

INTRODUCTION TO DATA FACTORY

THE THING THAT MOVES DATA AND ORCHESTRATES STUFF



Simon Whiteley
@MrSiWhiteley

SESSION AGENDA

Data Factory
Recap

Concepts
Components

ADFv2

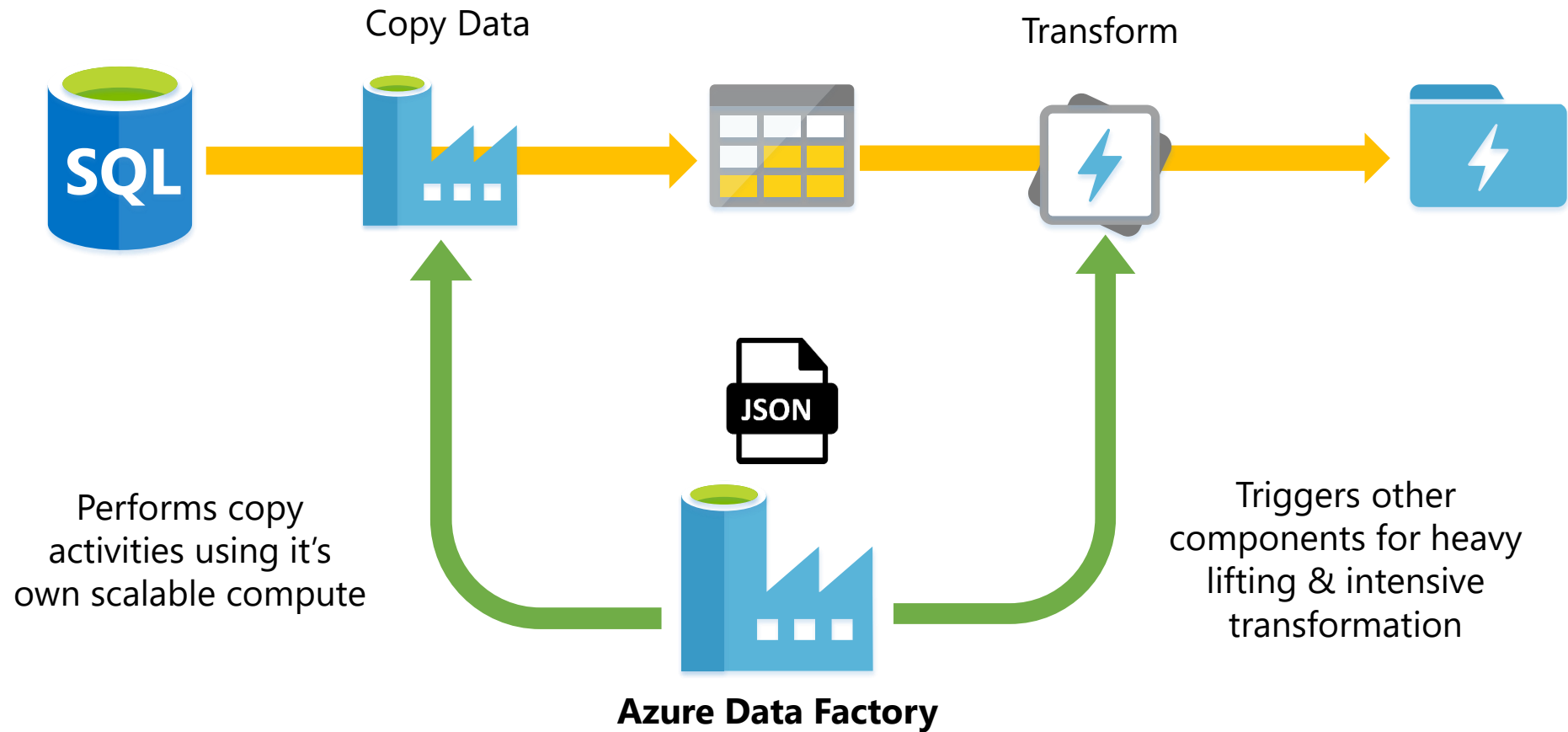
Features Update
The Integration
Runtime

Mapping Data
Flows

ADF DATA FLOWS

