

Azure Data Factory v2

SSIS Data Flows & Custom Extensibility

Paul Andrew | Senior Consultant

14/07/2018

 @MrPaulAndrew



Gold Data Analytics
Gold Data Platform
Gold Cloud Platform



Gold Sponsors

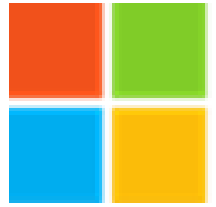


ROBERT WALTERS



@SQLSatMcr

Silver Sponsors



Microsoft



dbWatch
DATABASE CONTROL



@SQLSatMcr

Bronze Sponsors



@SQLSatMcr



<https://github.com/mrpaulandrew>

CommunityEvents

Demo code, content and slides from various community events.

● C++

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

Agenda

Data Factory
Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

Data Factory
Extensibility

SSIS, Functions,
Custom Activities

Conclusions

Design Patterns
ETL/ELT in Azure

Agenda

Data Factory
Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

Data Factory
Extensibility

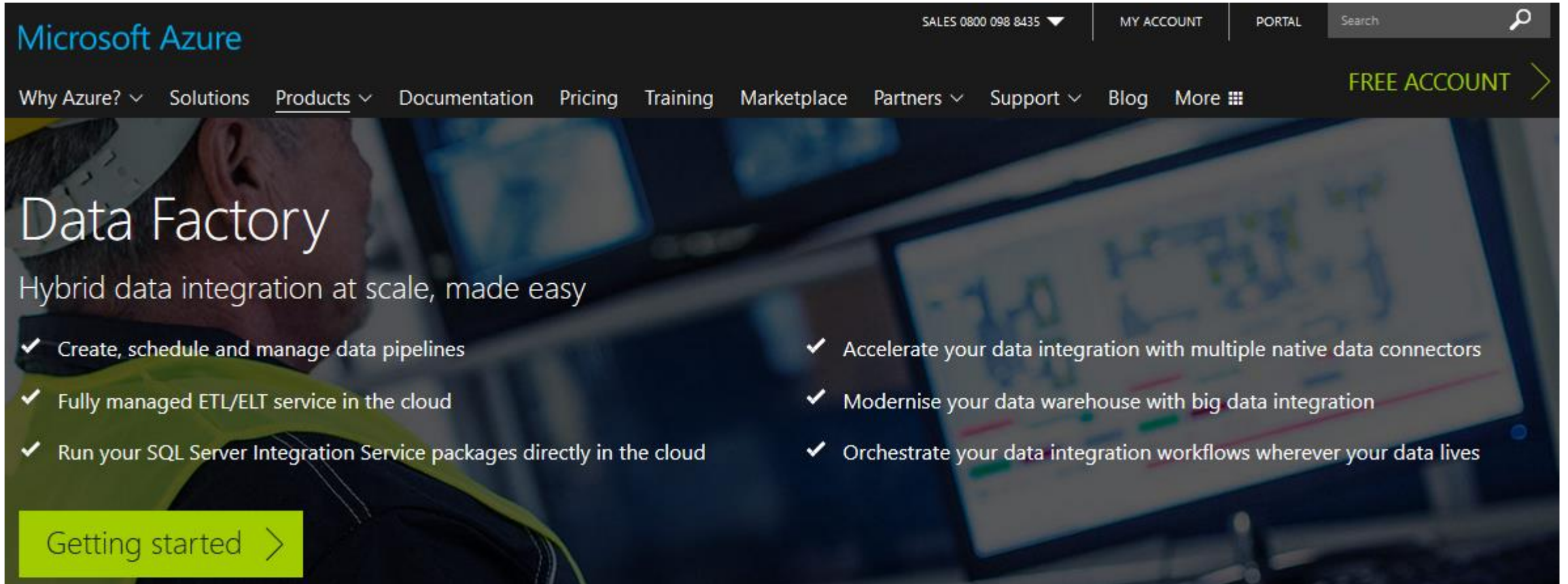
SSIS, Functions,
Custom Activities

Conclusions

Design Patterns
ETL/ELT in Azure

What is Azure Data Factory?

<https://azure.microsoft.com/en-gb/services/data-factory/>



Microsoft Azure

SALES 0800 098 8435 ▼ | MY ACCOUNT | PORTAL | Search

Why Azure? ▾ Solutions Products ▾ Documentation Pricing Training Marketplace Partners ▾ Support ▾ Blog More ☰

FREE ACCOUNT >

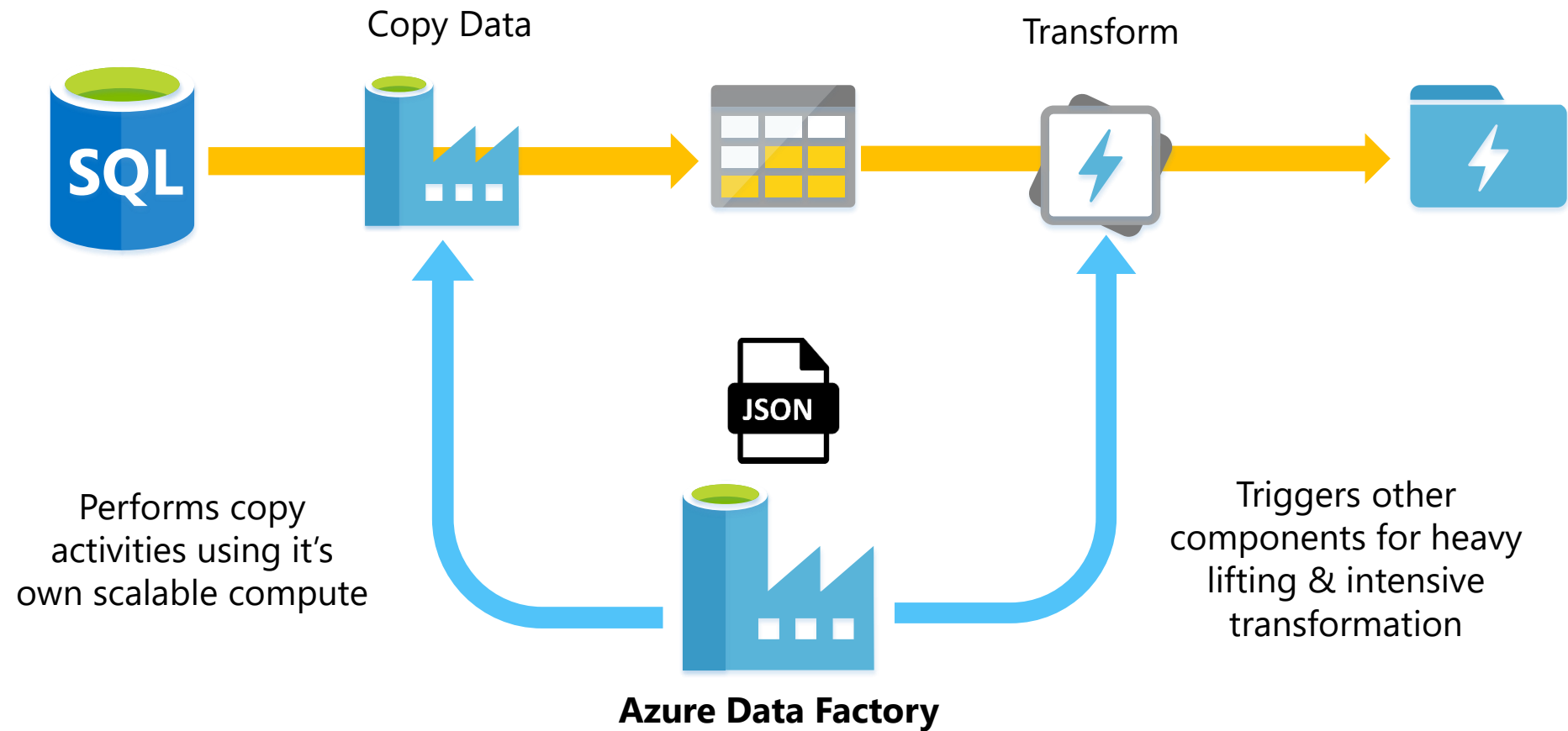
Data Factory

Hybrid data integration at scale, made easy

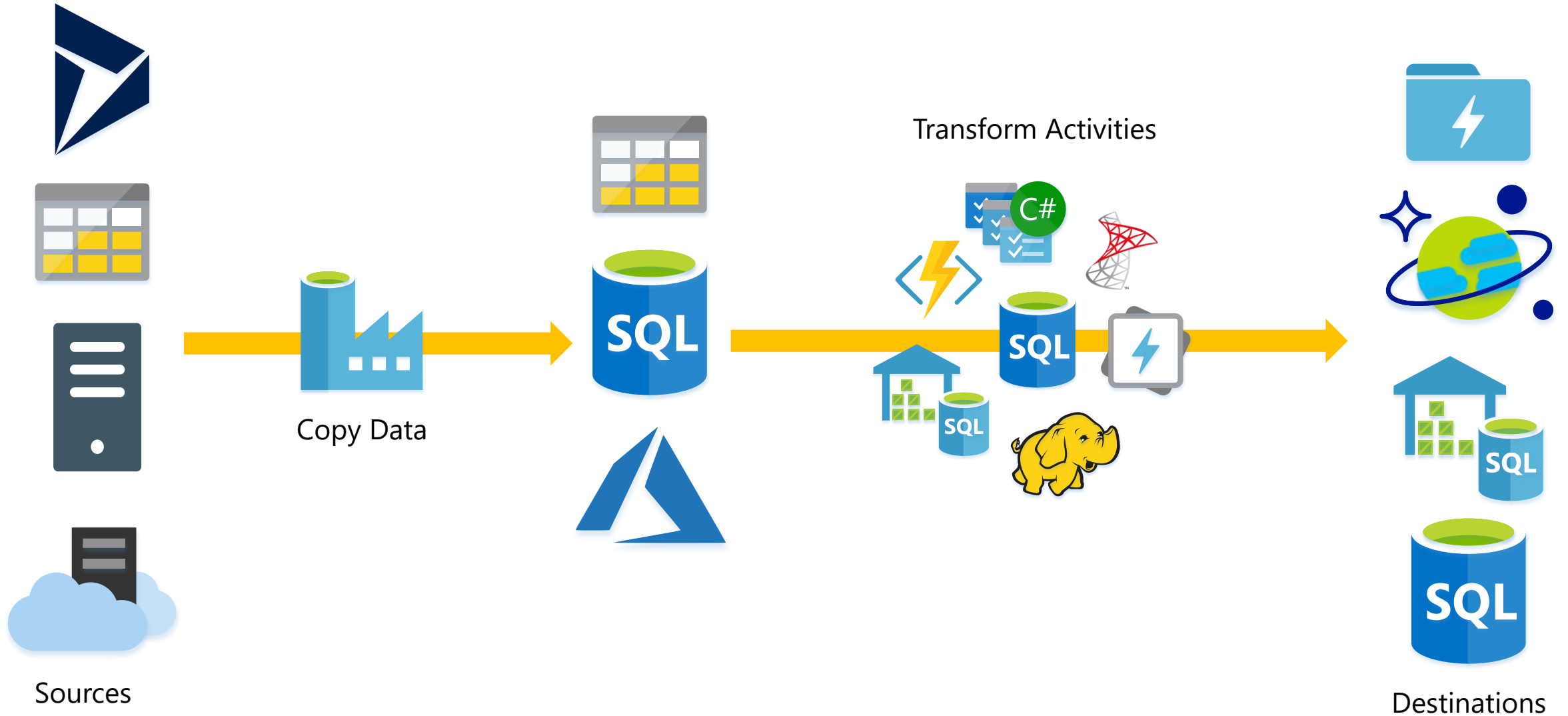
- ✓ Create, schedule and manage data pipelines
- ✓ Fully managed ETL/ELT service in the cloud
- ✓ Run your SQL Server Integration Service packages directly in the cloud
- ✓ Accelerate your data integration with multiple native data connectors
- ✓ Modernise your data warehouse with big data integration
- ✓ Orchestrate your data integration workflows wherever your data lives

Getting started >

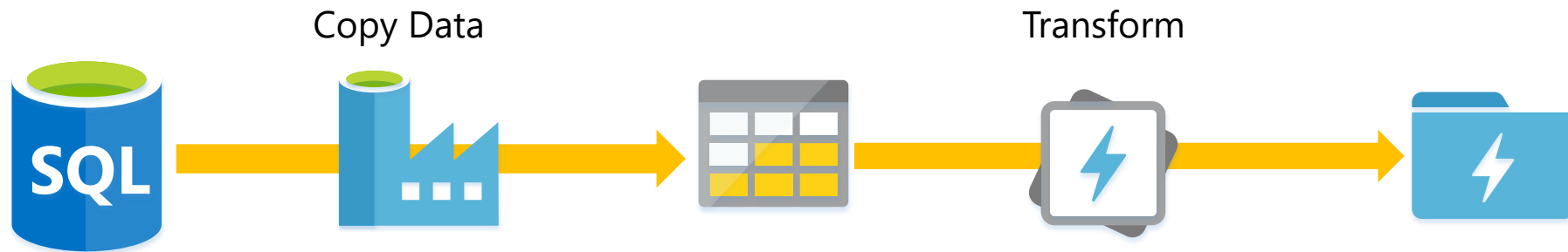
What is Azure Data Factory?



