Azure Data Factory v2

Data Integration in the Cloud

Paul Andrew | Adatis

10/03/2018

Microsoft® MVP Most Valuable Professional

@MrPaulAndrew

Microsoft Partner

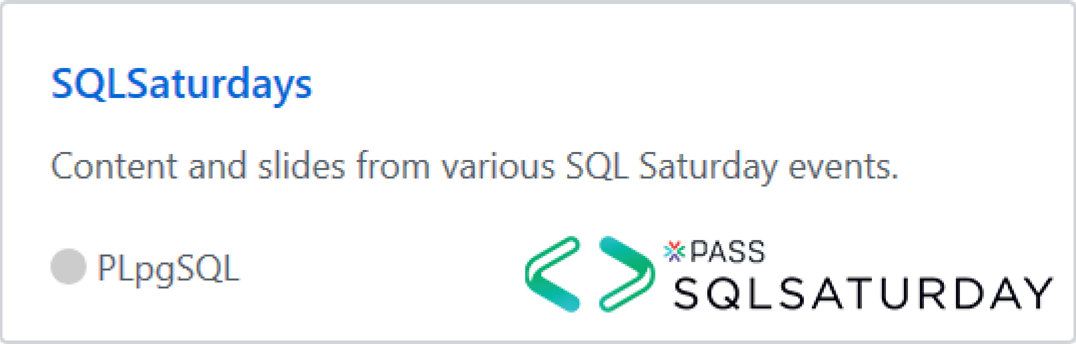
Gold Data Analytics
Gold Data Platform
Gold Cloud Platform

Microsoft

adatis

All Talk Content on GitHub

https://github.com/mrpaulandrew



SQLSaturdays

Content and slides from various SQL Saturday events.

● PLpgSQL

{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

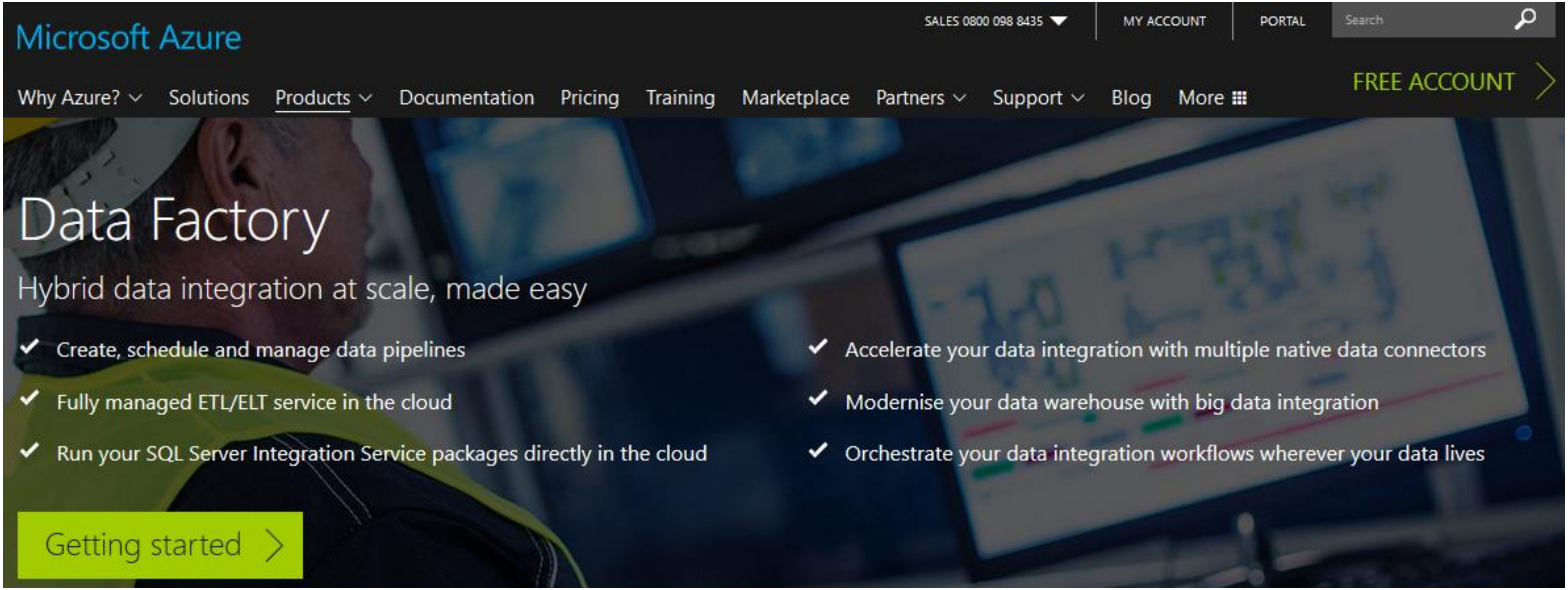
Conclusions

Design Patterns

ETL/ELT in Azure

What is Azure Data Factory?

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

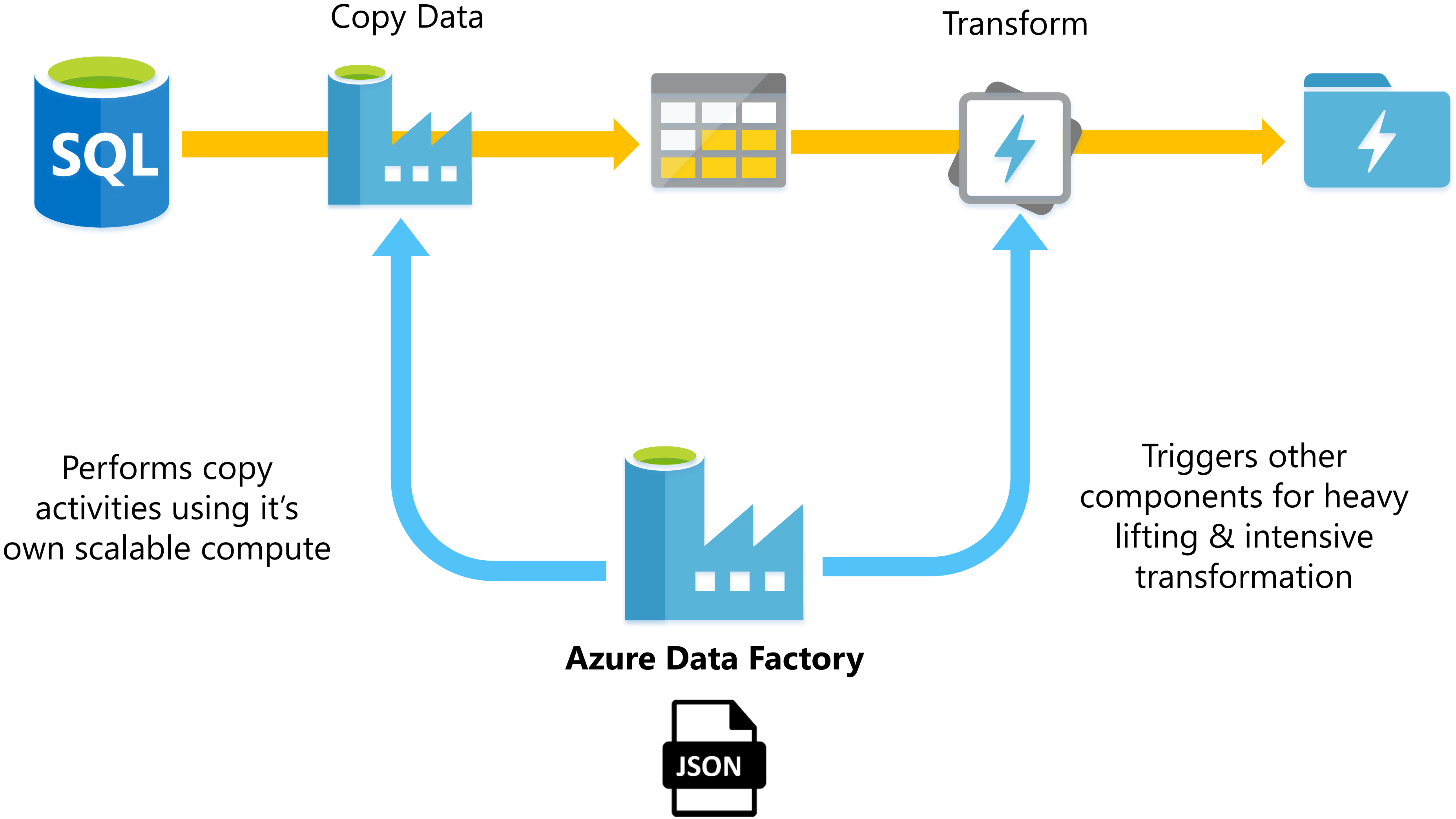


Data Factory

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

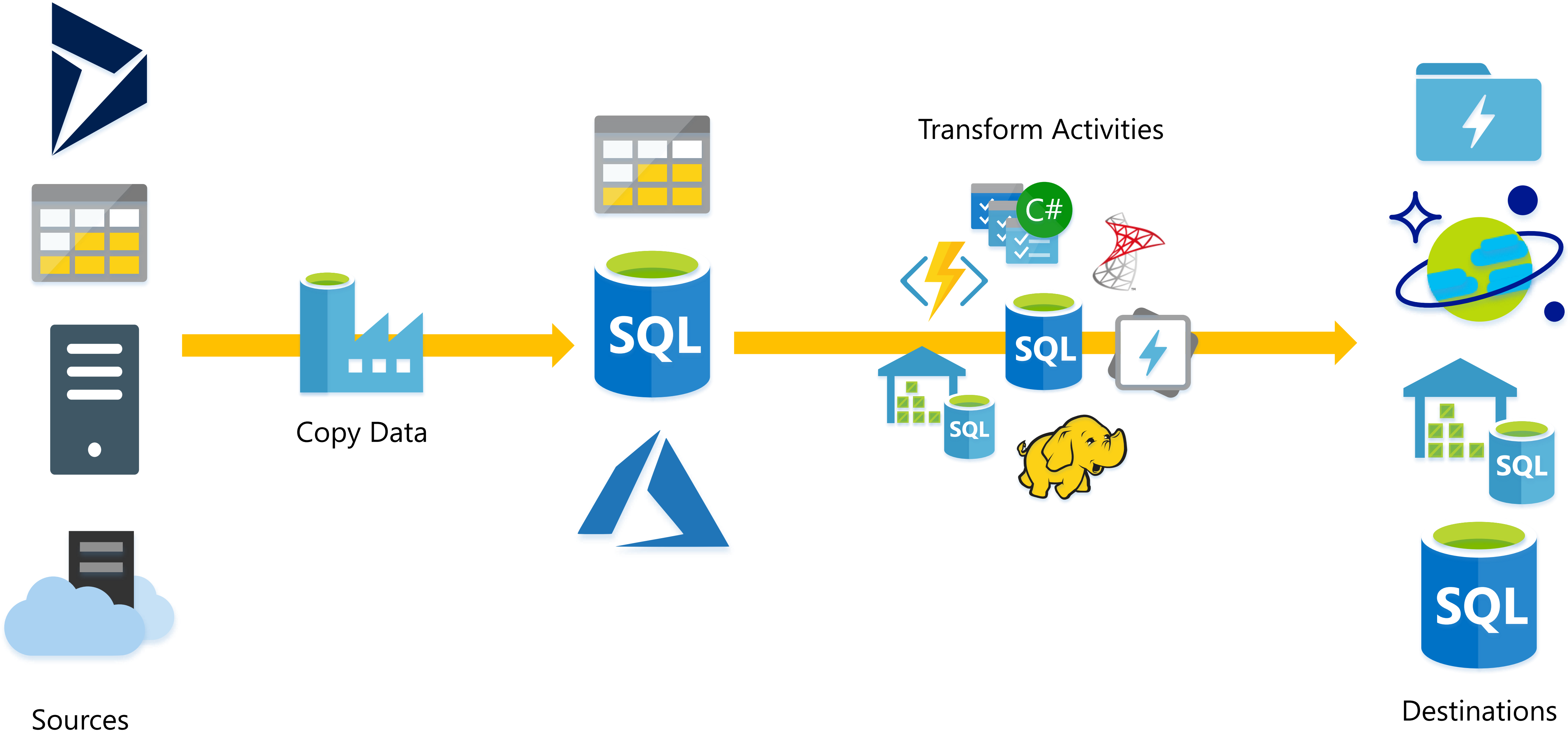
What is Azure Data Factory?



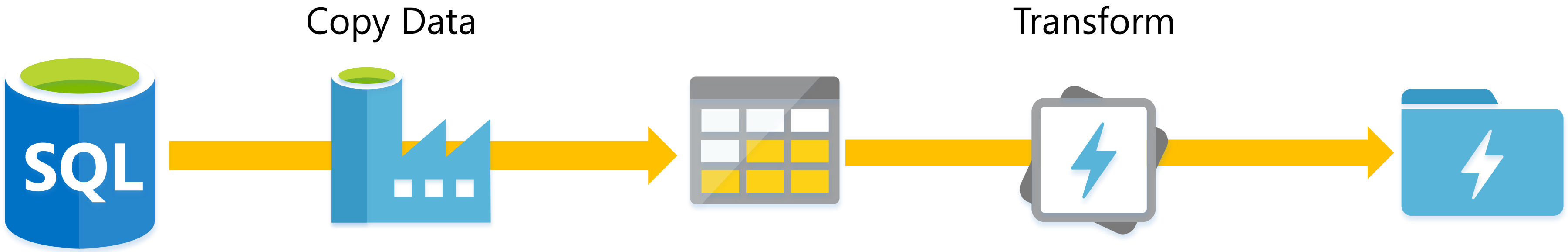
Performs copy activities using it’s own scalable compute

Triggers other components for heavy lifting & intensive transformation

What does Azure Data Factory do?