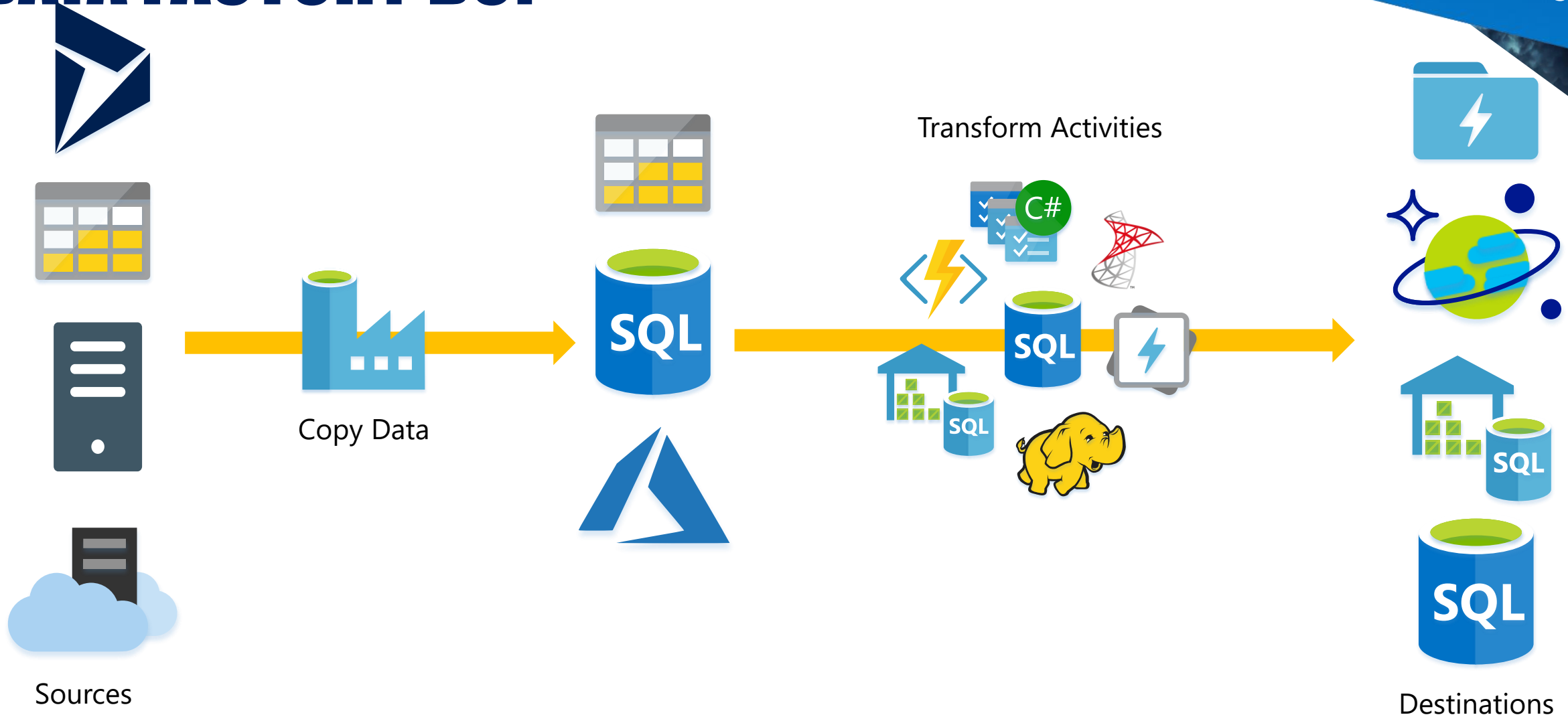
WHAT IS AZURE DATA FACTORY?

MODERN DATA WAREHOUSING



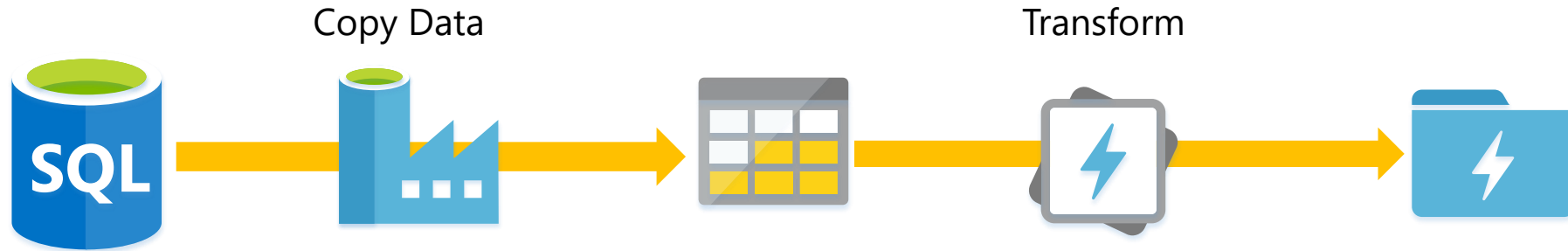
WHAT DOES AZURE DATA FACTORY DO?

MODERN DATA WAREHOUSING



DATA FACTORY COMPONENTS

MODERN DATA WAREHOUSING



1 Linked Services – How do I connect?

Like the SSIS Connection Manager!

