

STRATEGIC PARTNER



GOLD SPONSOR



SILVER SPONSOR



BRONZE SPONSOR





Azure Orchestration – Applying Data Factory in Production

Paul Andrew | Principal Consultant & Solution Architect



altius

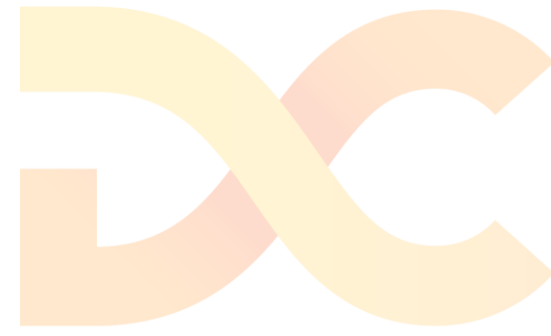


@MrPaulAndrew



In/MrPaulAndrew





<https://github.com/mrpaulandrew>

CommunityEvents

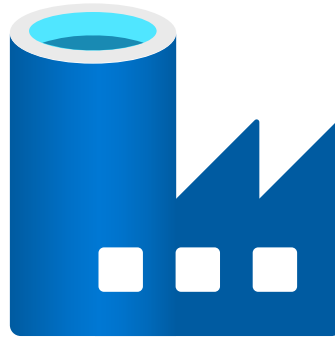
Demo code, content and slides from various community events.

● C++

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

AGENDA ??

Applying Data Factory in Production

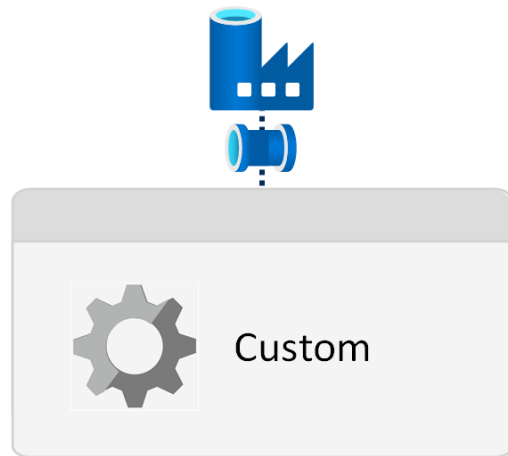




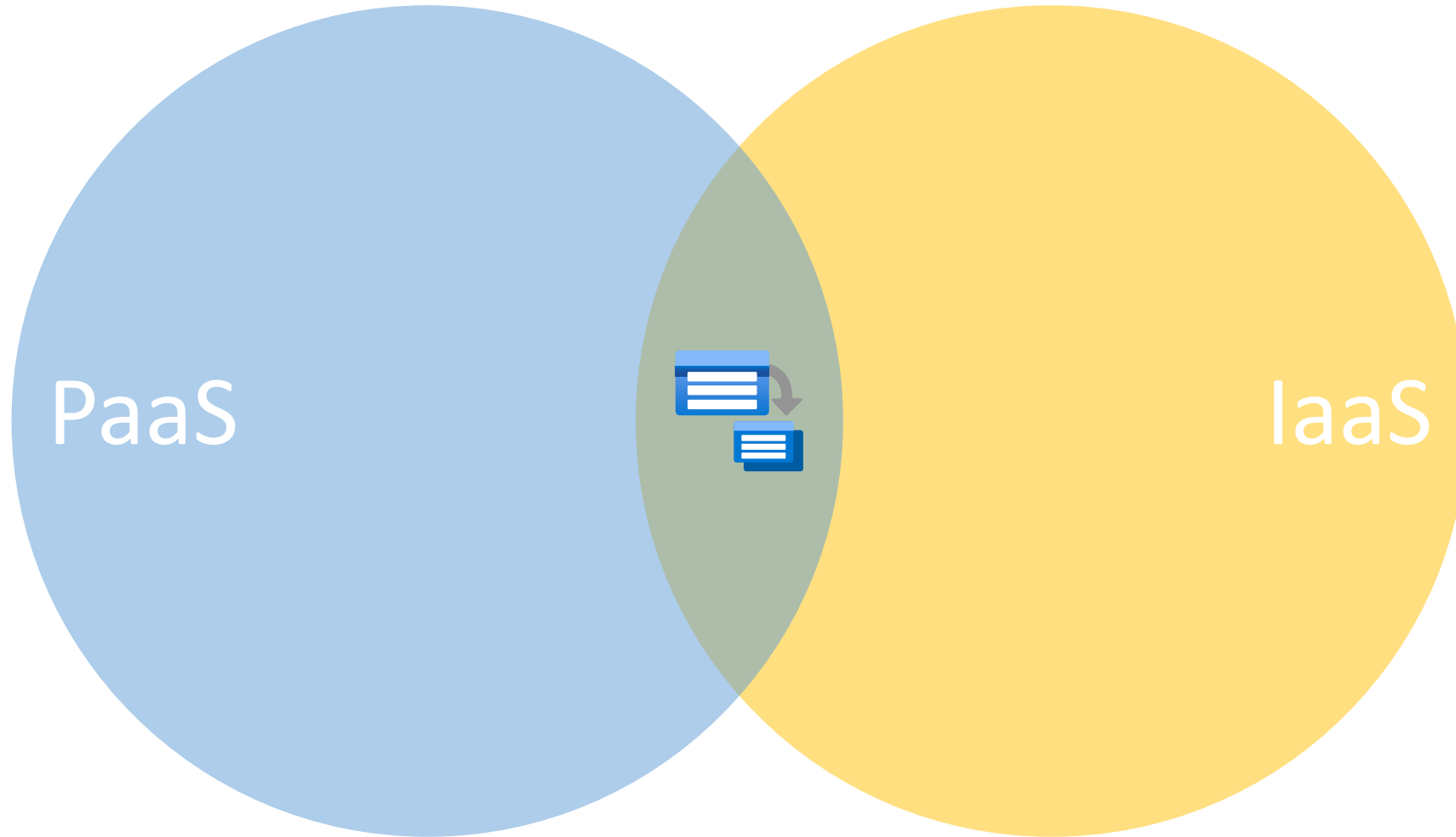
AGENDA – Short Stories

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

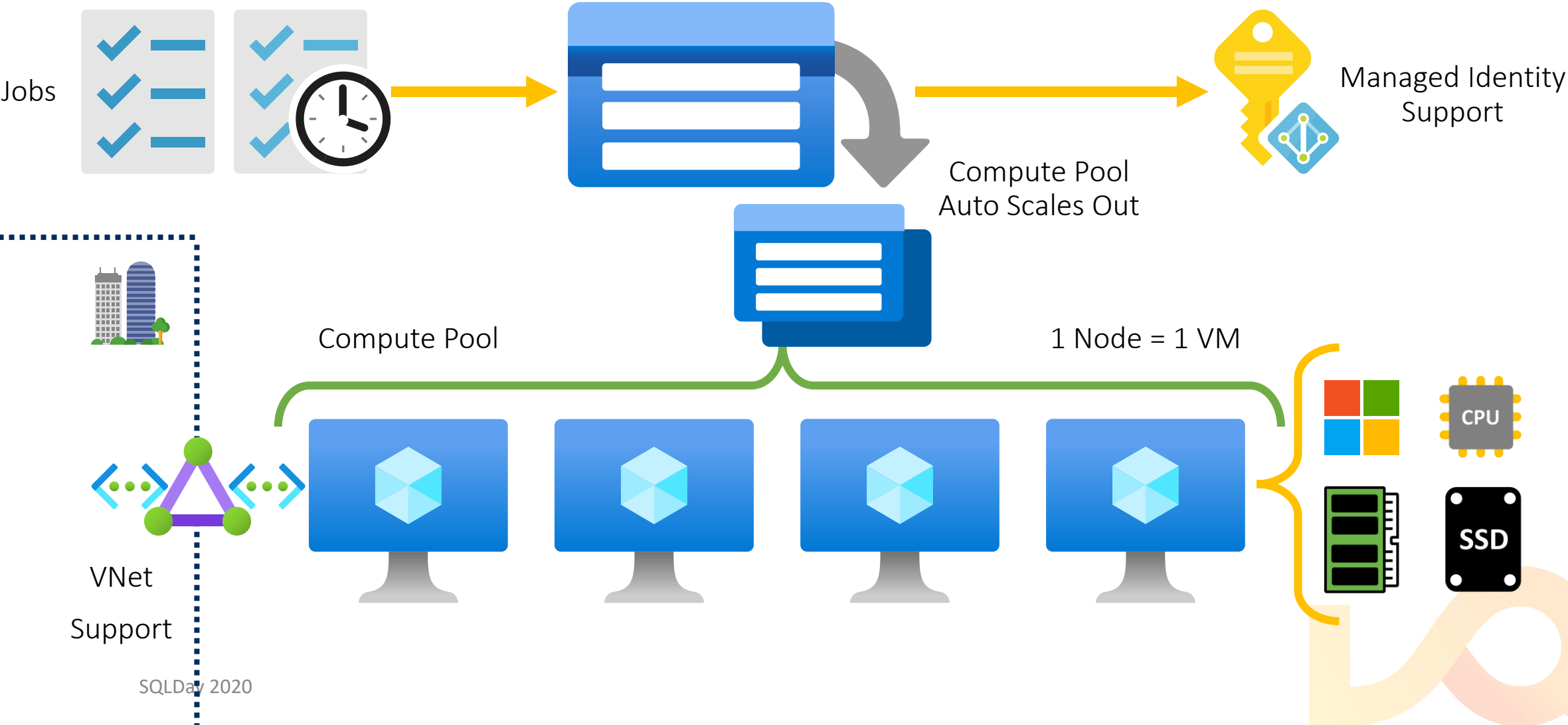
Custom Activities



Azure Batch Service



Azure Batch Service



Custom Activity

Extend Data Factory with Custom Code

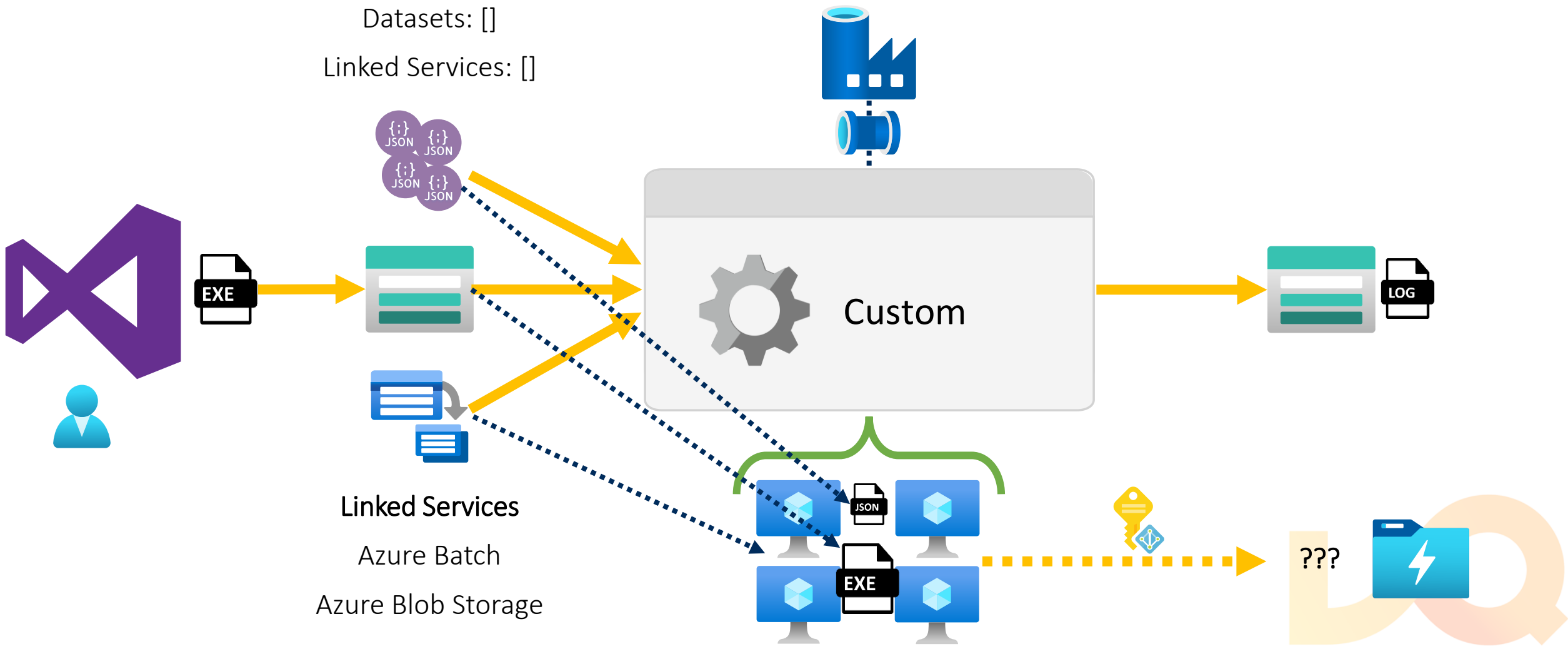


<https://mrpaulandrew.com/2018/11/12/creating-an-azure-data-factory-v2-custom-activity/>

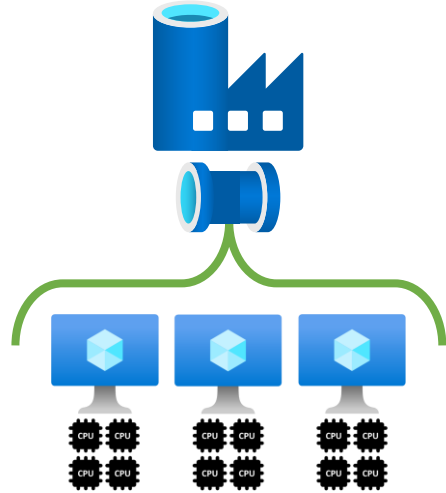
References Objects

Datasets: []

Linked Services: []



Controlling & Scaling Compute

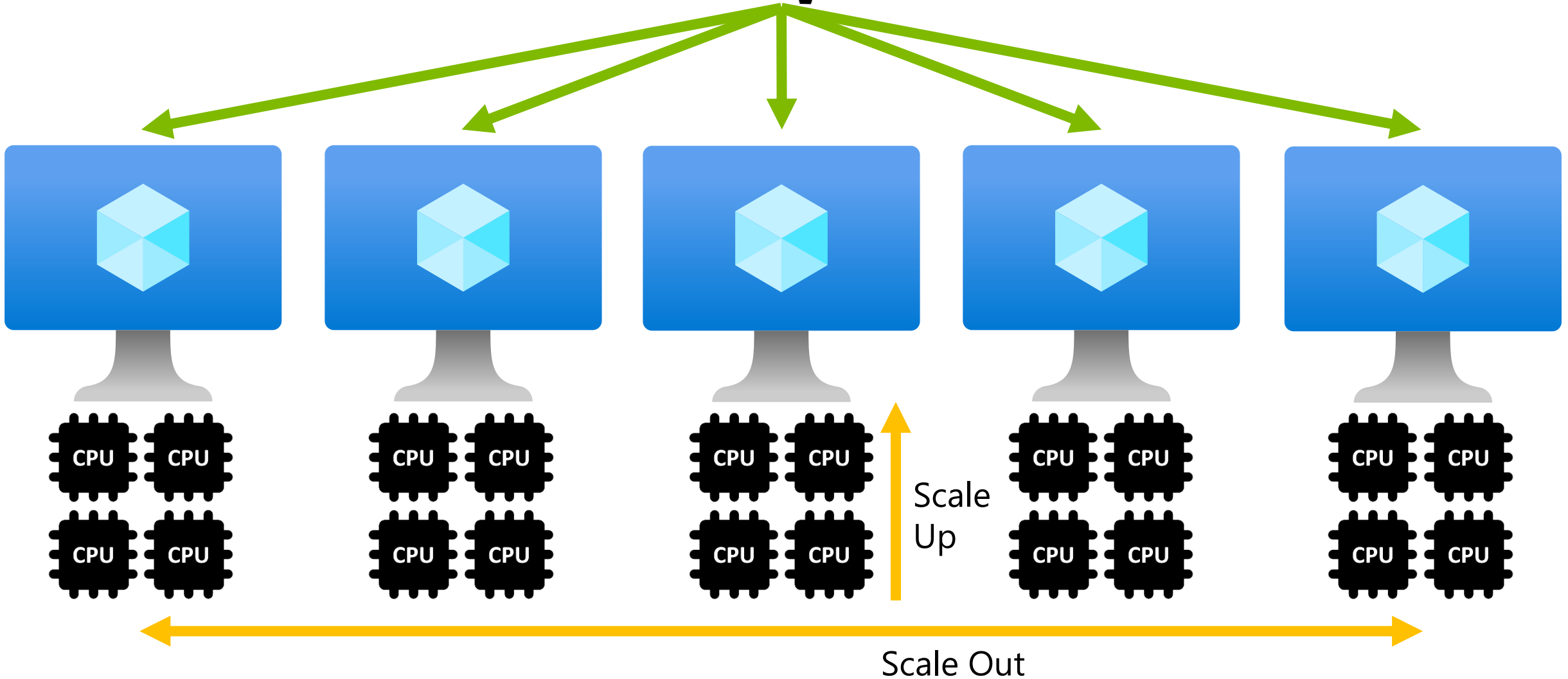


Scaling Up and/or Scaling Out



Workload:

Process 100TB of Data

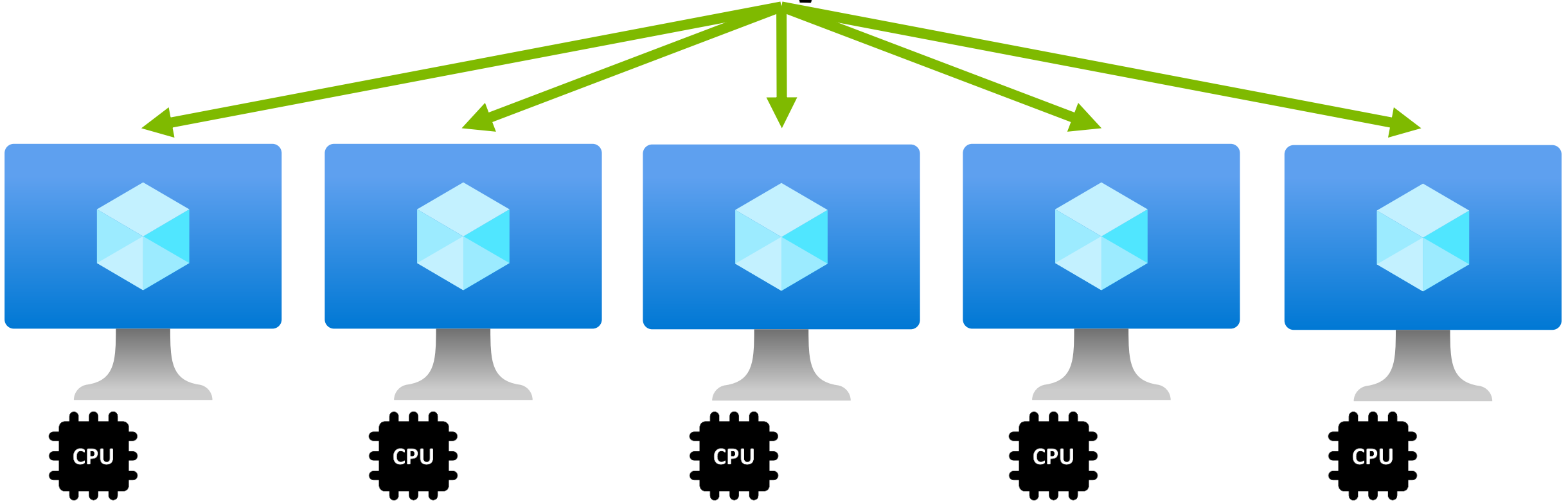


Scaling Up and/or Scaling Out



Workload:

Process 100TB of Data



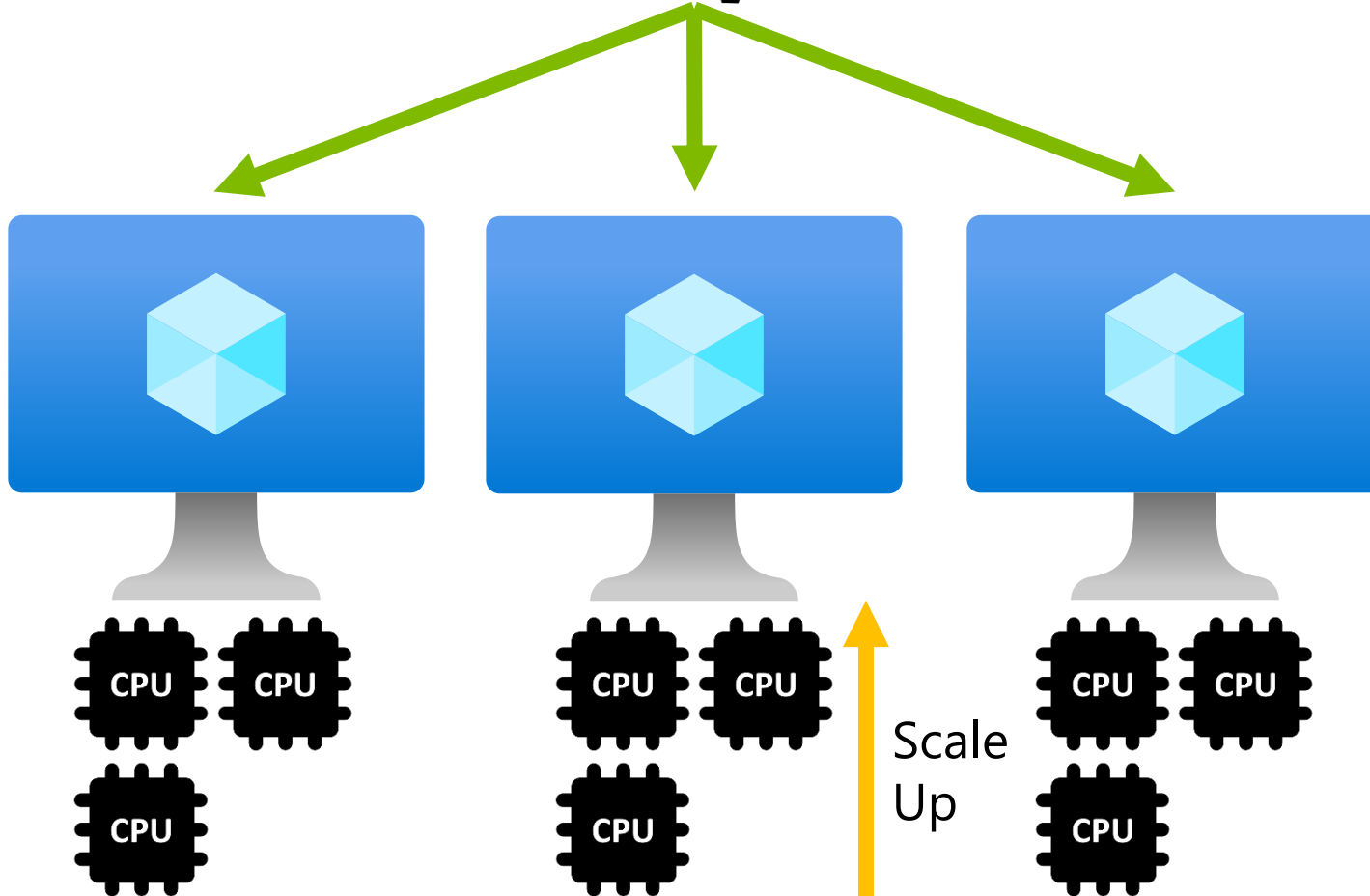
Scale Out

Scaling Up and/or Scaling Out



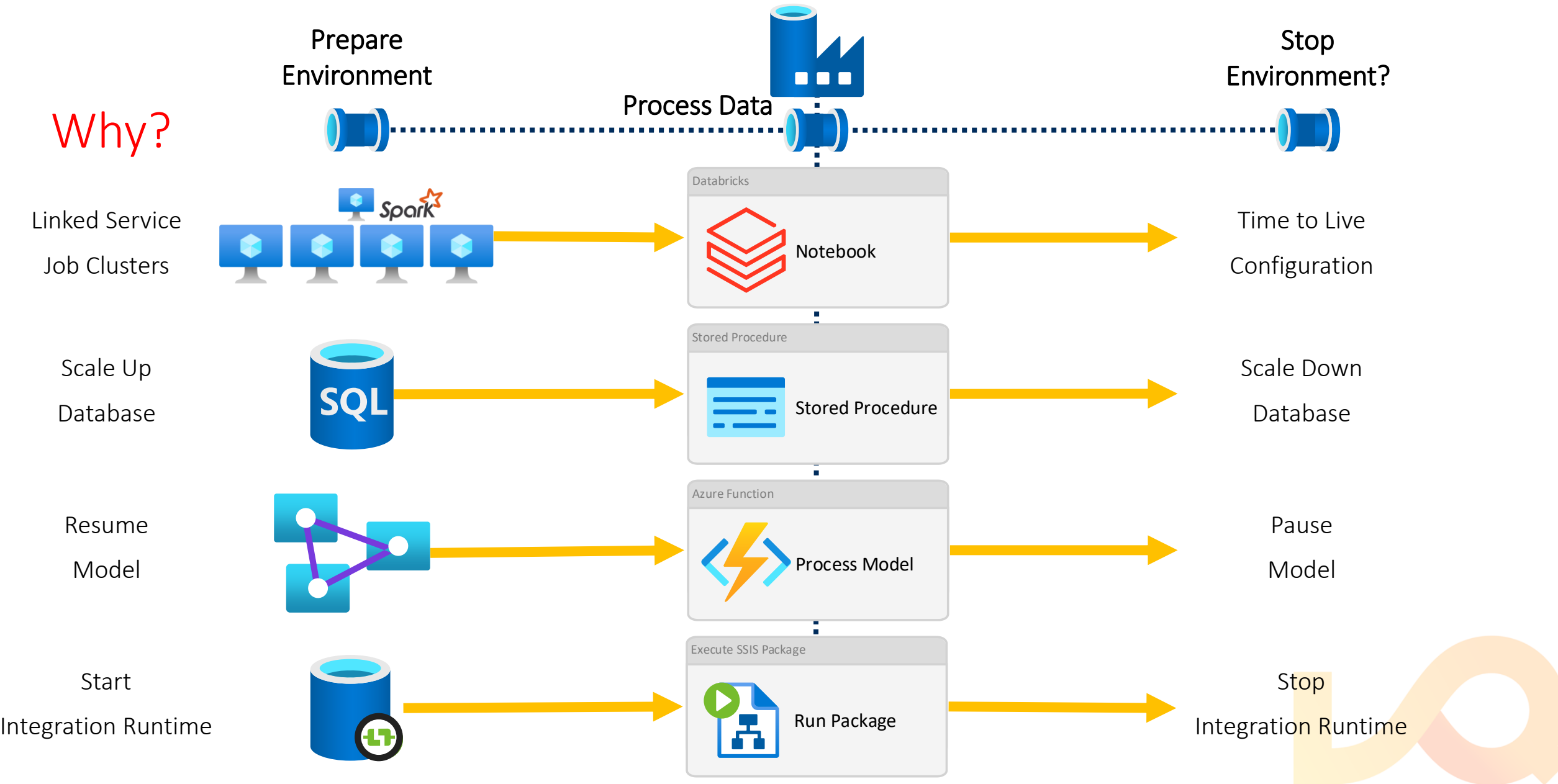
Workload:

Process 100TB of Data

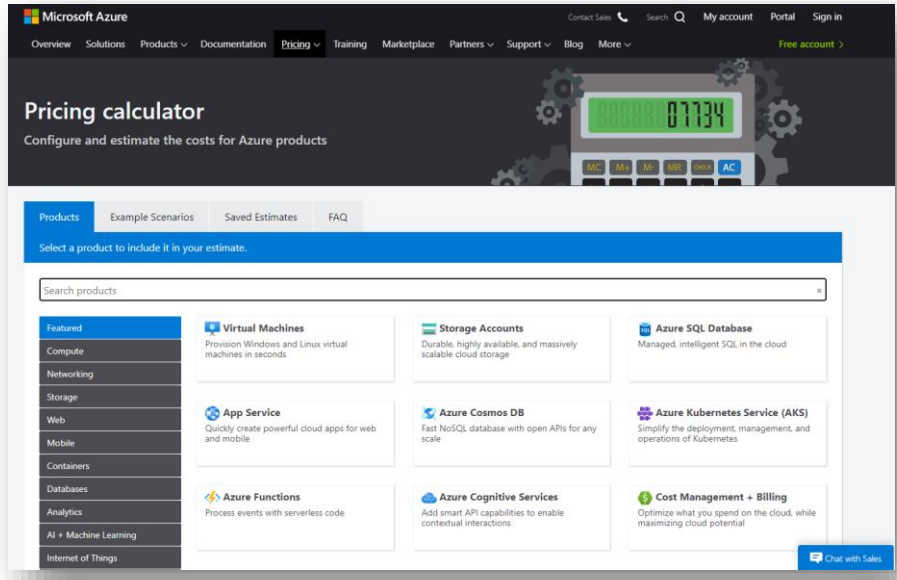
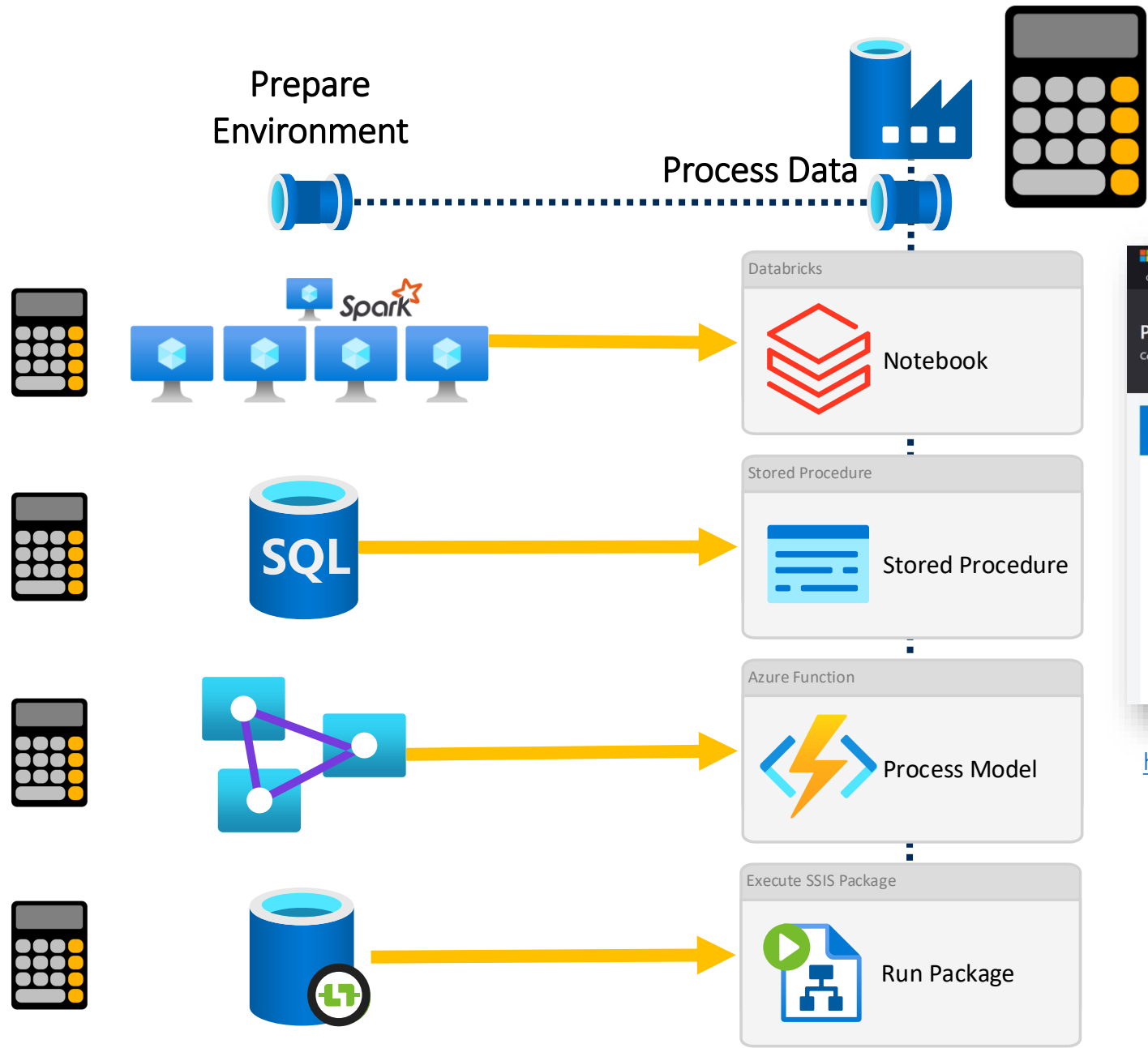


Scale Out

Resource Control



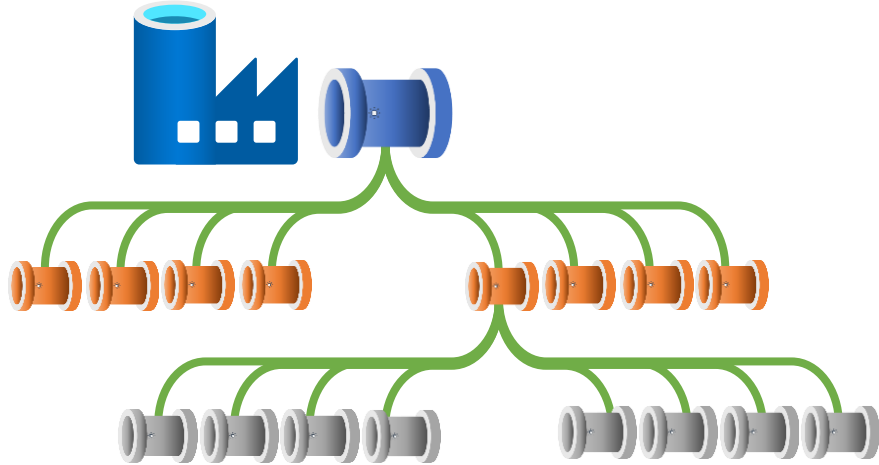
Resource Control - Cost



<https://azure.microsoft.com/en-us/pricing/calculator/>

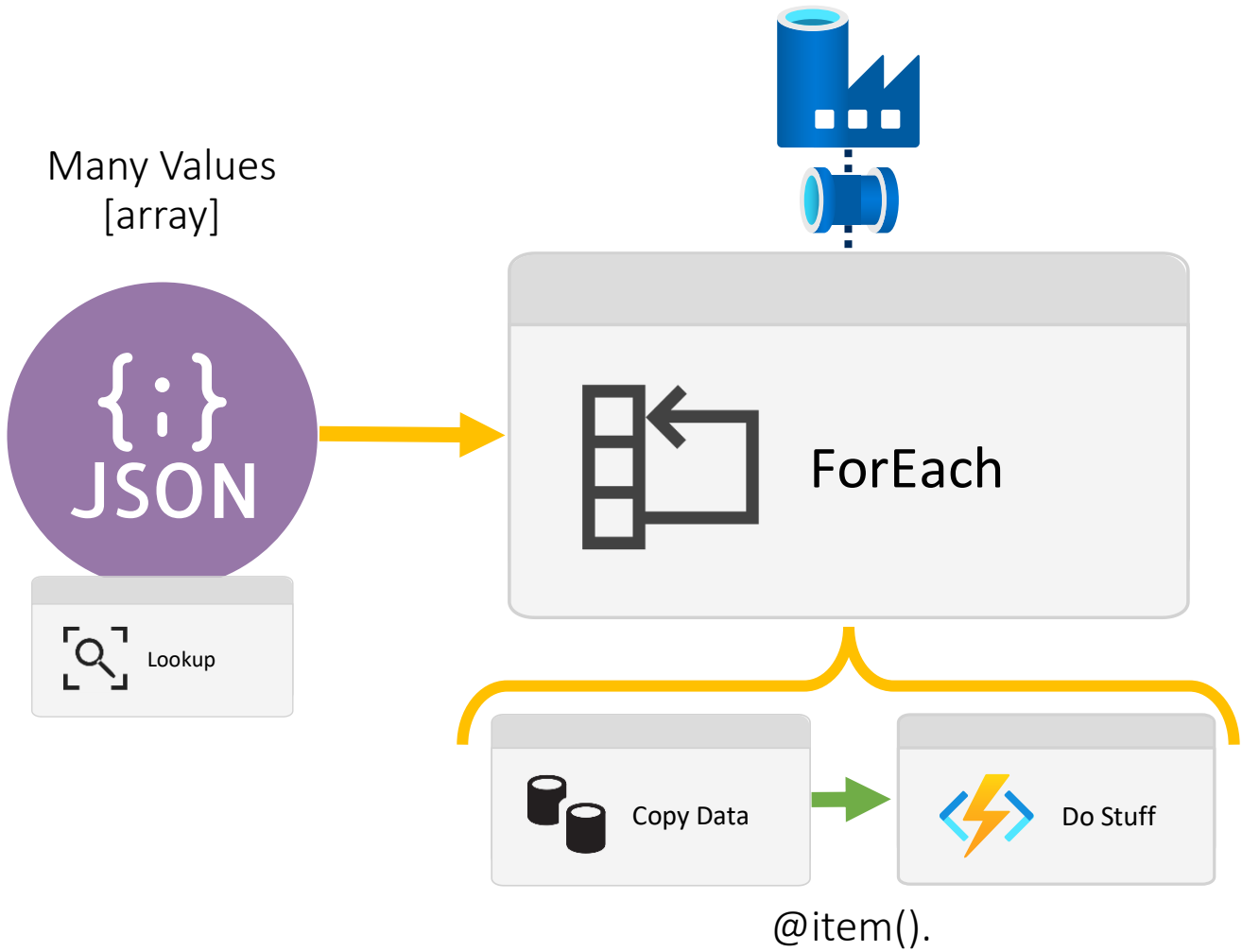


Scale Out Execution

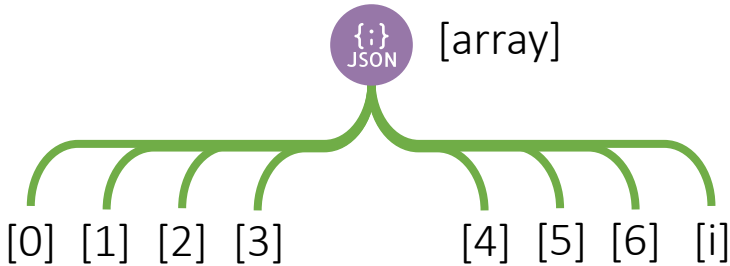
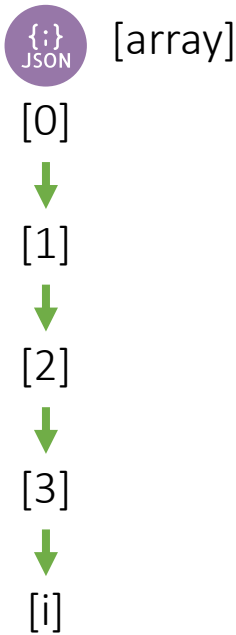


