

# Azure Data Factory v2

## Data Integration in the Cloud

Paul Andrew | Adatis

10/03/2018



Gold Data Analytics  
Gold Data Platform  
Gold Cloud Platform





## GOLD SPONSORS



## SILVER SPONSORS



## BRONZE SPONSOR



## STRATEGIC PARTNER





<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

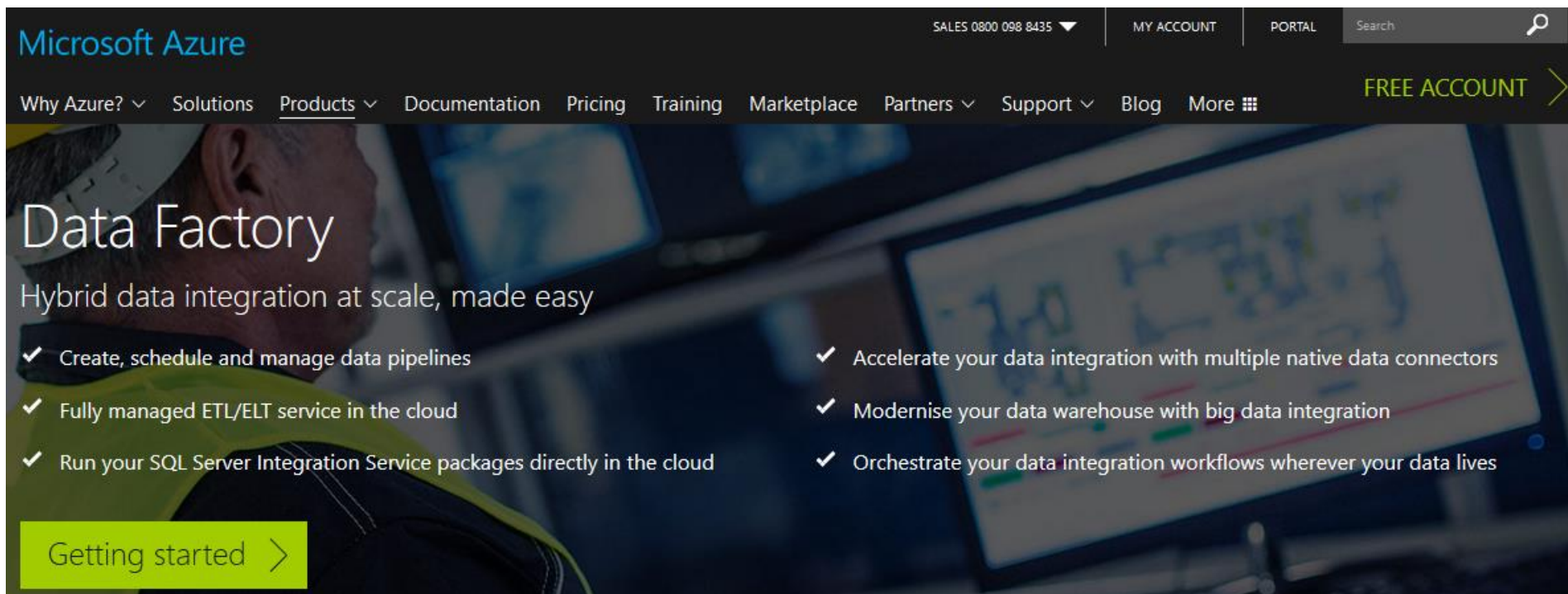
SSIS, Functions,  
Custom Activities

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 ☰

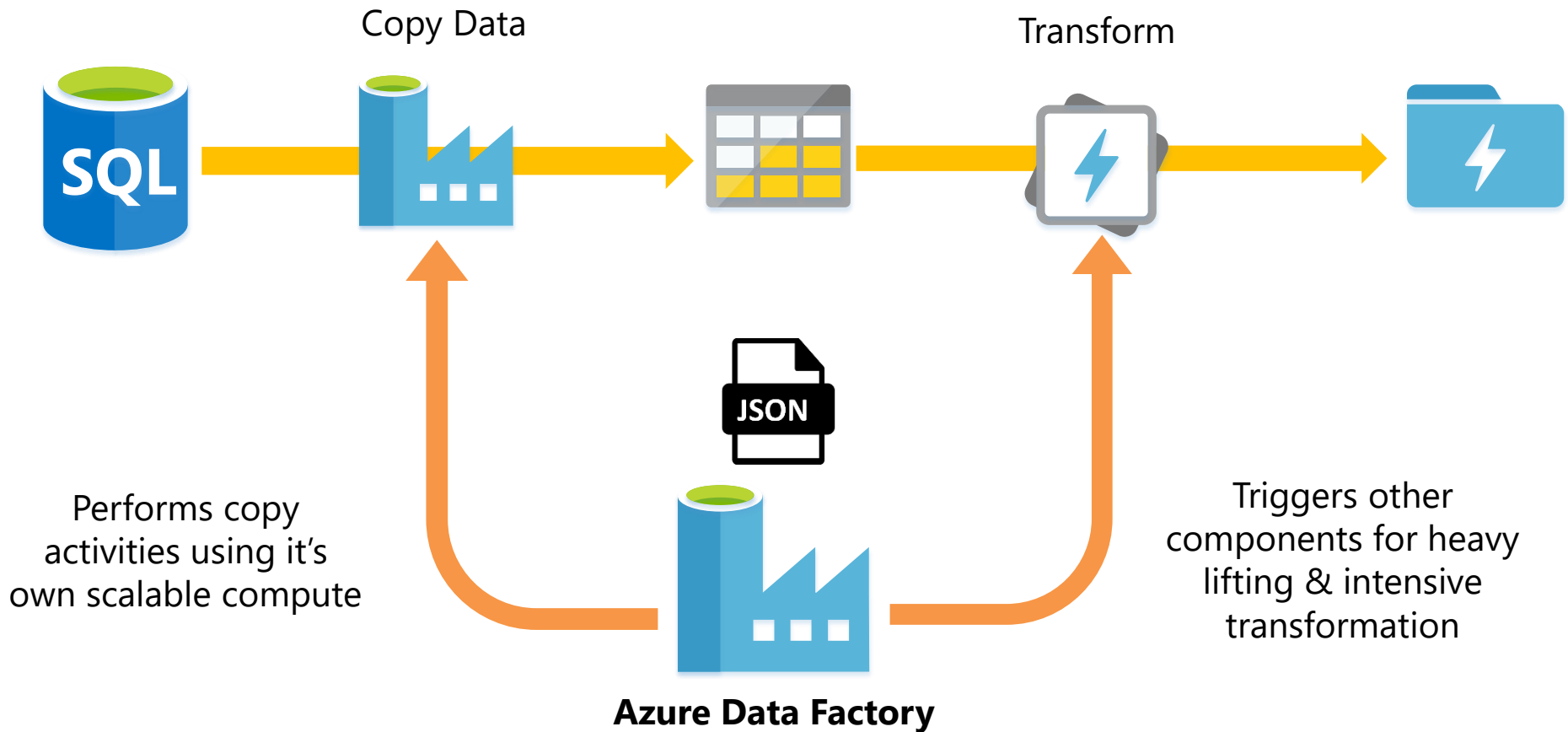
**Data Factory**

Hybrid data integration at scale, made easy

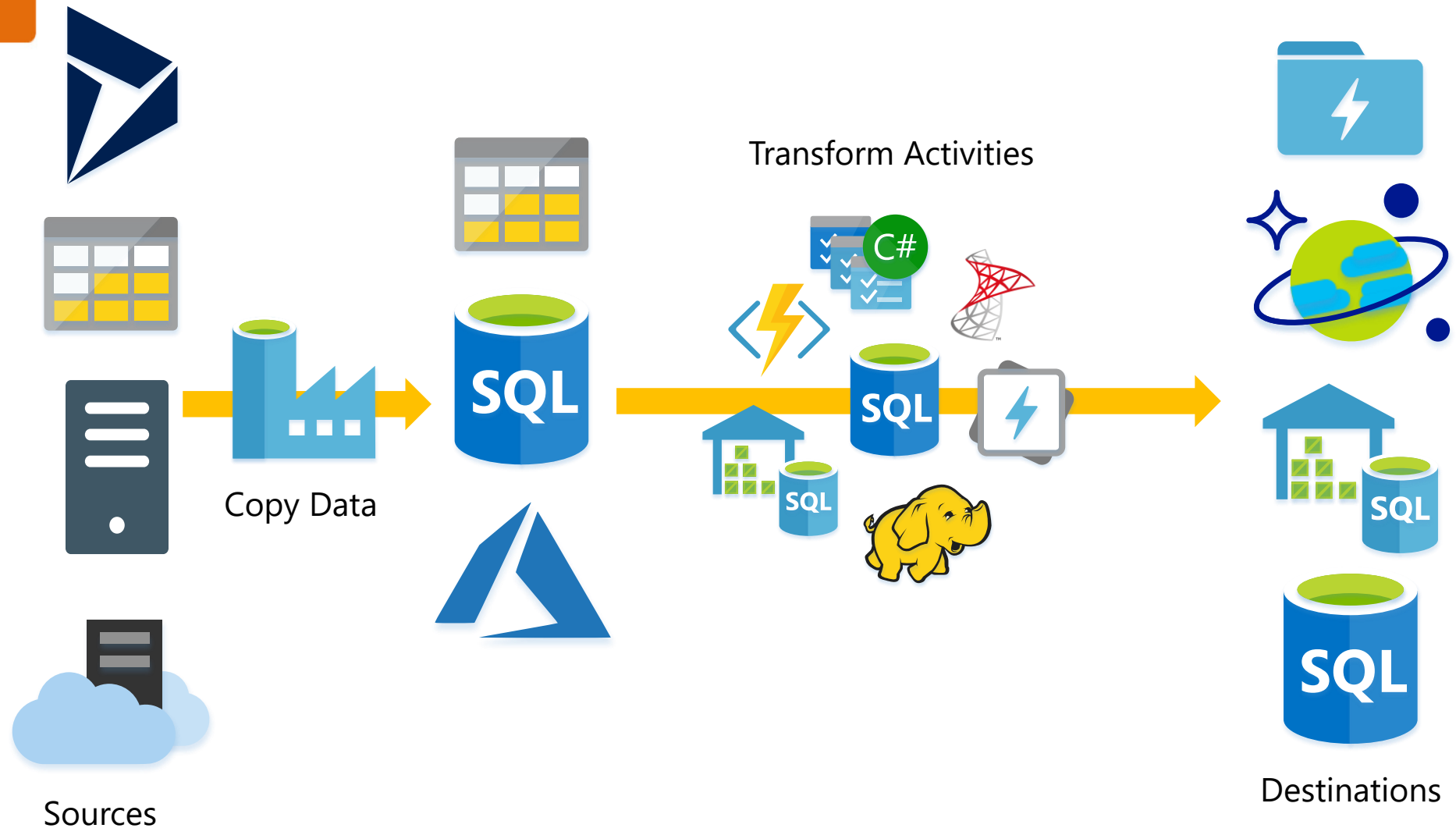
- ✓ Create, schedule and manage data pipelines
- ✓ Fully managed ETL/ELT service in the cloud
- ✓ Run your SQL Server Integration Service packages directly in the cloud
- ✓ Accelerate your data integration with multiple native data connectors
- ✓ Modernise your data warehouse with big data integration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

