Storage

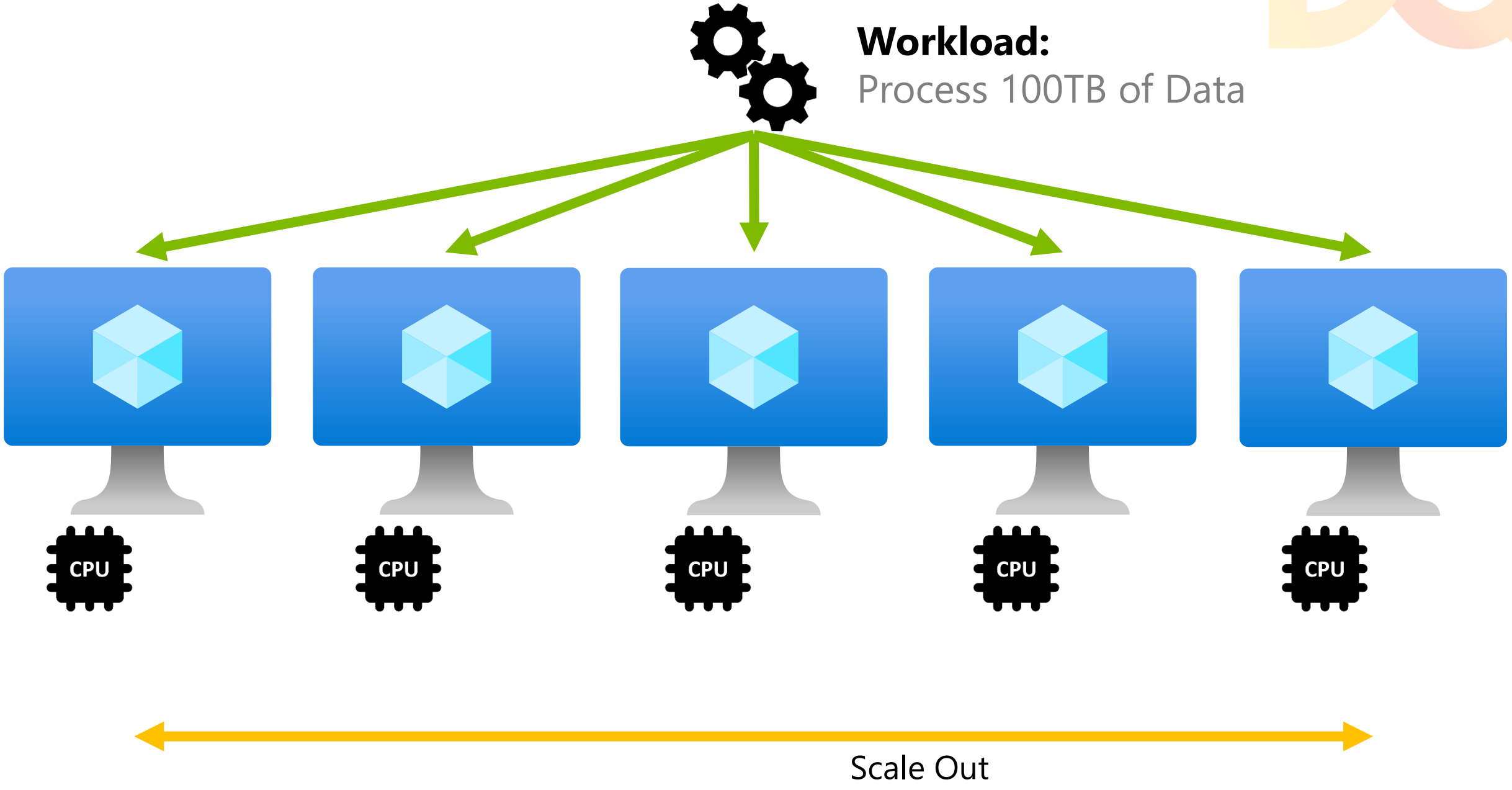
Controlling & Scaling Compute



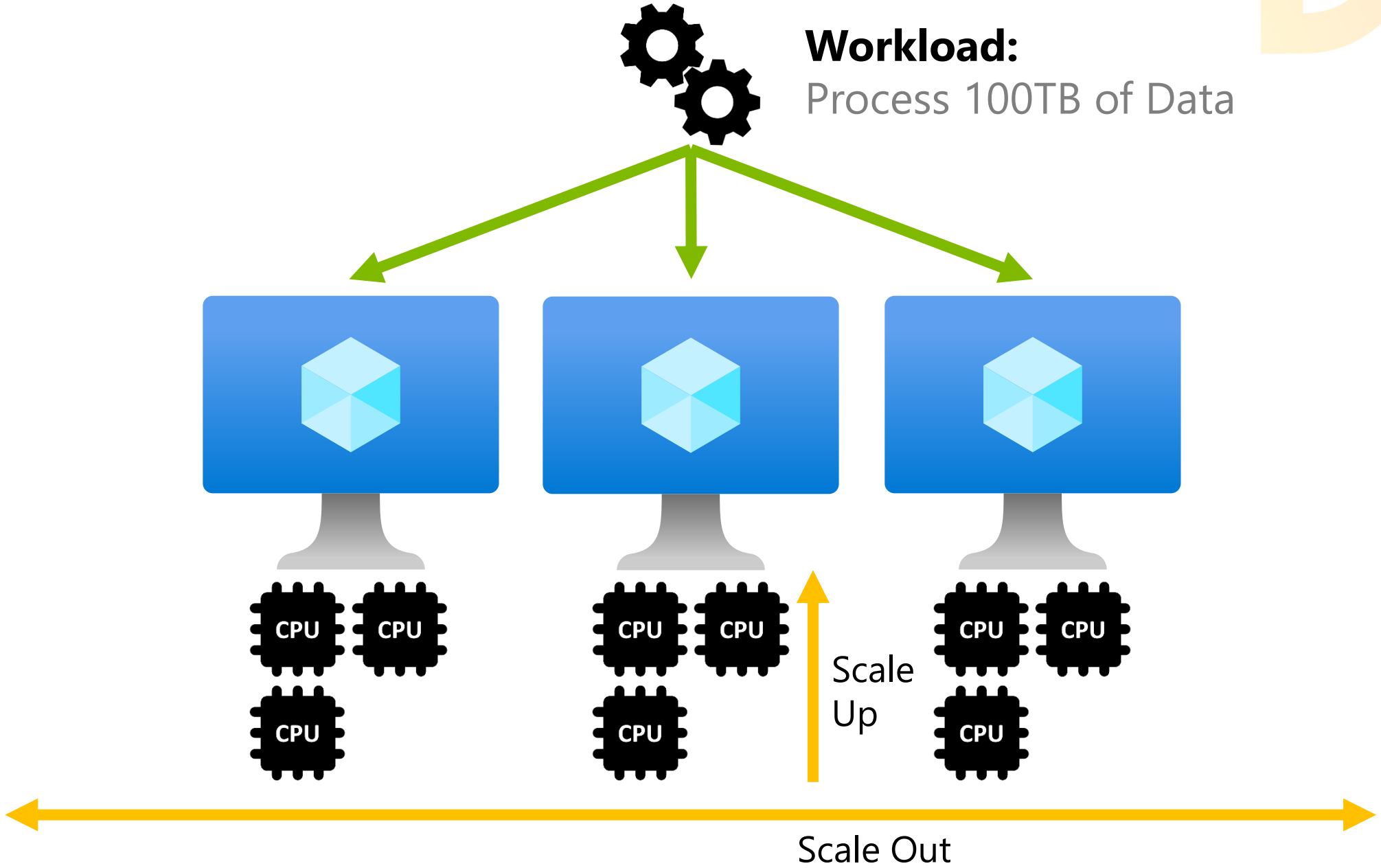
Scaling Up and/or Scaling Out



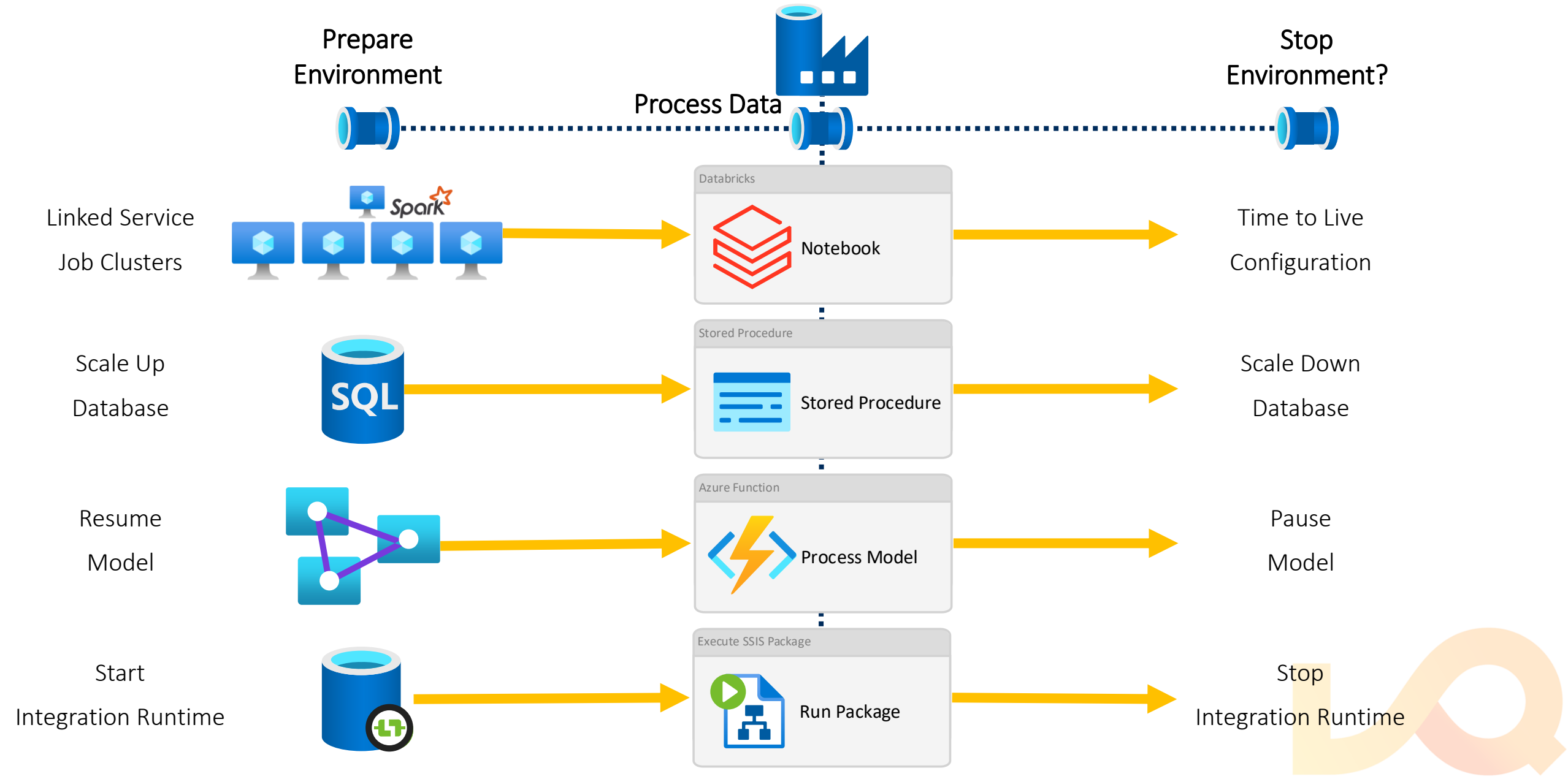
Scaling Up and/or Scaling Out



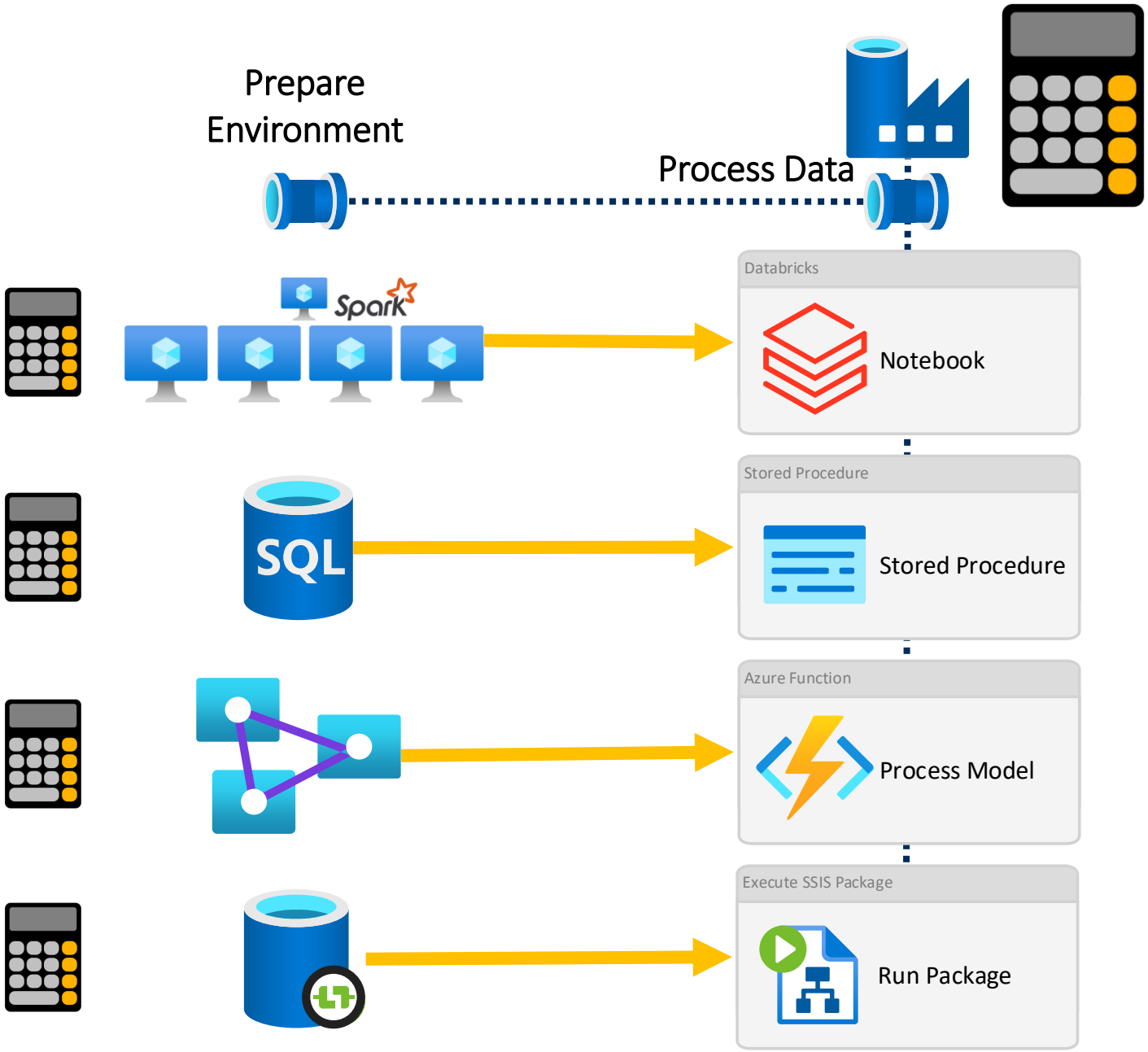
Scaling Up and/or Scaling Out



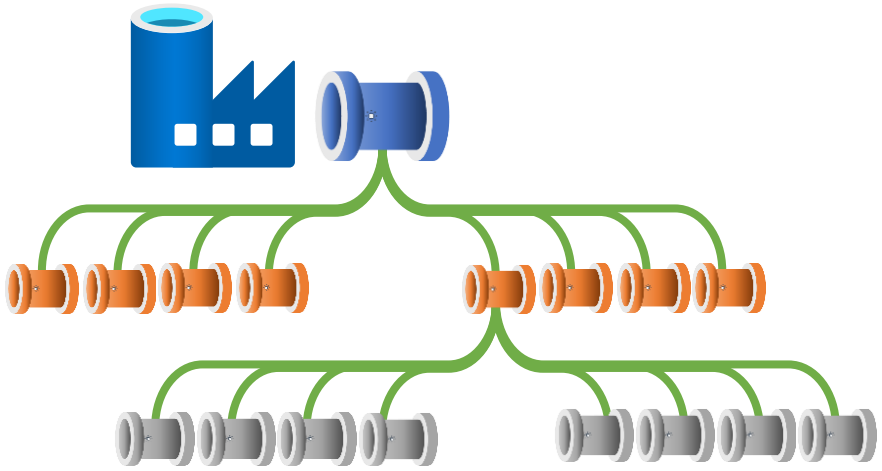
Resource Control



Resource Control - Cost

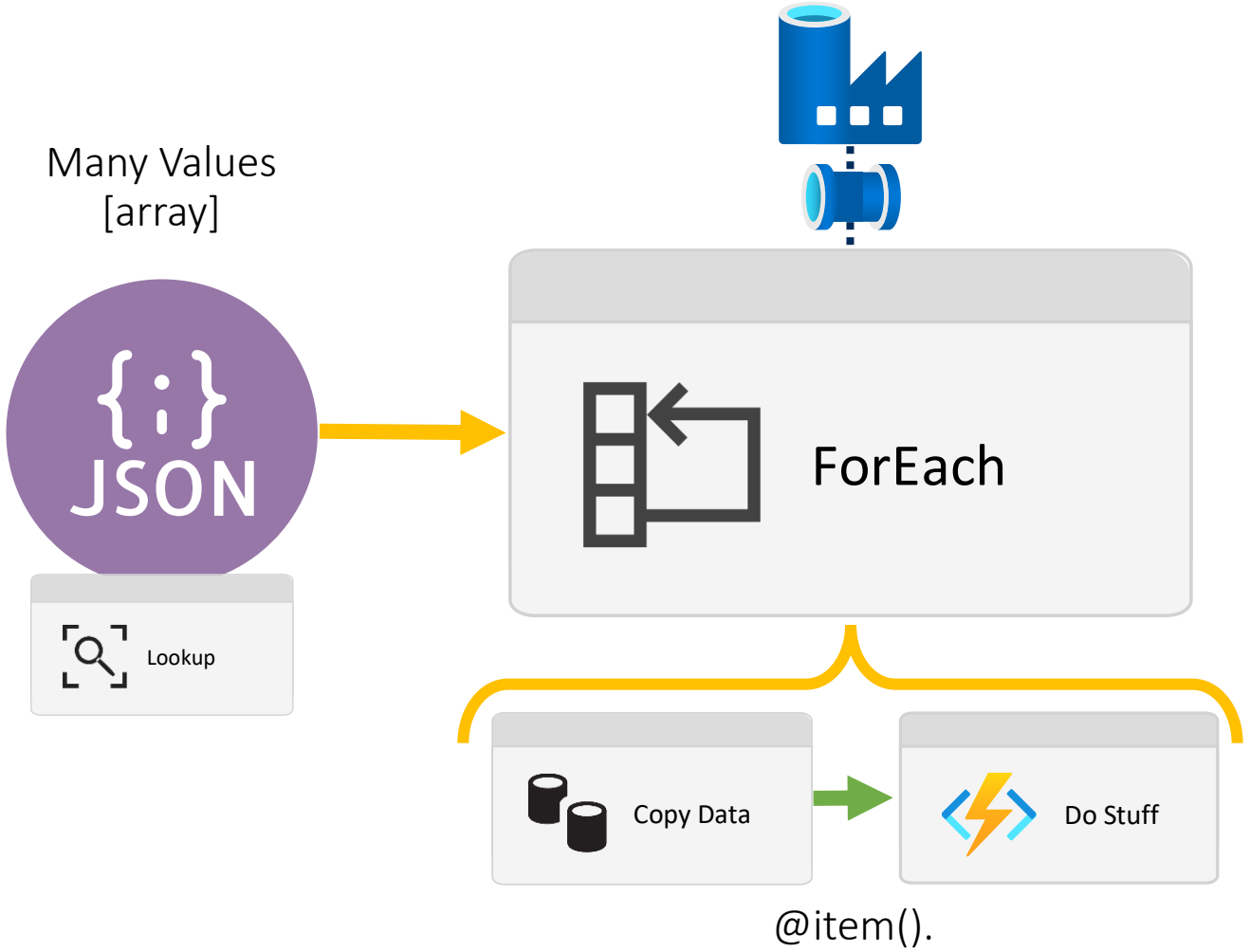


Scale Out Execution

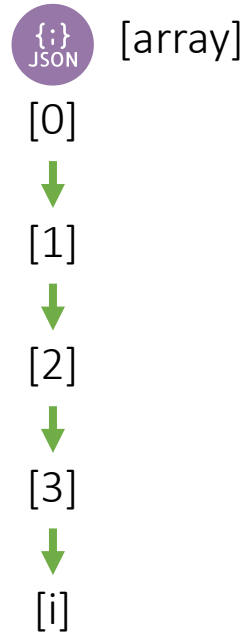


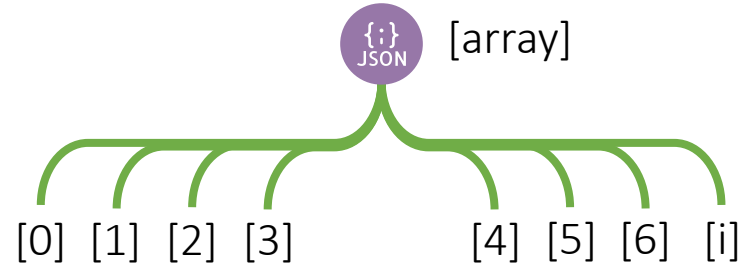
ForEach

Scaling Out Control Flow Activities