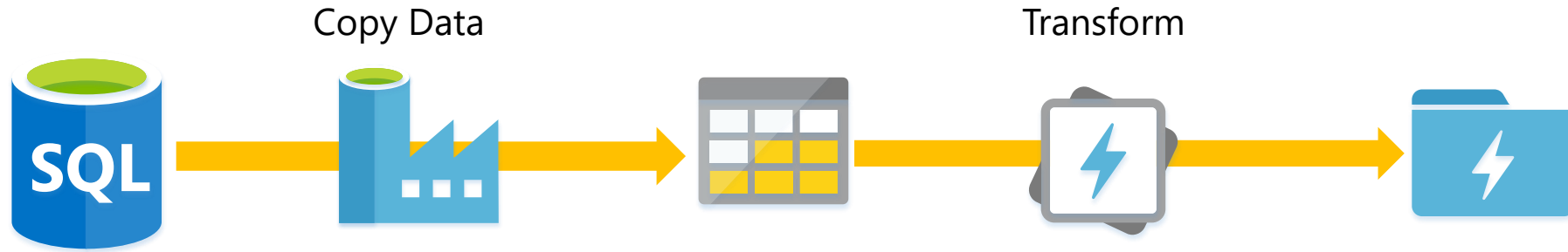


SQLDBLinkedService

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

DATA FACTORY COMPONENTS

MODERN DATA WAREHOUSING



1 Linked Services

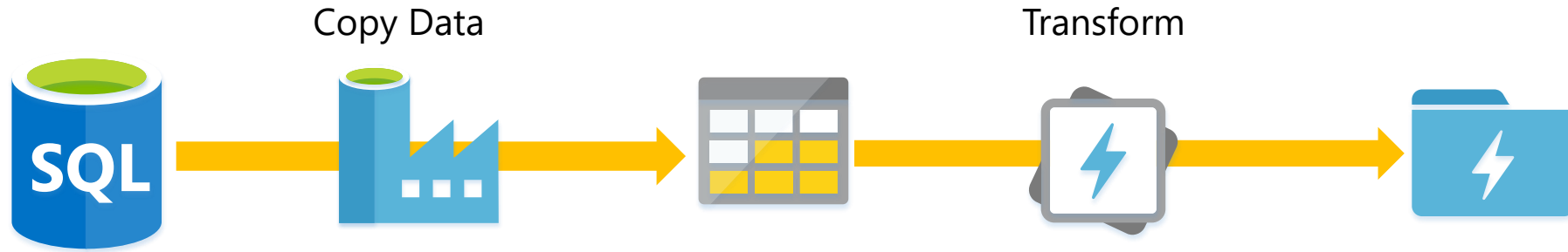
2 Data Sets – What slices/partitions does my data have?

 `dbo.DimCustomer`

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

DATA FACTORY COMPONENTS

MODERN DATA WAREHOUSING



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*

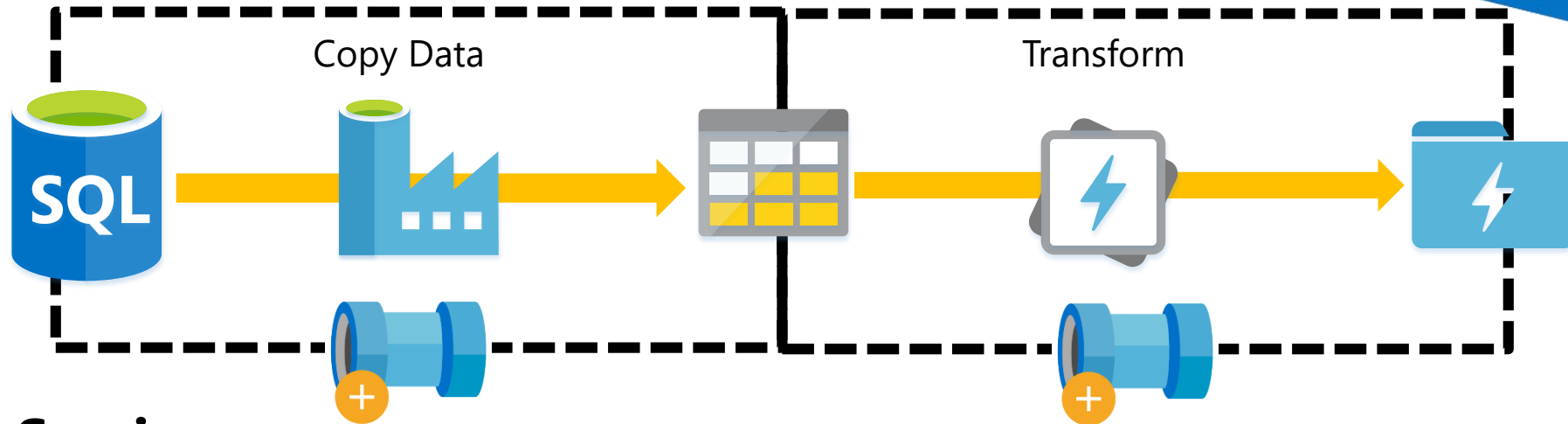
AUs: *5 units*

Priority: *1000*

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

DATA FACTORY COMPONENTS

MODERN DATA WAREHOUSING



1

Linked Services

2

Data Sets

3

Activities

4

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

DEVELOPER TOOLS


The screenshot displays the Microsoft Azure Data Factory (ADF) developer interface. The top navigation bar shows 'Microsoft Azure | Data Factory | dbricks' and a search bar. The left sidebar, titled 'Factory Resources', lists 'Pipelines' (3), 'DataFlows', '* Load Data' (selected), 'RunNotebooks', 'Datasets' (3), 'Data Flows (Preview)' (1), and 'B2E_TransformTaxi'. The main canvas shows a data flow pipeline named 'Load Data *'. The pipeline consists of several activities: 'Copy Data' (Move Data to Blob), 'Notebook' (Clean Data), 'Notebook' (Process Data), 'Azure Function' (Log Clean Failure), 'Stored Procedure' (Load to SQLDW), and 'Azure Function' (Log Process Failure). The 'Activities' pane on the left lists various activity types: Batch Service, Databricks (Notebook, Jar, Python), Move & Transform (Copy Data, Delete), Data Flow (Preview), Data Lake Analytics, and General (Append Variable, Azure Function). The bottom pane shows the 'General' tab for the selected activity, with fields for 'Name' (Load Data) and 'Description'.