ForEach

Scaling Out Control Flow Activities



IsSequential:
true

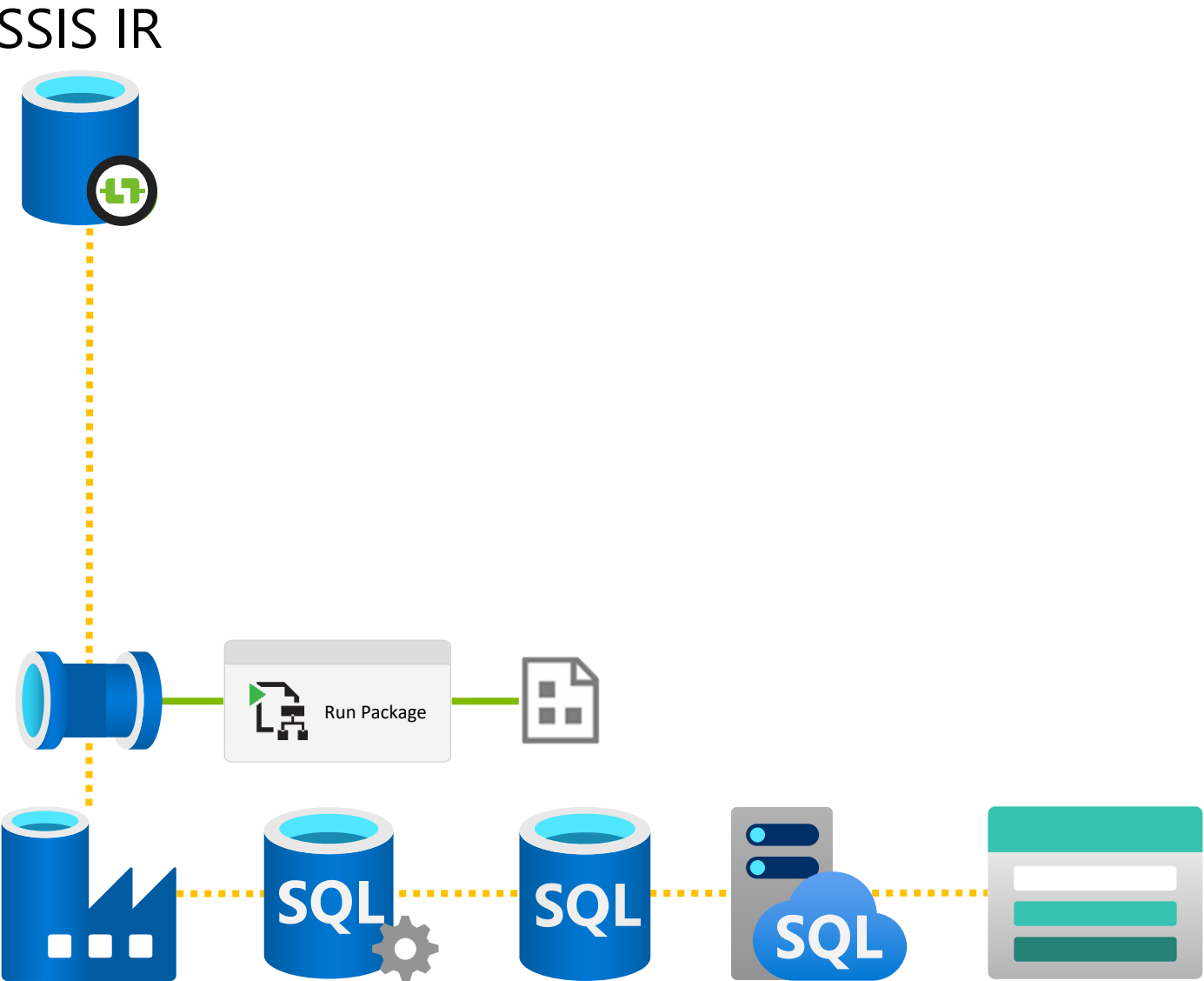


Batch Count Default: 20

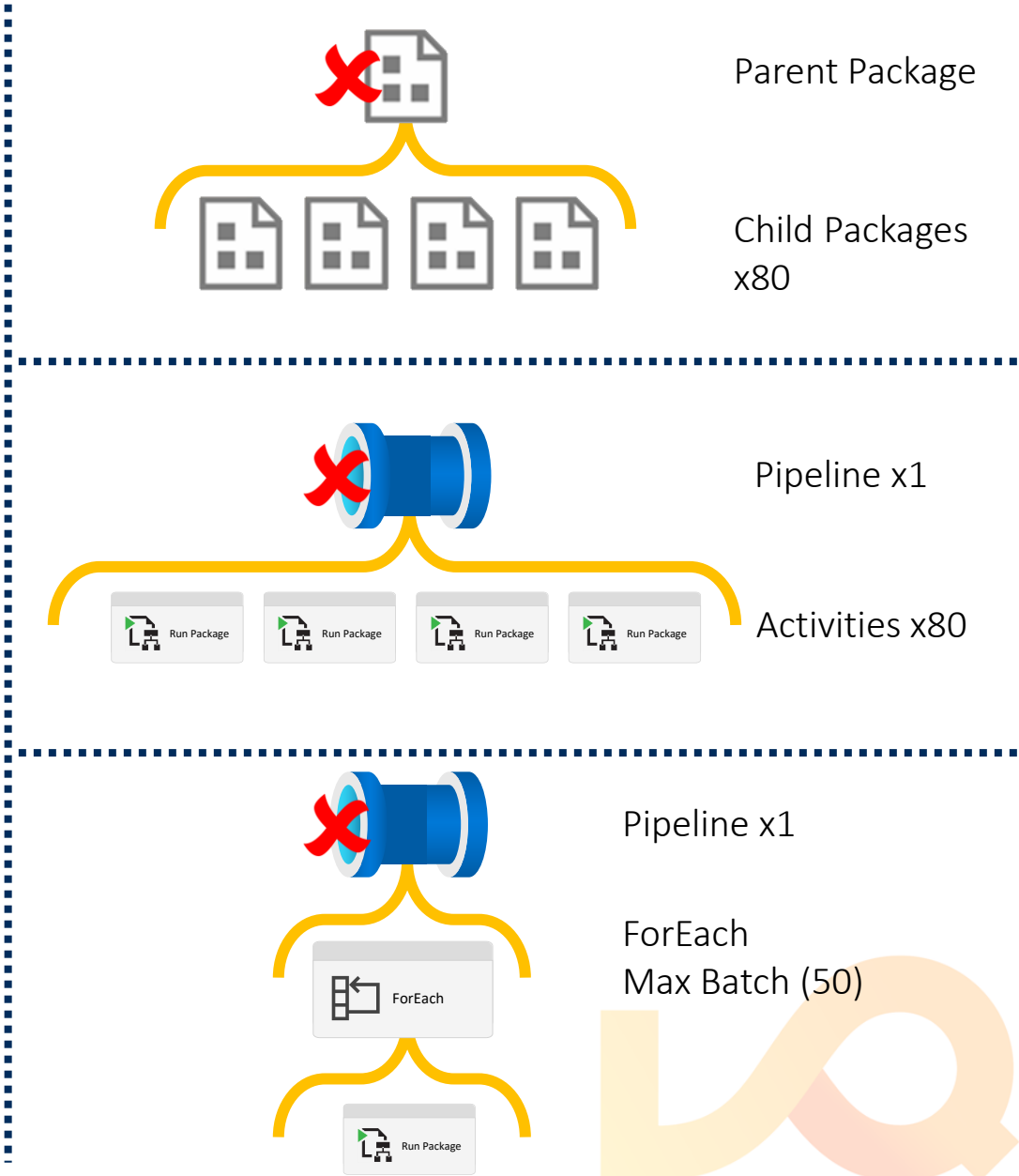
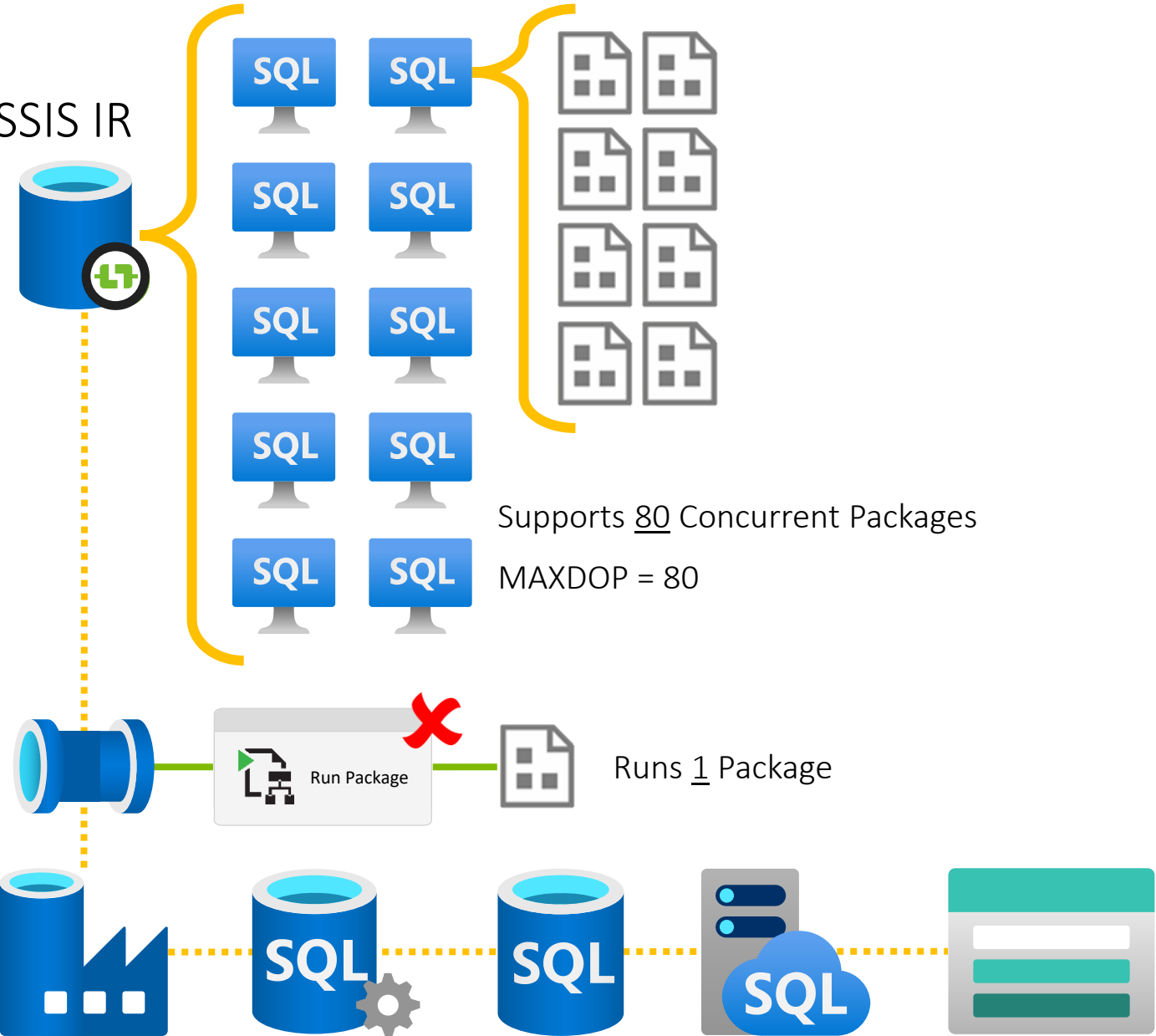
Batch Count Max: 50



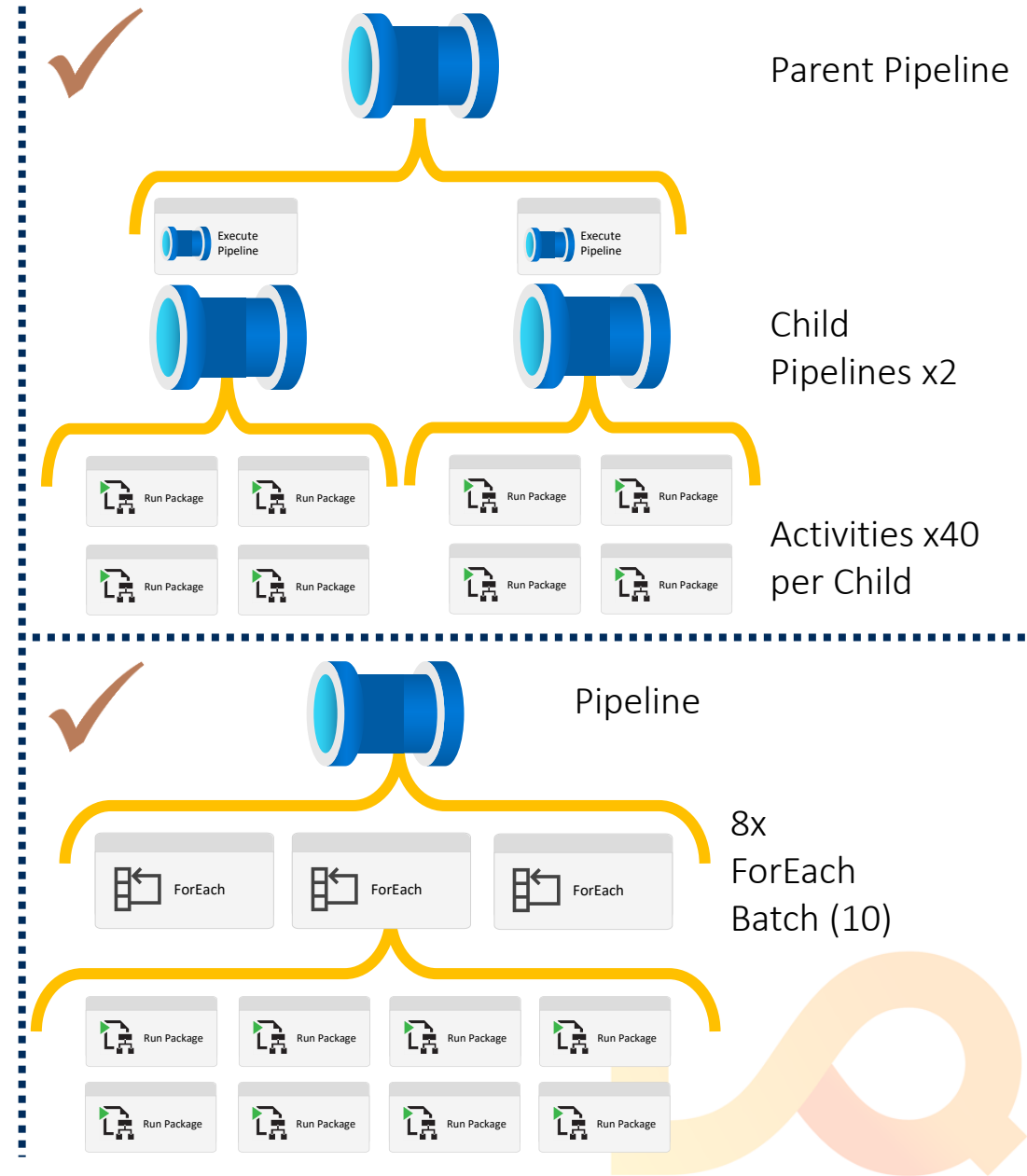
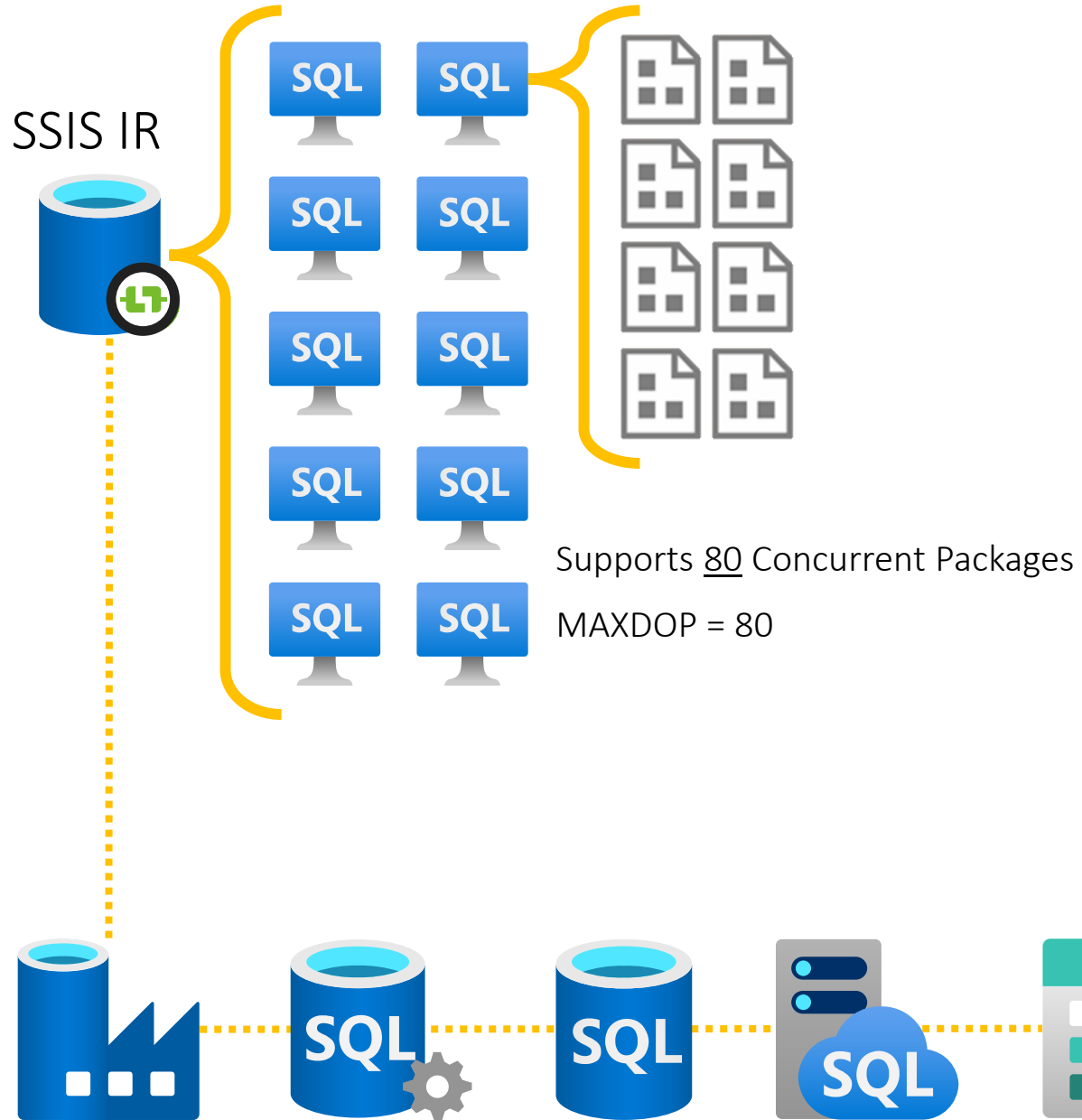
Problem: Using All Of The SSIS IR Compute



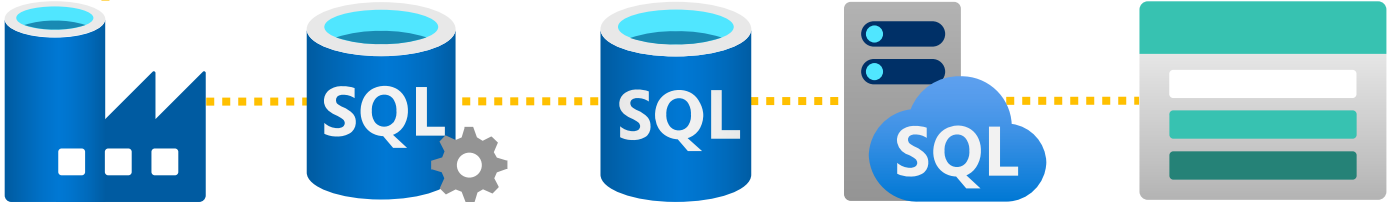
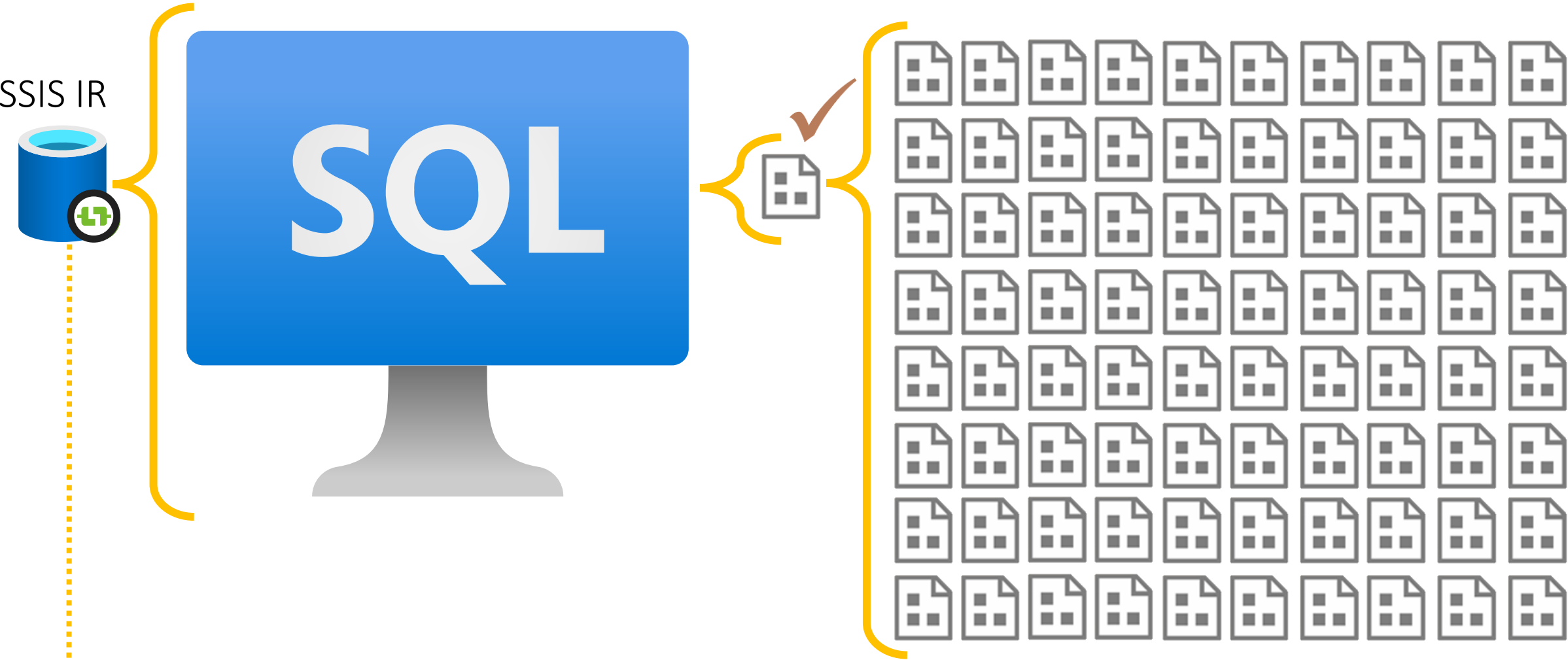
Problem: Using All Of The SSIS IR Compute