What does Azure Data Factory do?



Data Factory Components

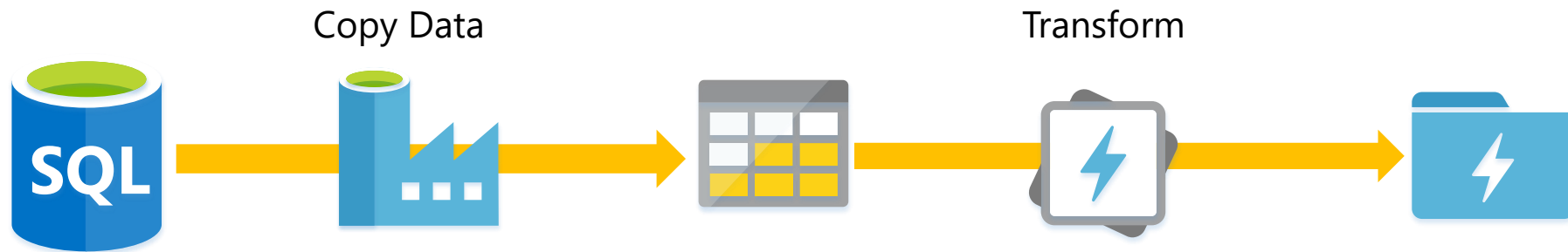


1 Linked Services – How do I connect?

Like the SSIS Connection Manager!



Data Factory Components



1

Linked Services

2

Data Sets – What slices/partitions does my data have?

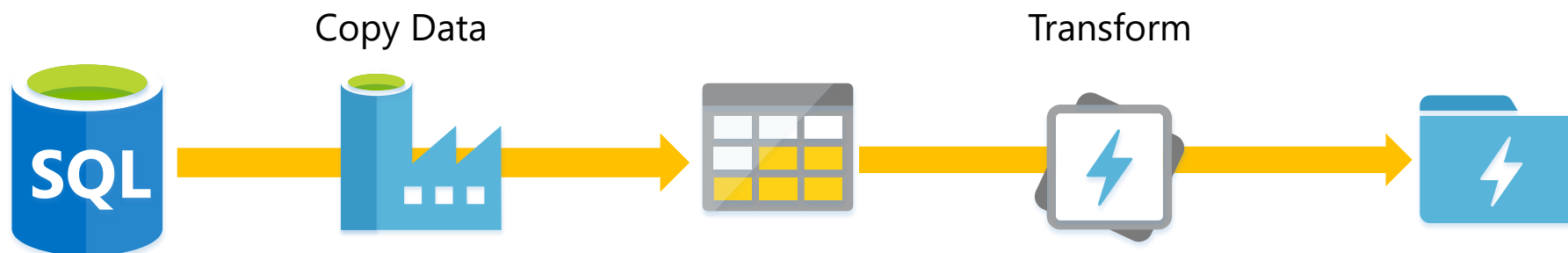


dbo.DimCustomer



/RAW/Orders/2018/01/01/Orders.csv

Data Factory Components



1

Linked Services

2

Data Sets

3

Activities – What do we want to happen?
With what conditions?



U-SQL Activity

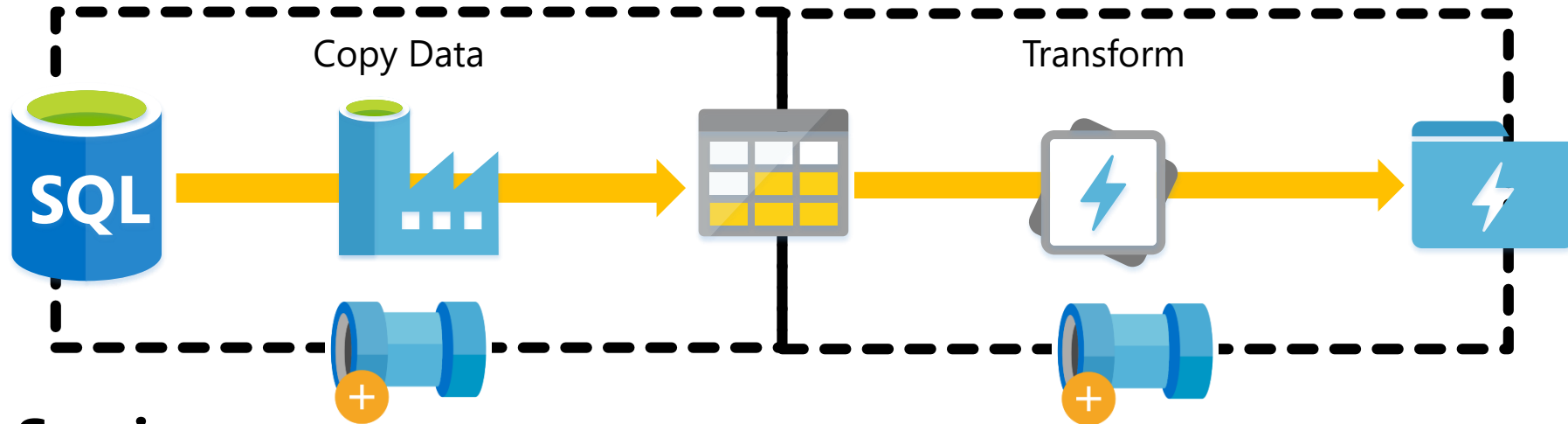
Script: *wasb://:myscripts/ProcessOrders.usql*

AUs: *5 units*

Priority: *1000*

Parameters: *@Output = "RAW/Orders/..."*

Data Factory Components



1

Linked Services

2

Data Sets

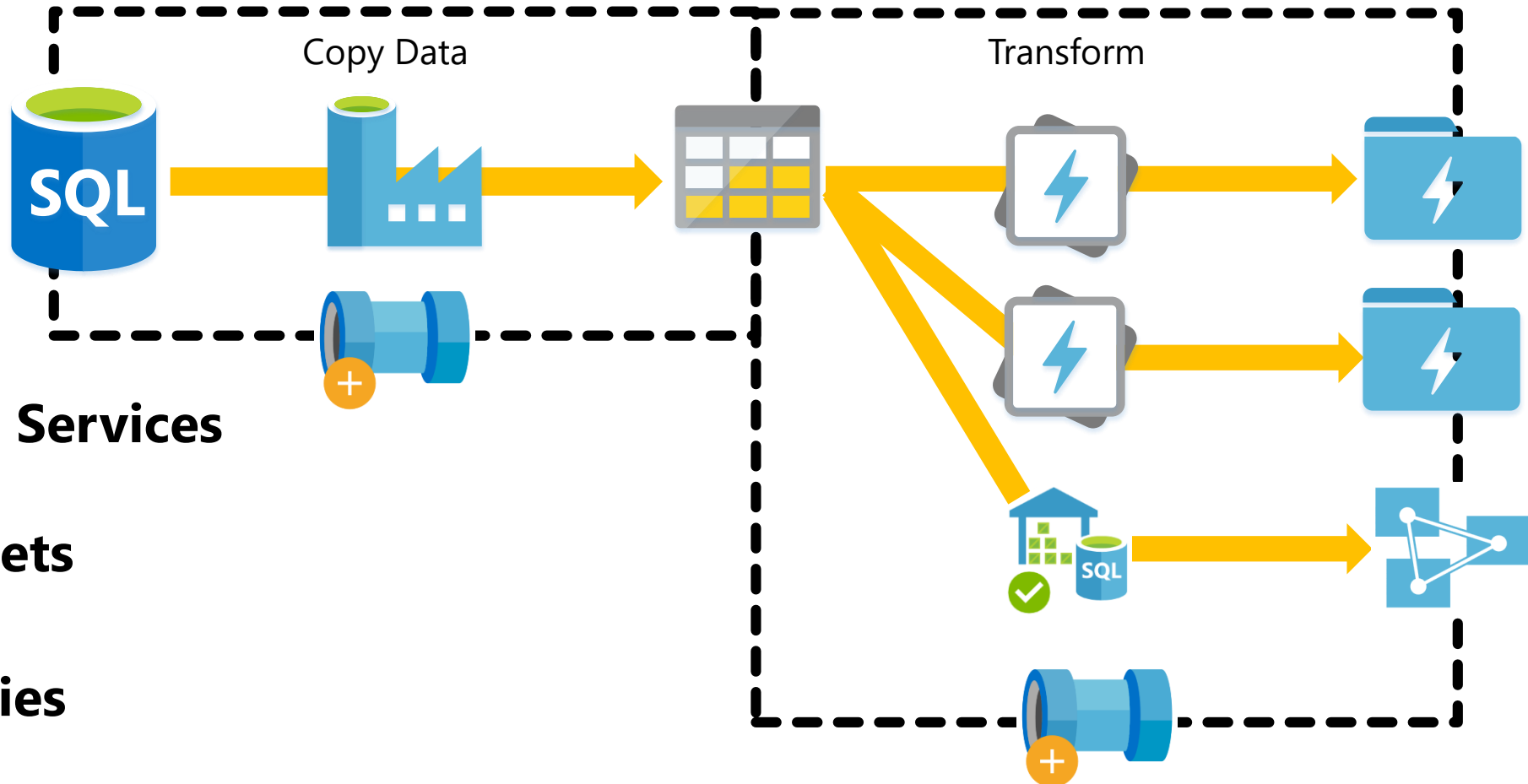
3

Activities

4

Pipelines – What groups of work do I want to do?

