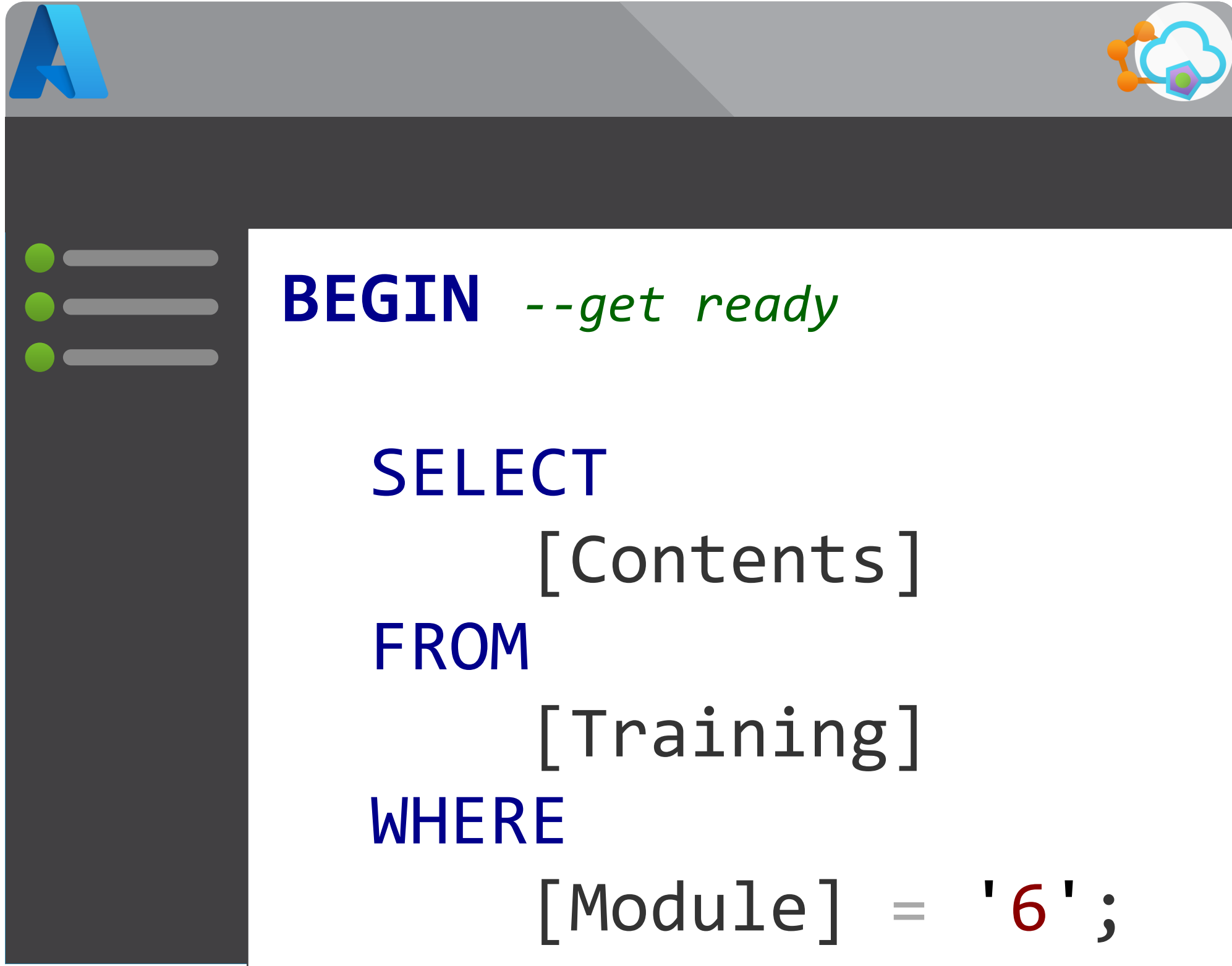


Module 6

Execution Parallelism

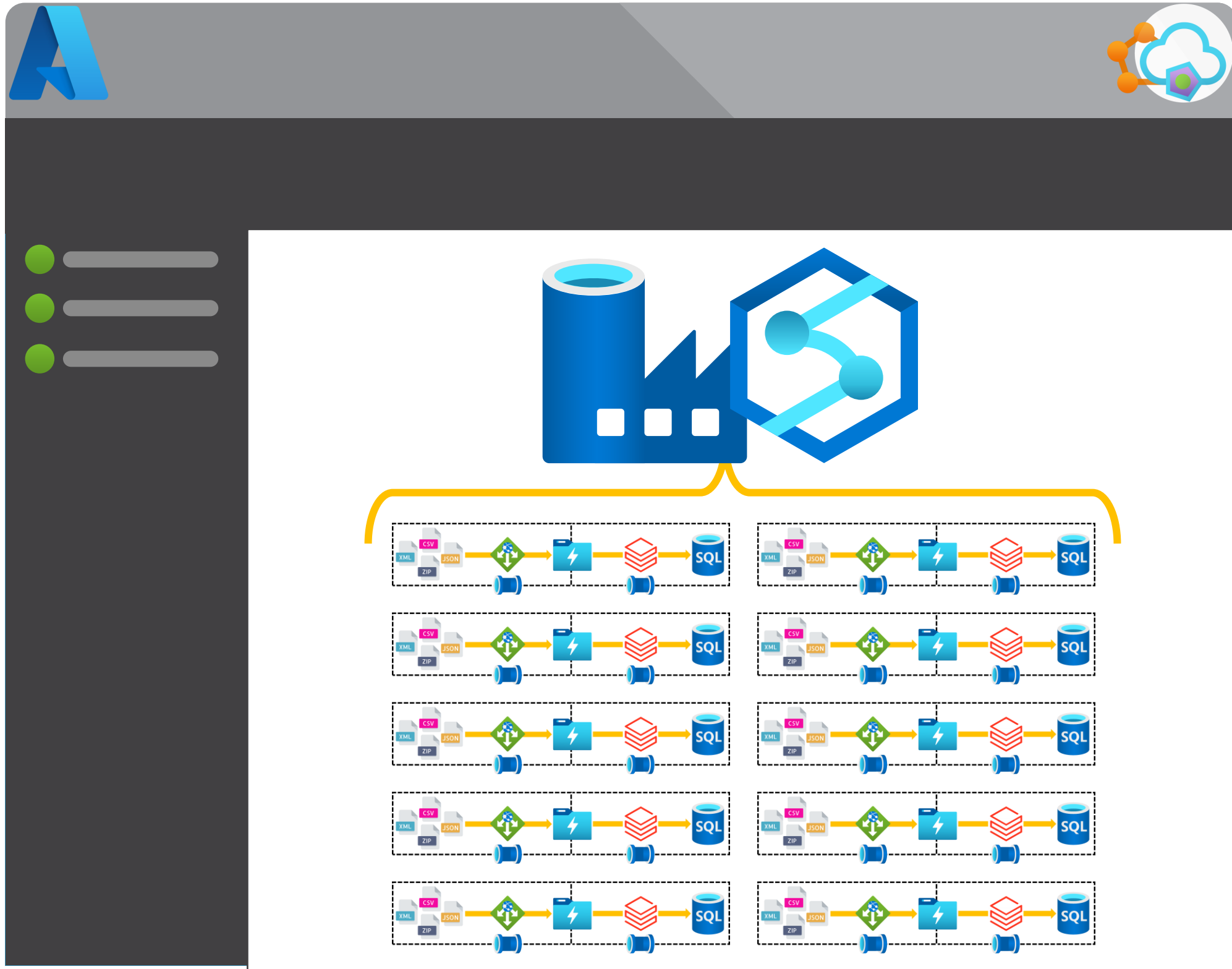


- Control Flow Scale Out
- Concurrency Limitations
- Internal vs External Activities
- Orchestration Frameworks

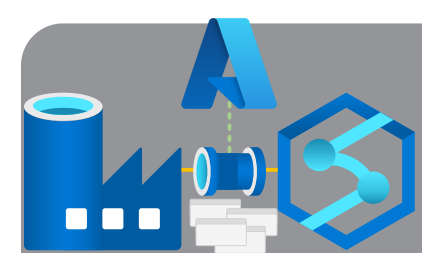
Note to self; Scale up
the database before
the demo!

Module 6

Execution Parallelism



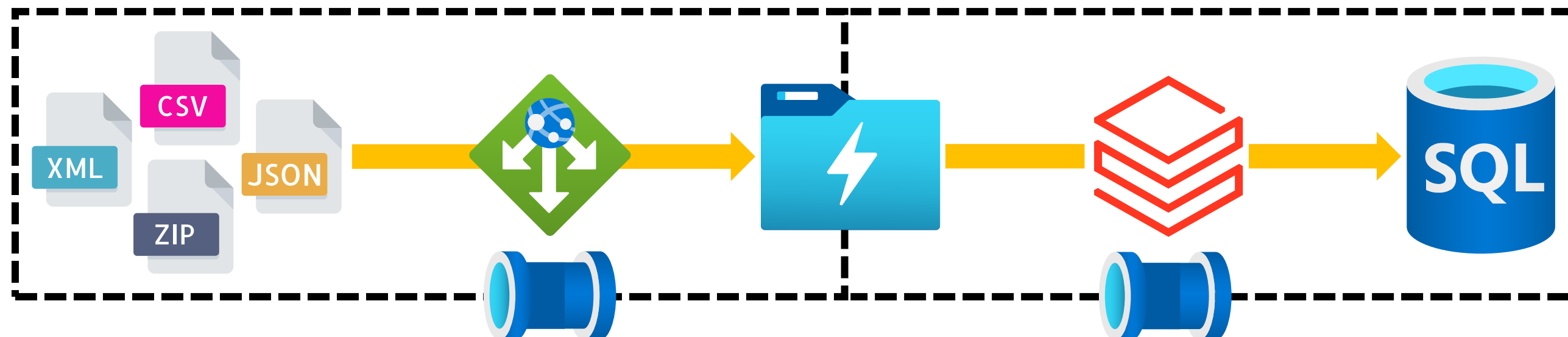
- Control Flow Scale Out
- Concurrency Limitations
- Internal vs External Activities
- Orchestration Framework - <http://procfwk.com>



Integration Pipelines as Data Engineers



Control Flow



1

Linked Services



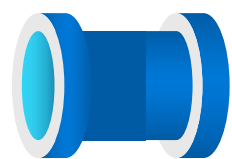
2

Datasets



3

Activities



4

Pipelines



5

Triggers



Add dynamic content [Alt+P]

Integration Runtimes

6



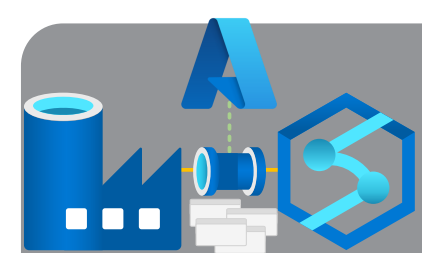
Azure IR



Hosted IR



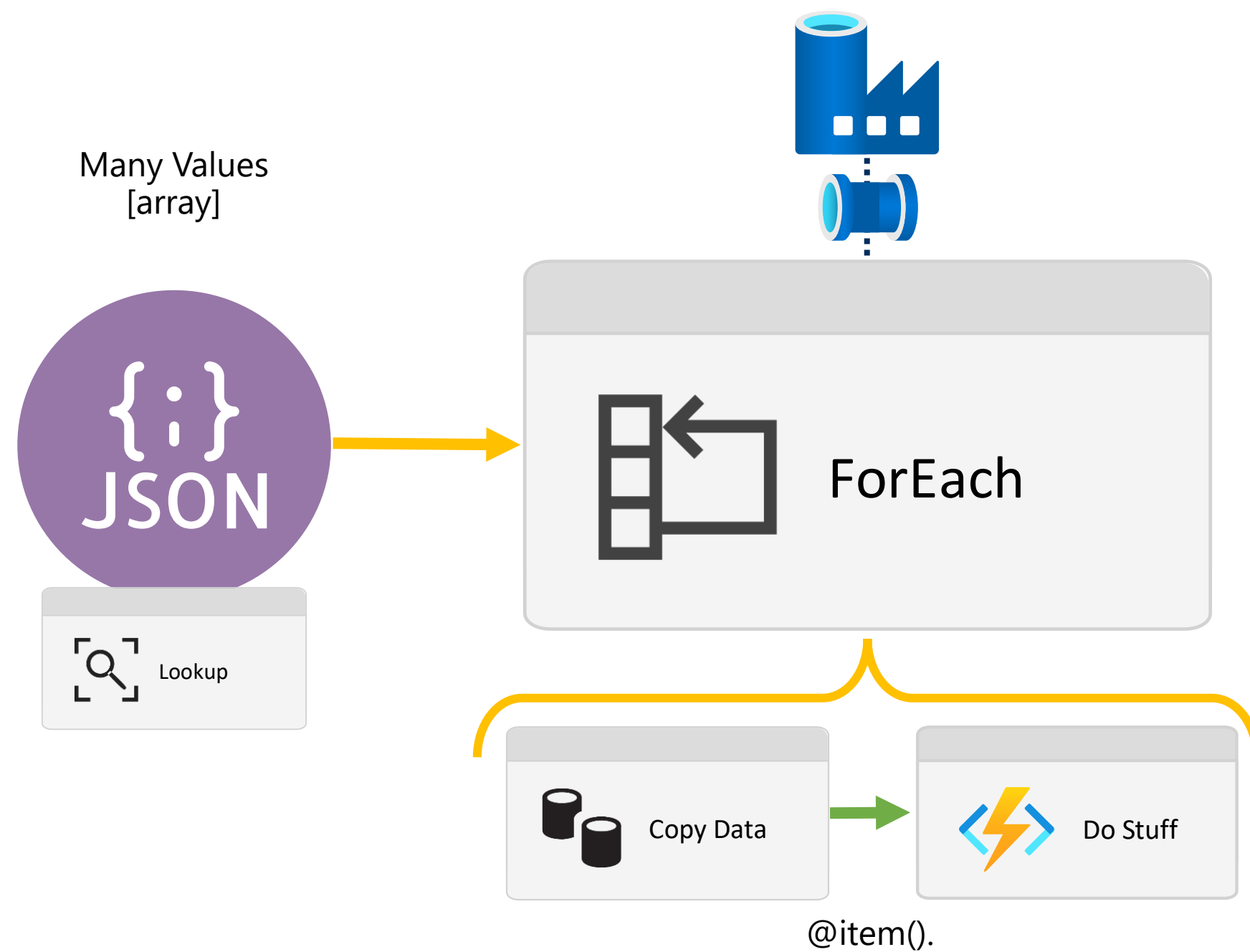
SSIS IR



For Each Activity



Scaling Out Control Flow Activities



IsSequential: true



[array]

[0]
↓
[1]
↓
[2]
↓
[3]
↓
[*i*]

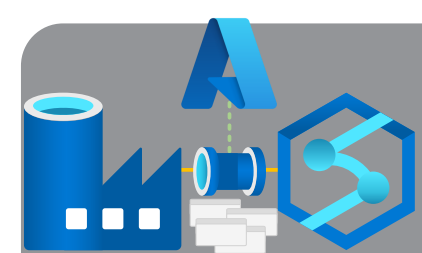


[array]

[0] [1] [2] [3] [4] [5] [6] [*i*]

Batch Count Default: 20

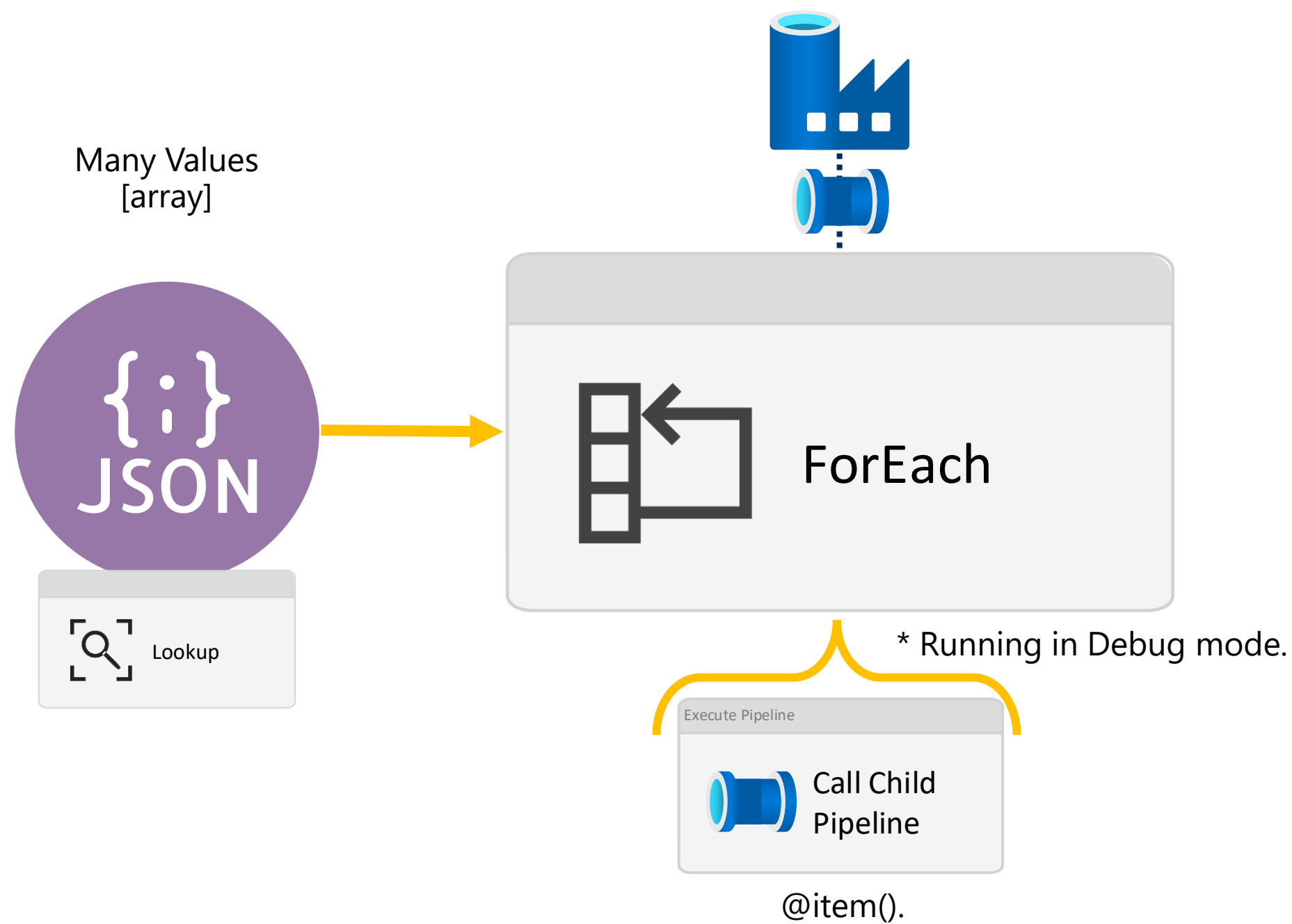
Batch Count Max: 50



For Each Activity



Scaling Out Control Flow Activities



IsSequential: true



[array]

[0]

[1]

[2]

[3]

[i]



[array]

[0]

[1]

[2]

[3]

[4]

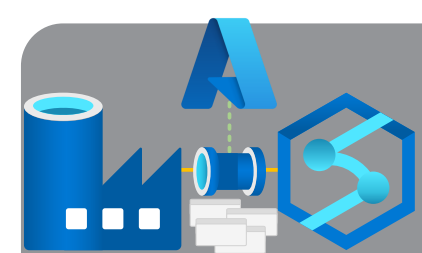
[5]

[6]