ADF DATA FLOWS

MONITORING

 Run  Cancel options  Refresh





 Custom Range 04/02/2019 8:00 AM - 04/07/2019 8:00 AM ▾

 Time Zone (UTC+00:00) Dublin, Edinburgh, Li... ▾

☐ View All Rerun History

 Filter

All Succeeded In Progress Queued Failed Cancelled

<input type="checkbox"/>	Pipeline Name ▾	Actions	Run Start ▾	Duration	Triggered By	Status	Parameters	Annotations ▾	Error	RunID
	RunNotebooks	 	04/03/2019, 6:08:10 PM	00:05:49	Manual trigger	✓ Succeeded				13ca5395-bb39-4559-a14
	RunNotebooks	 	04/03/2019, 4:03:46 PM	00:00:37	Manual trigger	✓ Succeeded				5b101cf8-e105-4e1c-8f82

 Custom Range 04/02/2019 8:00 AM - 04/07/2019 8:00 AM ▾

 Time Zone (UTC+00:00) Dublin, Edinburgh, Li... ▾

 Pipeline

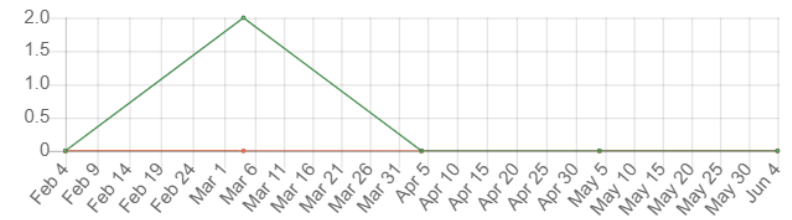
 



SUCCEEDED RUNS
2

 Activity



SUCCEEDED RUNS
2



ADF DATA FLOWS

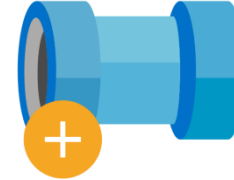
ADF IRS



1

Default
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

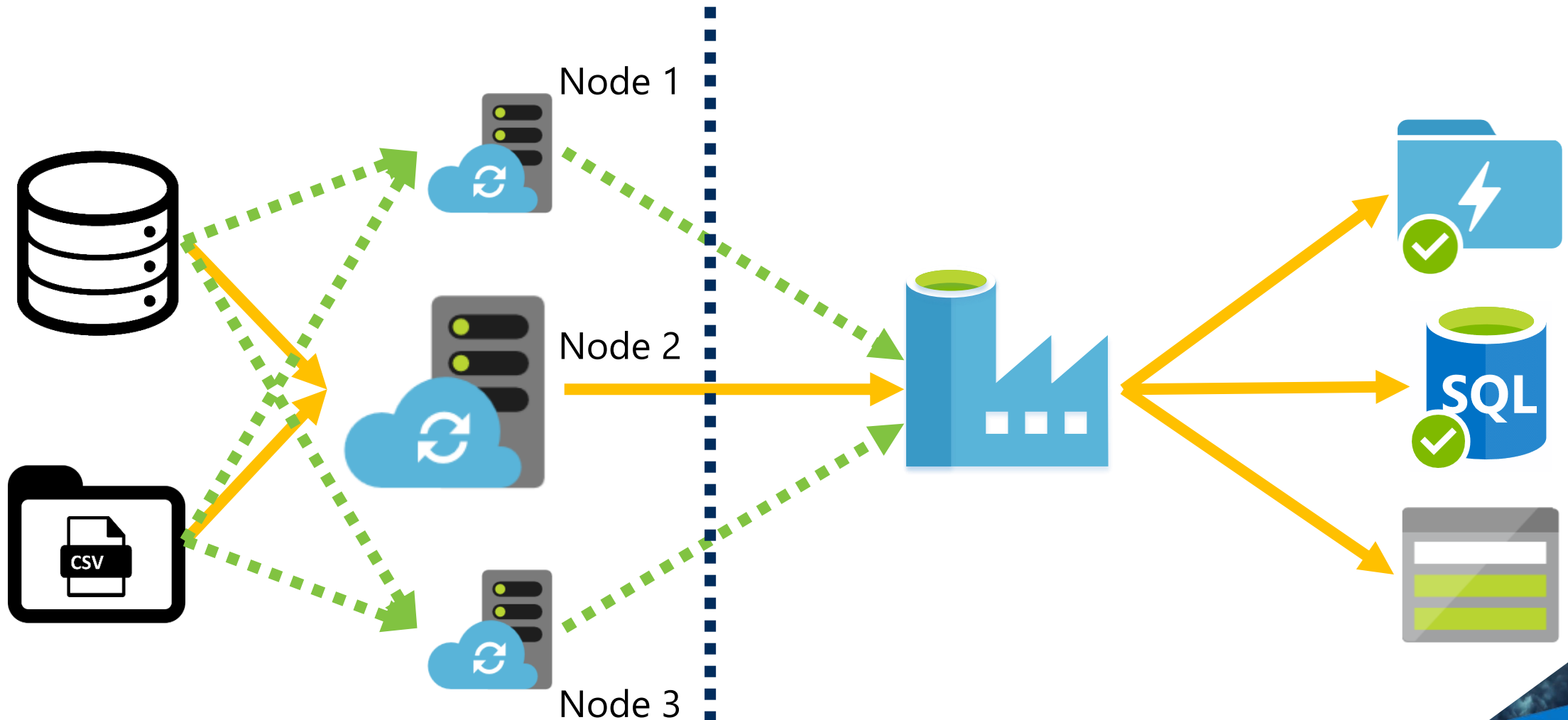


On-Prem Server



ADF FLOWS

THE INTEGRATION RUNTIME *(AKA THE DATA MANAGEMENT GATEWAY)*



ON PREMISES

Azure

ADF DATA FLOWS

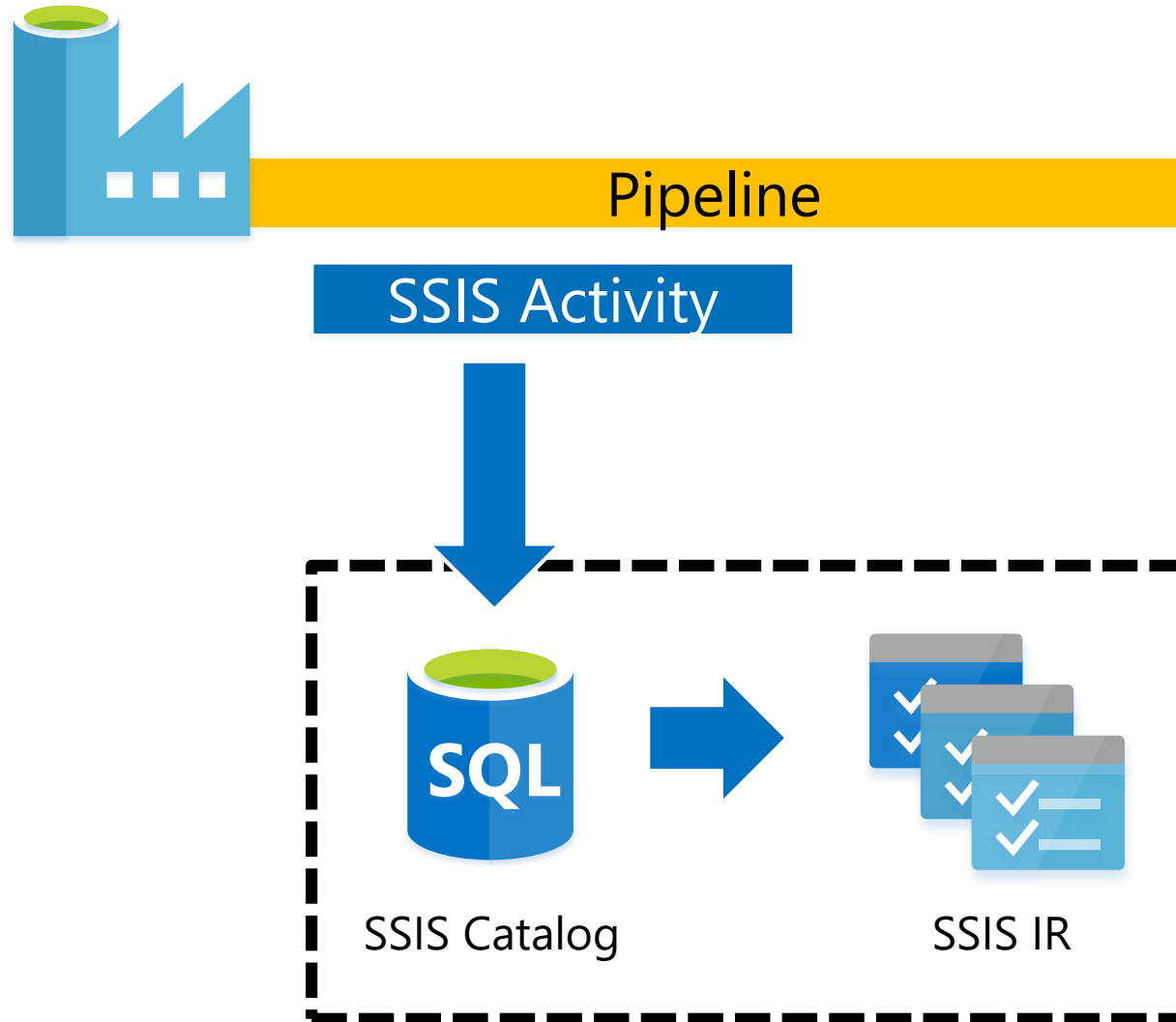


DEMO:

DATA FACTORY QUICK PEEK

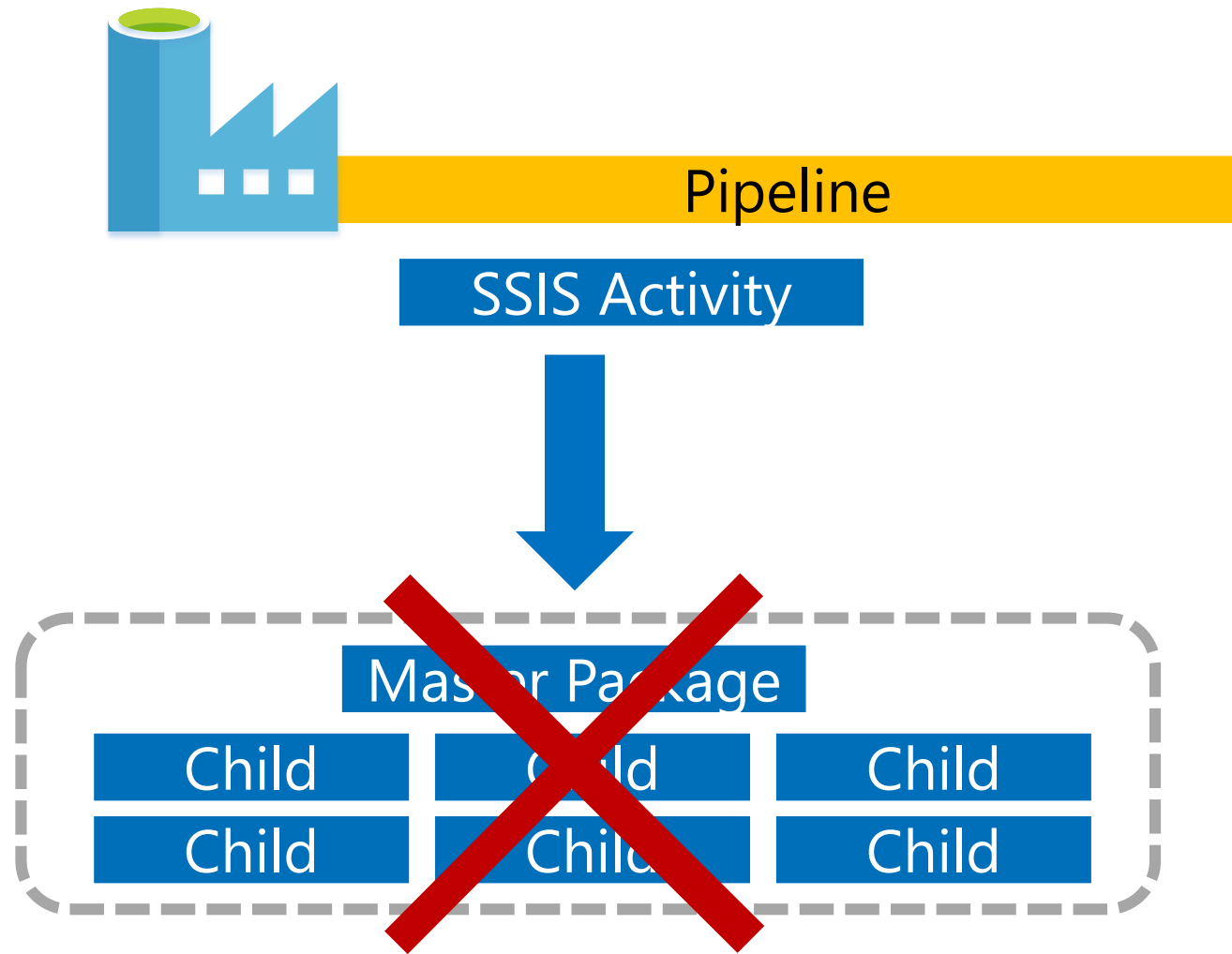
- Data Factory Workspace
- Linked Services, Activities & Pipelines