IsSequential: true



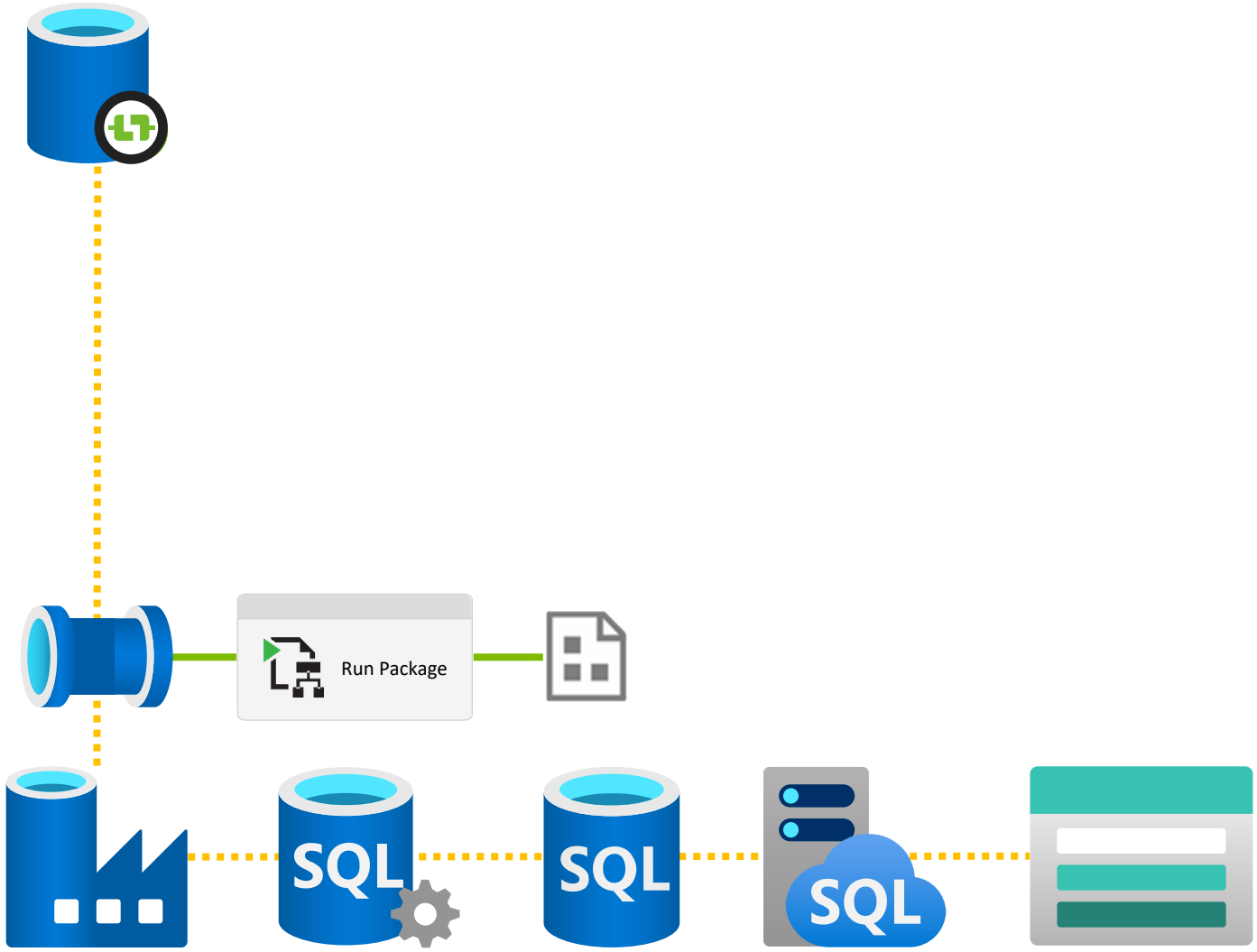


Batch Count Default: 20

Batch Count Max: 50

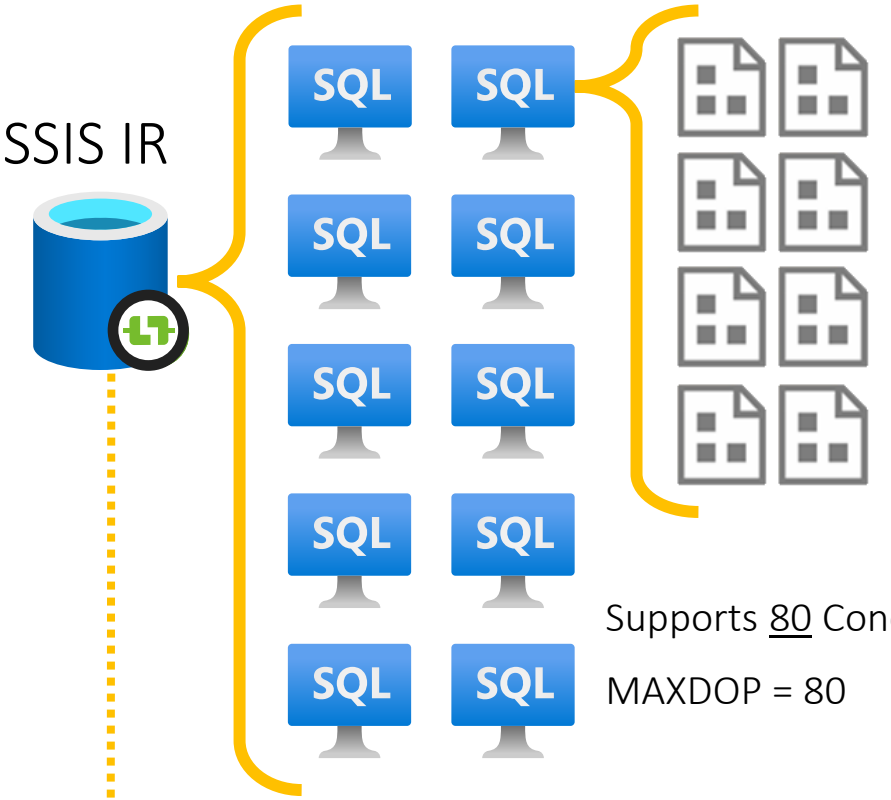
Problem: Using All Of The SSIS IR Compute

SSIS IR



Problem: Using All Of The SSIS IR Compute

SSIS IR



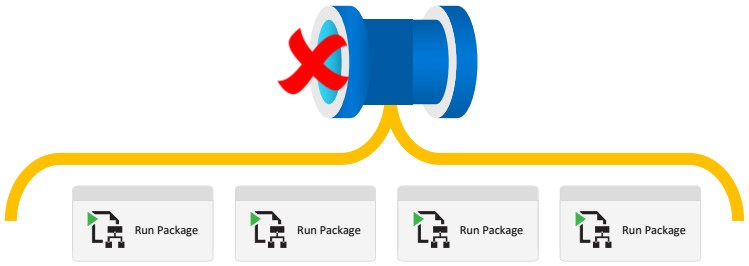
Supports 80 Concurrent Packages

MAXDOP = 80

Runs 1 Package

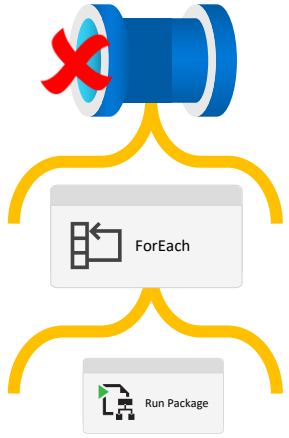
Parent Package

Child Packages x80



Pipeline x1

Activities x80

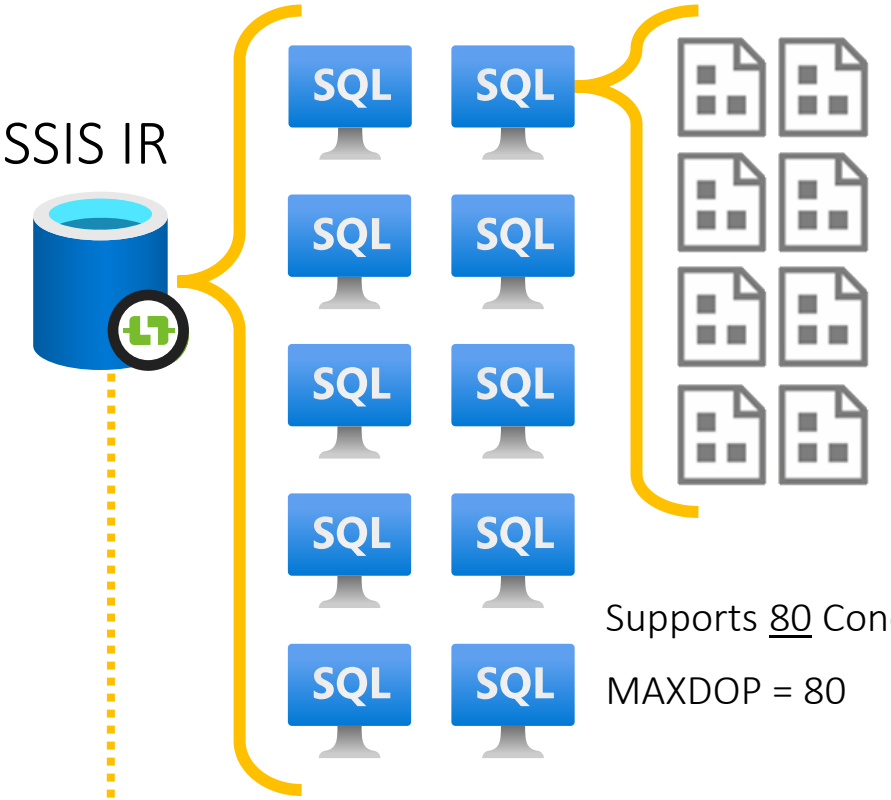


Pipeline x1

ForEach
Max Batch (50)