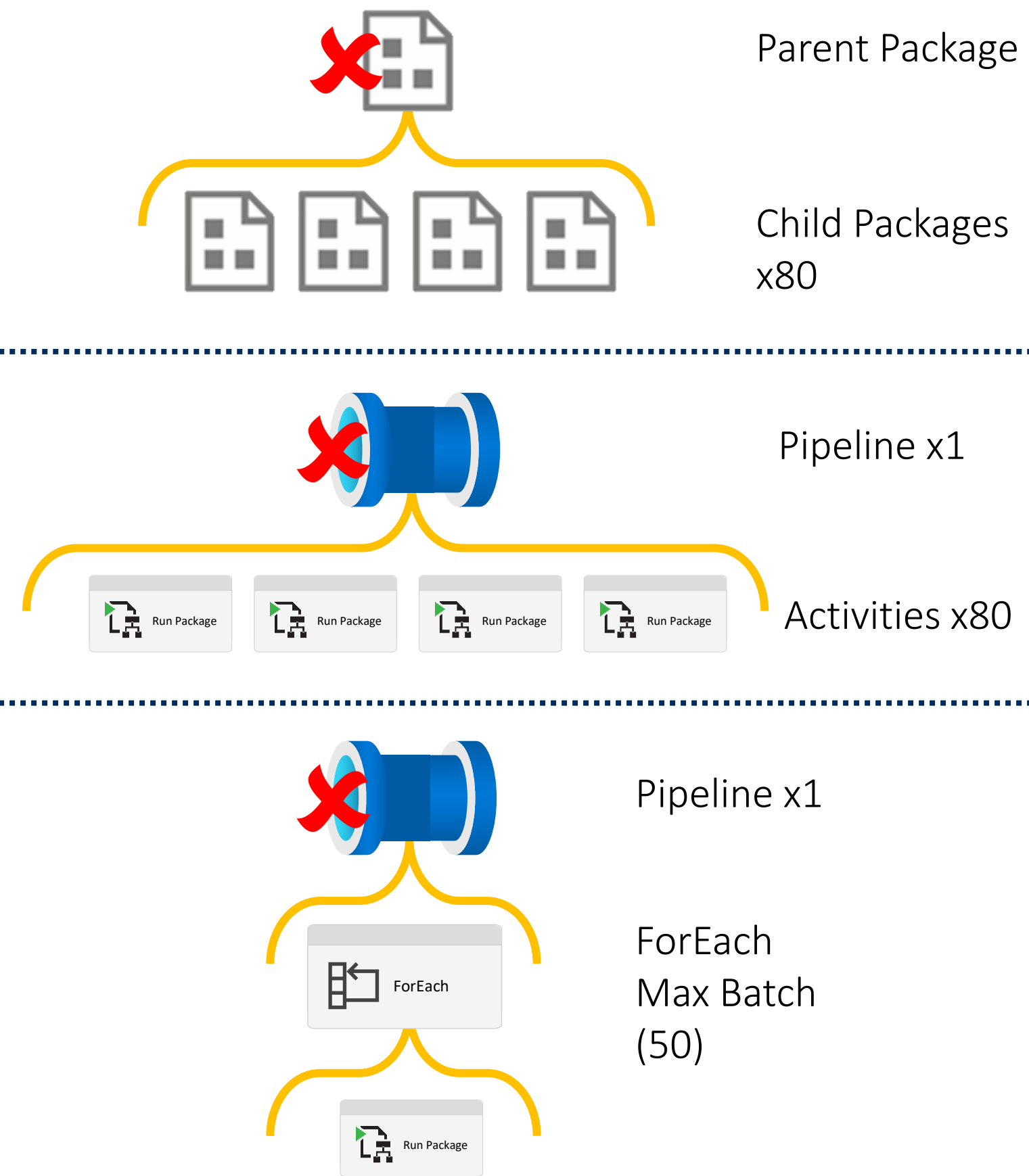
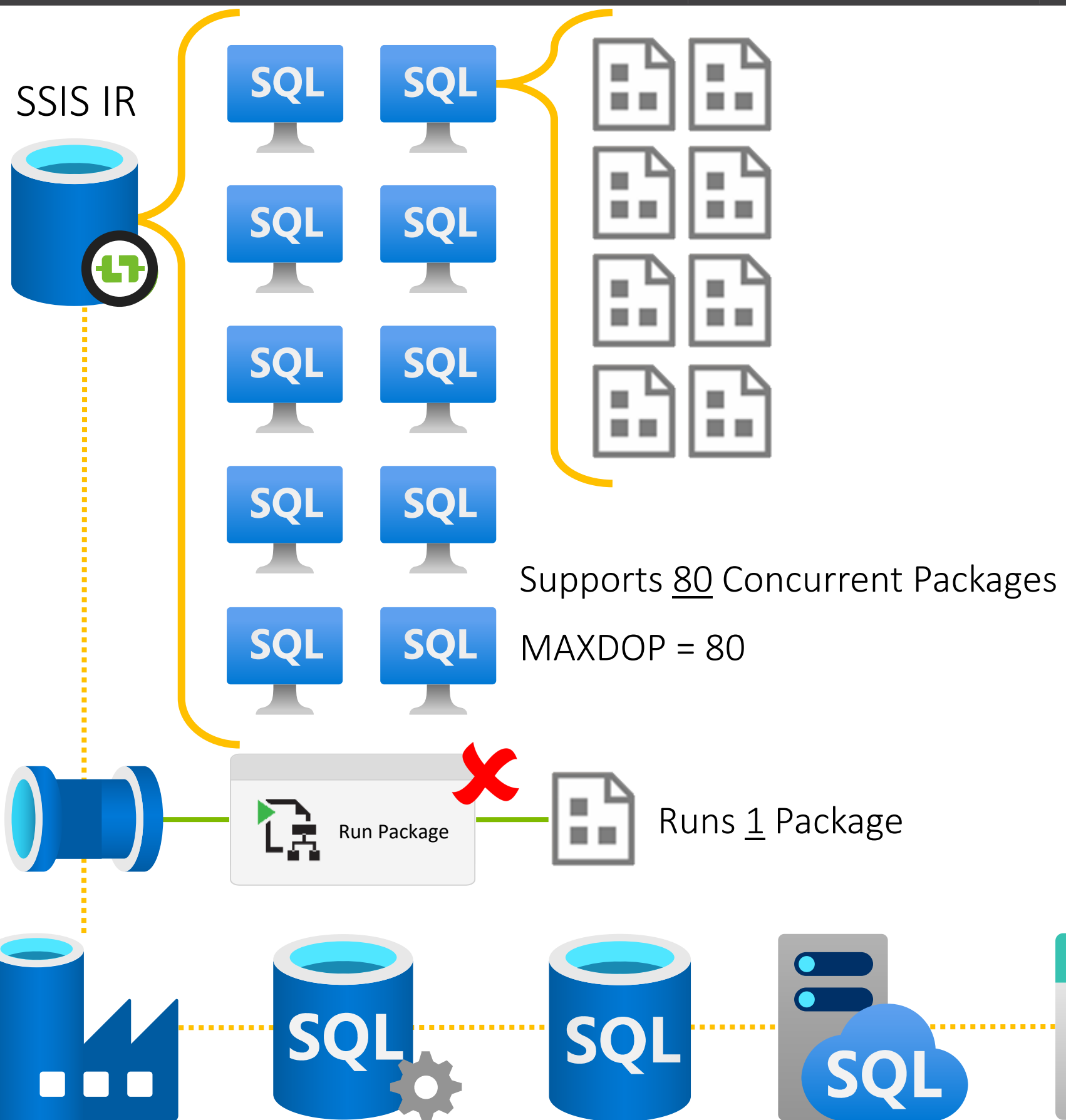
[i]

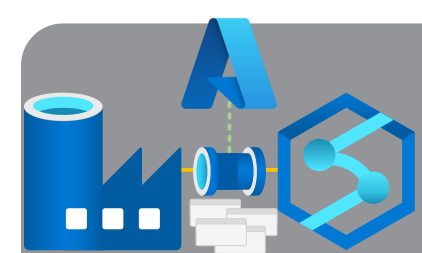
Batch Count Default: 20

Batch Count Max: 50



Problem: Using All Of The SSIS IR Compute

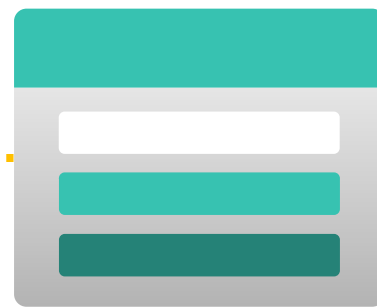
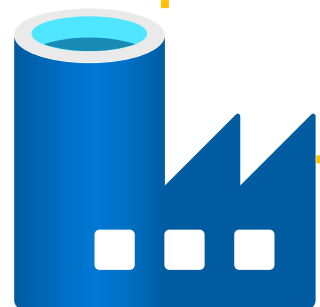
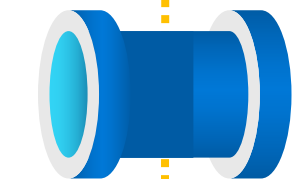
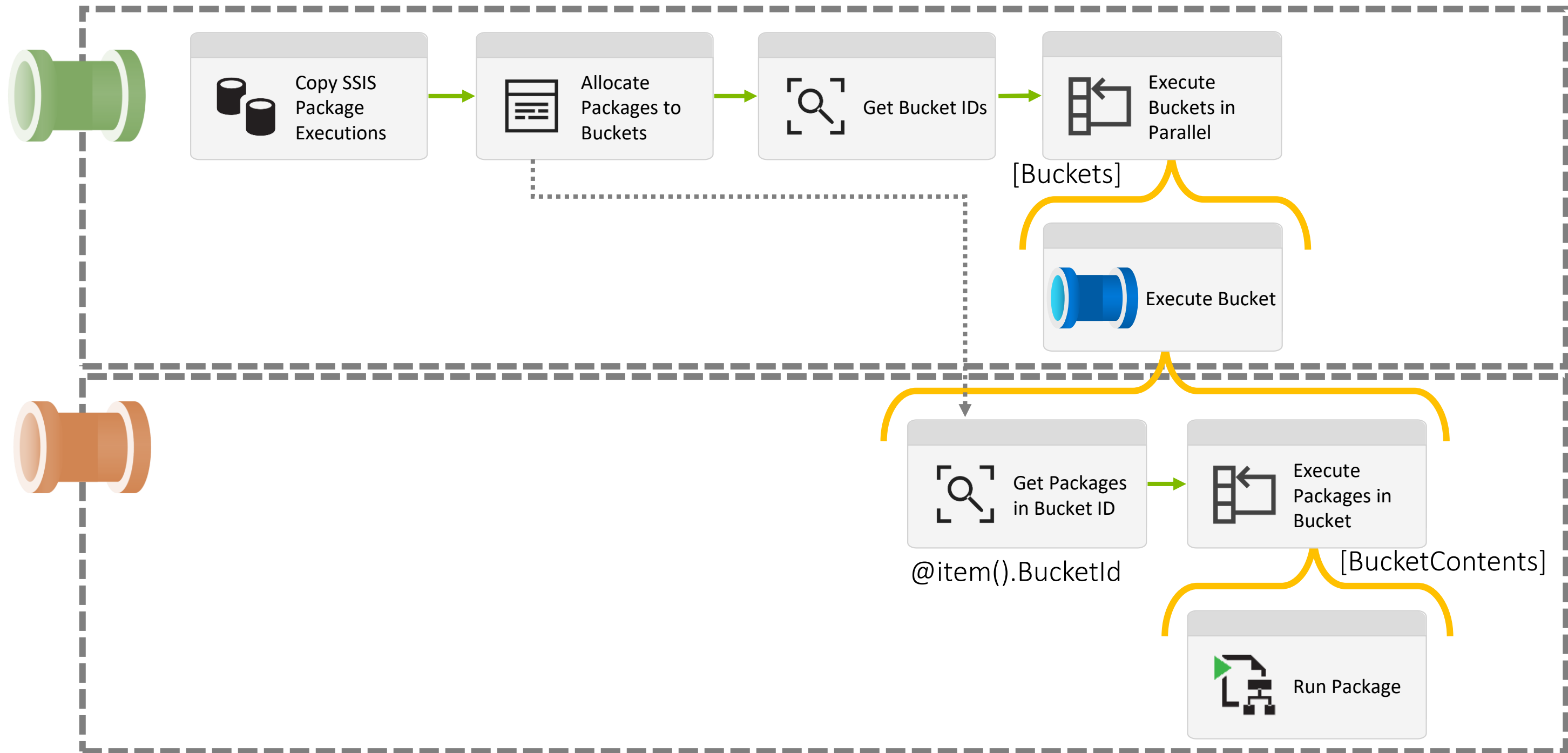


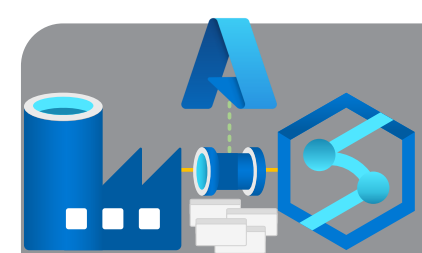


Nested ForEach Activities & Bucket Metadata



SSIS IR

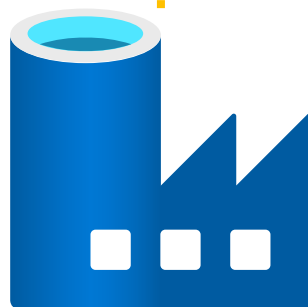
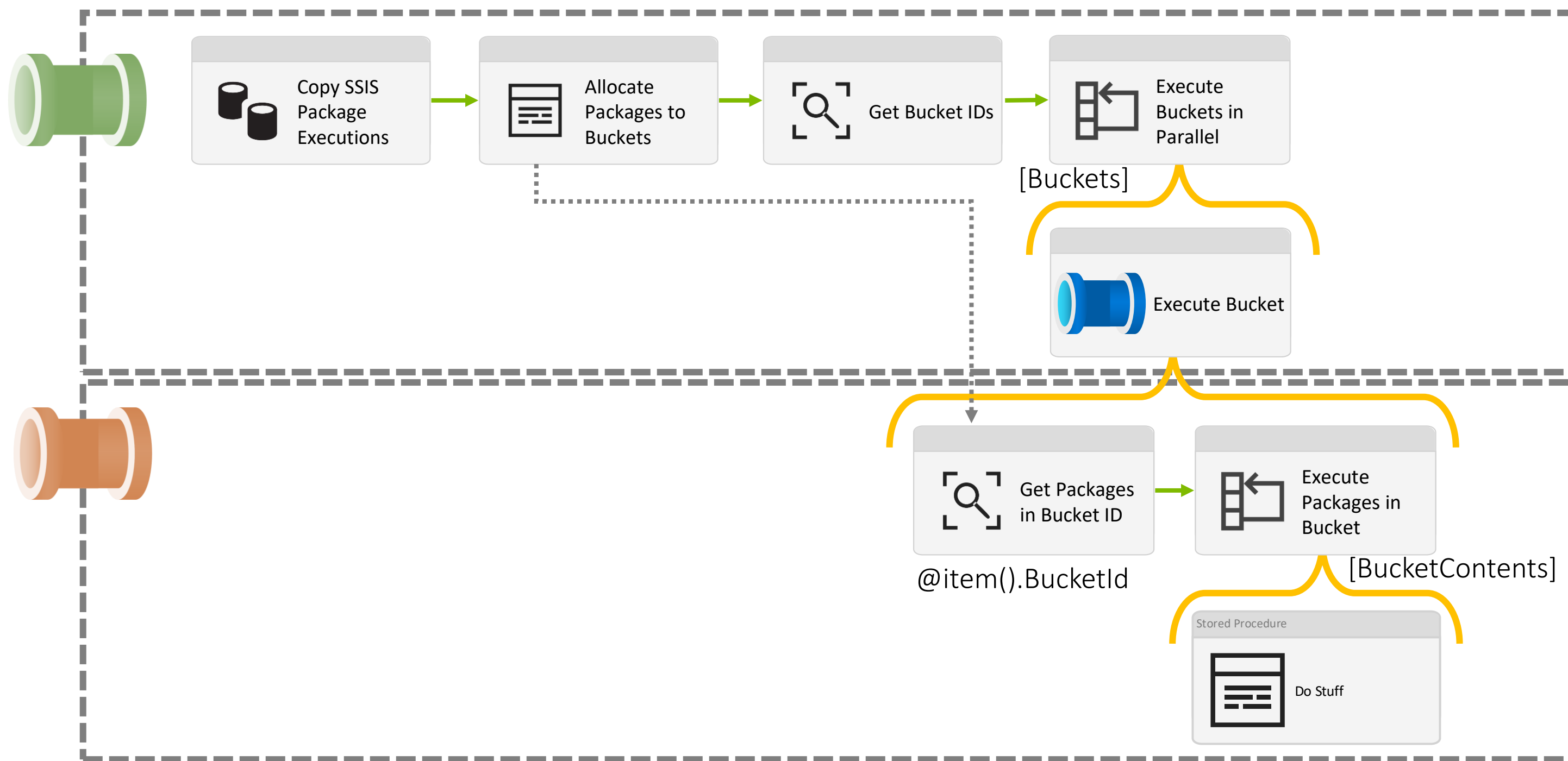




A General Pattern for Scaling Out

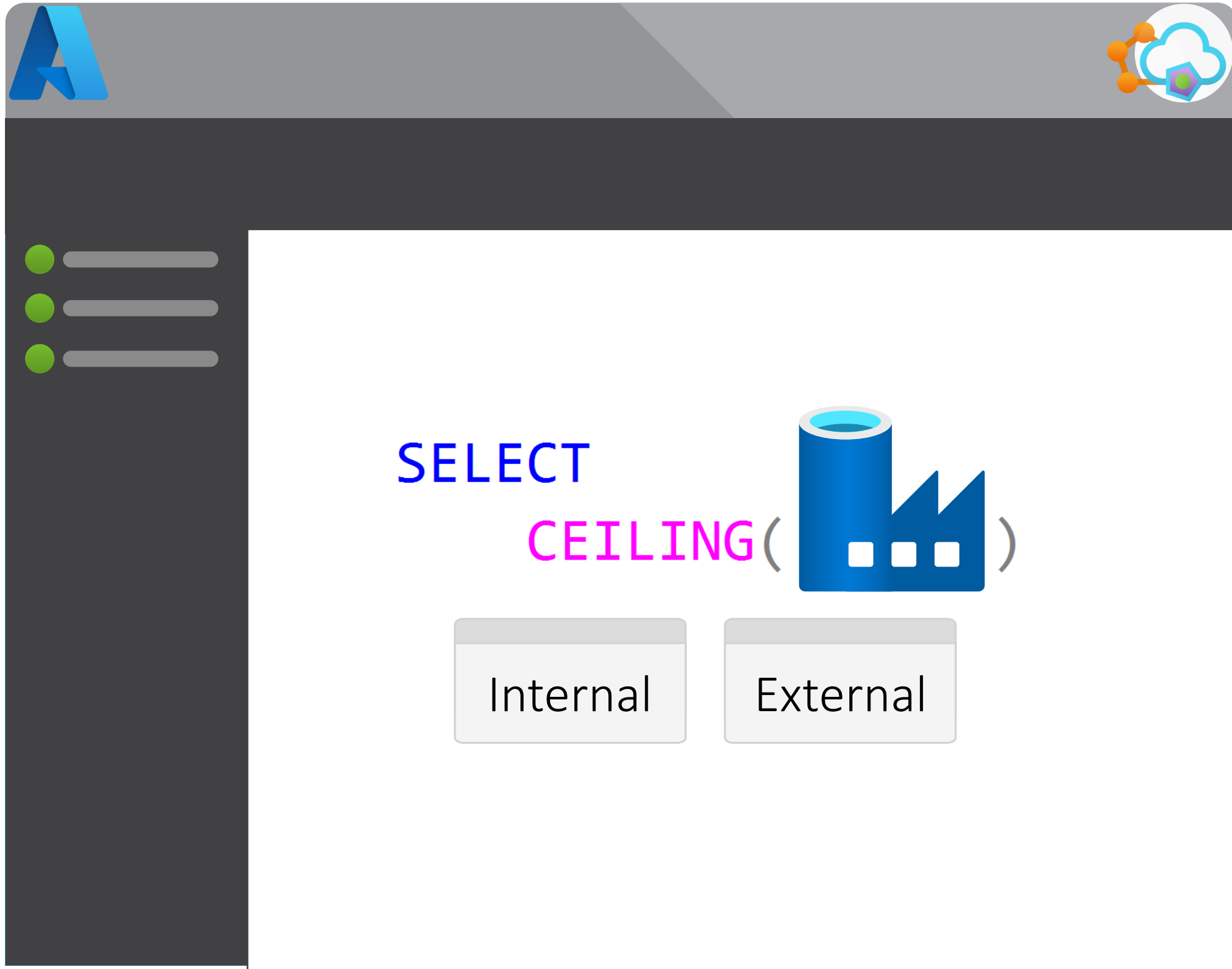


Azure IR

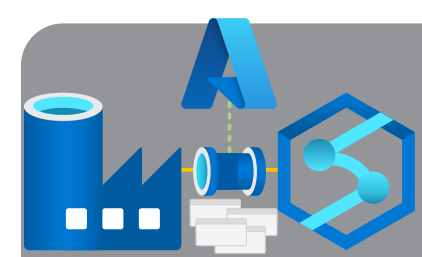


Module 6

Execution Parallelism



- Control Flow Scale Out
- Concurrency Limitations
- Internal vs External Activities
- Orchestration Framework - <http://procfwk.com>



