Creating a Metadata Driven Orchestration Framework

With Azure Data Factory

Paul Andrew | Principal Consultant & Solution

Microsoft® Most Valuable Professional

altius

@MrPaulAndrew In/MrPaulAndrew MrPaulAndrew.com

https://github.com/mrpaulandrew

CommunityEvents

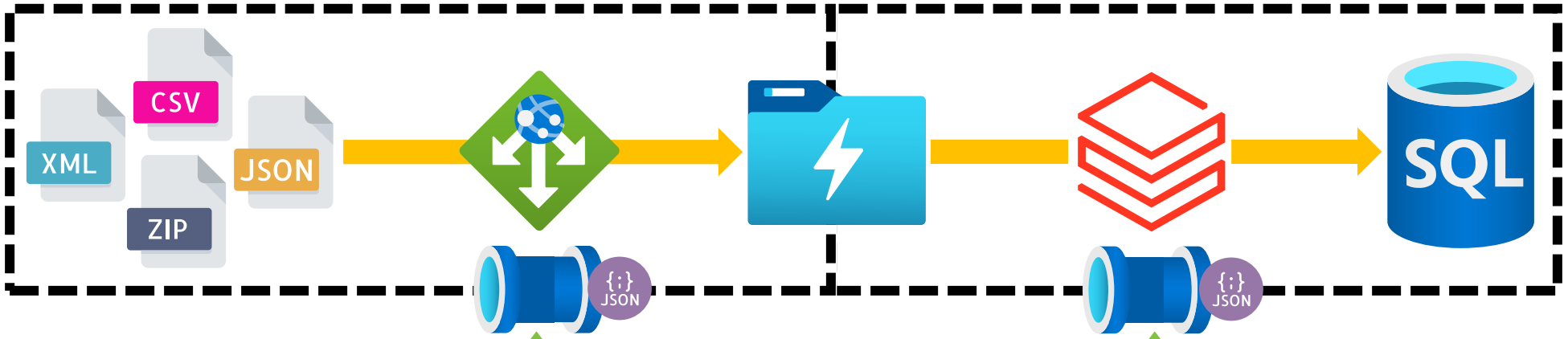
Demo code, content and slides from various community events.

C++

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

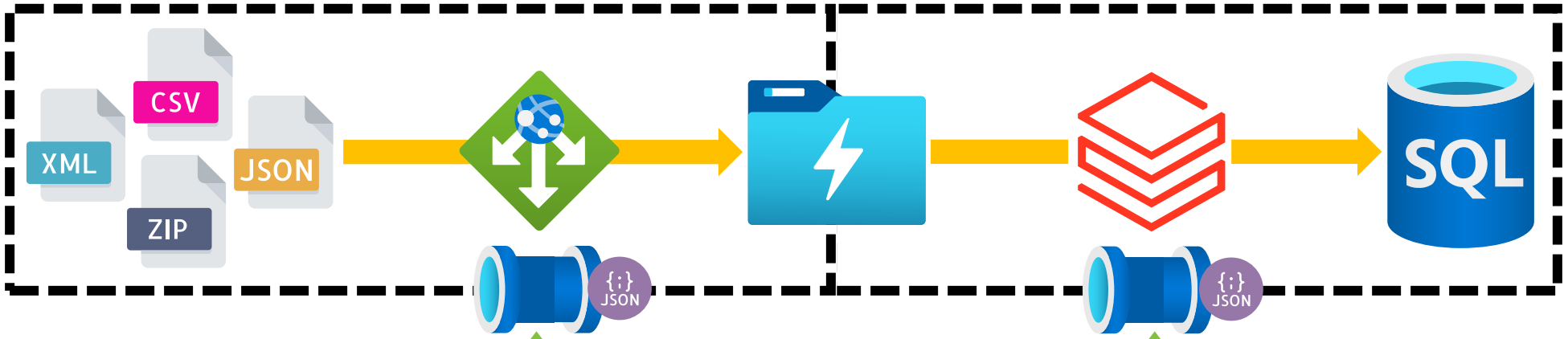
Data Factory

Data Factory A Quick Overview

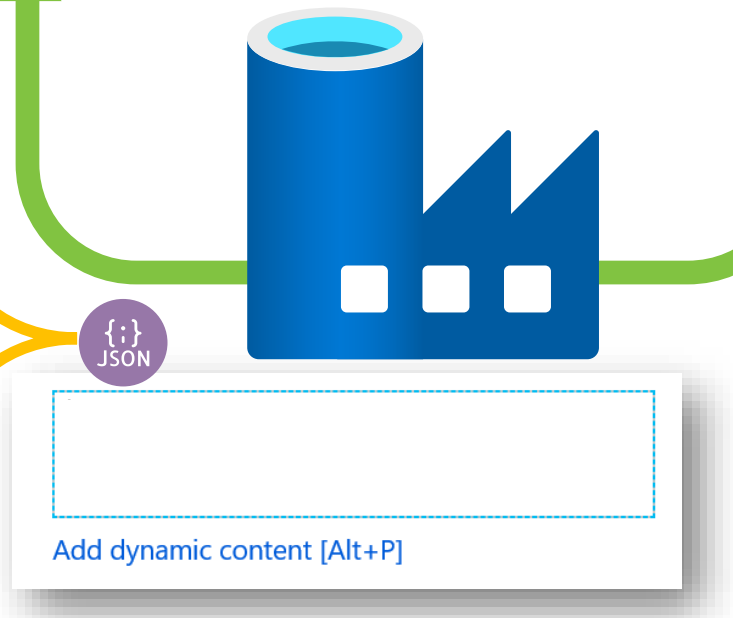


1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

Data Factory Components – Add Dynamic Content

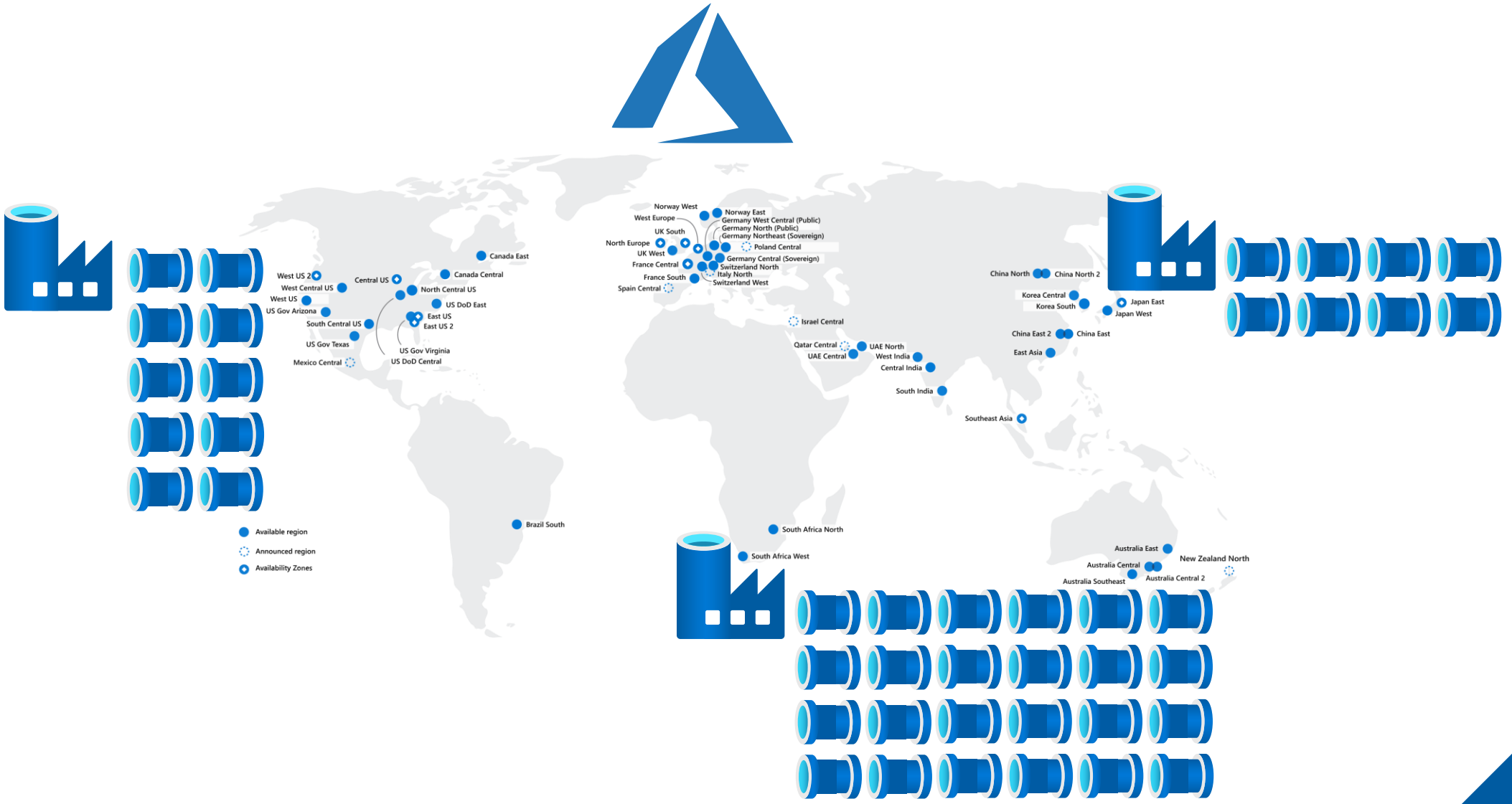


1. **Linked Services**
2. **Datasets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