Data Factory Components



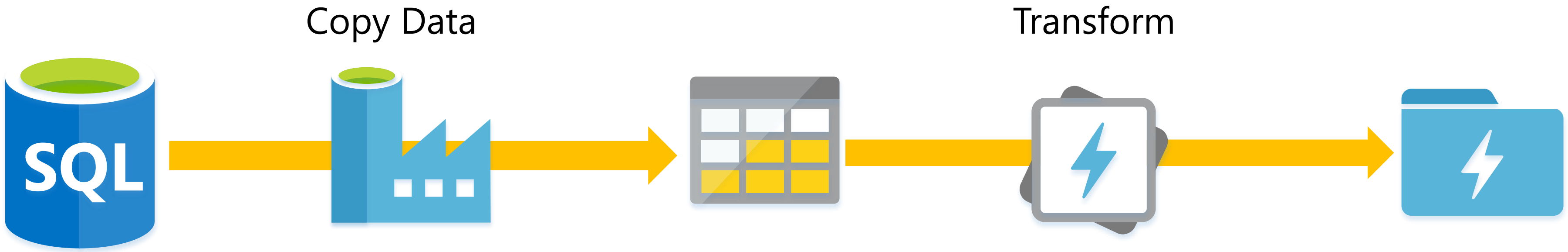
1 **Linked Services** – How do I connect?

Like the SSIS Connection Manager!

SQLDBLinkedService

ConnectionString: *Server=MyServer;Database=myDataBase*
UserName: “MrPaulAndrew”
Password: ****************

Data Factory Components

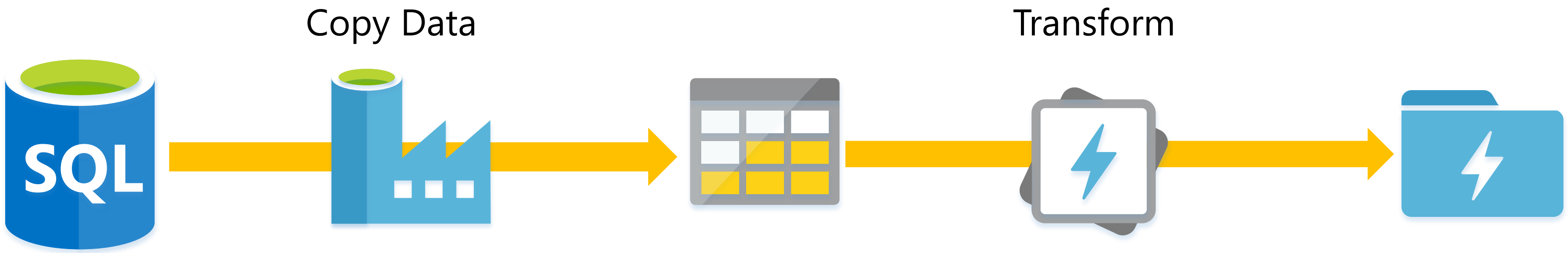


1. **Linked Services**
2. **Data Sets** – What slices/partitions does my data have?

   dbo.DimCustomer

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

Data Factory Components

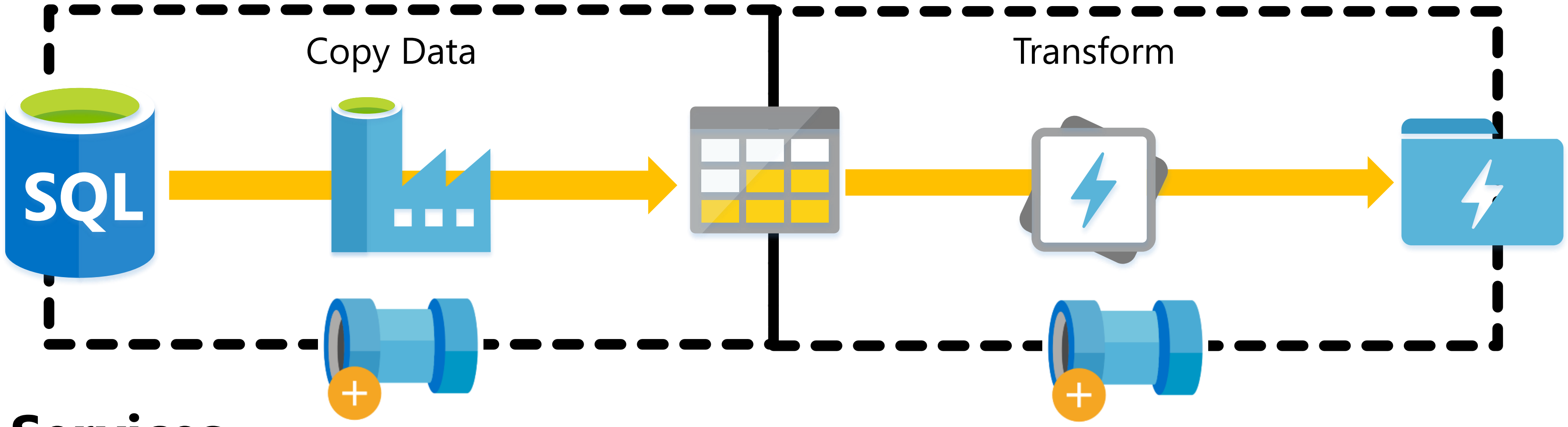


1. **Linked Services**
2. **Data Sets**
3. **Activities** – What do we want to happen?
   With what conditions?

U-SQL Activity

Script: *wasb//:myscripts/ProcessOrders.usql*
AUs: *5 units*
Priority: *1000*
Parameters: *@Output = "RAW/Orders/…"*

Data Factory Components



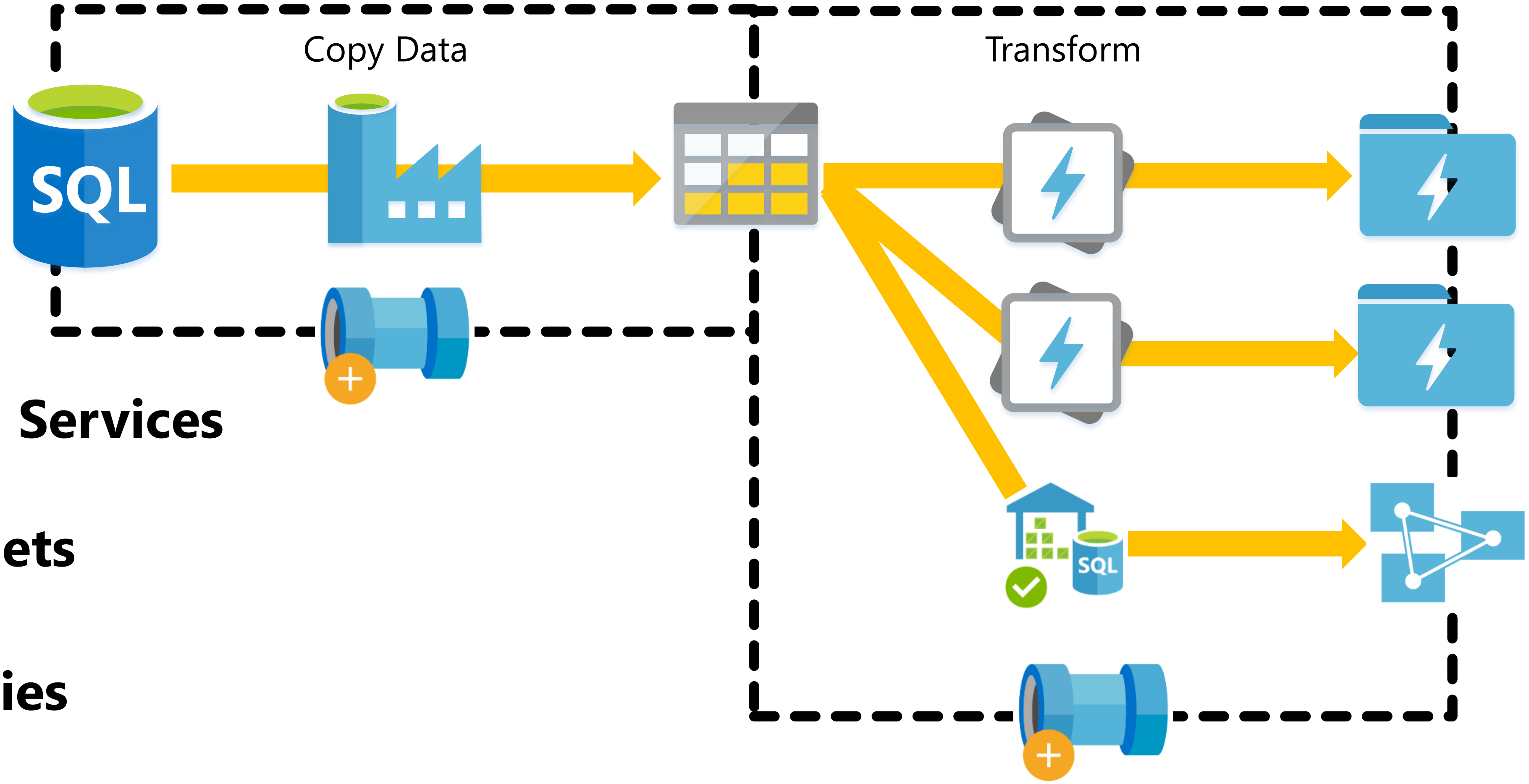
1. **Linked Services**

2. **Data Sets**

3. **Activities**

4. **Pipelines** – What groups of work do I want to do?