Data Factory Components



1

Linked Services

2

Data Sets

3

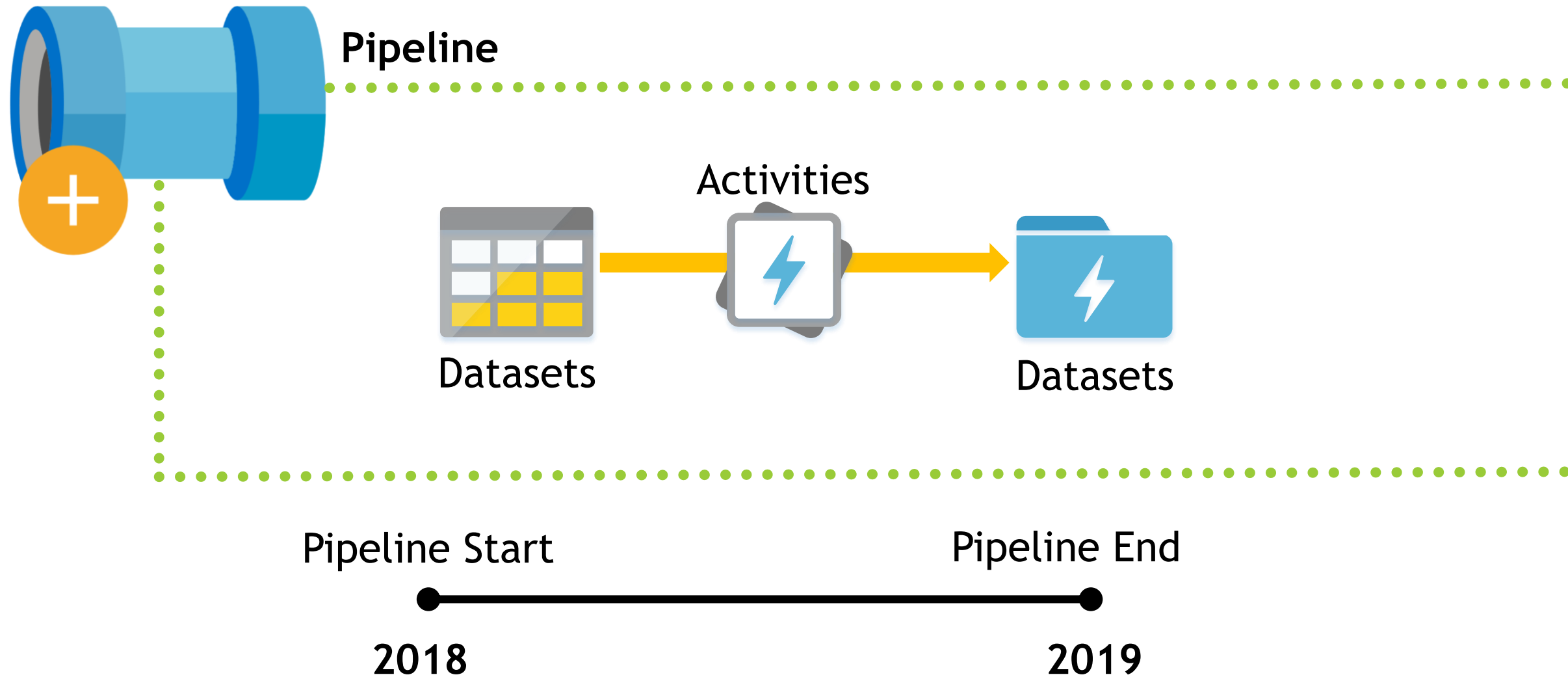
Activities

4

Pipelines – What groups of work do I want to do?

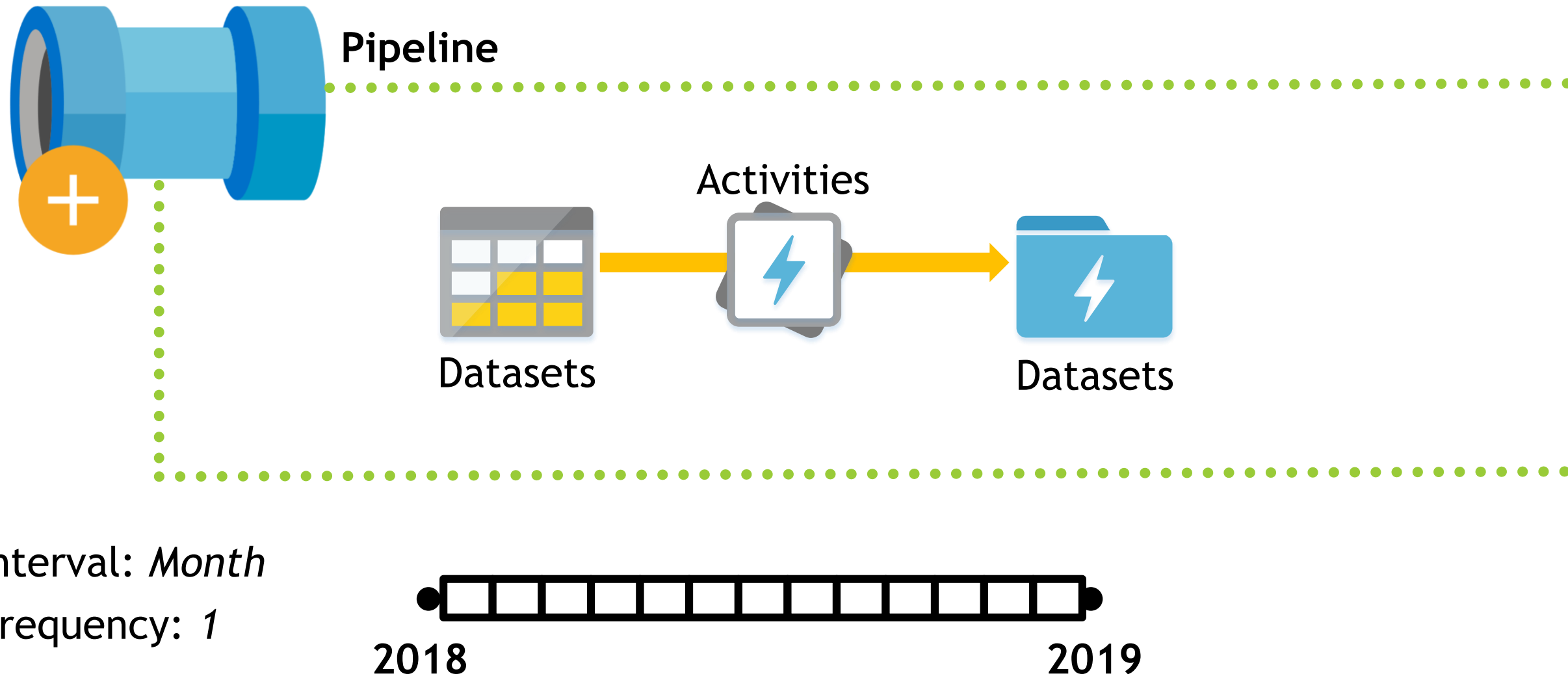
Azure Data Factory Concepts

Time Slices – triggering an activity execution.



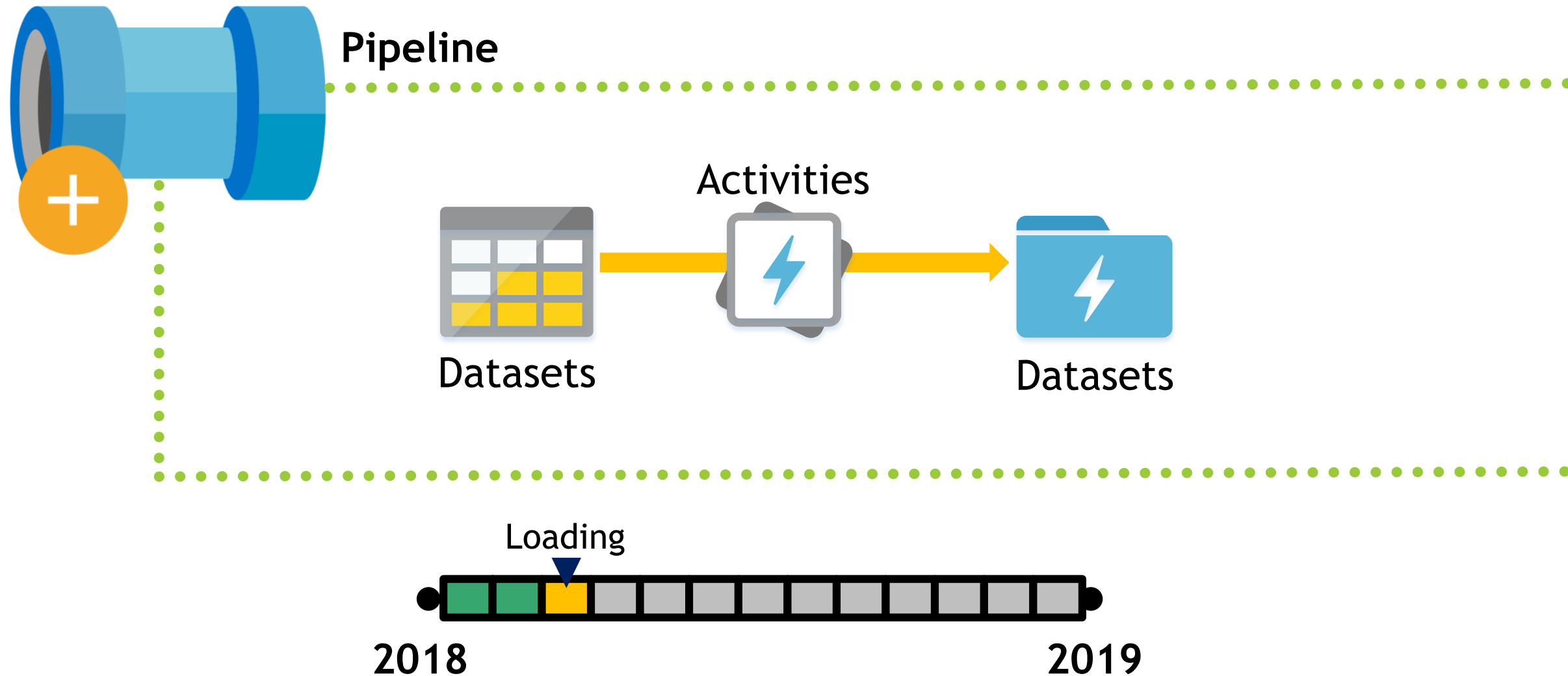
Azure Data Factory Concepts Continued

Time Slices – triggering an activity execution.

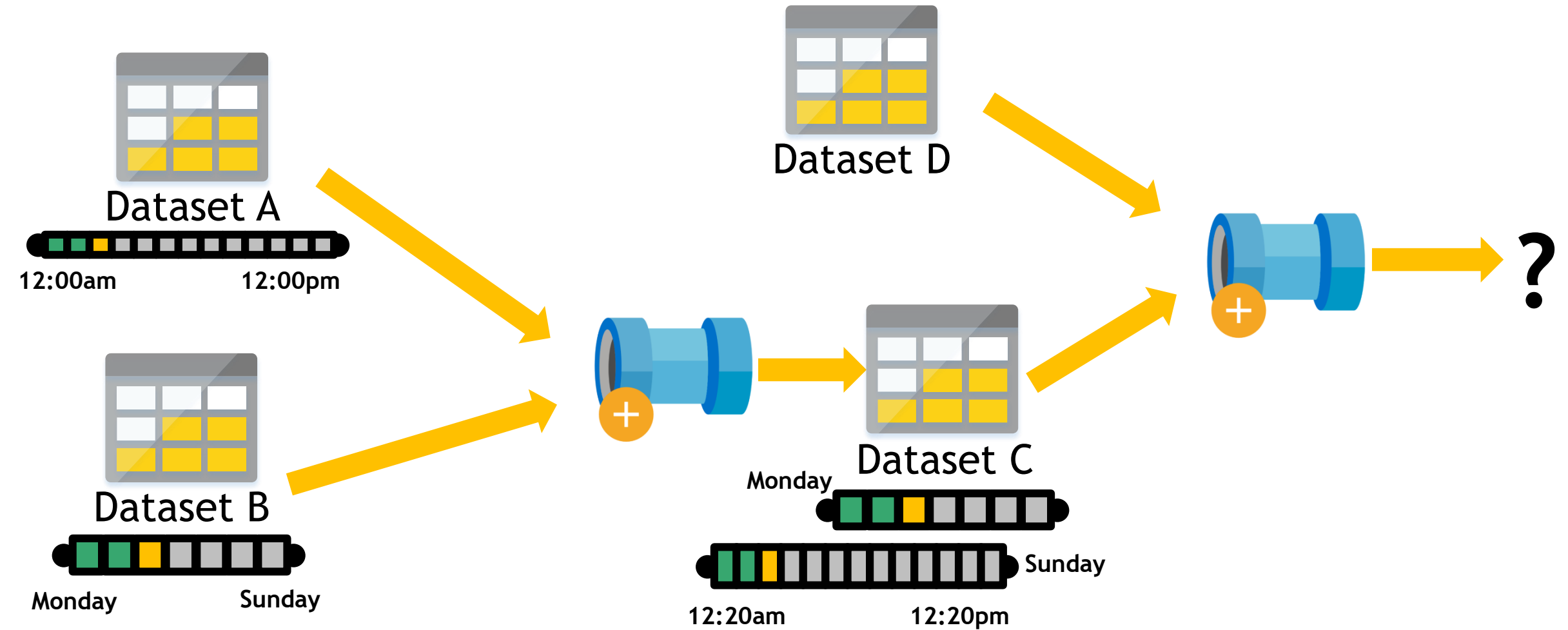


Azure Data Factory Concepts Continued

Time Slices – triggering an activity execution.