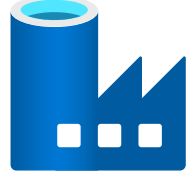
Data Factory Limitations

Resource Limits

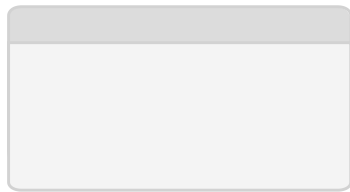
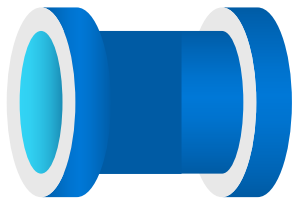
[mrpaulandrew.com/2020/01/29/azure-data-factory-resource-limitations/](https://github.com/MicrosoftDocs/azure-docs/blob/main/includes/azure-data-factory-limits.md)



<https://github.com/MicrosoftDocs/azure-docs/blob/main/includes/azure-data-factory-limits.md>



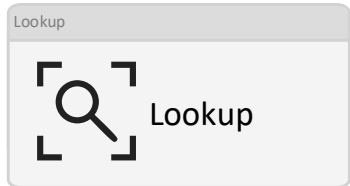
800 Data Factory Instances per Subscription



40 Activities per Data Factory Pipeline



3 Active Data Flow Debug Sessions per Data Factory



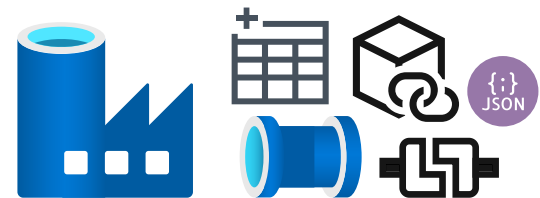
5,000 Rows or 4MB of Data Returned per Lookup (No Error if More)



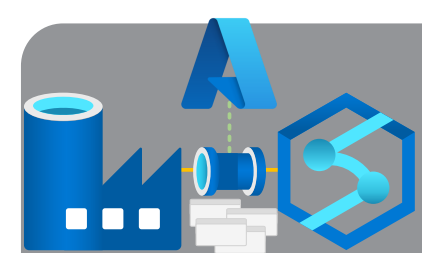
Minimum Tumbling Window Trigger – 15mins



4min Client Response Timeout Using Azure Functions Activity



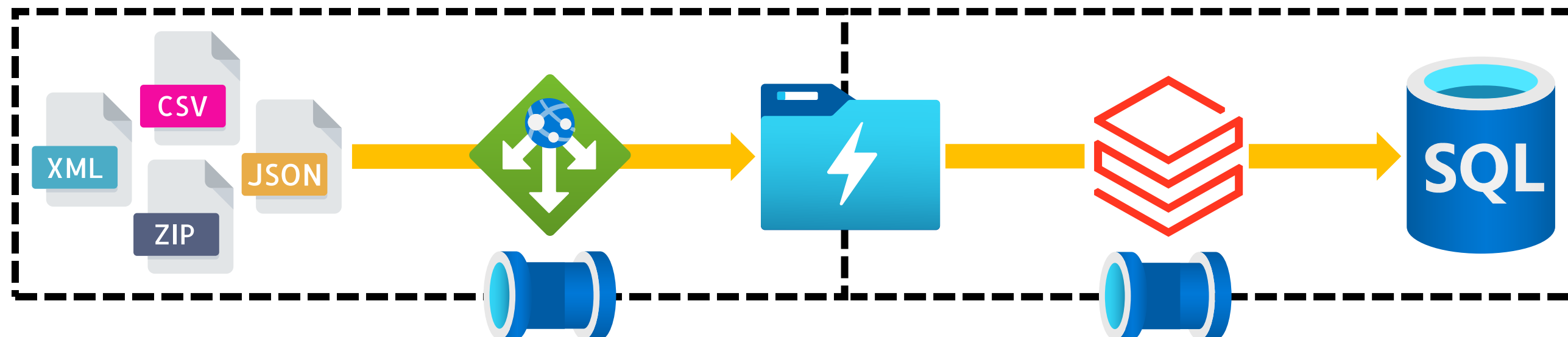
5,000 Entities (Components) per Data Factory Instance



Integration Pipelines as Data Engineers



Control Flow



1

Linked Services



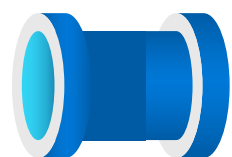
2

Datasets



3

Activities



4

Pipelines



5

Triggers



Add dynamic content [Alt+P]

Integration Runtimes

6



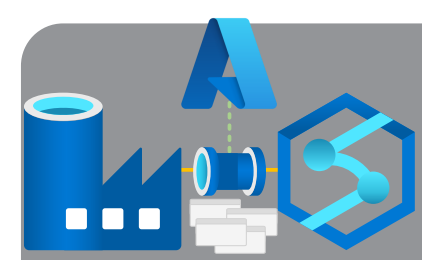
Azure IR



Hosted IR

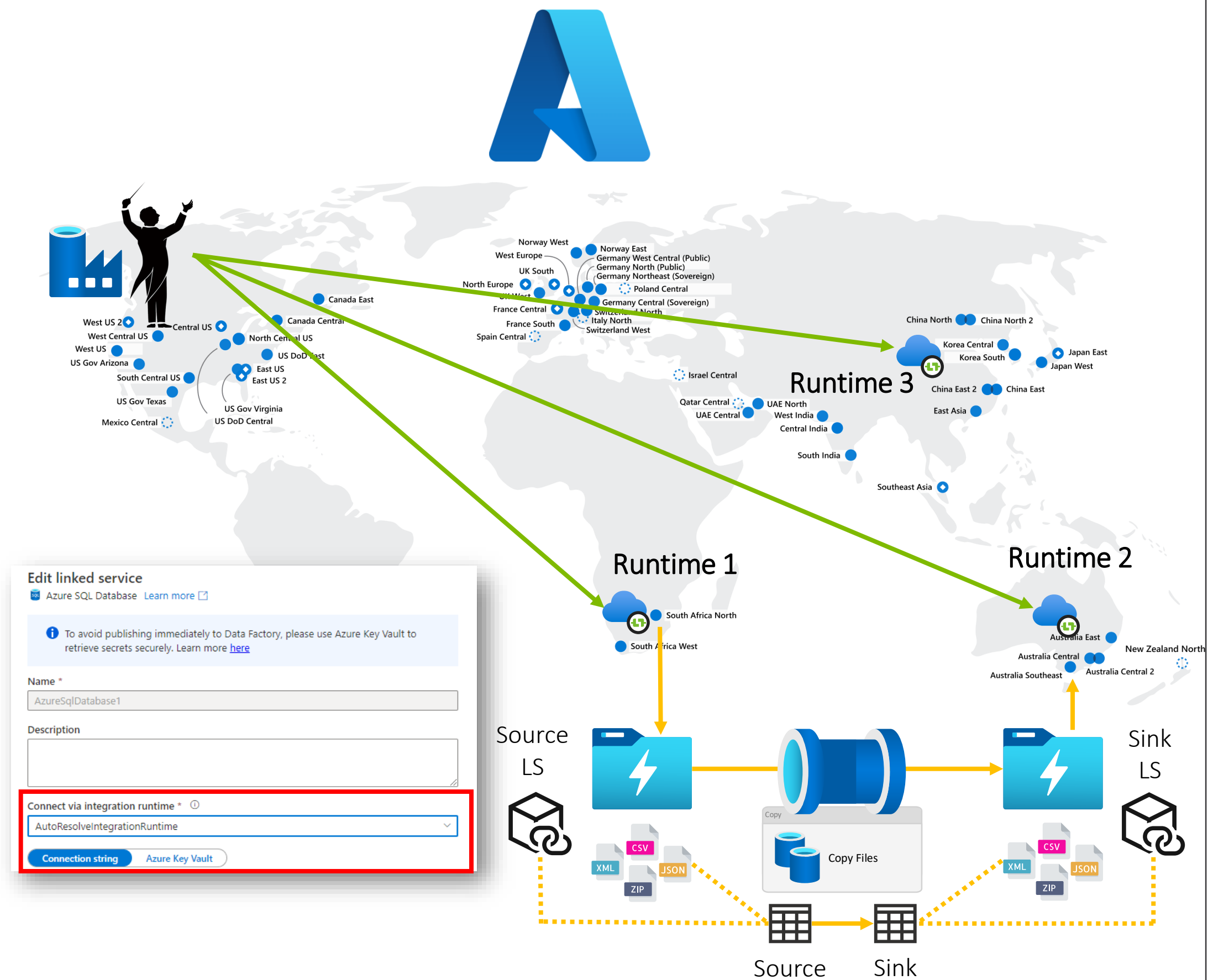


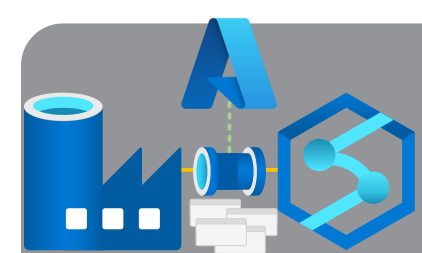
SSIS IR



Azure Integration Runtime







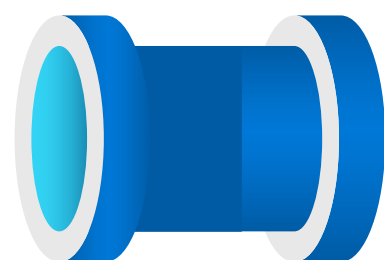
Compute Concurrency

Internal vs External Activities

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



Per Subscription, per IR Region



10,000

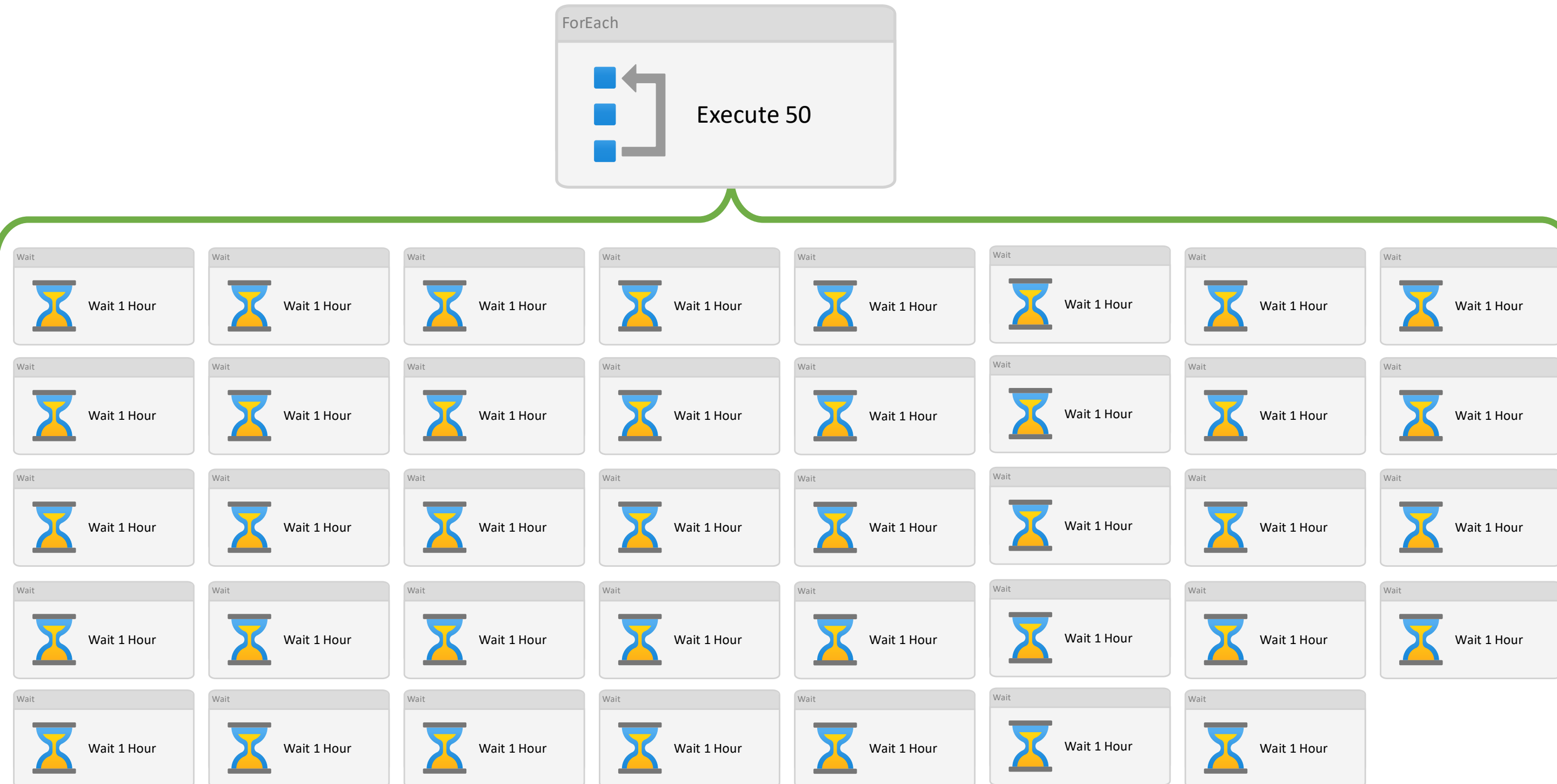
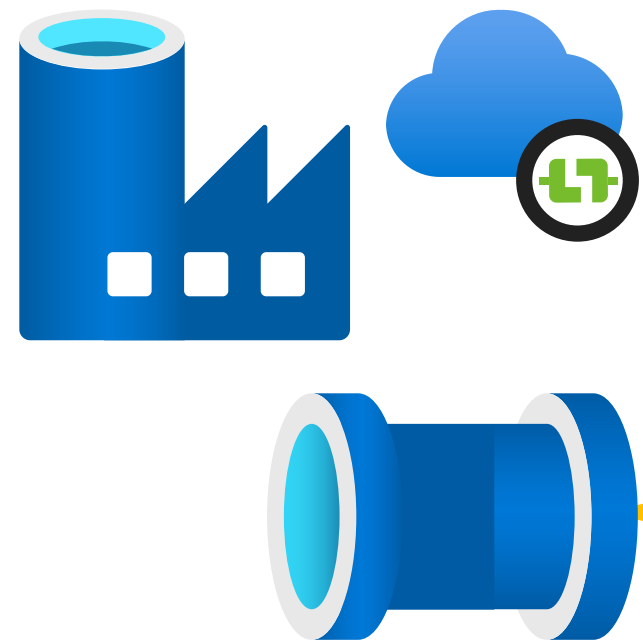
Internal

1,000

External

3,000

Example 1



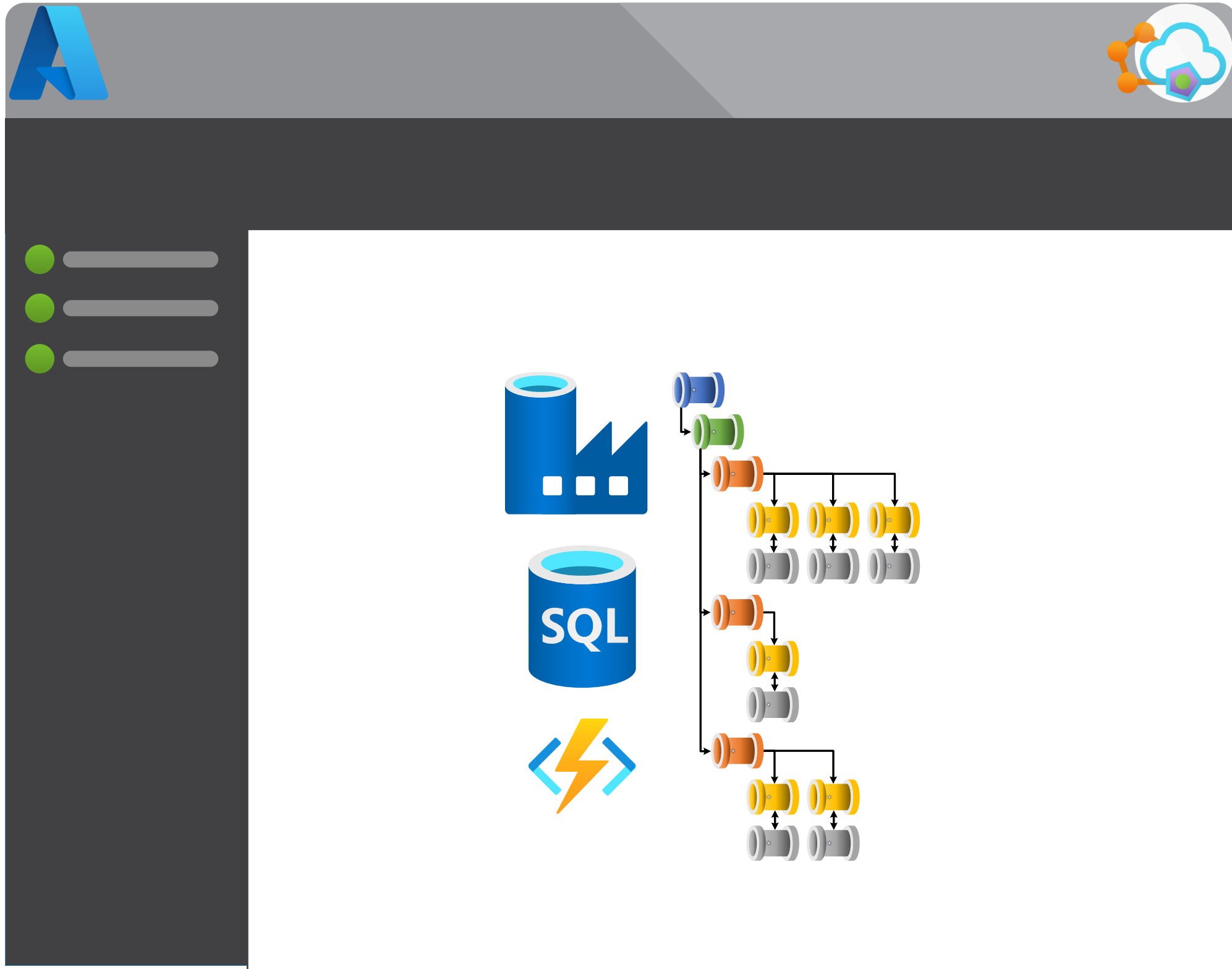
1
IR

1
Pipeline

$$1 \text{ ForEach} + \left(50 \text{ Batches} \times 39 \text{ Wait Activities} \right) = 1951 \text{ Concurrent Activities}$$