Data Factory Components



1 **Linked Services**

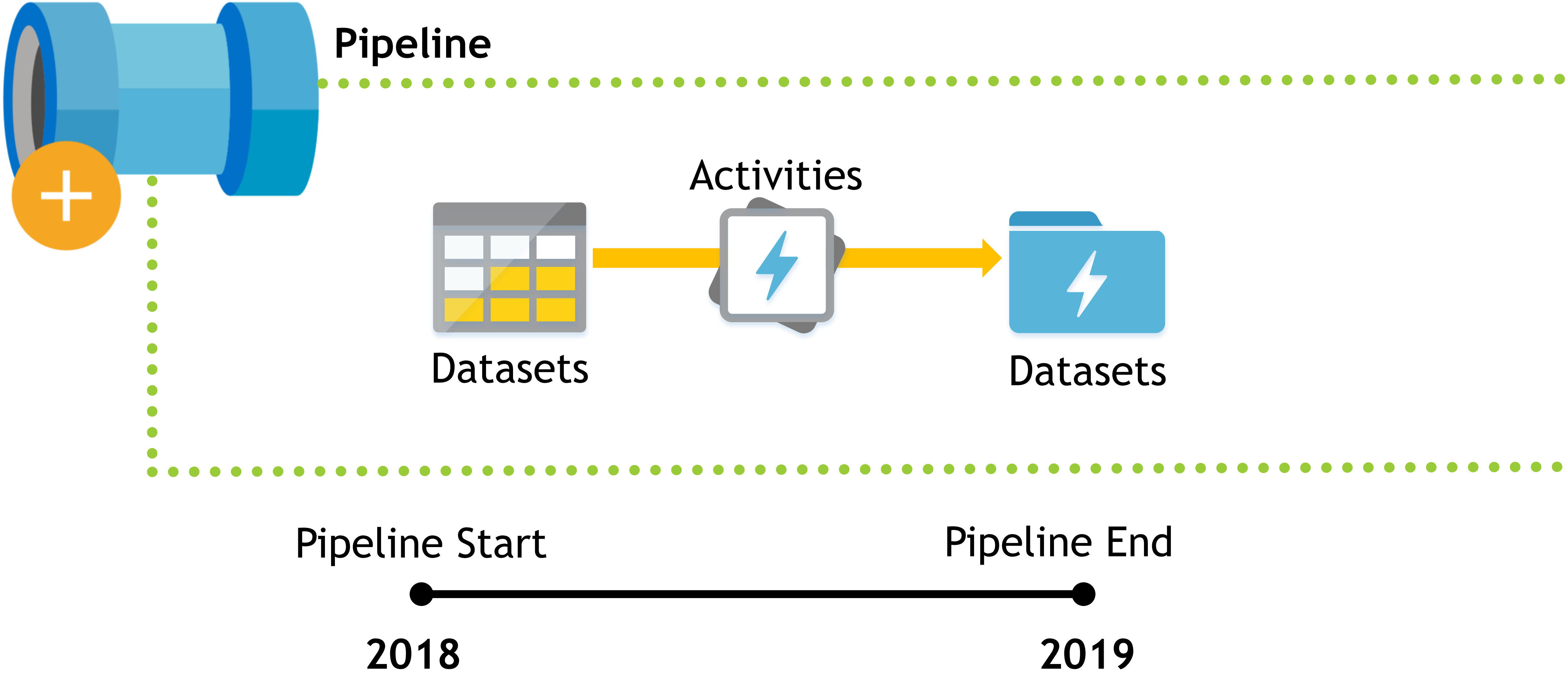
2 **Data Sets**

3 **Activities**

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

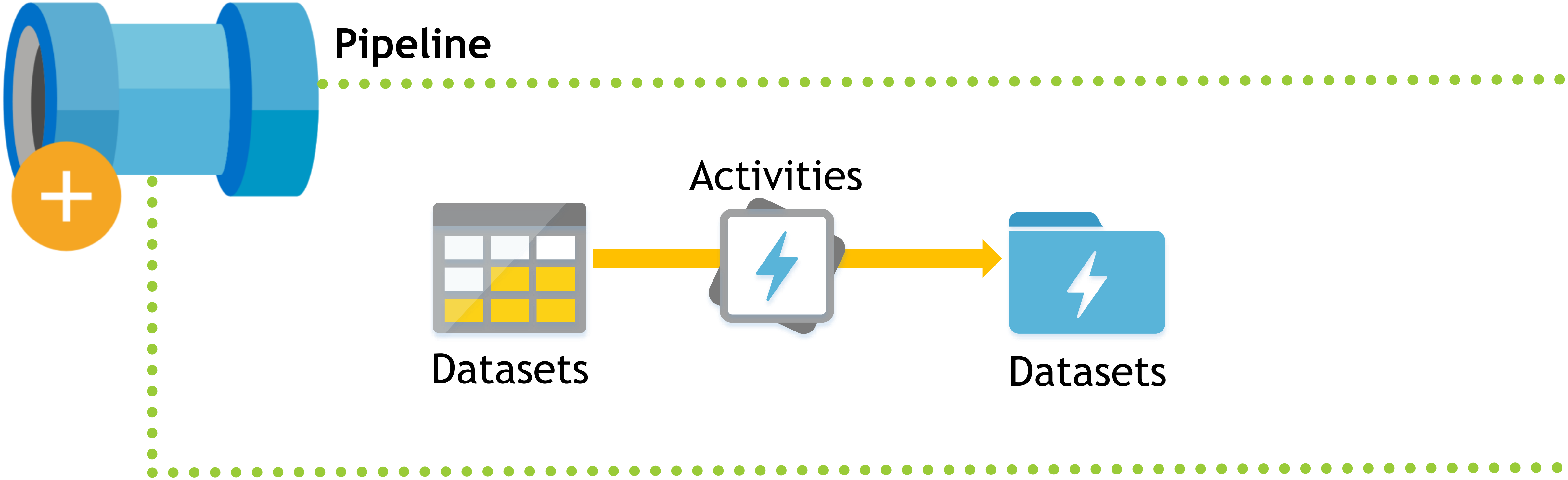
Azure Data Factory Concepts

Time Slices – triggering an activity execution.



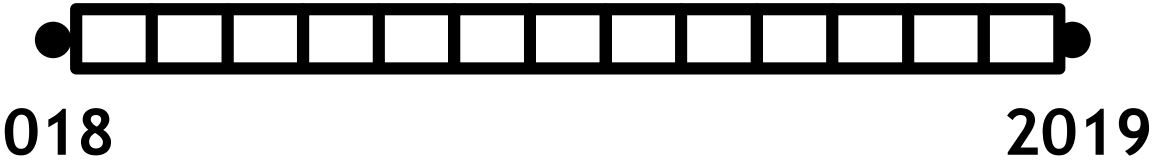
Azure Data Factory Concepts Continued

Time Slices – triggering an activity execution.



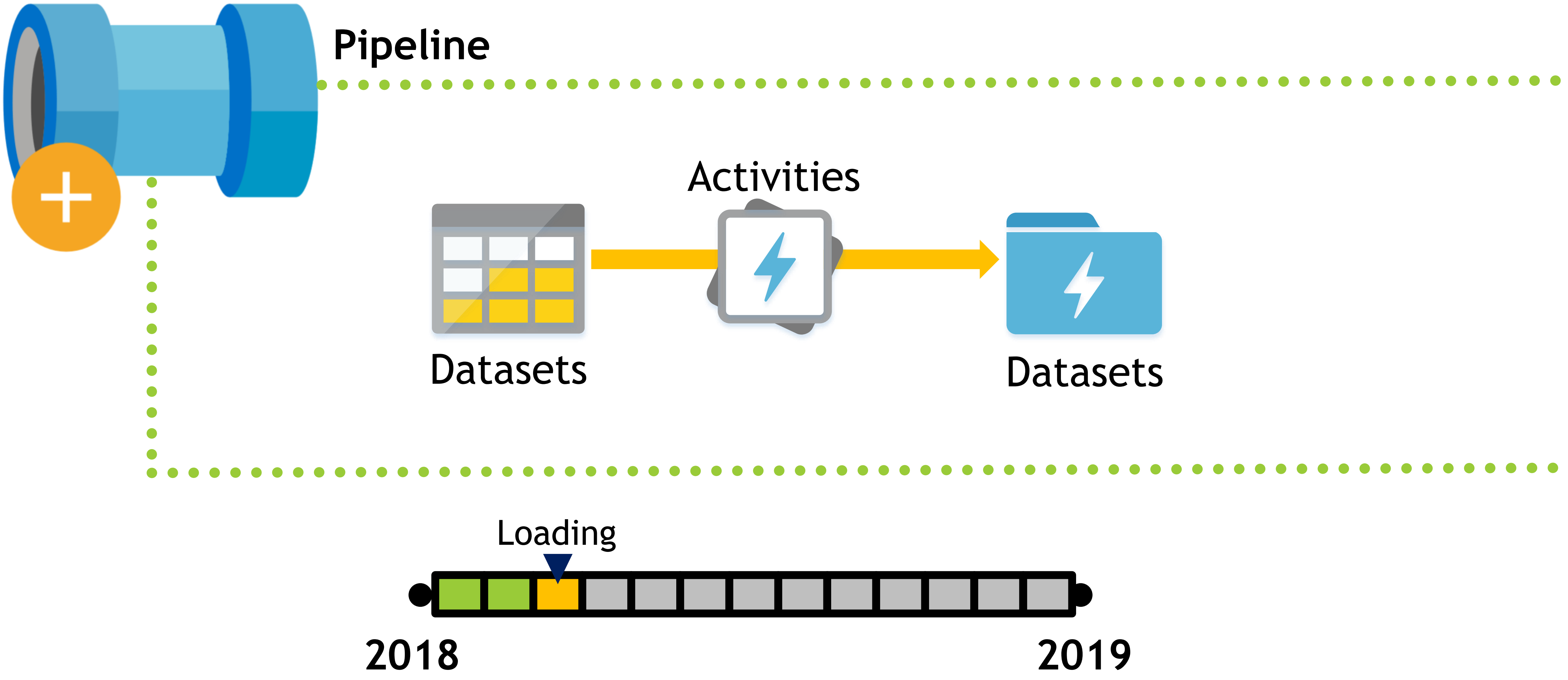
Interval: *Month*

Frequency: *1*

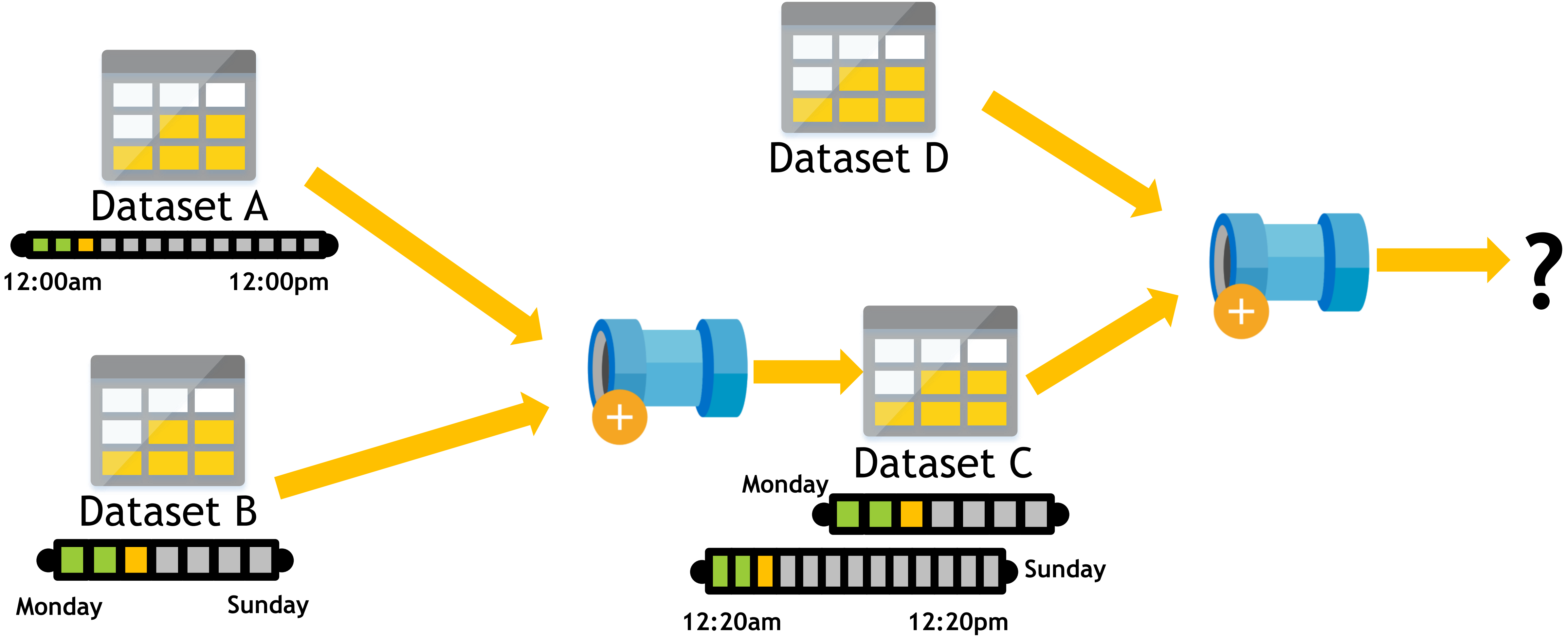


Azure Data Factory Concepts Continued

Time Slices – triggering an activity execution.



Time Slice Problems…



Agenda

Data Factory Recap

Concepts

Components

ADFv2

Features Update

The Integration Runtime

Data Factory Extensibility

SSIS, Functions, Custom Activities

Conclusions

Design Patterns

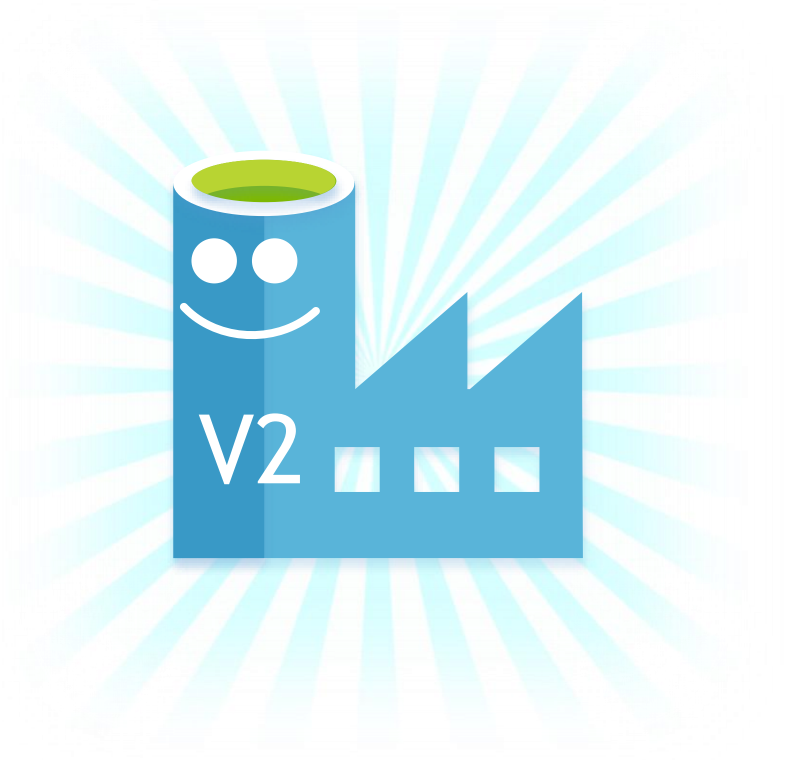
ETL/ELT in Azure

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

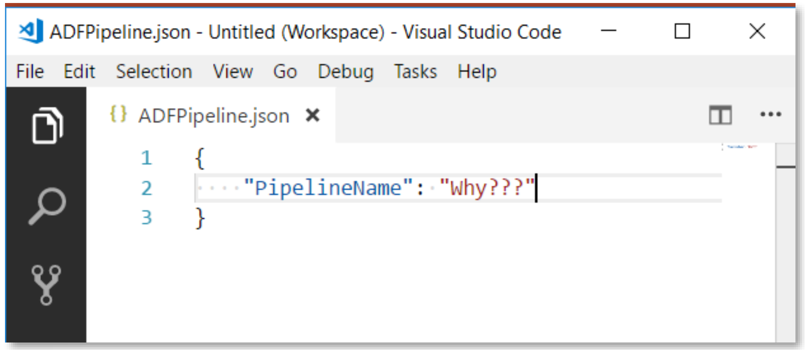
So What's Changed?