Time Slice Problems...



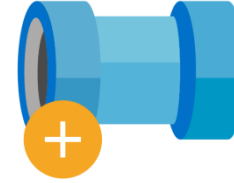
Integration Runtimes



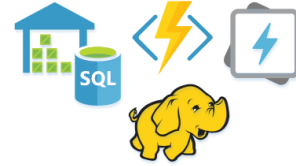
1

Azure
Integration Runtime

Movement Hours



Activity
Orchestration



Flexible Region



2

SSIS
Integration Runtime

SSIS Package
Execution



Specified Region



3

Self Hosted
Integration Runtime

Local Compute



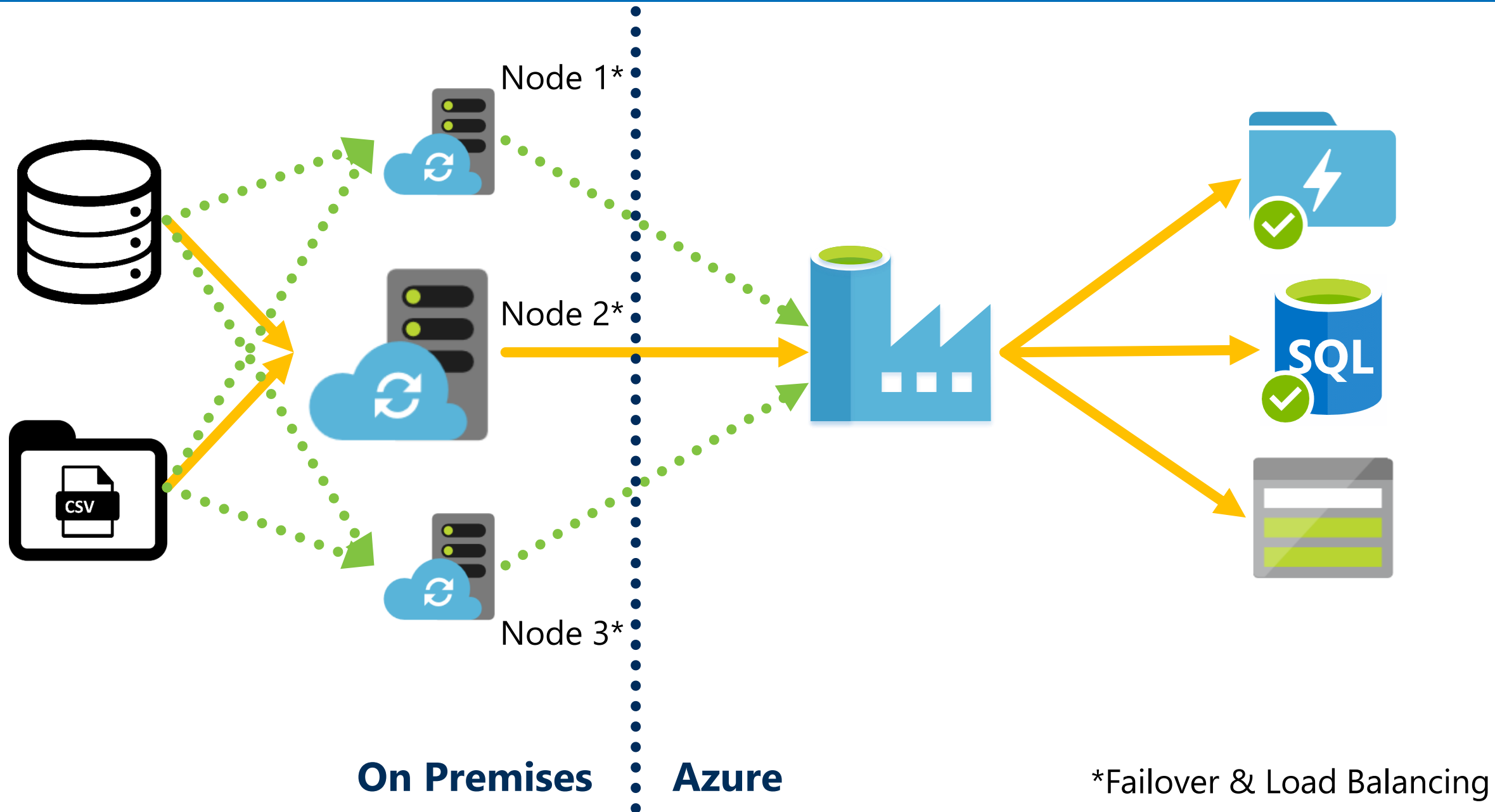
Activity
Orchestration



On-Prem Server



The Integration Runtime (AKA The Data Management Gateway)



Azure Data Factory Concepts & Components Recap

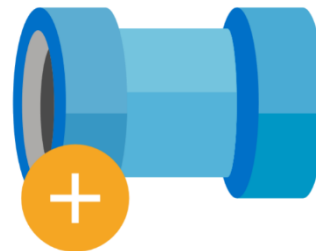


1 **Linked Services**

2 **Data Sets**

3 **Activities**

4 **Pipelines**



Time Slices



1 **Azure**
Integration Runtime

2 **SSIS**
Integration Runtime

3 **Self Hosted**
Integration Runtime

Agenda

Data Factory
Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

Data Factory
Extensibility

SSIS, Functions,
Custom Activities

Conclusions

Design Patterns
ETL/ELT in Azure

Data Factory Issues & Limitations

- ▶ **Time Slices** – Complex, difficult to change & provision
- ▶ **Pricing** – High and low frequency
- ▶ **Control Flow** – Not conditional. Pass or fail
- ▶ **Developer Tools** – VS or Portal JSON templates
- ▶ **Hard Coded Pipelines** – No dynamic values
- ▶ **C# Coding** – Often required for anything complex
- ▶ **Monitoring** – Times slice bound, focused on datasets
- ▶ **Connectivity** – Limited to Microsoft supported linked services



So What's Changed?



Data Movement (Copy)



Activities



Pipelines



Datasets



Time Slices/Tumbling Windows



Event Triggers



Recurring Schedules



Parameters



Expressions



Conditional Logic



Use of SSIS Packages



Graphic Developer Canvas

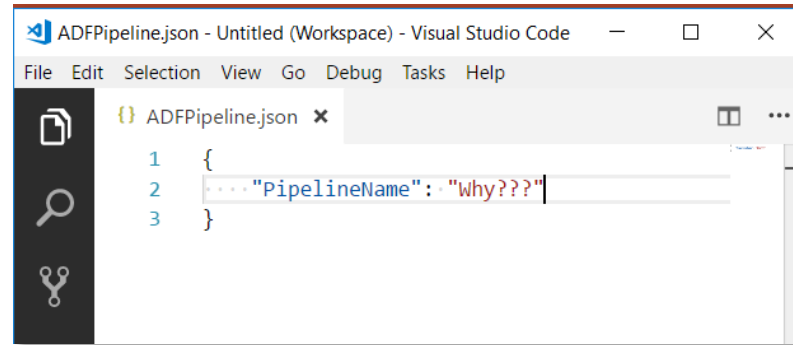


Drilldown Monitoring



Developer Tools

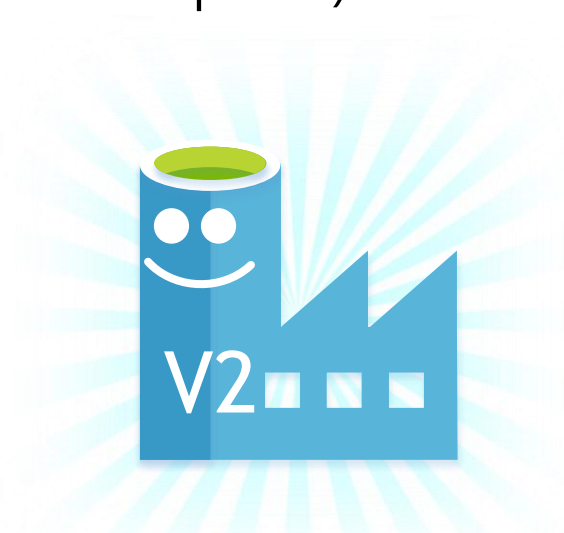
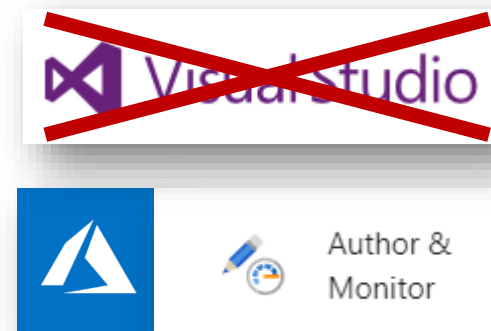
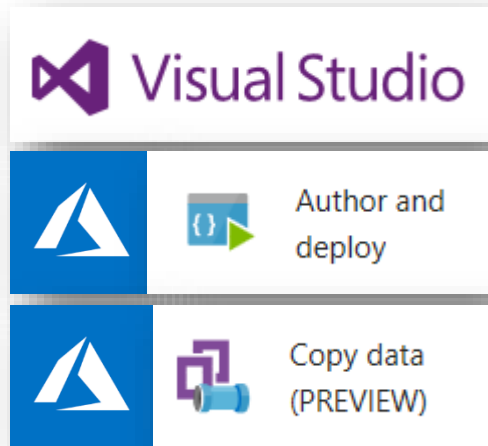
- ▶ JSON Templates
- ▶ Data Factory Wizard
- ▶ Reverse Engineer From Azure
- ▶ Deployment Wizard



```
ADFPipeline.json - Untitled (Workspace) - Visual Studio Code
File Edit Selection View Go Debug Tasks Help

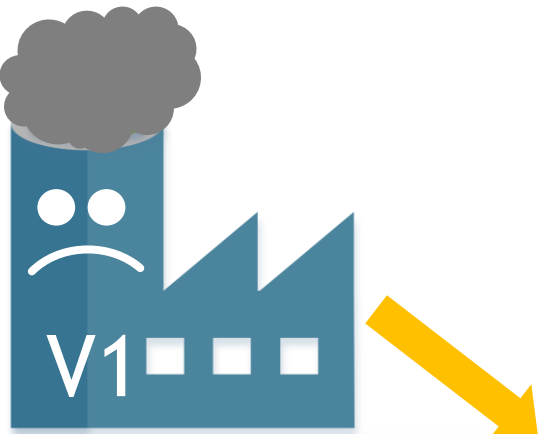
ADFPipeline.json x
1 {
2   ... "PipelineName": "Why???"
3 }
```

- ▶ No JSON Templates
- ▶ No Deployment Wizard
- ▶ Data Factory Wizard
- ▶ Reverse Engineer From Azure (via ARM Templates)



Only Visual Studio 2015
<http://bit.ly/2tsyD90>

Monitoring



Data factory

RESOURCE EXPLORER

- Data Factories
 - PaulsFunFactoryV1
 - Pipelines
 - FileCleaning
 - UploadFileToADLStore
 - Datasets
 - FakeOrdersClean
 - FakeOrdersLanding
 - FakeOrdersSourceFile
 - Linked services
 - BatchCompute
 - BlobStore
 - DataLakeStore
 - LaptopGateway
 - USQLEngine
 - Gateways
 - PaulsLappy
 - jhlhjkj

PaulsFunFactoryV1

Start time (UTC): 01/08/2018 01:32 pm End time (UTC): 01/16/2018 01:35 PM

FakeOrdersSourceFile (FILESHARE) → UploadFileToADLStore PIPELINE (1 activities) → FakeOrdersLanding (FREQUENCY: MONTH, INTERVAL: 1, AZURE DATA LAKE...) → FileCleaning PIPELINE (1 activities) → FakeOrdersClean (FREQUENCY: MONTH, INTERVAL: 1, AZURE DATA LAKE...)