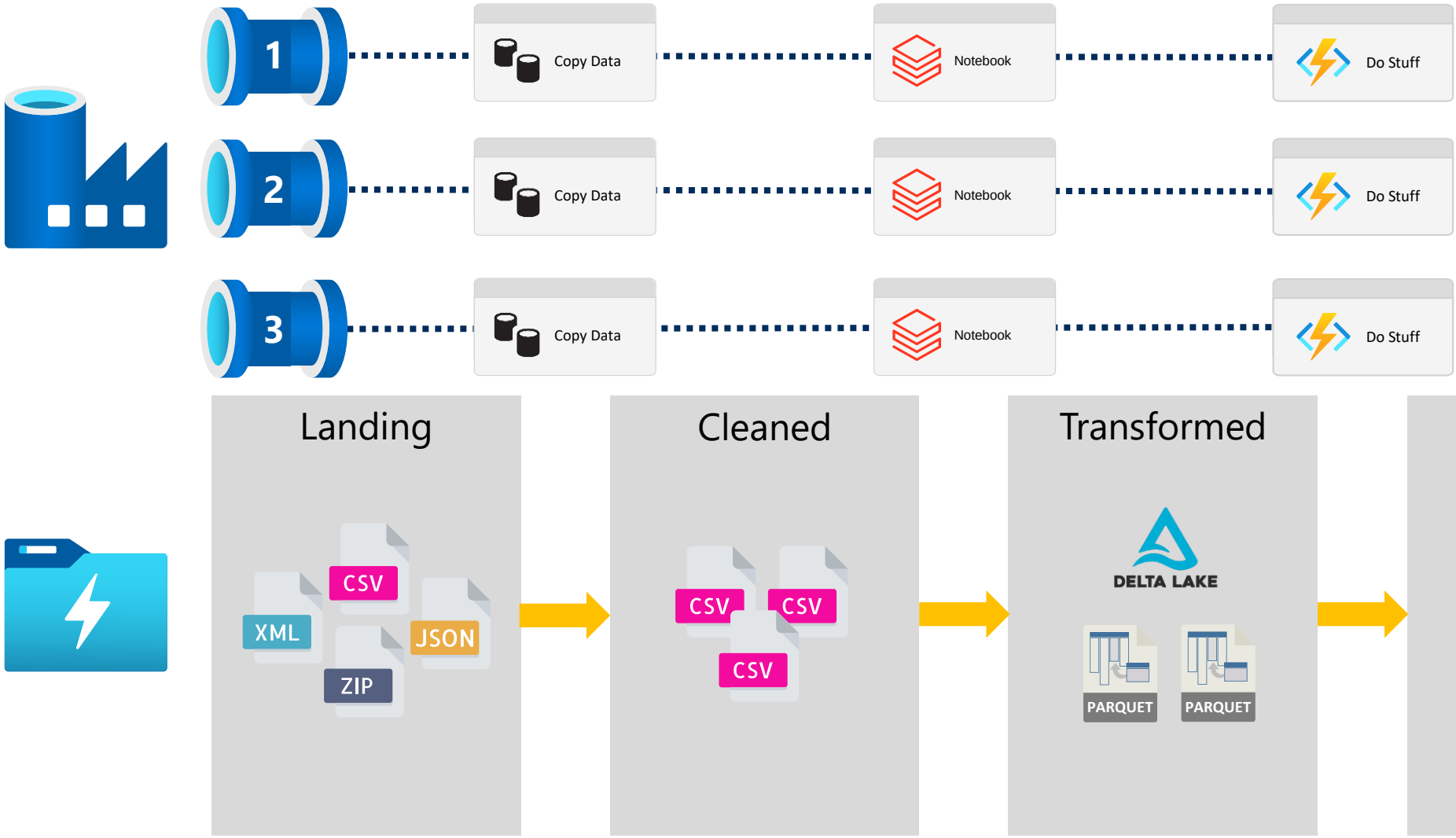


Problem

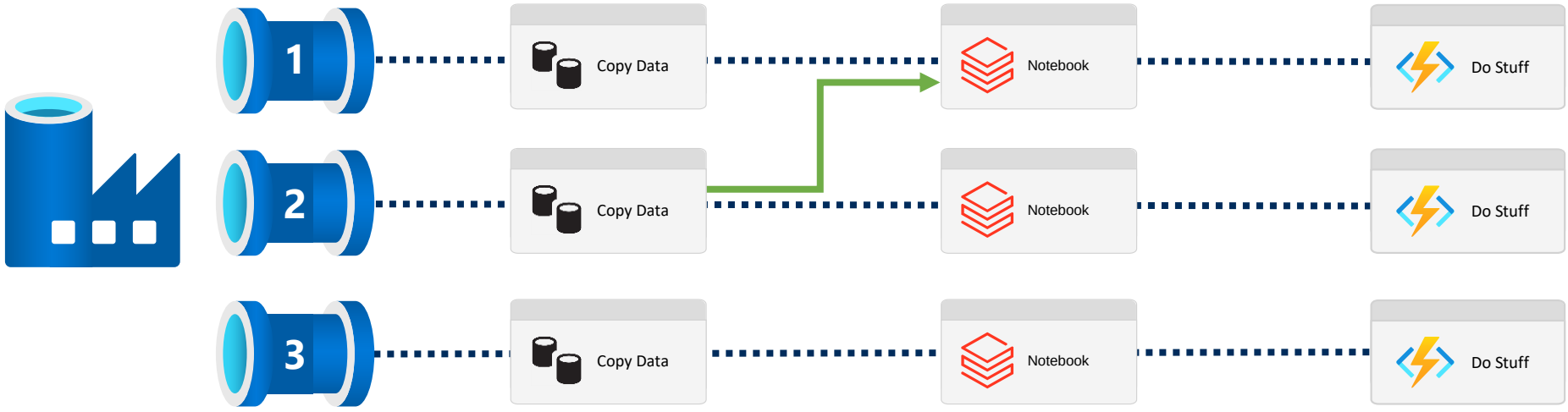
Problem: How should we structure our Data Factory Pipelines?

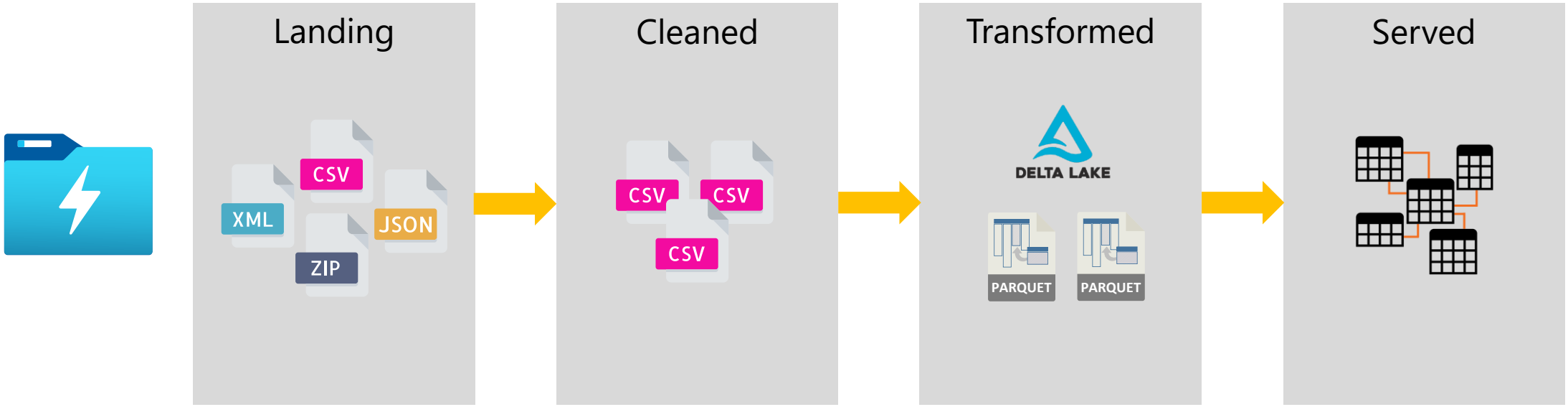


Problem

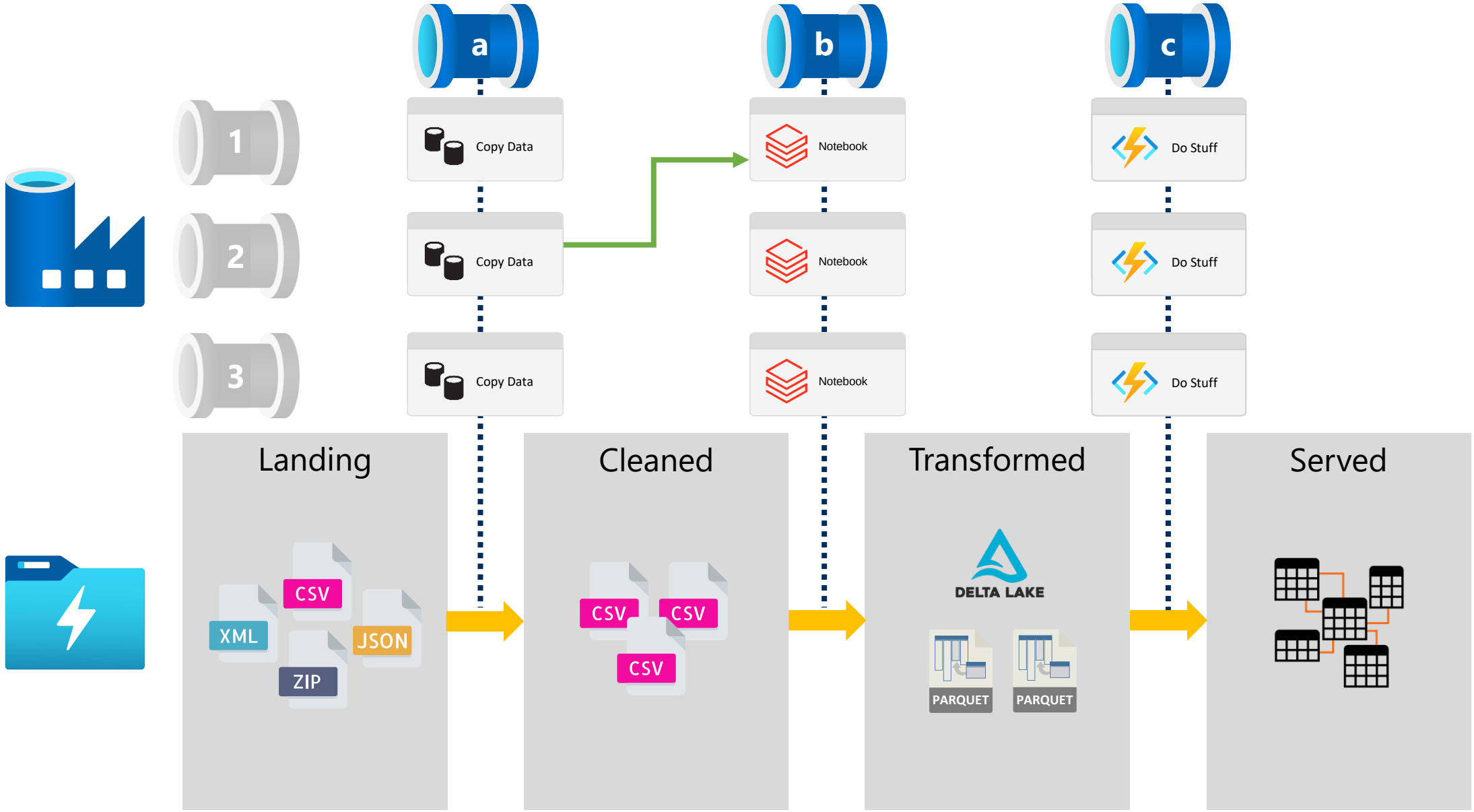

1
Copy Data
Notebook
Do Stuff
2
Copy Data
Notebook
Do Stuff
3
Copy Data
Notebook
Do Stuff
Landing
CSV
XML
JSON
ZIP
Cleaned
CSV
CSV
CSV
Transformed
DELTA LAKE
PARQUET
PARQUET
Served