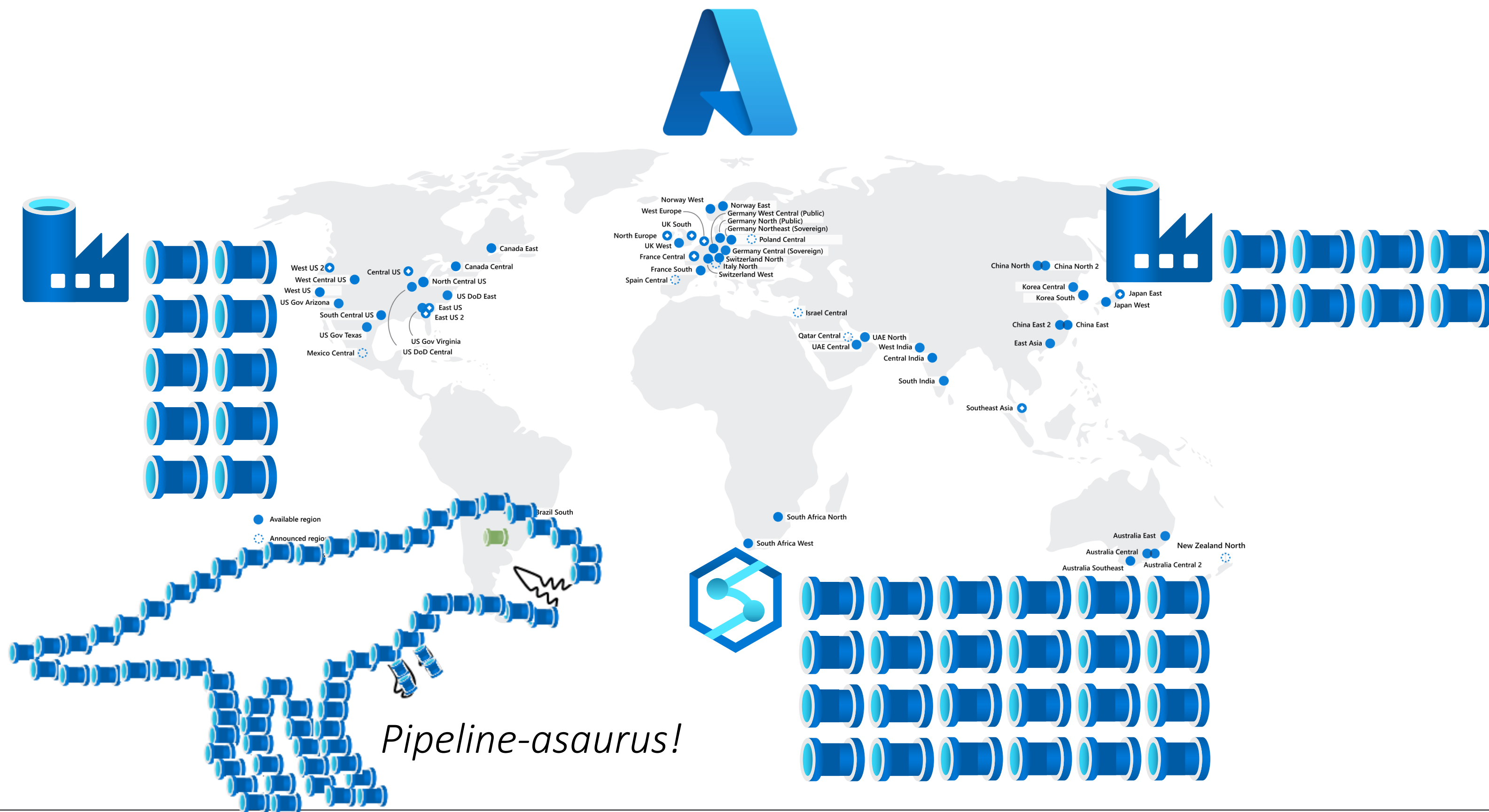
Module 6

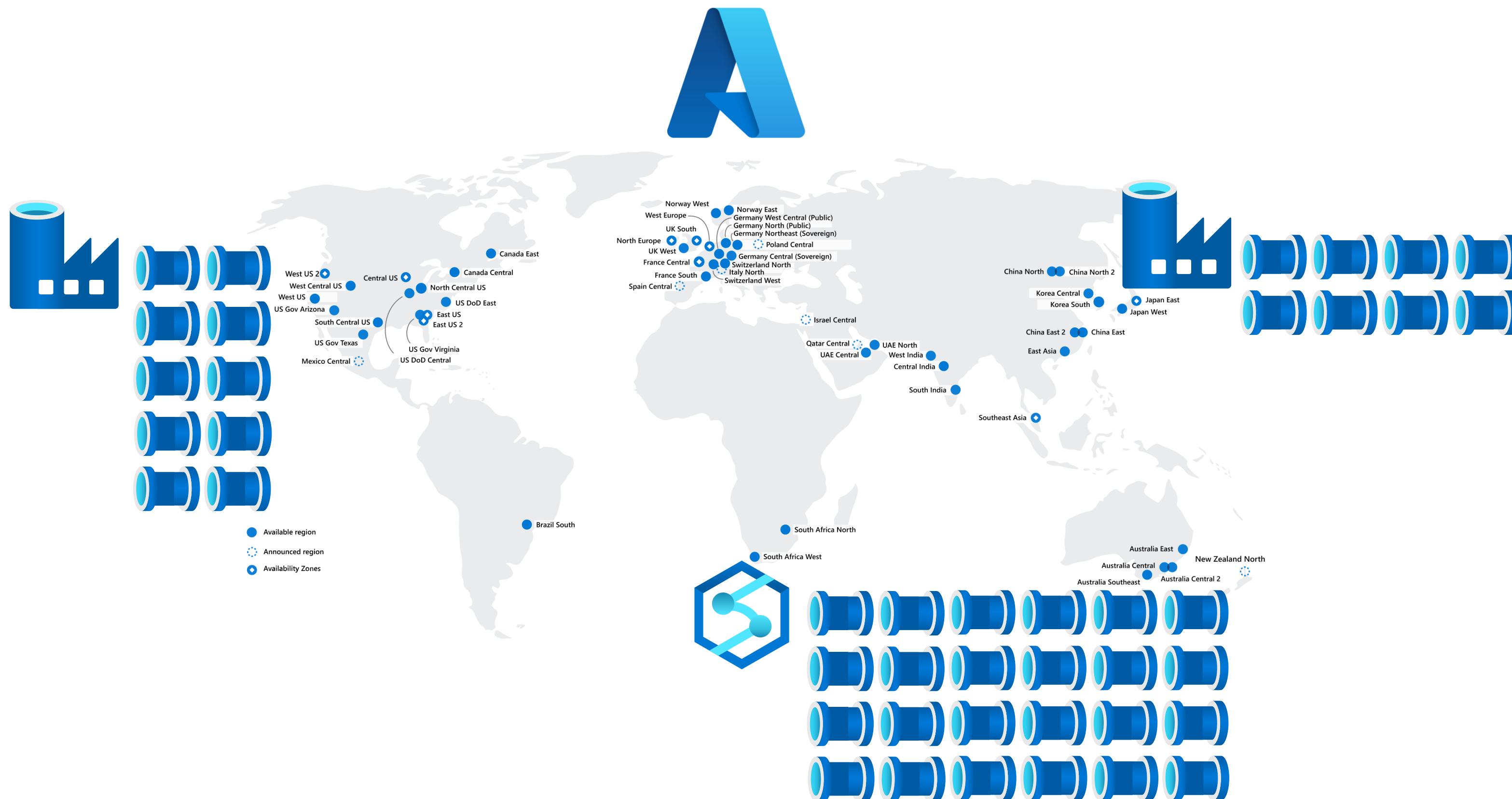
Execution Parallelism



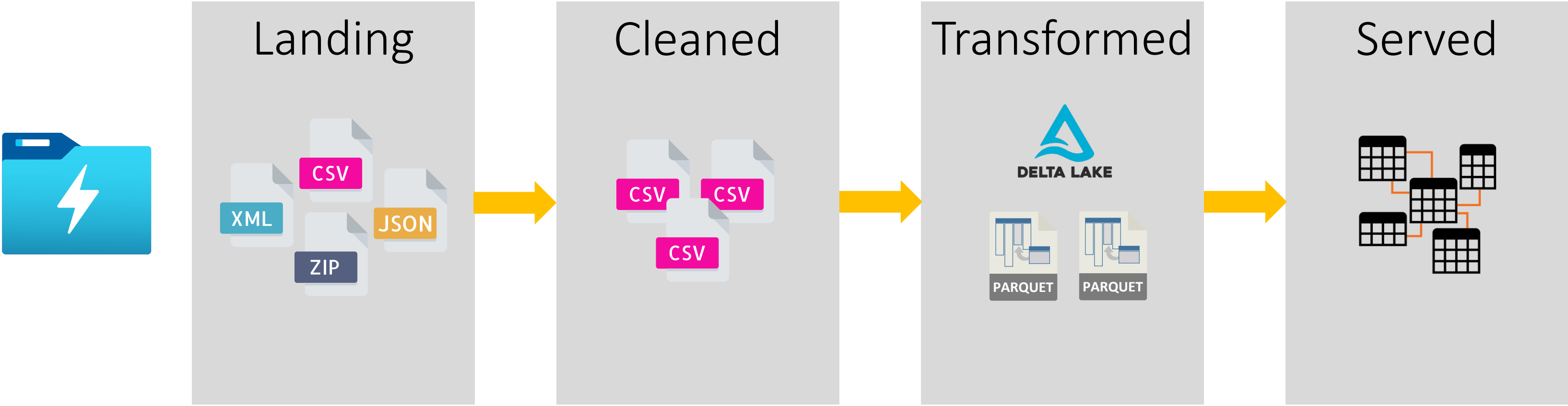
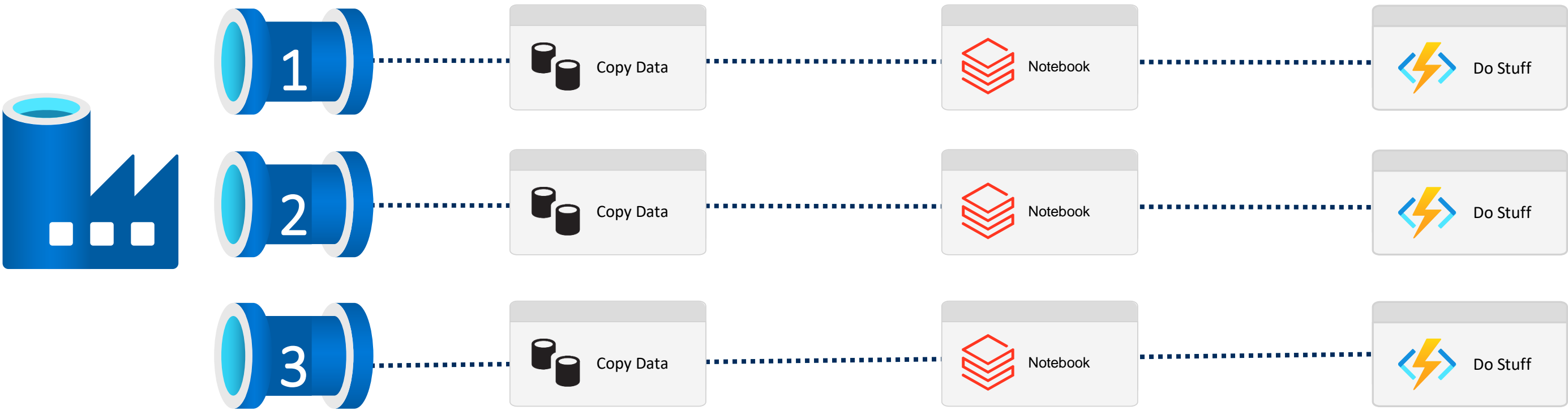
- Control Flow Scale Out
- Concurrency Limitations
- Internal vs External Activities
- Orchestration Framework - <http://procfwk.com>

How should we structure and trigger our Integration Pipelines?

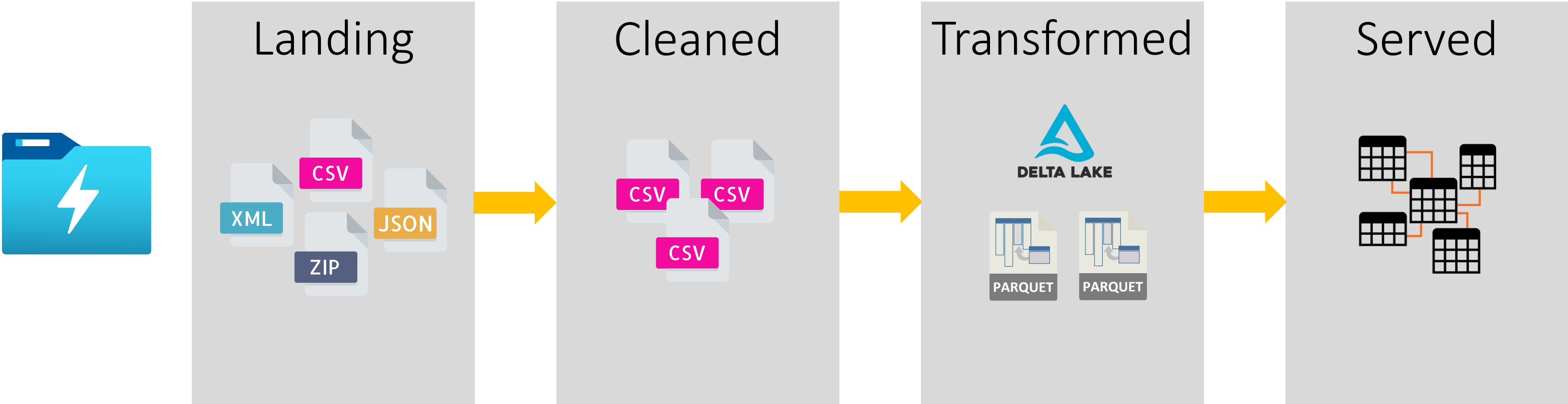
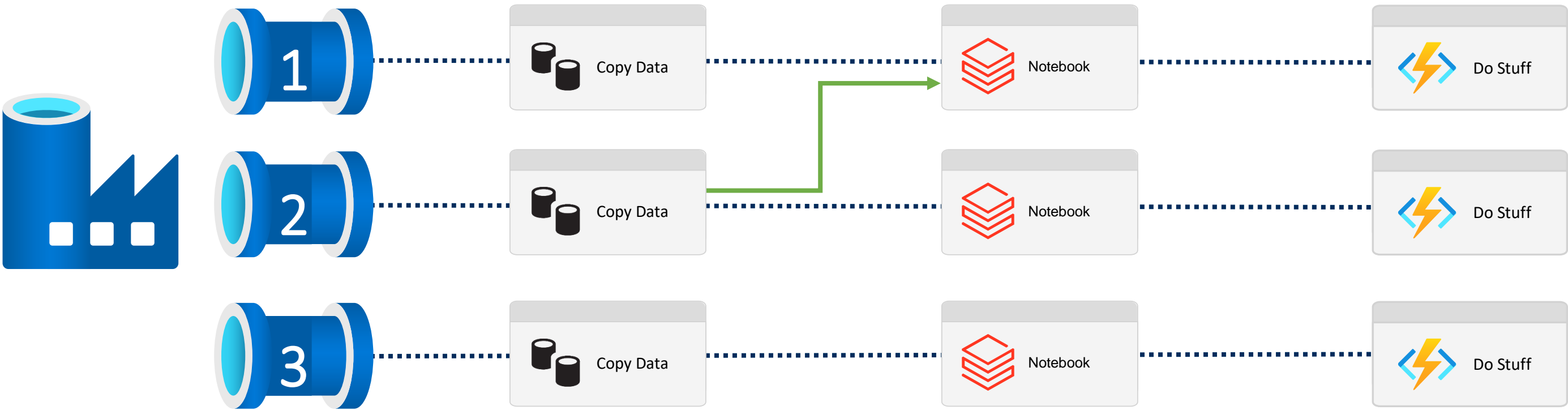