ACTIVITY WINDOWS

No filter applied.

Pipeline	Activity	Window Star...	Window End	Status	Type	Last Attempt...	Last Attempt...	Duration	Retry At
There are currently no activity windows to display.									

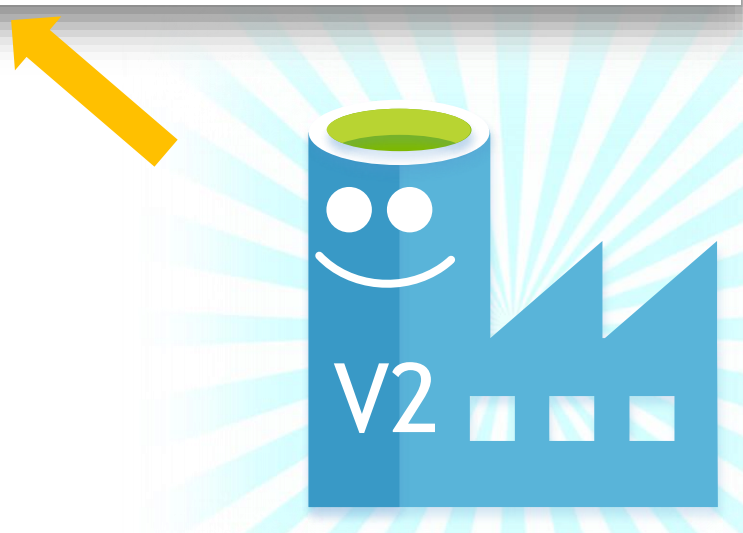
ADFv2DemoFactory01 | Monitor Pipeline Runs

Refresh

Last 24 Hours 01/14/2018 1:35 PM - 01/15/2018 1:35 PM Time Zone (UTC+00:00) London

All Succeeded In Progress Failed Cancelled

Pipeline Name	Actions	Run Start	Duration	Triggered By	Status	Parameters	Error
RunSSISPackage		01/15/2018, 1:36:42 PM	00:00:11	Manual trigger	Succeeded		



Agenda

Data Factory
Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

Data Factory
Extensibility

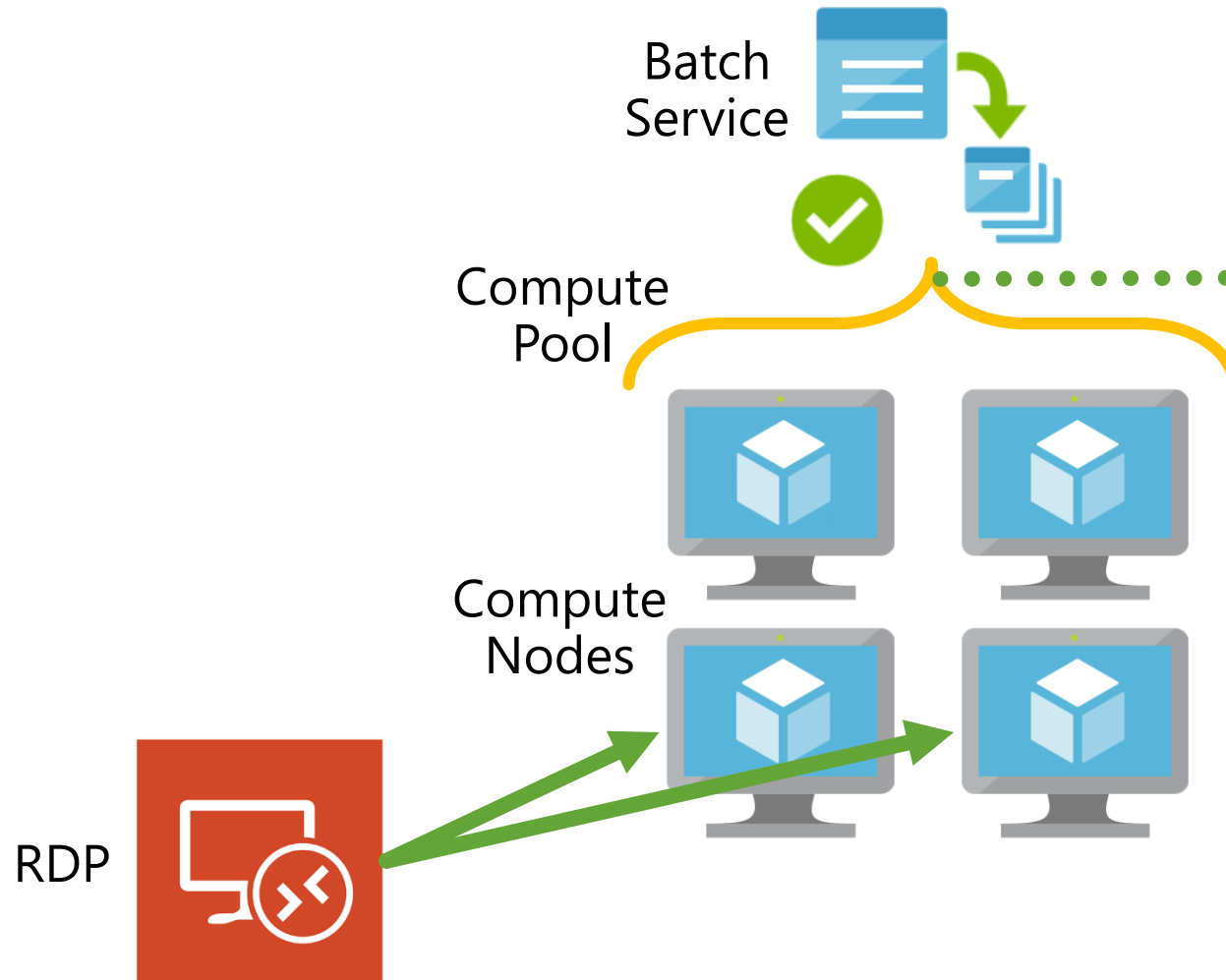
SSIS, Functions,
Custom Activities

Conclusions

Design Patterns
ETL/ELT in Azure

1

Custom Activities – A .Net Console App Executed Using Azure Batch Service



VM node size set per compute pool:

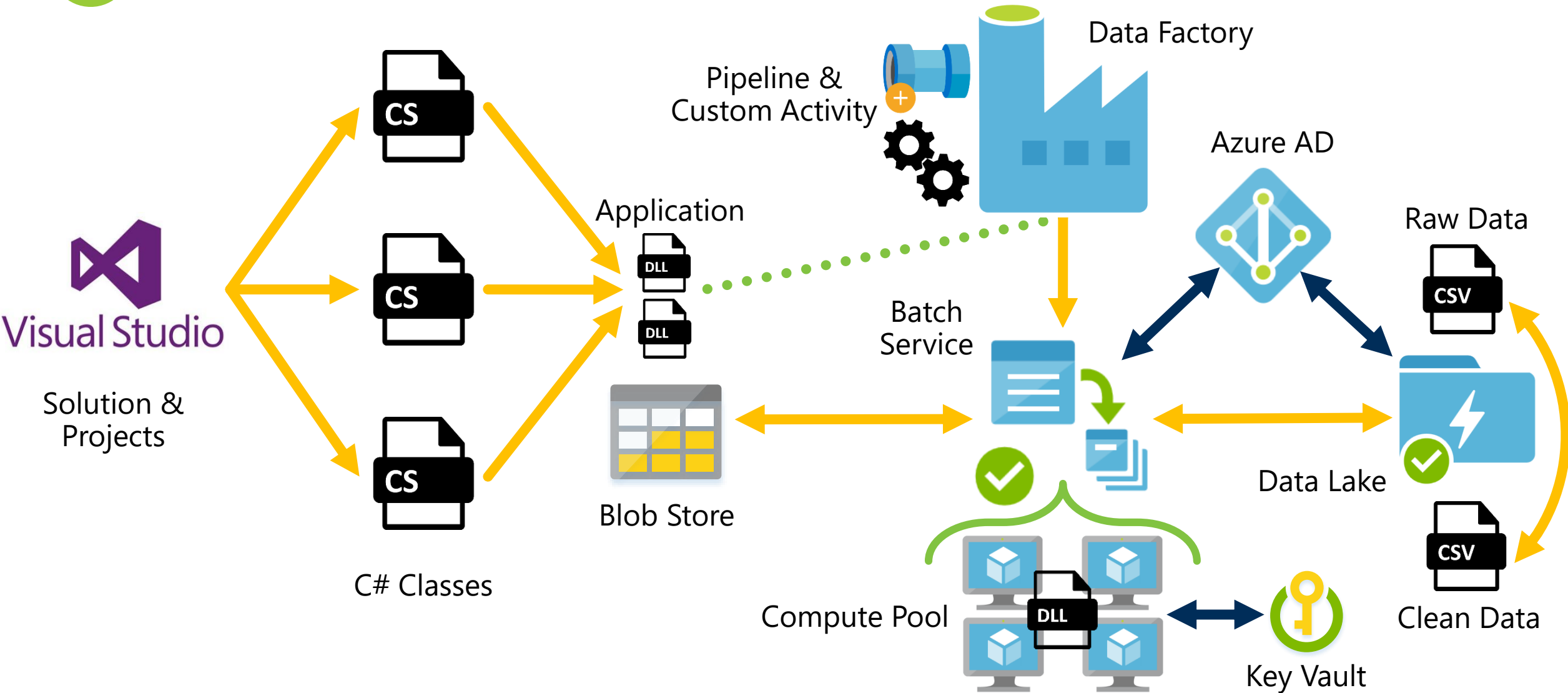
A1 Standard ★	A2 Standard ★	A3 Standard ★
1 Cores	2 Cores	4 Cores
1.8 GB	3.5 GB	7 GB
1 TB OS disk size	1 TB OS disk size	1 TB OS disk size
70 GB Resource disk size	135 GB Resource disk size	285 GB Resource disk size
2 Max data disk	4 Max data disk	8 Max data disk
Unable to display pricing	Unable to display pricing	Unable to display pricing

- ▶ 1 compute node = 1 virtual machine.
- ▶ 1 job per compute node.
- ▶ Max of 4 tasks per node.
- ▶ OS on D drive, not C.
- ▶ Special environment variables.

ADF Extensibility Continued

1

Custom Activities – A .Net Console App Executed Using Azure Batch Service

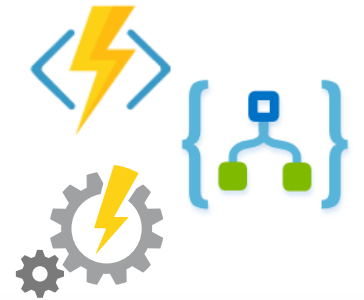


ADF Extensibility Continued

1 **Custom Activities** – A .Net Console App Executed Using Azure Batch Service

2 **Rest API Calls** – Eg. Web Activities Calling:

Azure Functions
Azure Logic Apps
Azure Automation



General Settings² Parameters Advanced

Name * Web1

Description

Timeout 7.00:00:00

Retry 0

Retry interval 20

General Settings² Parameters Advanced

URL *

Method * Select API method...
Select API method...
GET
POST
PUT

Headers

General Settings² Parameters Advanced

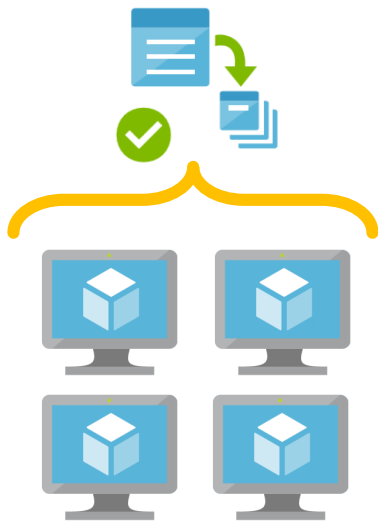
Use [expressions](#), [functions](#) or refer to [system variables](#) in the 'value' column.