SSIS INTEGRATION RUNTIMES



ADF DATA FLOWS

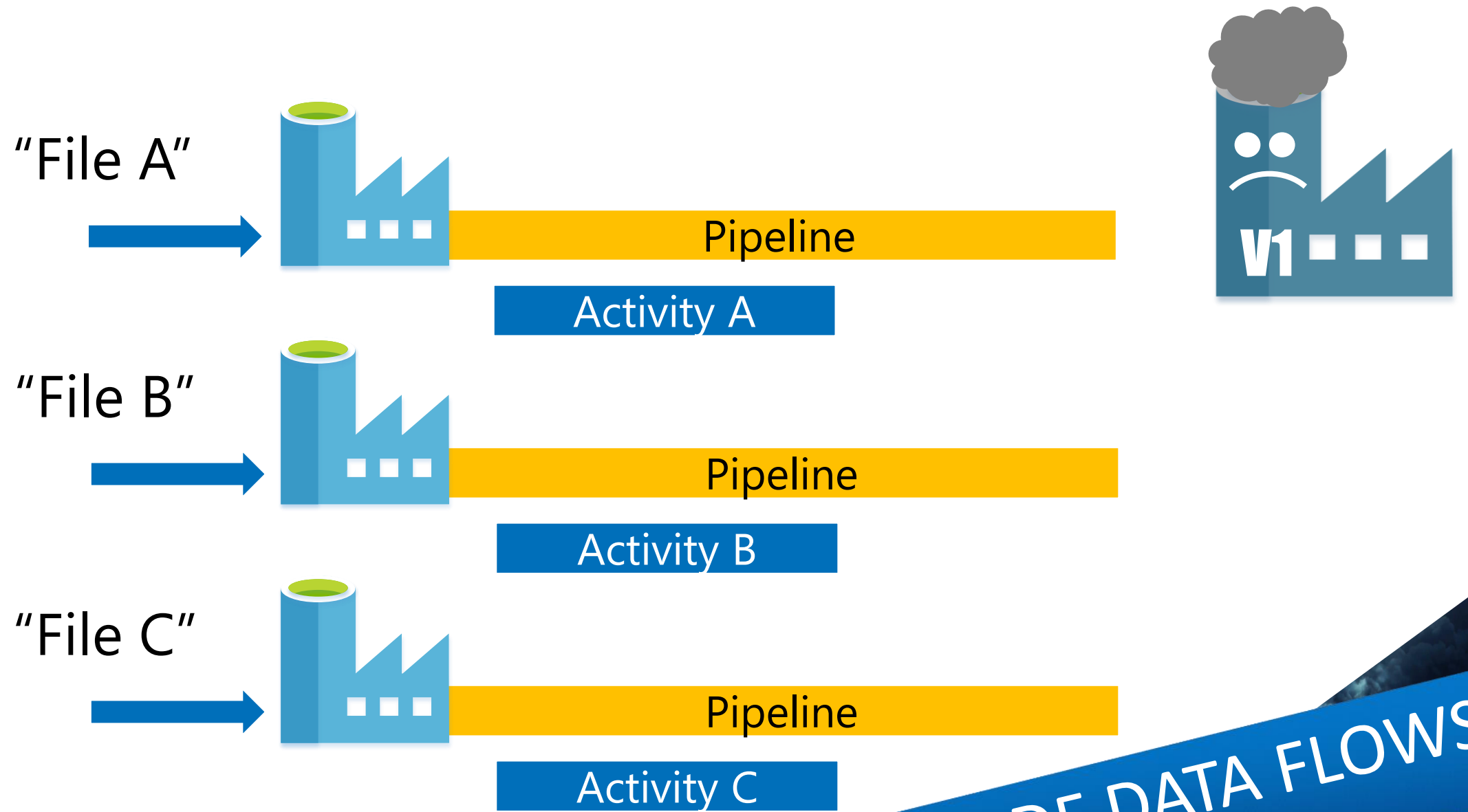
CHILD EXECUTIONS



This will execute the packages on a single node, rather than scale out across multiple!