✓	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

```
{
    "PipelineName": "Why???"
}
```

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

Visual Studio

Author and deploy

Copy data (PREVIEW)

~~Visual Studio~~

Author & Monitor

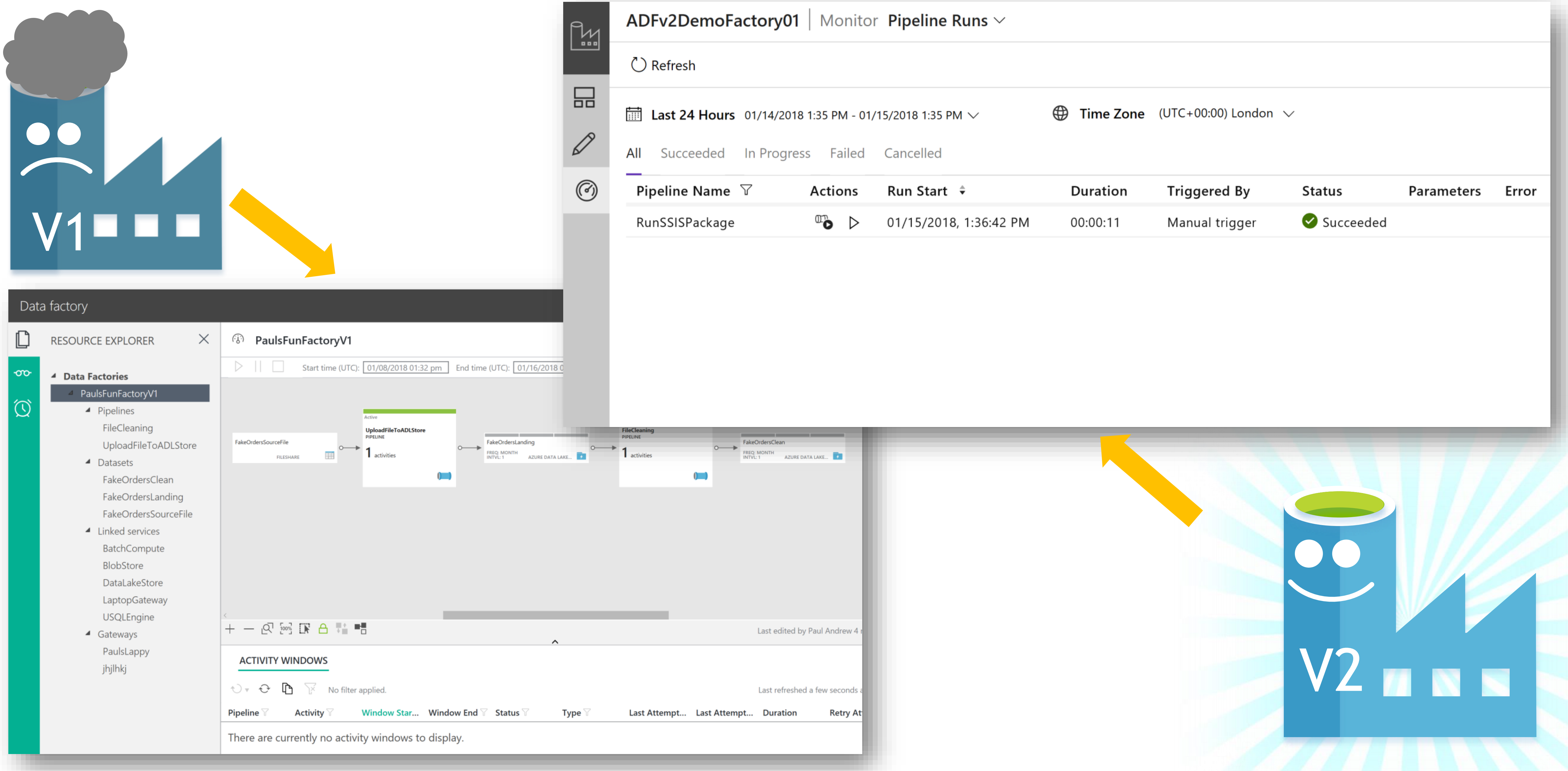
Microsoft .NET

PowerShell

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
      - jhjlhkj

PaulsFunFactoryV1

Start time (UTC): 01/08/2018 01:32 pm End time (UTC): 01/16/2018 0

Last edited by Paul Andrew 4 r

ACTIVITY WINDOWS

No filter applied.

Last refreshed a few seconds a

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

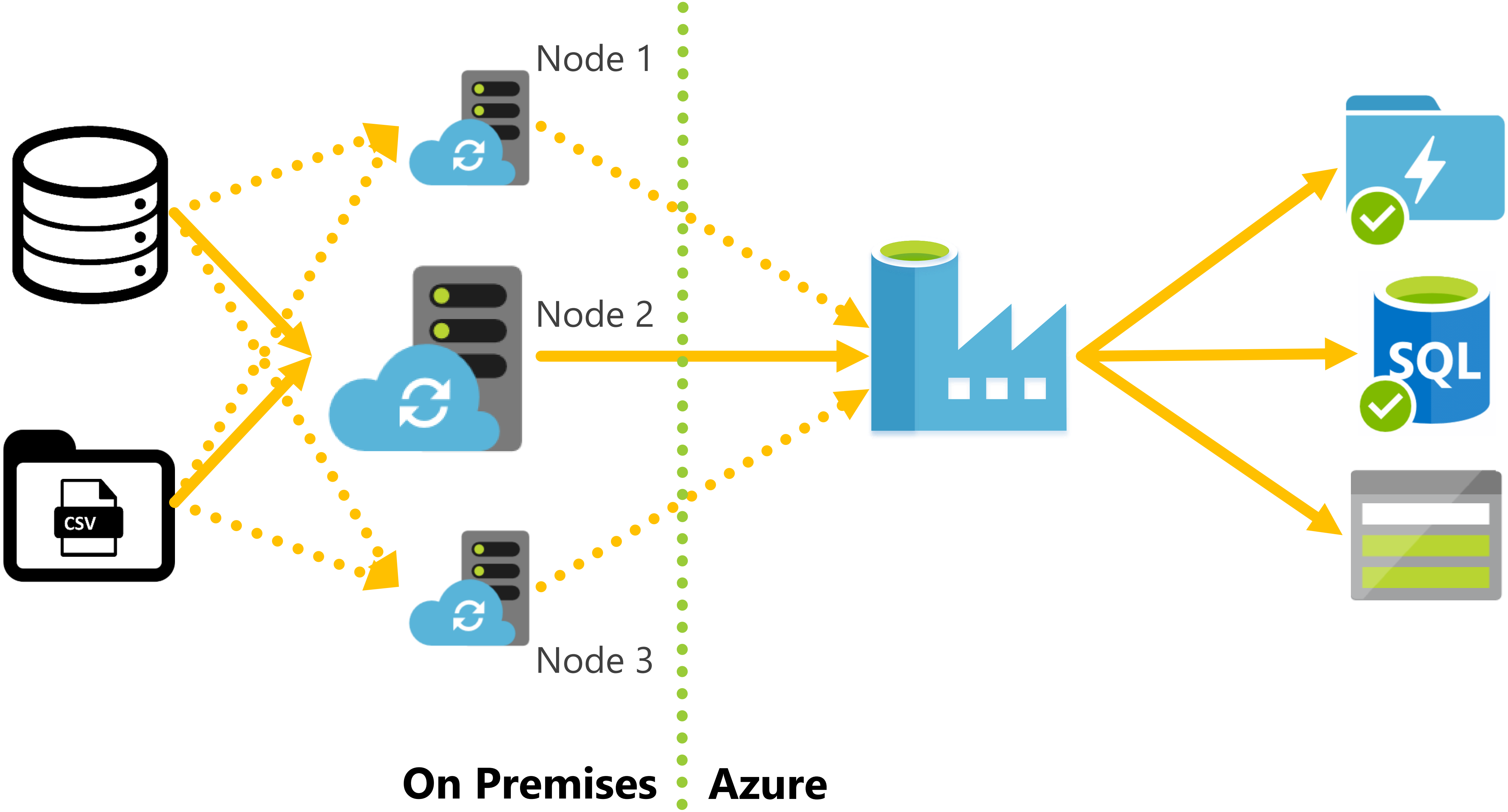
Last 24 Hours 01/14/2018 1:35 PM - 01/15/2018 1:35 PM

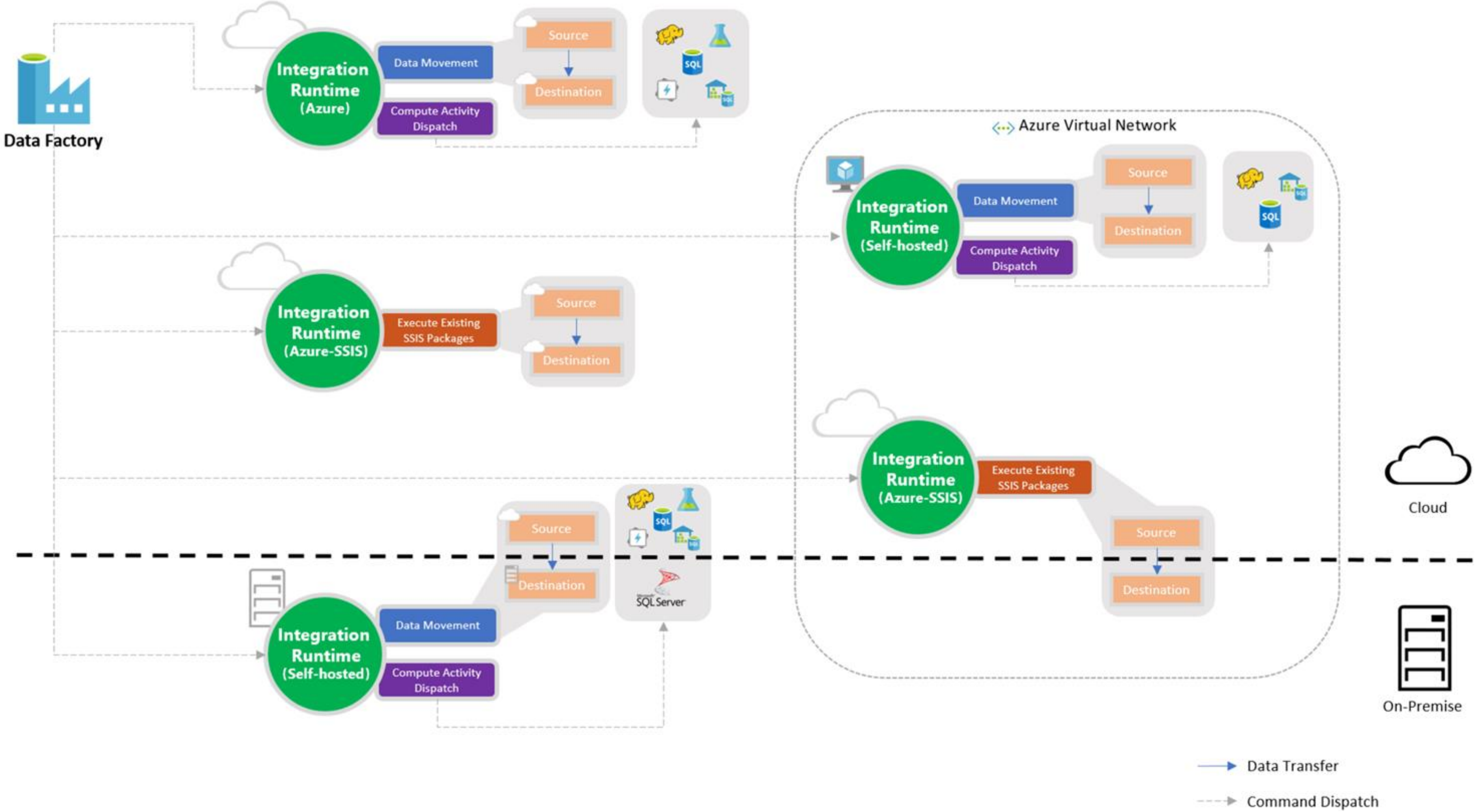
Time Zone (UTC+00:00) London

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		

The Integration Runtime *(AKA The Data Management Gateway)*




Data Factory
Integration Runtime (Azure)
Data Movement
Compute Activity Dispatch
Source
Destination
Azure Virtual Network
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
Source
Destination
SQL
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Cloud
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
Source
Destination
SQL Server
On-Premise
Data Transfer
Command Dispatch