Parameterizable properties ⓘ

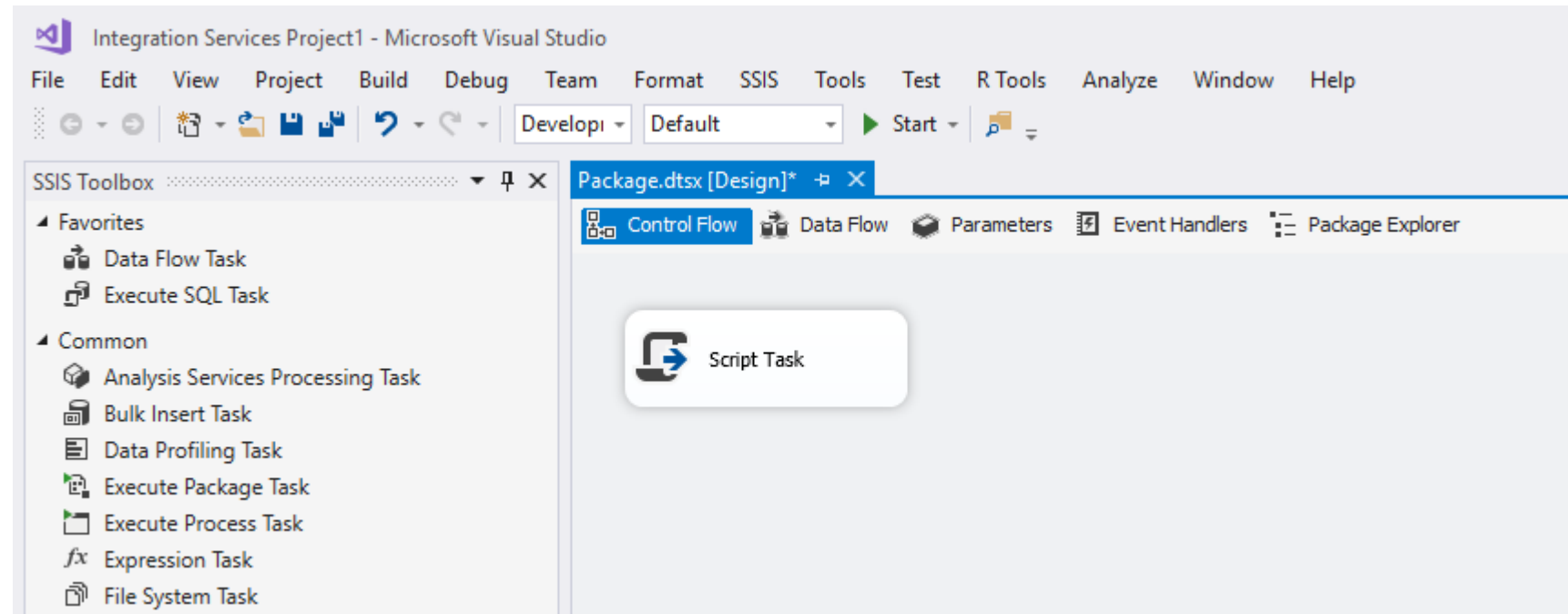
NAME	VALUE
url	<input type="text" value="Value"/>
body	<input type="text" value="Value"/>
Timeout	<input type="text" value="Value"/>
Retry	<input type="text" value="Value"/>

ADF Extensibility Continued

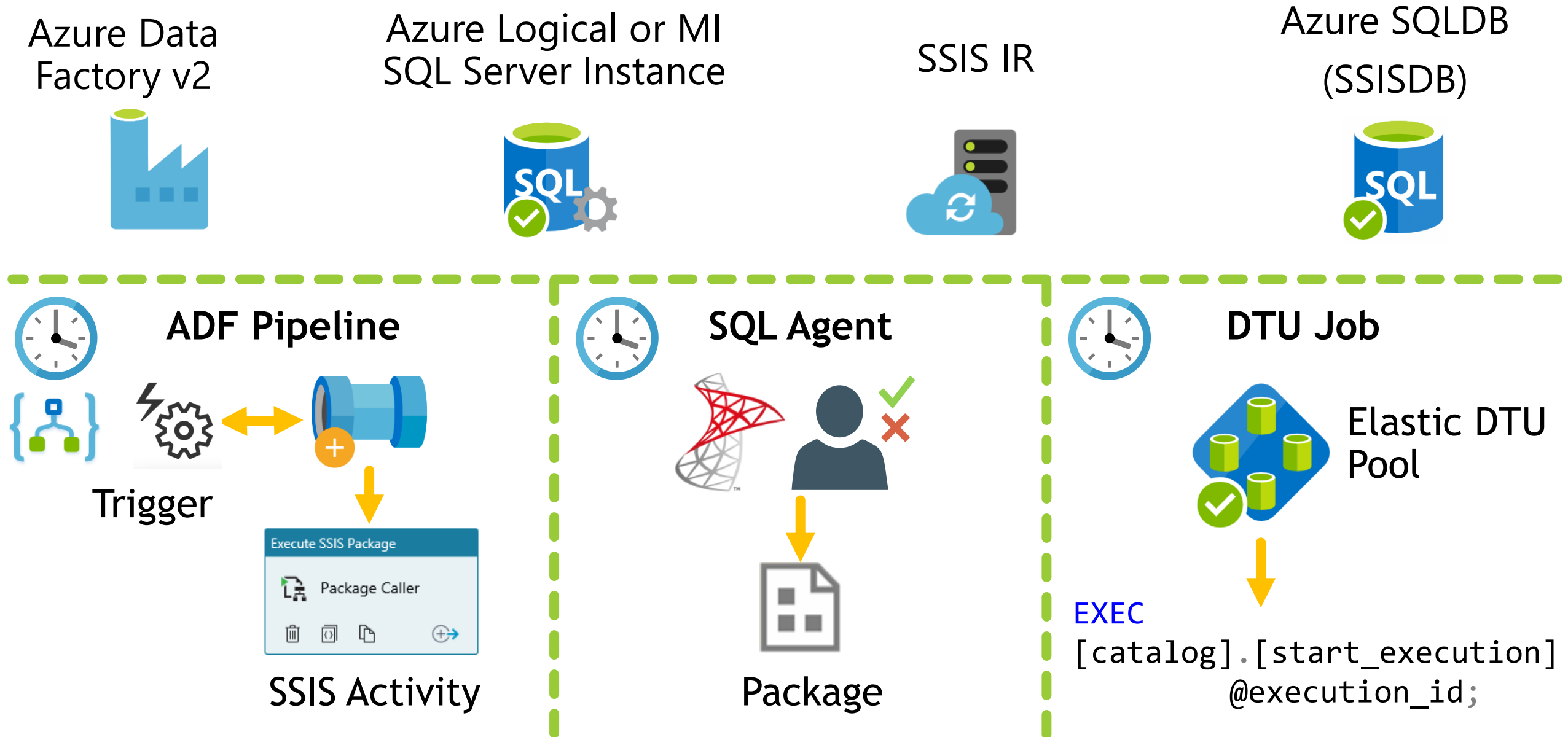
- 1 **Custom Activities**
- 2 **Rest API Calls**
- 3 **SSIS** – Packages with Control Flows and Data Flows



ADF SSIS IR



How do we schedule an SSIS Package in Azure?





Agenda

Data Factory
Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

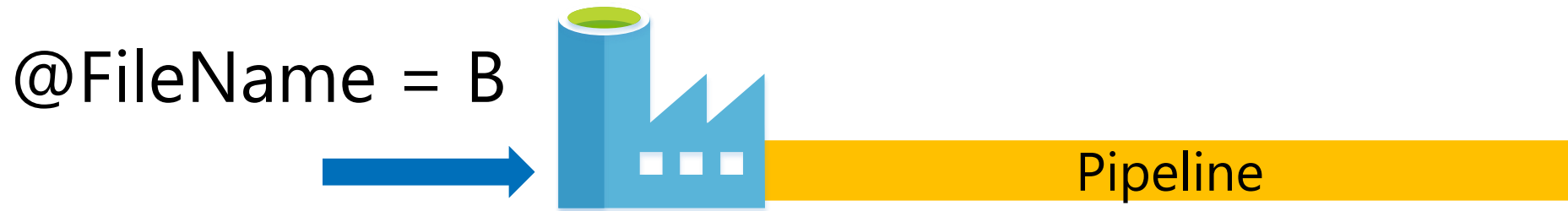
Data Factory
Extensibility

SSIS, Functions,
Custom Activities

Conclusions

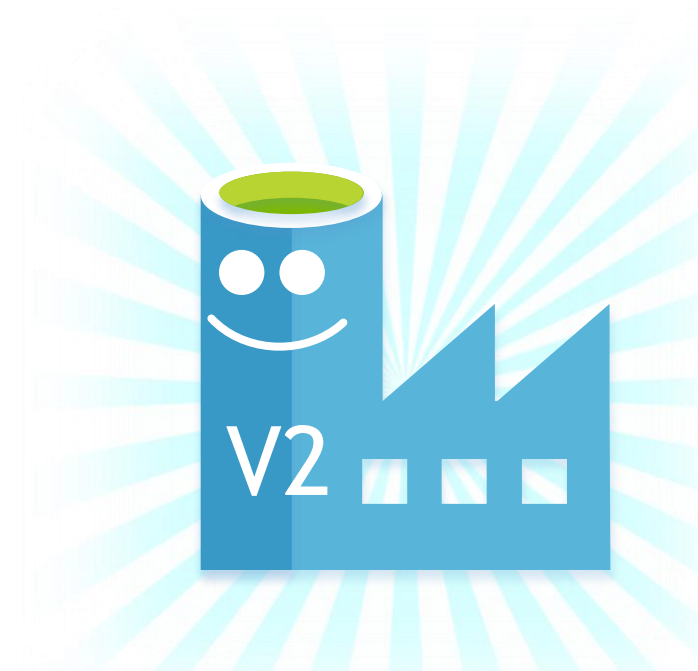
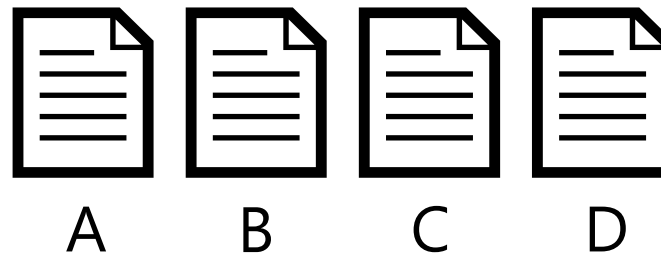
Design Patterns
ETL/ELT in Azure