Problem

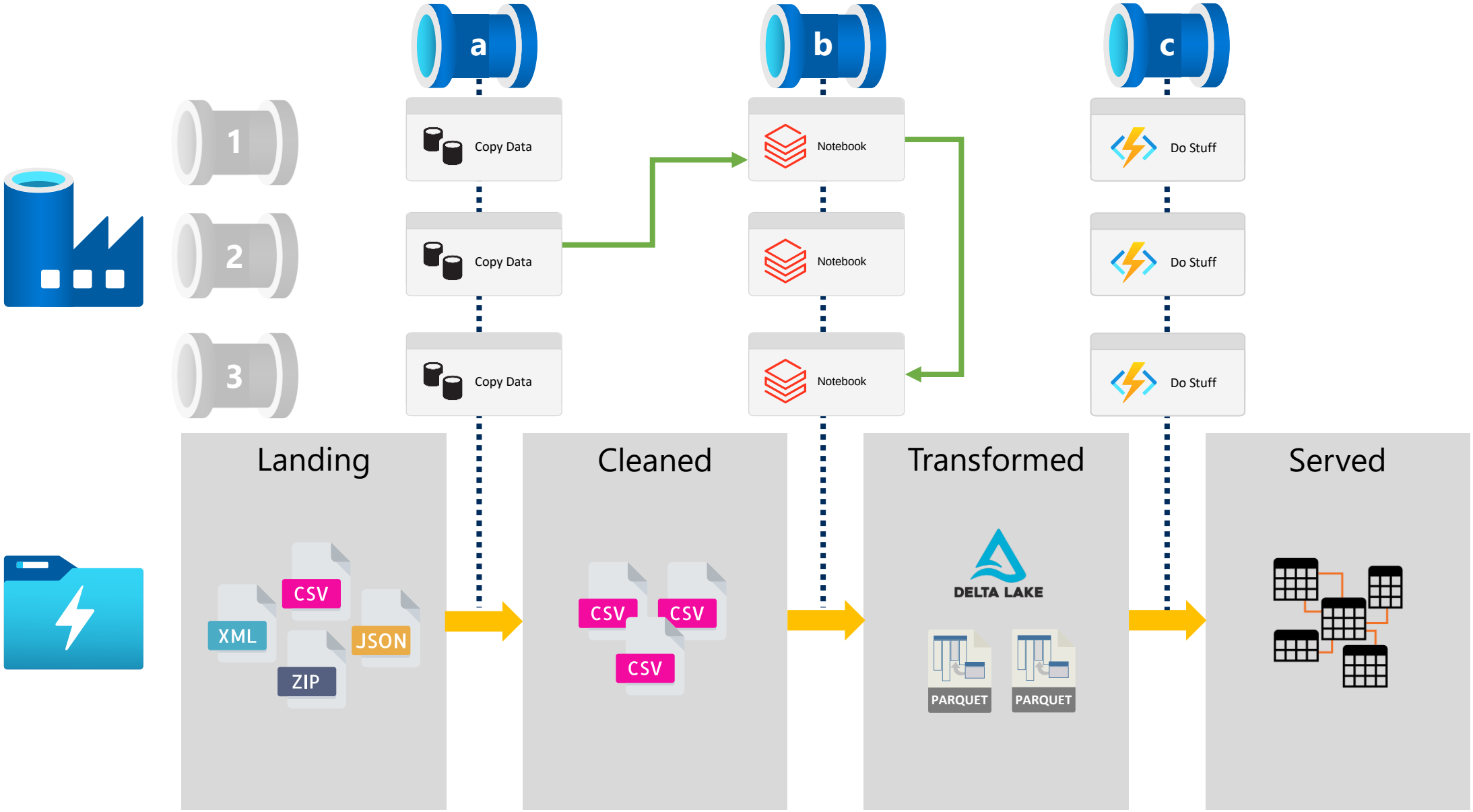

1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff


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

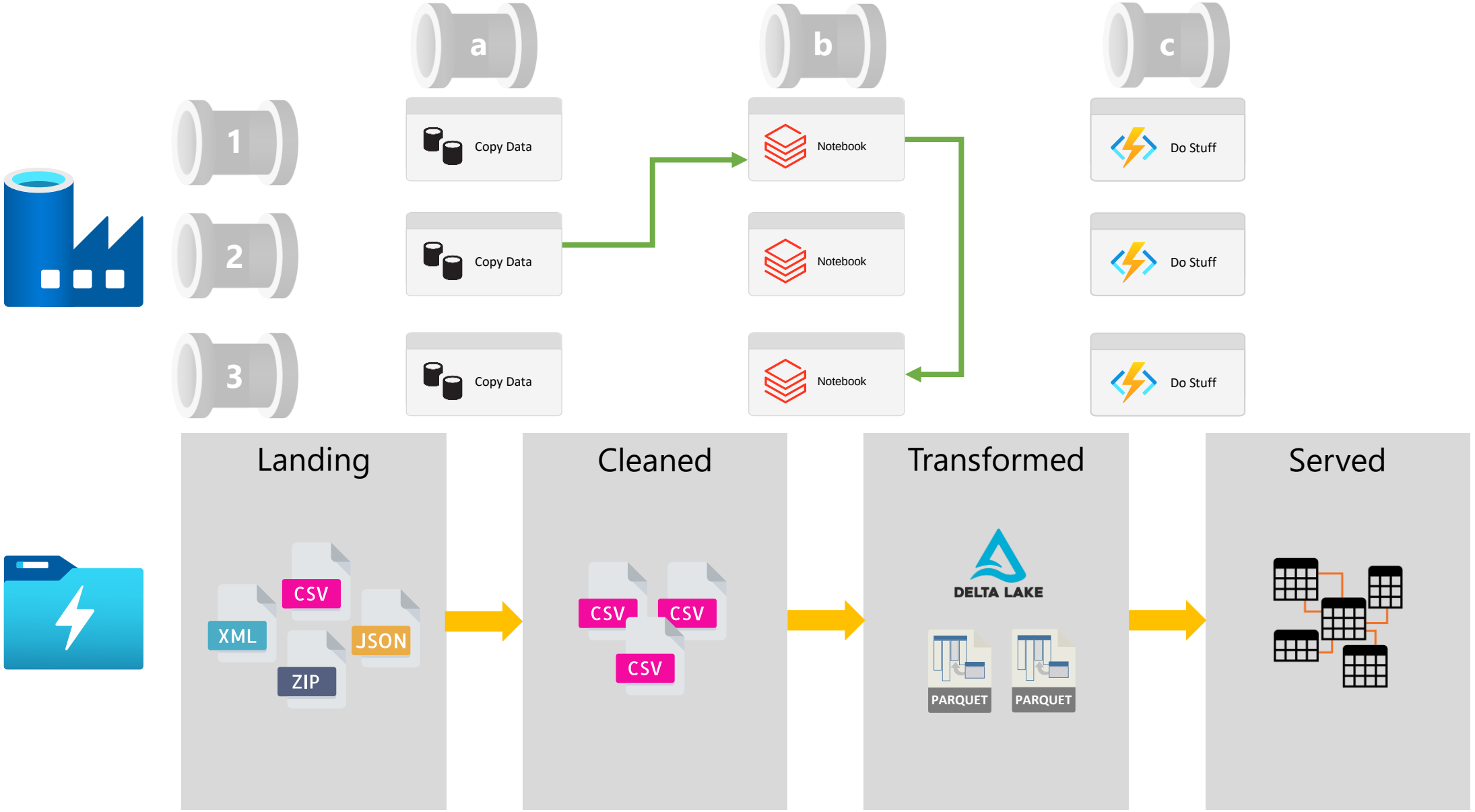
Problem


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

Problem

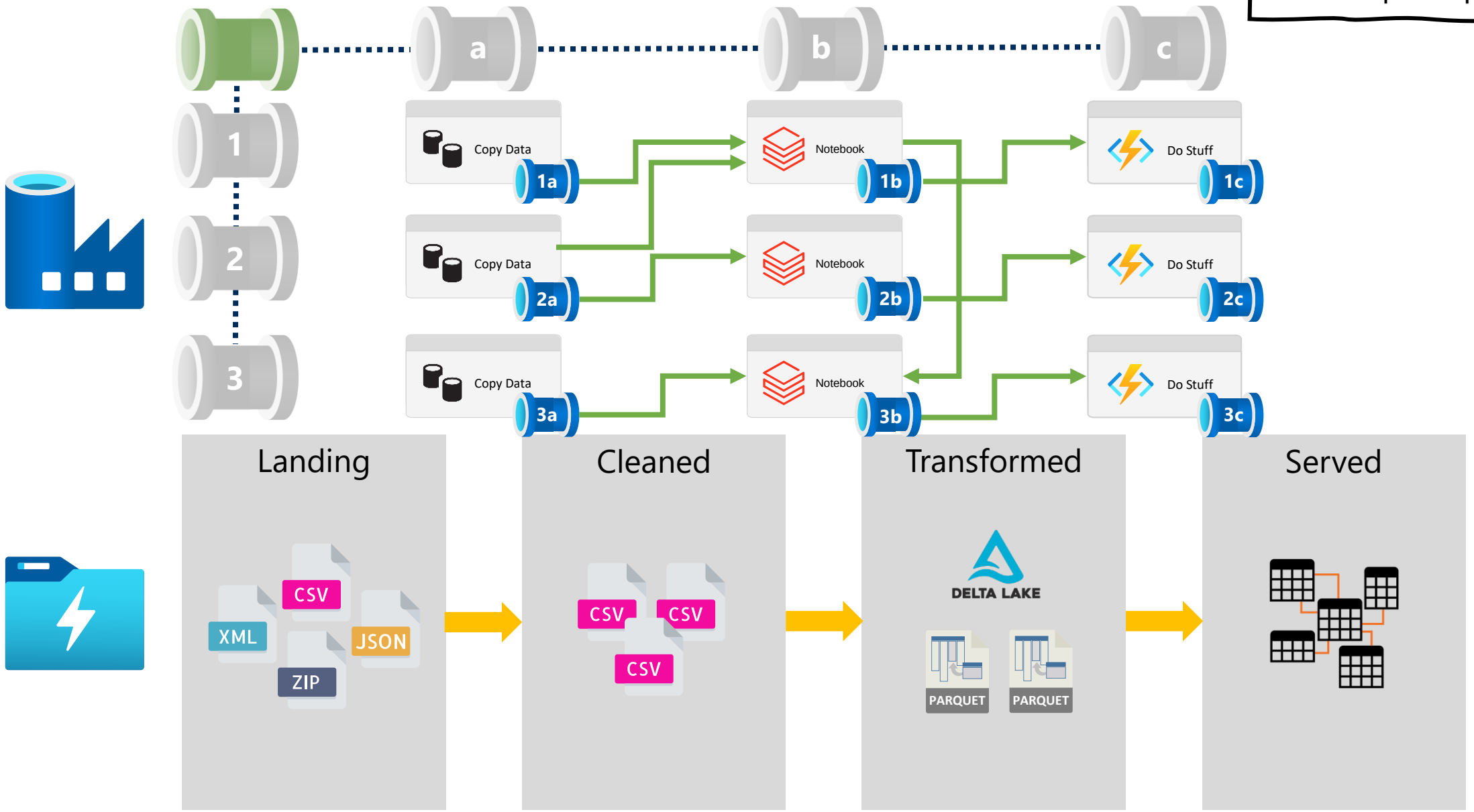

a
b
c
1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff
Landing
Cleaned
Transformed
Served
CSV
XML
JSON
ZIP
CSV
CSV
CSV
DELTA LAKE
PARQUET
PARQUET

Problem

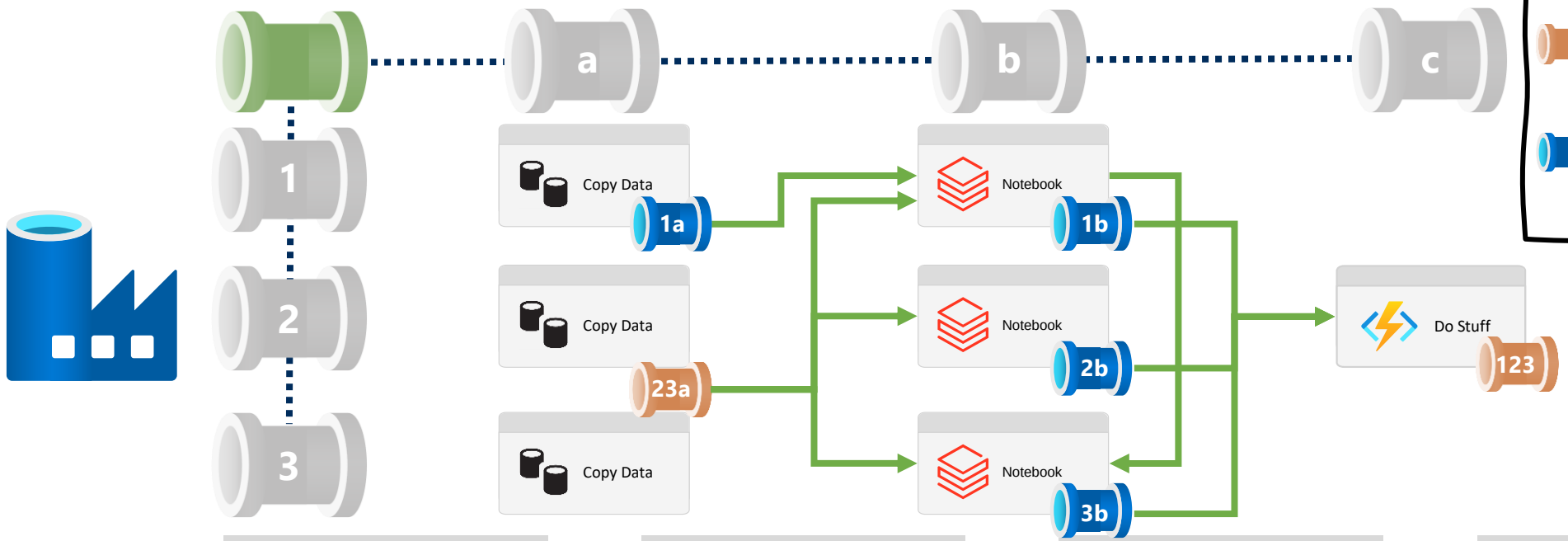

a
b
c
1
2
3
Copy Data
Copy Data
Copy Data
Notebook
Notebook
Notebook
Do Stuff
Do Stuff
Do Stuff
Landing
Cleaned
Transformed
Served
CSV
XML
JSON
ZIP
CSV
CSV
CSV
DELTA LAKE
PARQUET
PARQUET

Problem

Only 40 Activities per Pipeline.



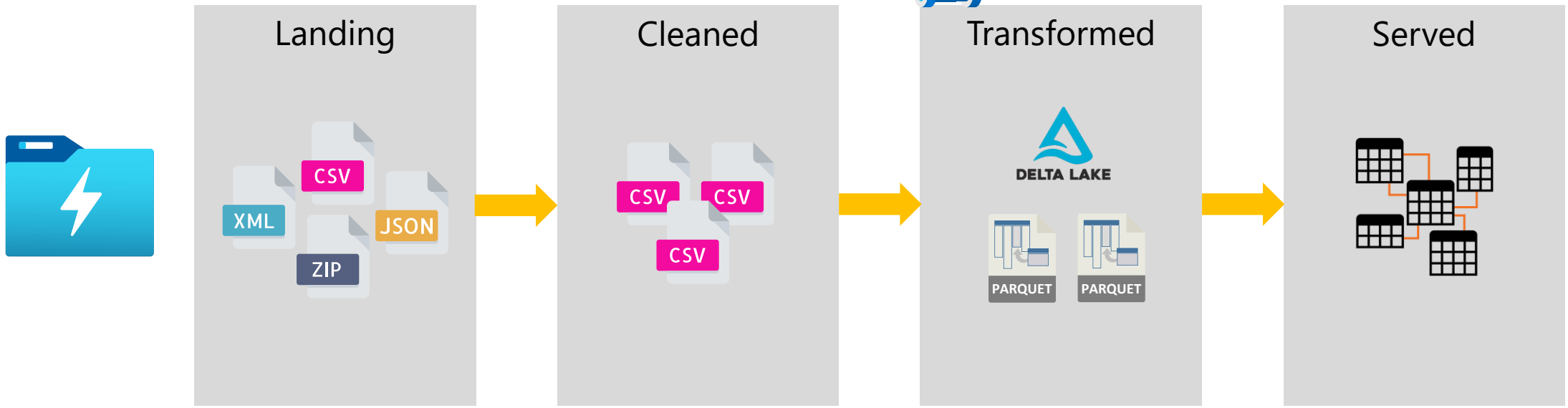
Problem



Grandparent pipeline for all processing.