Agenda

Data Factory Recap

Concepts
Components

ADFv2

Features Update
The Integration Runtime

Data Factory Extensibility

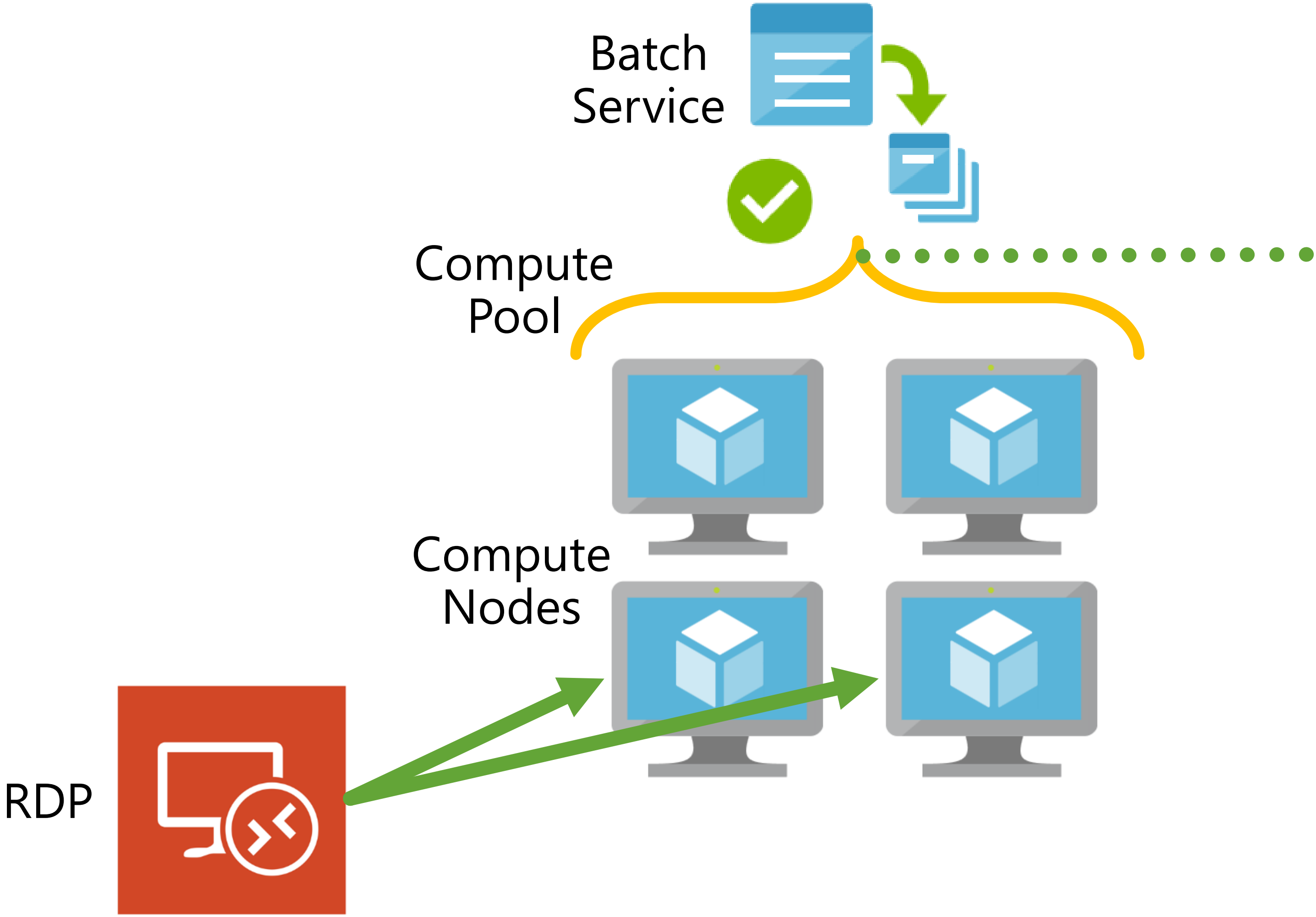
SSIS, Functions, Custom Activities

Conclusions

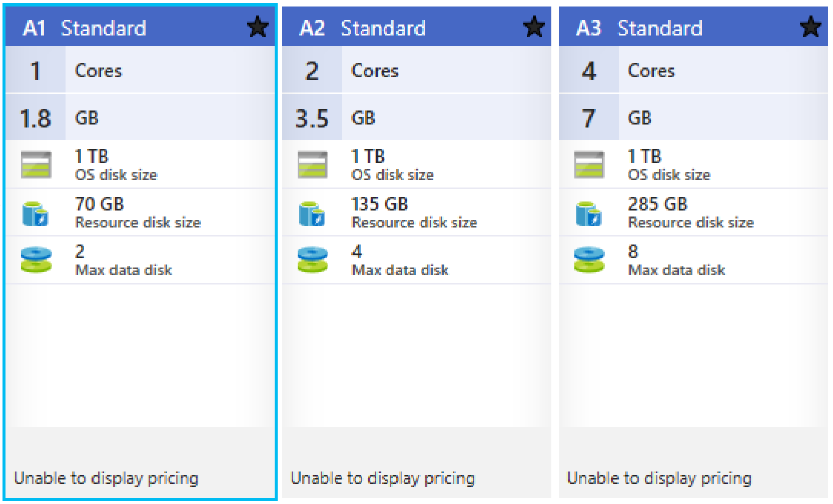
Design Patterns
ETL/ELT in Azure

ADF Extensibility

1 **Custom Activities** – C# classes called by the Azure Batch Service



VM node size set per compute pool:

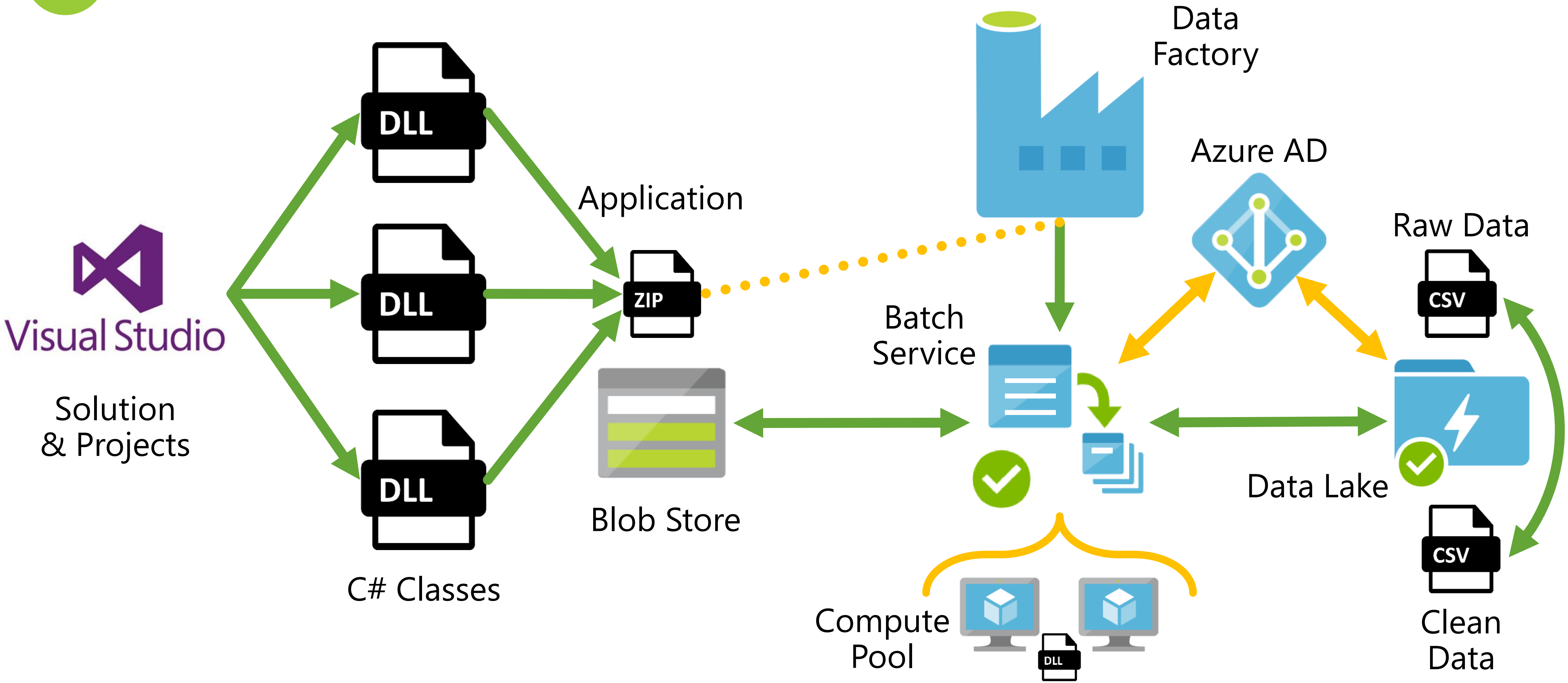


A1 Standard ★	A2 Standard ★	A3 Standard ★
1 Cores	2 Cores	4 Cores
1.8 GB	3.5 GB	7 GB
1 TB OS disk size	1 TB OS disk size	1 TB OS disk size
70 GB Resource disk size	135 GB Resource disk size	285 GB Resource disk size
2 Max data disk	4 Max data disk	8 Max data disk
Unable to display pricing	Unable to display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

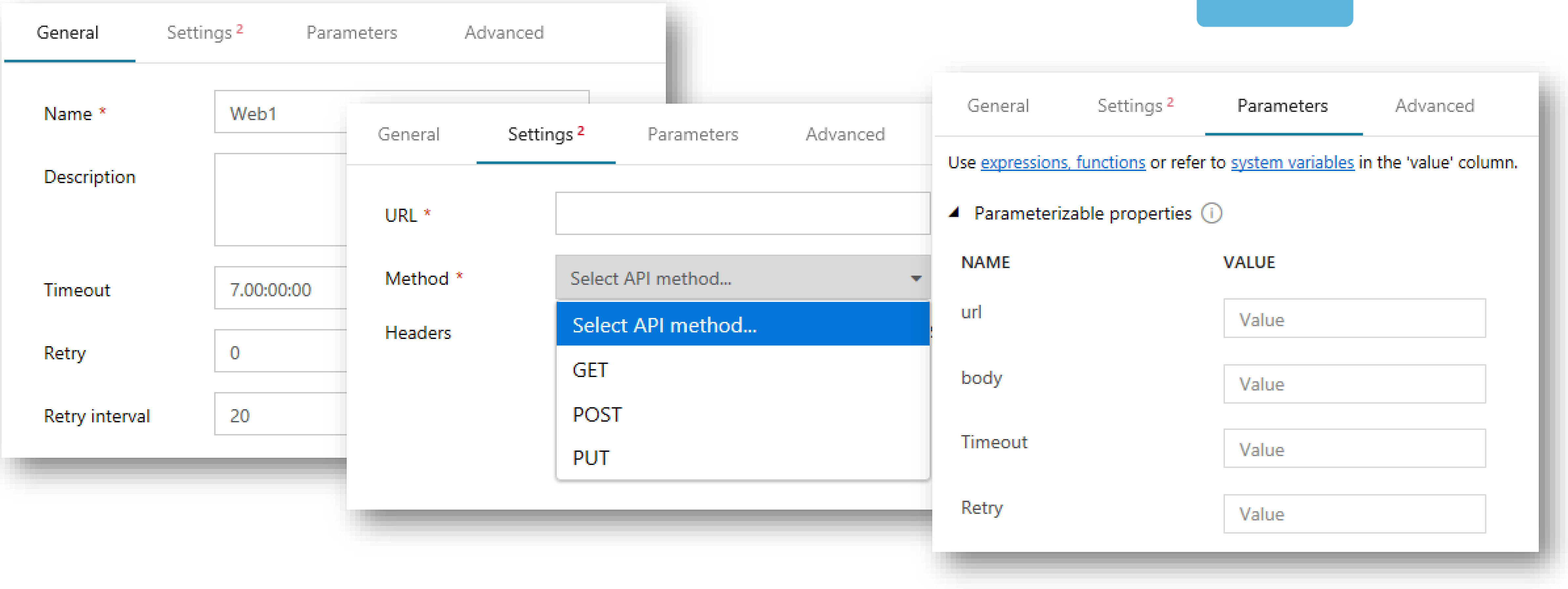
ADF Extensibility Continued

1 **Custom Activities** – C# classes called by the Azure Batch Service



ADF Extensibility Continued

1. **Custom Activities**
2. **Rest API Calls** – Web Activities Calling Azure Functions



General | Settings 2 | Parameters | Advanced

Name * Web1

Description

Timeout 7.00:00:00

Retry 0

Retry interval 20

General | Settings 2 | Parameters | Advanced

URL *

Method * Select API method...

Headers

General | Settings 2 | Parameters | Advanced

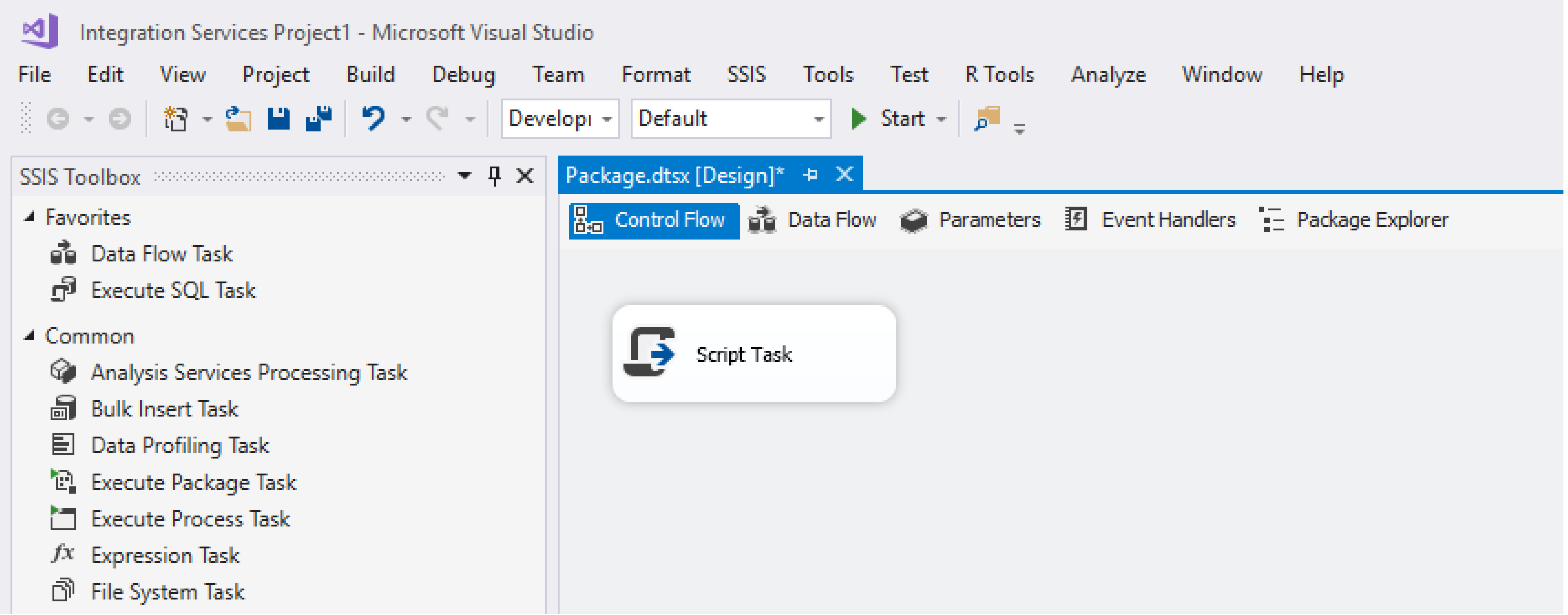
Use expressions, functions or refer to system variables in the 'value' column.