Getting started >

# What is Azure Data Factory?

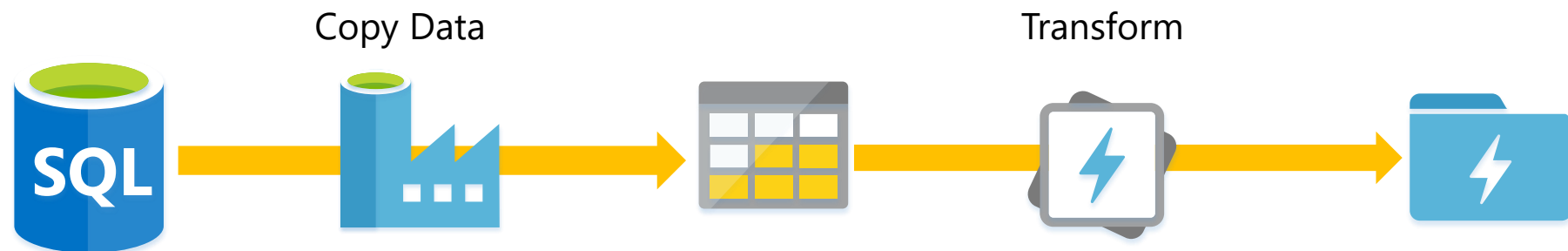


# What does Azure Data Factory do?





# Data Factory Components

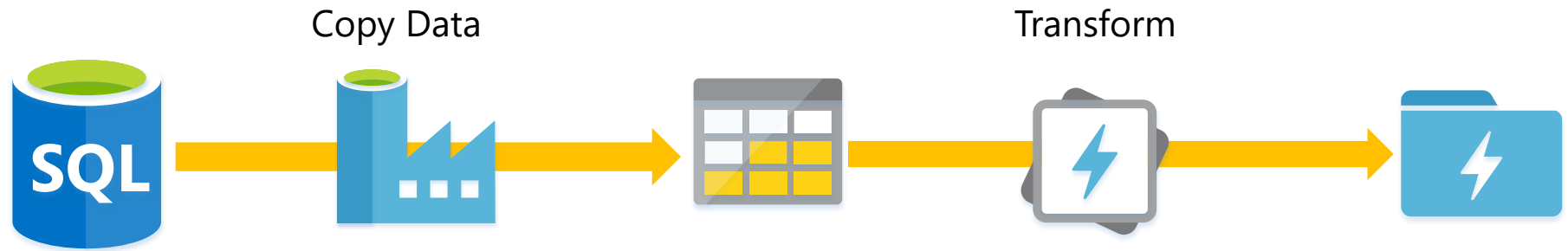


## 1 Linked Services – How do I connect?

Like the SSIS Connection Manager!



# Data Factory Components



## 1 Linked Services

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

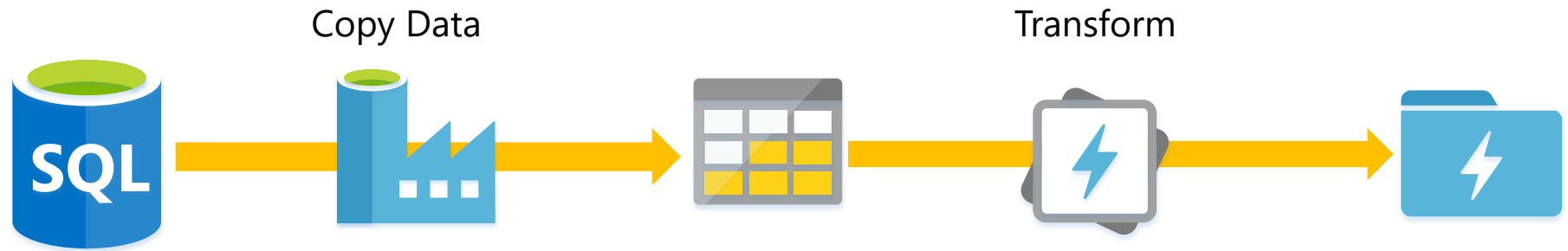


dbo.DimCustomer



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

# Data Factory Components



1 **Linked Services**

2 **Data Sets**

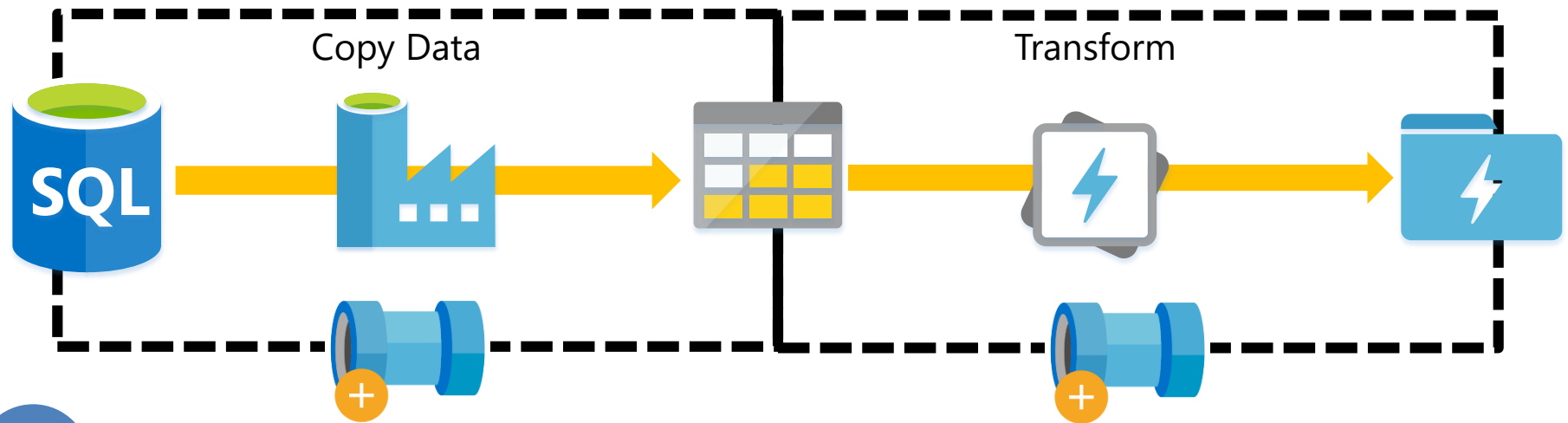
3 **Activities** –  
What do we  
want to happen?  
With what  
conditions?



## U-SQL Activity

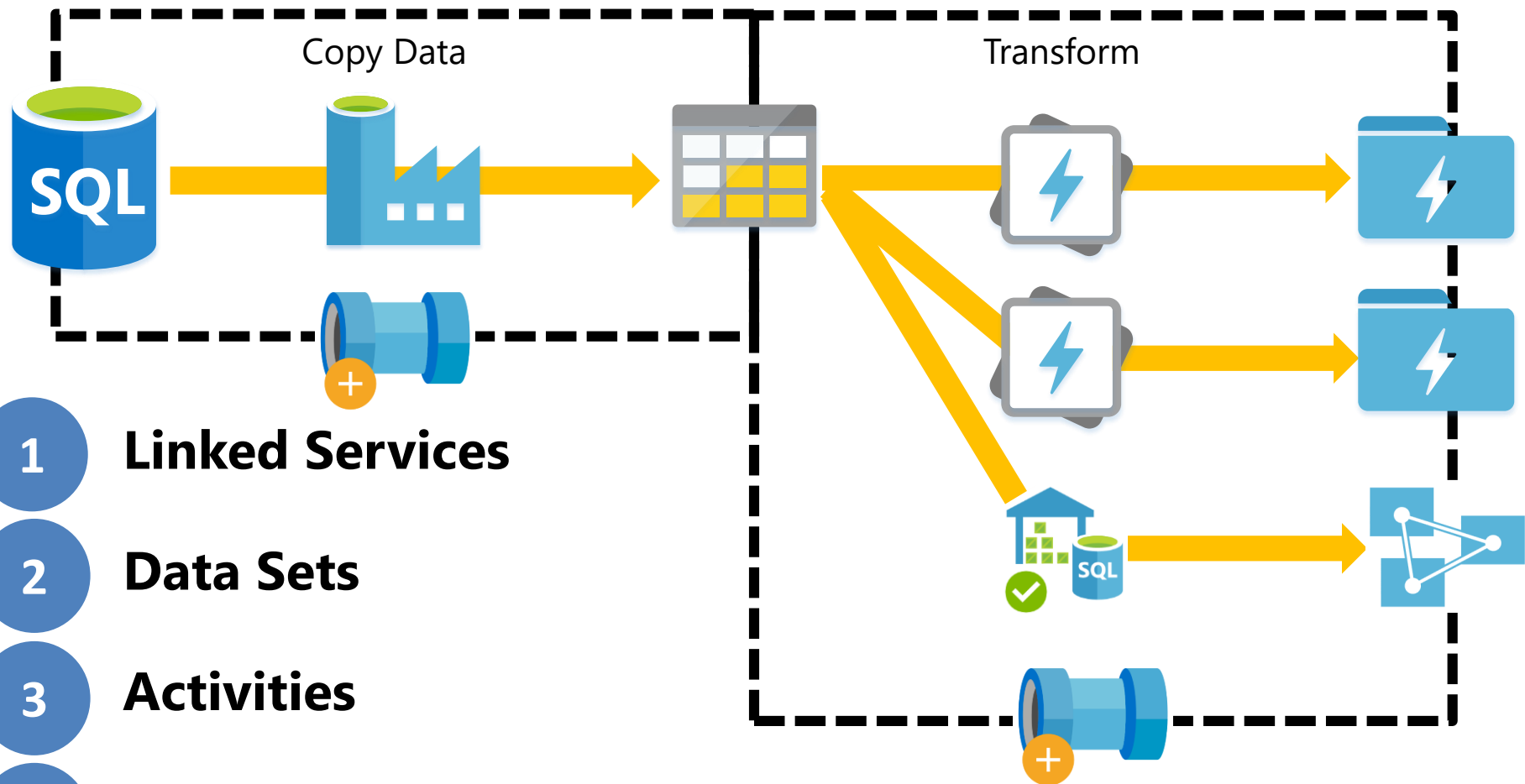
Script: *wasb//:myscripts/ProcessOrders.usql*  
AUs: *5 units*  
Priority: *1000*  
Parameters: *@Output = "RAW/Orders/..."*