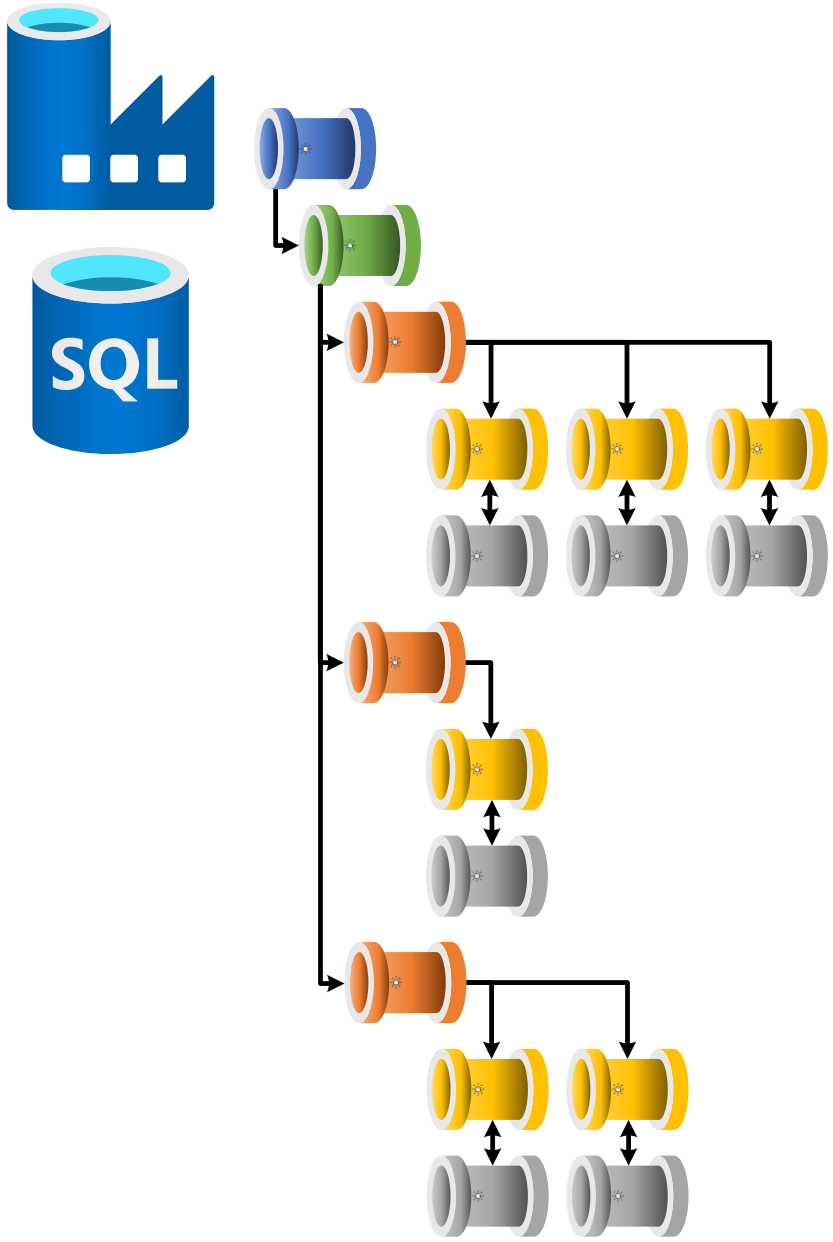
Parent pipeline to consolidate work.

Child pipelines for low level dependencies.

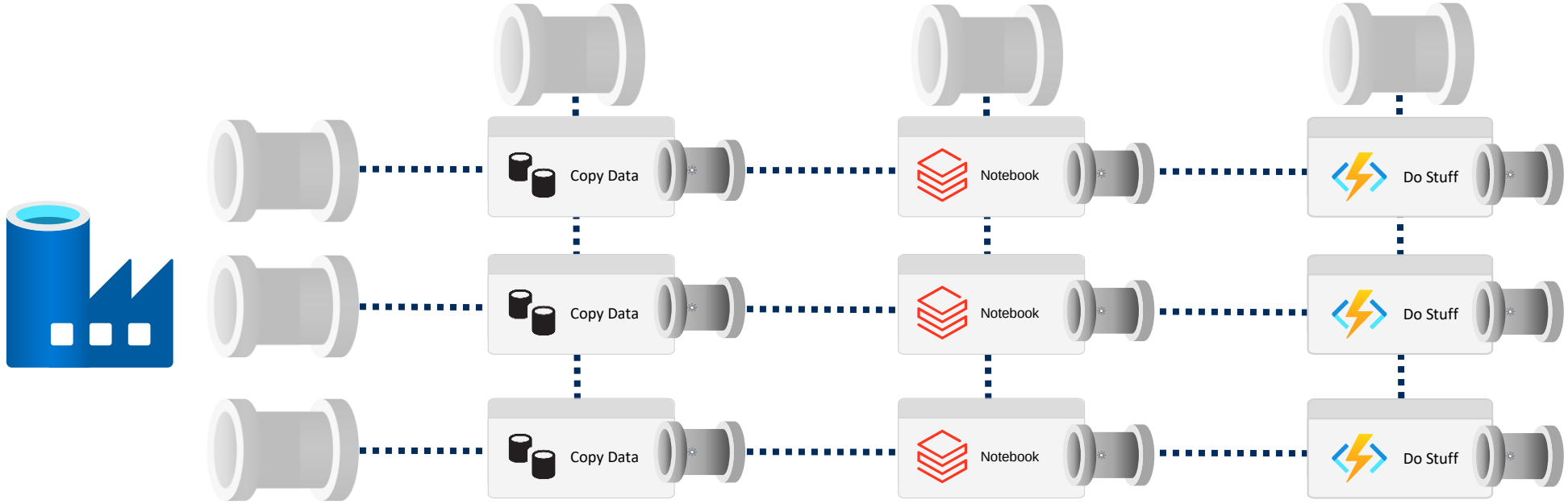


Solution

Solution: Use Metadata to Drive Data Factory Pipelines



Solution

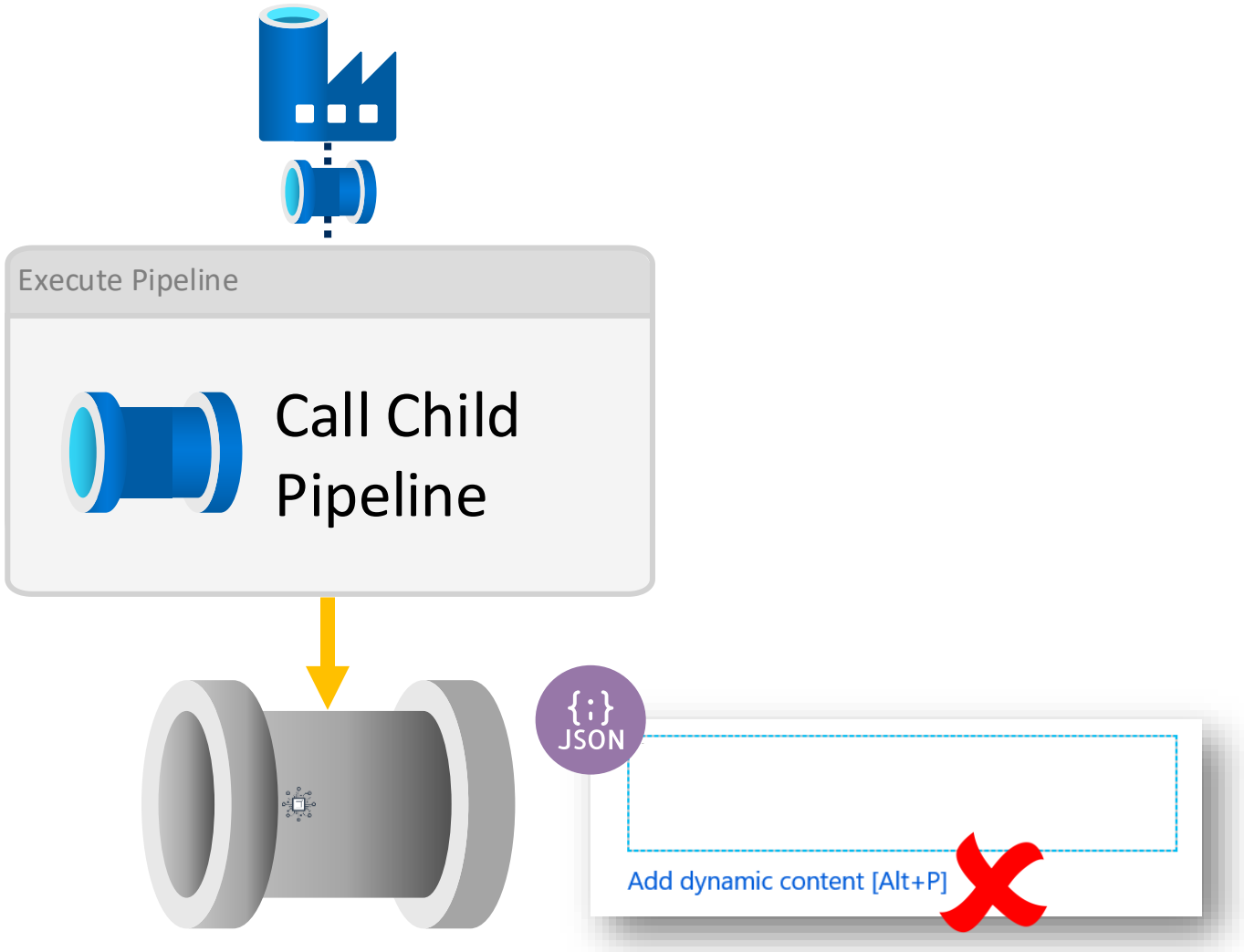


Stages
1
2
3

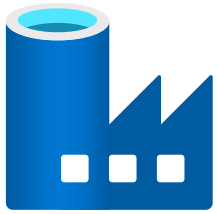
Pipelines
a
b
c
d
e
f
g
h
i

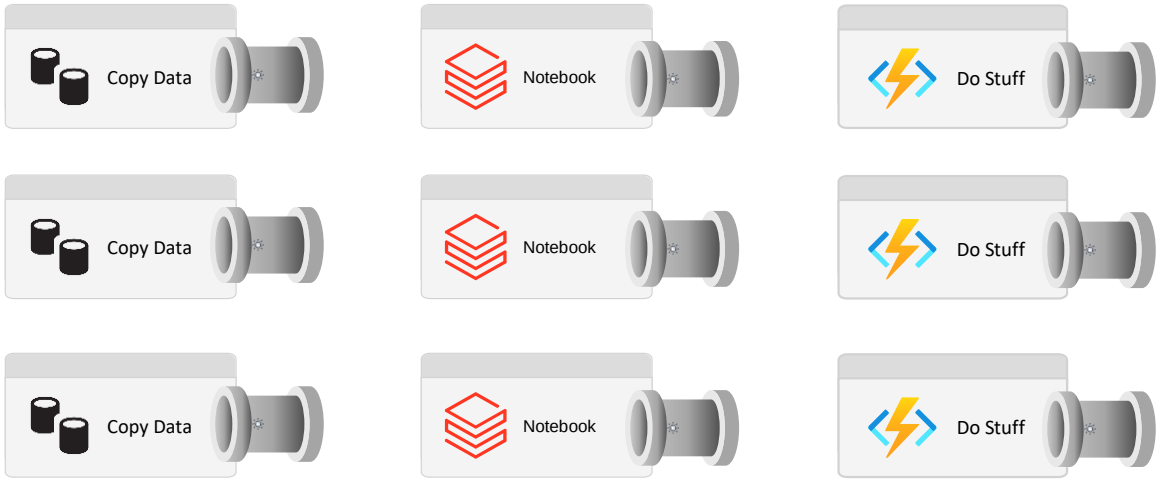
Stage	Pipeline
1	a
1	b
1	c
2	d
2	e
3	f
3	g
3	h
3	i