◢ Parameterizable properties ⓘ

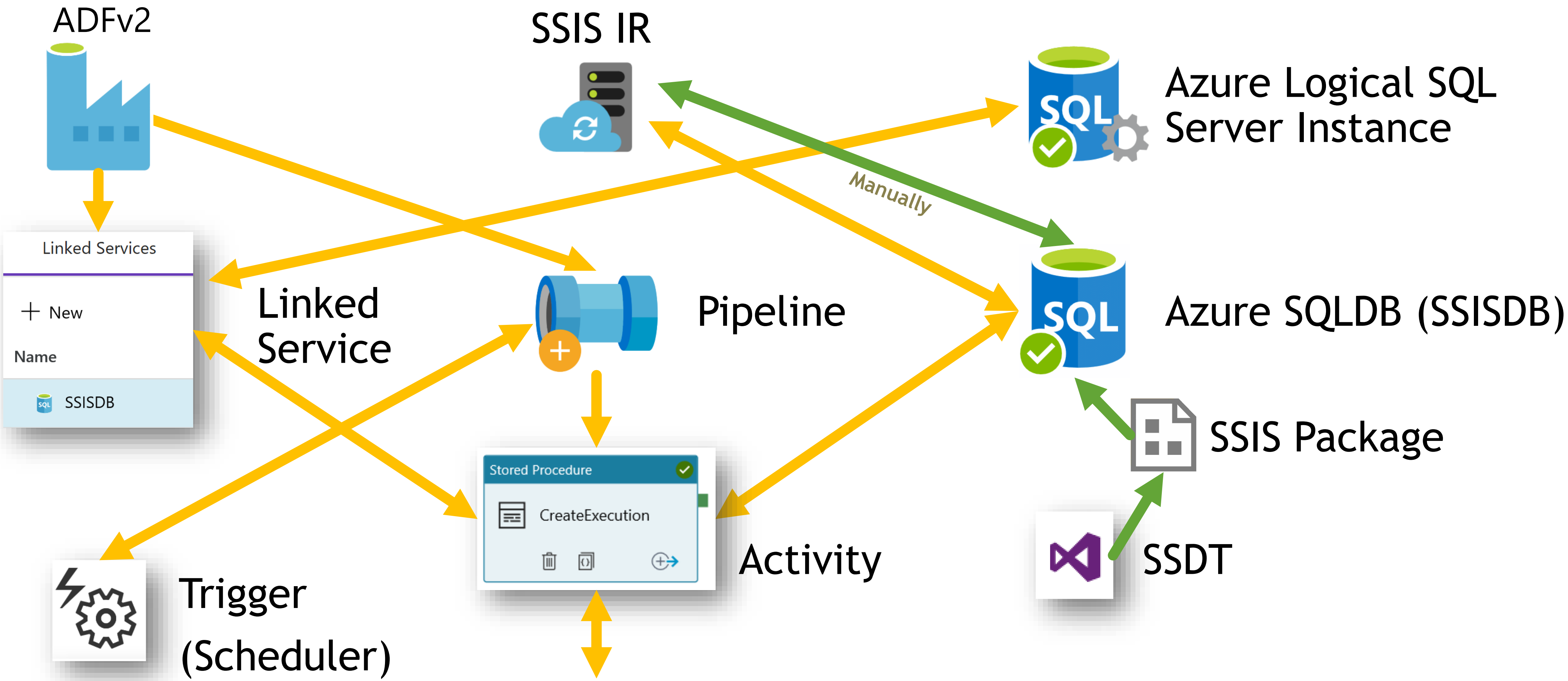
NAME	VALUE
url	Value
body	Value
Timeout	Value
Retry	Value

ADF Extensibility Continued

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



What do we need to schedule an SSIS Package in Azure?



ADFv2

SSIS IR

Azure Logical SQL Server Instance

Linked Services

+ New

Name

SSISDB

Linked Service

Pipeline

Manually

Azure SQLDB (SSISDB)

SSIS Package

Stored Procedure

CreateExecution

Activity

SSDT

Trigger (Scheduler)

```
EXEC [catalog].[start_execution] @execution_id;
```

Demo: ADFV2

Agenda

Data Factory Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

Data Factory Extensibility

SSIS, Functions,
Custom Activities

Conclusions

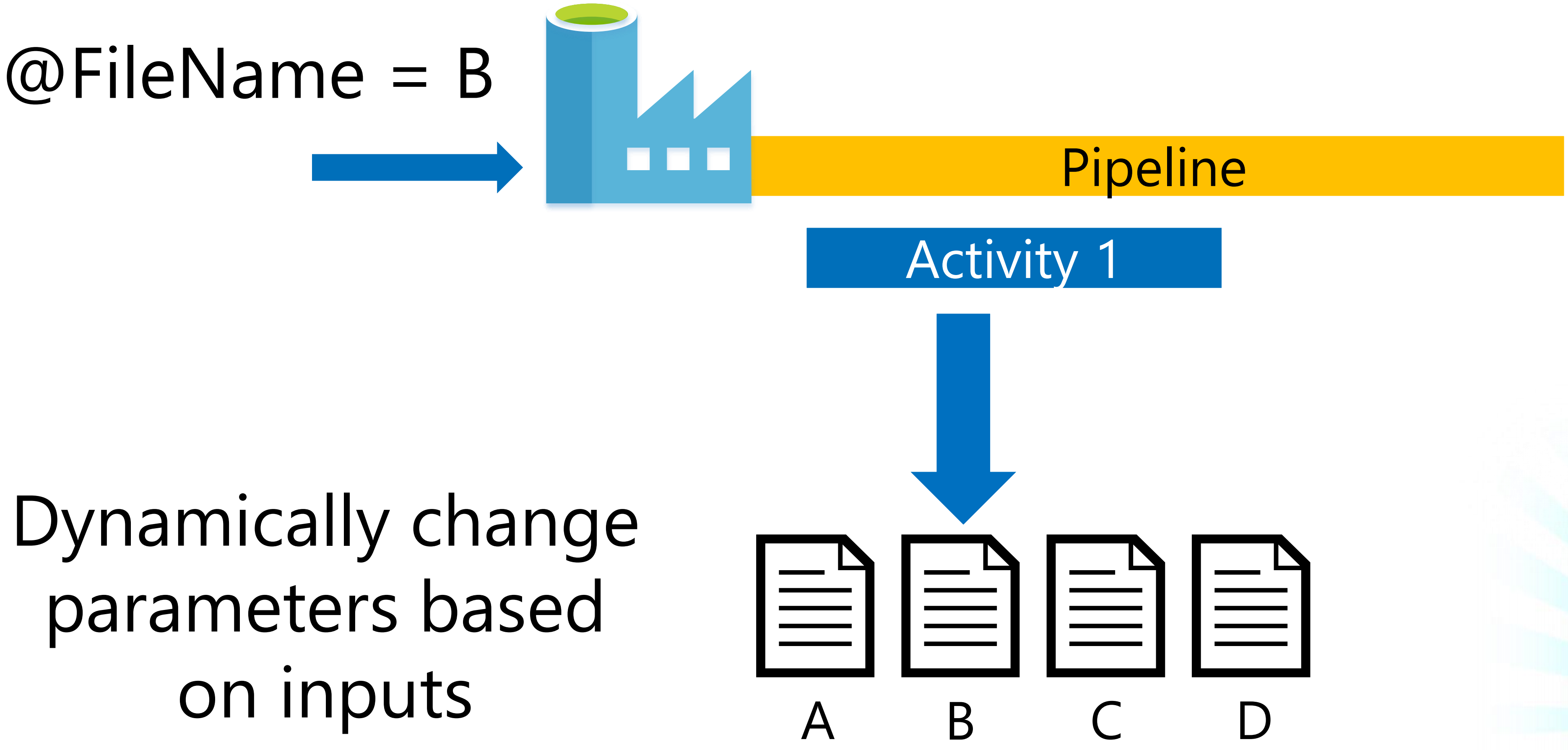
Design Patterns
ETL/ELT in Azure

Is ADF the right tool for our data integration?

Maybe, limited use.

Yes, definitely.

Dynamic Pipelines using Parameters



@FileName = B

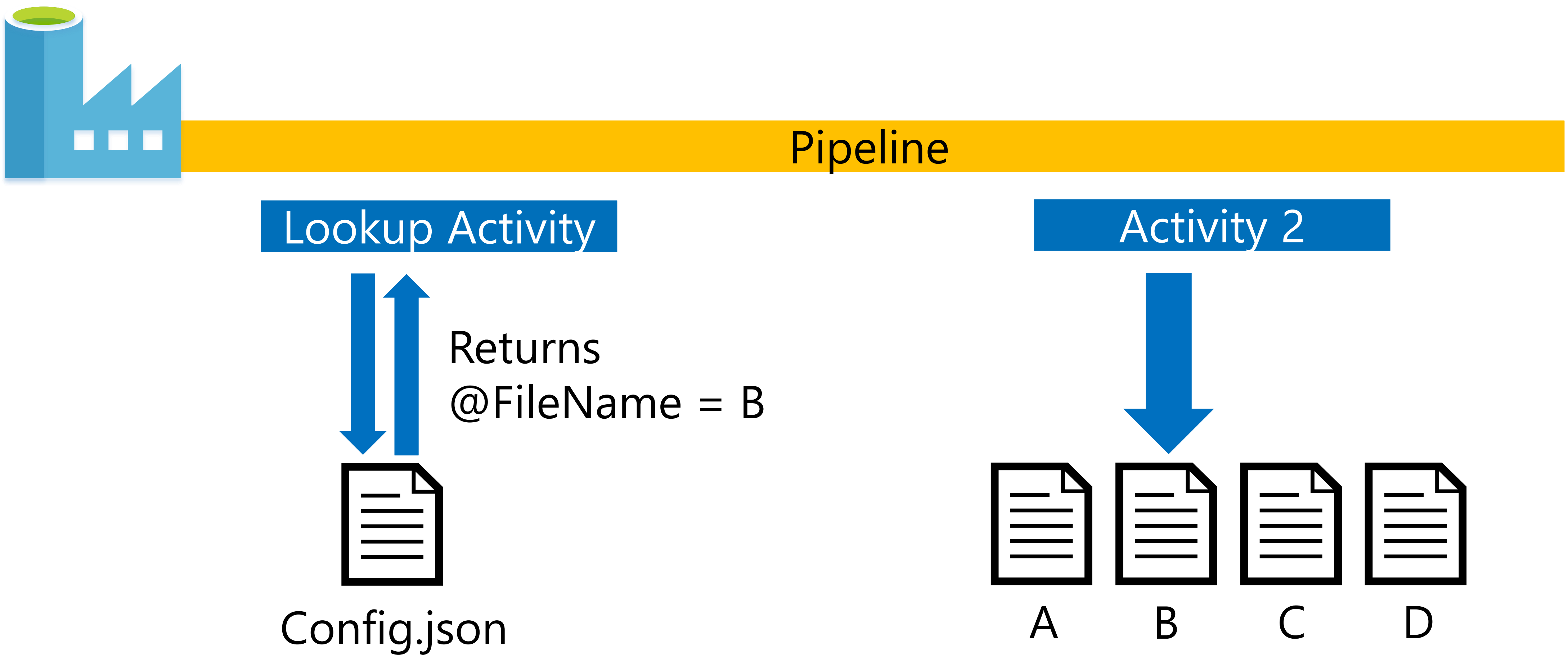
Pipeline

Activity 1

A B C D

Dynamically change parameters based on inputs

Dynamic Pipelines using Lookup Activity



Pipeline

Lookup Activity

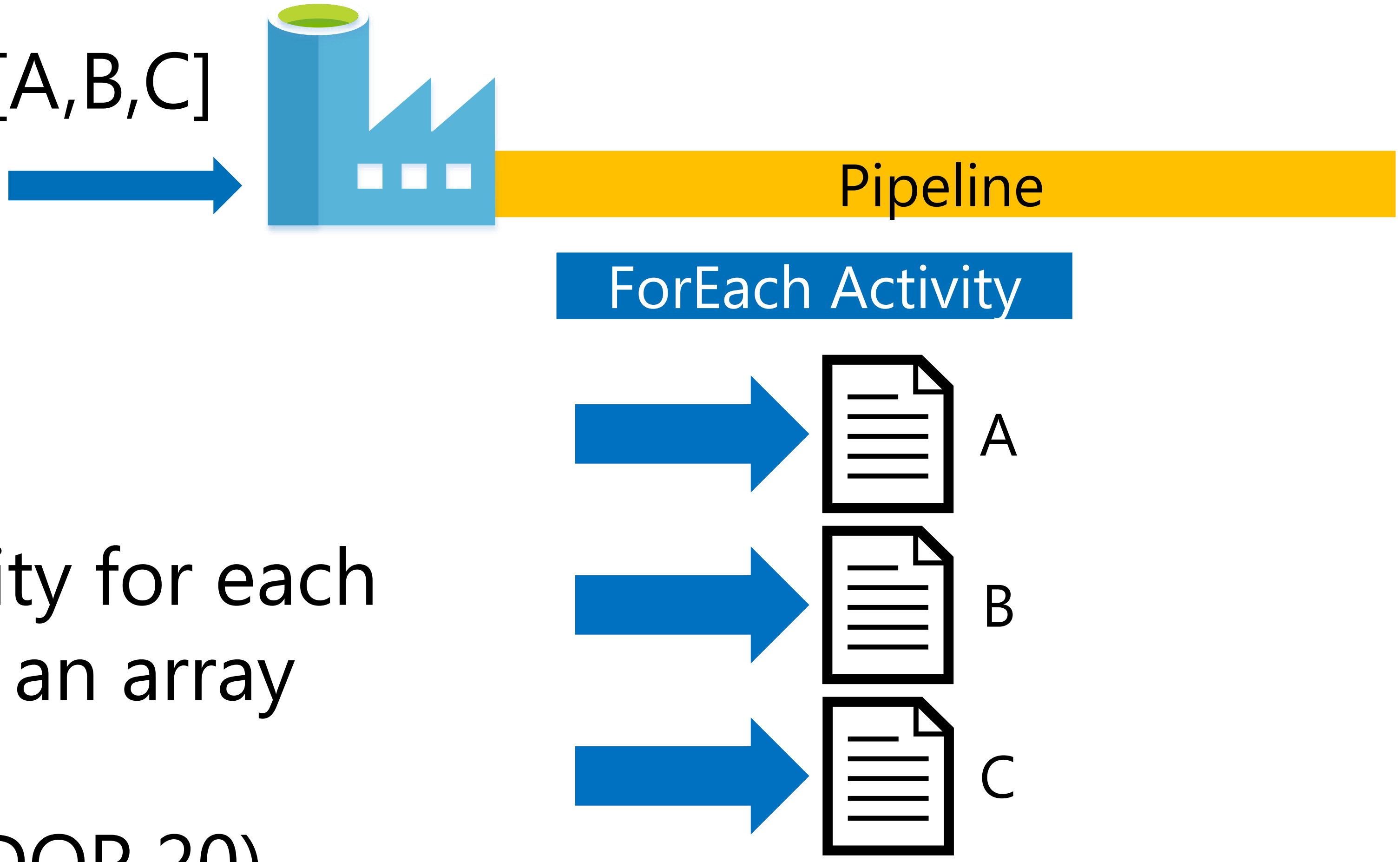
Activity 2

Returns
@FileName = B

Config.json

ForEach Pipelines

@Files = [A,B,C]