# 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

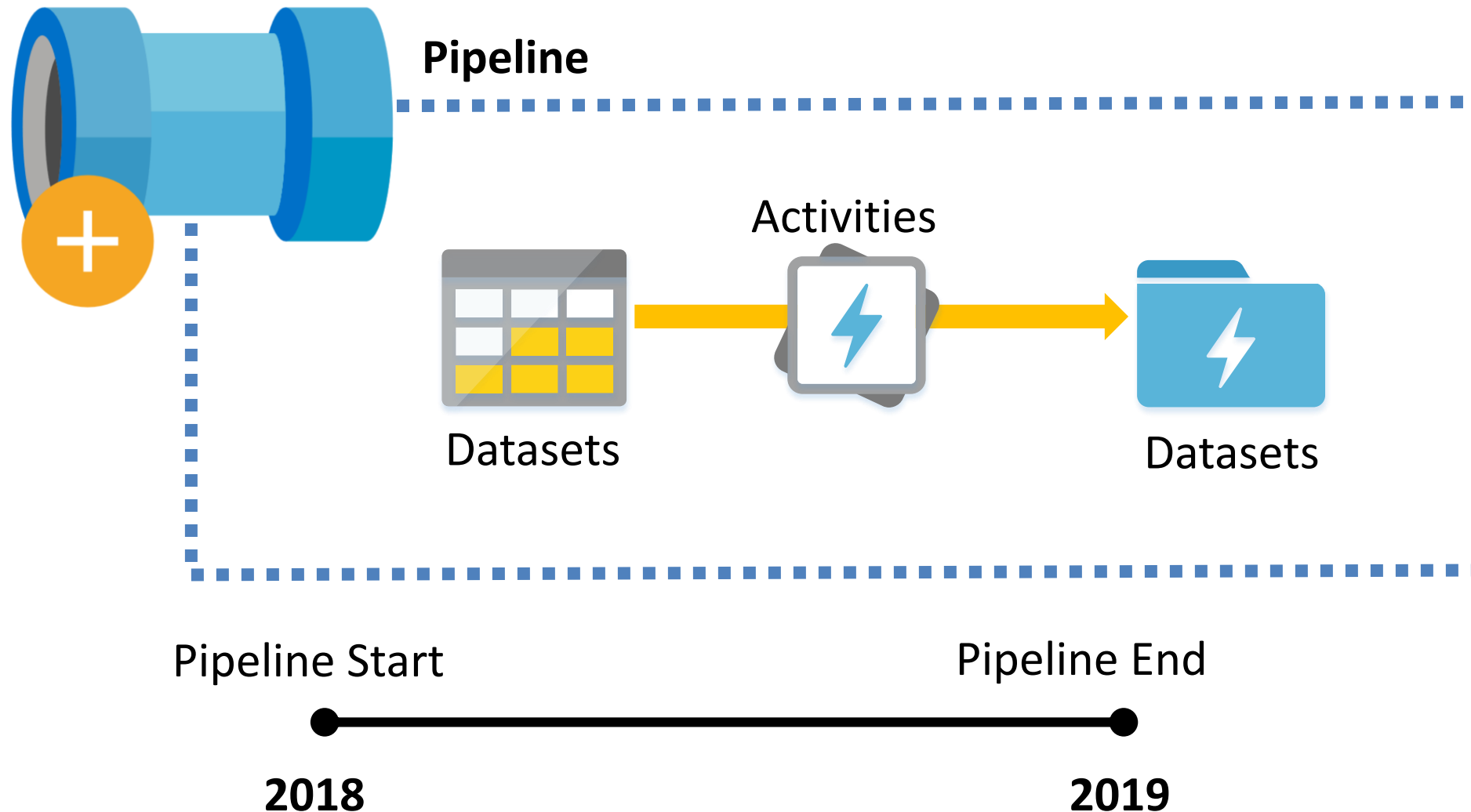
**Activities**

4

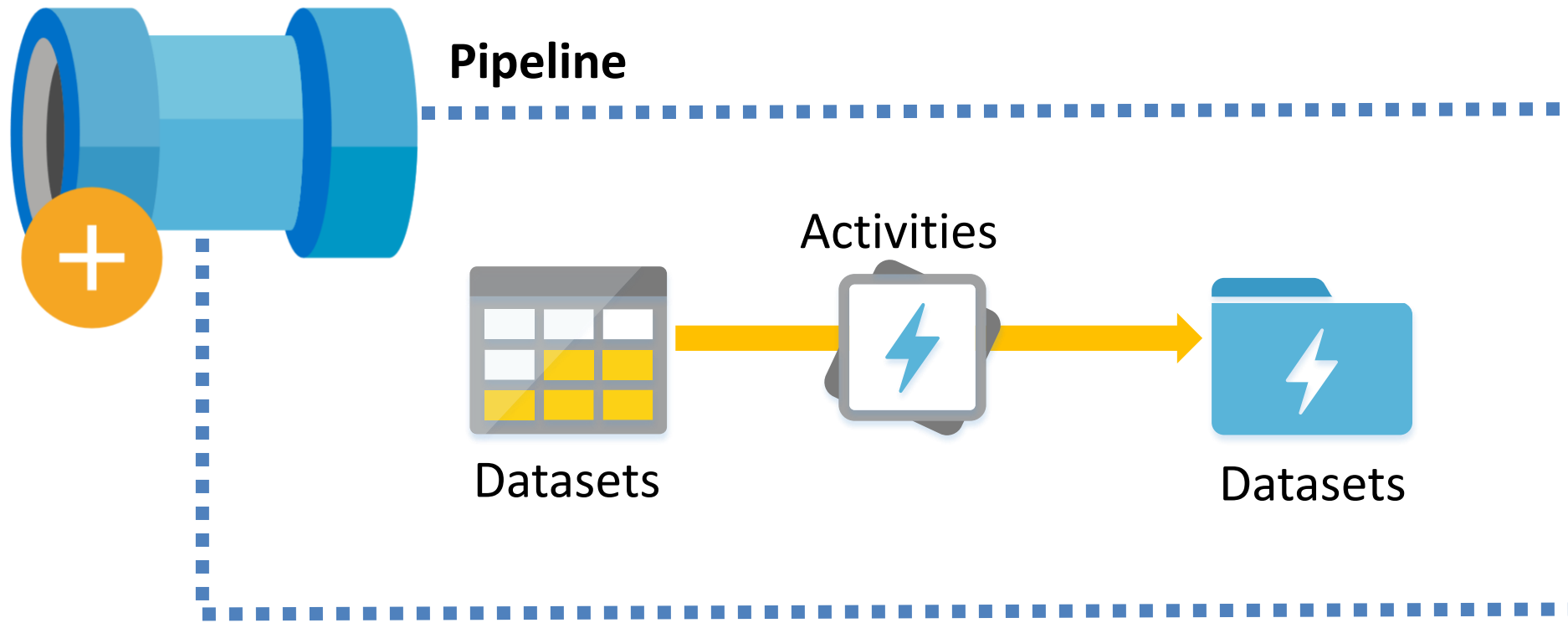
**Pipelines** – What groups of work do I want to do?