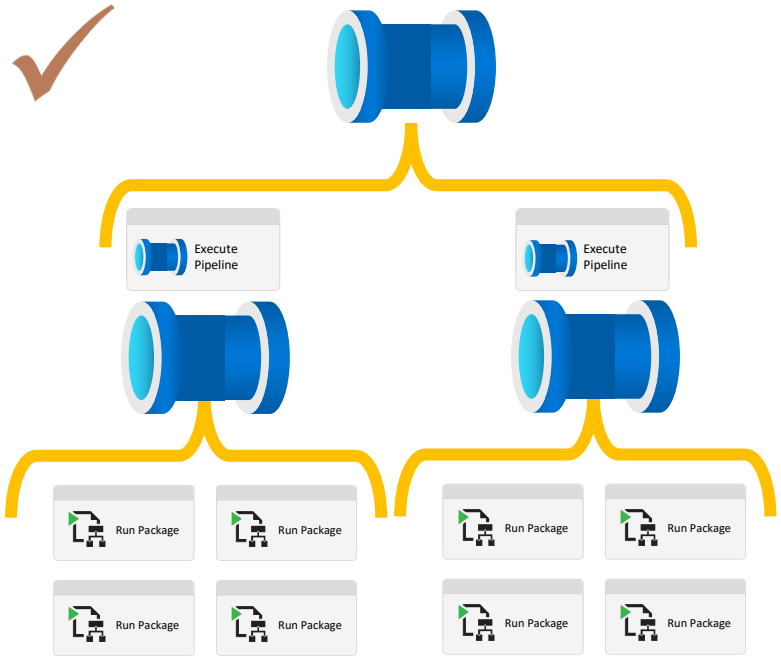
Solution 1 & 2: Static Pipelines

SSIS IR



Supports 80 Concurrent Packages

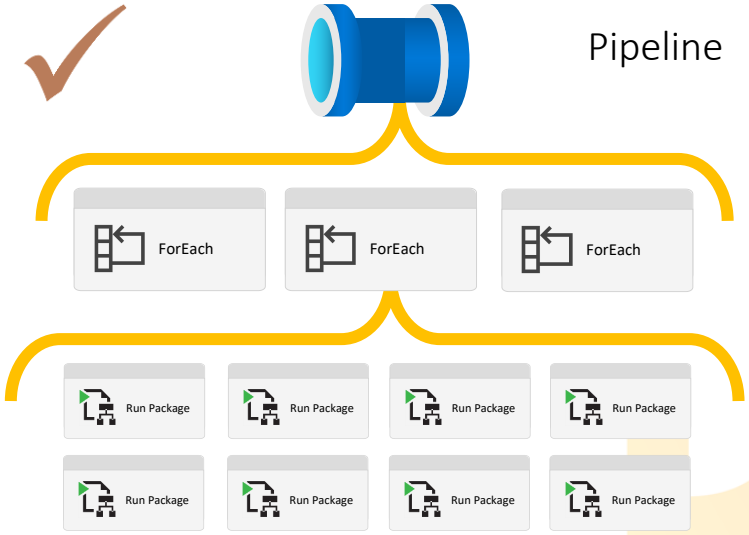
MAXDOP = 80



Parent Pipeline

Child Pipelines x2

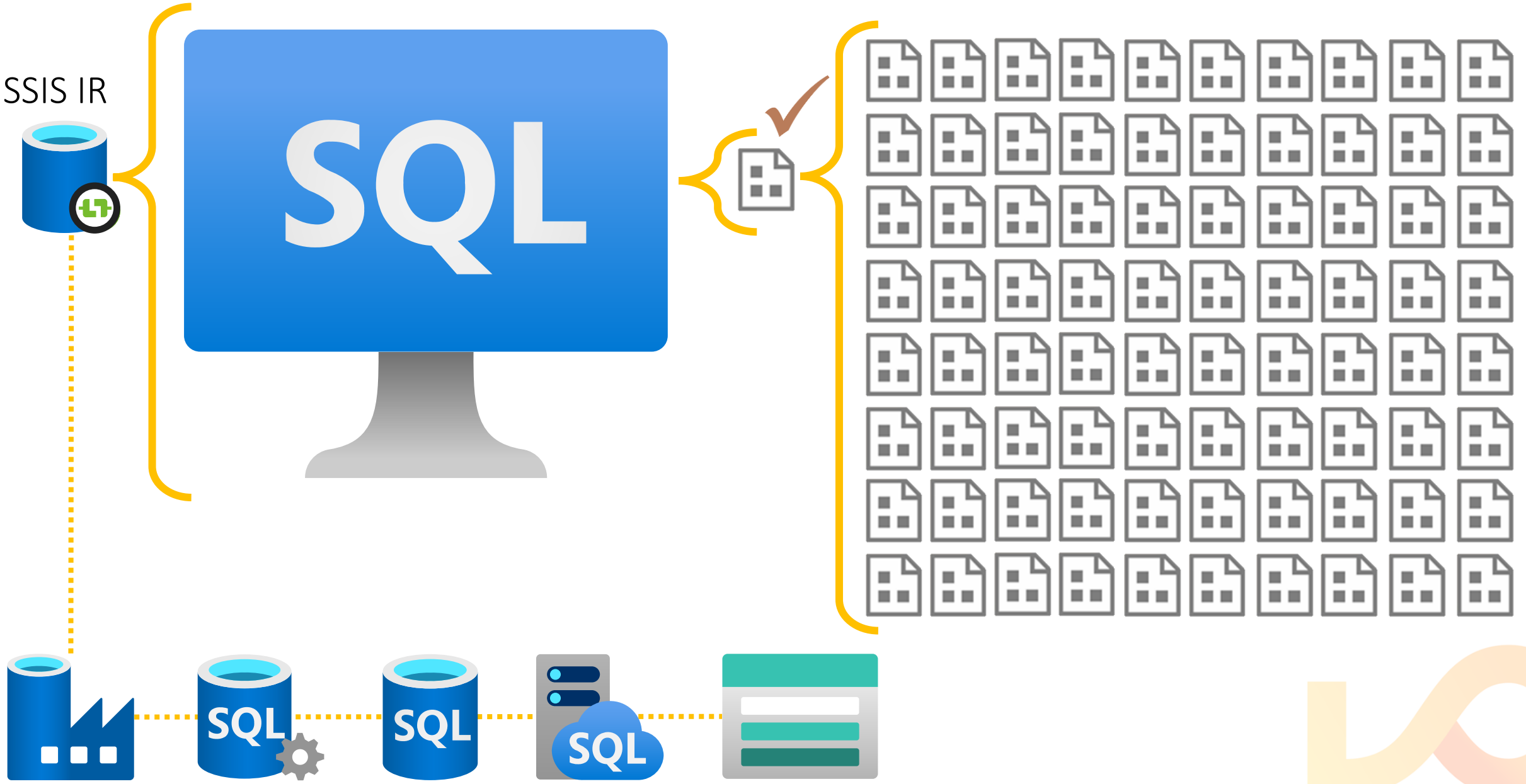
Activities x40 per Child



Pipeline

8x ForEach Batch (10)

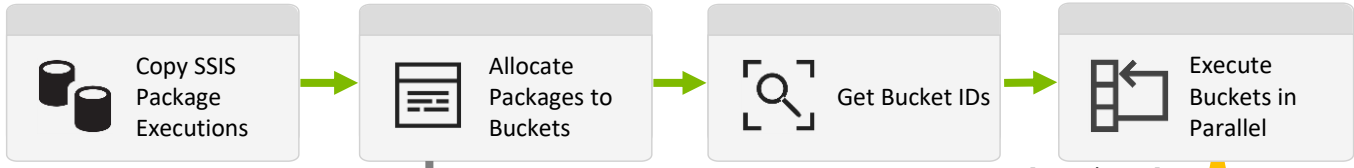
Solution 3: Packages Refactored on a Single Node IR



Solution 4: Nested ForEach Activities & Bucket Metadata

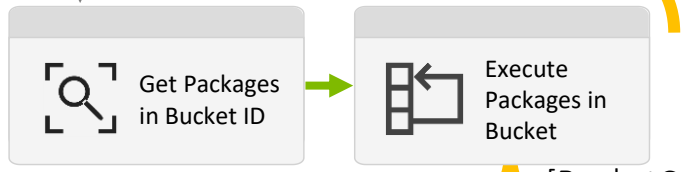
(FE L1) x (FE L2) = NEW MAXDOP
50 x 50 = 2500

SSIS IR



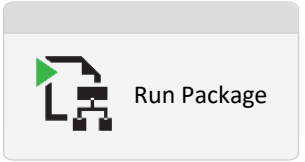
[Buckets]

FE L1
MAXDOP 50



@item().BucketId

[BucketContents]

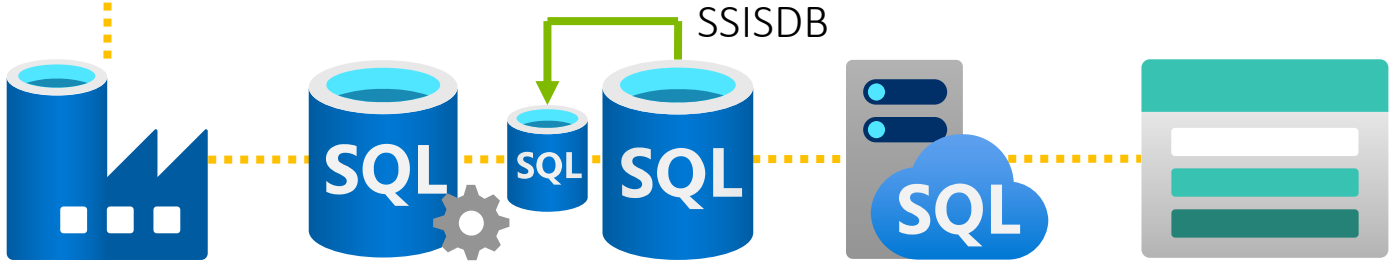


FE L2
MAXDOP 50

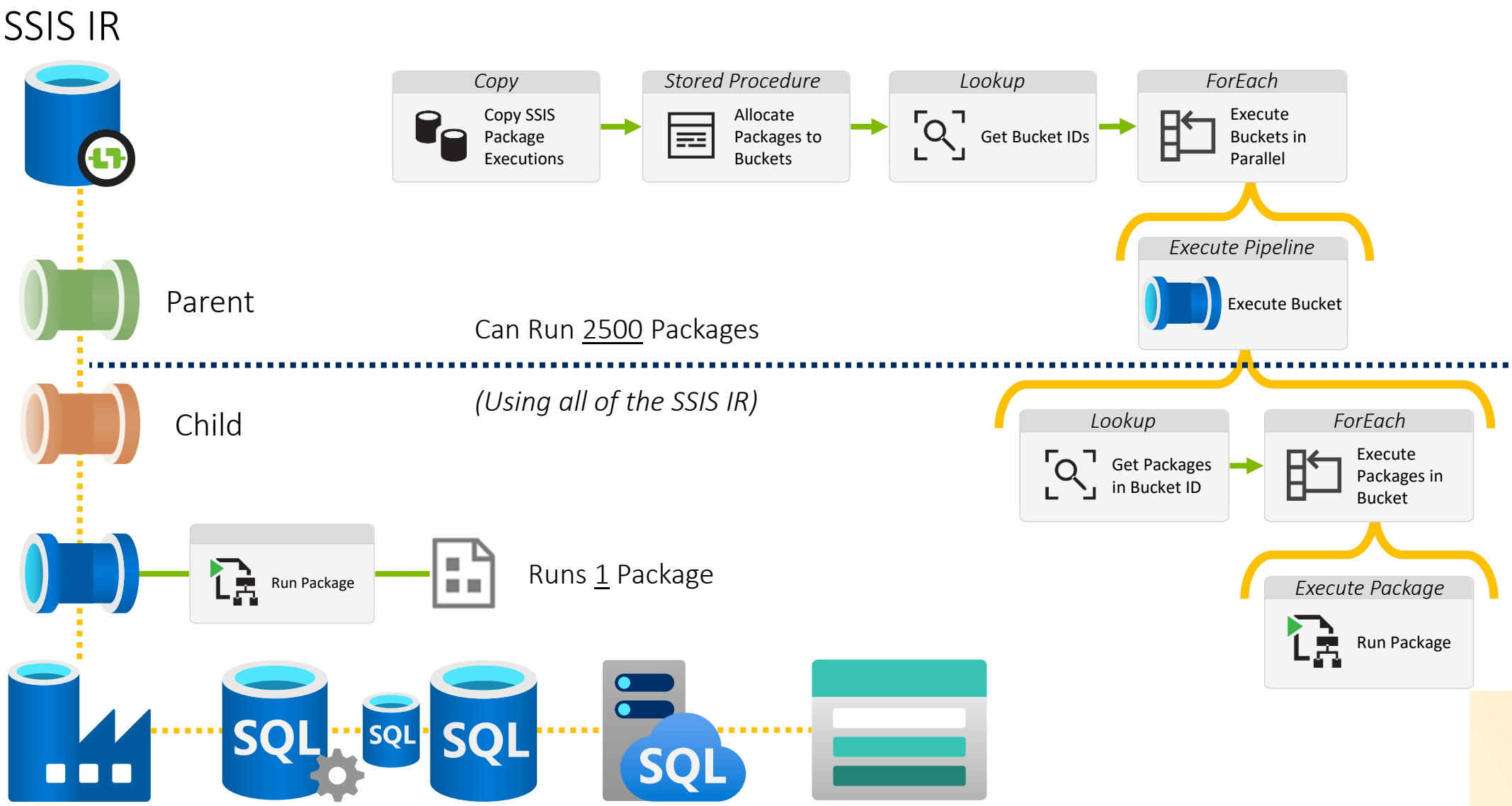
Compute size	S4	S6	S7	S9	S12
Max DTUs	200	400	800	1600	3000
Included storage (GB) [1]	250	250	250	250	250
Max storage (GB)	1024	1024	1024	1024	1024
Max in-memory OLTP storage (GB)	N/A	N/A	N/A	N/A	N/A
Max concurrent workers (requests)	400	800	1600	3200	6000
Max concurrent sessions	4800	9600	19200	30000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

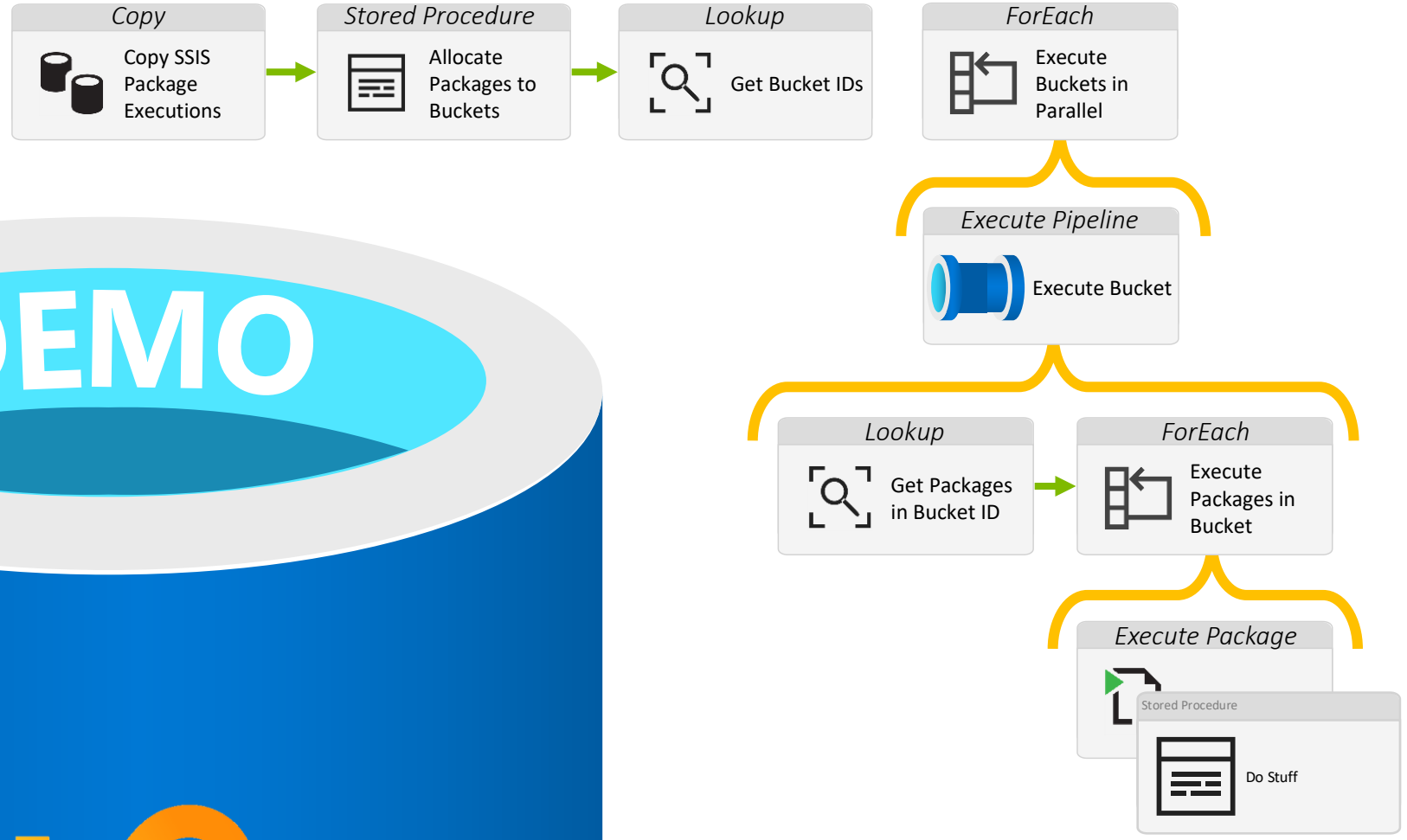
https://docs.microsoft.com/en-us/azure/azure-sql/database/resource-limits-dtu-single-databases