Dynamic Pipelines using Parameters & Expressions

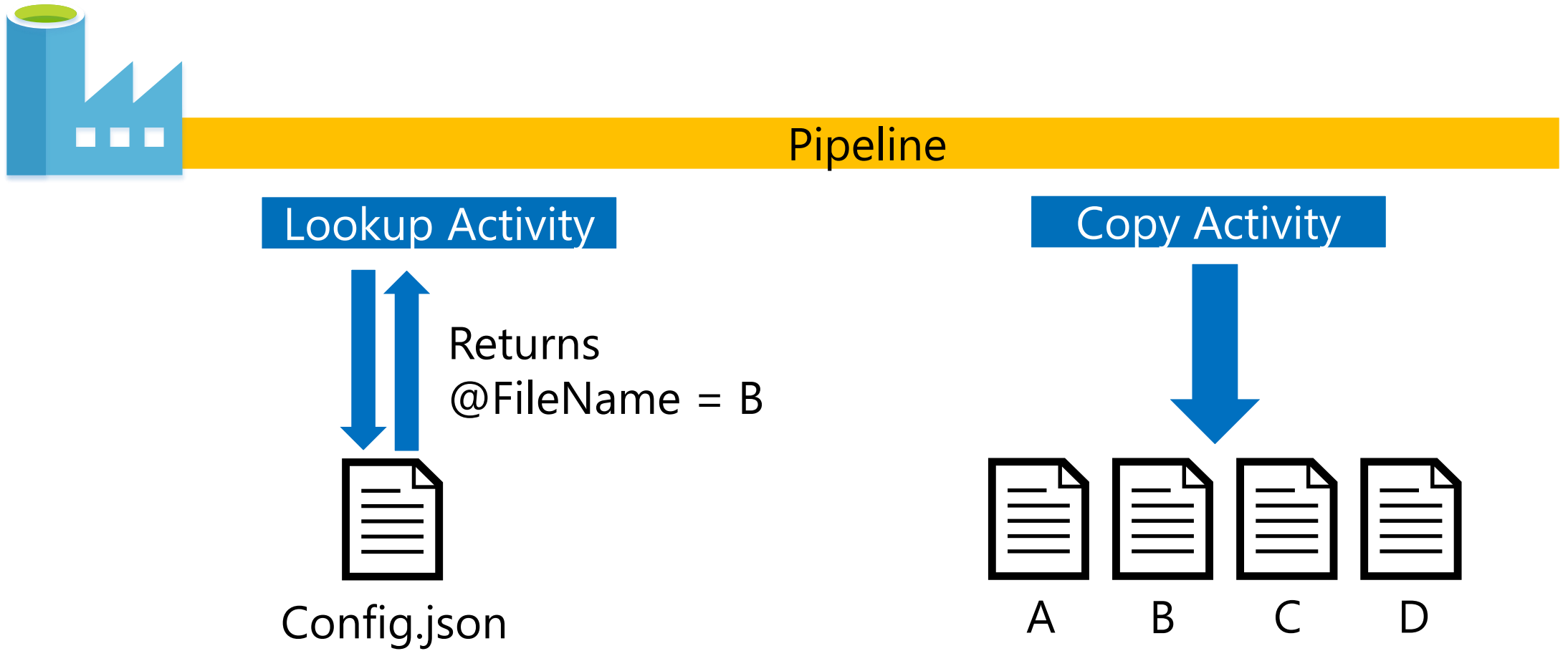


Activity 1

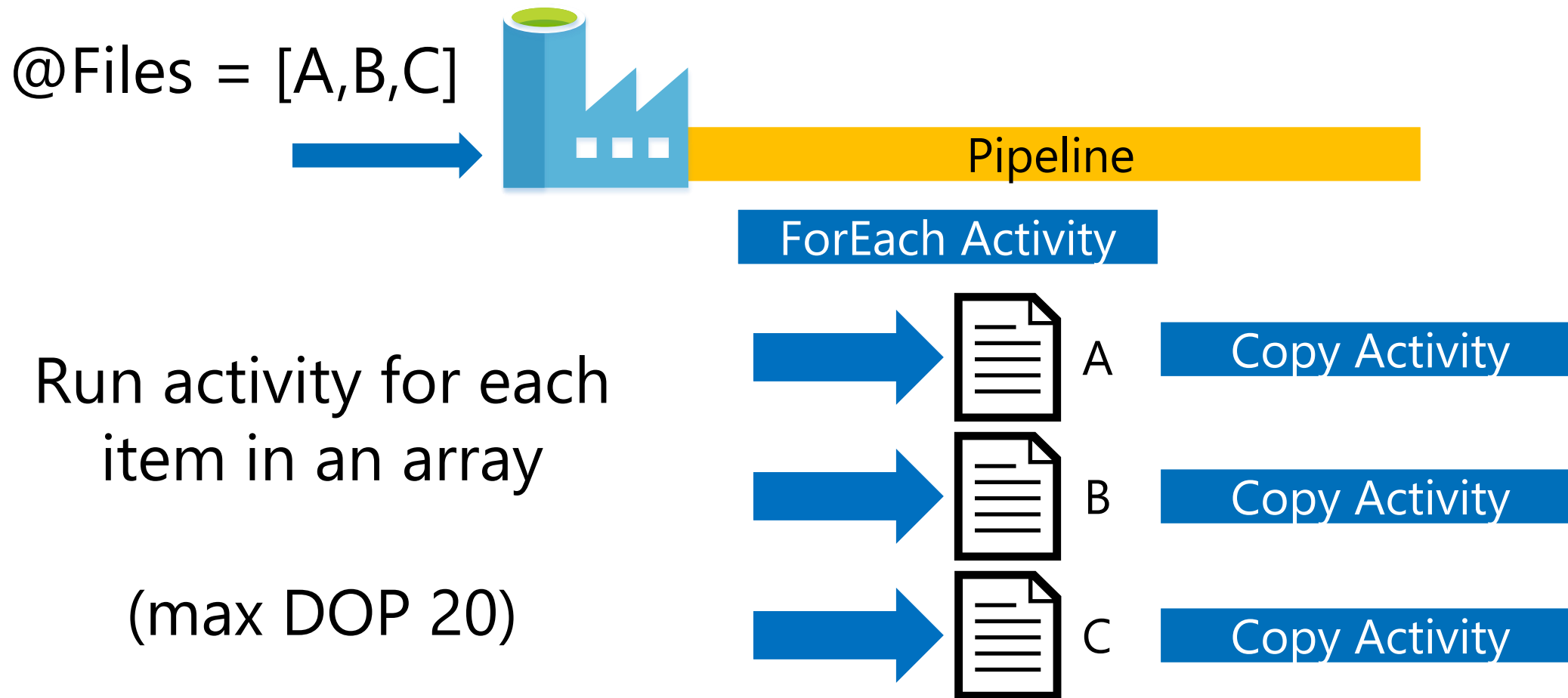
Dynamically change
parameters based
on inputs