# Azure Data Factory Concepts

**Time Slices** – triggering an activity execution.

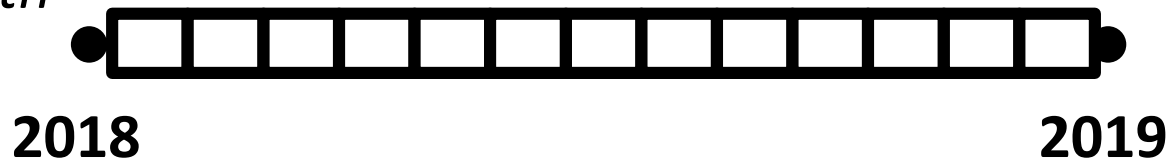


# Azure Data Factory Concepts Continued

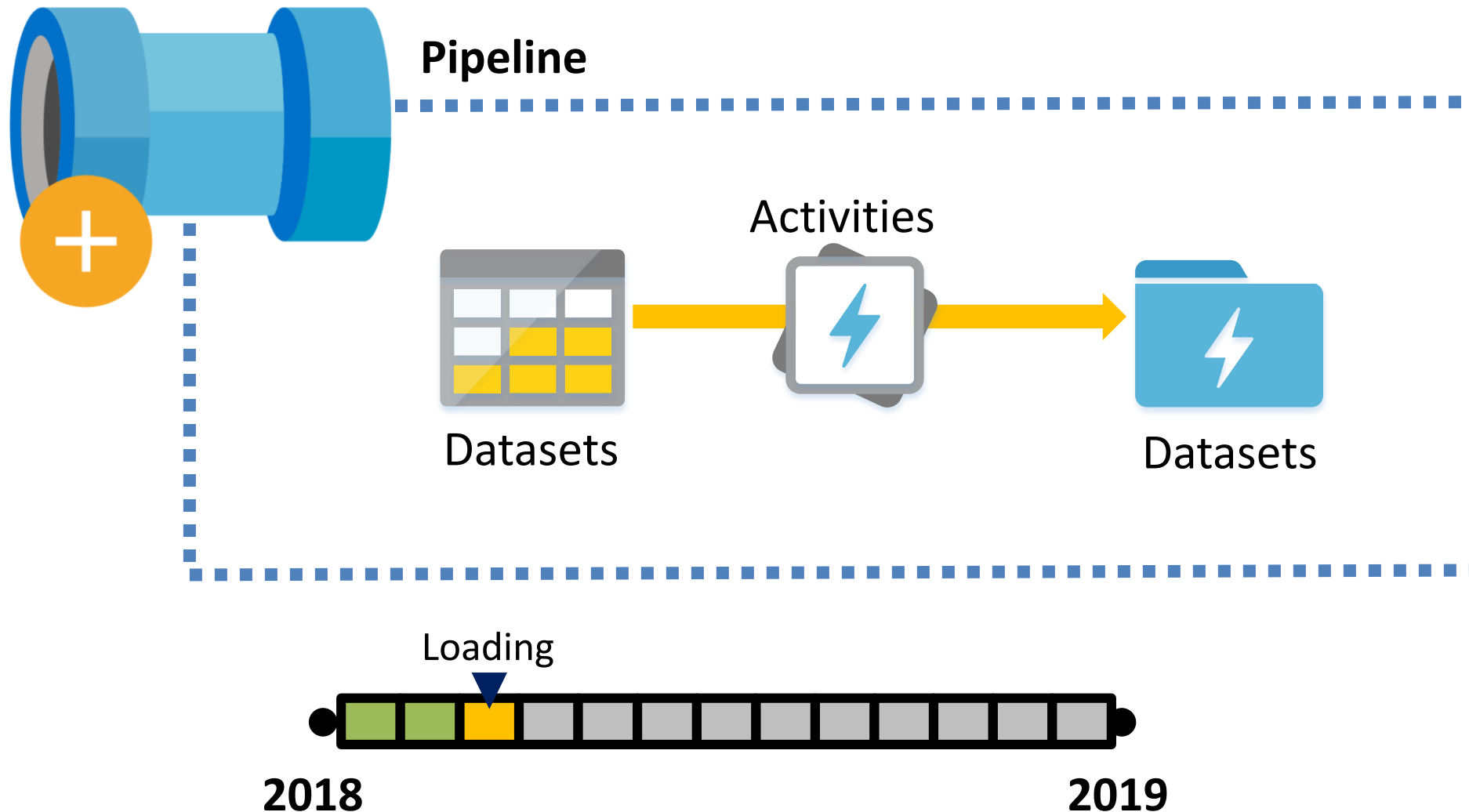


Interval: *Month*

Frequency: *1*

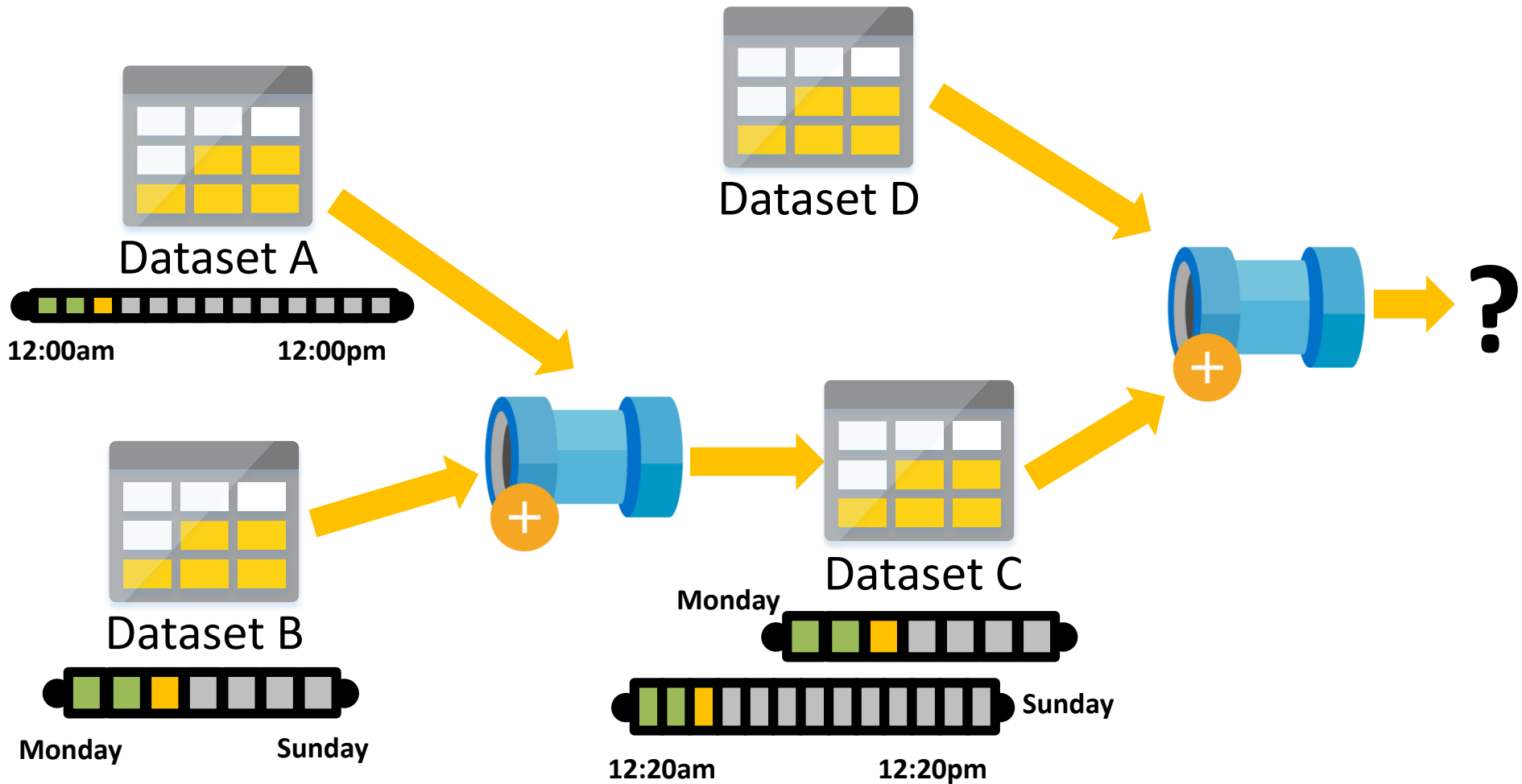


# Azure Data Factory Concepts Continued





# Time Slice Problems...



# Agenda

Data Factory  
Recap

Concepts  
Components

ADFv2

Features Update  
The Integration  
Runtime

Data Factory  
Extensibility

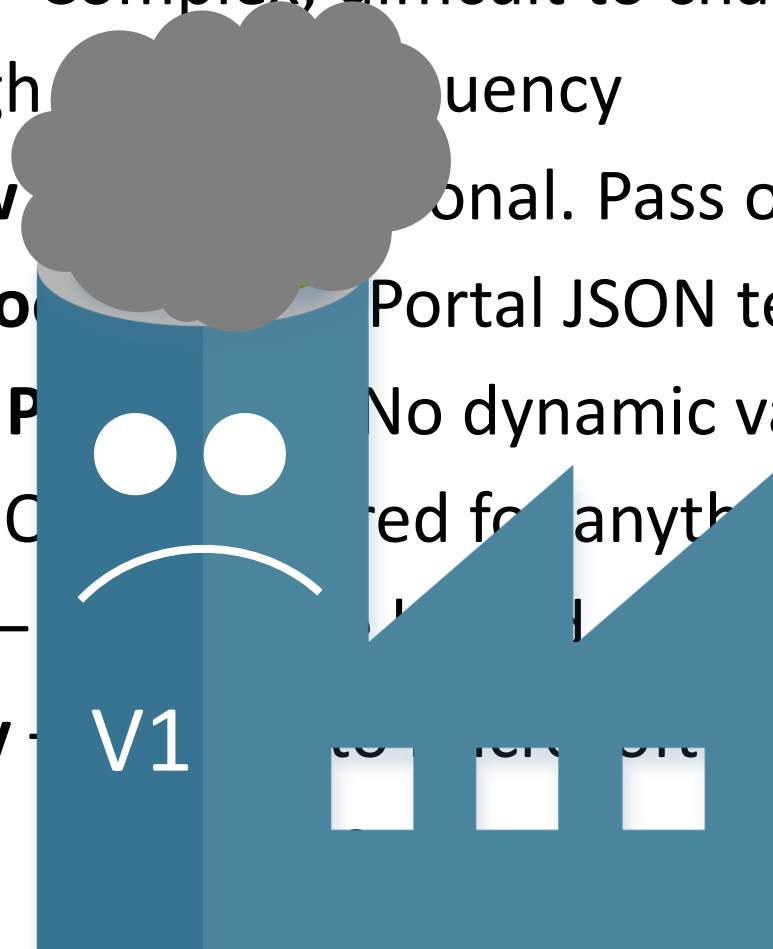
SSIS, Functions,  
Custom Activities

Conclusions

Design Patterns  
ETL/ELT in Azure

# Data Factory Issues & Limitations

- **Time Slices** – Complex, difficult to change & provision
- **Pricing** – High frequency
- **Control Flow** – Conditional. Pass or fail
- **Developer Tool** – Portal JSON templates
- **Hard Coded Parameters** – No dynamic values
- **C# Coding** – Created for anything complex
- **Monitoring** – Based on datasets
- **Connectivity** – Supported linked



# So What's Changed?

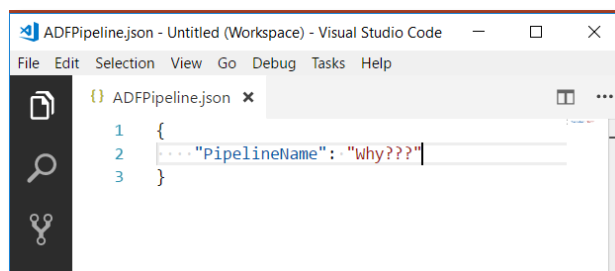


- |   |                              |   |
|---|------------------------------|---|
| ✓ | Data Movement (Copy)         | ✓ |
| ✓ | Activities                   | ✓ |
| ✓ | Pipelines                    | ✓ |
| ✓ | Datasets                     | ✓ |
| ✓ | Time Slices/Tumbling Windows | ✓ |
|   | Event Triggers               | ✓ |
| ✗ | Recurring Schedules          | ✓ |
| ✗ | Parameters                   | ✓ |
| ✗ | Expressions                  | ✓ |
| ✗ | Conditional Logic            | ✓ |
| ✗ | Use of SSIS Packages         | ✓ |
| ✗ | Graphic Developer Canvas     | ✓ |
| ✗ | Drilldown Monitoring         | ✓ |

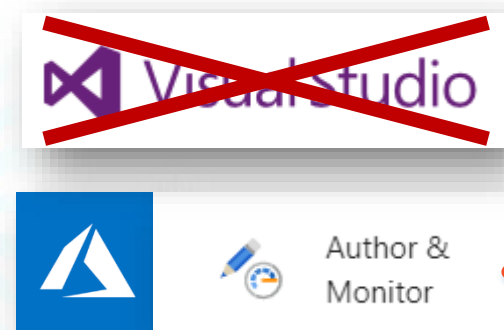
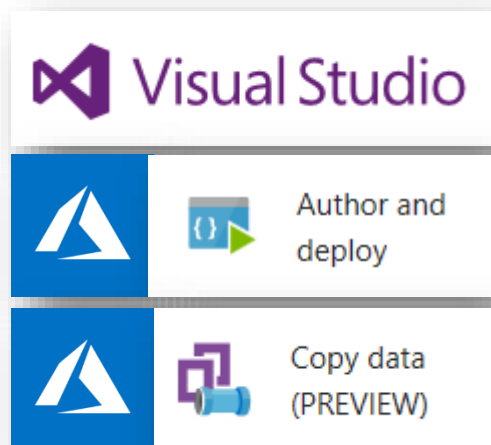


# Developer Tools

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



- ▶ 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



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
RunSSISPackage		01/15/2018, 1:36:42 PM	00:00:11	Manual trigger	✅ Succeeded

Data factory

RESOURCE EXPLORER

Data Factories

PaulsFunFactoryV1

Pipelines

FileCleaning

UploadFileToADLStore

Datasets

FakeOrdersClean

FakeOrdersLanding

FakeOrdersSourceFile

Linked services

BatchCompute

BlobStore

DataLakeStore

LaptopGateway

USQLEngine

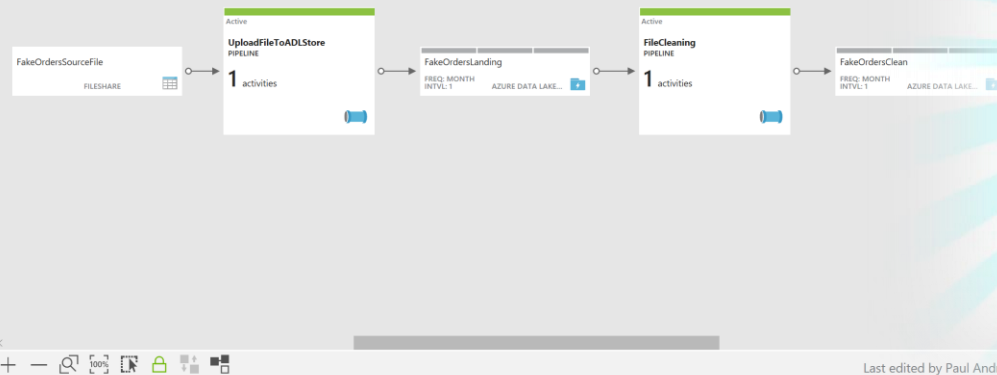
Gateways

PaulsLappy

jhlhkhj

PaulsFunFactoryV1

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



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.