SSISDB



Solution 4: Nested ForEach Activities & Bucket Metadata

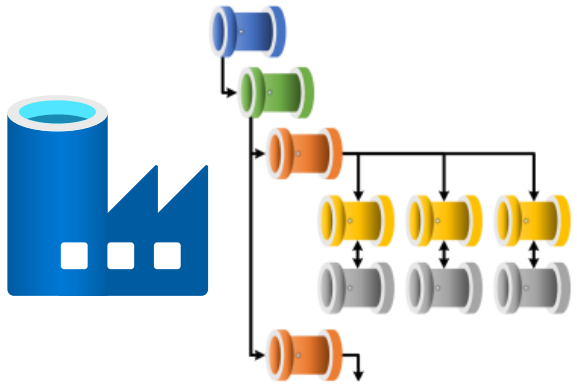




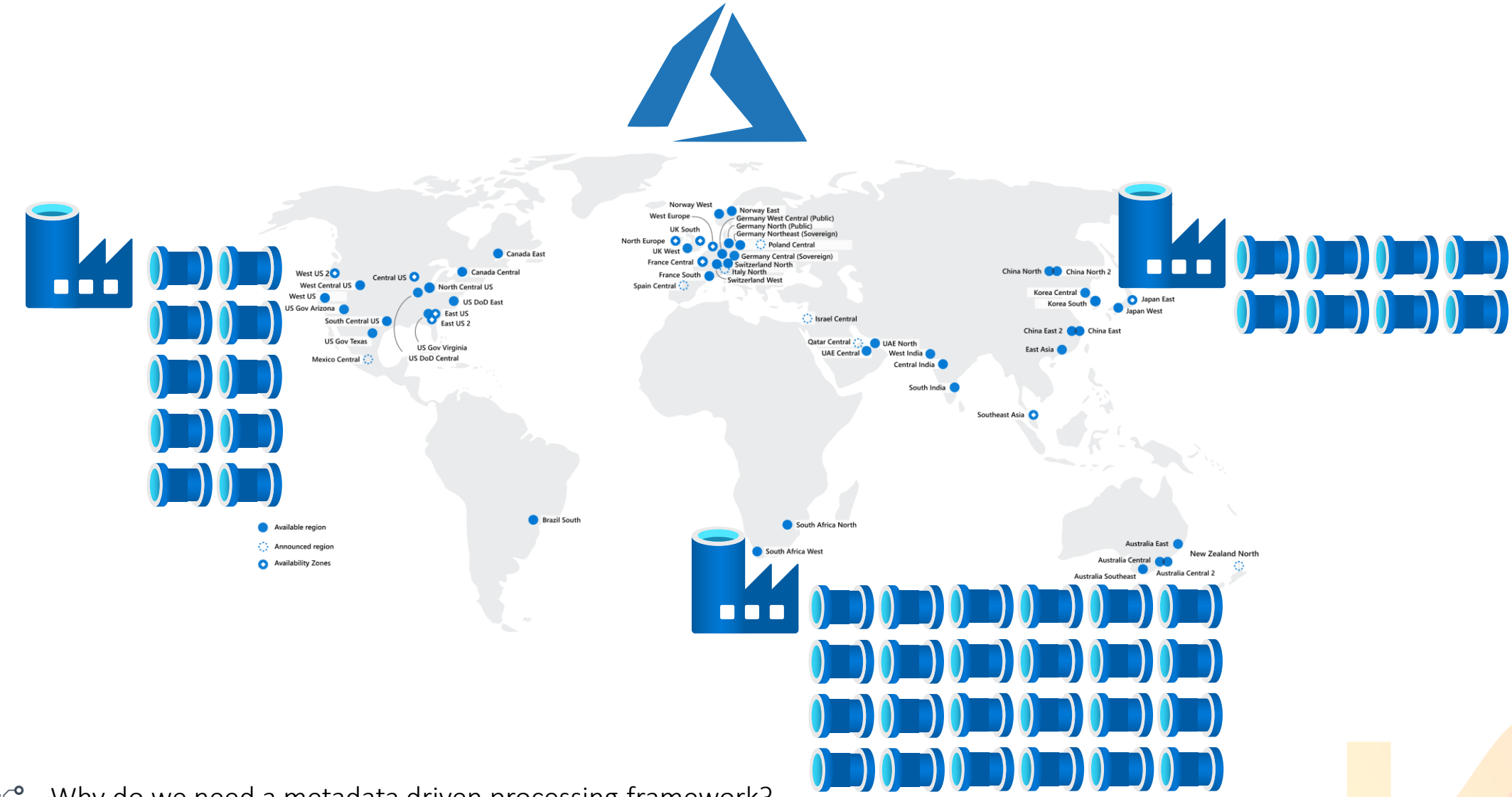
Demo Data Factory and code here:

https://github.com/mrpaulandrew/A-Day-Full-of-Azure-Data-Factory

Metadata Driven Pipelines



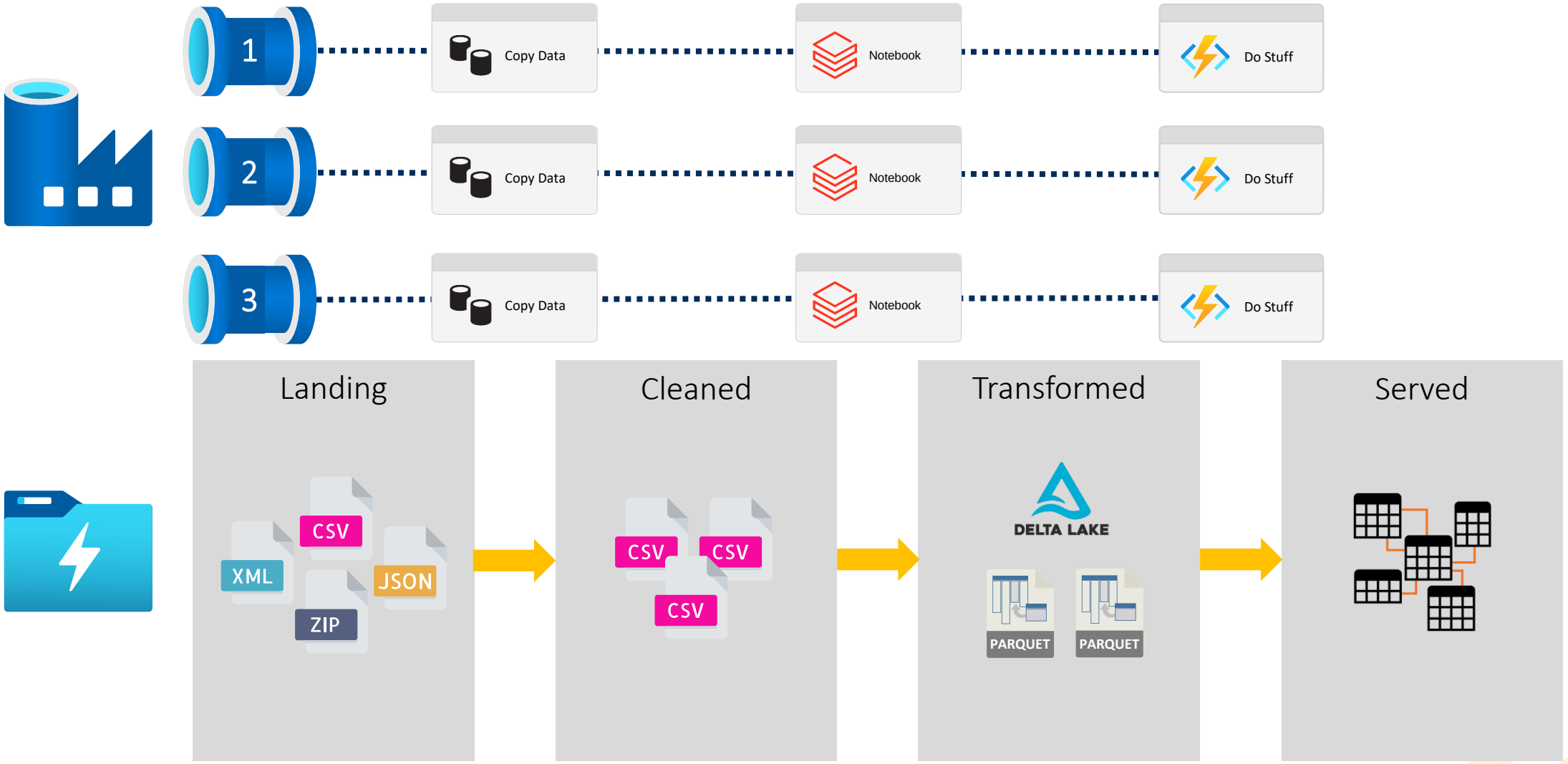
Problem: How should we structure our Data Factory Pipelines?



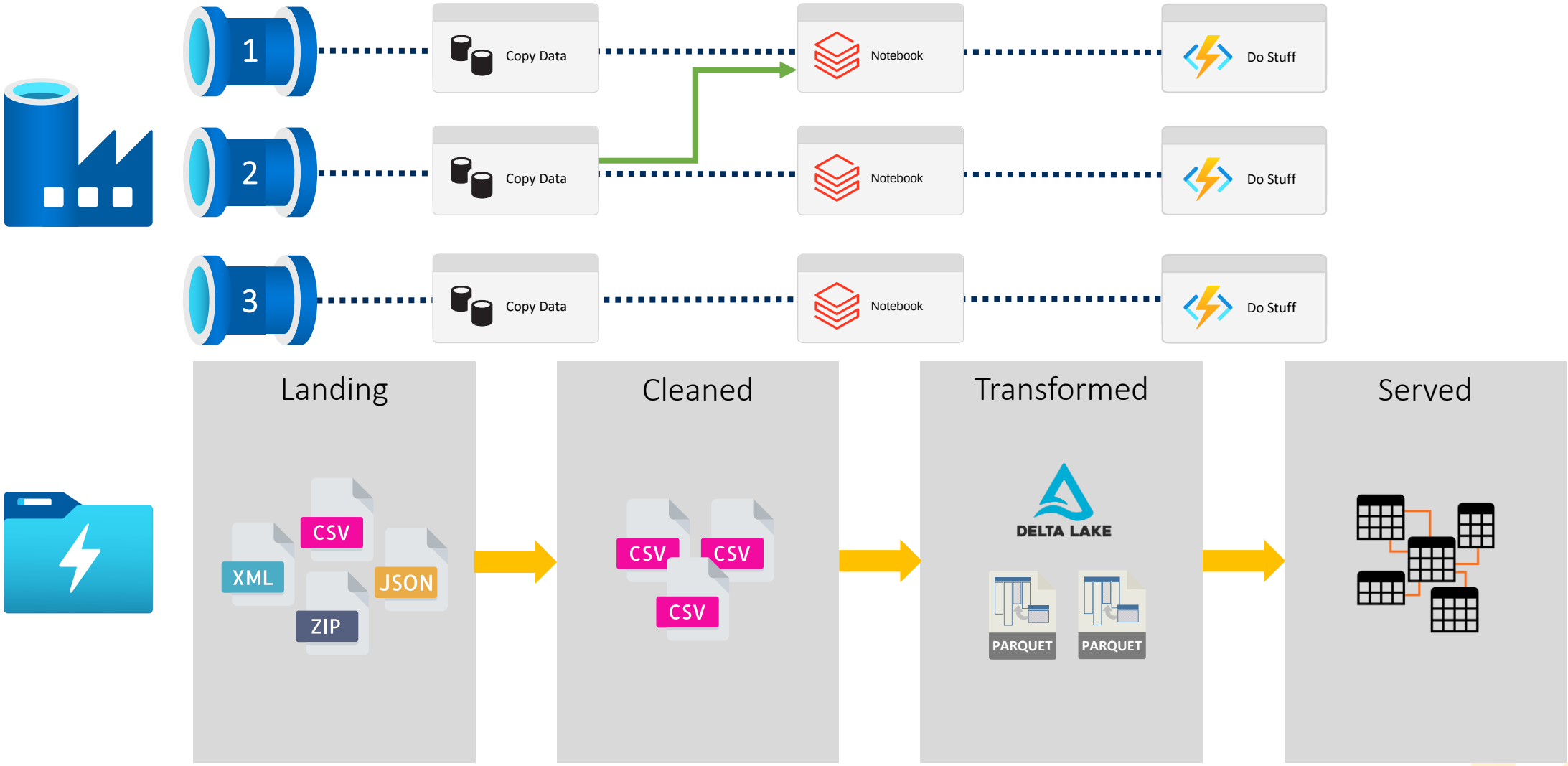
Why do we need a metadata driven processing framework?