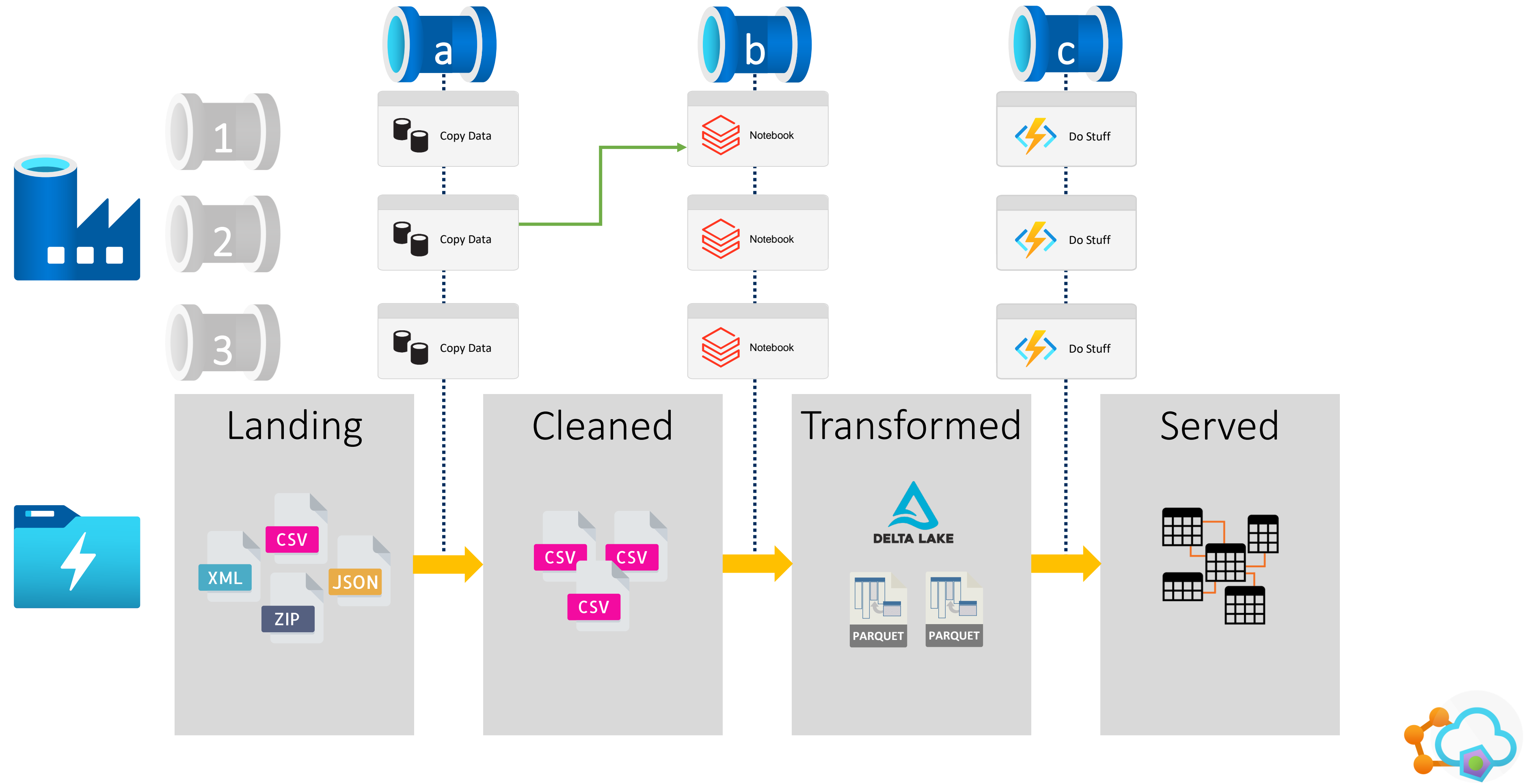
Problem



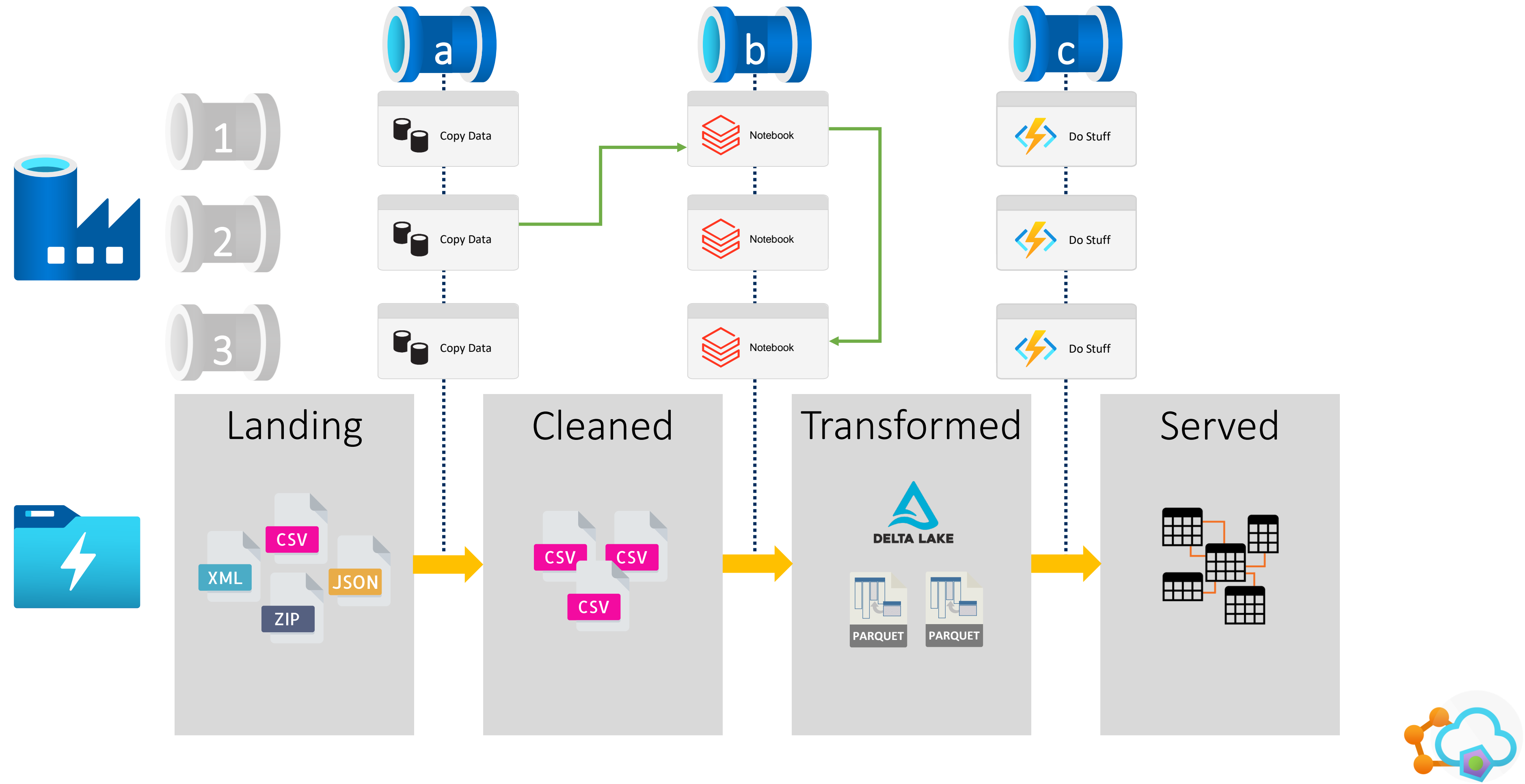
Problem



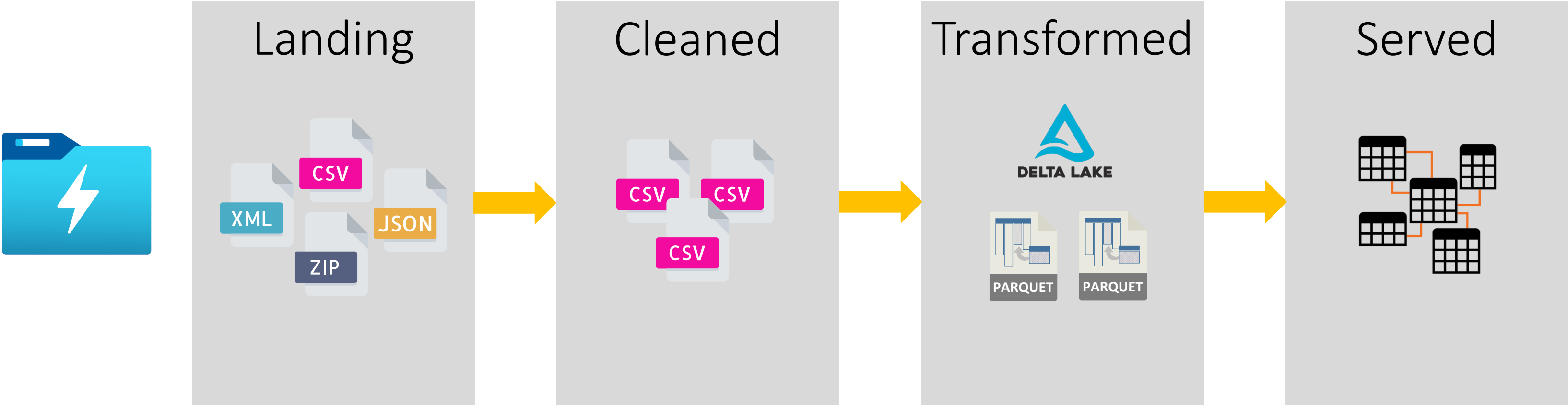
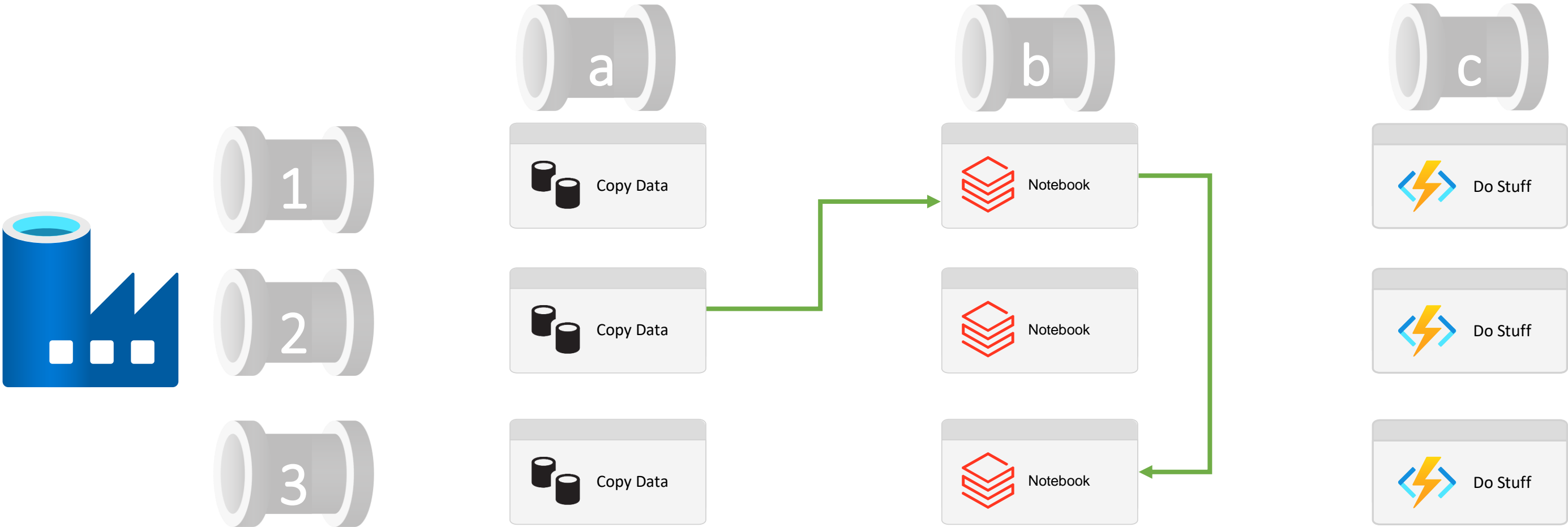
Problem




Problem

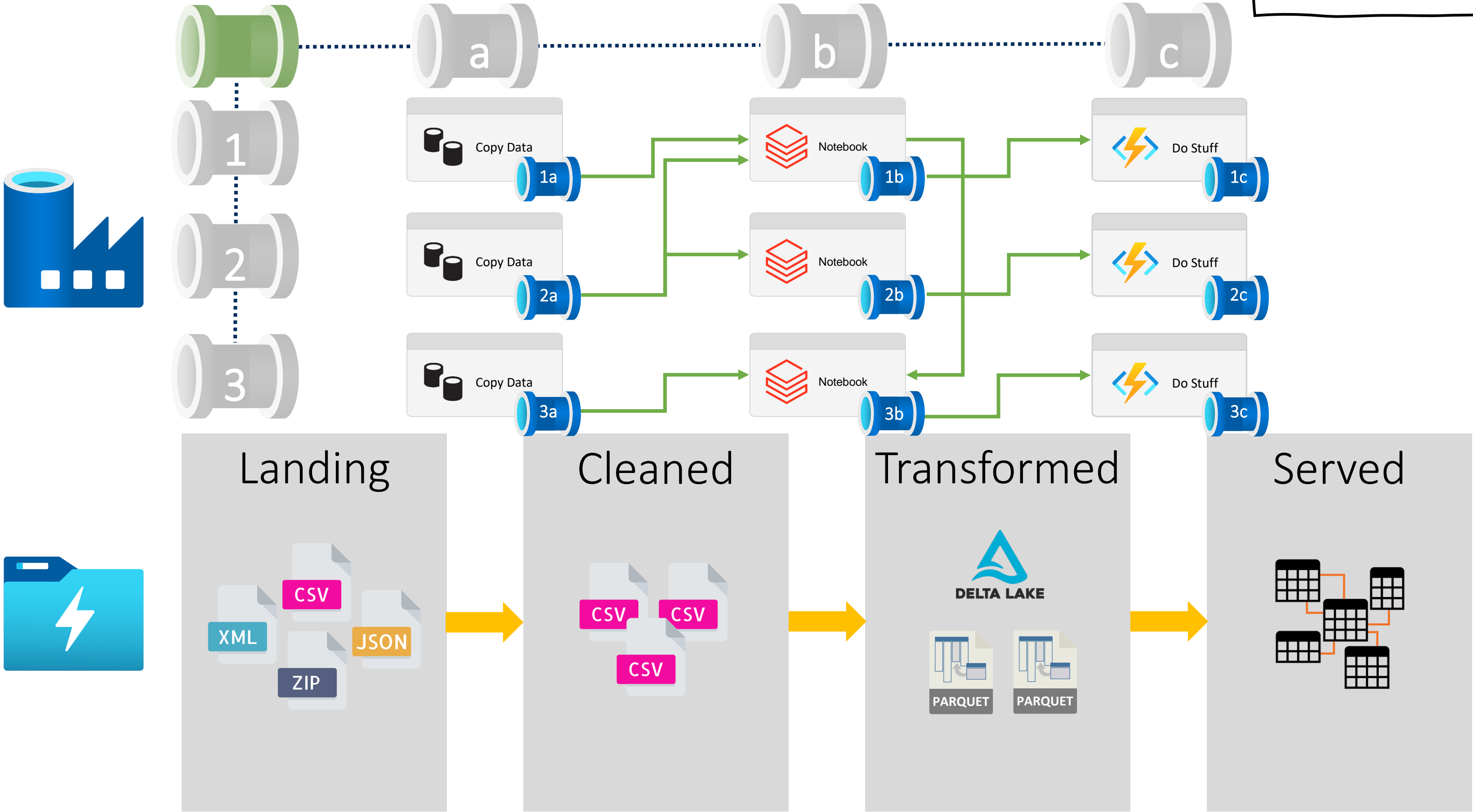


Problem

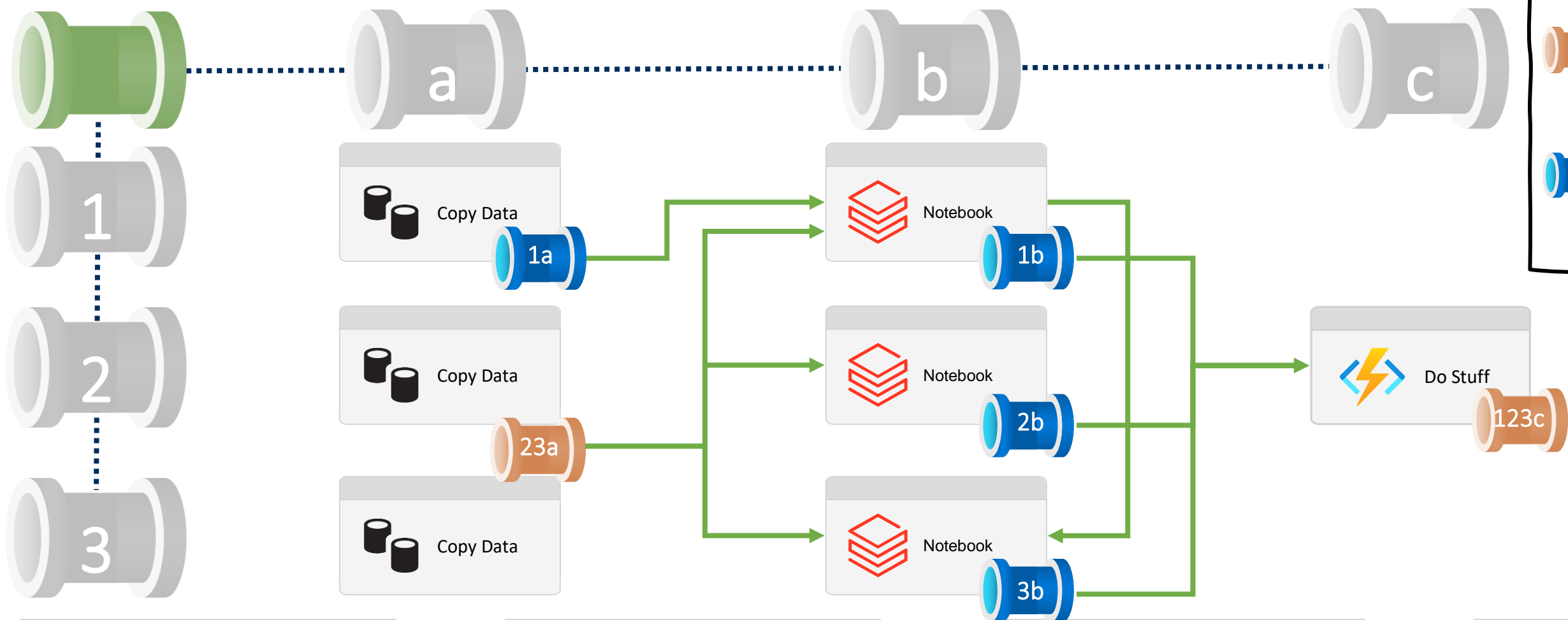
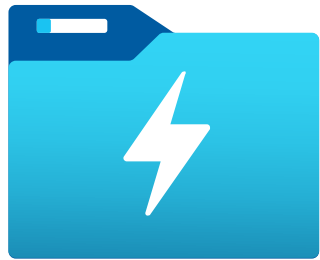
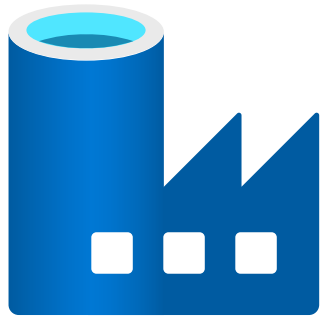




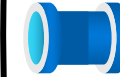
Problem

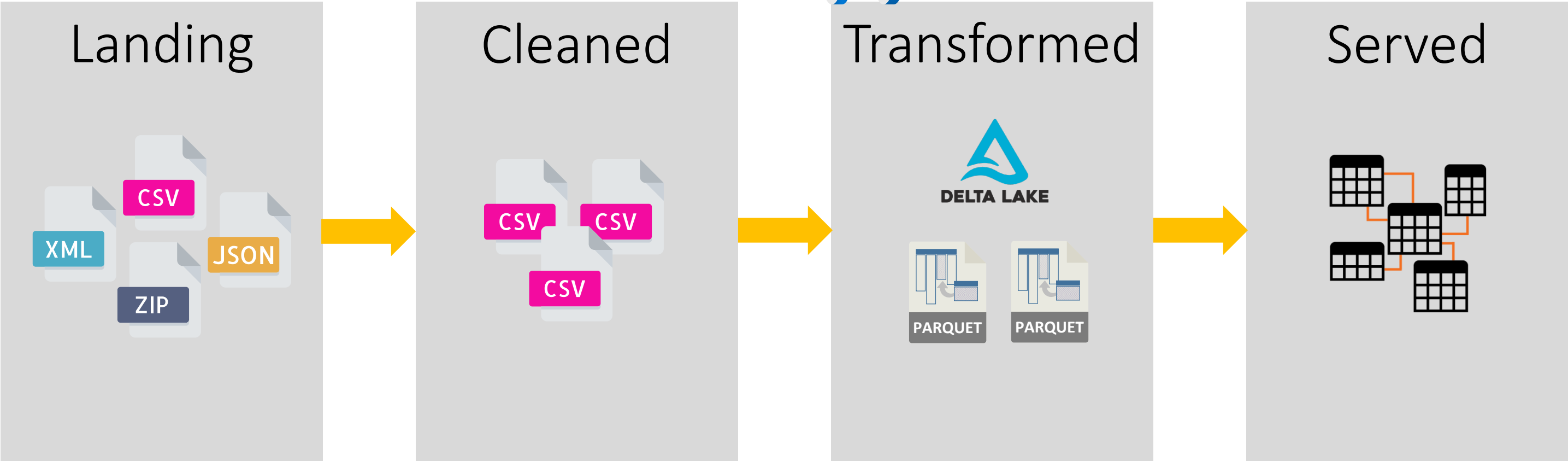
 Only 40 Activities per Pipeline.



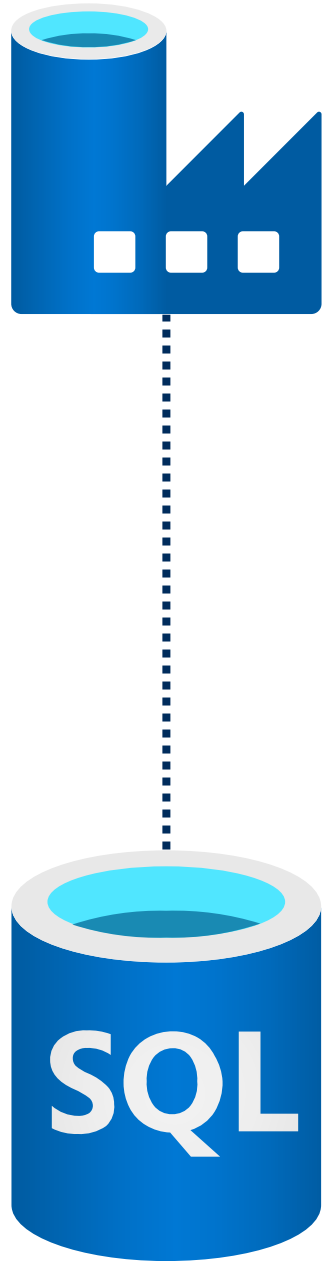
Problem



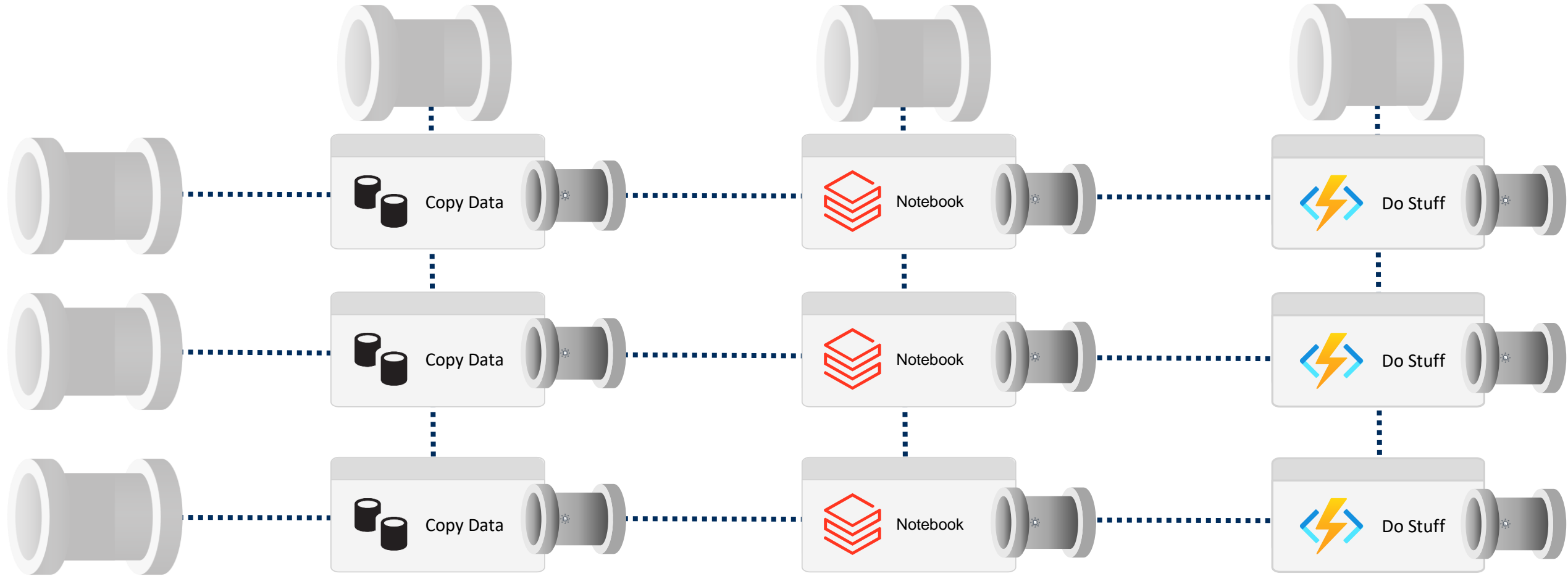
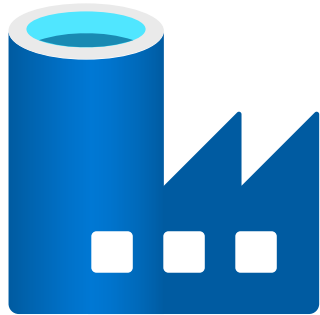
-  Grandparent pipeline for all processing.
-  Parent pipeline to consolidate work.
-  Child pipelines for low level dependencies.