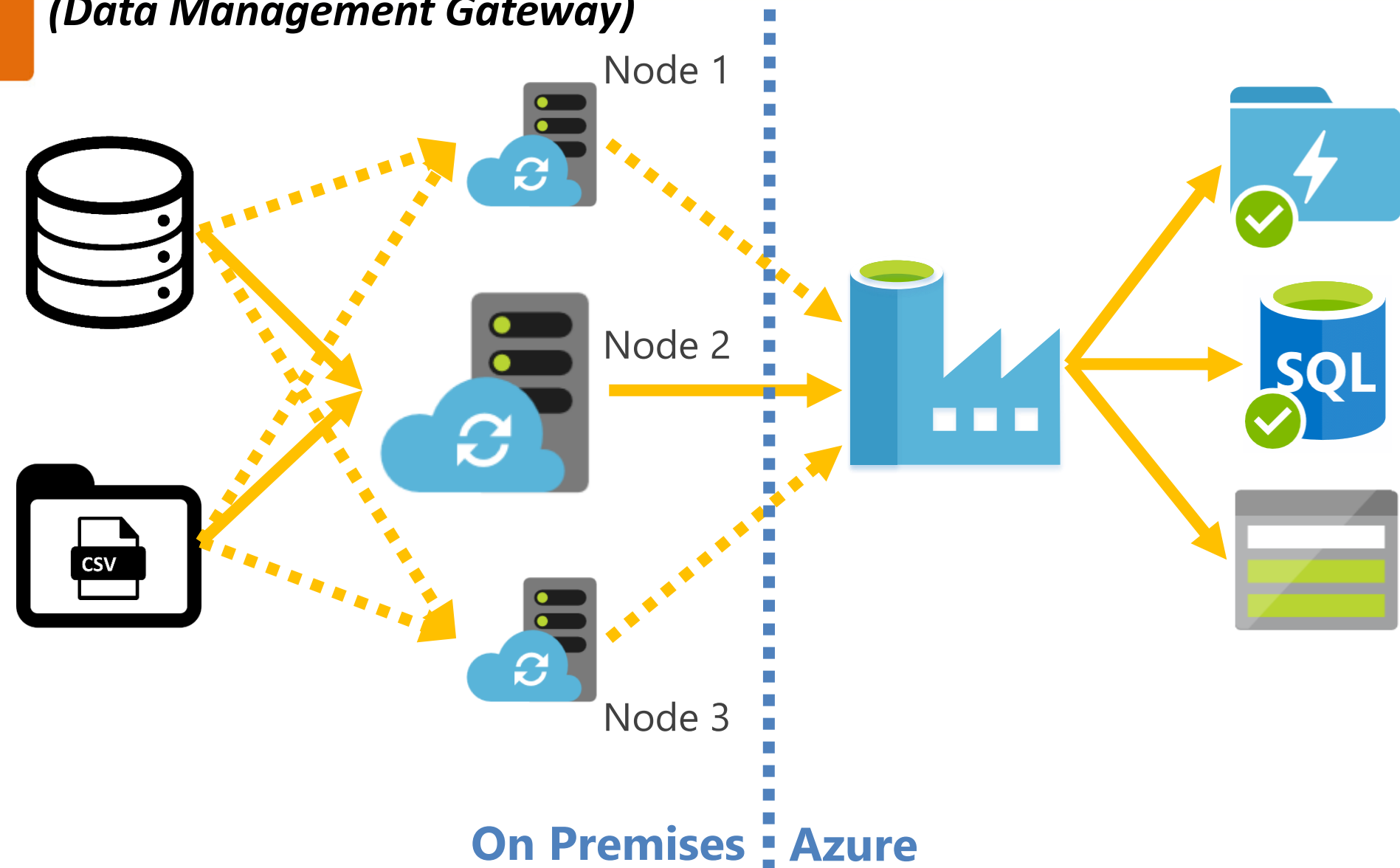


PaulAndrew



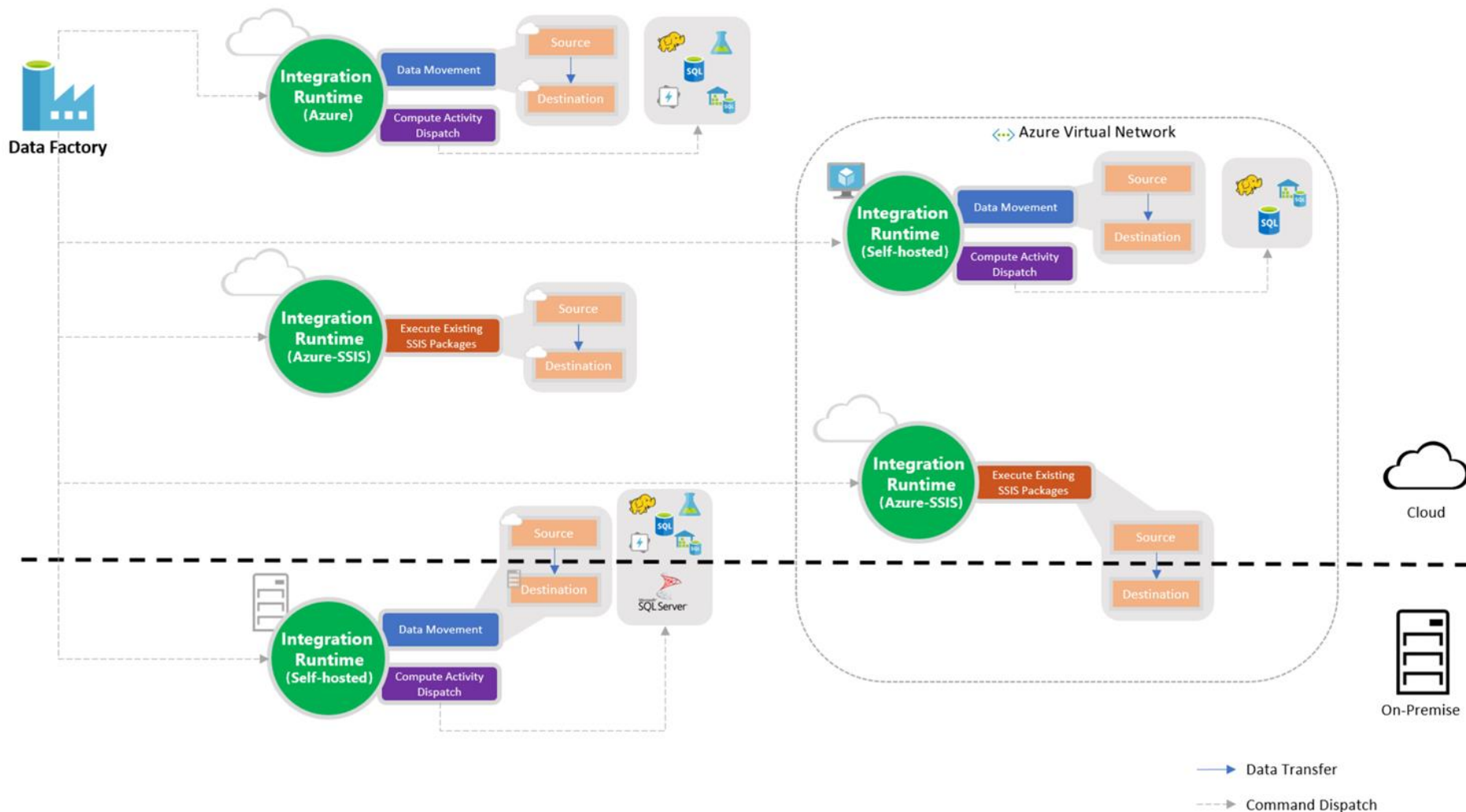
# The Integration Runtime

*(Data Management Gateway)*



# The Integration Runtime

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime>





# Agenda

Data Factory  
Recap

Concepts  
Components

ADFv2

Features Update  
The Integration  
Runtime

Data Factory  
Extensibility

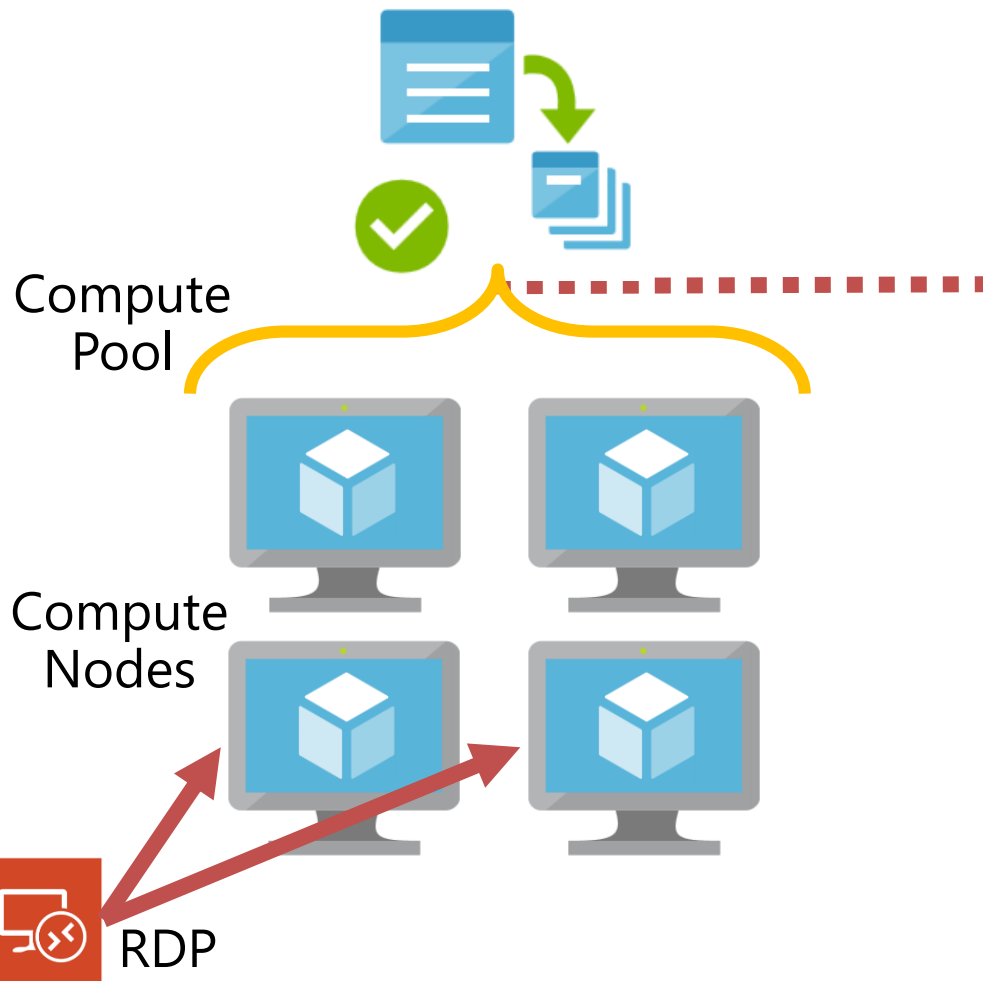
SSIS, Functions,  
Custom Activities

Conclusions

Design Patterns  
ETL/ELT in Azure

# Extending ADFv2

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



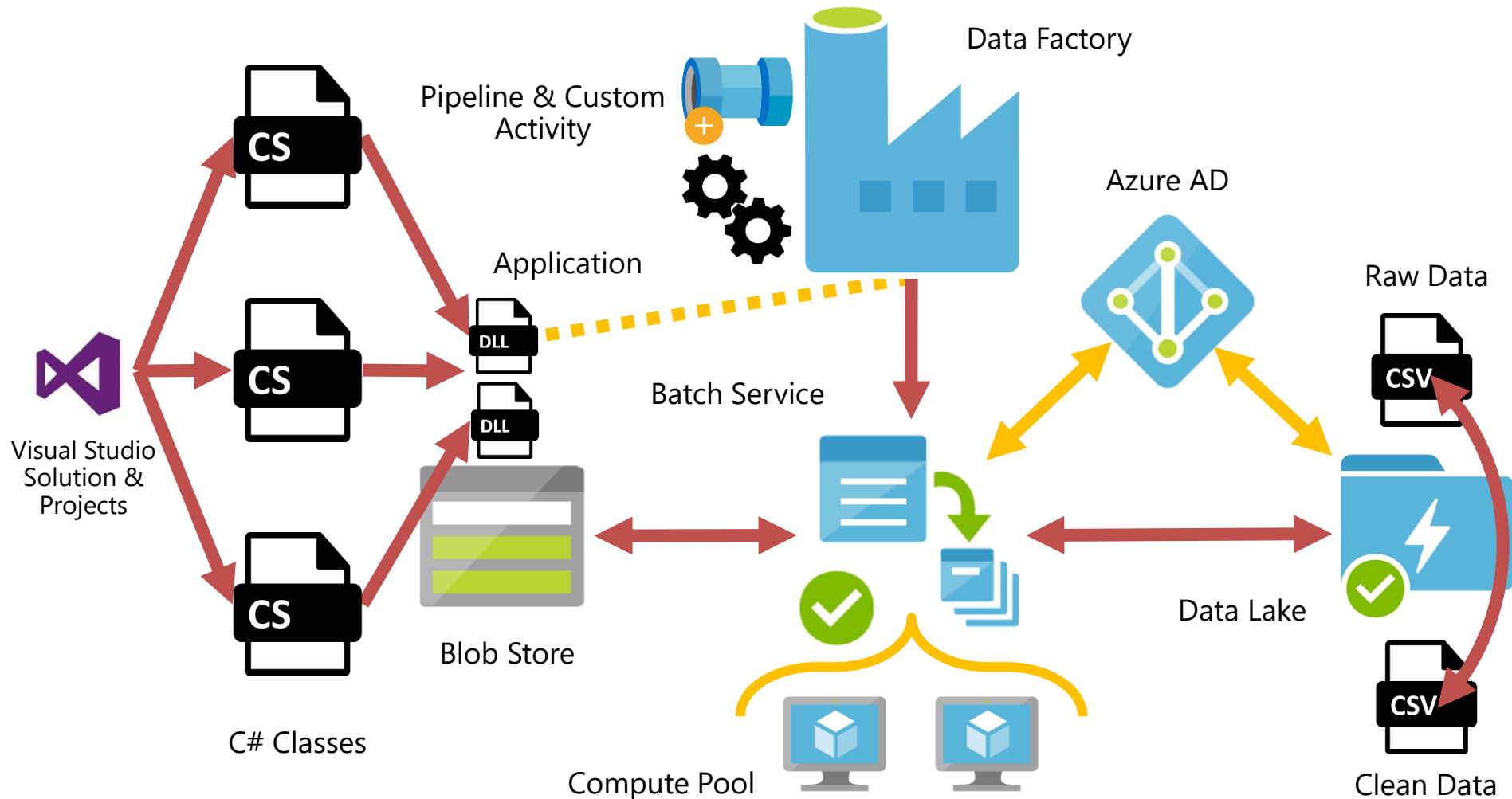
VM node size set per compute pool:

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.