Solution 1 & 2: Static Pipelines



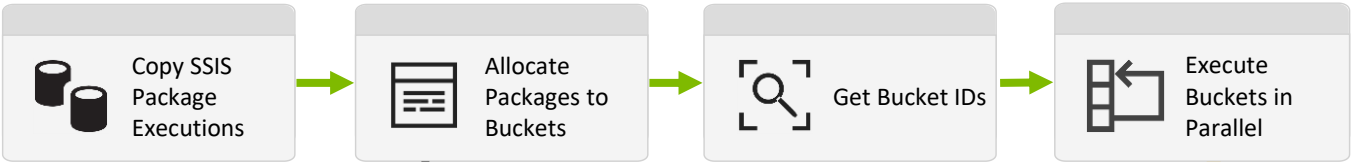
Solution 3: Packages Refactored on a Single Node IR



Solution 4: Nested ForEach Activities & Bucket Metadata

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

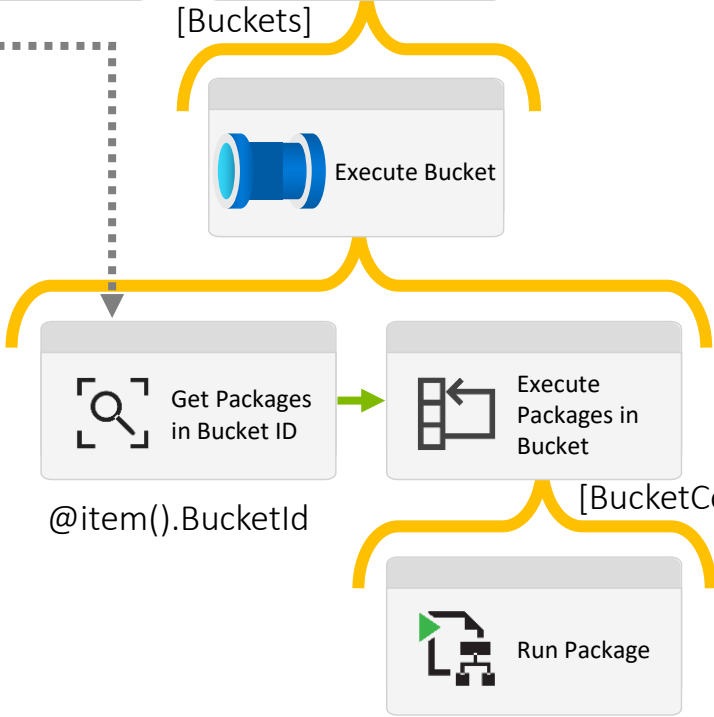
SSIS IR



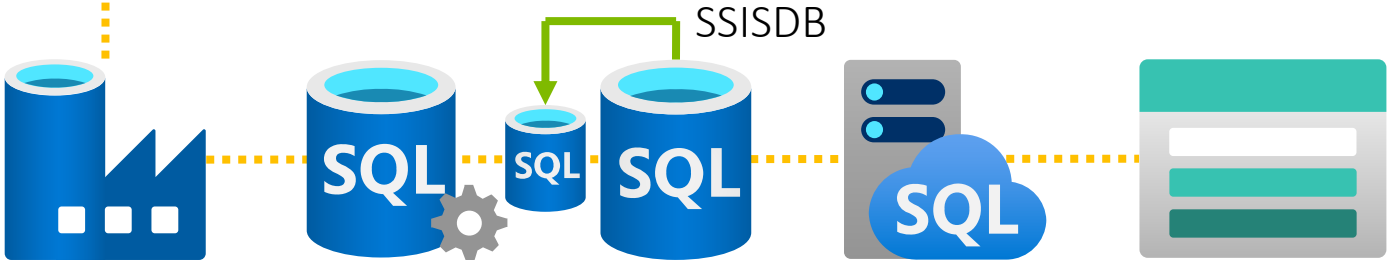
FE L1
MAXDOP 50

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>

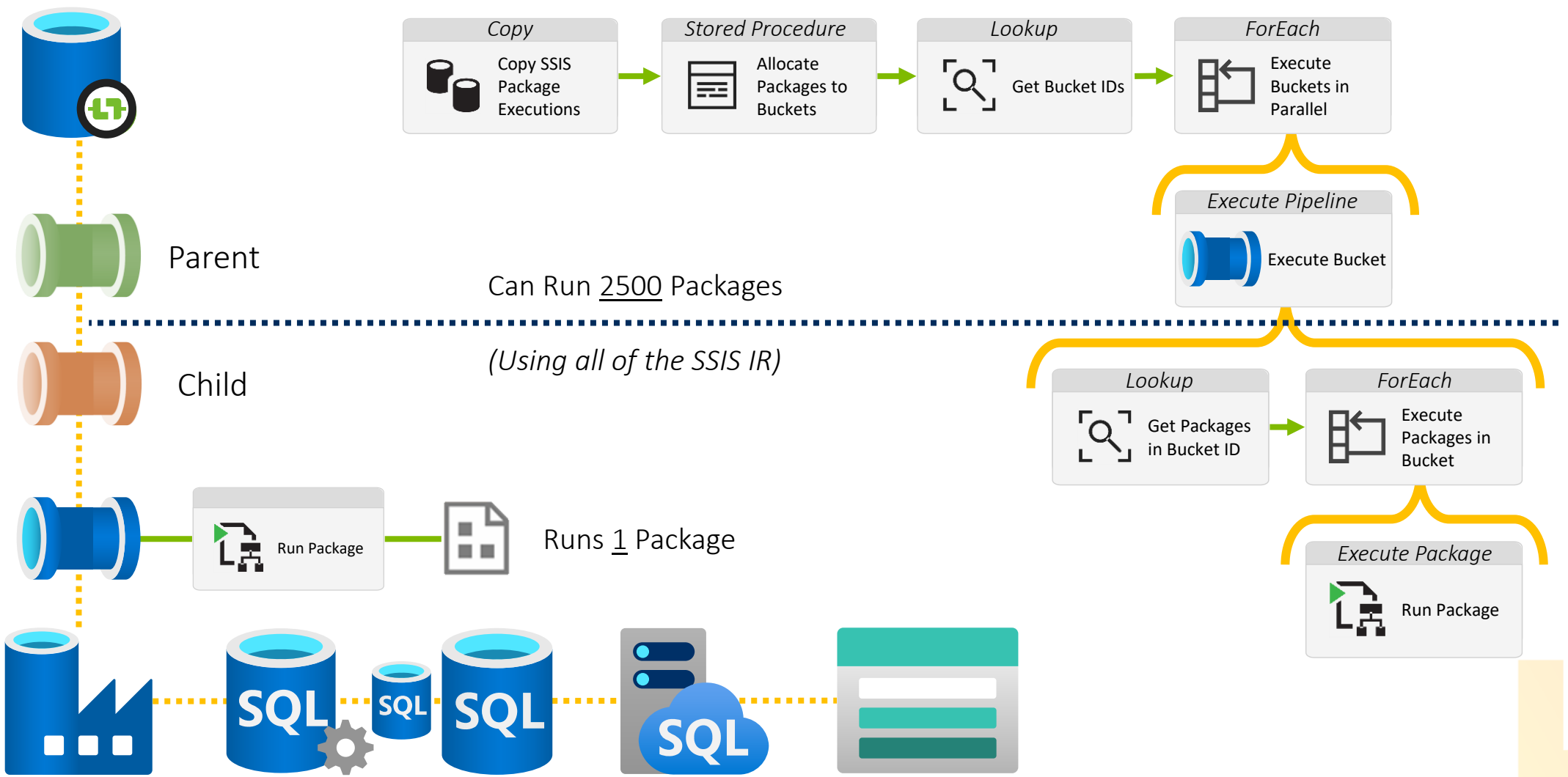


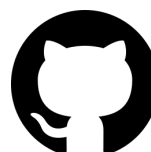
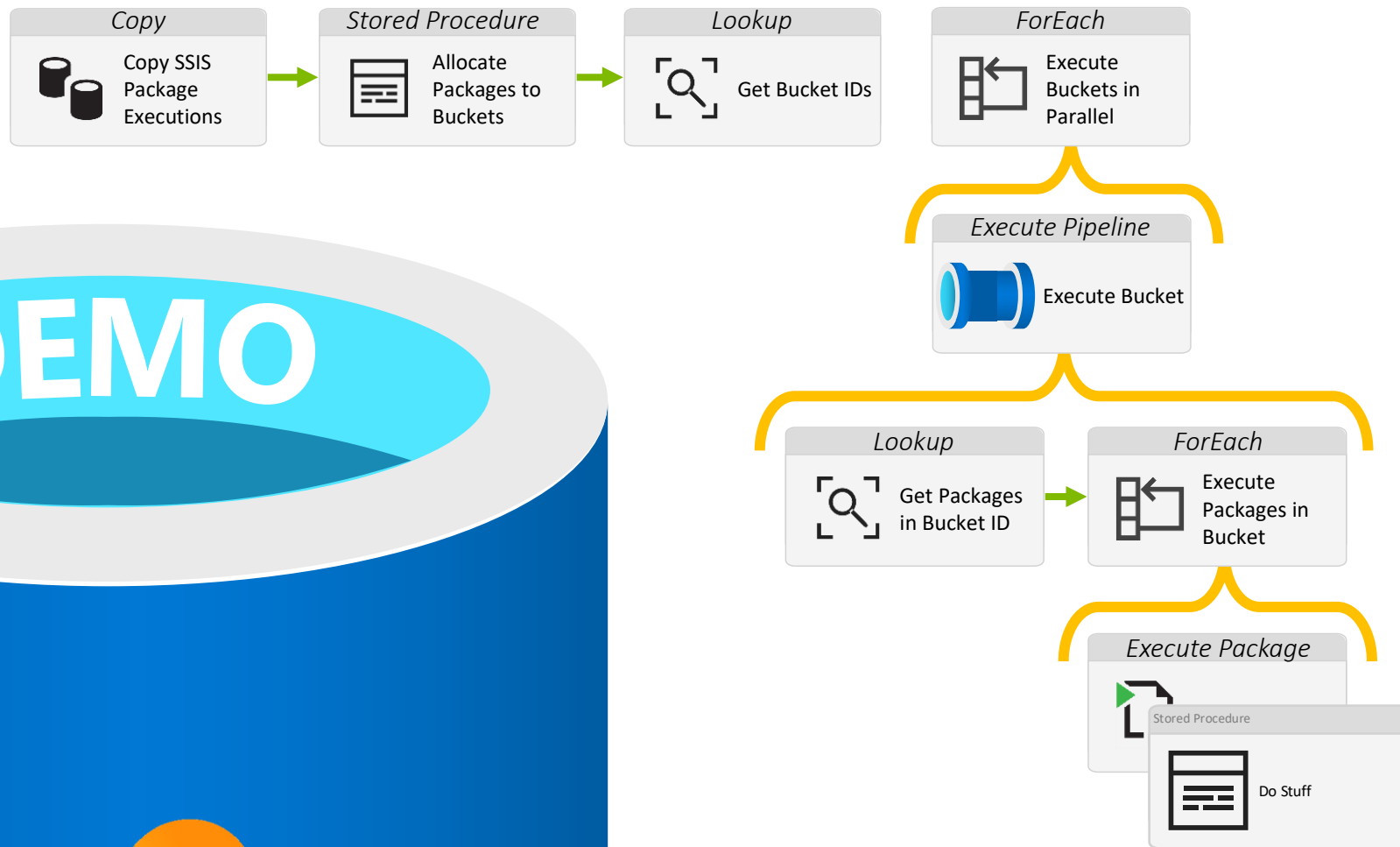
FE L2
MAXDOP 50



Solution 4: Nested ForEach Activities & Bucket Metadata

SSIS IR

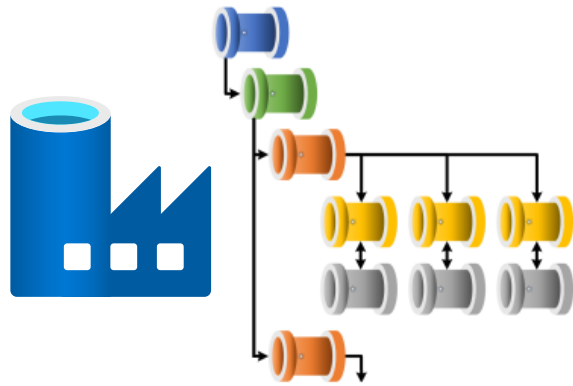




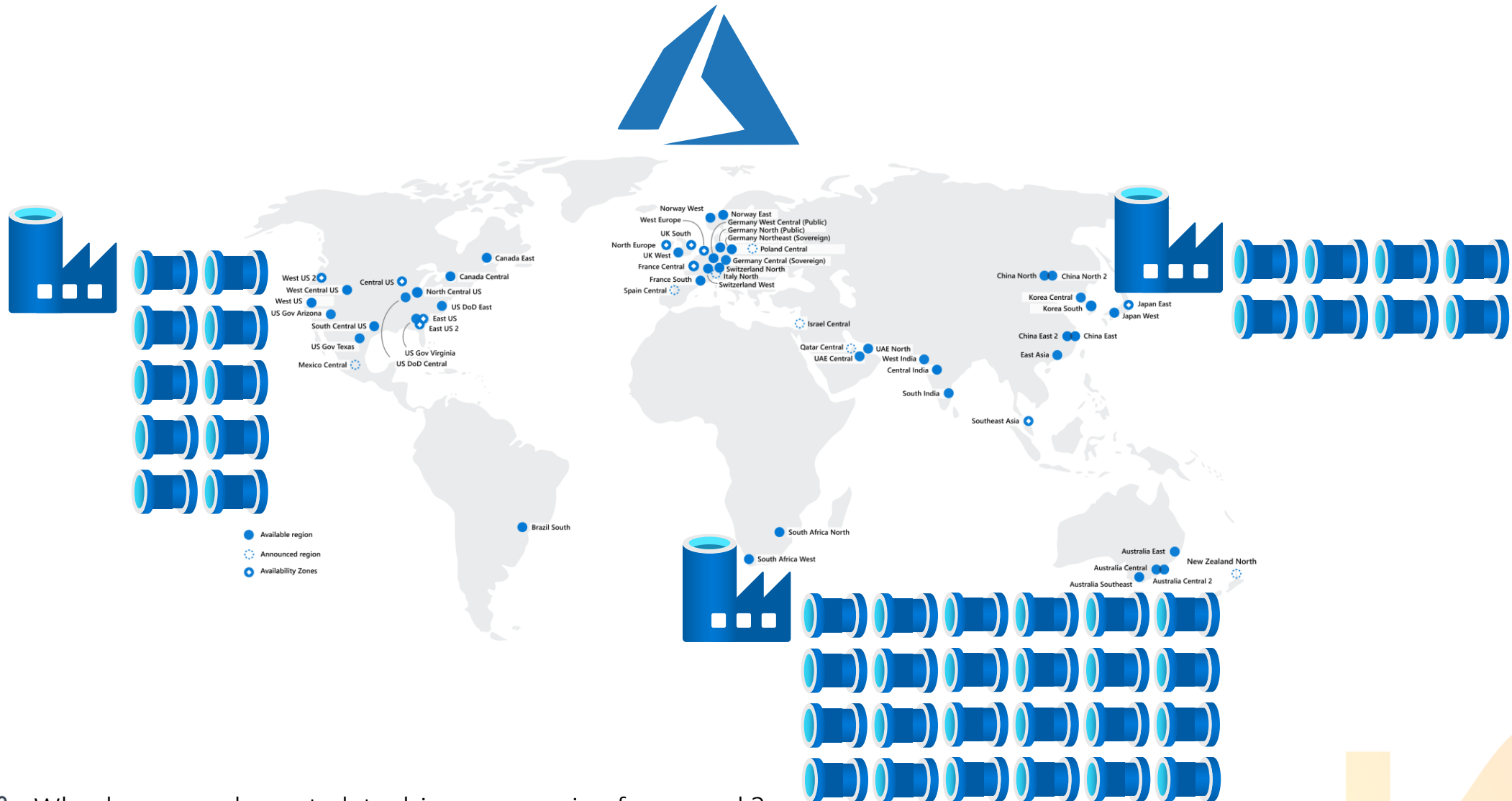
Demo Data Factory and code here:

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

Metadata Driven Pipelines



Problem: How should we structure our Data Factory Pipelines?

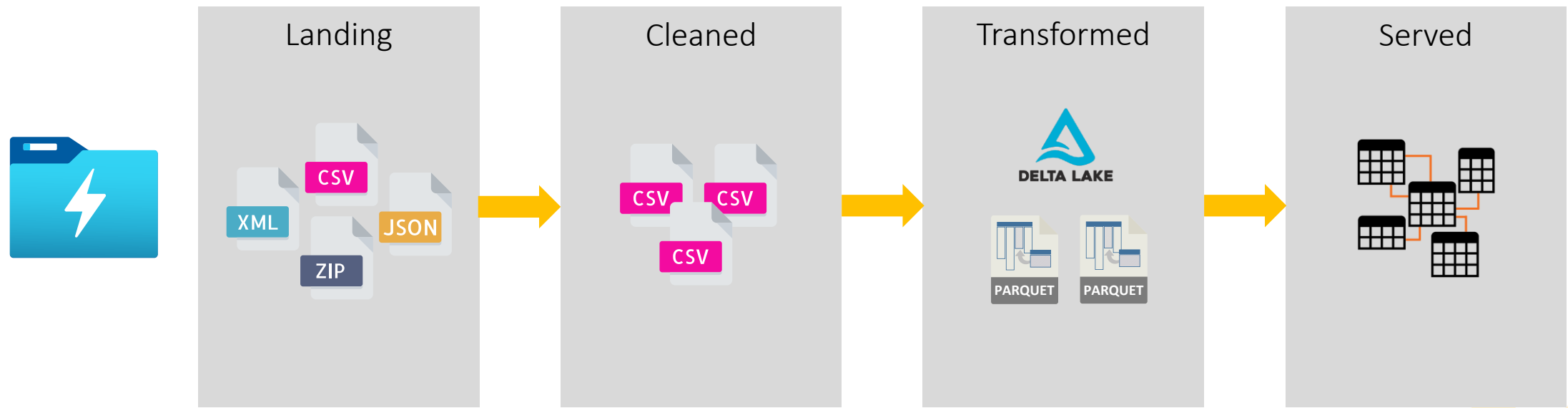
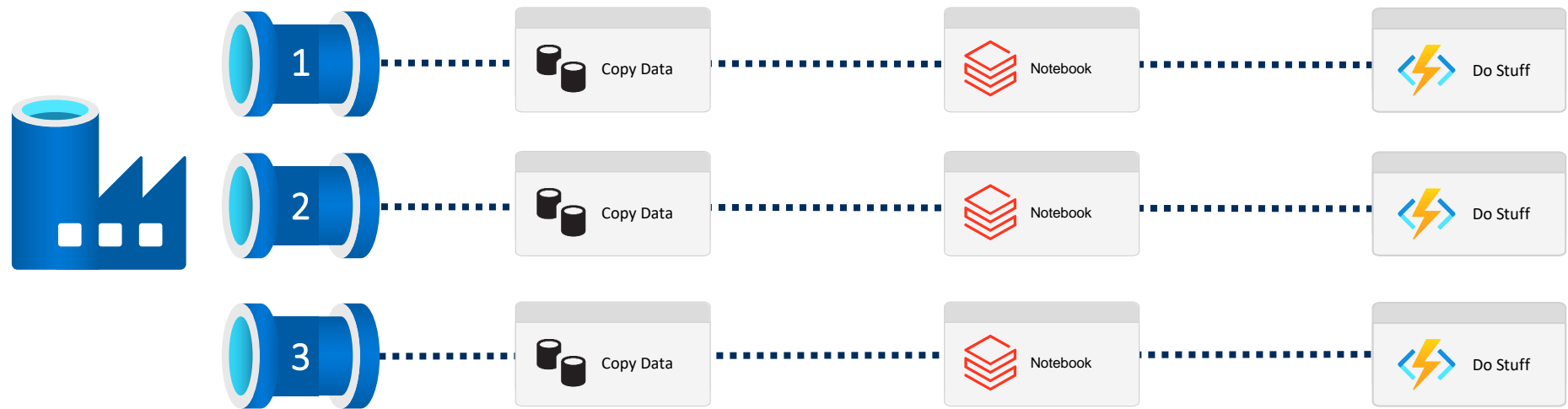


Why do we need a metadata driven processing framework?