Use Metadata to Drive Data Factory Pipelines

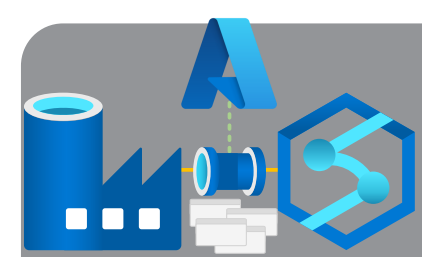


Solution



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

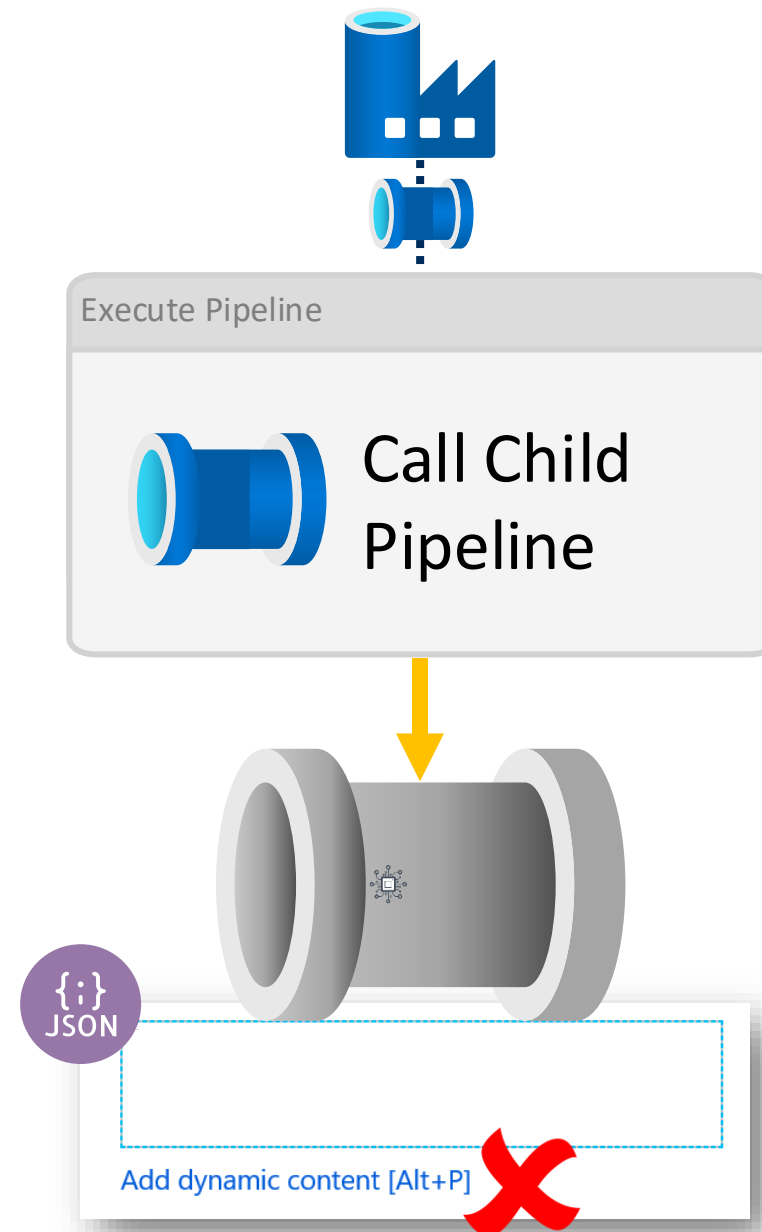
Stage	Pipeline
1	a
1	b
1	c
2	d
2	e
3	f
3	g
3	h
3	i



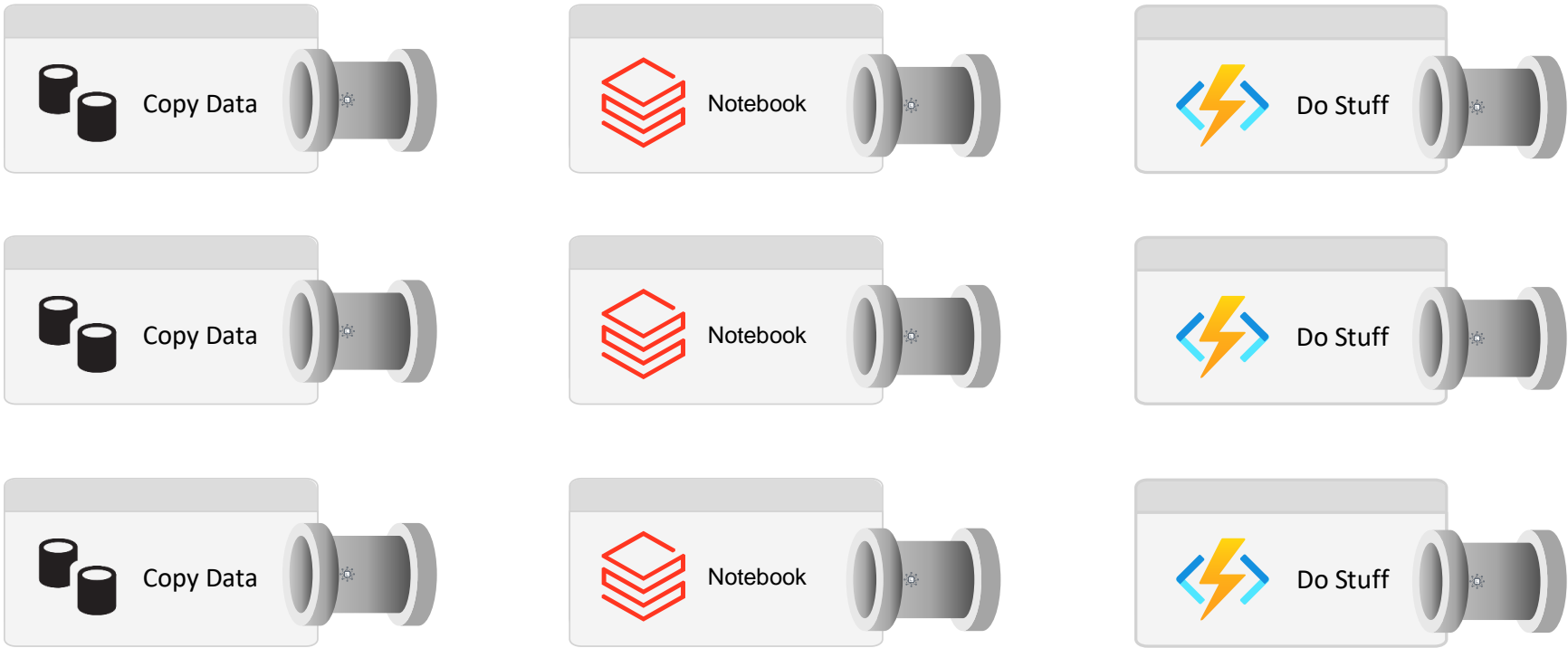
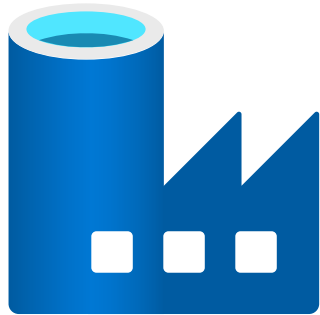
Calling Our Worker Pipelines



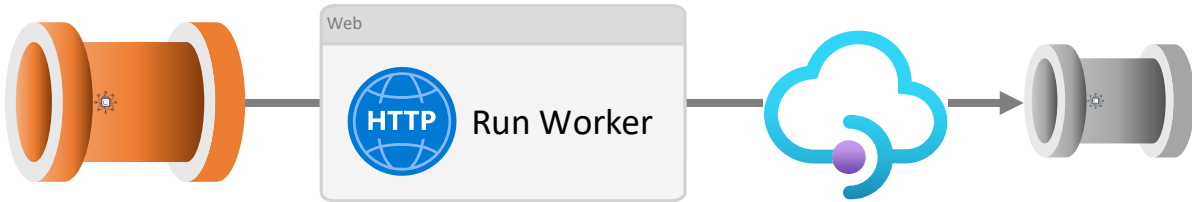
One More Problem to Consider



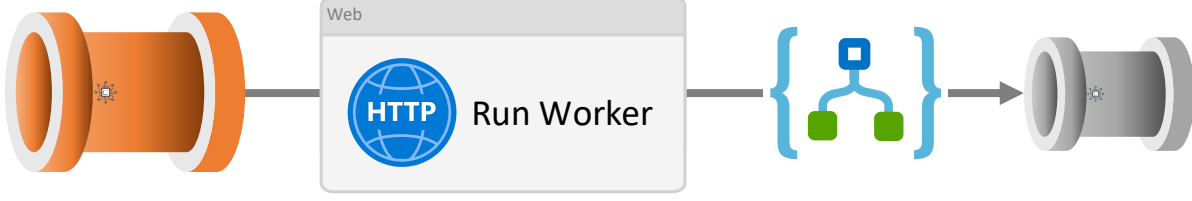
Calling Our Worker Pipelines



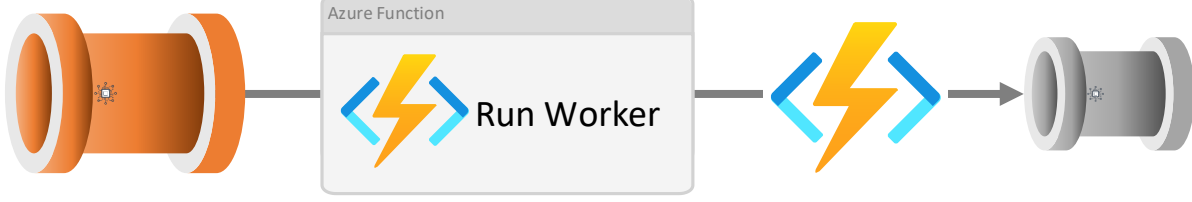
Option 1:



Option 2:



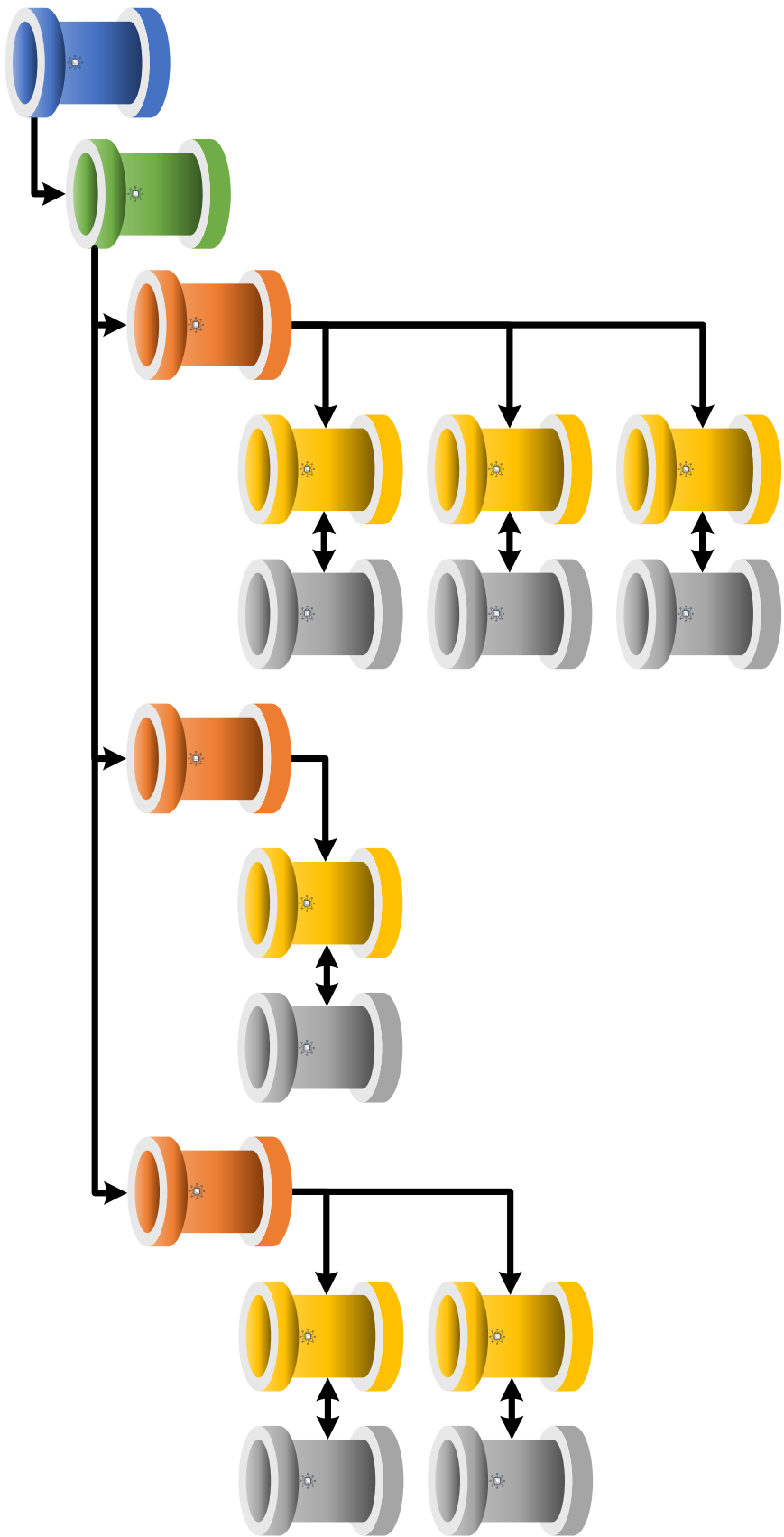
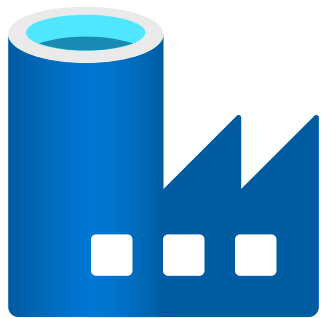
Option 3:

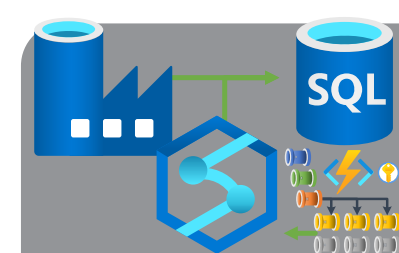


Stages	Pipelines
1	a
2	b
3	c
	d
	e
	f
	g
	h
	i

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

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



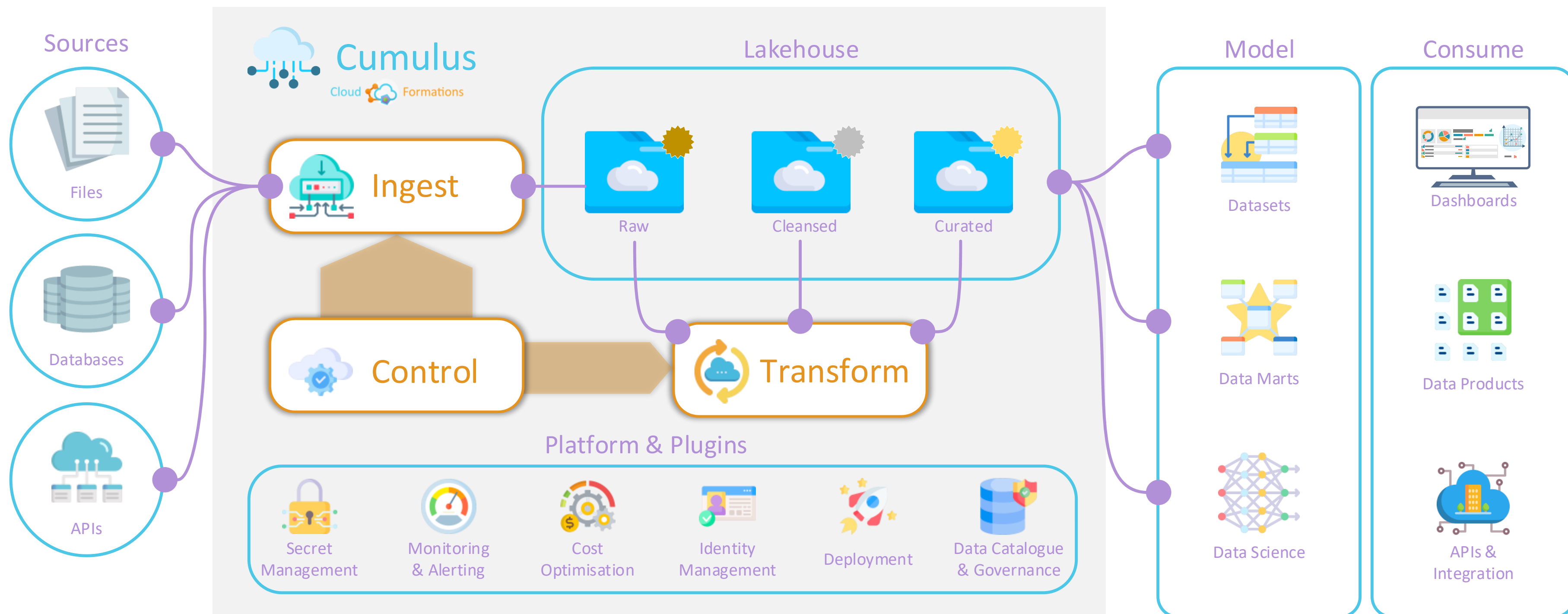


Introducing CF.Cumulus



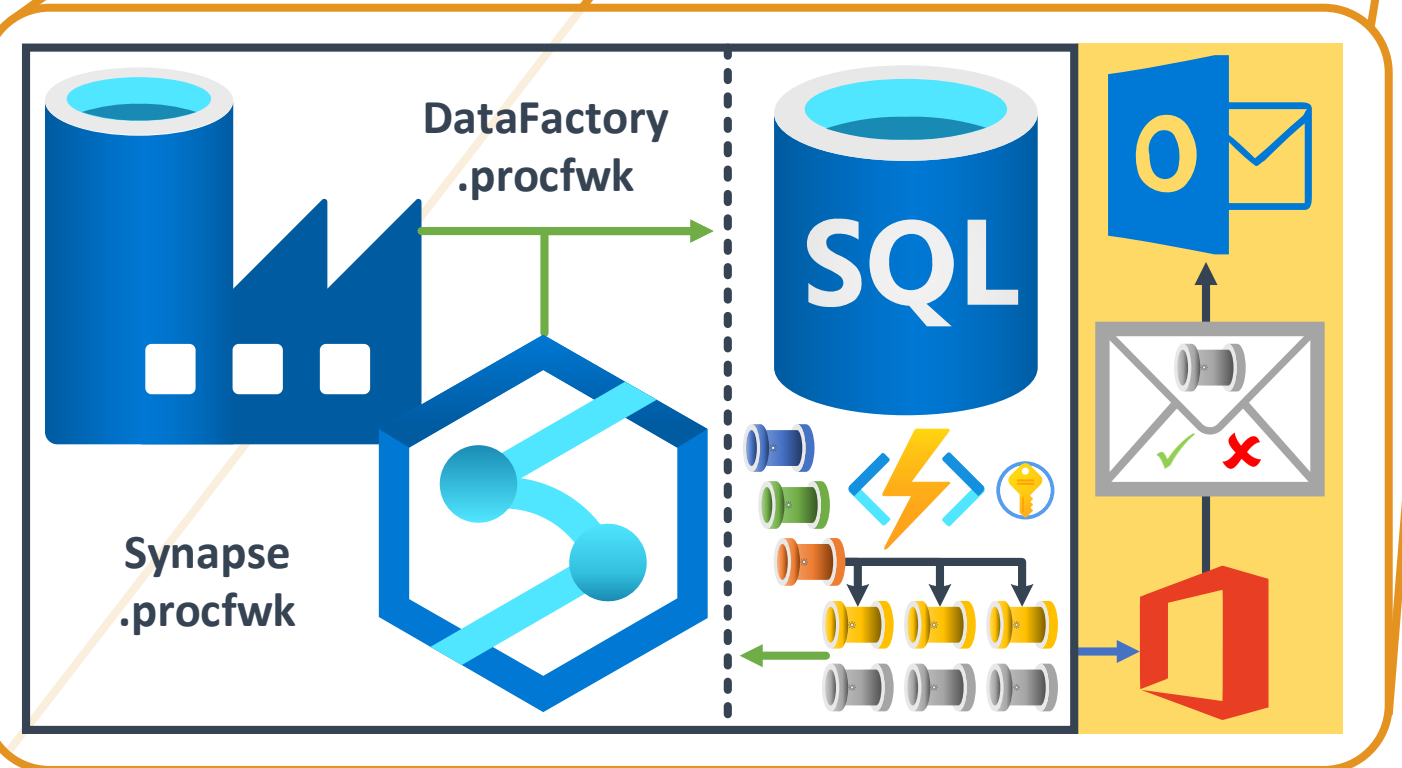
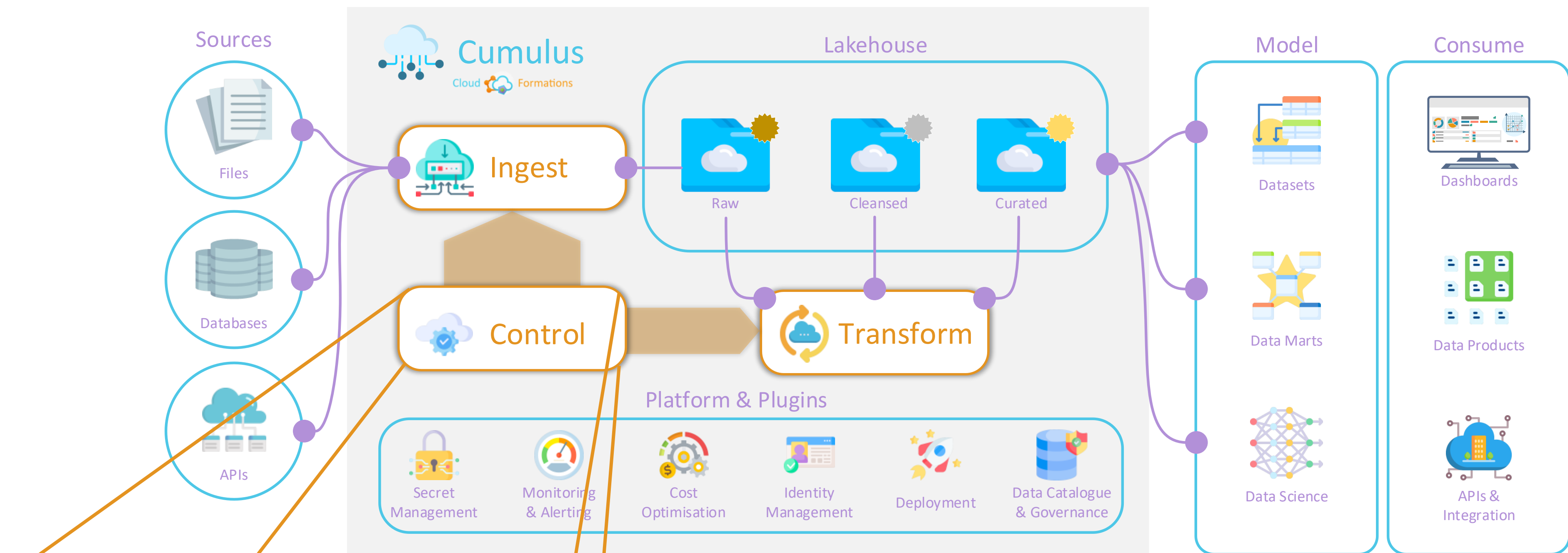
cloudformations.org/cumulus

A cloud data platform product to accelerate time to insights. Our open-source framework is designed for the real world. Stripping away the complexity, giving you the power to build, scale, and manage your dataflows with ease, accelerating data delivery.



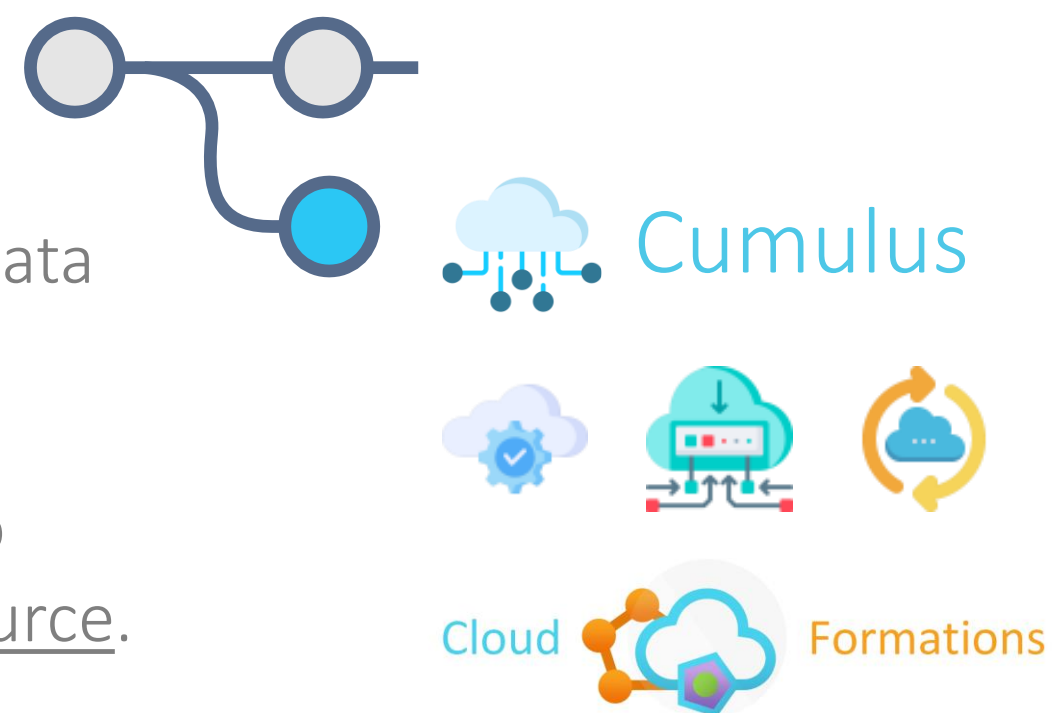
Underlying Technology: **Spark** **DELTA LAKE** **C#** **JSON**

Works With:

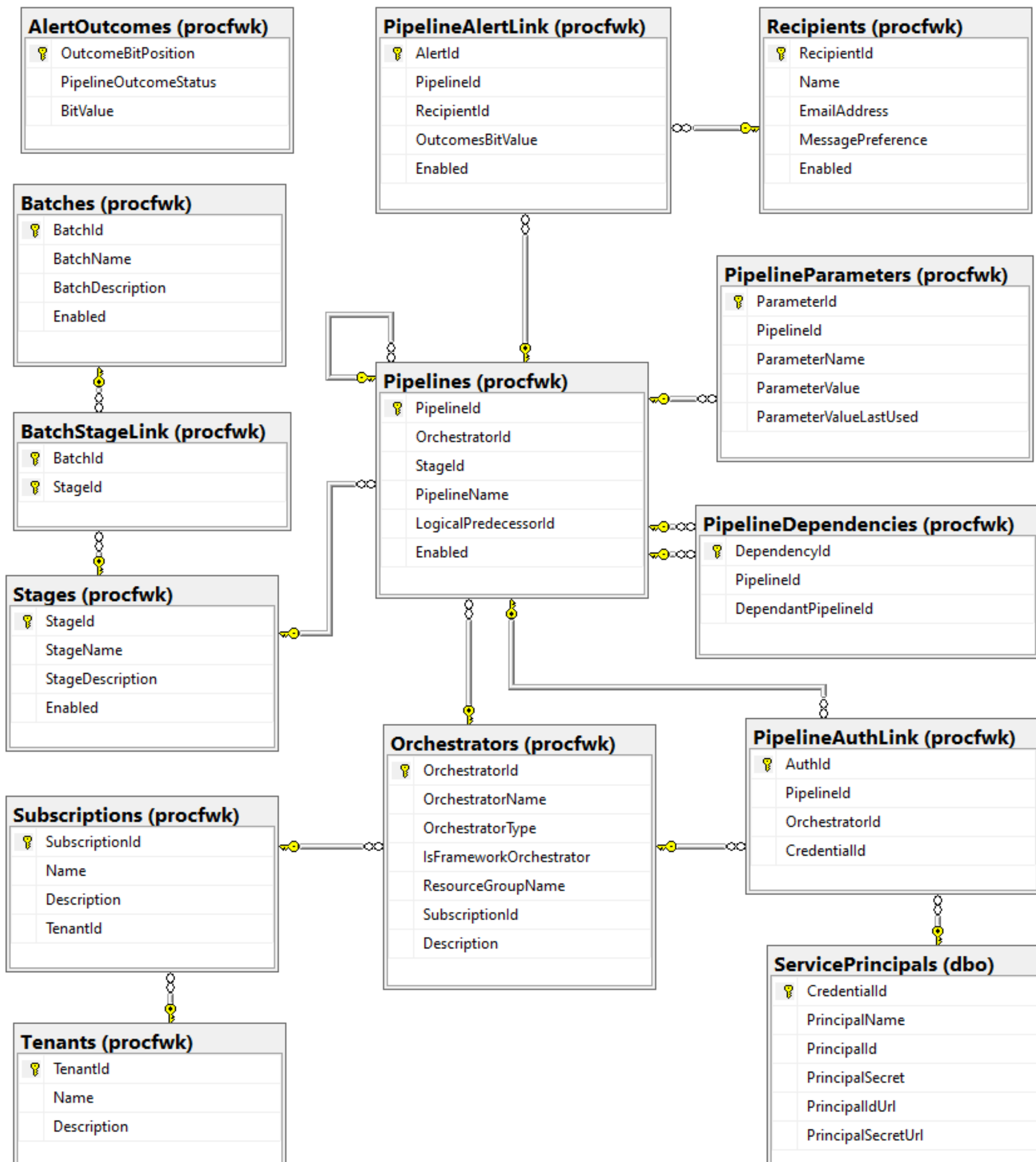


 **ProcFwk.com**

An evolution from metadata driven orchestration to complete data delivery. Accelerating your time to insight. And still open-source.



Framework Database



Properties (procfwk)	
PropertyId	
PropertyName	
PropertyValue	
Description	
ValidFrom	
ValidTo	

BatchExecution (procfwk)	
BatchId	
ExecutionId	
BatchName	
BatchStatus	
StartDateTime	
EndDateTime	

ExecutionLog (procfwk)	
LogId	
LocalExecutionId	
StageId	
PipelineId	
CallingDataFactoryName	
ResourceGroupName	
DataFactoryName	
PipelineName	
StartDateTime	
PipelineStatus	
EndDateTime	
AdfPipelineRunId	
PipelineParamsUsed	

CurrentExecution (procfwk)	
LocalExecutionId	
StageId	
PipelineId	
CallingDataFactoryName	
ResourceGroupName	
DataFactoryName	
PipelineName	
StartDateTime	
PipelineStatus	
LastStatusCheckDateTime	
EndDateTime	
IsBlocked	
AdfPipelineRunId	
PipelineParamsUsed	

ErrorLog (procfwk)	
LogId	
LocalExecutionId	
AdfPipelineRunId	
ActivityRunId	
ActivityName	
ActivityType	
ErrorCode	
ErrorType	
ErrorMessage	

Framework Database

Configuration &
Behaviour

Core Metadata

Execution Handling

Location &
Authentication

Email Alerting

Runtime & Logging

AlertOutcomes (procfwk)	
OutcomeBitPosition	
PipelineOutcomeStatus	
BitValue	

PipelineAlertLink (procfwk)	
AlertId	
PipelineId	
RecipientId	
OutcomesBitValue	
Enabled	

Recipients (procfwk)	
RecipientId	
Name	
EmailAddress	
MessagePreference	
Enabled	

Batches (procfwk)	
BatchId	
BatchName	
BatchDescription	
Enabled	

BatchStageLink (procfwk)	
BatchId	
StageId	

Stages (procfwk)	
StageId	
StageName	
StageDescription	
Enabled	

Subscriptions (procfwk)	
SubscriptionId	
Name	
Description	
TenantId	

Tenants (procfwk)	
TenantId	
Name	
Description	

Pipelines (procfwk)	
PipelineId	
OrchestratorId	
StageId	
PipelineName	
LogicalPredecessorId	
Enabled	

PipelineParameters (procfwk)	
ParameterId	
PipelineId	
ParameterName	
ParameterValue	
ParameterValueLastUsed	

PipelineDependencies (procfwk)	
DependencyId	
PipelineId	
DependantPipelineId	

Orchestrators (procfwk)	
OrchestratorId	
OrchestratorName	
OrchestratorType	
IsFrameworkOrchestrator	
ResourceGroupName	
SubscriptionId	
Description	

PipelineAuthLink (procfwk)	
AuthId	
PipelineId	
OrchestratorId	
CredentialId	

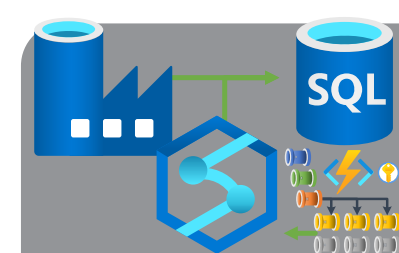
ServicePrincipals (dbo)	
CredentialId	
PrincipalName	
PrincipalId	
PrincipalSecret	
PrincipalIdUrl	
PrincipalSecretUrl	

BatchExecution (procfwk)	
BatchId	
ExecutionId	
BatchName	
BatchStatus	
StartDateTime	
EndDateTime	

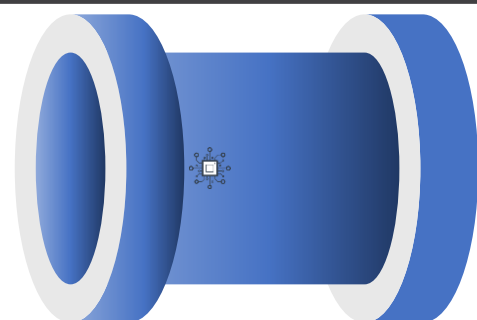
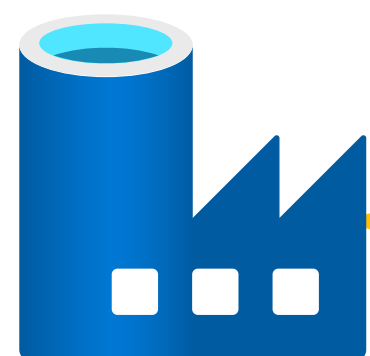
CurrentExecution (procfwk)	
LocalExecutionId	
StageId	
PipelineId	
CallingDataFactoryName	
ResourceGroupName	
DataFactoryName	
PipelineName	
StartDateTime	
PipelineStatus	
LastStatusCheckDateTime	
EndDateTime	
IsBlocked	
AdfPipelineRunId	
PipelineParamsUsed	

ExecutionLog (procfwk)	
LogId	
LocalExecutionId	
StageId	
PipelineId	
CallingDataFactoryName	
ResourceGroupName	
DataFactoryName	
PipelineName	
StartDateTime	
PipelineStatus	
EndDateTime	
AdfPipelineRunId	
PipelineParamsUsed	

ErrorLog (procfwk)	
LogId	
LocalExecutionId	
AdfPipelineRunId	
ActivityRunId	
ActivityName	
ActivityType	
ErrorCode	
ErrorType	
ErrorMessage	