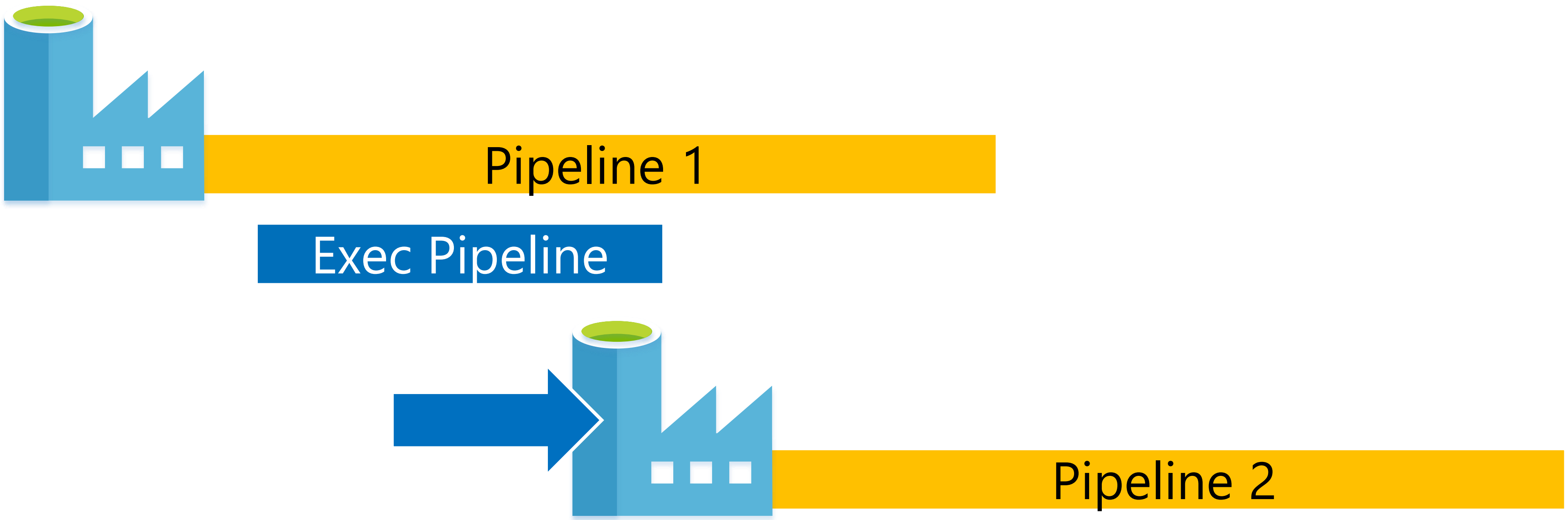
Pipeline

ForEach Activity

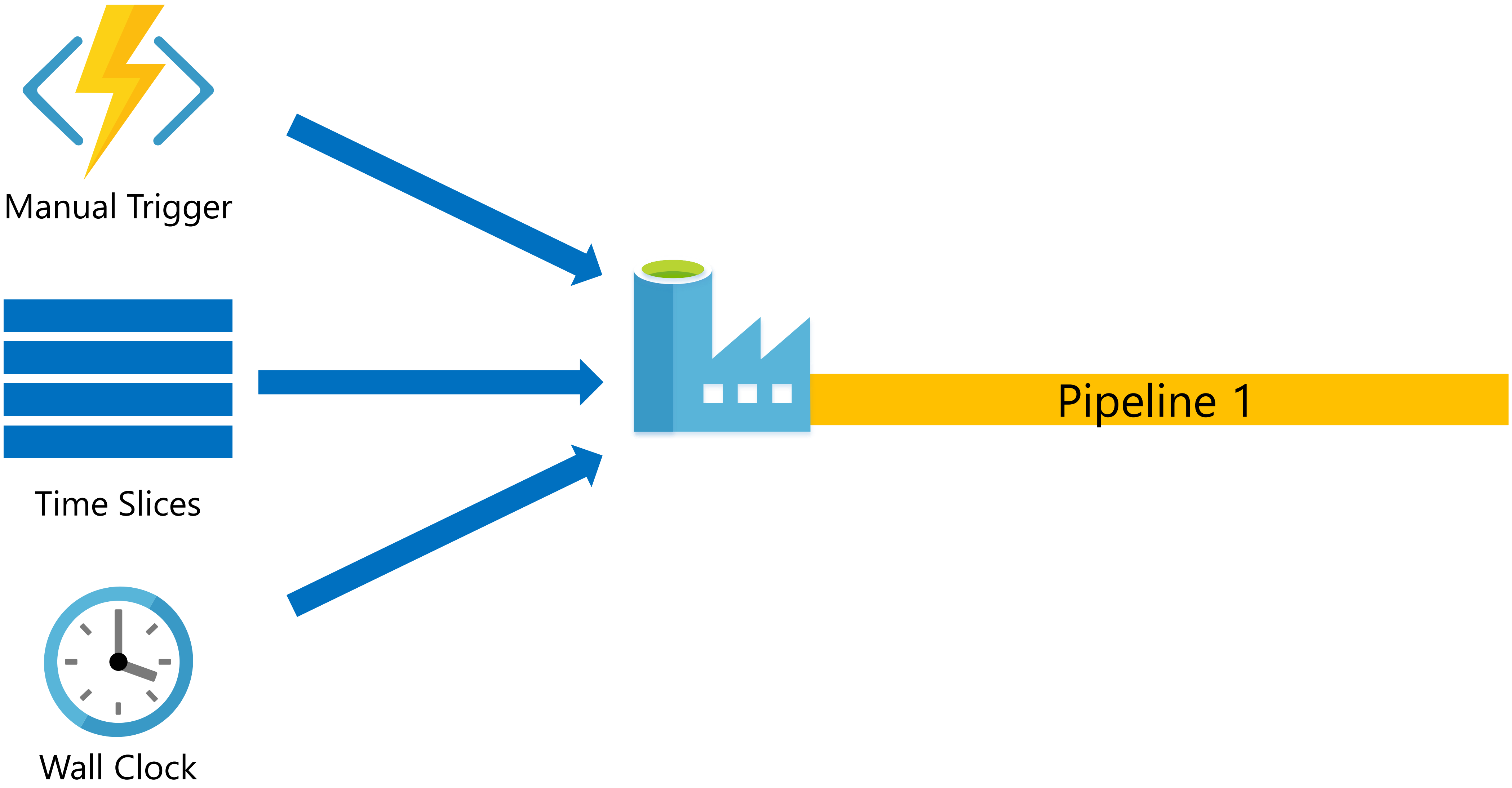
Run activity for each item in an array

(max DOP 20)