https://youtu.be/rVlc-GBpNnc

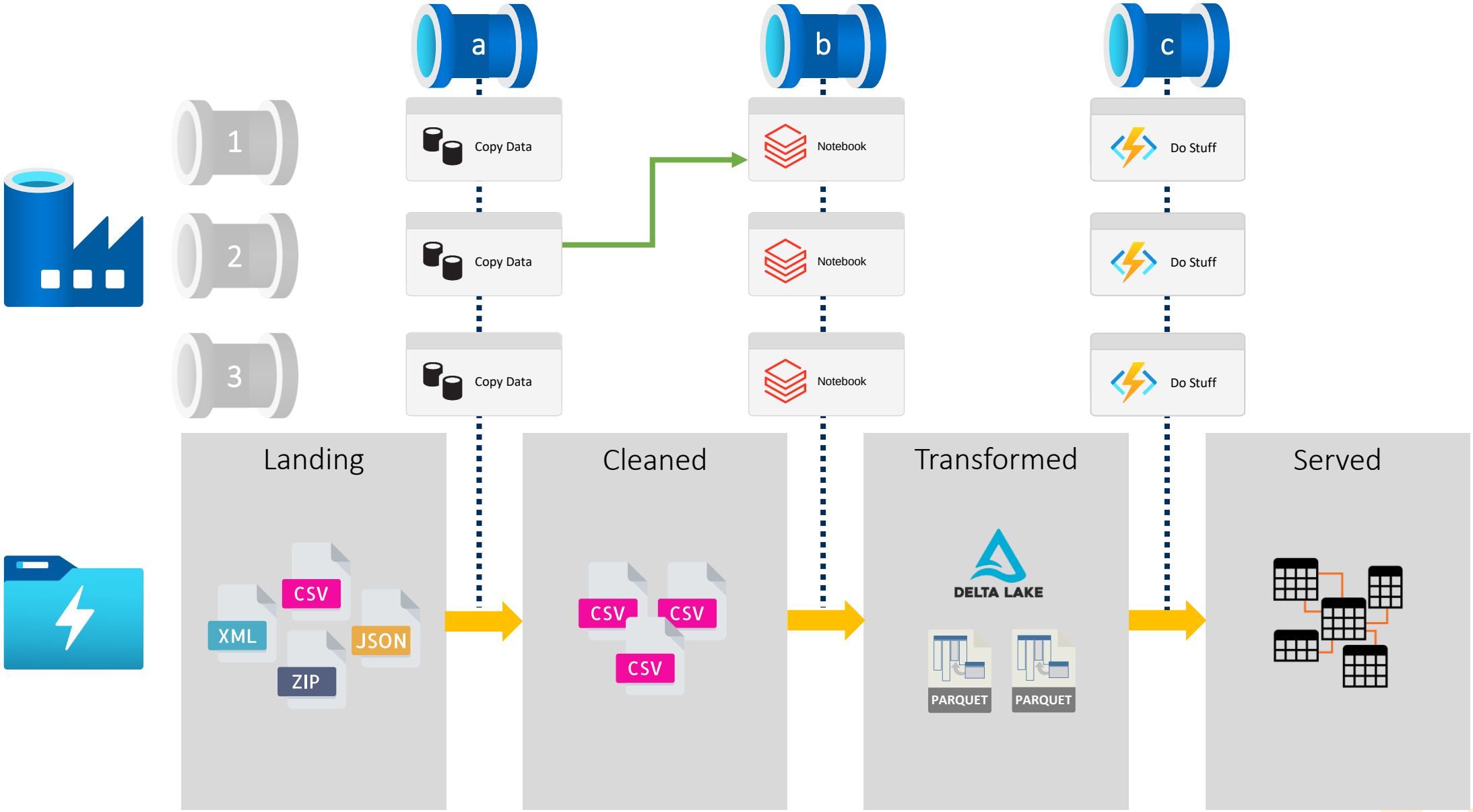
Problem


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

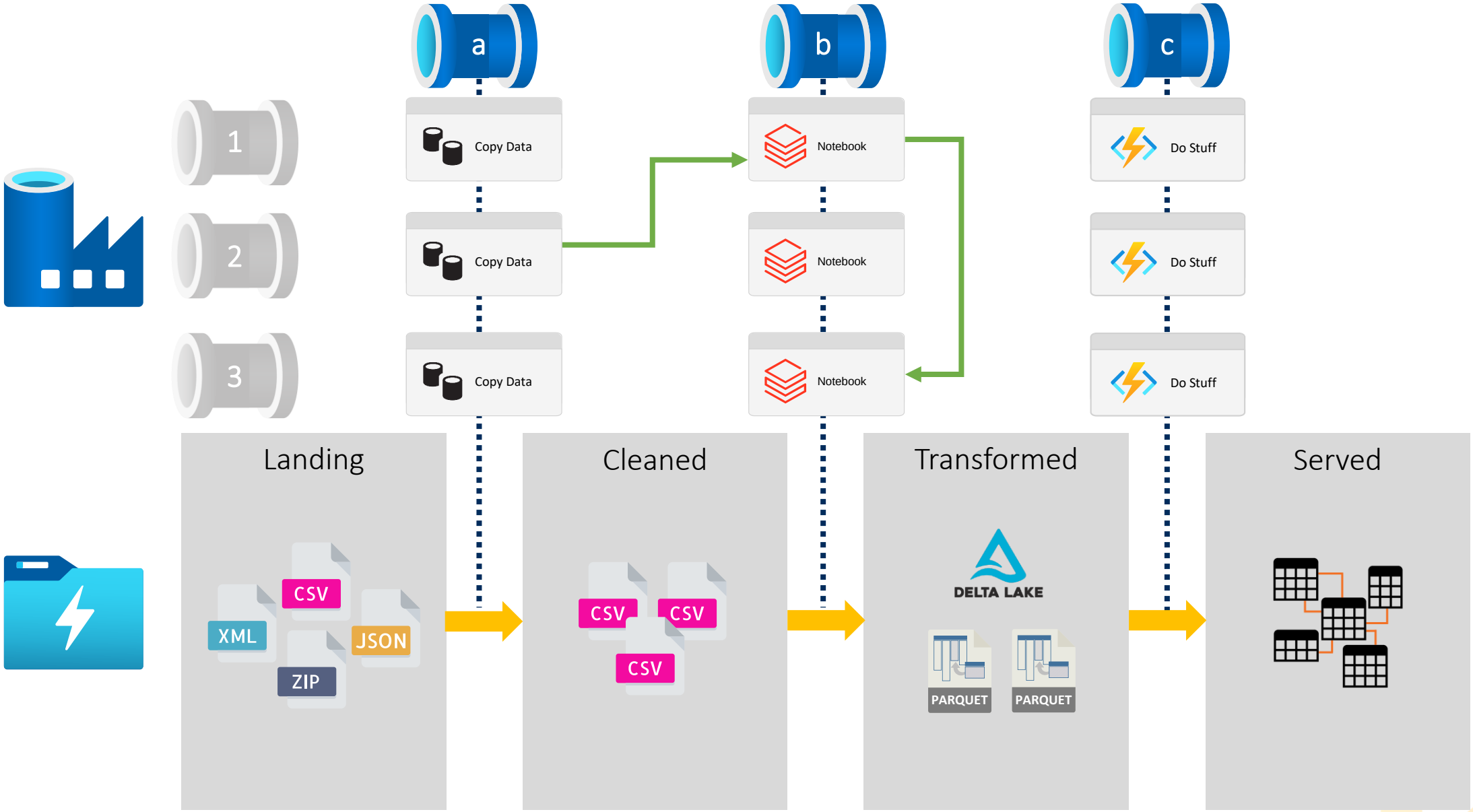
Problem


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

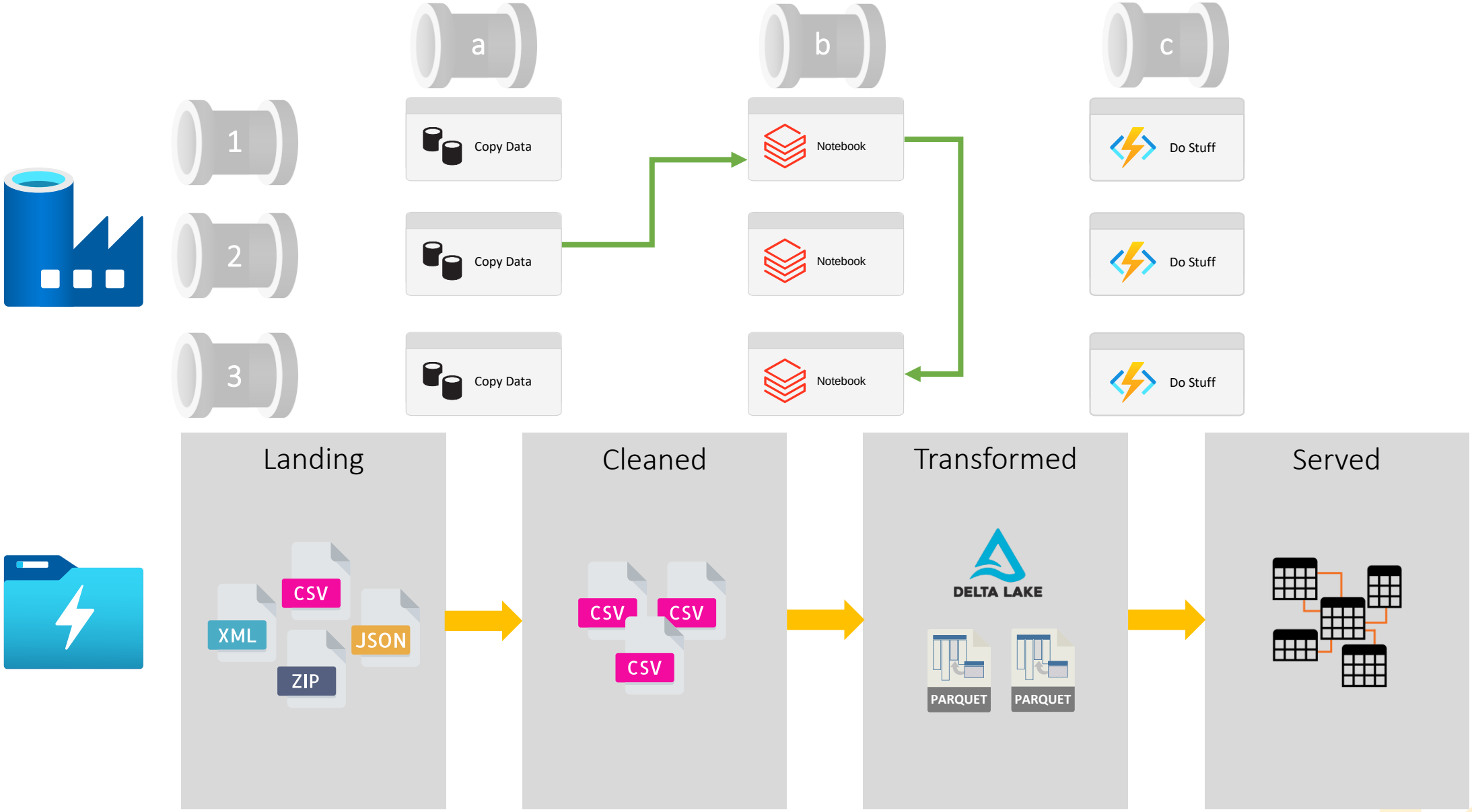
Problem


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

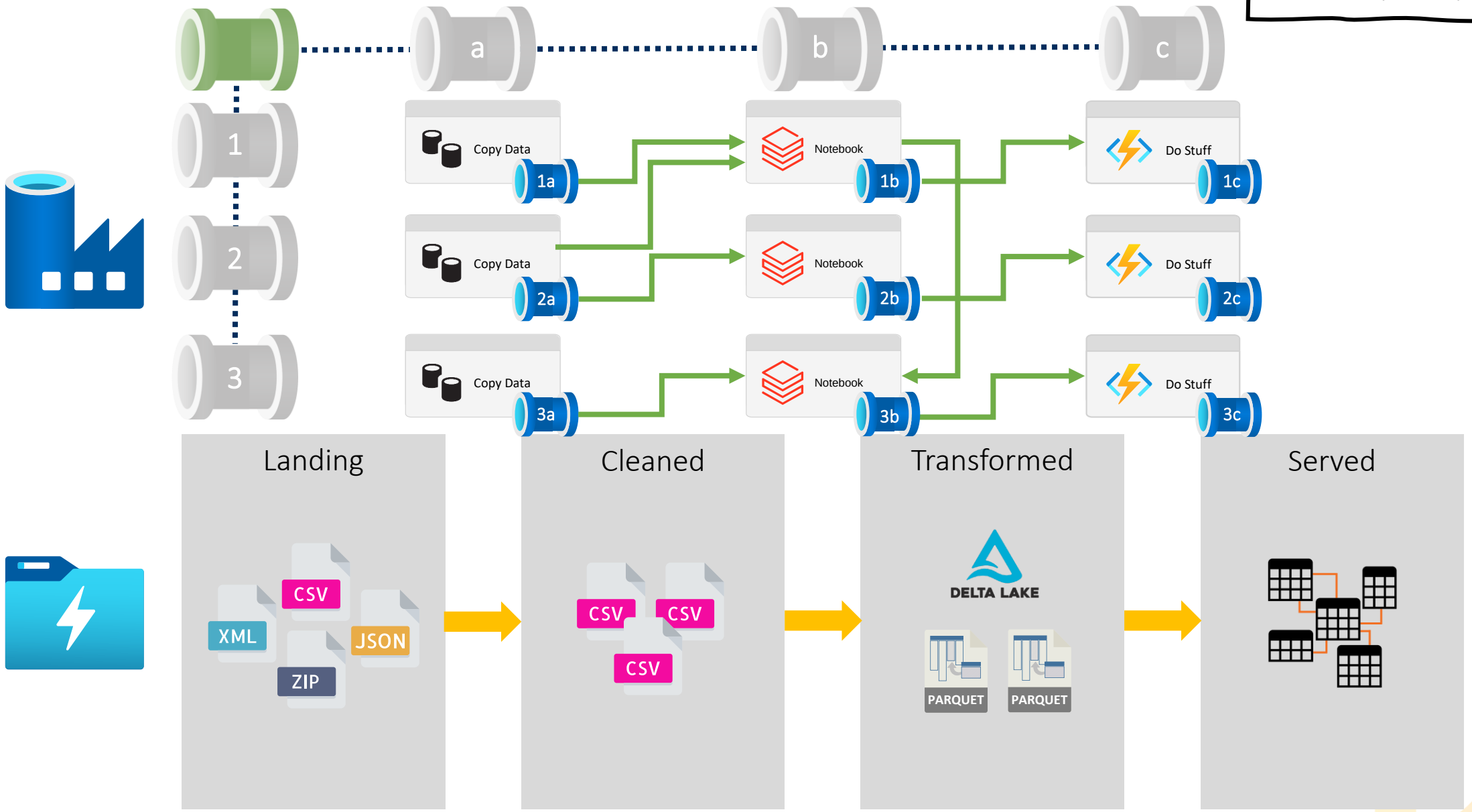

a
b
c
1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff
Landing
Cleaned
Transformed
Served
CSV
XML
JSON
ZIP
CSV
CSV
CSV
DELTA LAKE
PARQUET
PARQUET

Problem

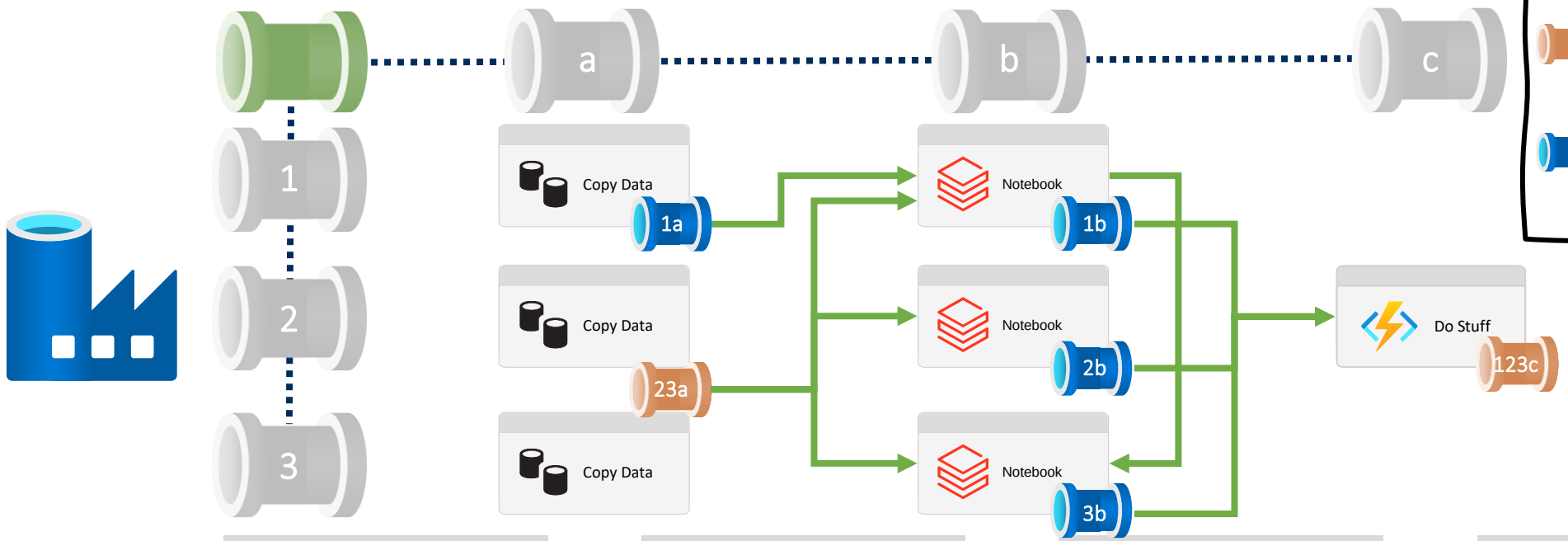

a
b
c
1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff
Landing
Cleaned
Transformed
Served
CSV
XML
JSON
ZIP
CSV
CSV
CSV
DELTA LAKE
PARQUET
PARQUET

Problem

Only 40 Activities per Pipeline.