One More Problem



Calling Our Worker Pipelines



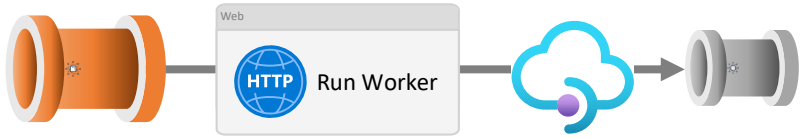


Stages
1
2
3

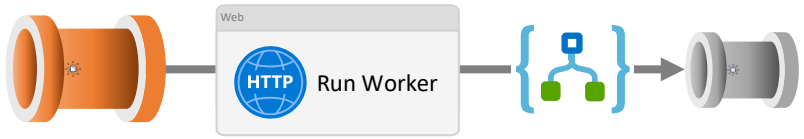
Pipelines
a
b
c
d
e
f
g
h
i

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

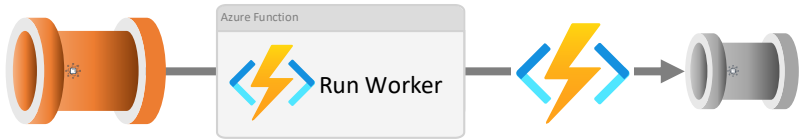
Option 1:



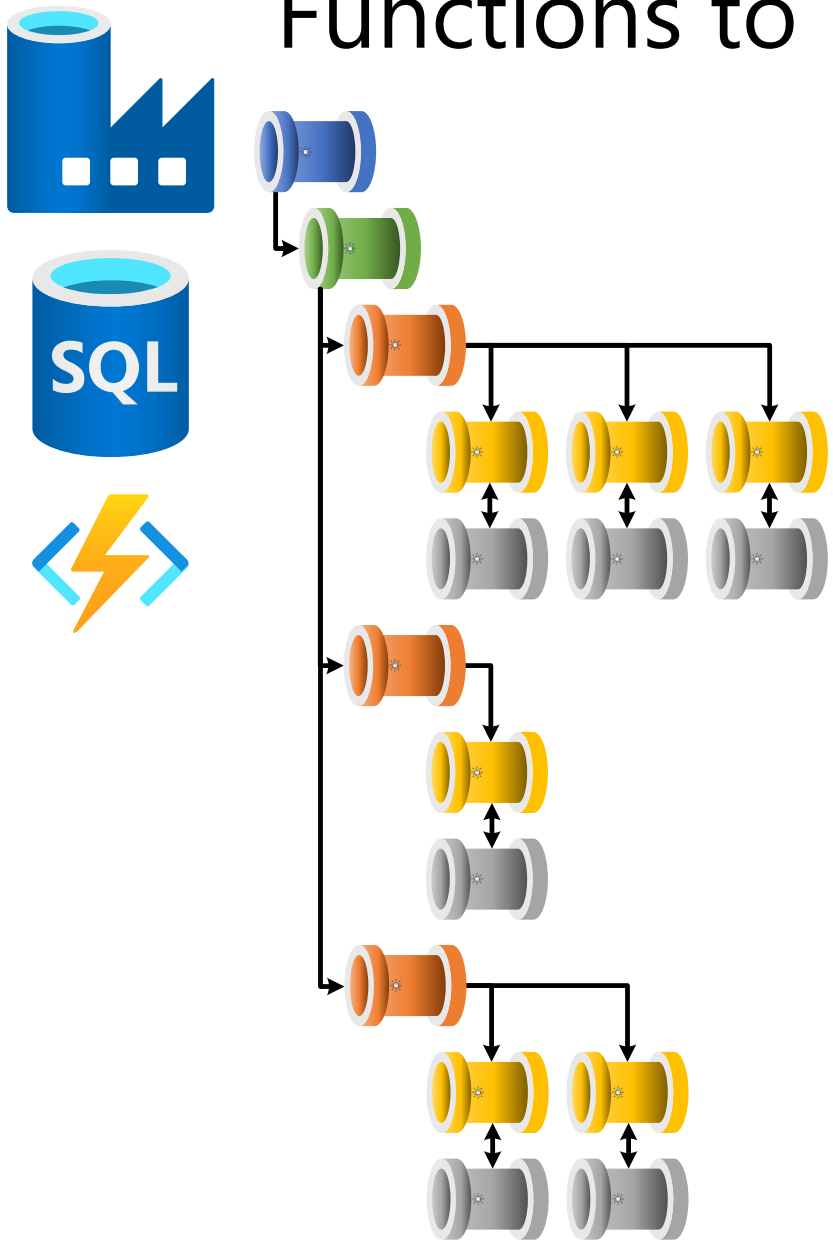
Option 2:



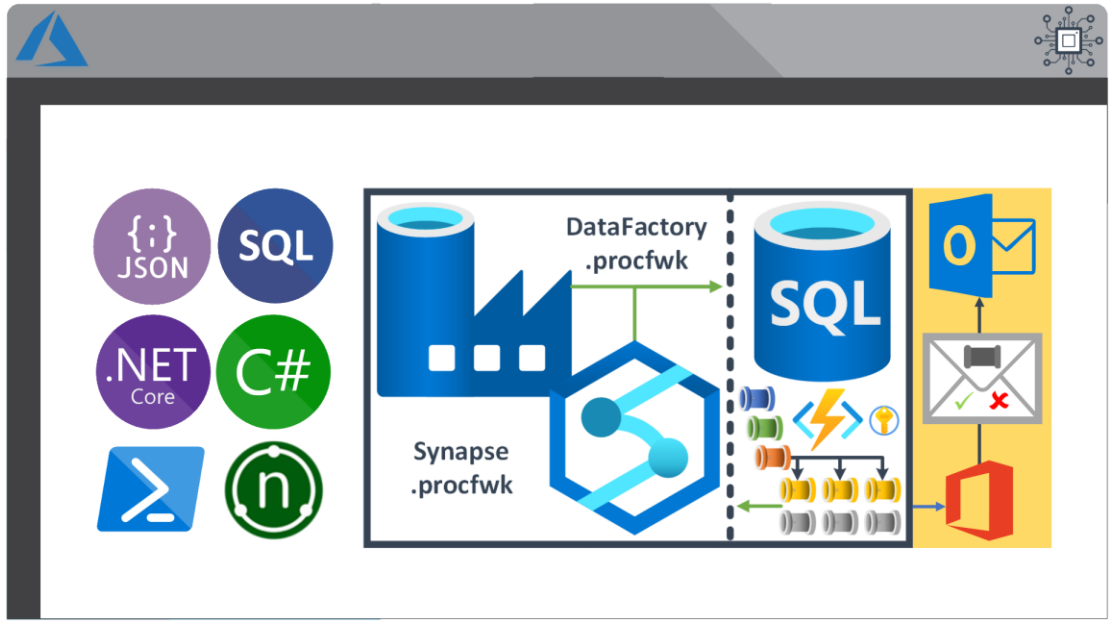
Option 3:



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

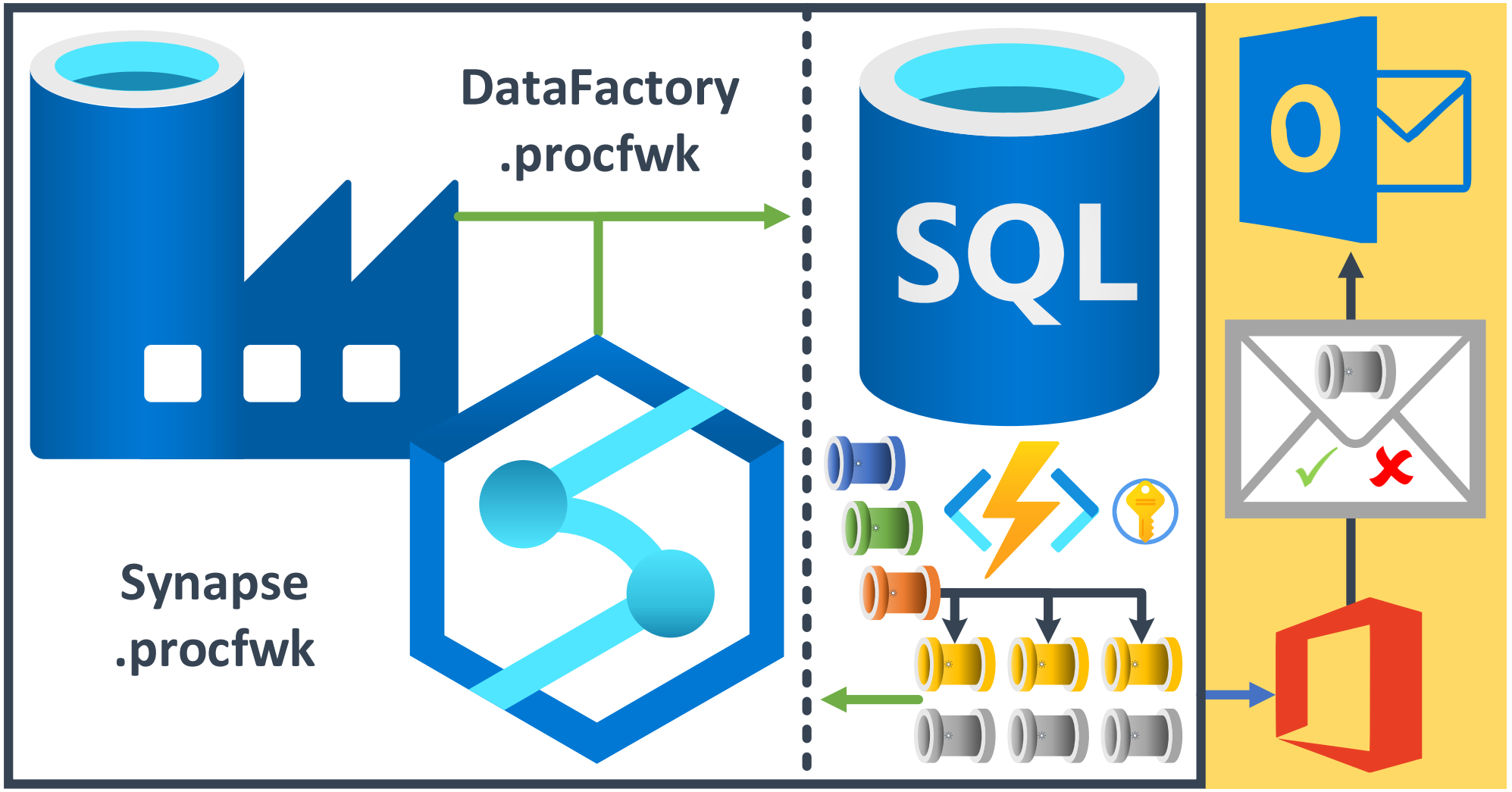


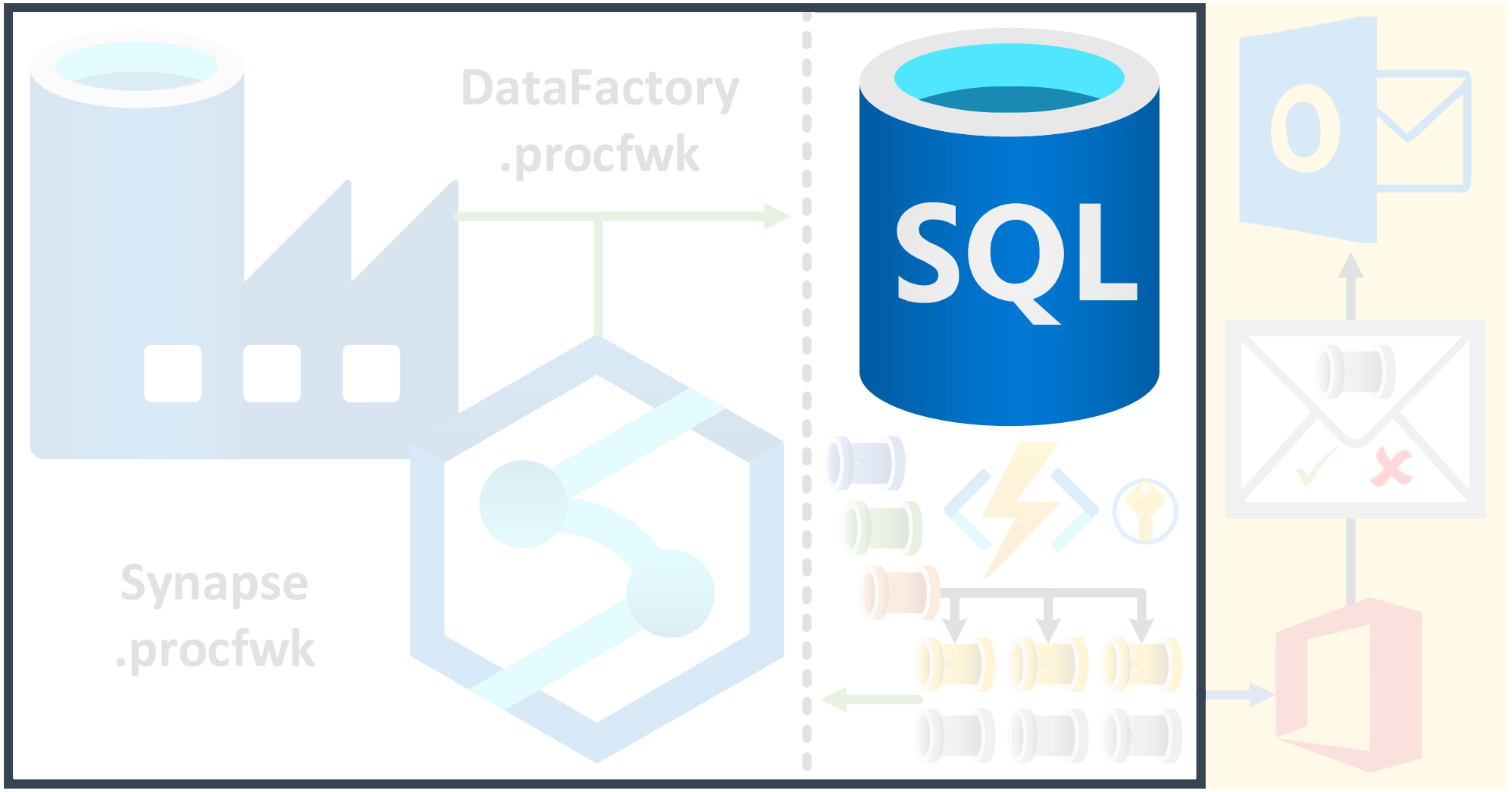
Introducing procfwk.com



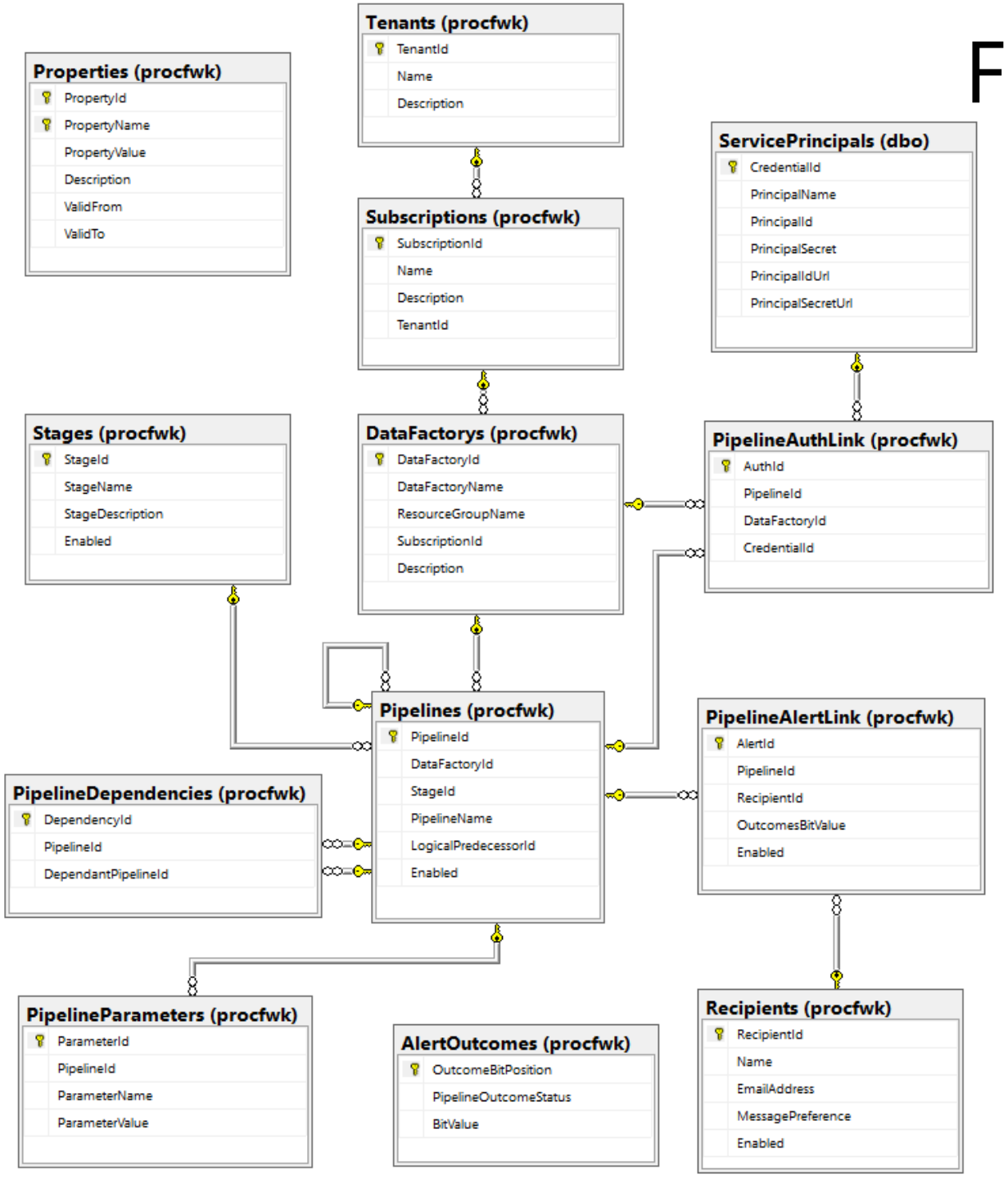
procfwk Features

- Granular metadata control.
- Metadata integrity checking.
- Global properties.
- Complete pipeline dependency chains.
- Execution restart-ability.
- Parallel execution.
- Full execution and error logs.
- Operational dashboards.
- Low cost orchestration.
- Disconnection between framework and worker pipelines.
- Cross Tenant/Subscription/Data Factory control flows.
- Pipeline parameter support.
- Simple troubleshooting.
- Easy deployment.
- Email alerting.
- Automated testing.
- Azure Key Vault integration.