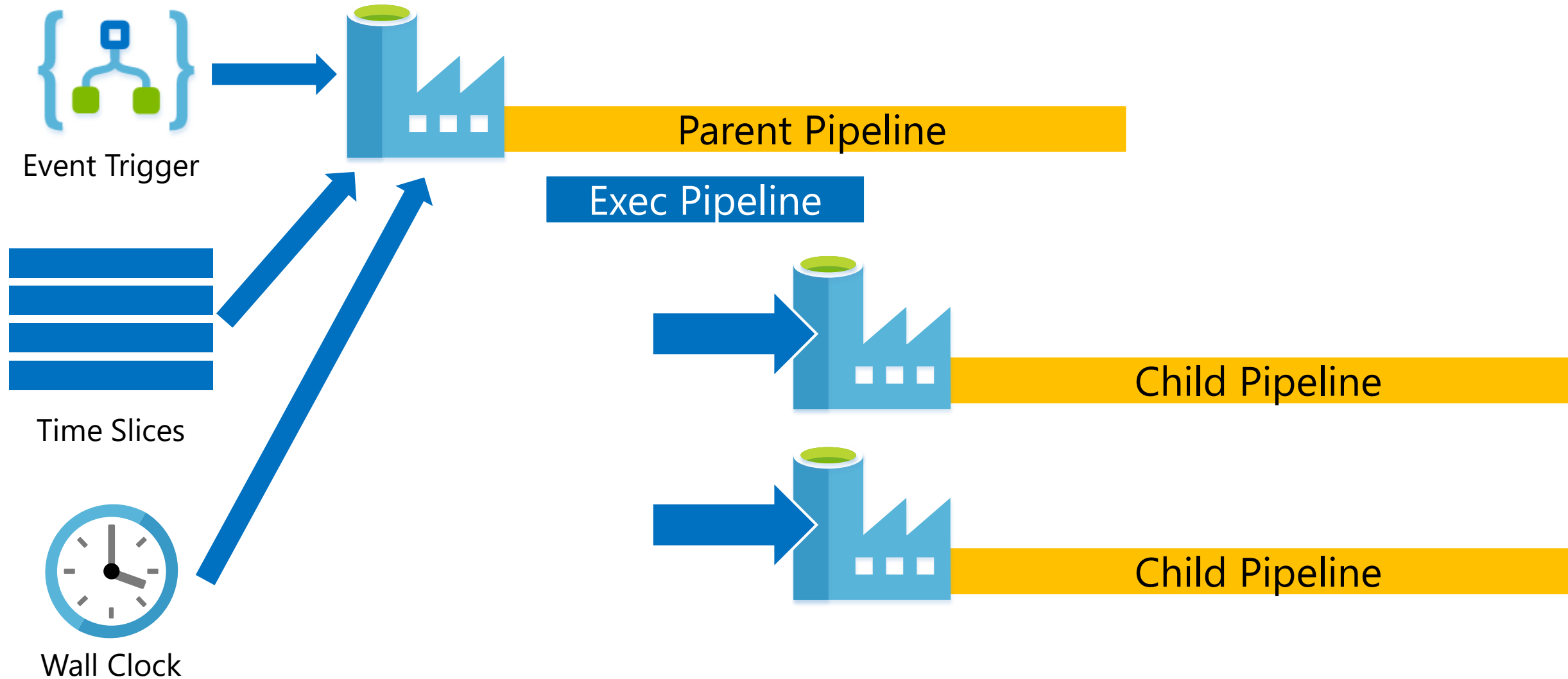
Dynamic Pipelines using Lookup Activity



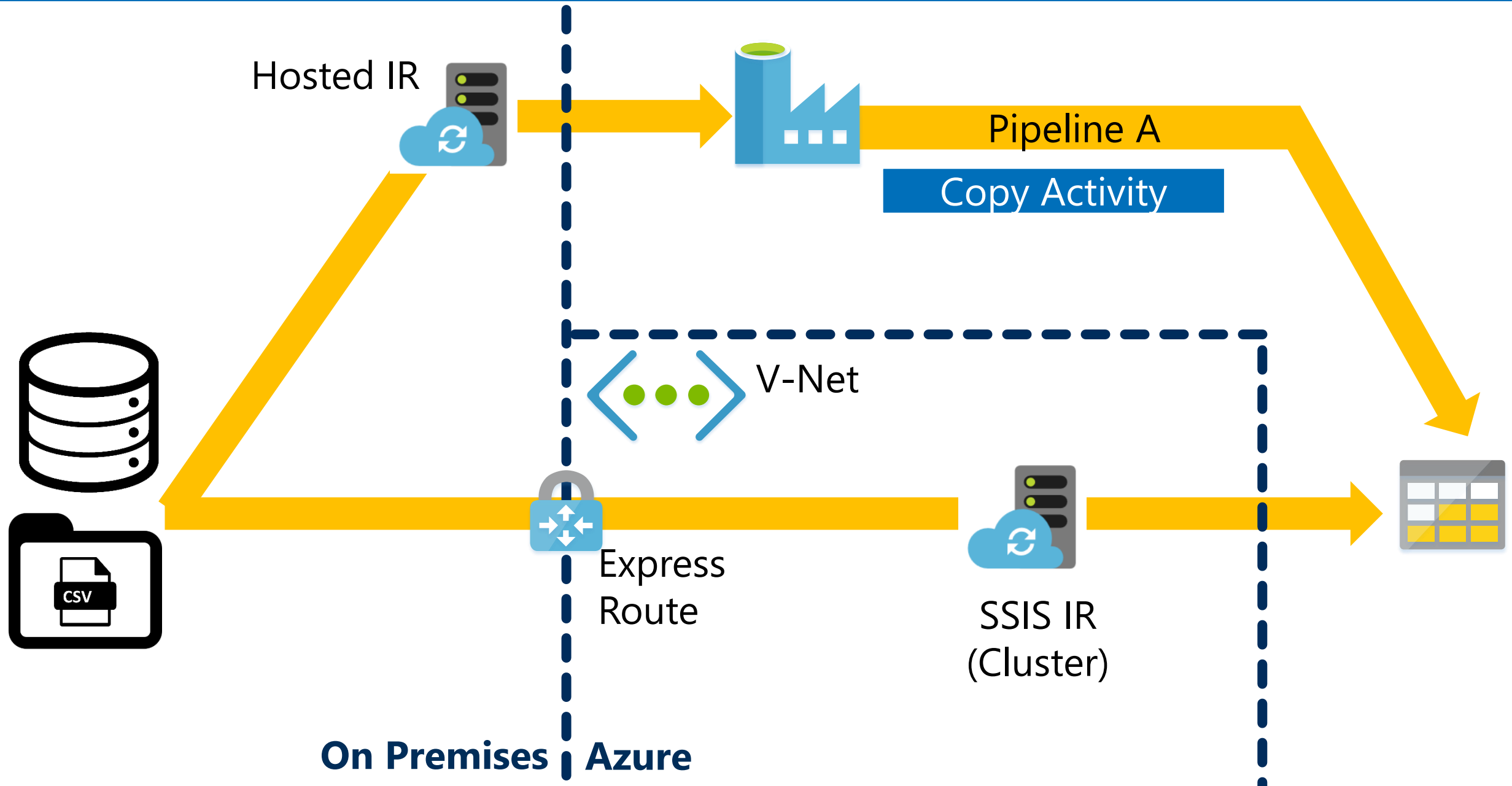
For Each Pipelines



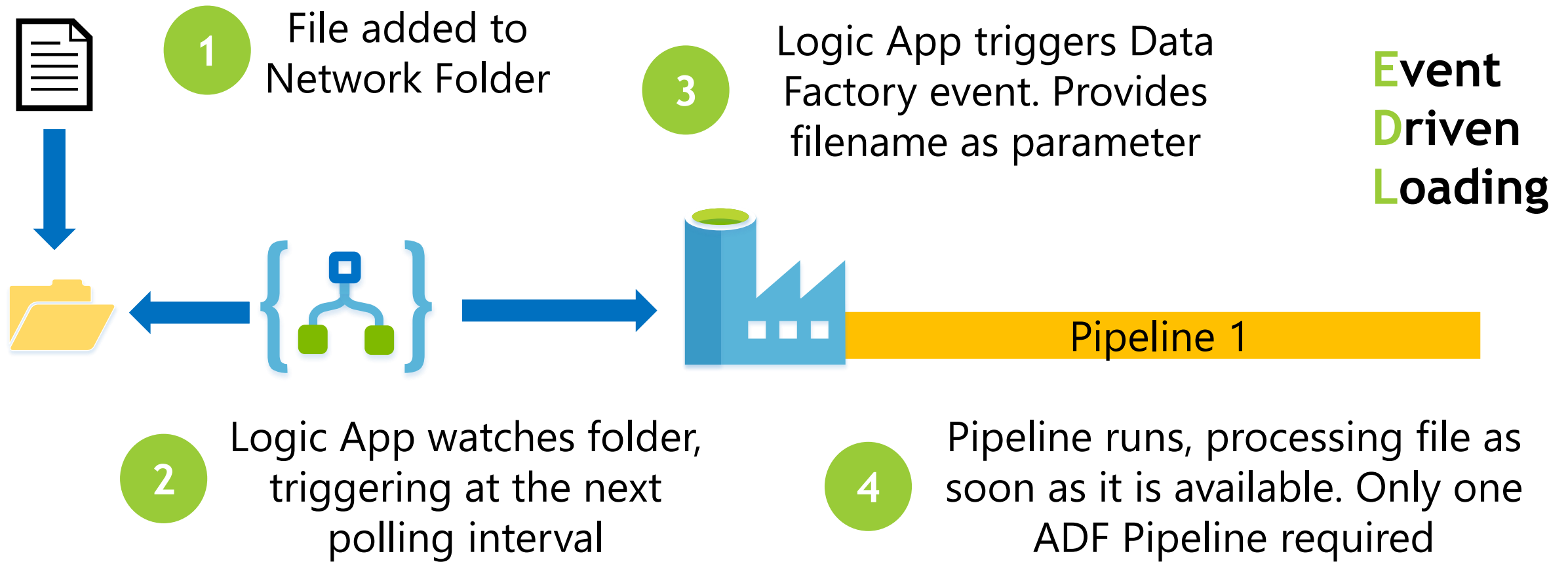
Parent/Child Pipelines & Triggering



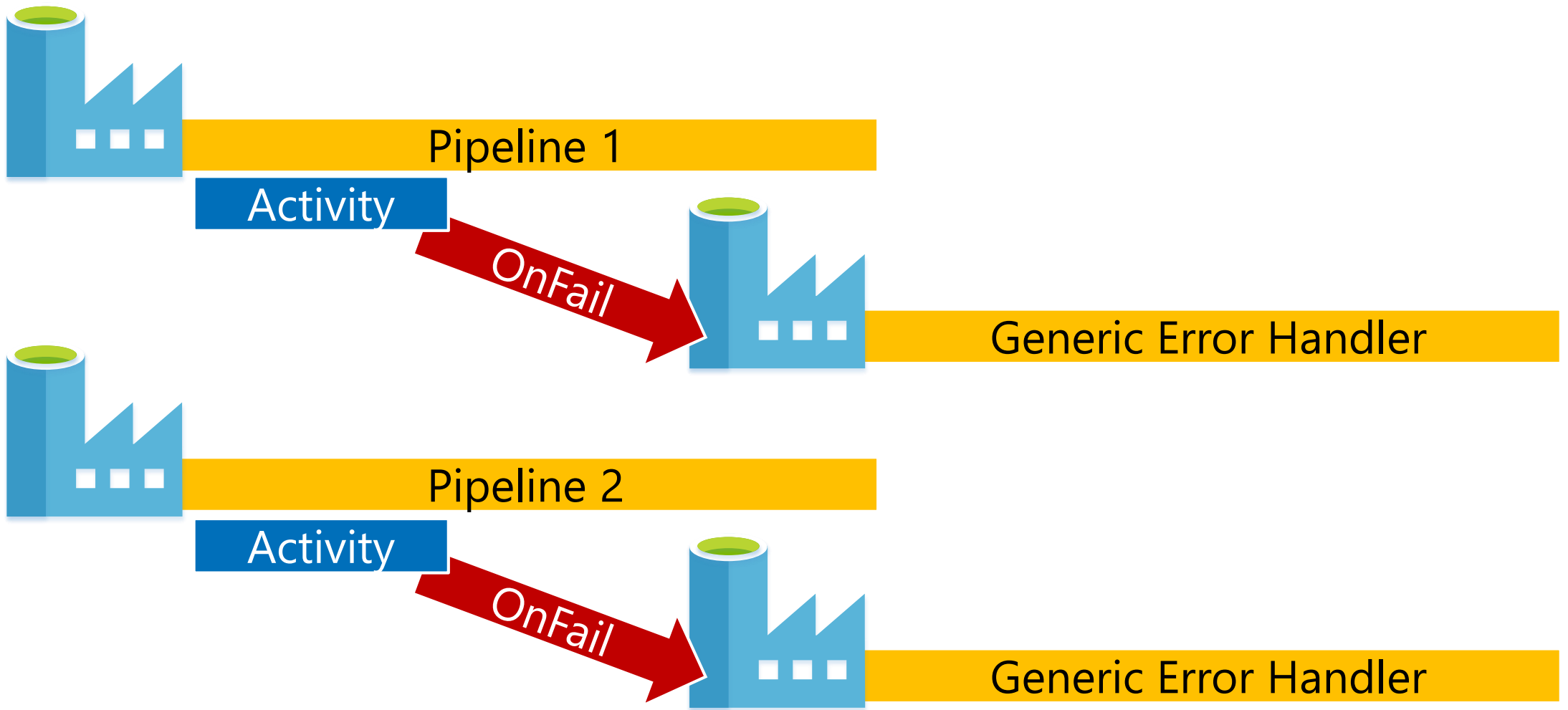
The SSIS IR with Azure V-Net Access



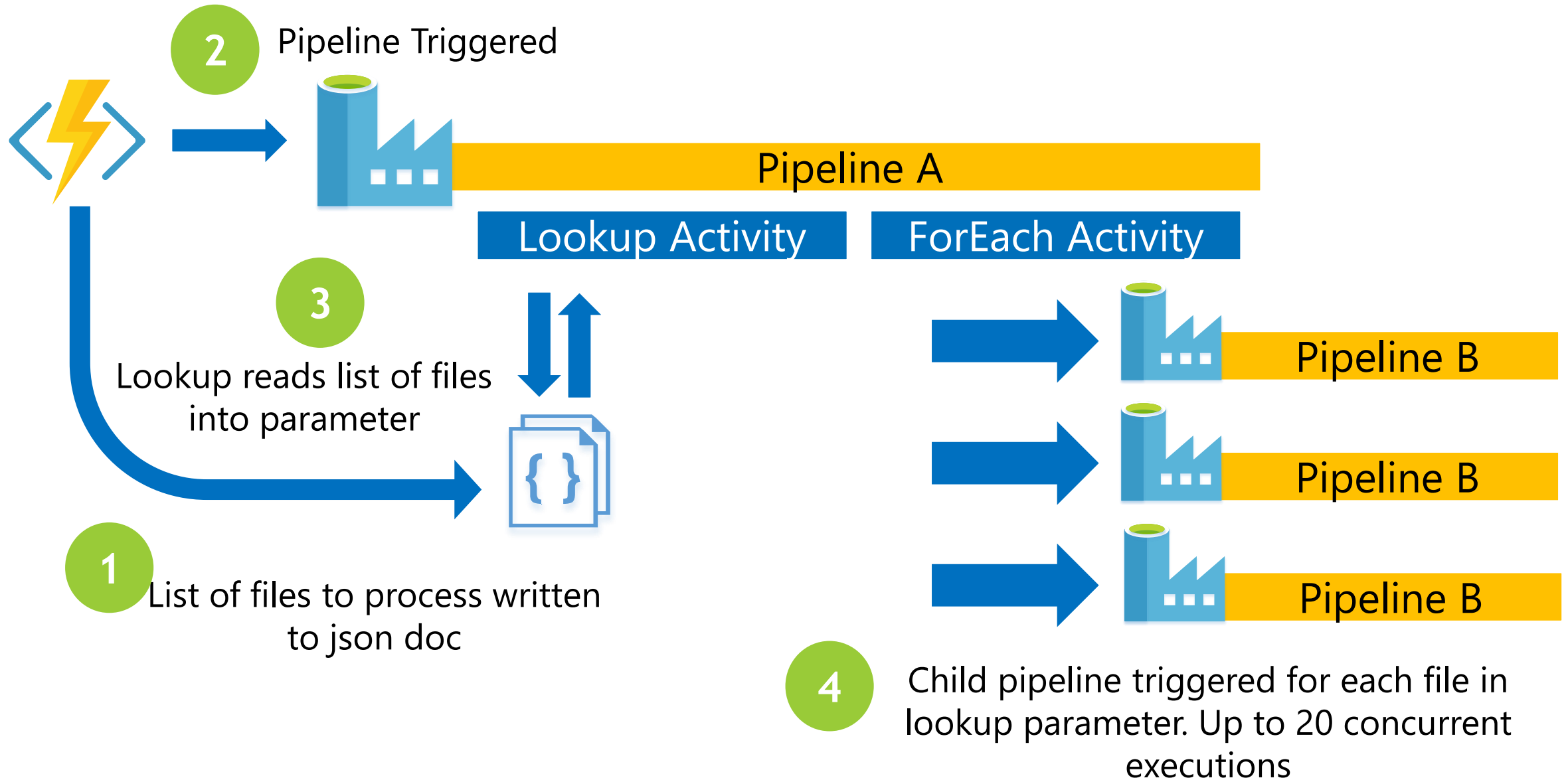
Event Driven Loading



Reusable Pipelines with Conditional Logic



Design Pattern Combinations





Is ADF the right tool for our data integration & orchestration in Azure?



Maybe, limited use.



Yes, definitely.

Thanks for Listening

Paul Andrew



@MrPaulAndrew



Blog: <http://mrpaulandrew.com>

Email: paul@mrpaulandrew.com