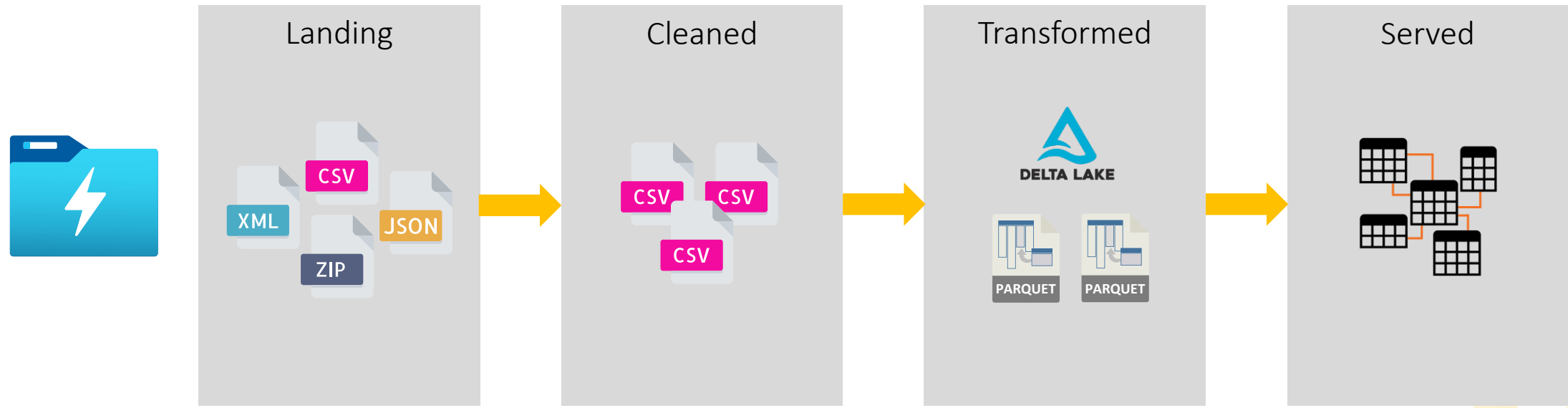
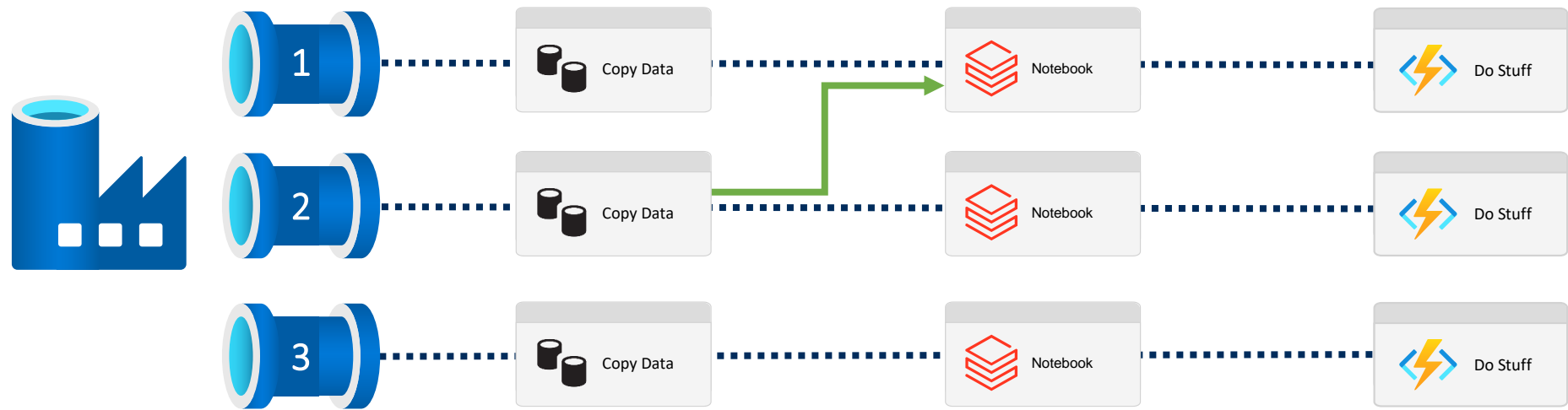
<https://youtu.be/rVlc-GBpNnc>



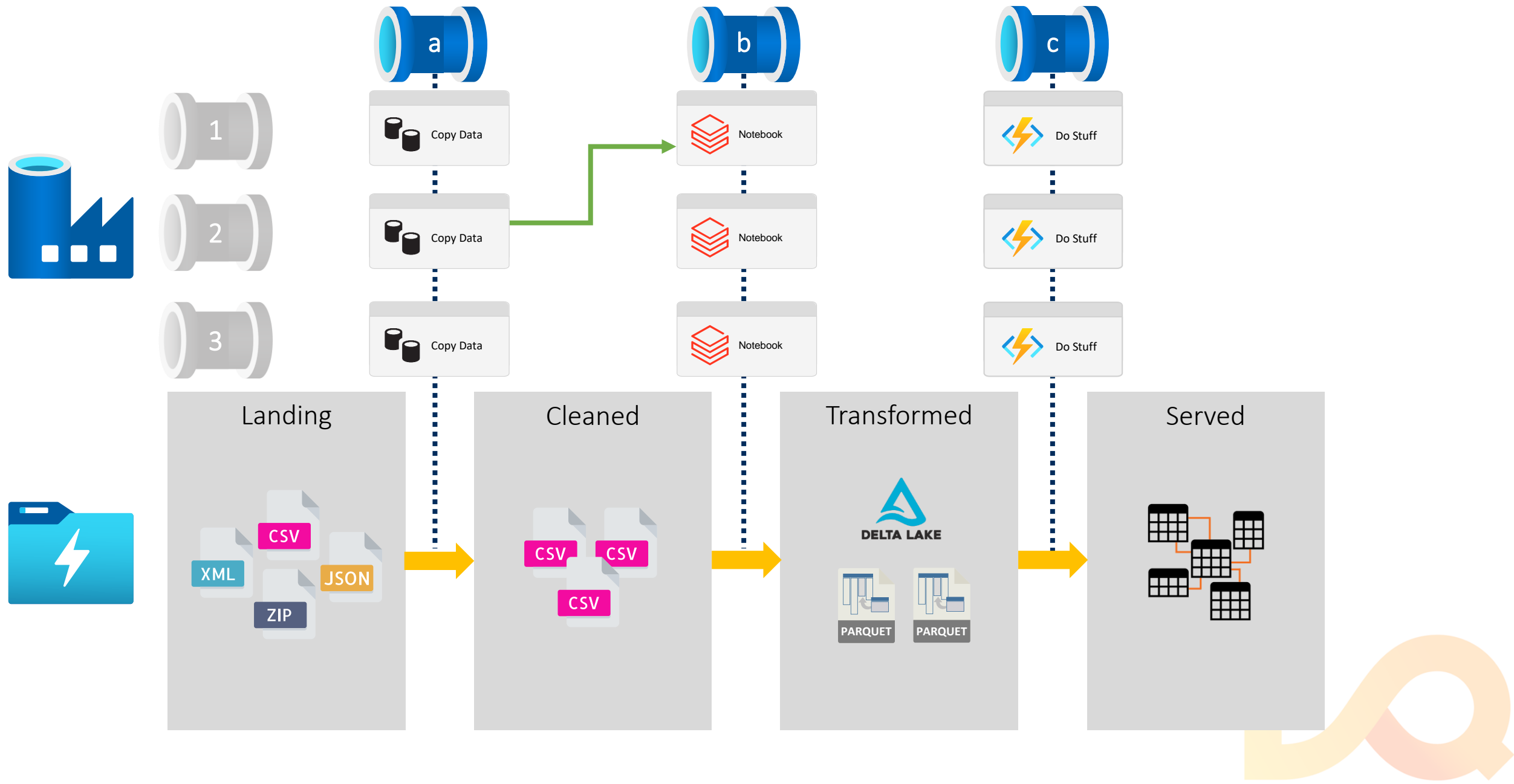
Problem



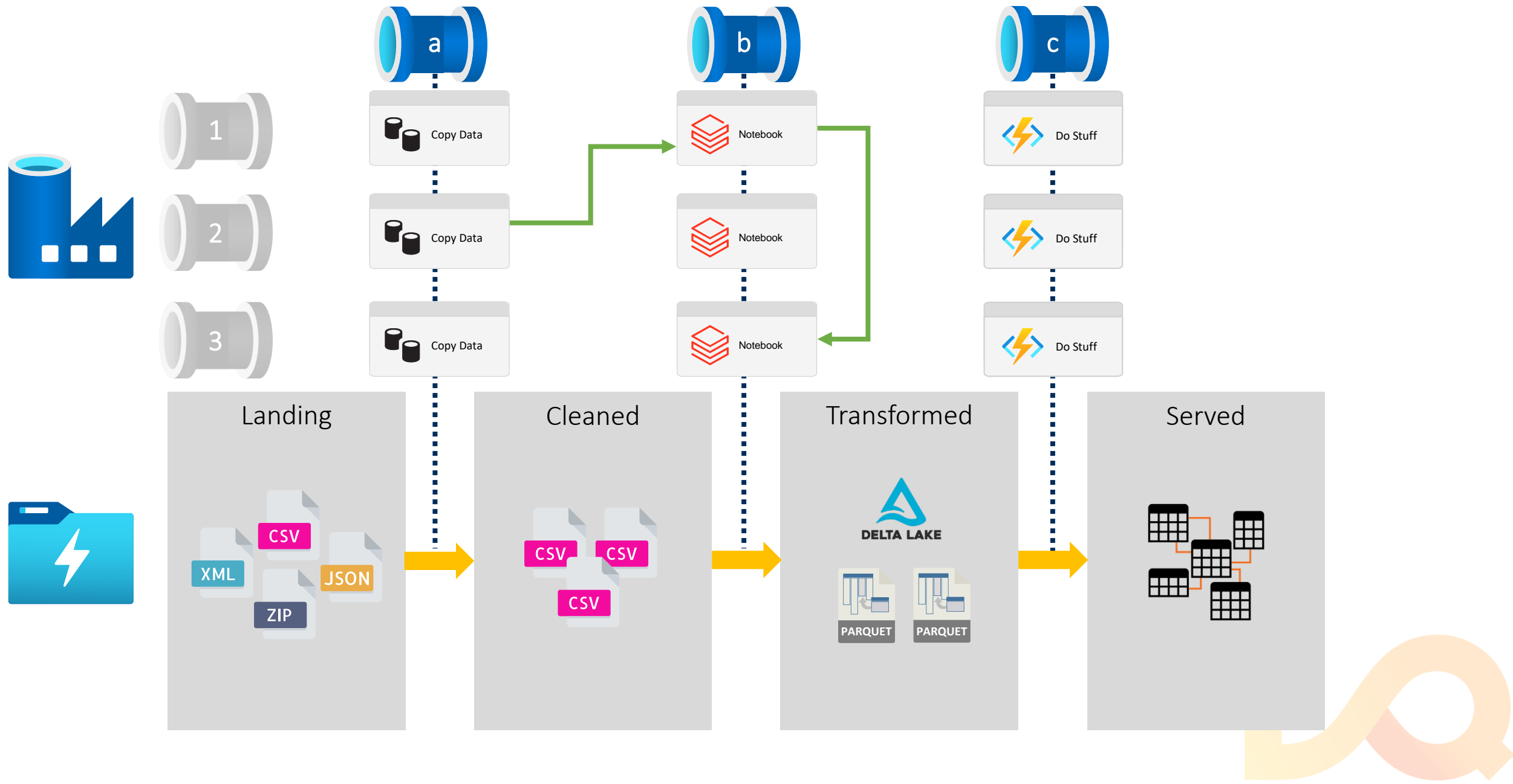
Problem



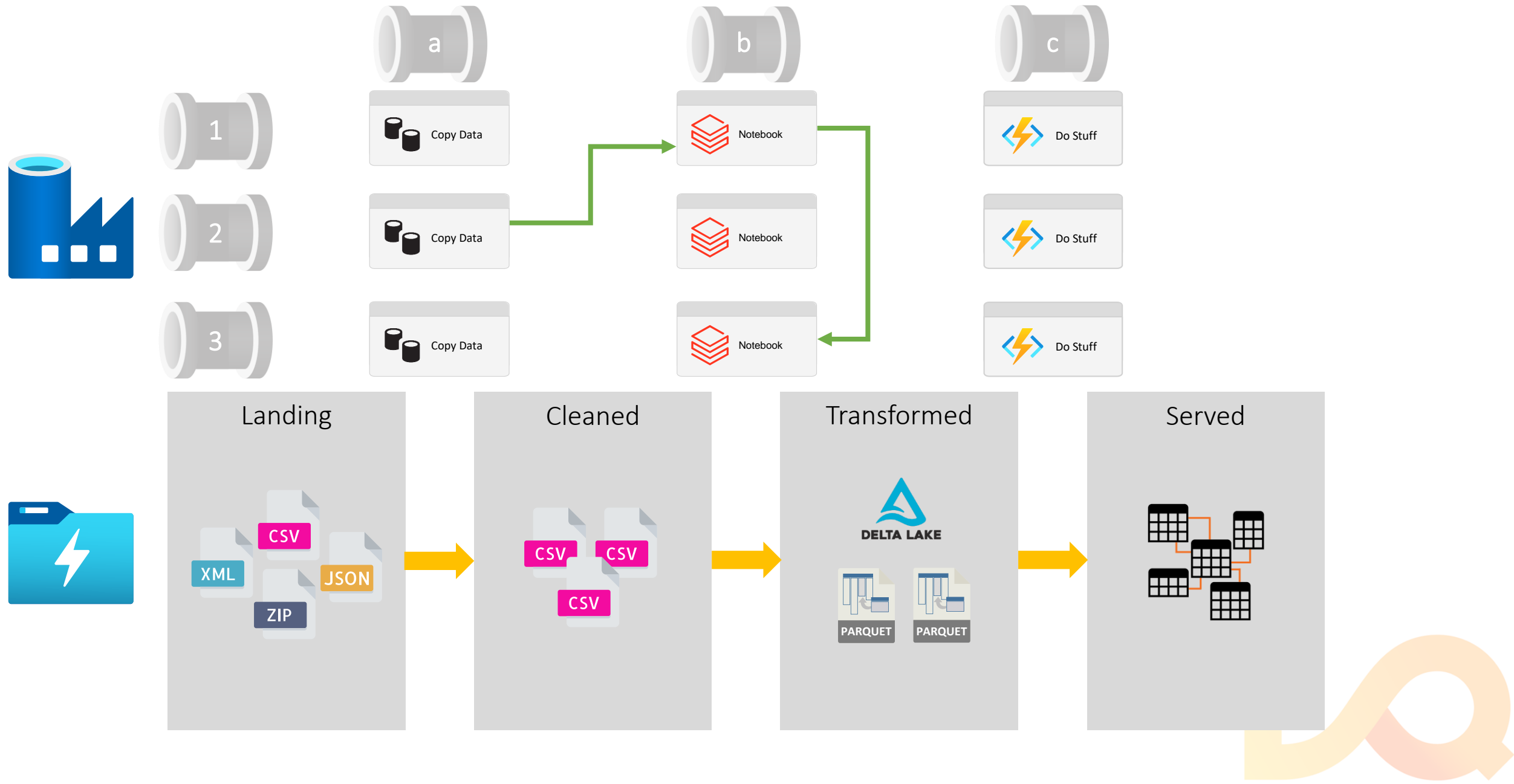
Problem




Problem

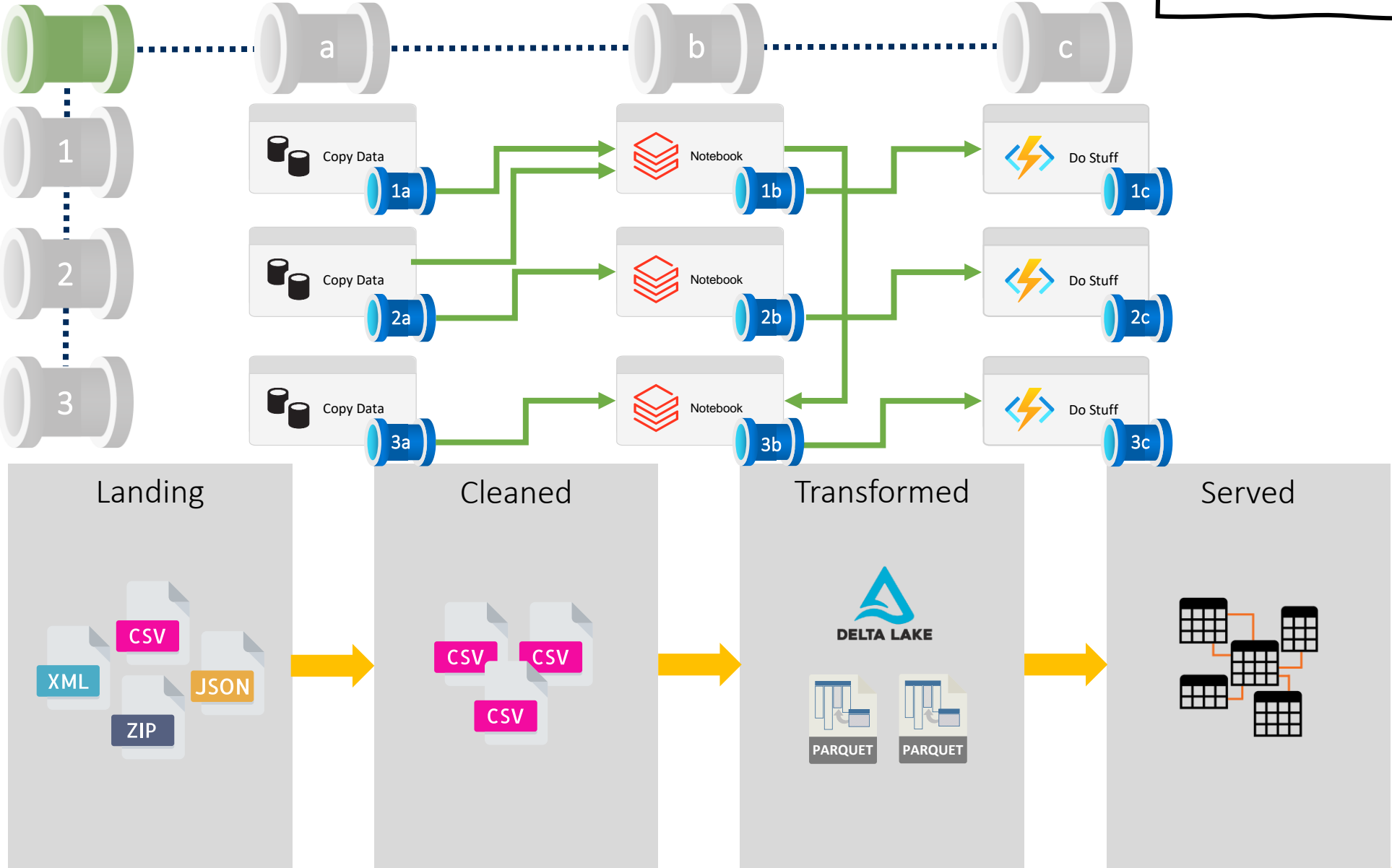
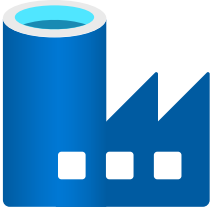