# Extending ADFv2

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



# Extending ADFv2

## 1 Custom Activities

## 2 Rest API Calls – Eg. Web Activities Calling Azure Functions or Azure Logic Apps



General Settings<sup>2</sup> Parameters Advanced

Name \* Web1

Description

Timeout

Retry

Retry interval

URL \*

Method \* Select API method...

Headers

GET

POST

PUT

General Settings<sup>2</sup> Parameters Advanced

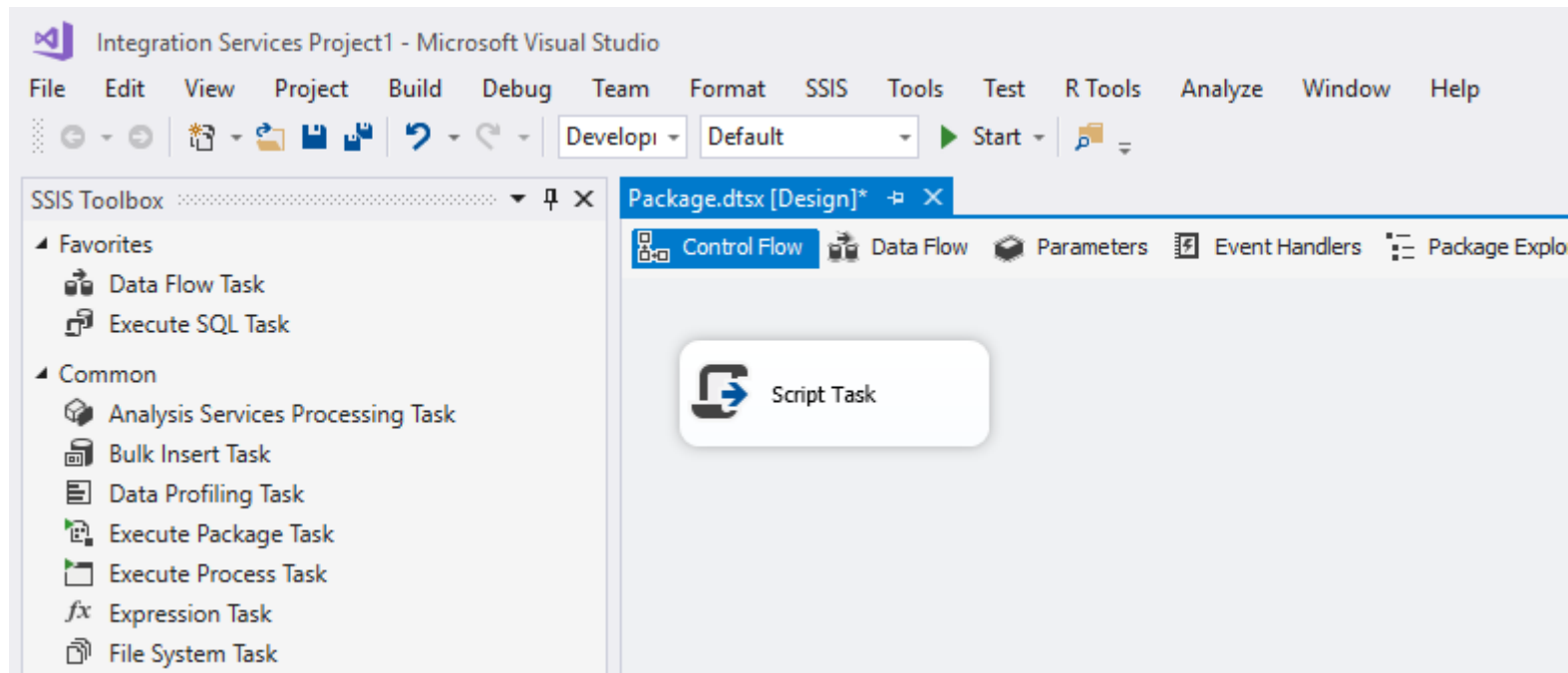
Use [expressions, functions](#) or refer to [system variables](#) in the 'value' column.

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"/>

# Extending ADFv2

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



# How do we schedule an SSIS Package in Azure?

Azure Data  
Factory v2



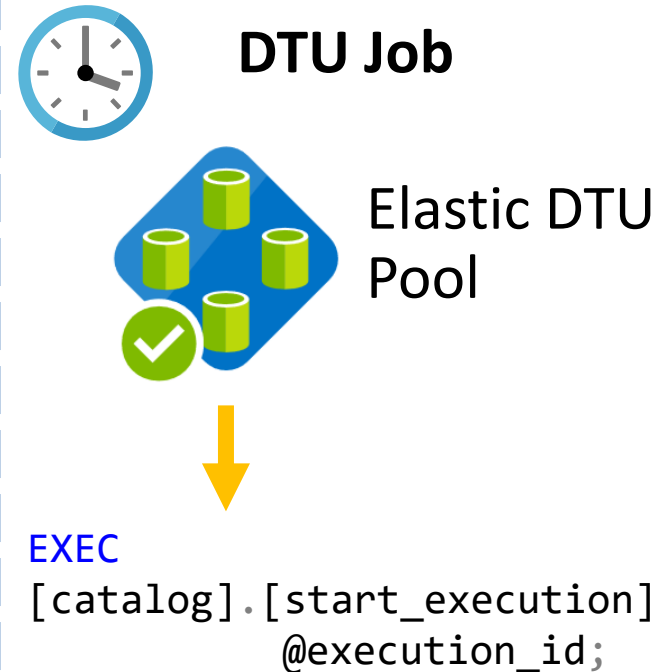
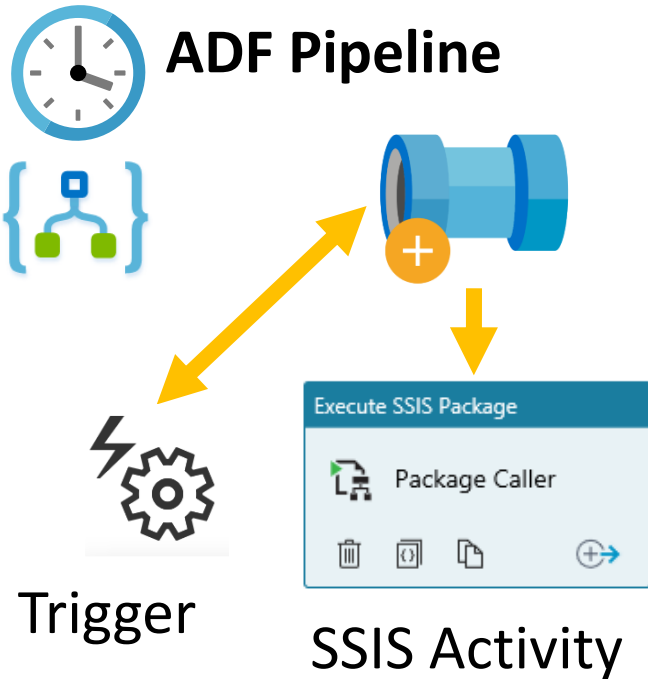
Azure SQL Logical or  
Managed Instance



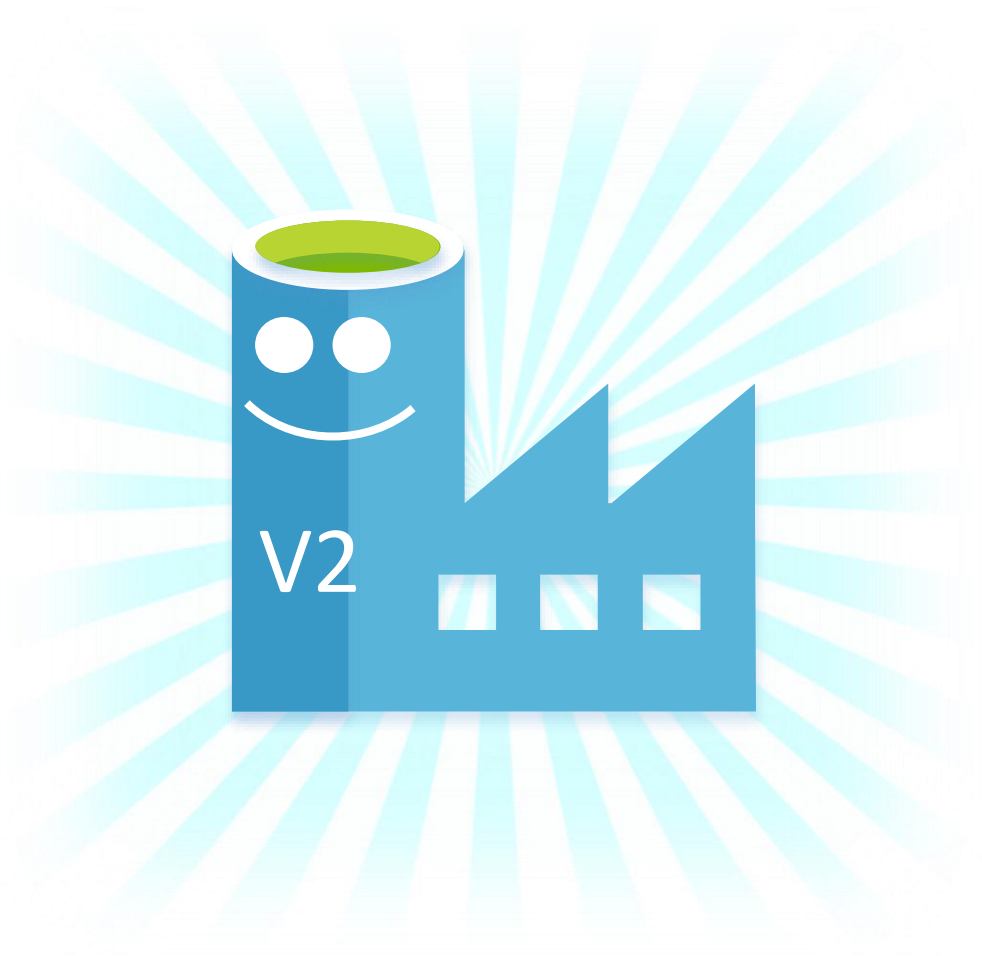
SSIS IR



Azure SQLDB  
(SSISDB)