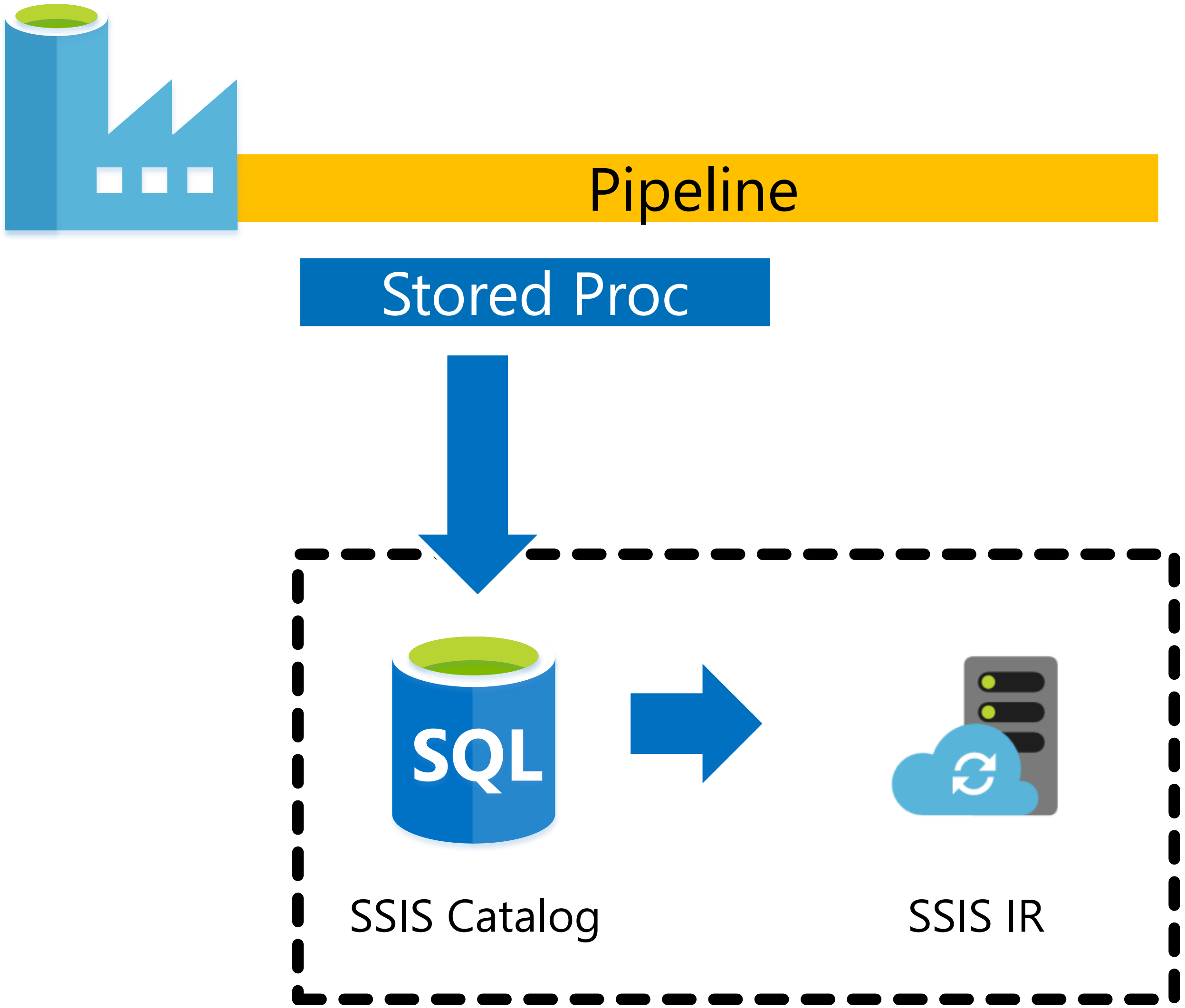
Execute Pipeline Activity



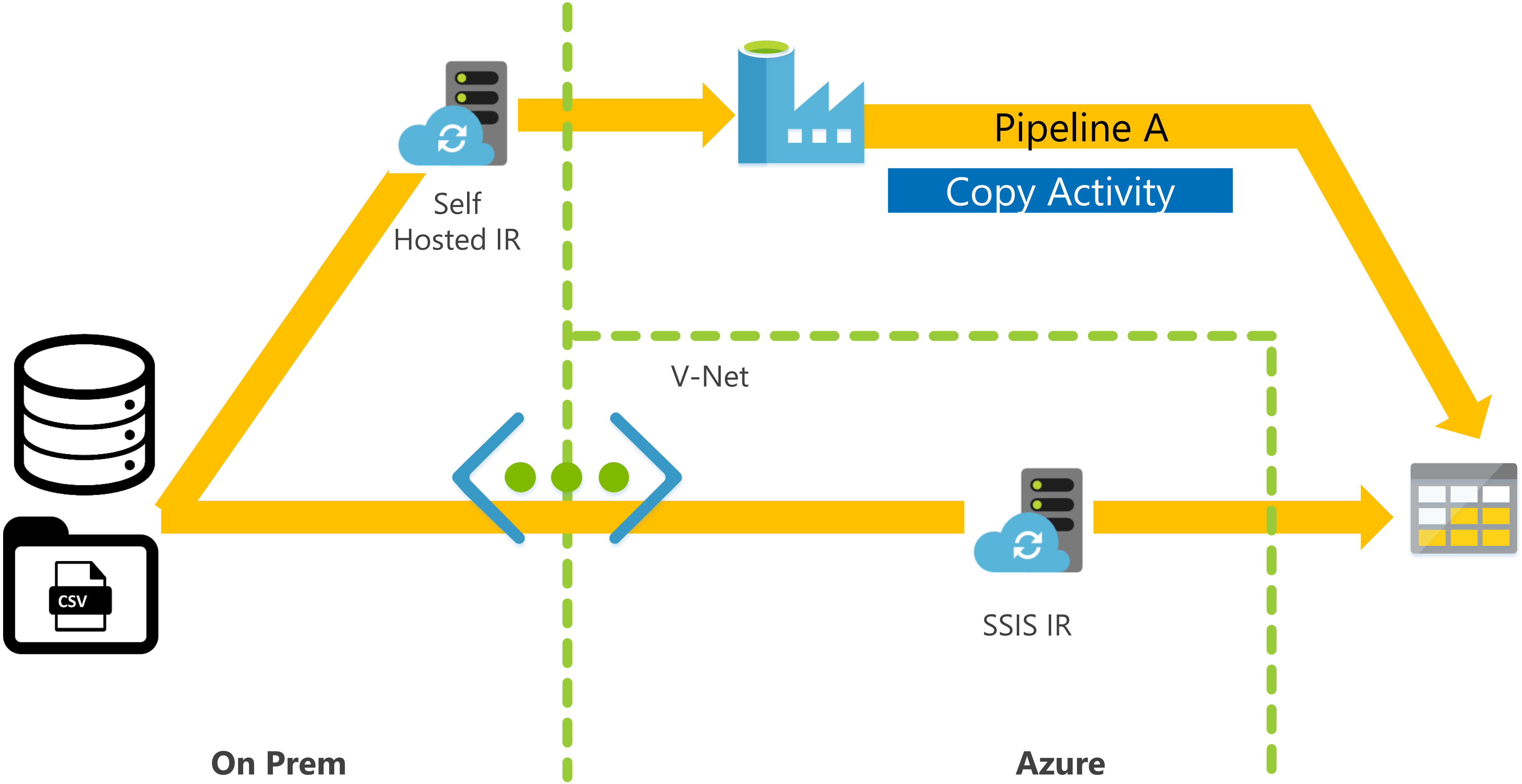
Pipeline Triggers



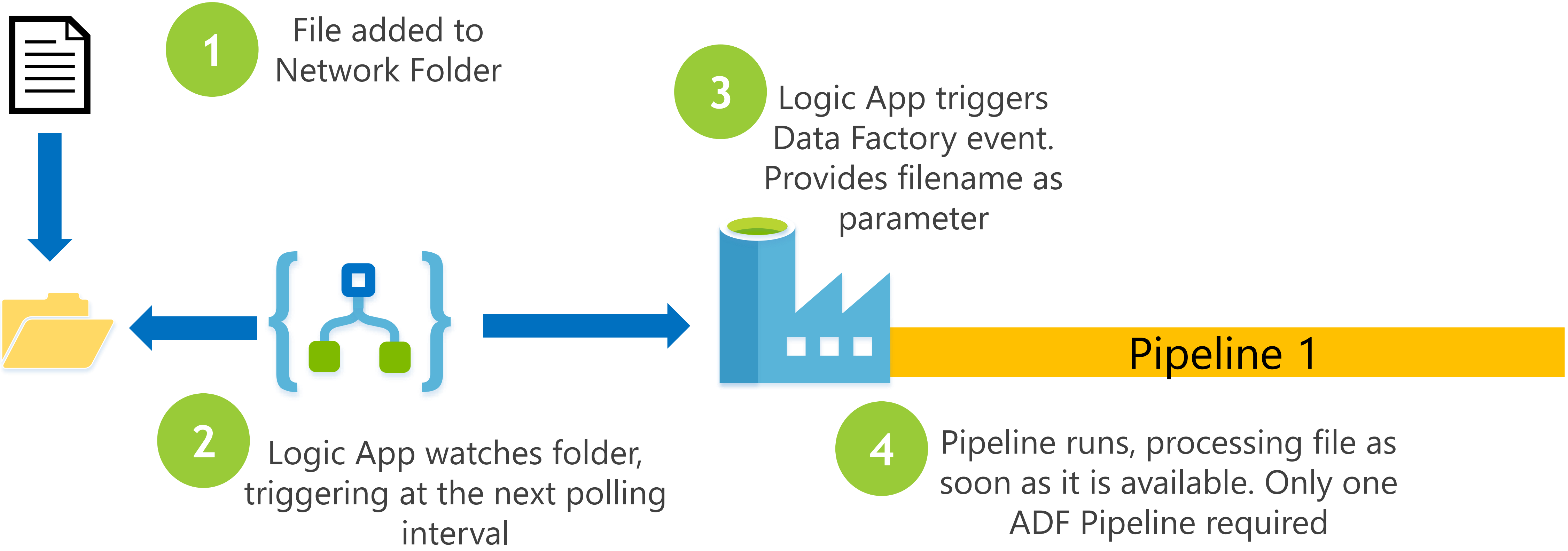
SSIS Integration Runtimes



The SSIS IR with Azure V-Net Access



Event-Triggered Loading (The New ETL?)



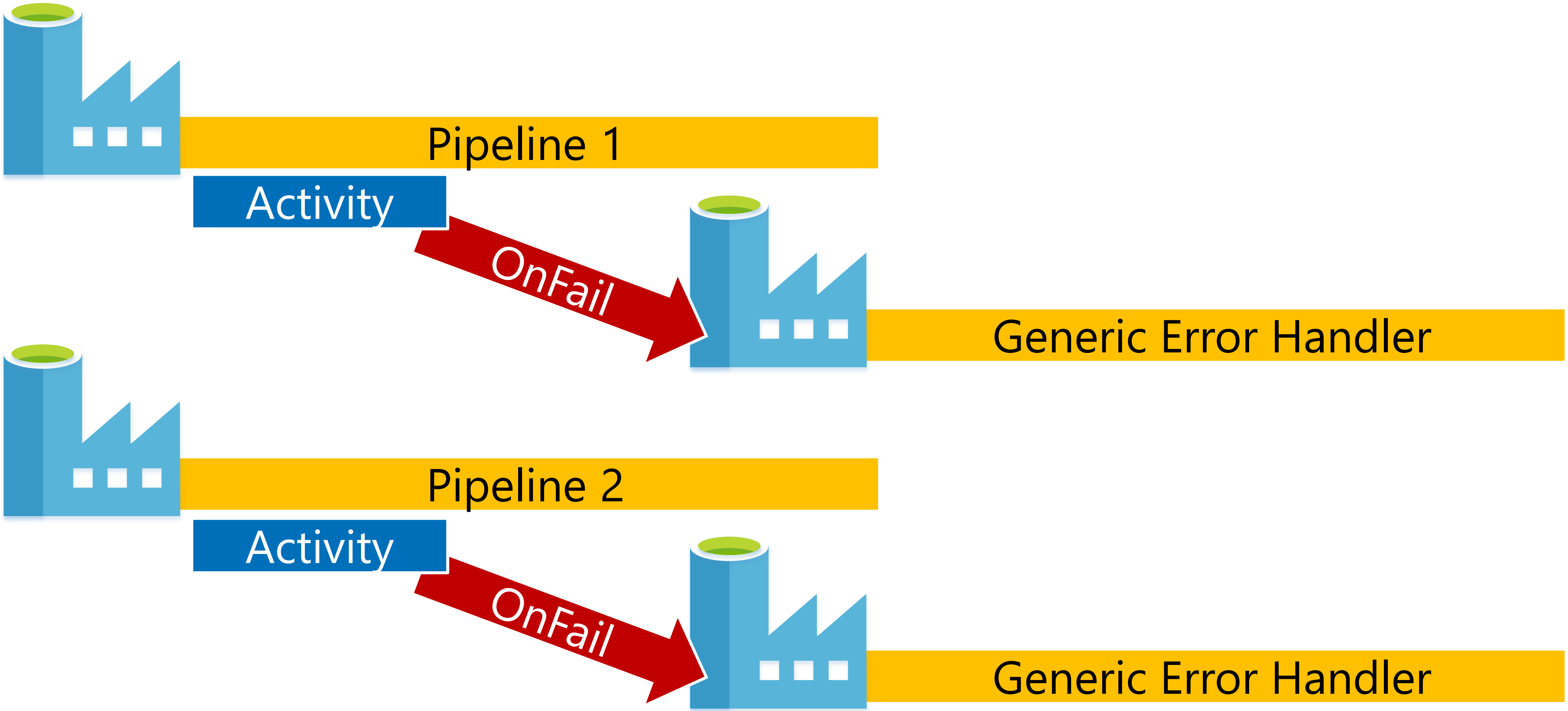
1 File added to Network Folder

2 Logic App watches folder, triggering at the next polling interval

3 Logic App triggers Data Factory event. Provides filename as parameter

4 Pipeline runs, processing file as soon as it is available. Only one ADF Pipeline required

Reusable Pipelines



Pipeline 1

Activity

OnFail

Generic Error Handler

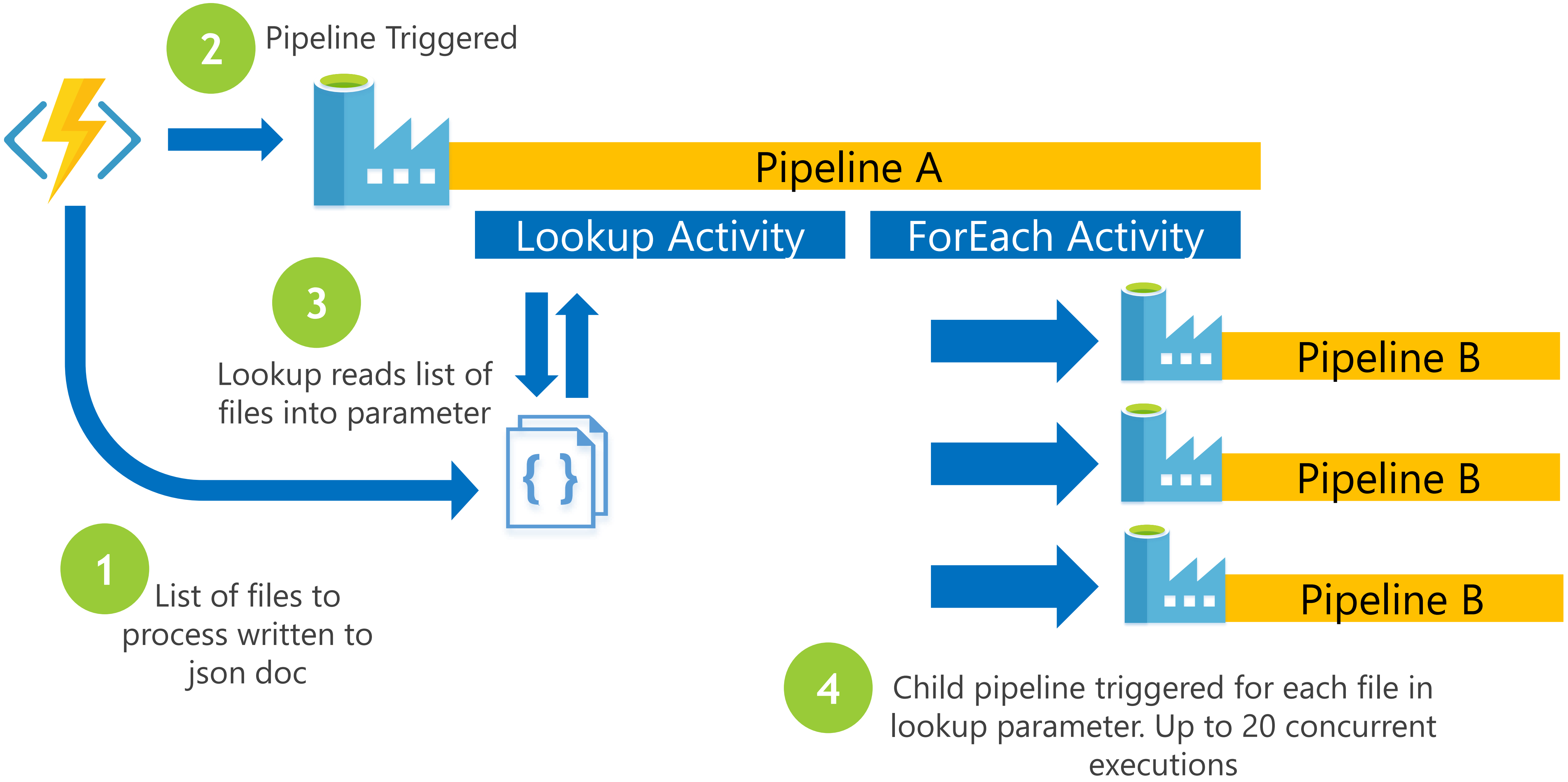
Pipeline 2

Activity

OnFail

Generic Error Handler

Batch Queuing



2 Pipeline Triggered

Pipeline A

Lookup Activity

ForEach Activity

3 Lookup reads list of files into parameter

Pipeline B

Pipeline B

Pipeline B

1 List of files to process written to json doc

4 Child pipeline triggered for each file in lookup parameter. Up to 20 concurrent executions

Thanks for Listening

Paul Andrew

@MrPaulAndrew

MVP Microsoft® Most Valuable Professional

adatis

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