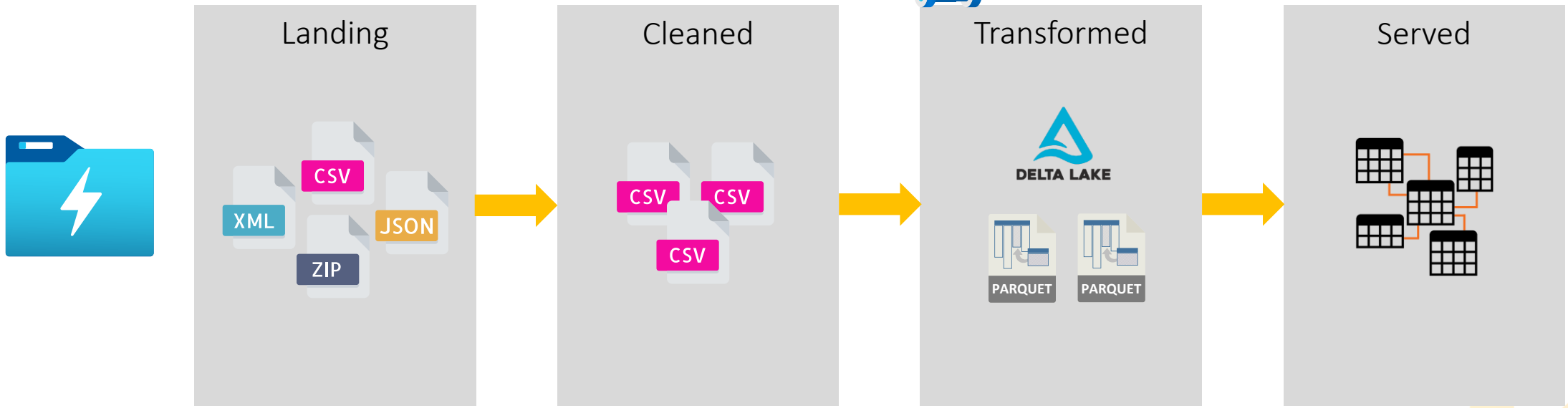
Problem



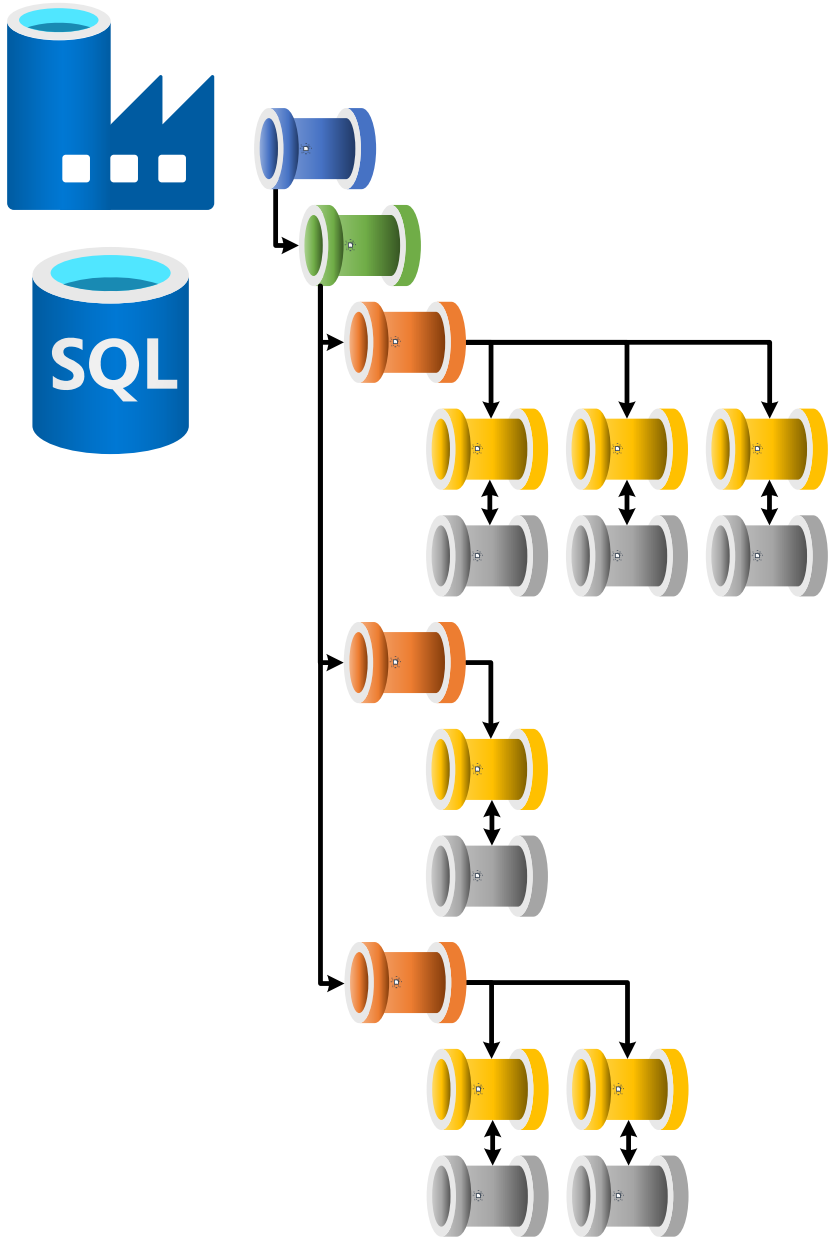
Grandparent pipeline for all processing.

Parent pipeline to consolidate work.

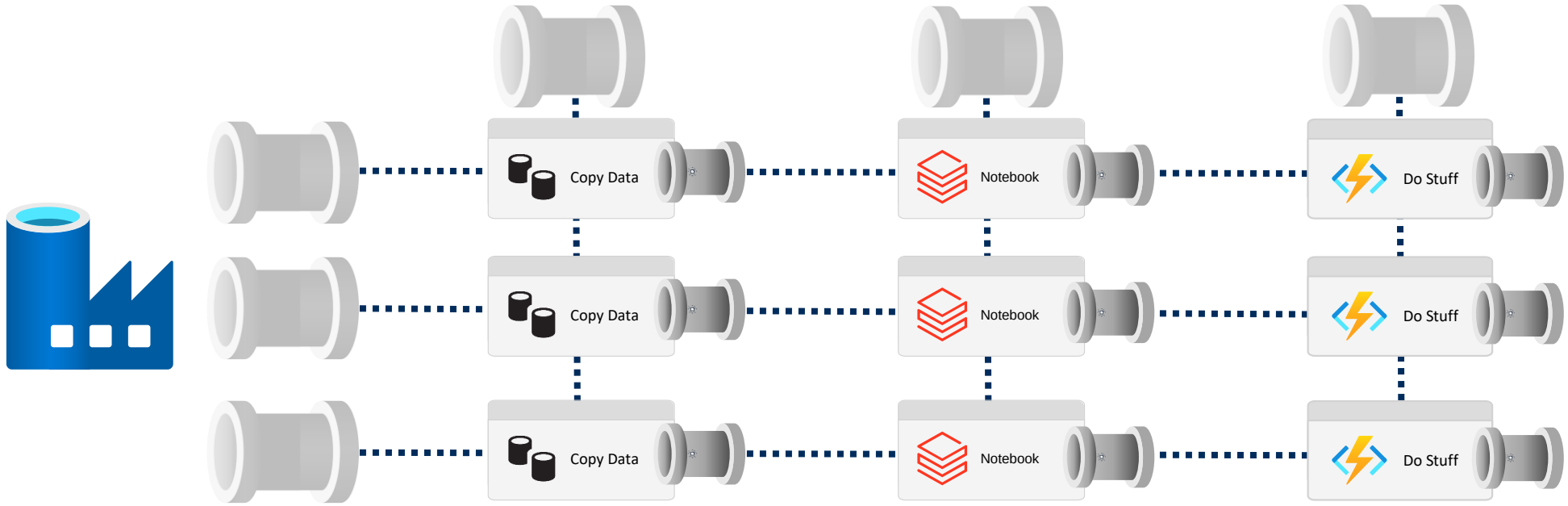
Child pipelines for low level dependencies.



Solution: Use Metadata to Drive Data Factory Pipelines



Solution

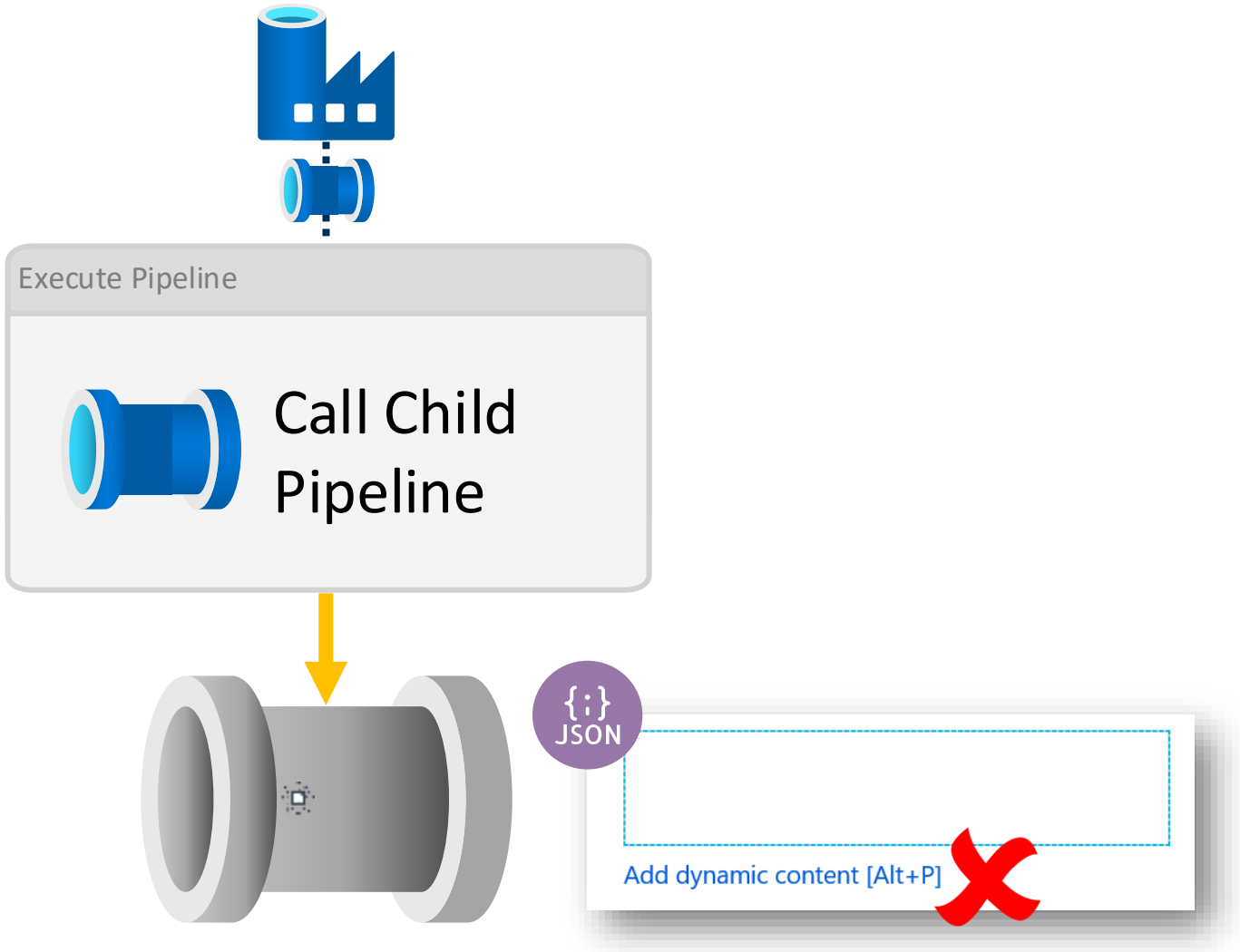


Stages
1
2
3

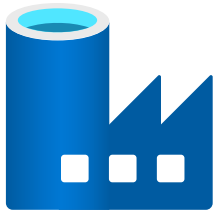
Pipelines
a
b
c
d
e
f
g
h
i

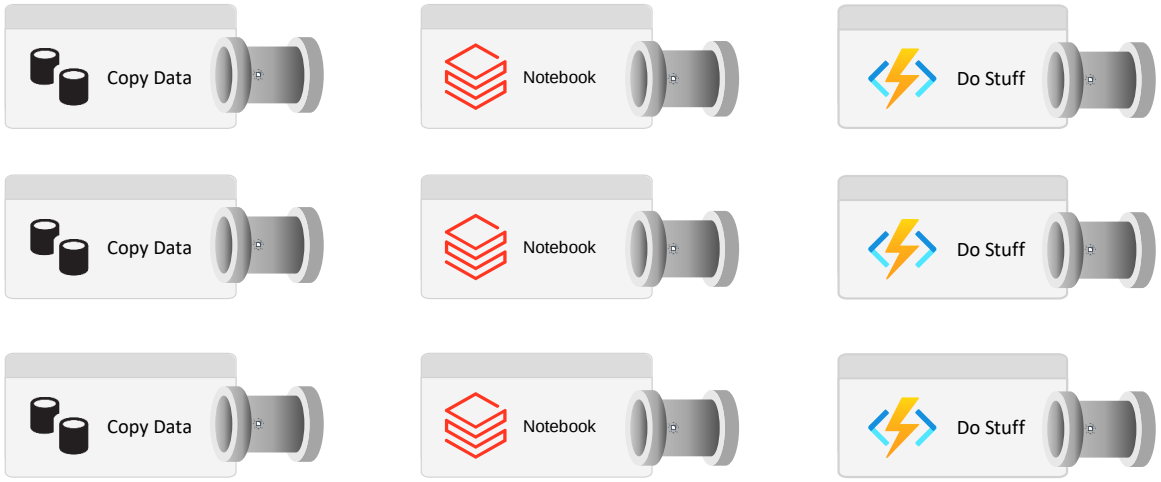
Stage	Pipeline
1	a
1	b
1	c
2	d
2	e
3	f
3	g
3	h
3	i

One More Problem



Calling Our Worker Pipelines



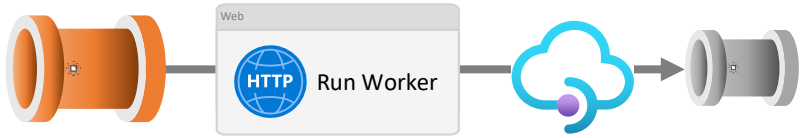


Stages
1
2
3

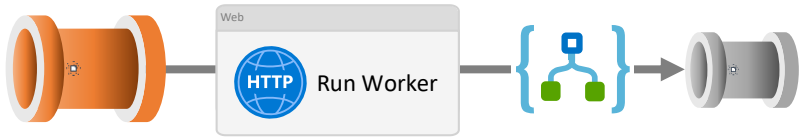
Pipelines
a
b
c
d
e
f
g
h
i

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

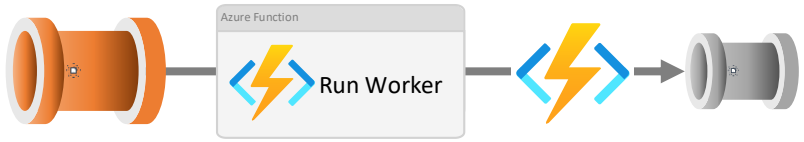
Option 1:



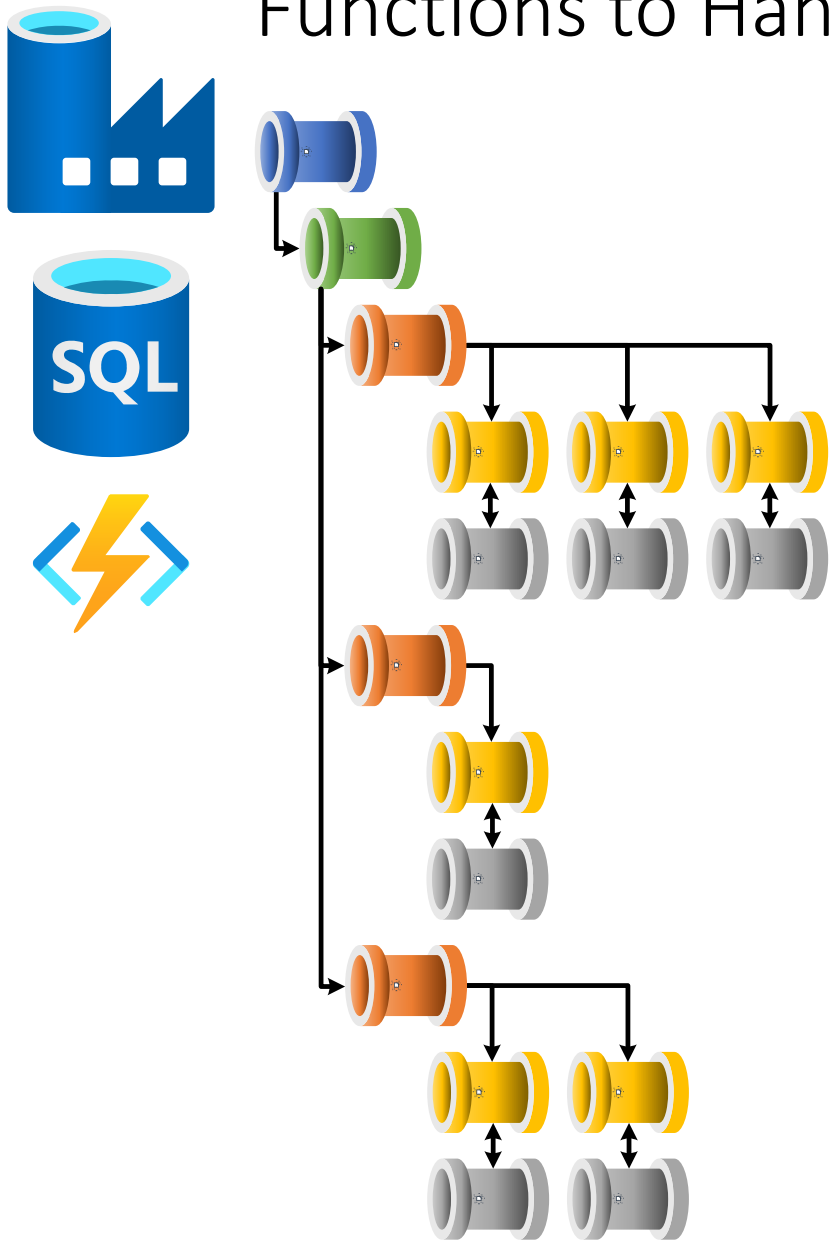
Option 2:



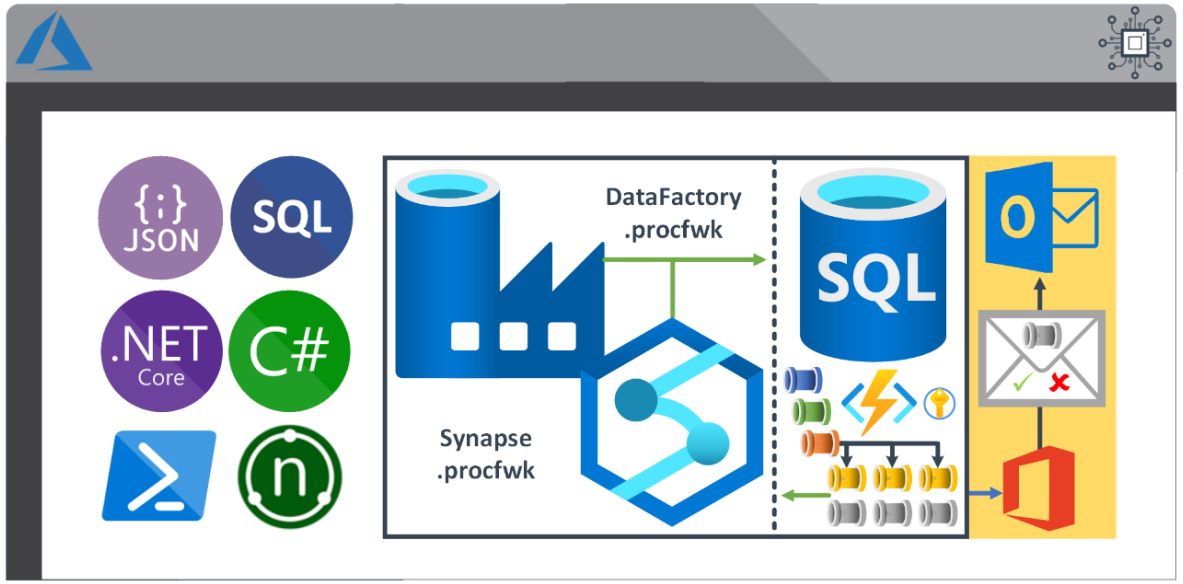
Option 3:



Solution: Use Metadata to Drive Data Factory Pipelines & Functions to Handle the Worker Pipeline Interactions

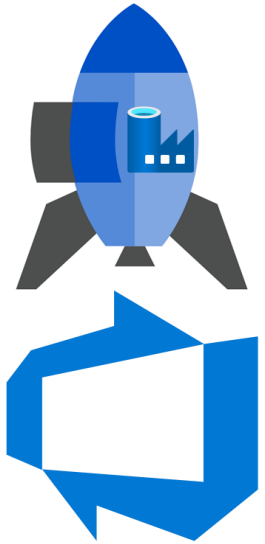


procfwk.com

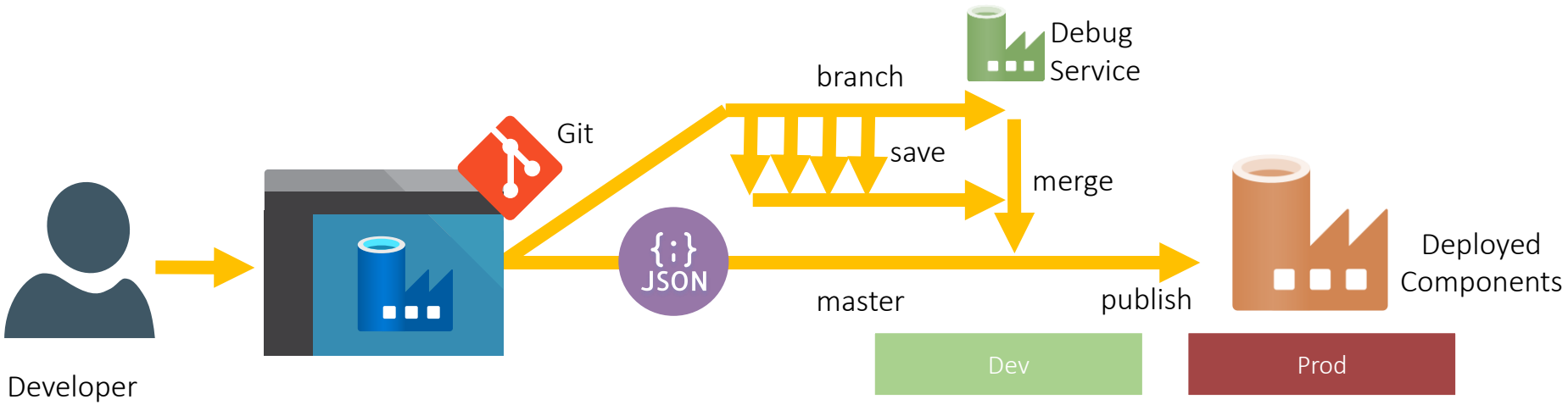


github.com/mrpaulandrew/procfwk

Deploying Data Factory

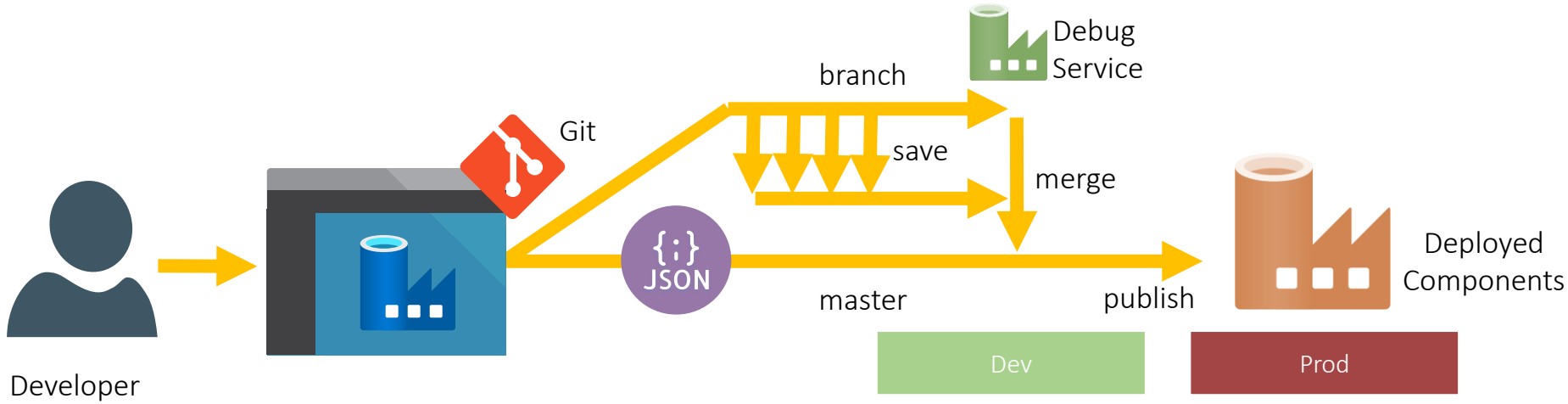


Deploying Data Factory

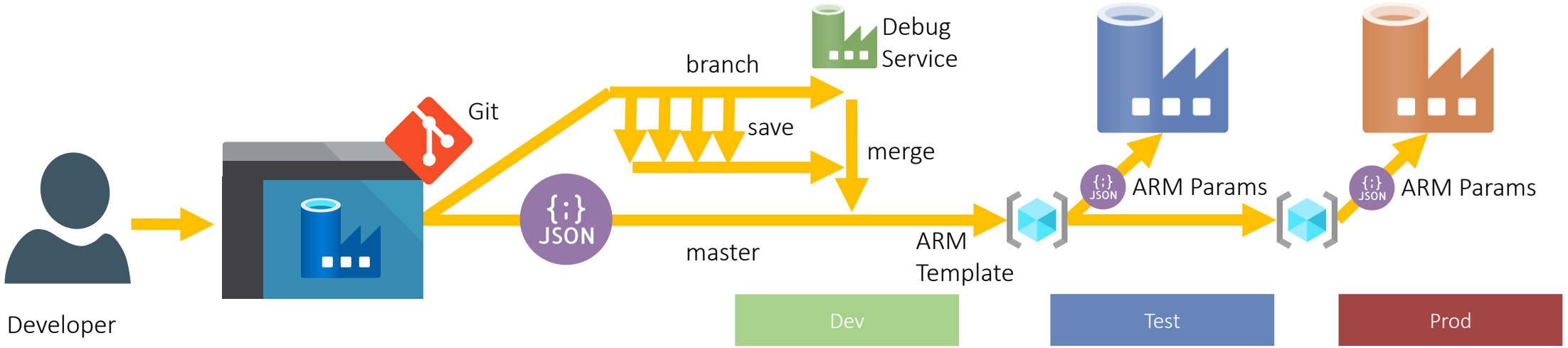


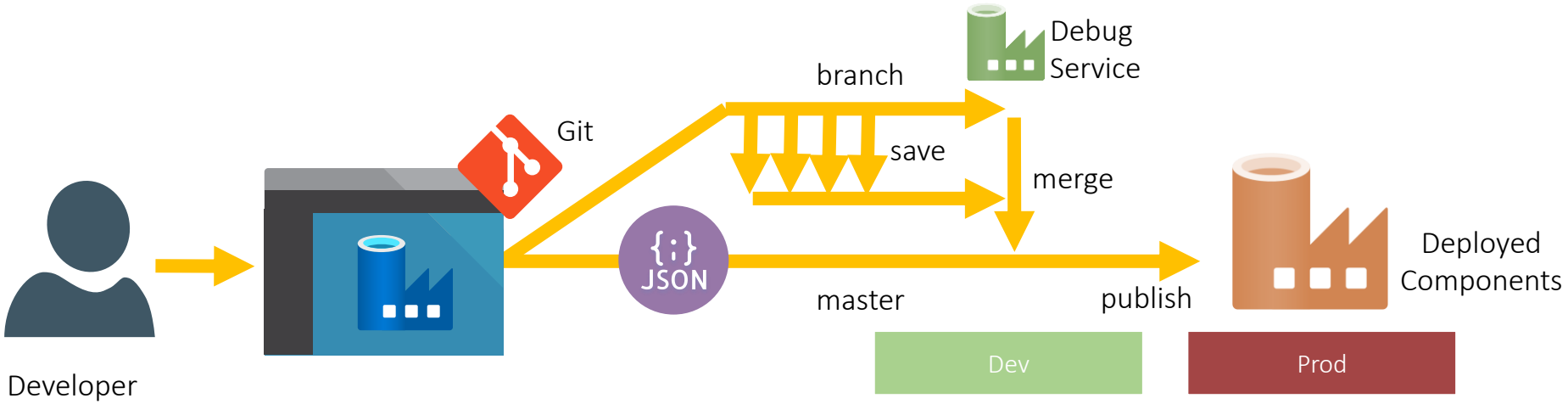
Deploying Data Factory

Option 1 – Single Data Factory Service

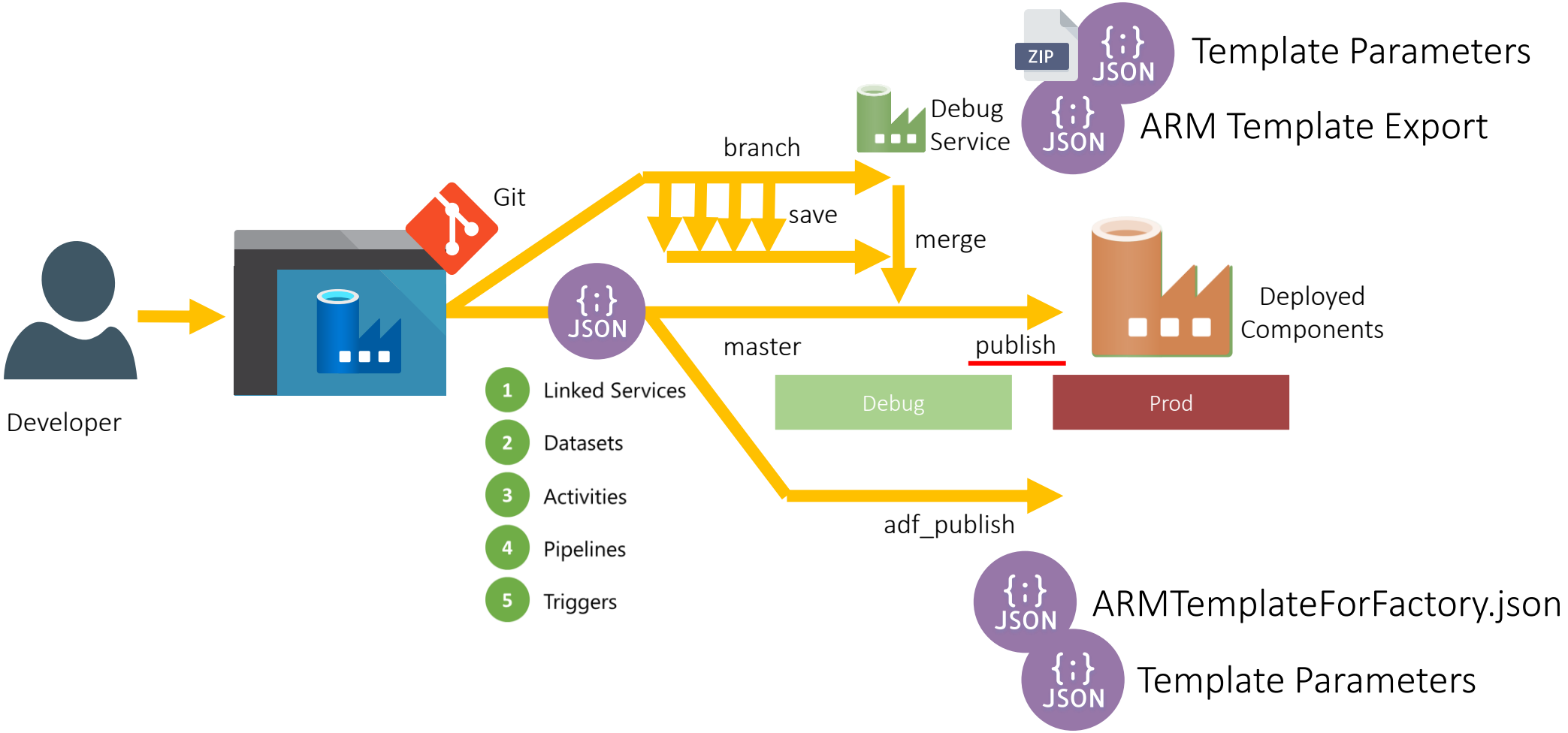


Option 2 – ARM Templates for Multiple Data Factory Services

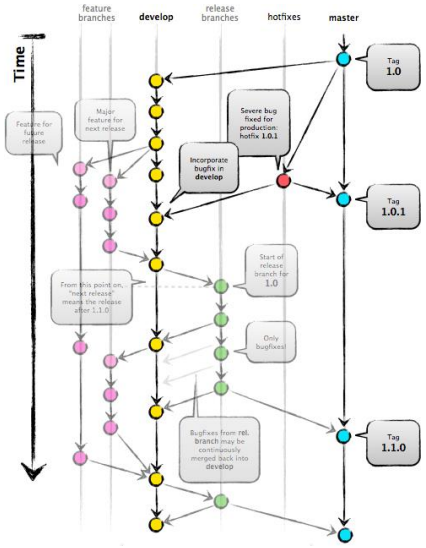



Debug
Service
branch
Git
save
merge
{;}
JSON
Deployed
Components
master
publish
Dev
Prod
Developer