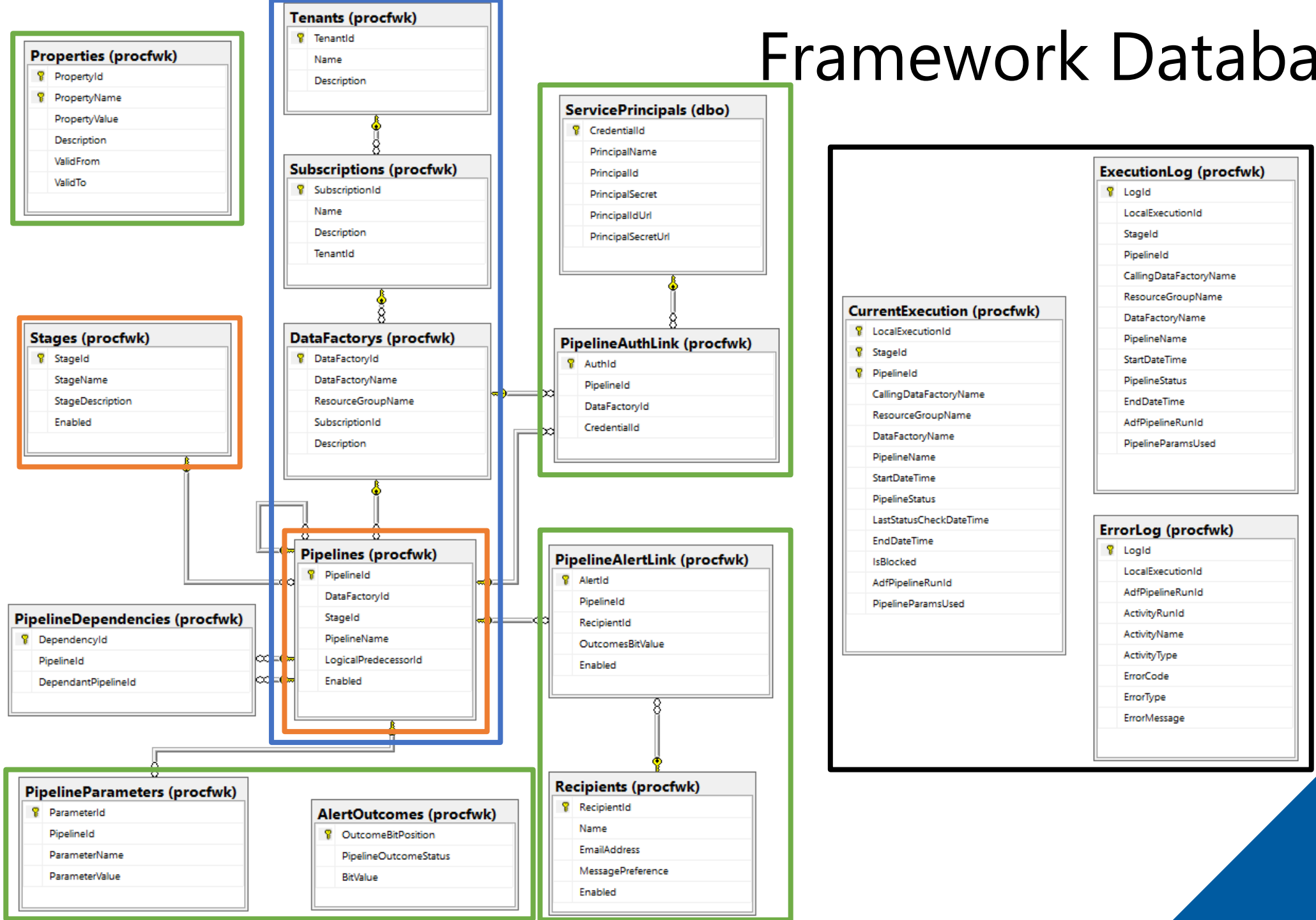

DataFactory
.procfwk
SQL
Synapse
.procfwk

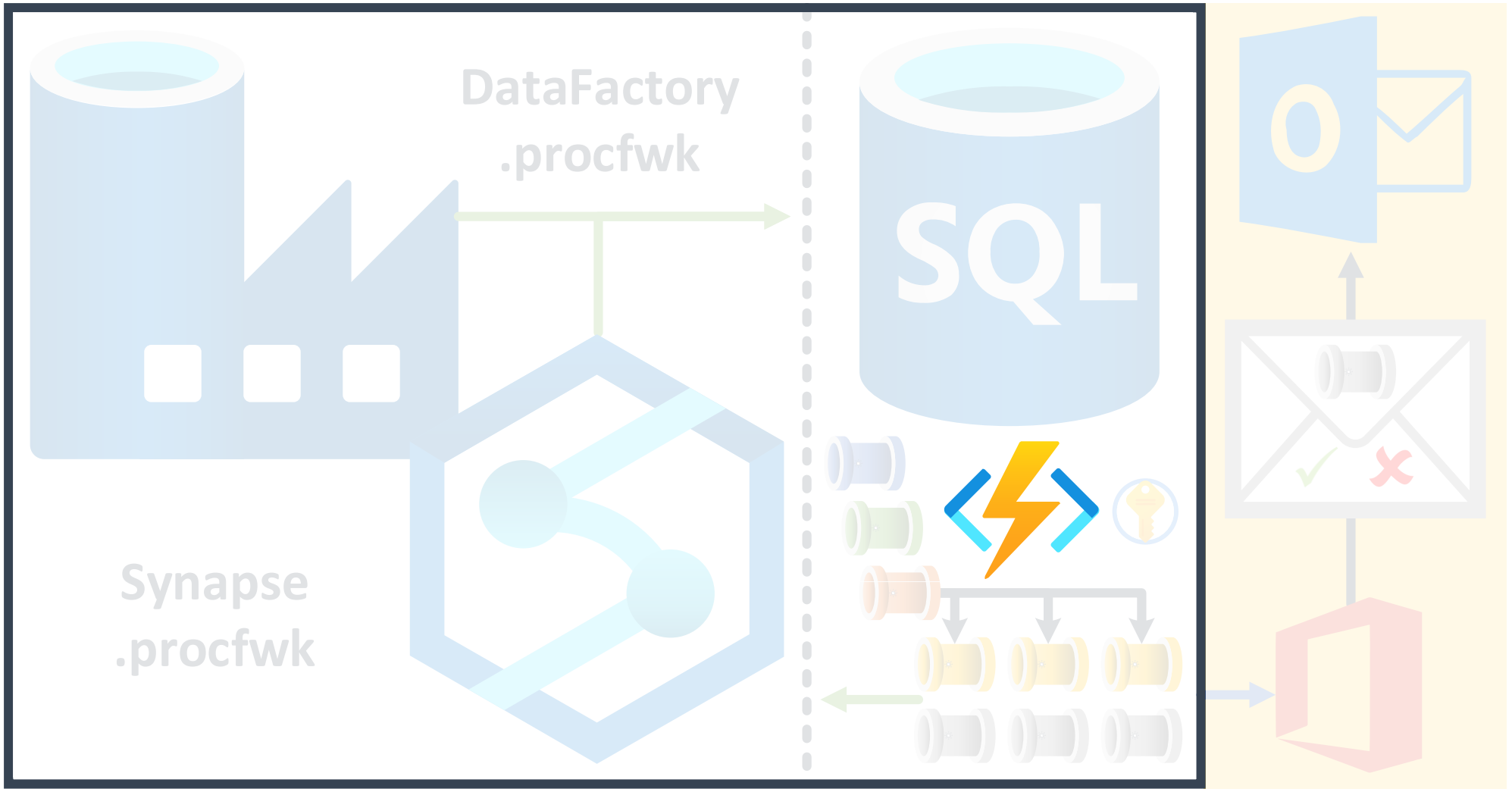

DataFactory
.procfwk
SQL
Synapse
.procfwk

Framework Database

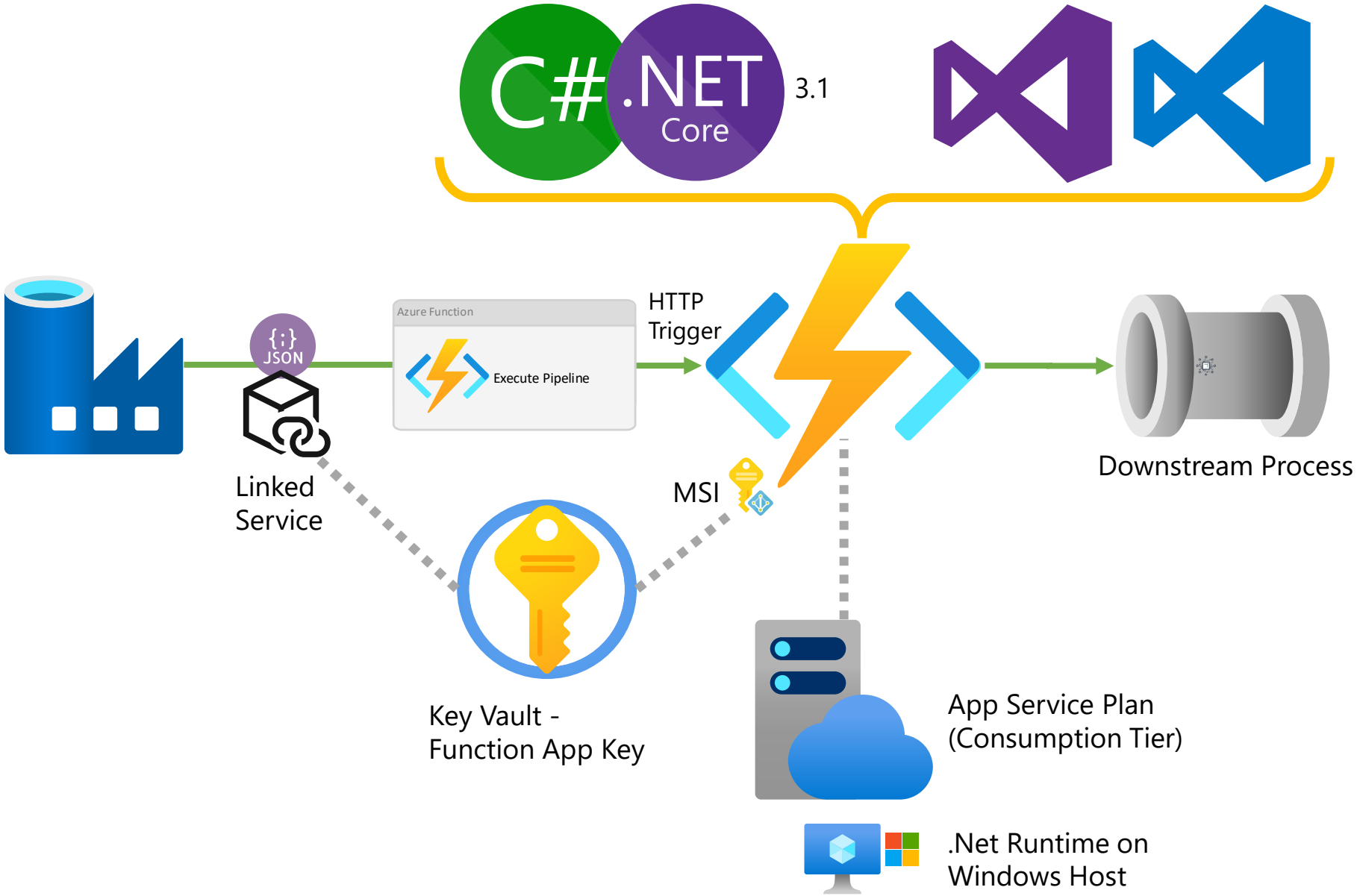


Framework Database

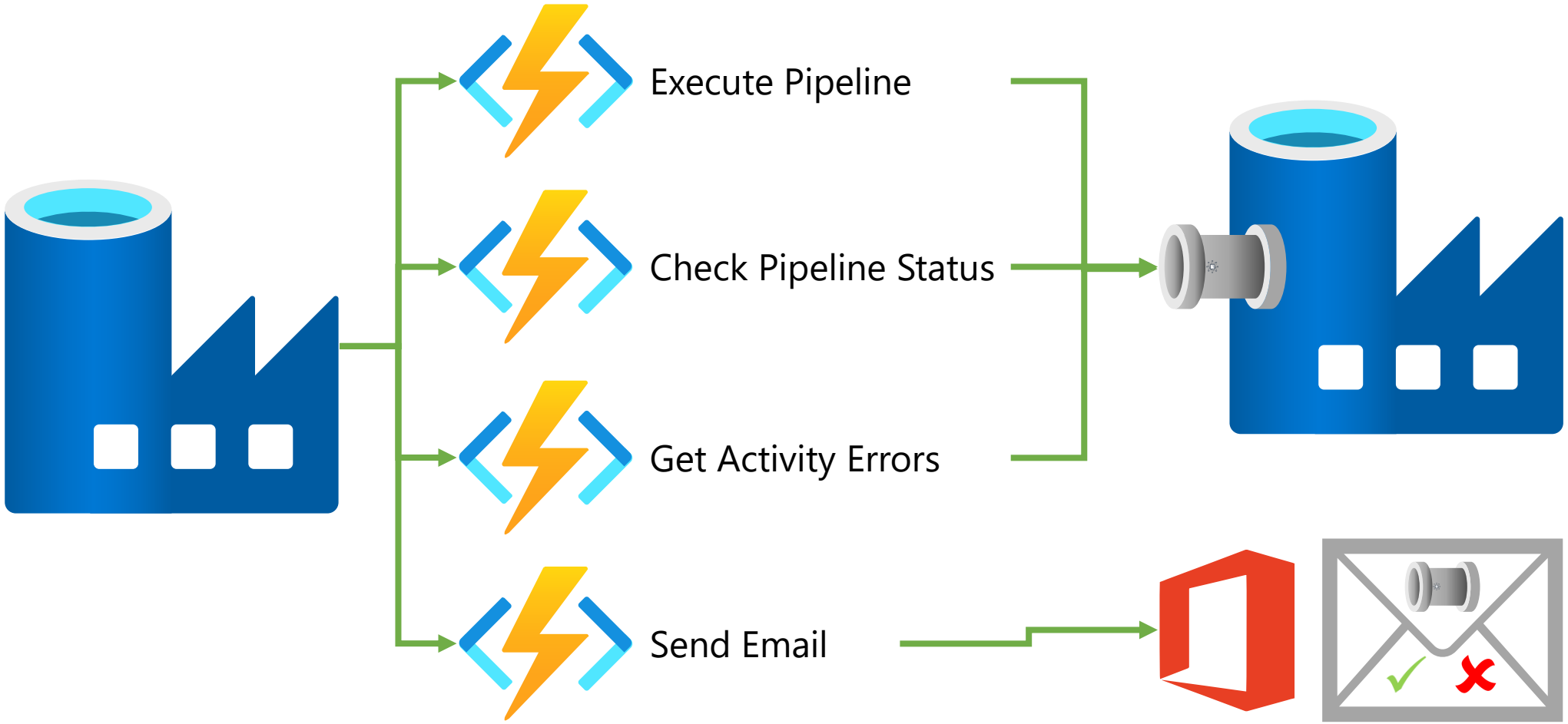


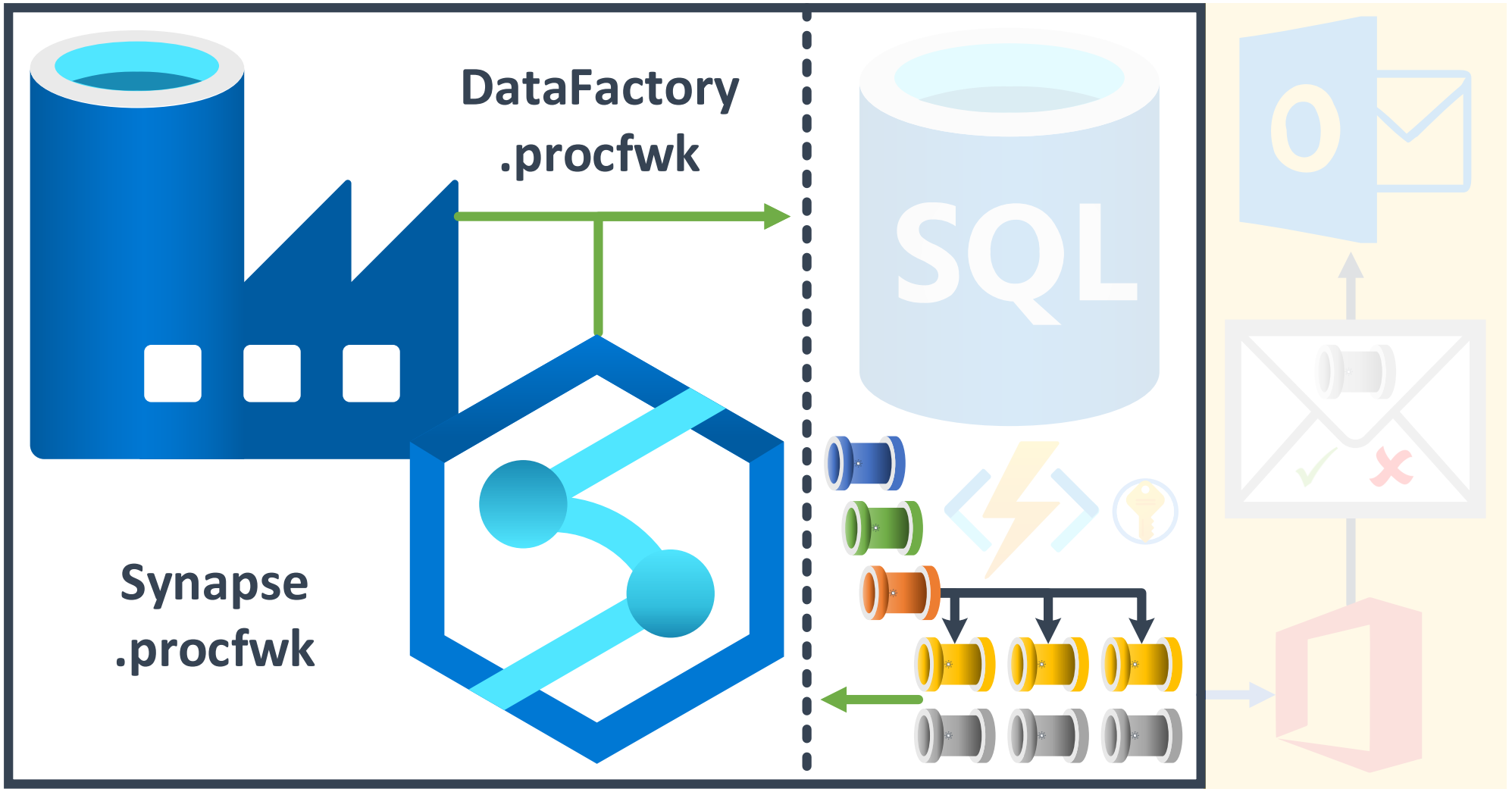

DataFactory
.procfwk
SQL
Synapse
.procfwk

Functions Creation & Configuration

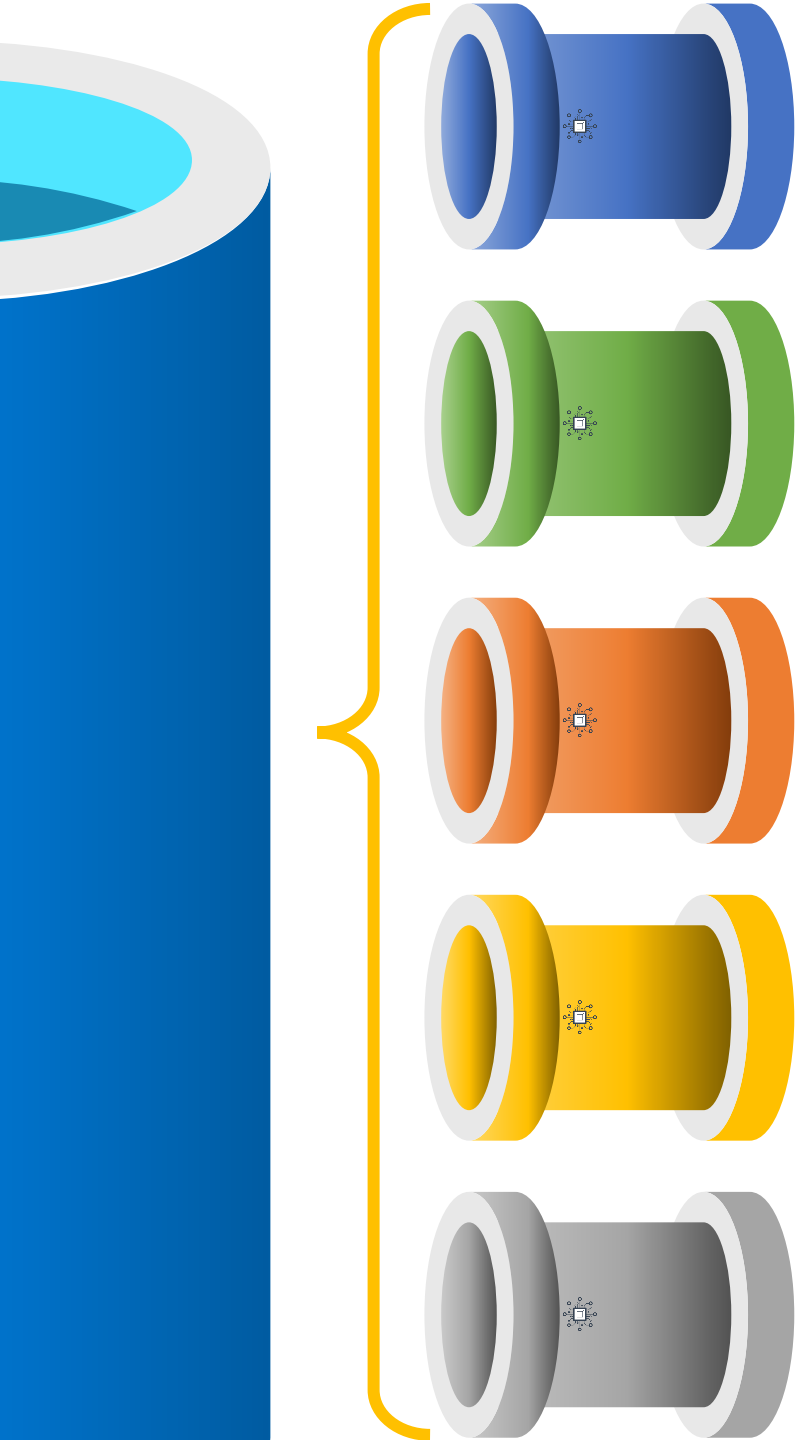


procfwk Functions




DataFactory
.procfwk
Synapse
.procfwk
SQL

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 and execution stage iterator.

- Child

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

- Infant