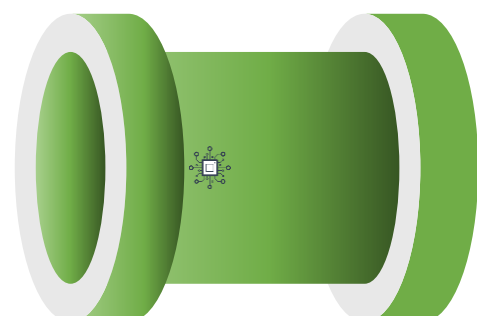


Framework Pipeline Hierarchy



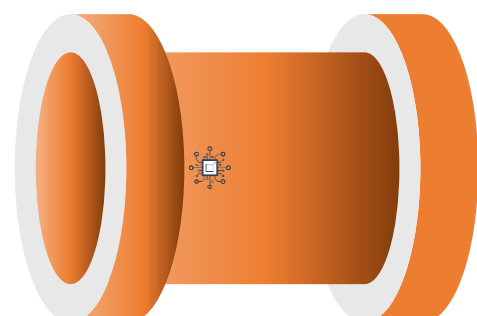
- Grandparent

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



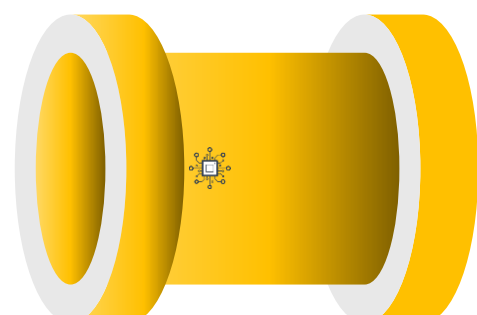
- Parent

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



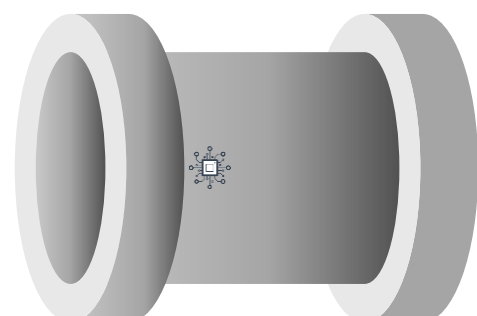
- Child

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



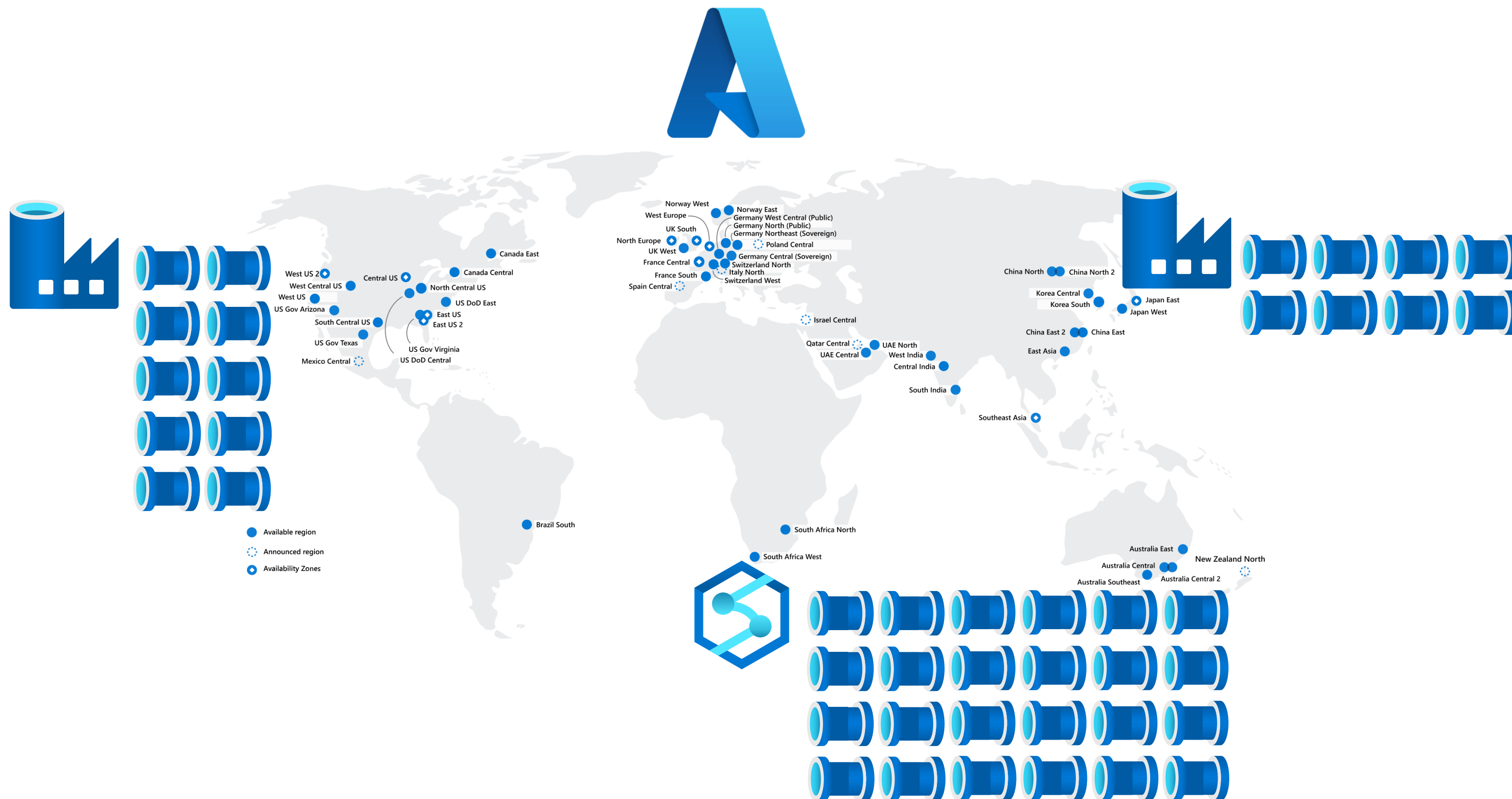
- Infant

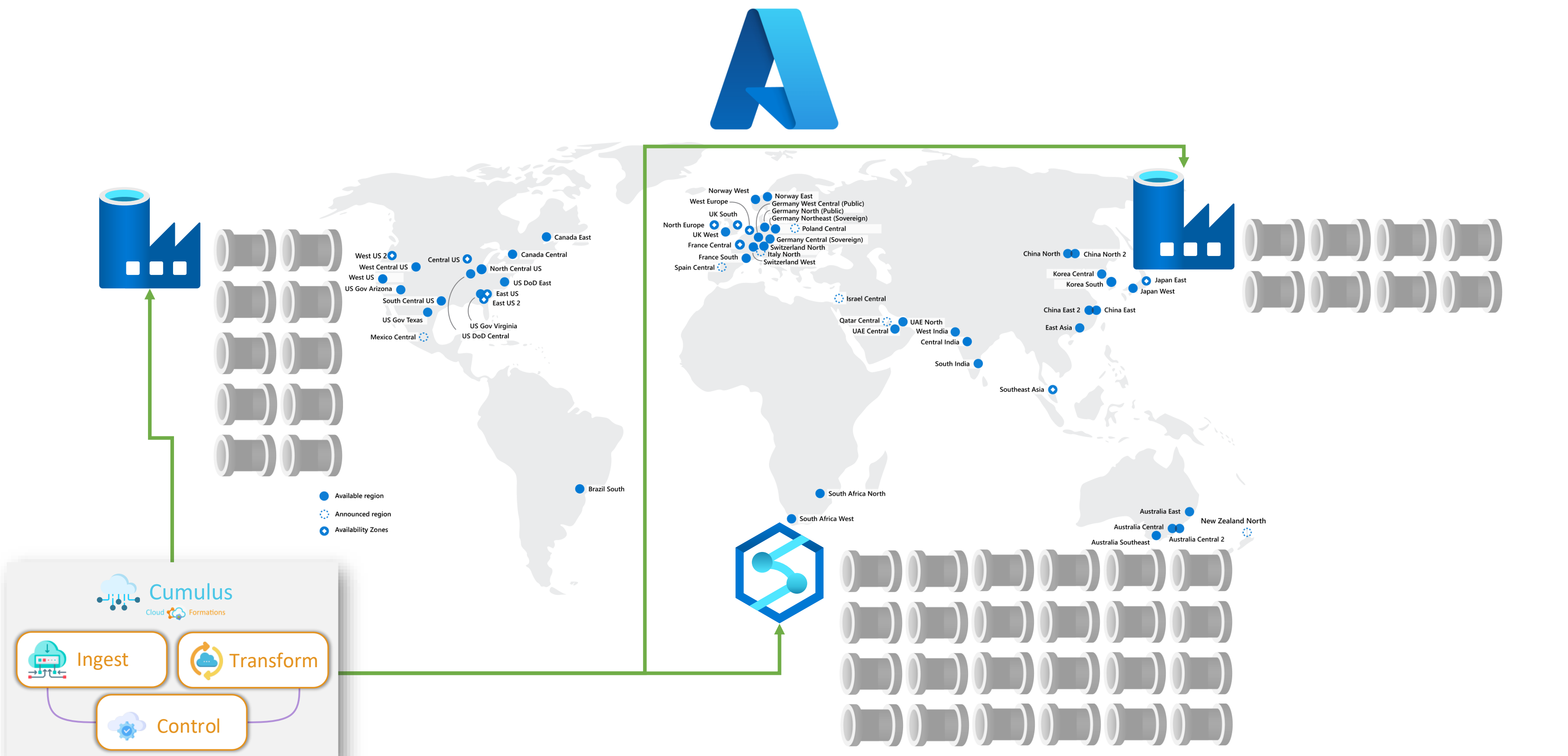
Role: Worker validator, executor, monitor and reporting of the outcome for the single worker pipeline.



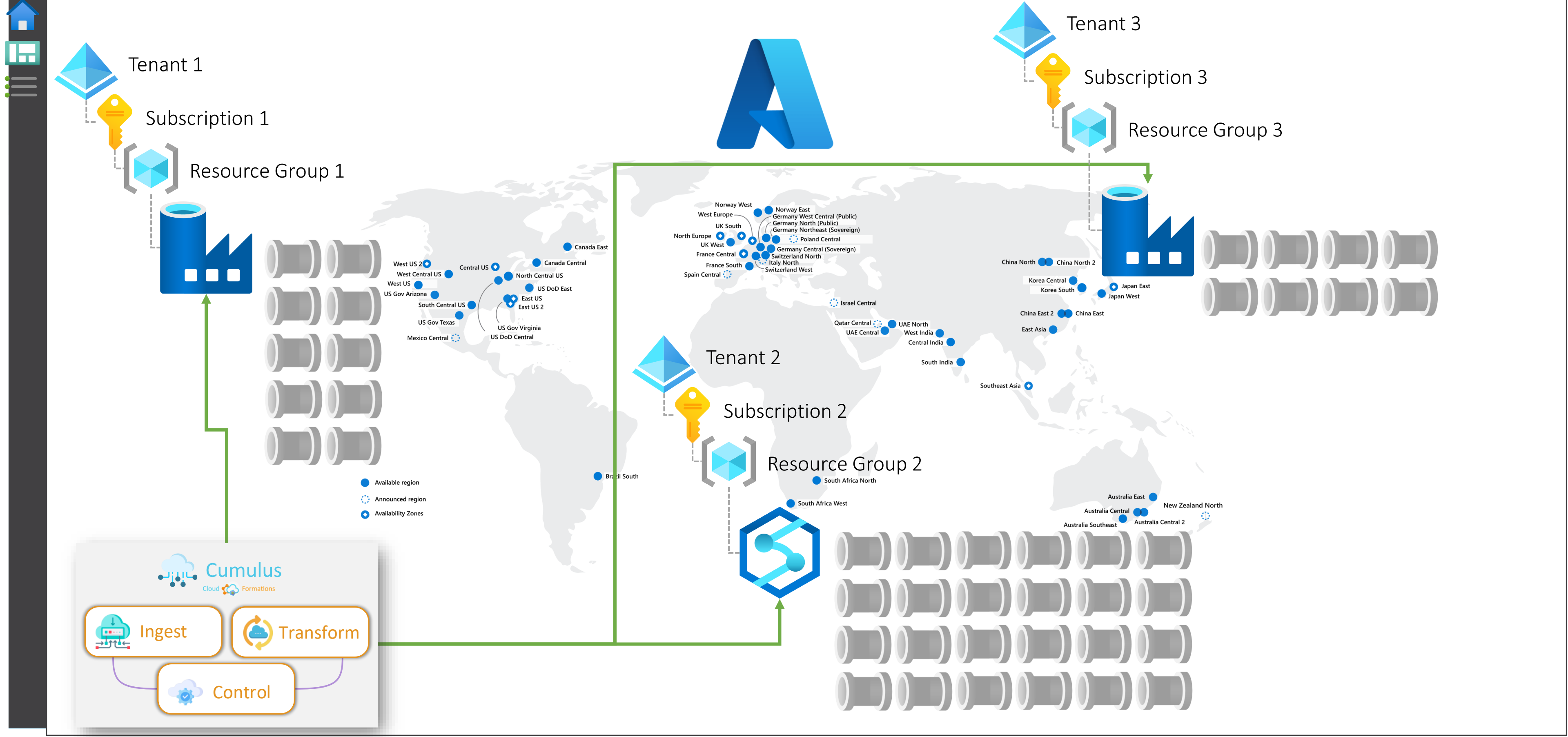
- Worker

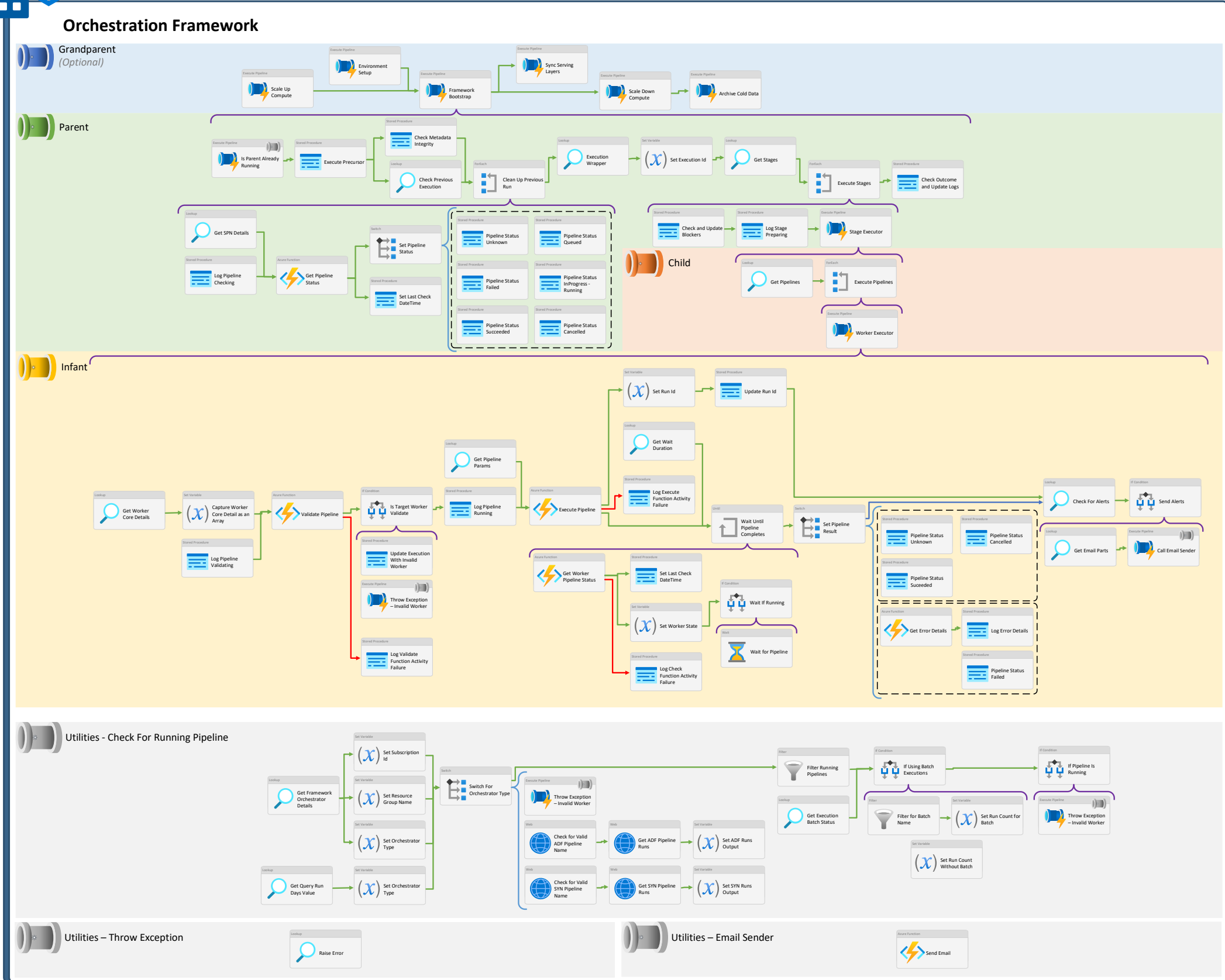
Role: Anything specific to the process needing to be performed.



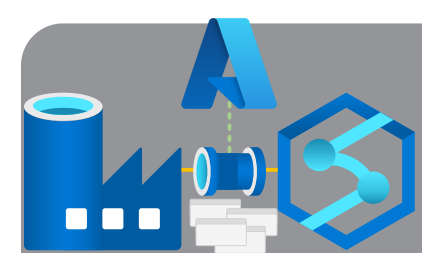


Use Metadata to Drive Integration Pipeline execution

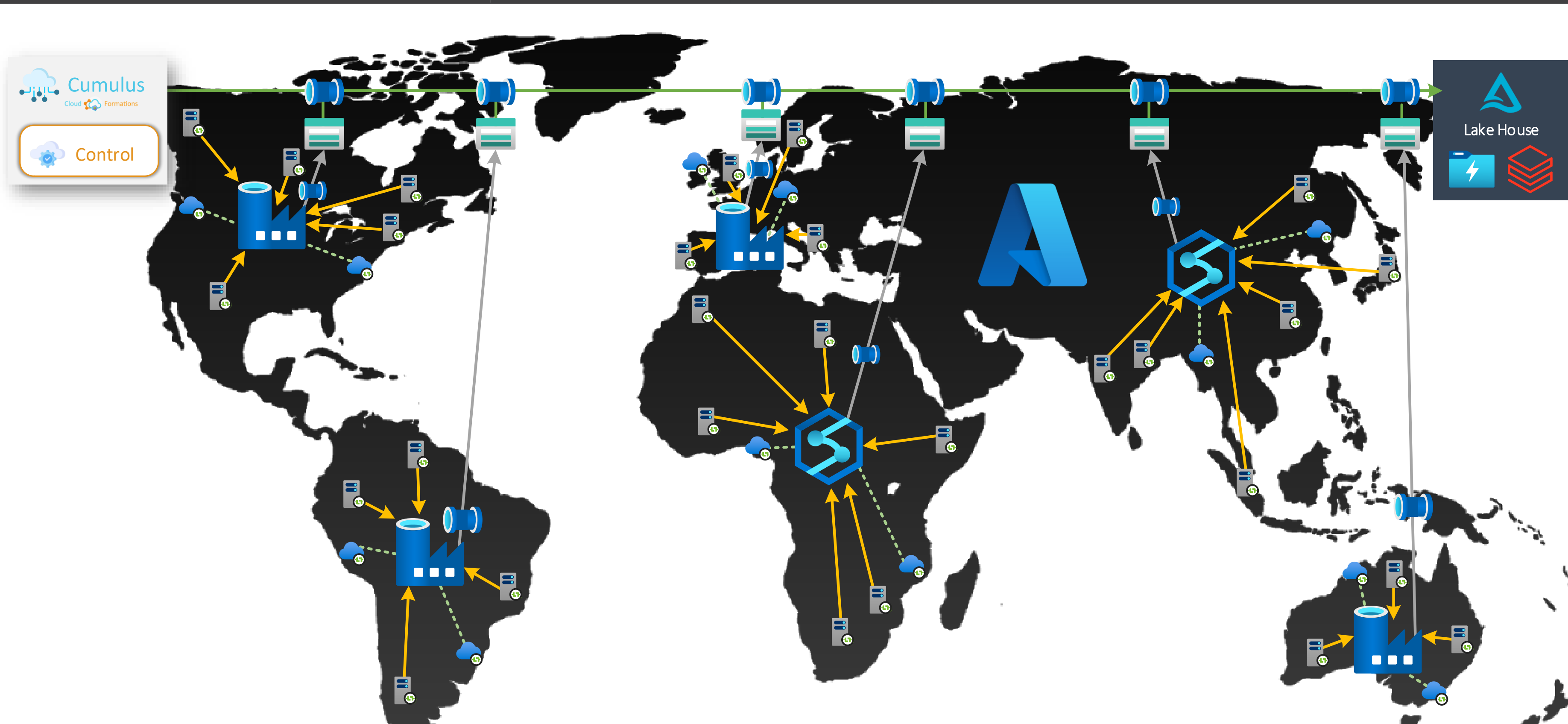




- ### Worker Pipelines
- Worker 1 - Extract
 - Worker 2 - Clean
 - Worker 3 - Transform
 - Worker 4 - Load
 - Worker 5 - Serve
 - Worker n -

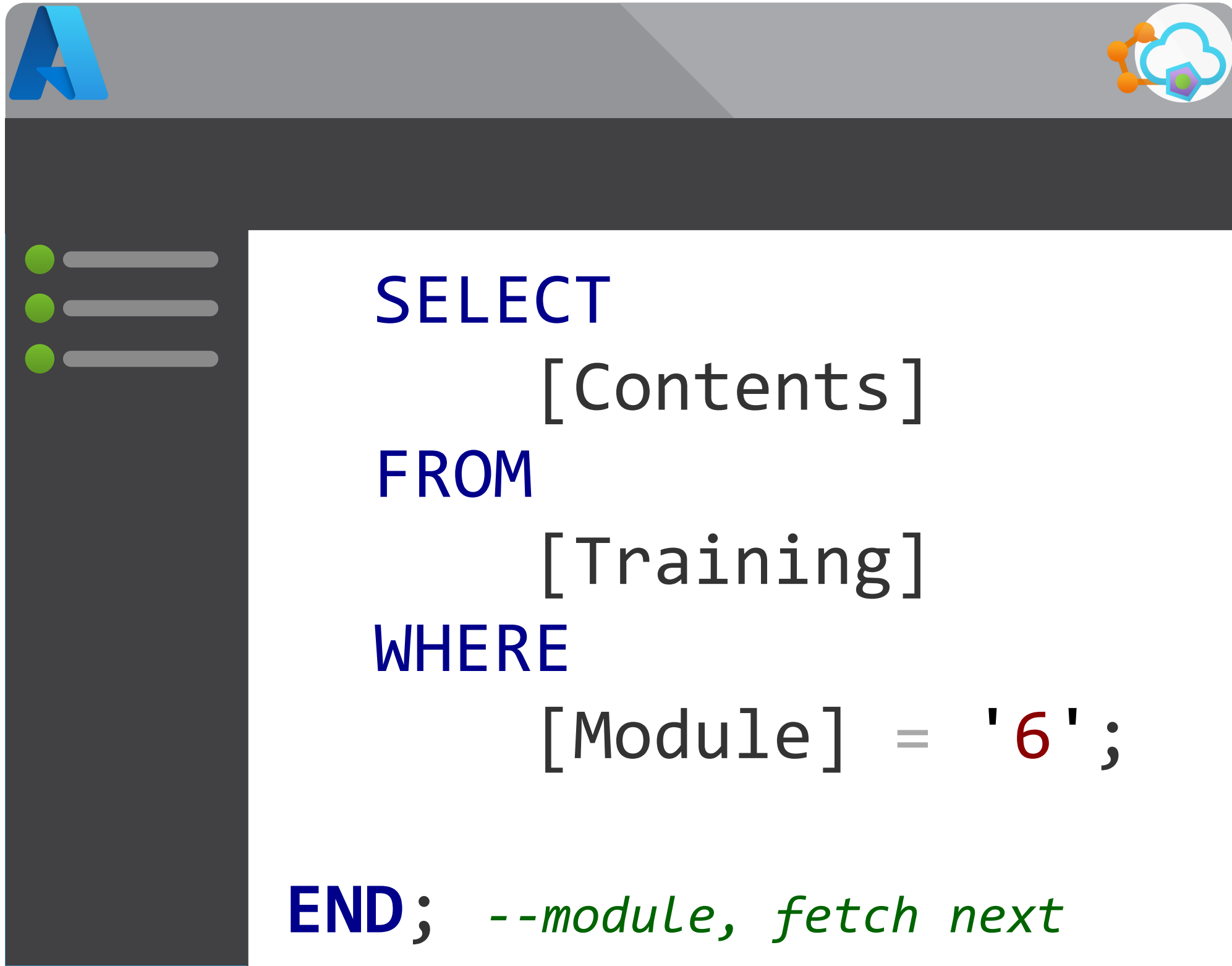


Hub & Spoke Integration Architecture



Module 6

Execution Parallelism



- Control Flow Scale Out
- Concurrency Limitations
- Internal vs External Activities
- Orchestration Frameworks

Note to self; Scale
down the database
because £££££!!!