# Demo



# Agenda

Data Factory  
Recap

Concepts  
Components

ADFv2

Features Update  
The Integration  
Runtime

Data Factory  
Extensibility

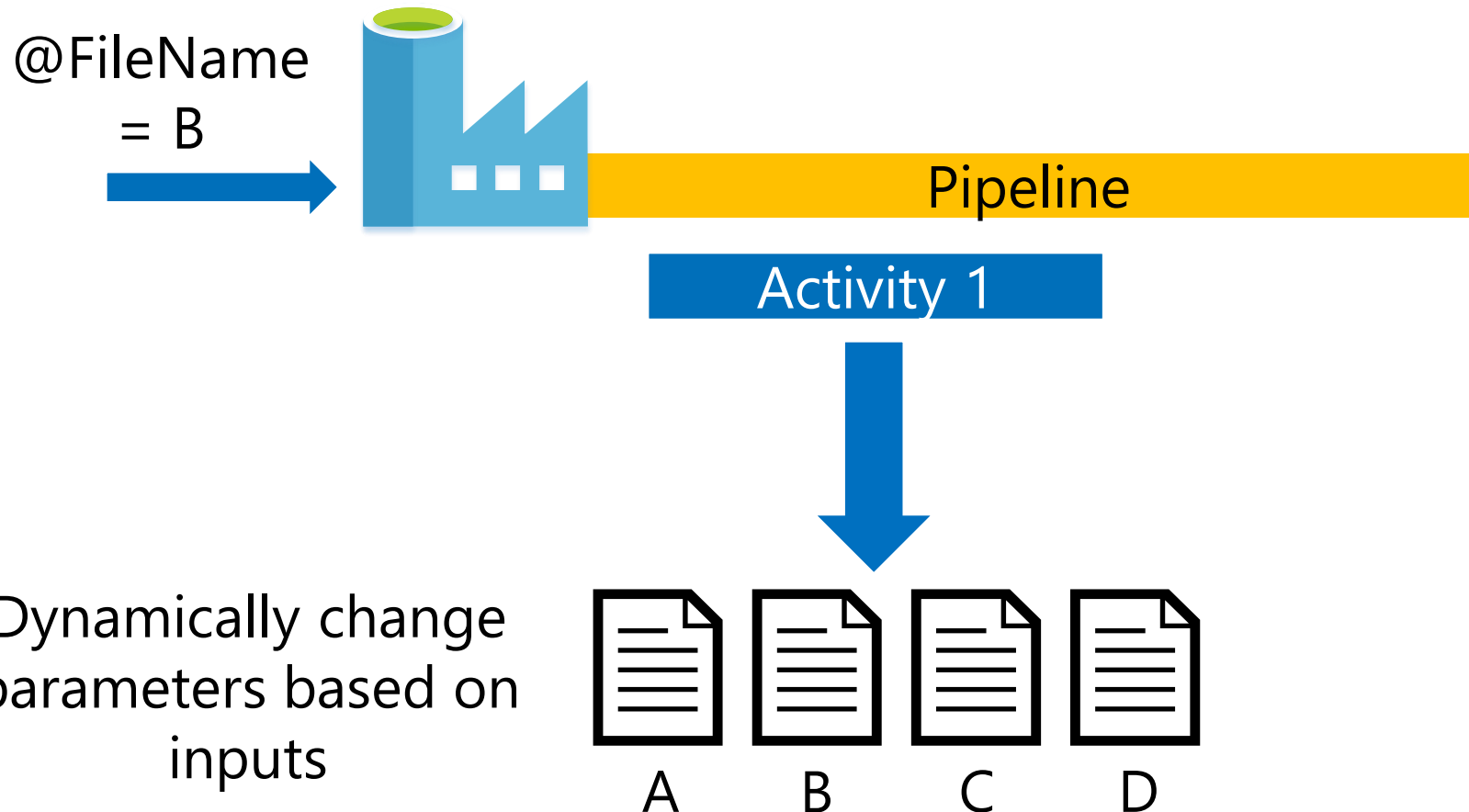
SSIS, Functions,  
Custom Activities

Conclusions

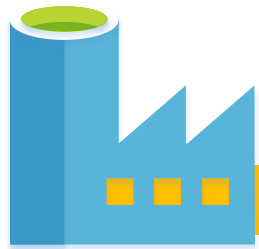
Design Patterns  
ETL/ELT in Azure



# Dynamic Pipelines using Parameters & Expressions



# Dynamic Pipelines using Lookup Activity



Pipeline

Lookup Activity



Returns  
@FileName = B



Config.json

Copy Activity



A



B



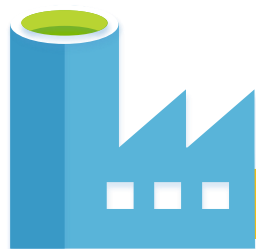
C



D

# For Each Pipelines

@Files =  
[A,B,C]



Pipeline

ForEach Activity

Run activity for each  
item in an array

(max DOP 20)