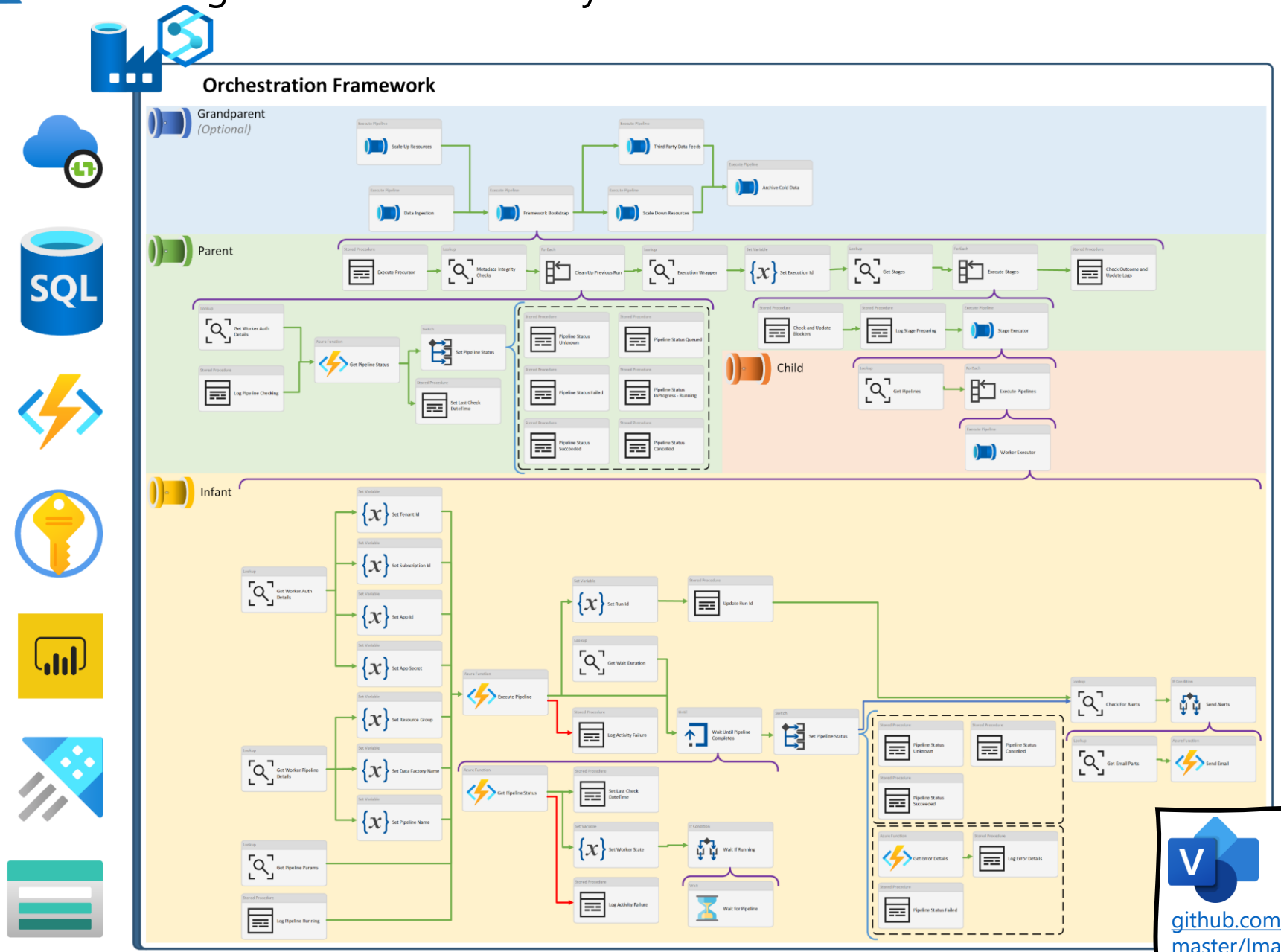
Role: Worker executor, monitor and reporting of the outcome for the single worker pipeline.

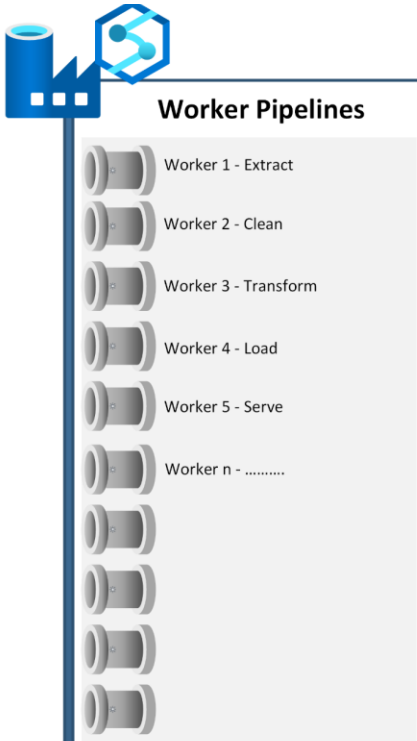
- Worker

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

Processing Framework - Activity Chain

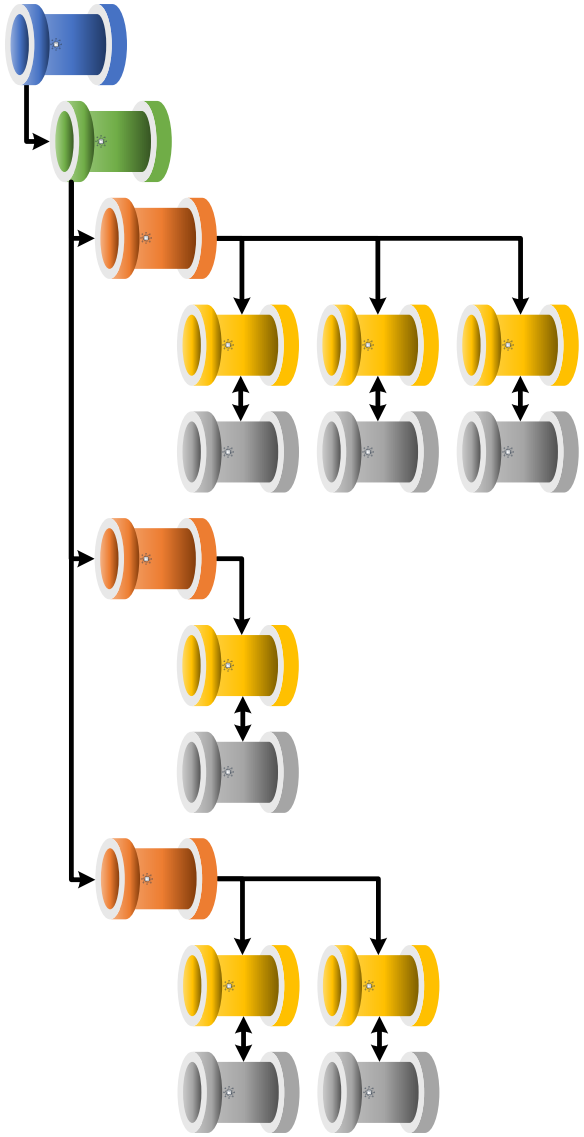
Orchestration Framework Support Resources





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


DEMO



Thank you for listening…

Paul Andrew

Microsoft MVP Most Valuable Professional

altius

Blog: mrpaulandrew.com
Email: paul@mrpaulandrew.com

Twitter: @mrpaulandrew
LinkedIn: In/mrpaulandrew

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