Getting Our ADF Source Code



Data Factory Continuous Delivery - Option 2



Gitflow

Azure DevOps

Build Artifacts

Releases

Scoped Variables

Test

Azure Resource Group Deployment

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

ARMTemplateForFactory.json

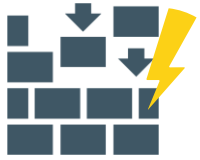
Release Approver

Production

Azure Resource Group Deployment

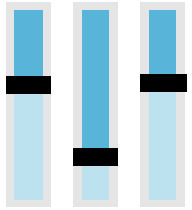
Data Factory Continuous Delivery - Option 2

Azure DevOps



Build Artifacts

Releases



Scoped Variables

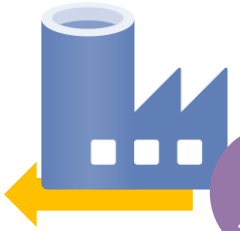
Test

Azure Resource Group Deployment

- Override Template Parameters

  -factoryName "$(DataFactory.Name)"

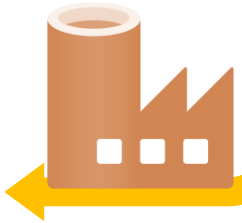
Release Approver



ARMTemplateForFactory.json

Production

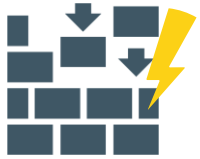
Azure Resource Group Deployment



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

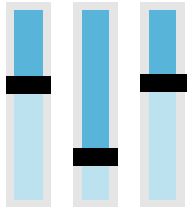
Data Factory Continuous Delivery - Option 2

Azure DevOps



Build Artifacts

Releases



Scoped Variables

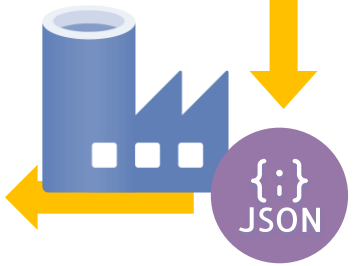
Test

Release Approver

Azure Resource Group Deployment

- Override Template Parameters

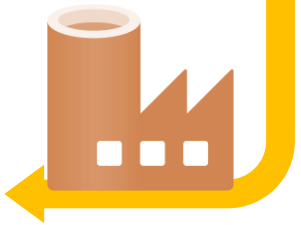
  -factoryName "$(DataFactory.Name)"



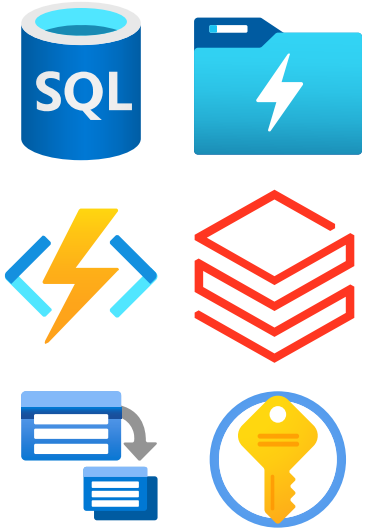
ARMTemplateForFactory.json

Production

Azure Resource Group Deployment

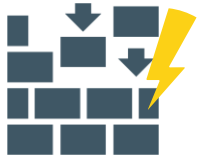


1 Linked Services



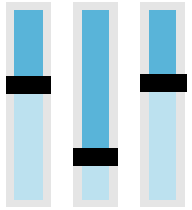
Data Factory Continuous Delivery - Option 3

Azure DevOps



Build Artifacts

Releases



Scoped Variables

1 Linked Services

2 Datasets

3 Activities

4 Pipelines

5 Triggers

Test

Run PowerShell

params.json

Azure Resource Group Deployment

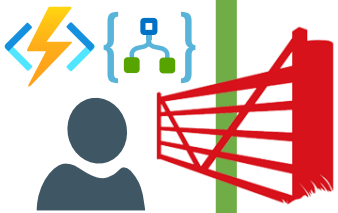
Override Template Parameters

-factoryName "$(DataFactory.Name)"

Set-AzDataFactoryV2Pipeline

Set-AzDataFactoryV2Dataset

Set-AzDataFactoryV2LinkedService



Release Approver

Production

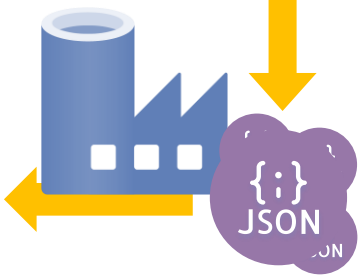
Run PowerShell

params.json

Azure Resource Group Deployment

Set-AzDataFactoryV2IntegrationRuntime

Set-AzDataFactoryV2Trigger



linkedservices.json

pipelines & activites.json

datasets.json

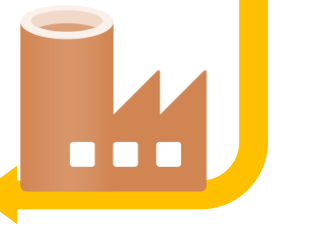
triggers.json

ARMTemplateForFactory.json

and_par

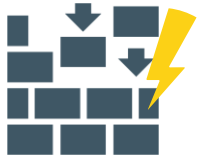
adf_pu

publish



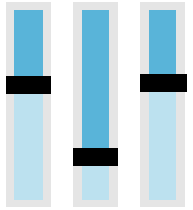
Data Factory Continuous Delivery - Option 3

Azure DevOps



Build Artifacts

Releases



Scoped Variables

Test

Run PowerShell

params.json

Set-AzDataFactoryV2Pipeline

Set-AzDataFactoryV2Dataset

Set-AzDataFactoryV2LinkedService

Release Approver

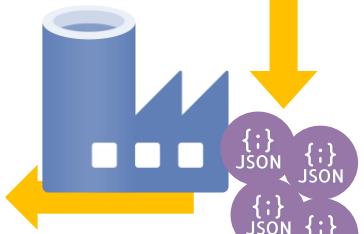
Production

Run PowerShell

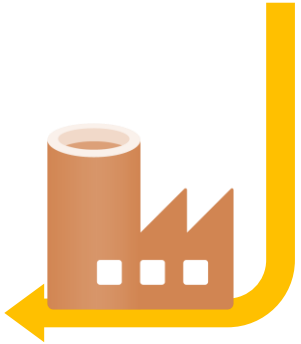
params.json

Set-AzDataFactoryV2IntegrationRuntime

Set-AzDataFactoryV2Trigger



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



1 Linked Services

2 Datasets

3 Activities

4 Pipelines

5 Triggers

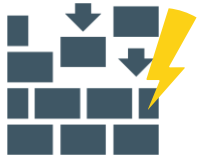
Caveats

1) Handle own dependencies.

2) Handle own removals.

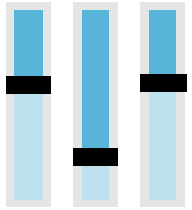
Data Factory Continuous Delivery - Option 4

Azure DevOps



Build Artifacts

Releases



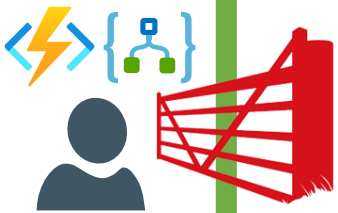
Scoped Variables

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

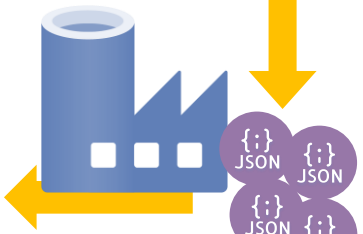
Test

Run PowerShell

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



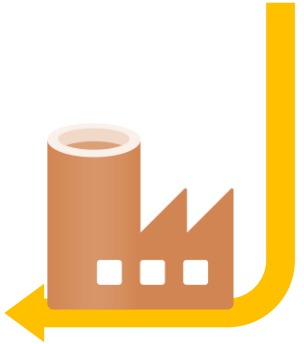
Release Approver



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

Production

Publish Azure Data Factory



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

Deployment Options Summary

Option 1 – Use a single Data Factory service.

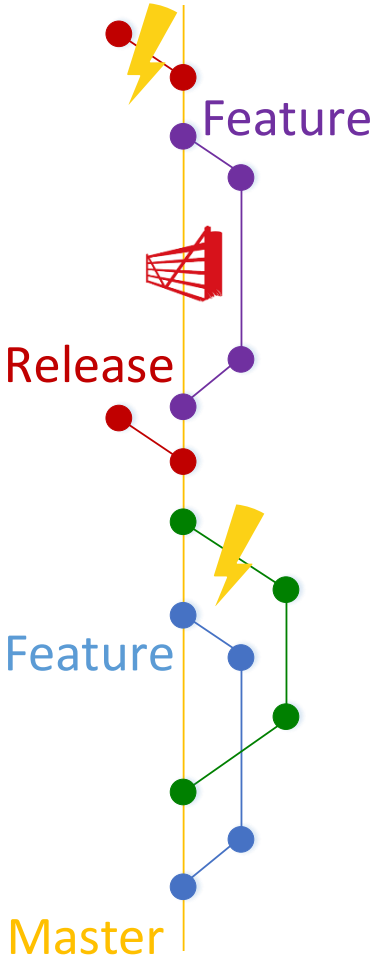
Option 2 – ARM Templates for multiple Data Factory services (environments).

Option 3 – Use PowerShell cmdlets for each Data Factory JSON artifact.

Option 4 – Use a PowerShell module or custom Azure DevOps task.

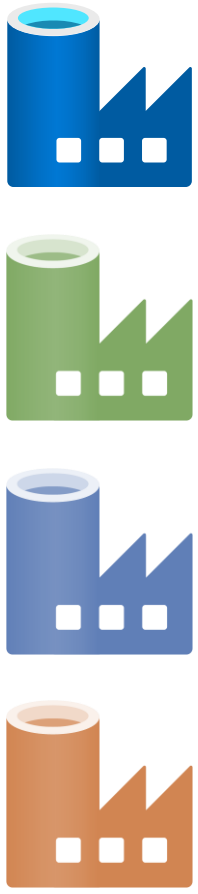
Data Factory DevOps Story Summary

What is your code branching strategy?



Which source control tool to use?

How many environments do we want?



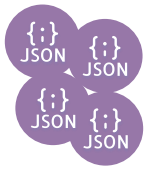
What deployment method do we want to use?

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

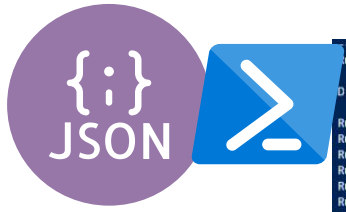
Best Practices

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

Key Points

- 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 Templates
- Documentation

https://mrpaulandrew.com/2020/11/09/best-practices-for-implementing-azure-data-factory-auto-checker-script-v0-1/




```
Running checks for Data Factory ARM template:

D:\Stuff\arm_template2.json

Running check...  Pipeline(s) without any triggers attached. Directly or indirectly.
Running check...  Pipeline(s) with an impossible AND/OR activity execution chain.
Running check...  Pipeline(s) without a description value.
Running check...  Pipeline(s) not organised into folders.
Running check...  Pipeline(s) without annotations.
Running check...  Data Flow(s) without a description value.
Running check...  Activitie(s) with timeout values still set to the service default value of 7 days.
Running check...  Activitie(s) without a description value.
Running check...  Activitie(s) ForEach iteration without a batch count value set.
Running check...  Activitie(s) ForEach iteration with a batch count size that is less than the service maximum.
Running check...  Linked Service(s) not using Azure Key Vault to store credentials.
Running check...  Linked Service(s) not used by any other resource.
Running check...  Linked Service(s) without a description value.
Running check...  Linked Service(s) without annotations.
Running check...  Dataset(s) not used by any other resource.
Running check...  Dataset(s) without a description value.
Running check...  Dataset(s) not organised into folders.
Running check...  Dataset(s) without annotations.
Running check...  Trigger(s) not used by any other resource.
Running check...  Trigger(s) without a description value.
Running check...  Trigger(s) without annotations.

Results Summary:

Checks ran against template: 21
Checks with issues found: 21
Total issue count: 264

Issue Count Check Details                                                                                  Severity
----------- -------------                                                                                  --------
         14 Pipeline(s) without any triggers attached. Directly or indirectly.                             Medium
          2 Pipeline(s) with an impossible AND/OR activity execution chain.                                High
         14 Pipeline(s) without a description value.                                                       Low
          2 Pipeline(s) not organised into folders.                                                        Low
         16 Pipeline(s) without annotations.                                                               Low
          6 Data Flow(s) without a description value.                                                      Low
         21 Activitie(s) with timeout values still set to the service default value of 7 days.             High
         35 Activitie(s) without a description value.                                                      Low
          2 Activitie(s) ForEach iteration without a batch count value set.                                High
          3 Activitie(s) ForEach iteration with a batch count size that is less than the service maximum.  Medium
         21 Linked Service(s) not using Azure Key Vault to store credentials.                              High
          9 Linked Service(s) not used by any other resource.                                              Medium
         24 Linked Service(s) without a description value.                                                 Low
         16 Linked Service(s) without annotations.                                                         Low
          9 Dataset(s) not used by any other resource.                                                     Medium
         30 Dataset(s) without a description value.                                                        Low
          7 Dataset(s) not organised into folders.                                                         Low
         30 Dataset(s) without annotations.                                                                Low
          1 Trigger(s) not used by any other resource.                                                     Medium
          1 Trigger(s) without a description value.                                                        Low
          1 Trigger(s) without annotations.                                                                Low
```

Thank you for listening…

Paul Andrew

altius

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

Twitter: @mrpaulandrew
LinkedIn: In/mrpaulandrew

GitHub: github.com/mrpaulandrew /CommunityEvents /ContentCollateral /procfwk

STRATEGIC PARTNER

Microsoft

GOLD SPONSOR

elitmind think bright

Future Processing

SILVER SPONSOR

dbWatch DATABASE CONTROL

nationale nederlanden

BRONZE SPONSOR

predica.