A

Copy Activity



B

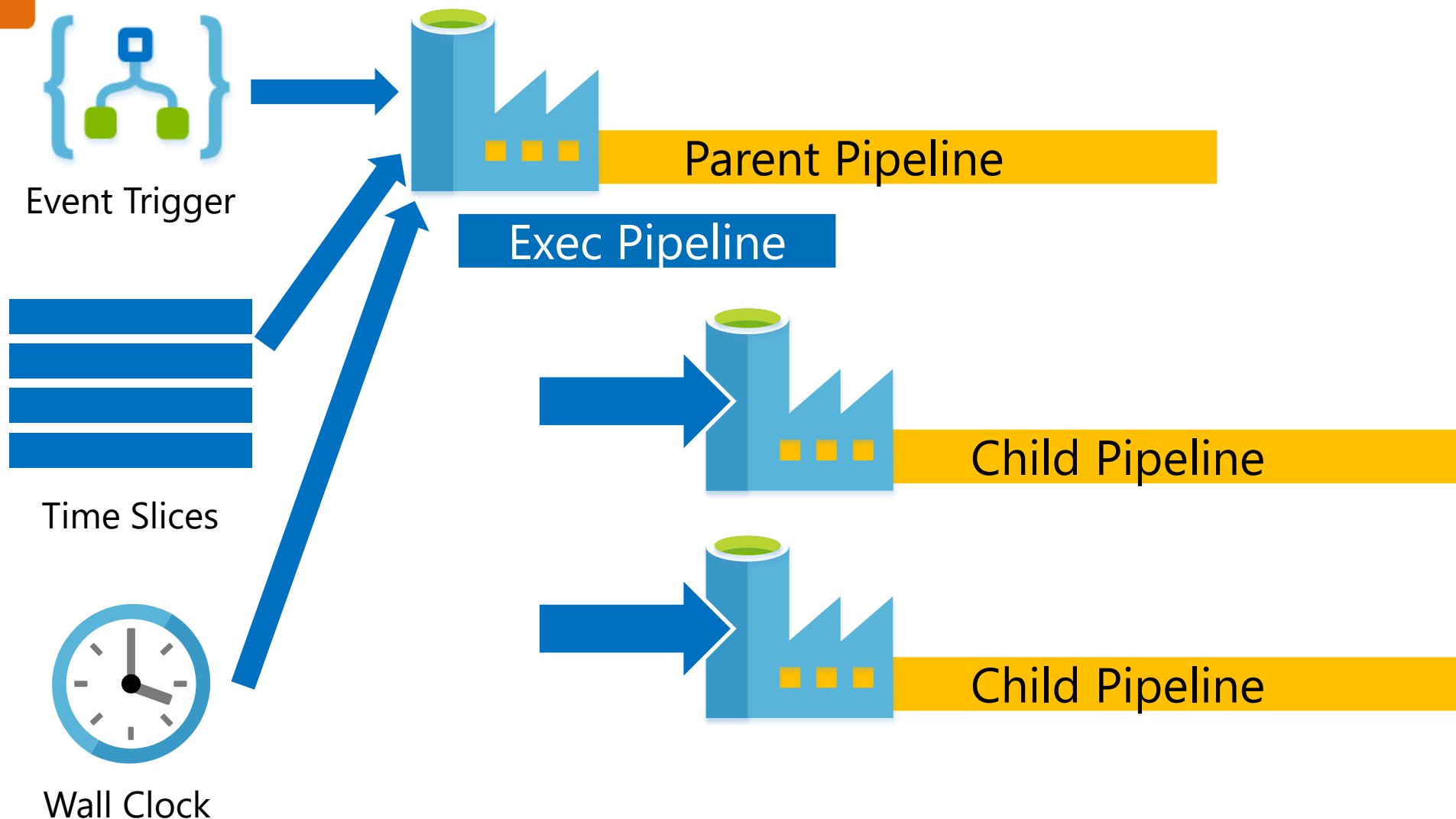
Copy Activity



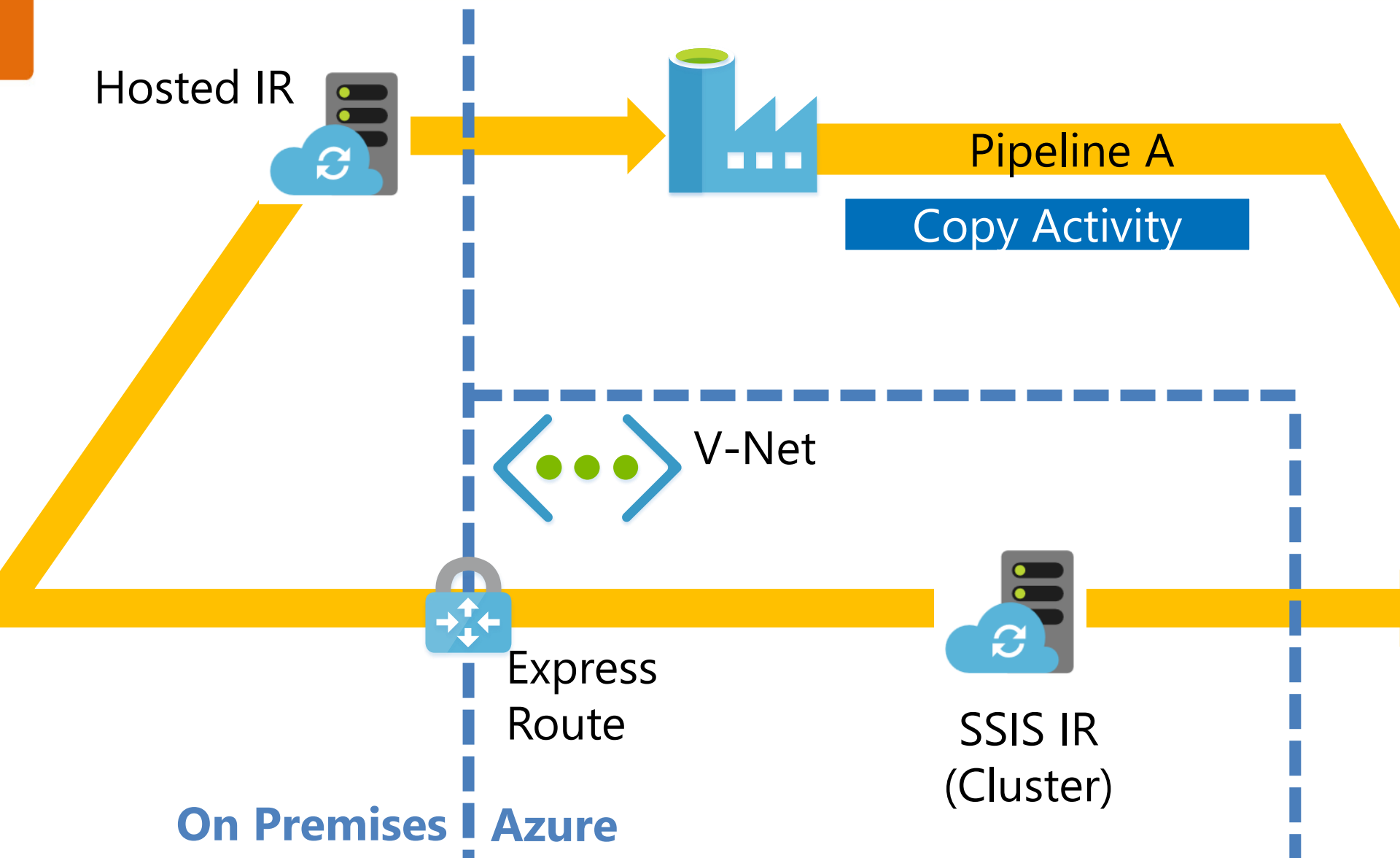
C

Copy Activity

# Parent/Child Pipelines & Triggering

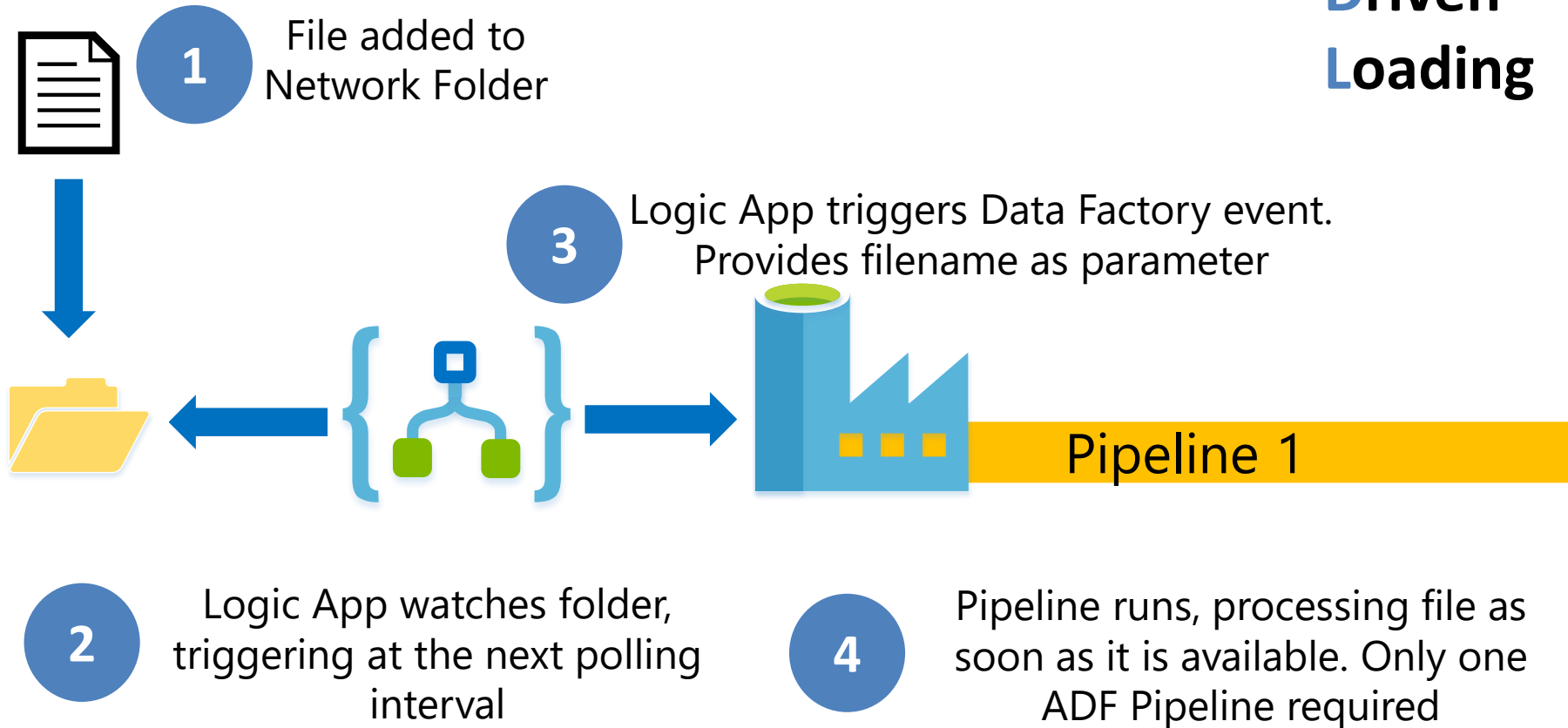


# SSIS IR with Azure V-Net Access

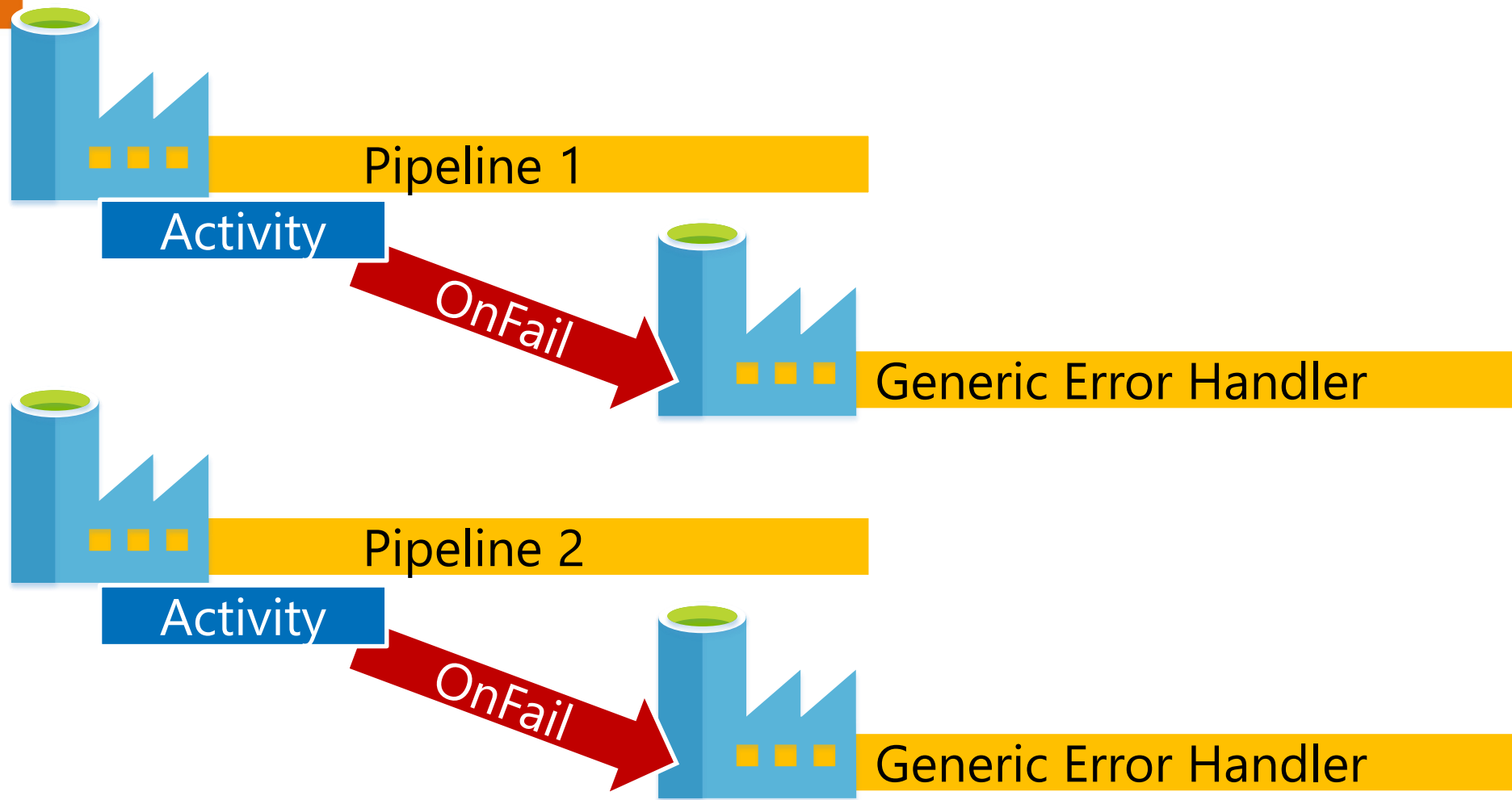


# Event Driven Loading

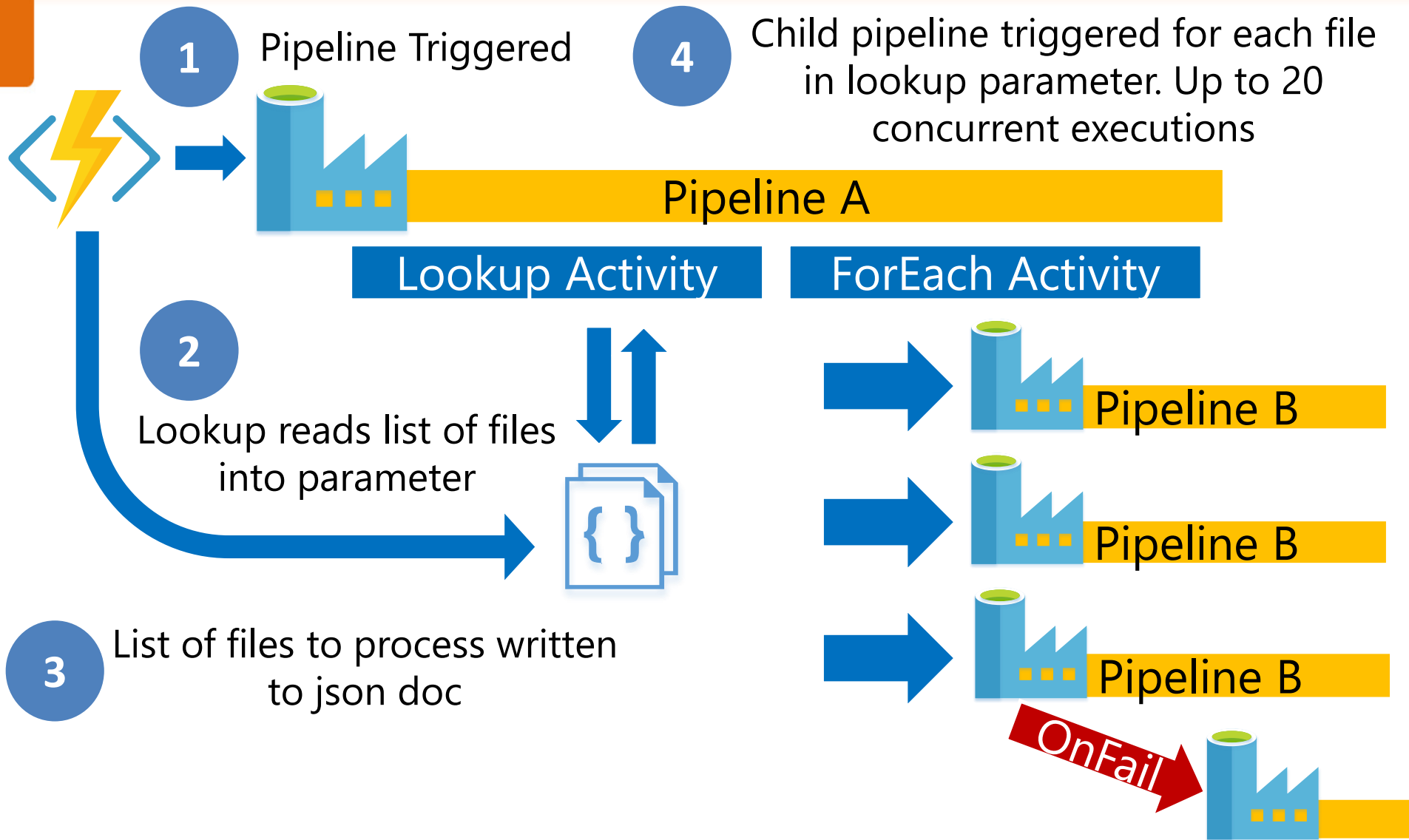
## Event Driven Loading



# Reusable Pipelines with Conditional Logic



# Design Pattern Combinations





# Is ADF the right tool for our cloud data integration?



**Maybe, limited use.**



**Yes, definitely.**

# Thanks for Listening

## Paul Andrew

 @MrPaulAndrew



**Blog:** <http://mrpaulandrew.com>

**Email:** [paul@mrpaulandrew.com](mailto:paul@mrpaulandrew.com)