Problem

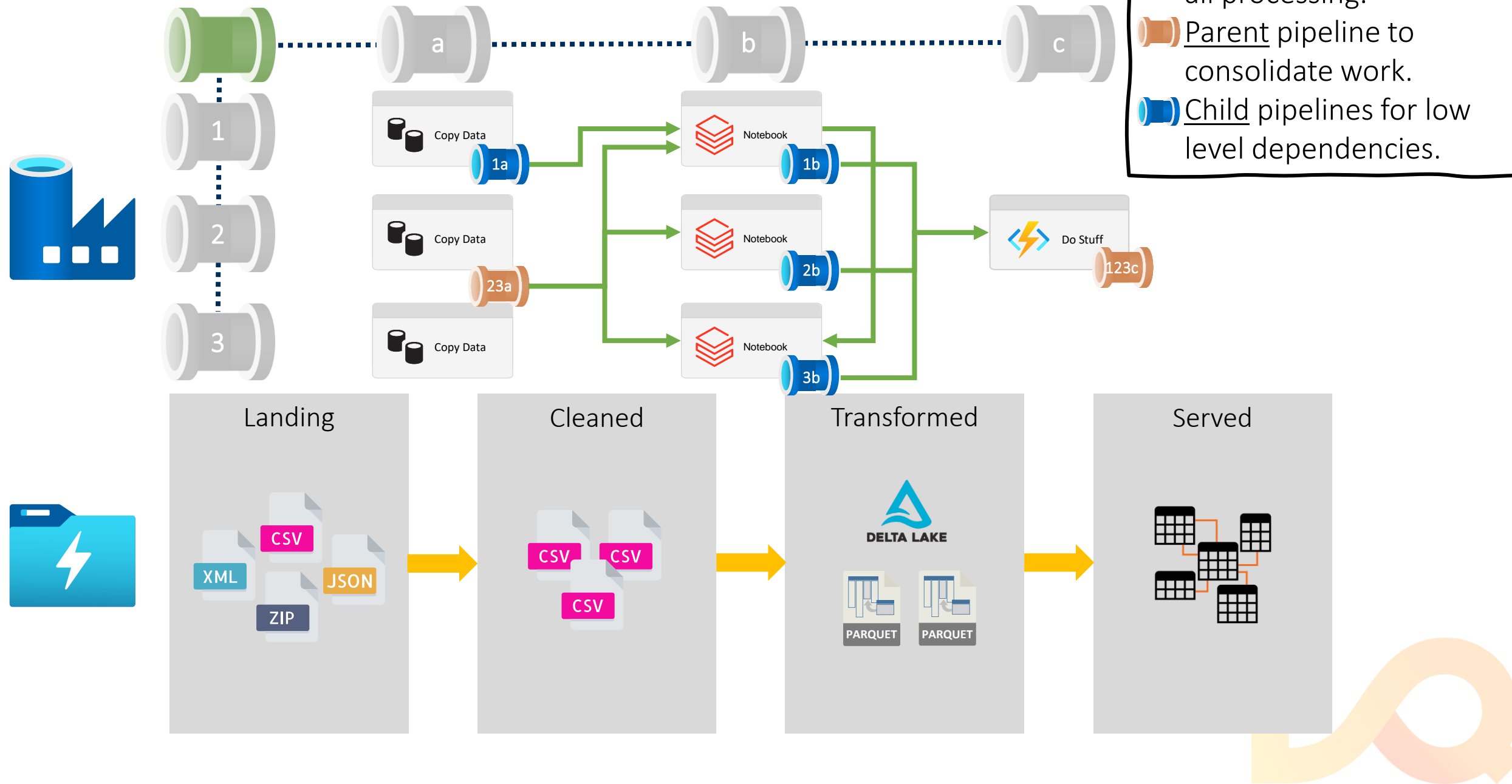


Problem

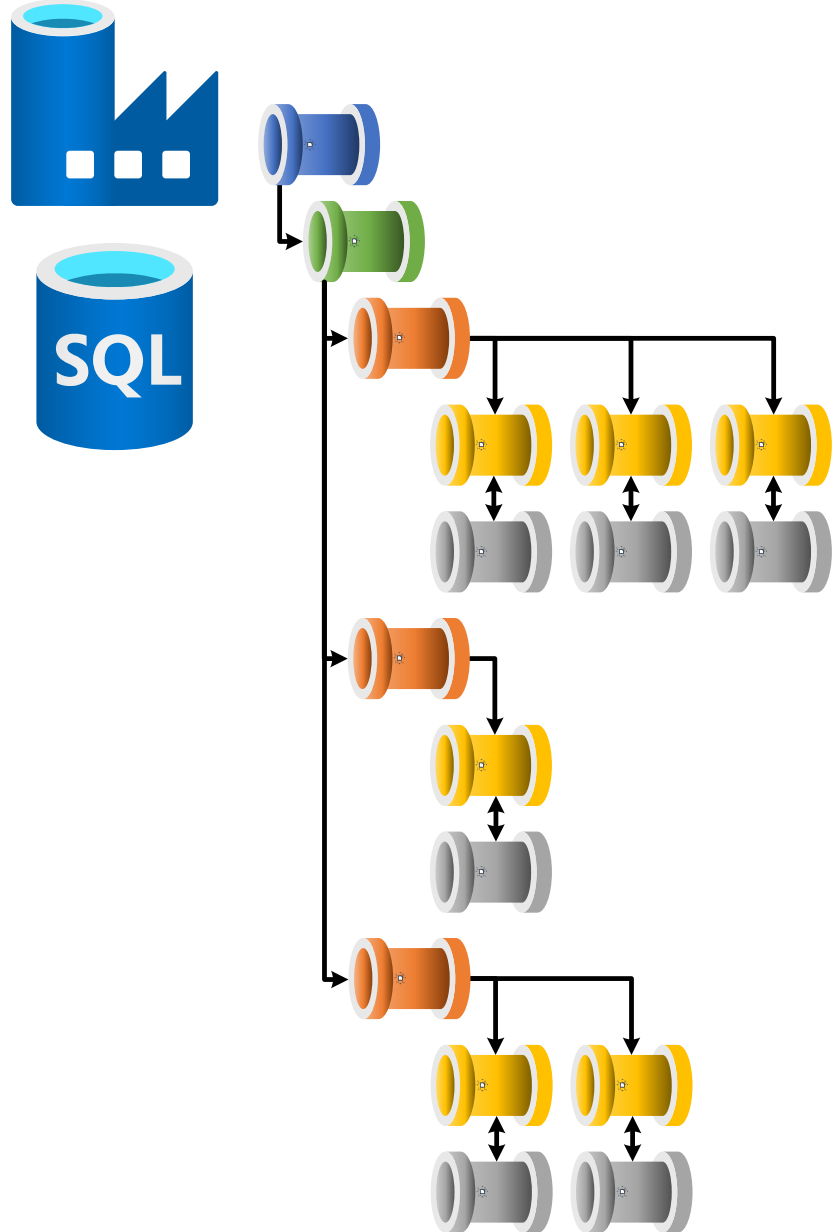
 Only 40 Activities per Pipeline.



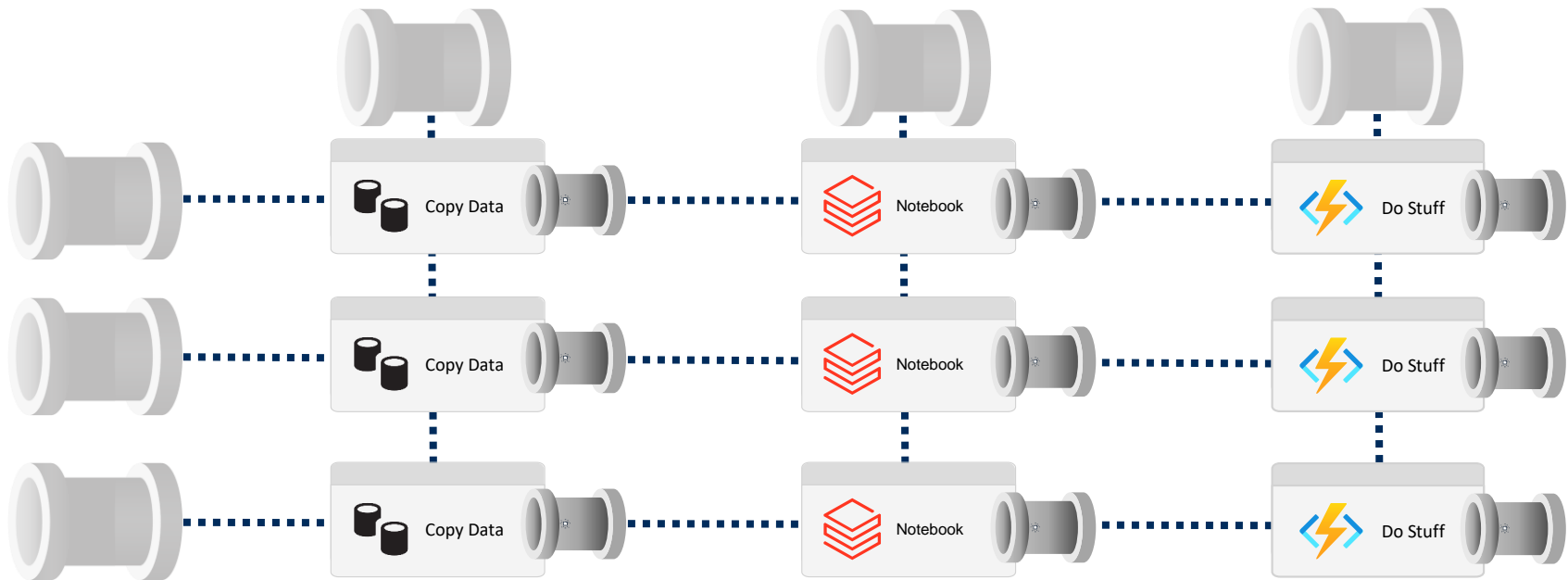
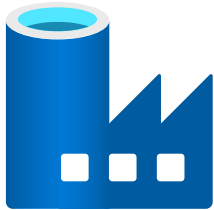
Problem



Solution: Use Metadata to Drive Data Factory Pipelines



Solution

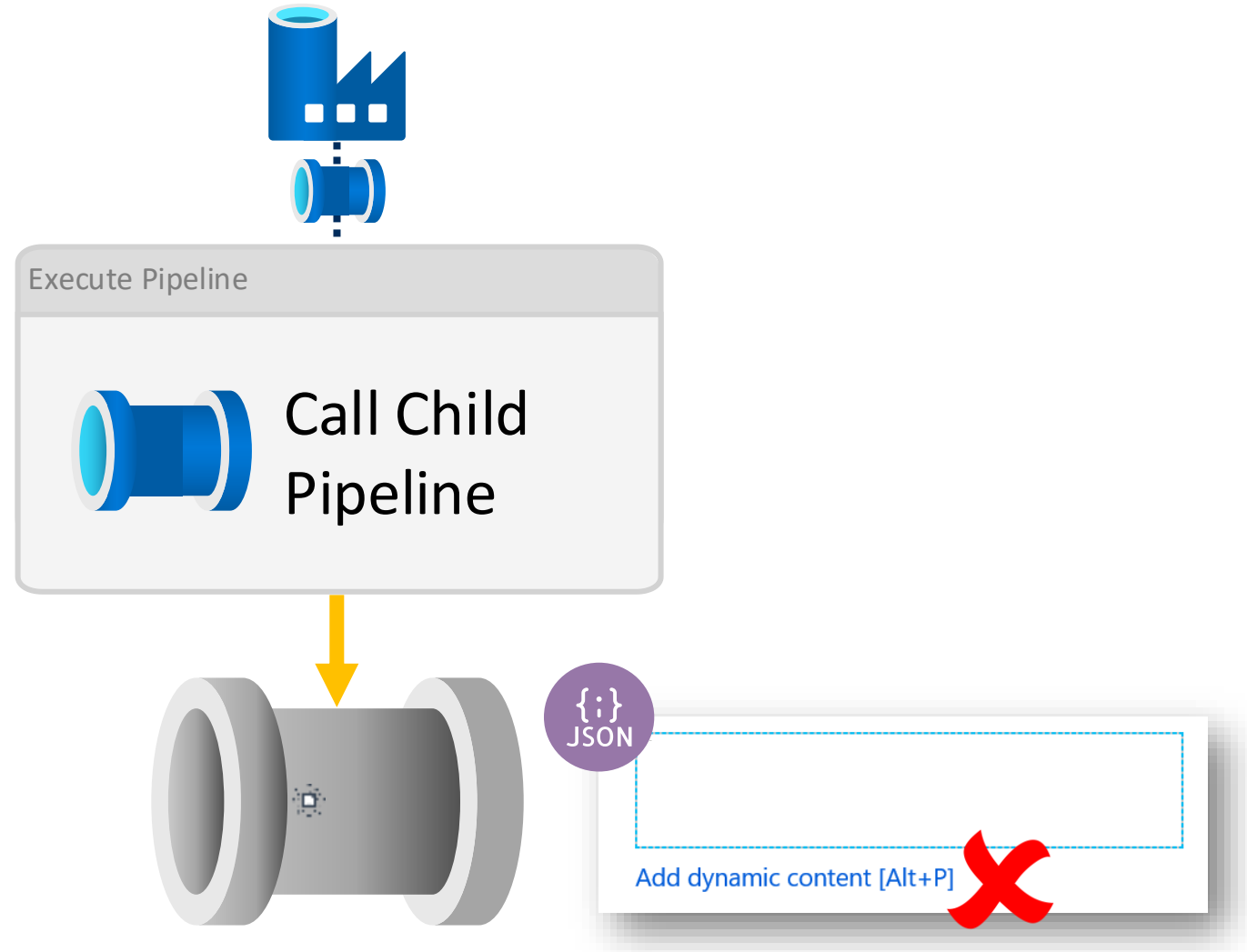


Stages	Pipelines
1	a
2	b
3	c
	d
	e
	f
	g
	h
	i

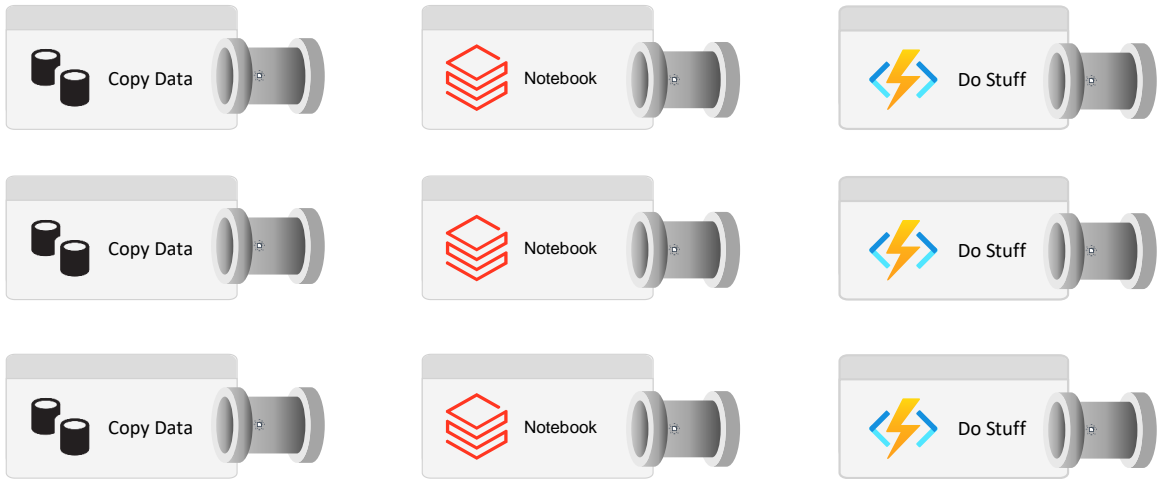
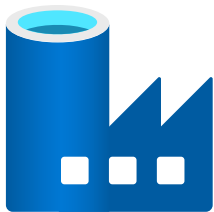
Stage	Pipeline
1	a
1	b
1	c
2	d
2	e
3	f
3	g
3	h
3	i



One More Problem



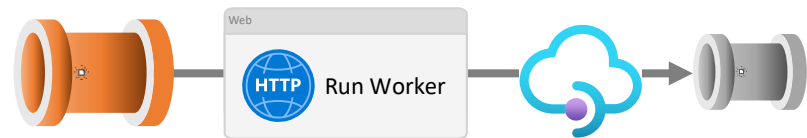
Calling Our Worker Pipelines



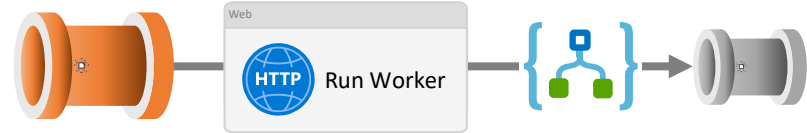
Stages	Pipelines
1	a
2	b
3	c
	d
	e
	f
	g
	h
	i

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

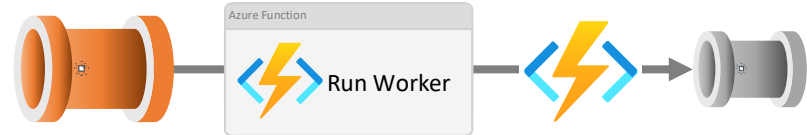
Option 1:



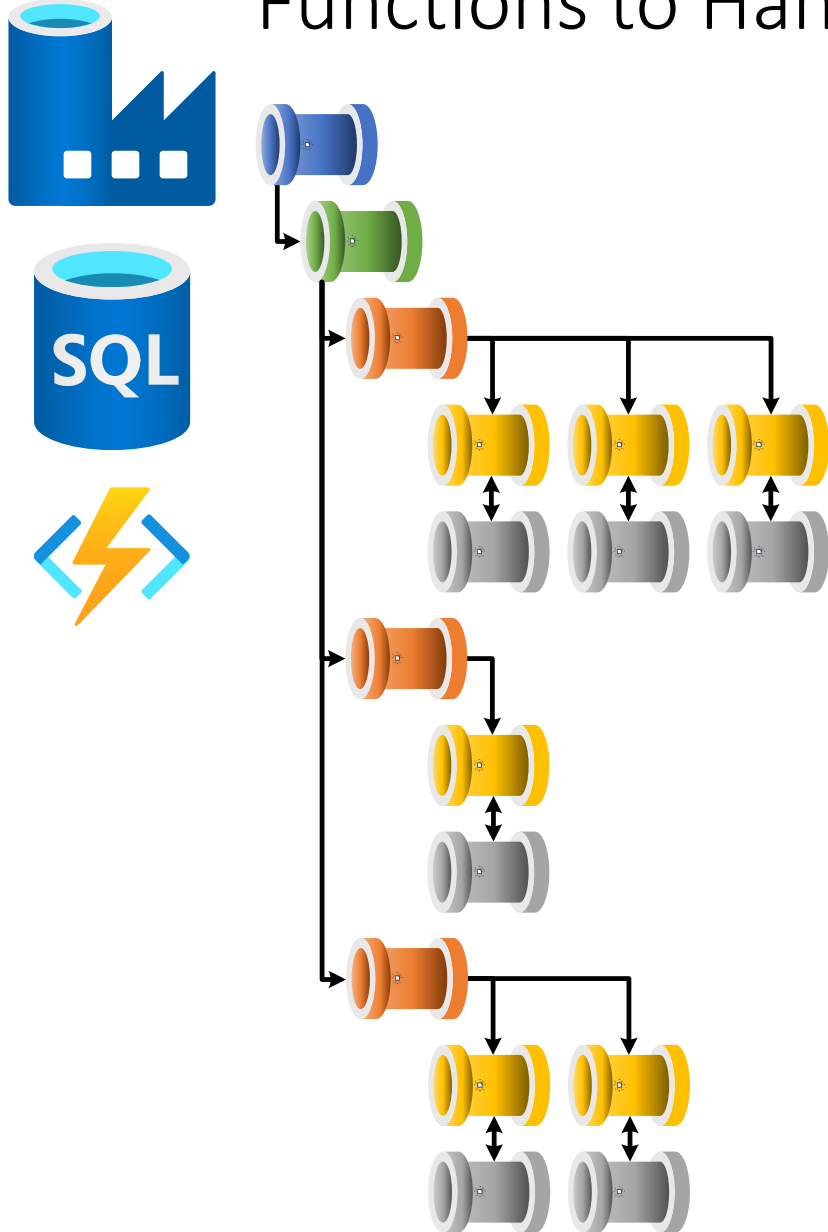
Option 2:



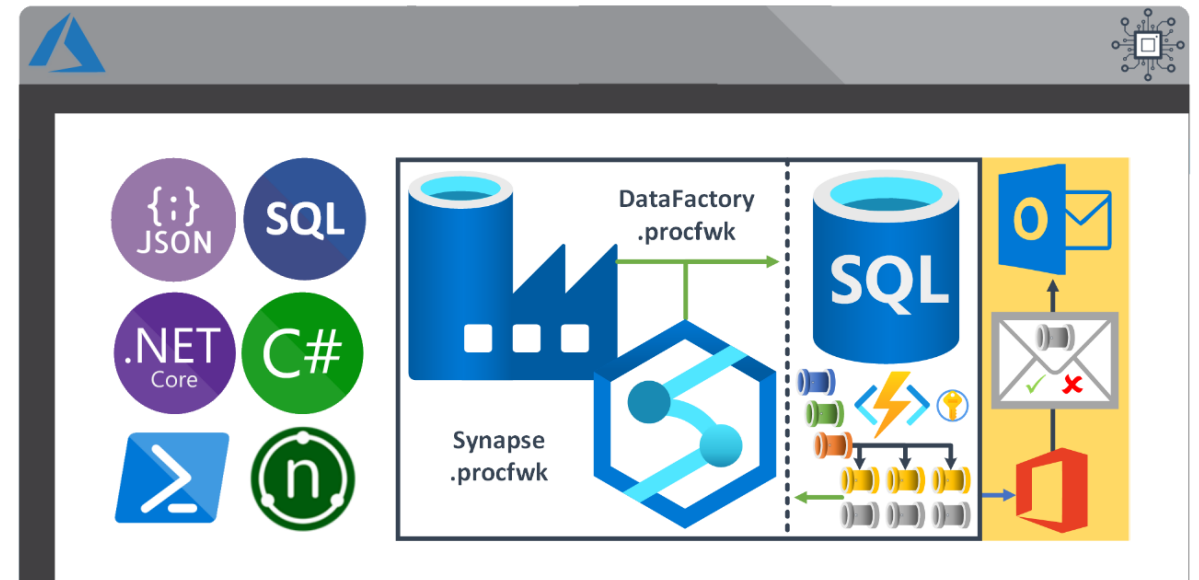
Option 3:



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



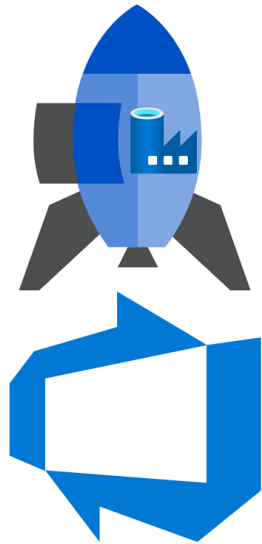
procfwk.com



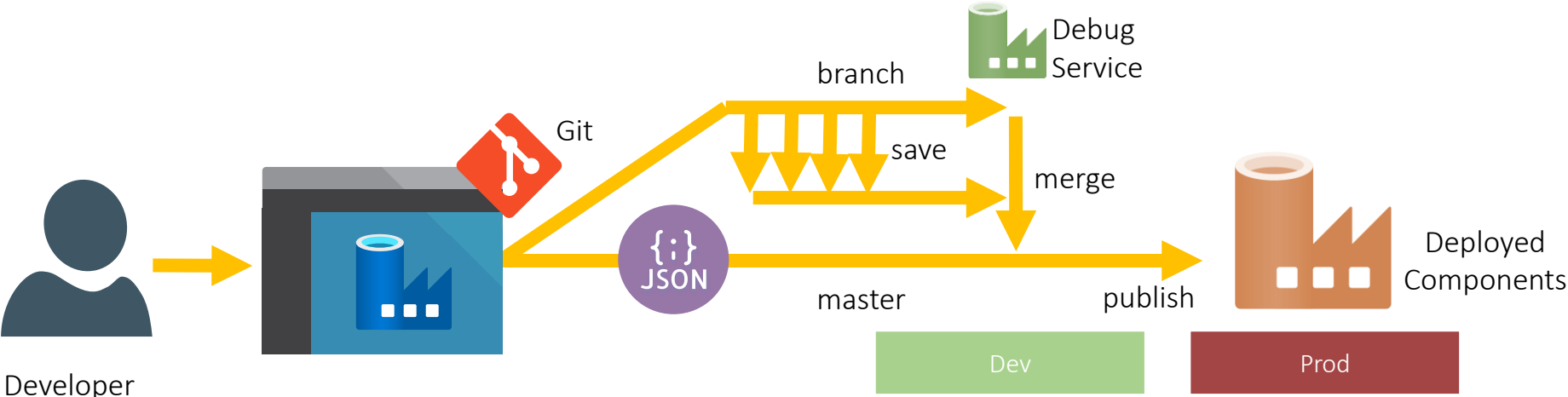
github.com/mrpaulandrew/procfwk



Deploying Data Factory



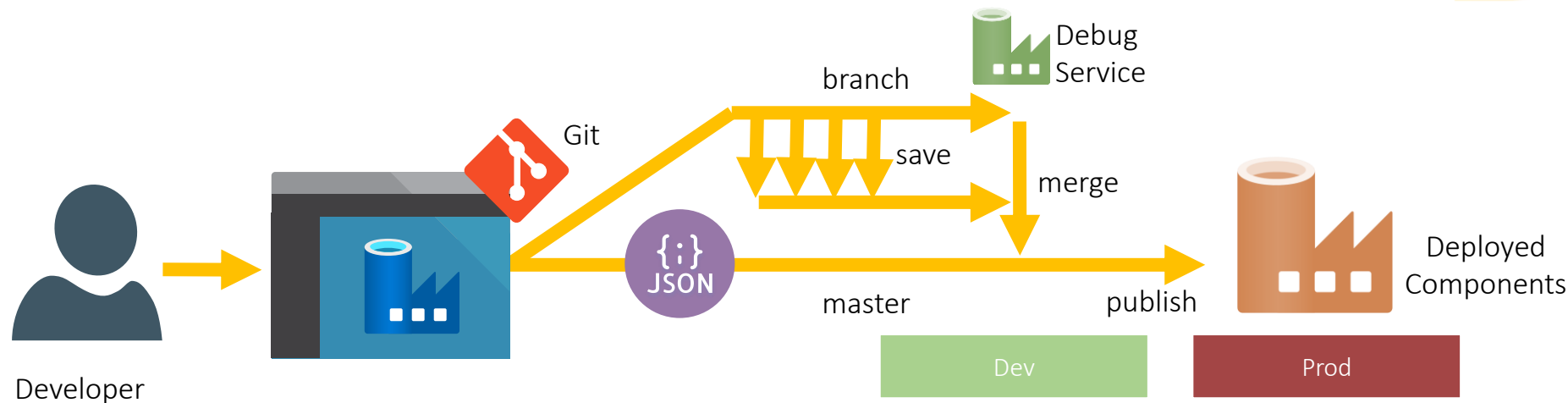
Deploying Data Factory



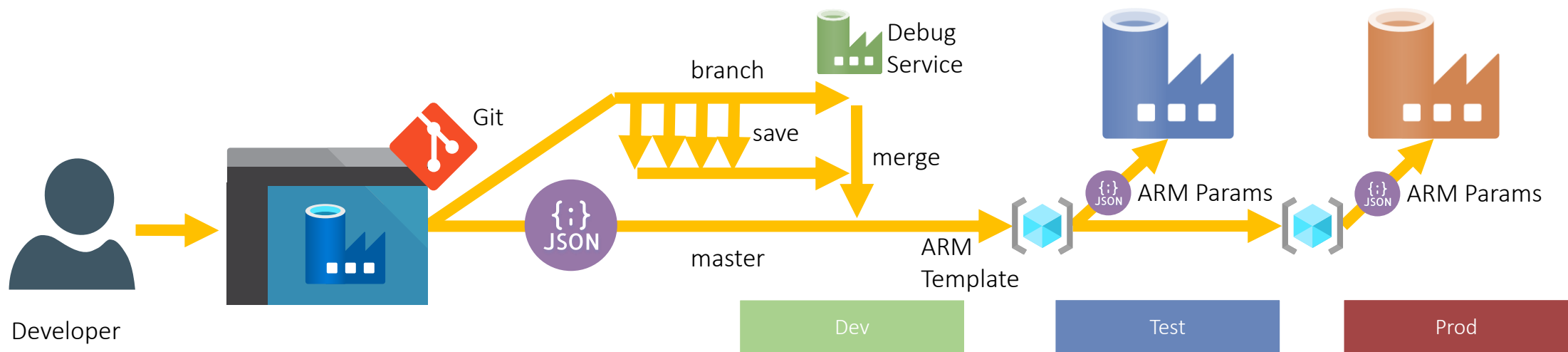
Deploying Data Factory

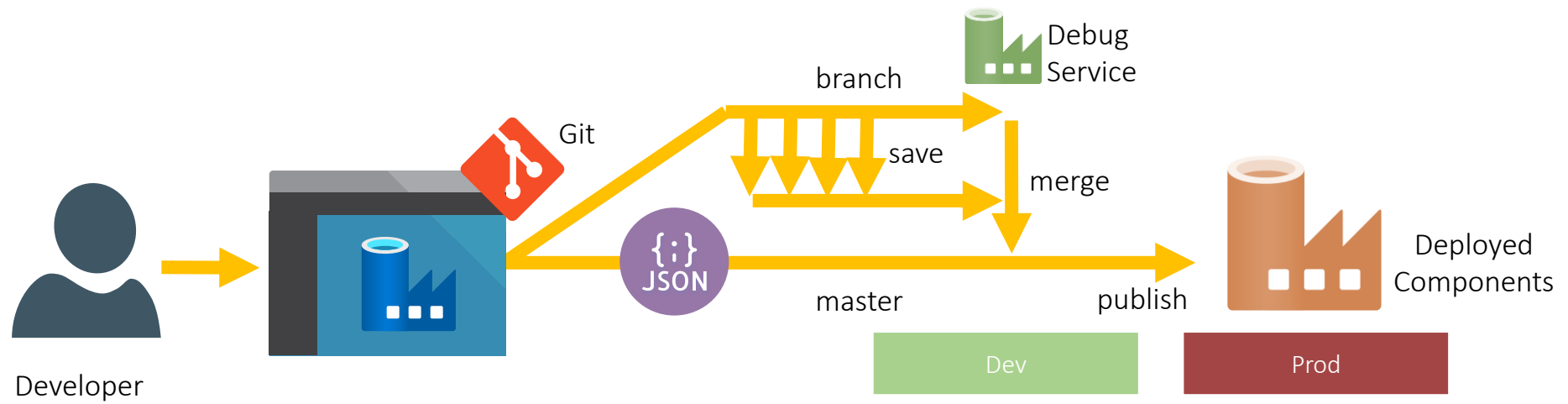


Option 1 – Single Data Factory Service

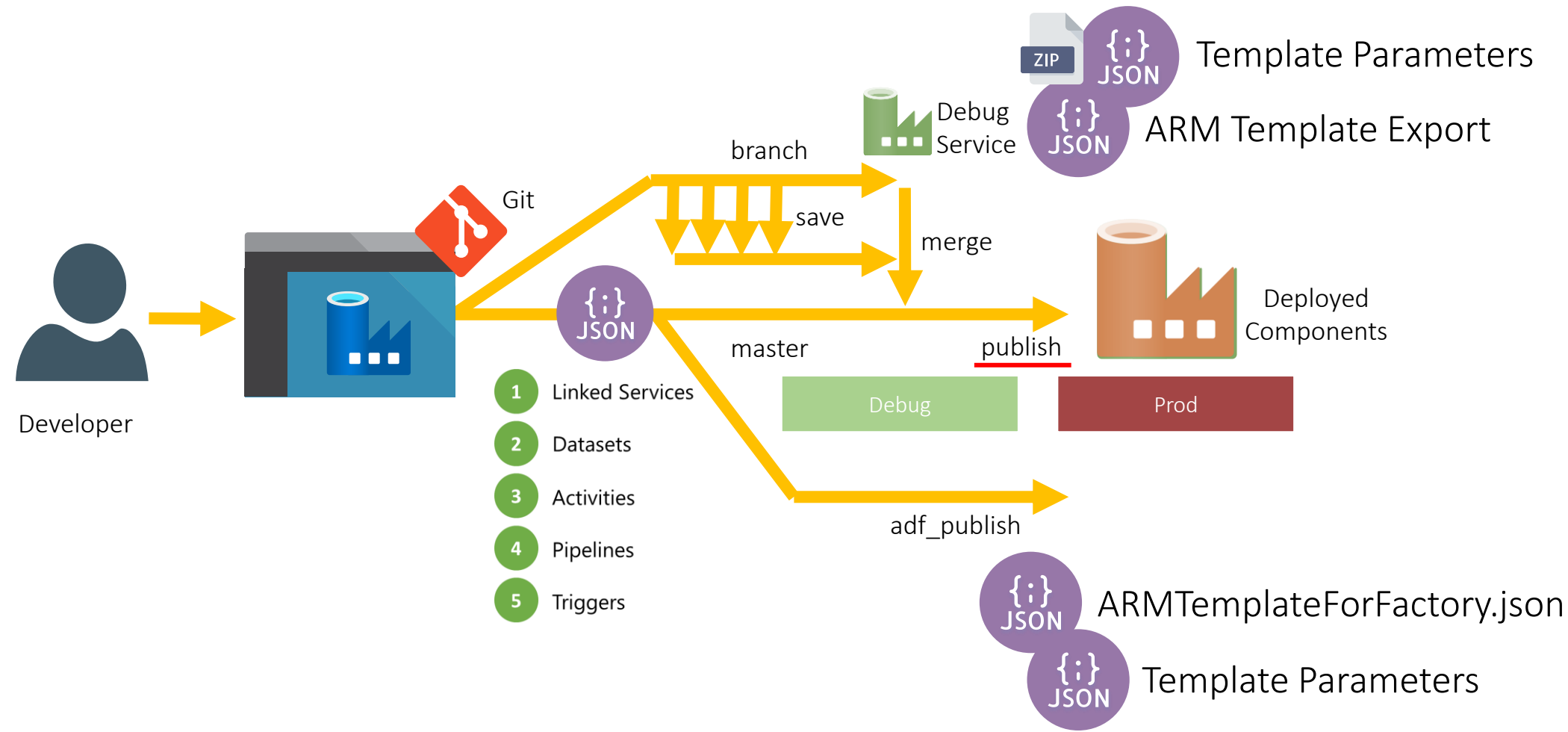


Option 2 – ARM Templates for Multiple Data Factory Services

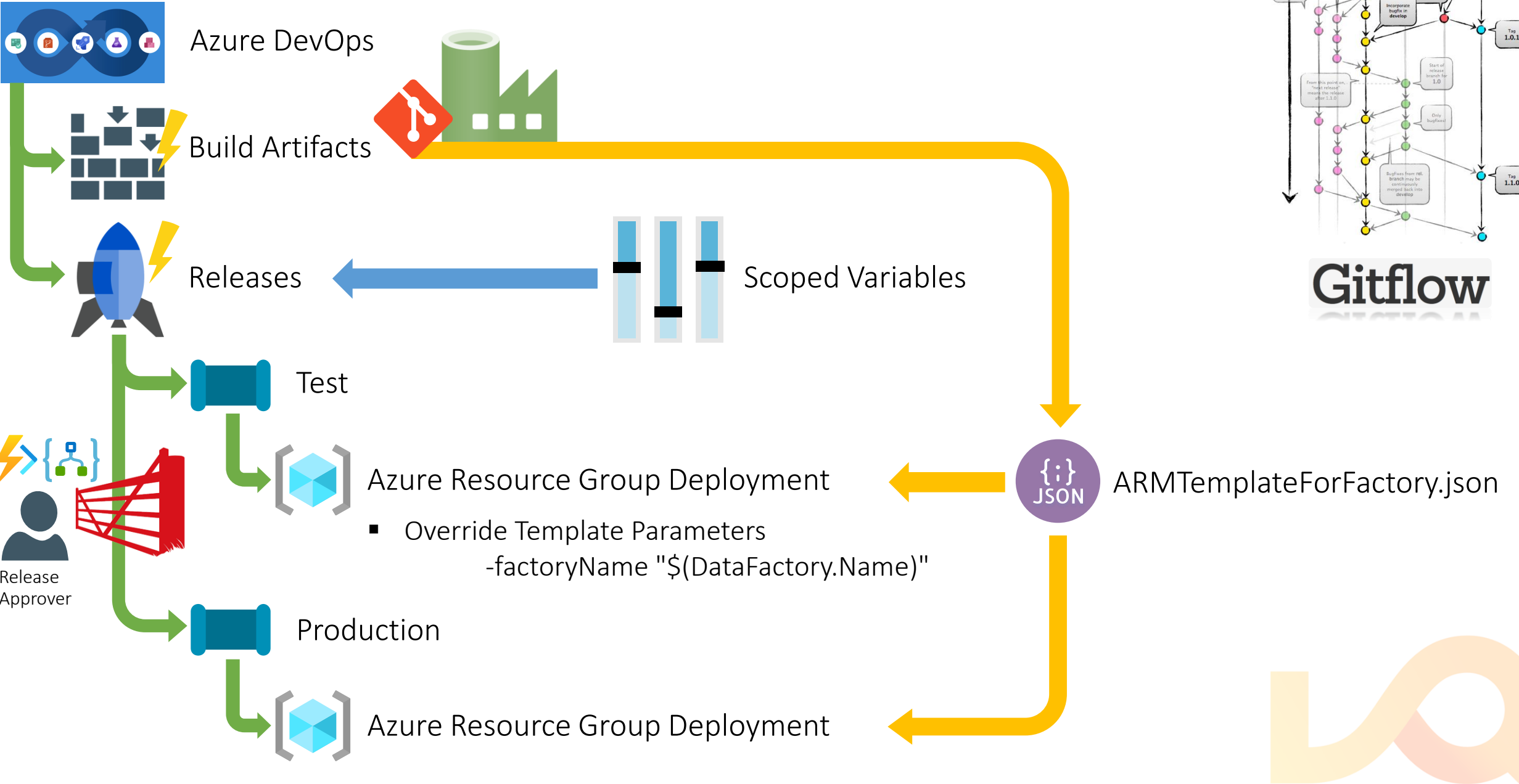




Getting Our ADF Source Code

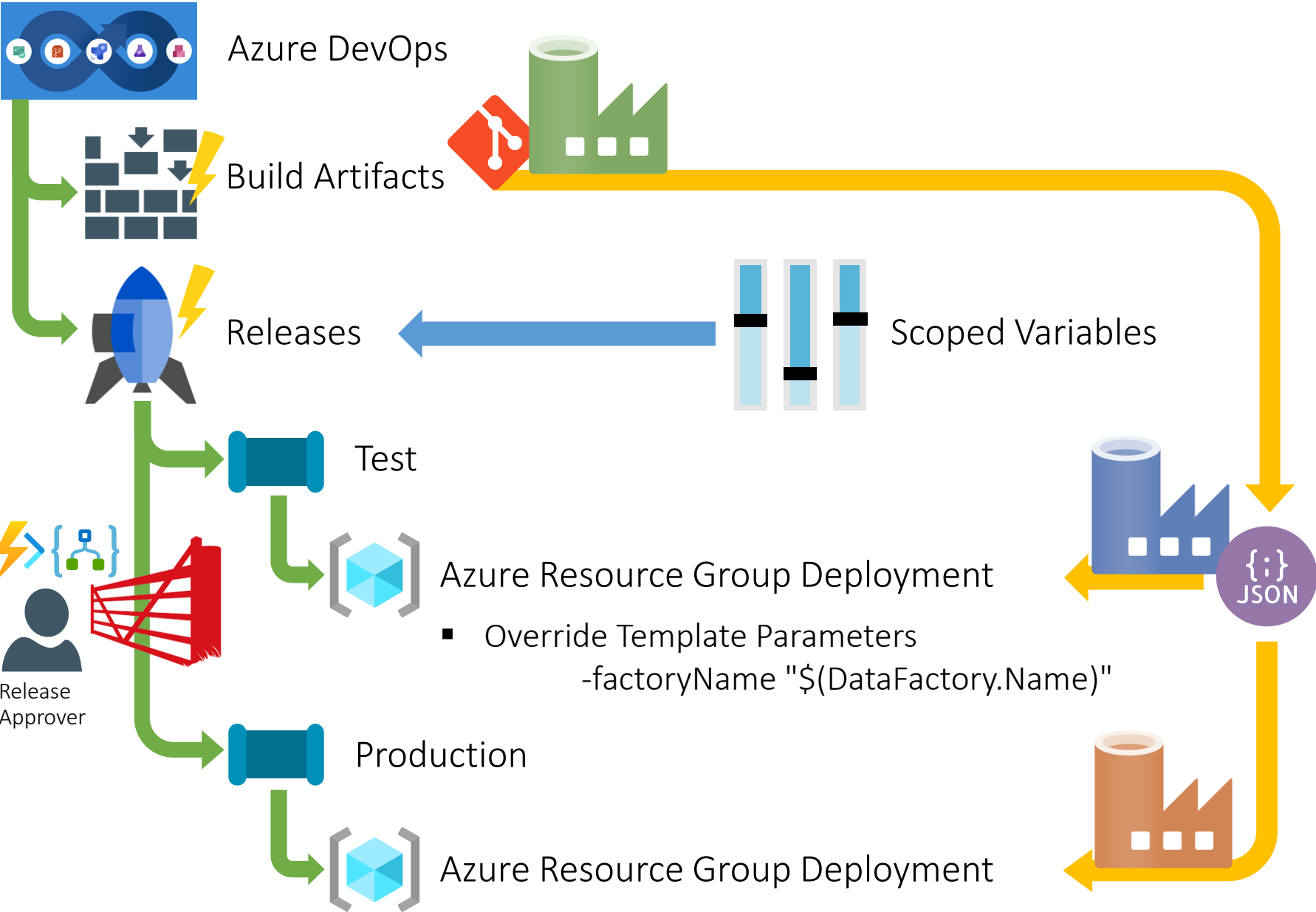


Data Factory Continuous Delivery - Option 2



Data Factory Continuous Delivery - Option 2

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



Azure Resource Group Deployment

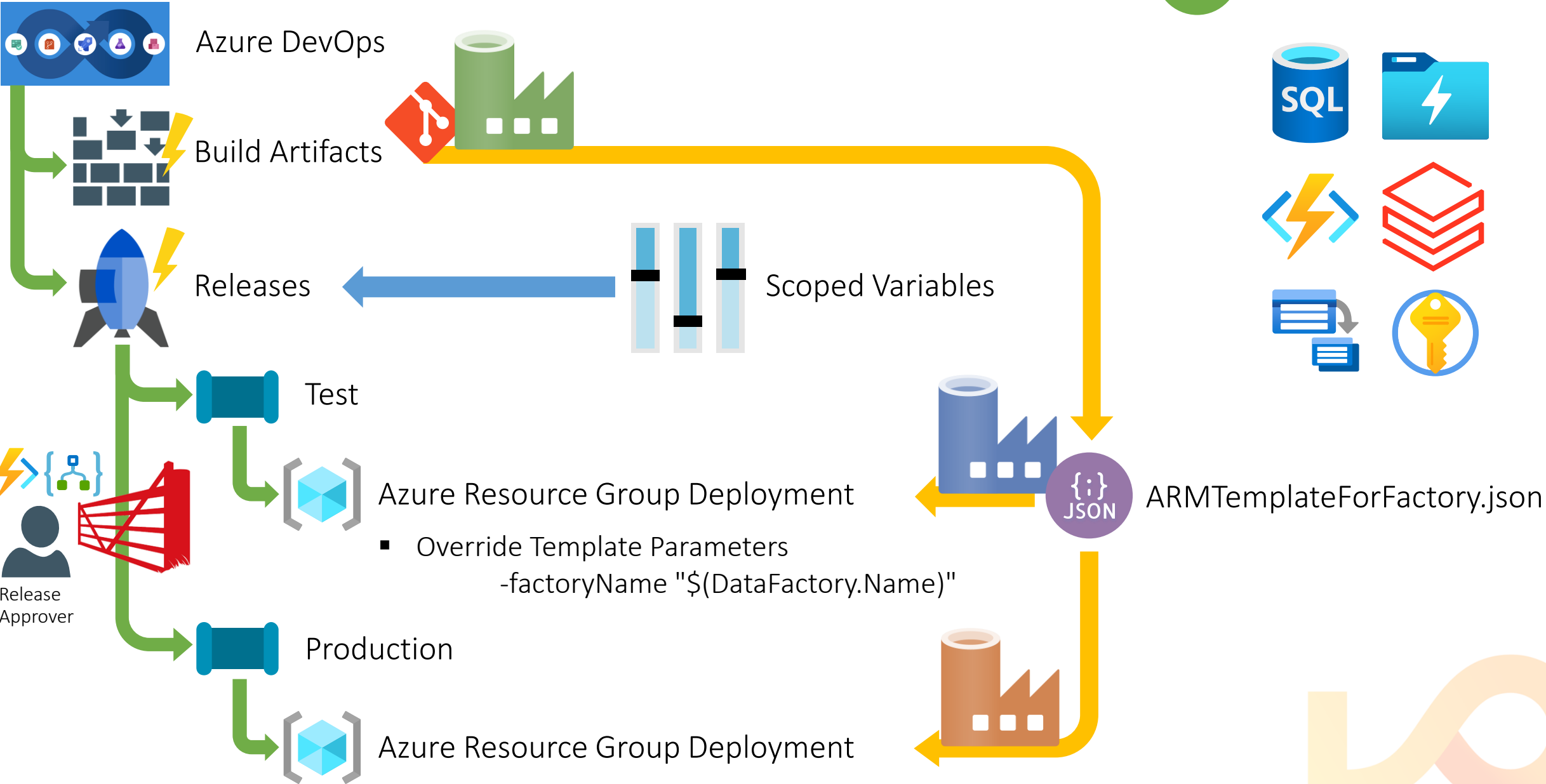
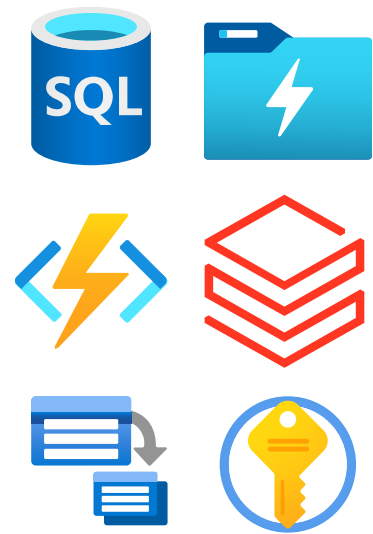
- Override Template Parameters
-factoryName "\$(DataFactory.Name)"



Data Factory Continuous Delivery - Option 2

1

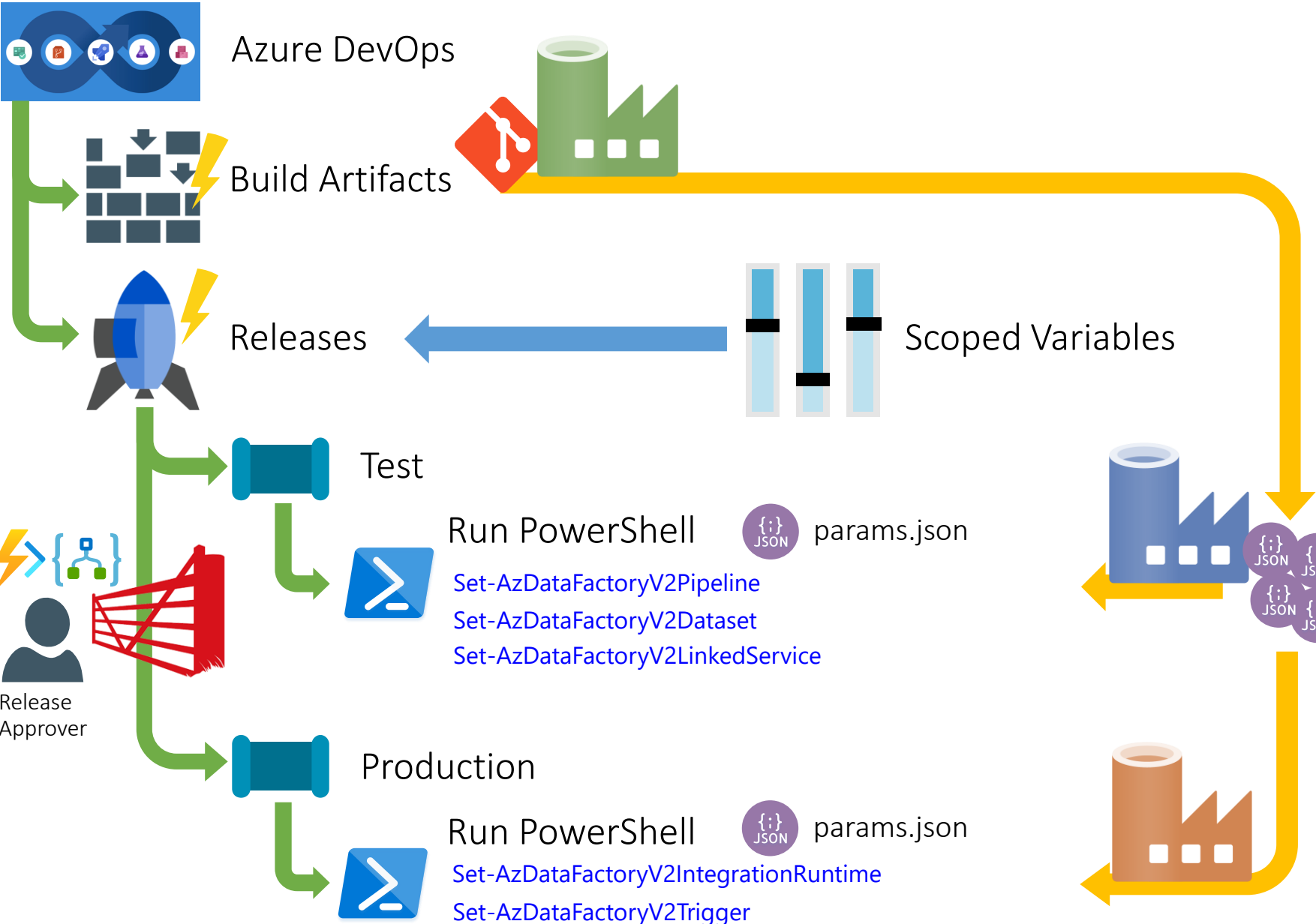
Linked Services



- Option 3

Data Factory Continuous Delivery - Option 3

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

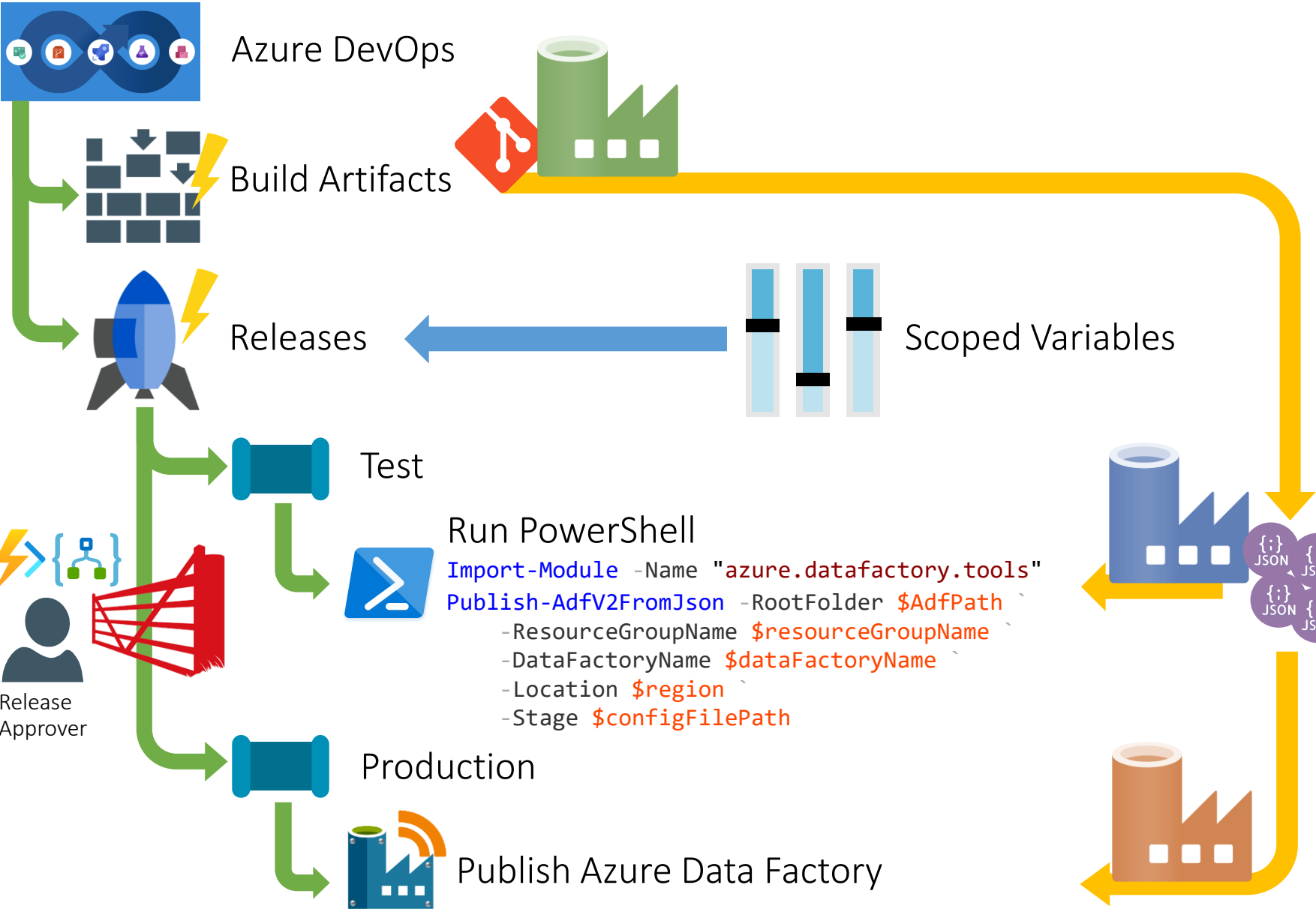


Caveats

- 1) Handle own dependencies.
- 2) Handle own removals.

Data Factory Continuous Delivery - Option 4

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



Deployment Options Summary

Option 1 – Use a single Data Factory service.

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

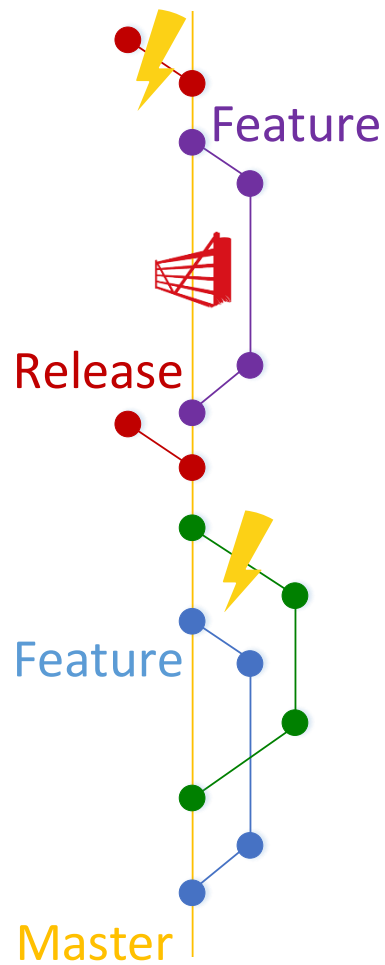
Option 3 – Use PowerShell cmdlets for each Data Factory JSON artifact.

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



Data Factory DevOps Story Summary

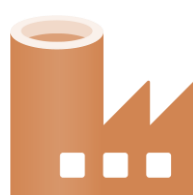
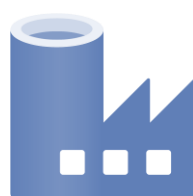
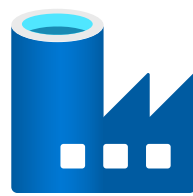
What is your code branching strategy?



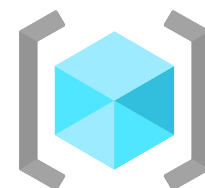
Which source control tool to use?



How many environments do we want?




What deployment method do we want to use?



What artifacts are we going to use?...

OR

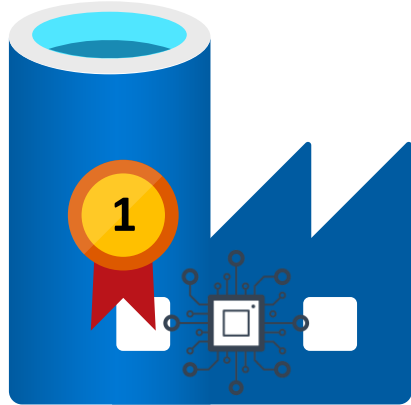
How much control do you want?

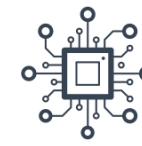
 ARMTemplate
ForFactory.json

 linkedservices.json
pipelines &
activities.json
datasets.json
triggers.json



Best Practices





Key Points

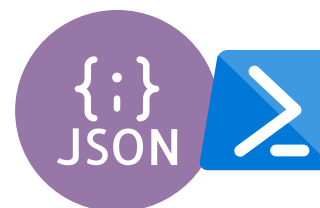
- 00 Environment Setup & Developer Debugging
- 00 Deployments
- 00 Automated Testing
- 00 Naming Conventions
- 00 Pipeline Hierarchies
- 00 Pipeline & Activity Descriptions
- 00 Factory Component Folders
- 00 Linked Service Security via Azure Key Vault
- 00 Dynamic Linked Services
- 00 Generic Datasets

- 00 Metadata Driven Processing
- 00 Parallel Execution
- 00 Hosted Integration Runtimes
- 00 Azure Integration Runtimes
- 00 Wider Platform Orchestration
- 00 Custom Error Handler Paths
- 00 Monitoring via Log Analytics
- 00 Service Limitations
- 00 Using Templates
- 00 Documentation





DEMO



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... Activity(ies) with timeout values still set to the service default value of 7 days.
Running check... Activity(ies) without a description value.
Running check... Activity(ies) ForEach iteration without a batch count value set.
Running check... Activity(ies) 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	Activity(ies) with timeout values still set to the service default value of 7 days.	High
35	Activity(s) without a description value.	Low
2	Activity(s) ForEach iteration without a batch count value set.	High
3	Activity(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

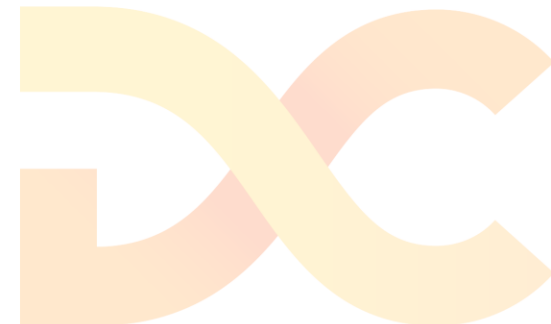


Thank you for listening...

Paul Andrew



altius



Blog: mrpaulandrew.com

YouTube: [c/mrpaulandrew](https://www.youtube.com/c/mrpaulandrew)

Email: paul@mrpaulandrew.com

Twitter: [@mrpaulandrew](https://twitter.com/mrpaulandrew)

LinkedIn: [In/mrpaulandrew](https://www.linkedin.com/in/mrpaulandrew)

GitHub: github.com/mrpaulandrew

[/CommunityEvents](#)

[/ContentCollateral](#)

[/procfwk](#)

STRATEGIC PARTNER



GOLD SPONSOR



SILVER SPONSOR



BRONZE SPONSOR