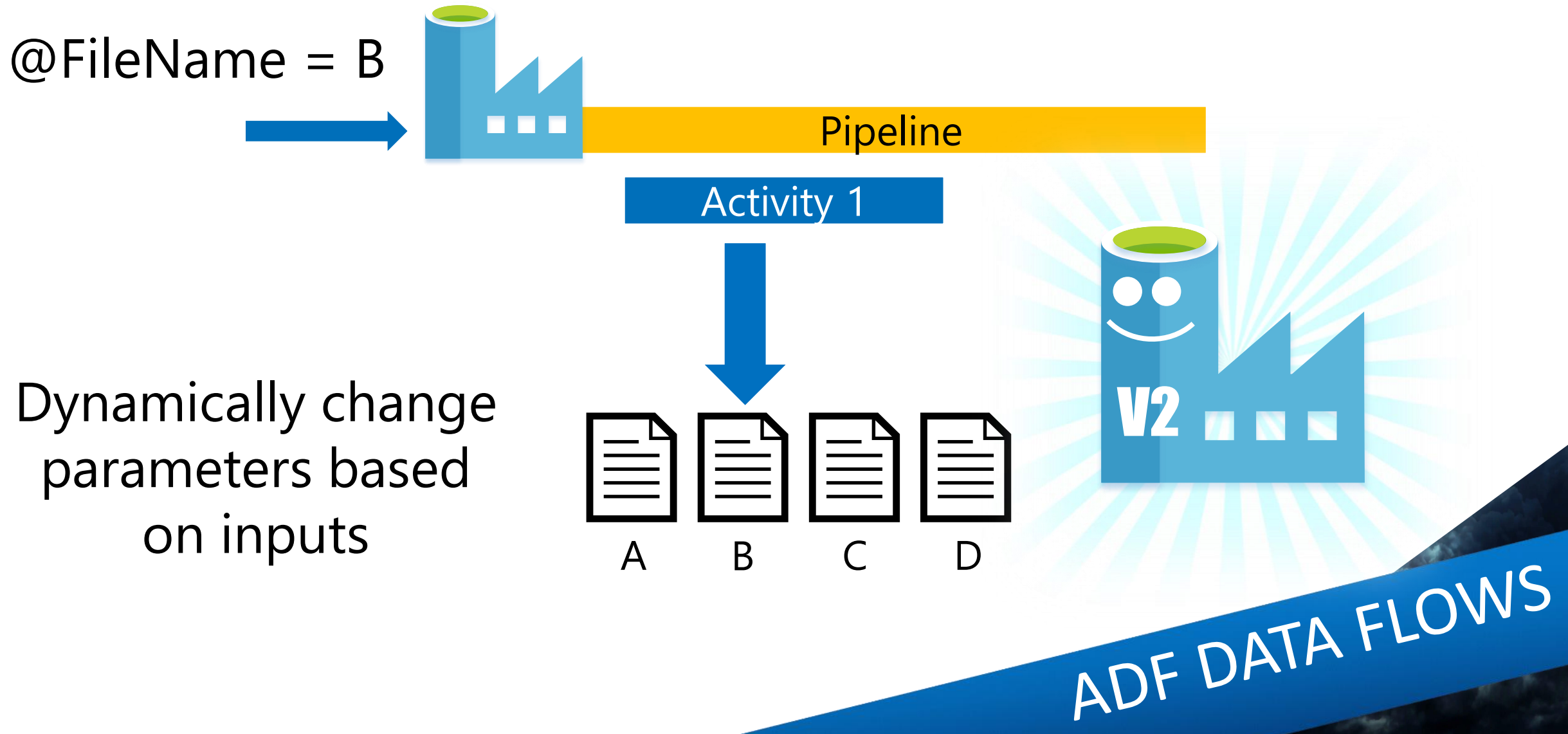
ADF DATA FLOWS

HARDCODED PIPELINES

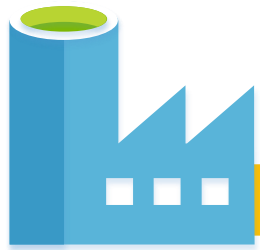


ADF DATA FLOWS

DYNAMIC PIPELINES USING PARAMETERS



DYNAMIC PIPELINES USING LOOKUP ACTIVITY



Pipeline

Lookup Activity



Returns
@FileName = B



Config.json

Activity 2



A



B



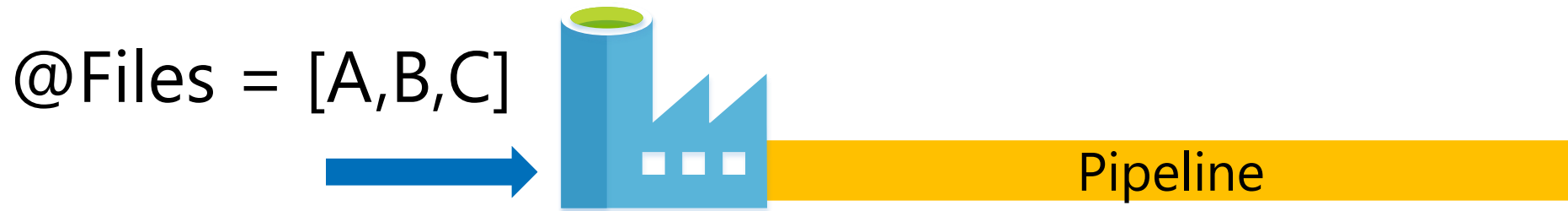
C



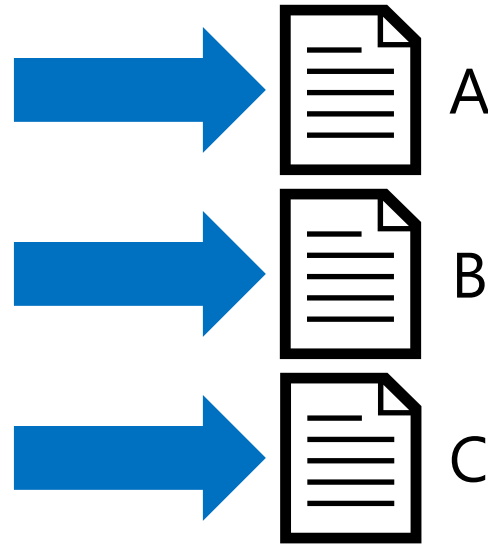
D

ADF DATA FLOWS

FOREACH PIPELINES



ForEach Activity

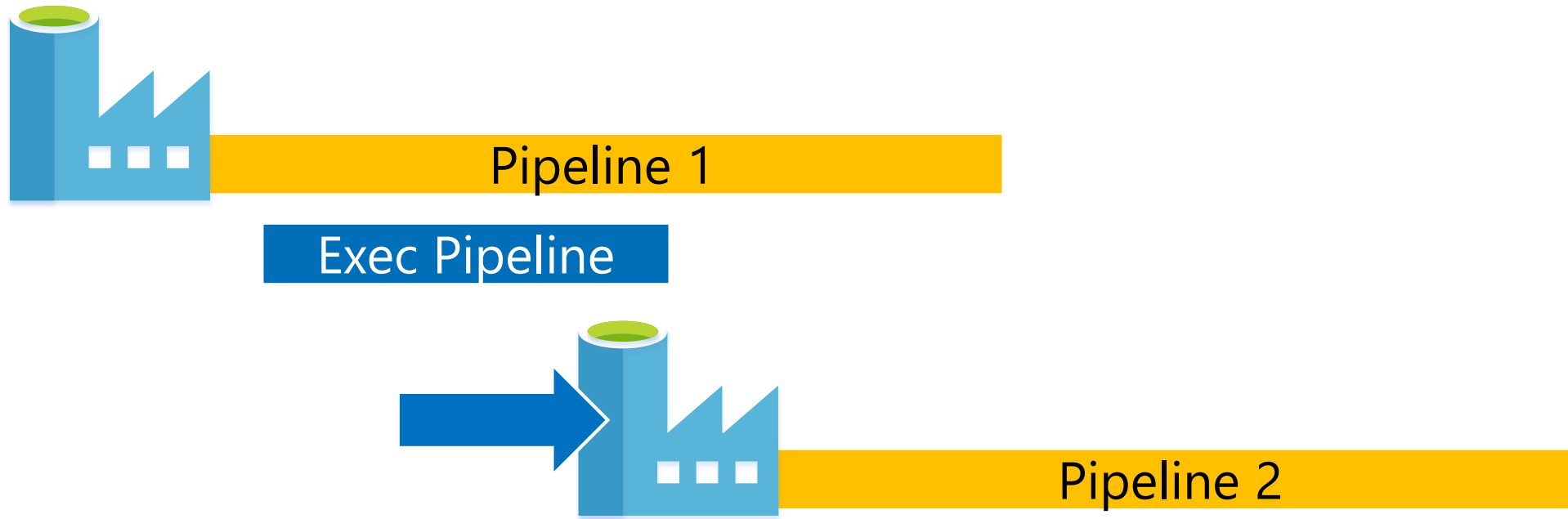


Run activity for each
item in an array

(max DOP 20)

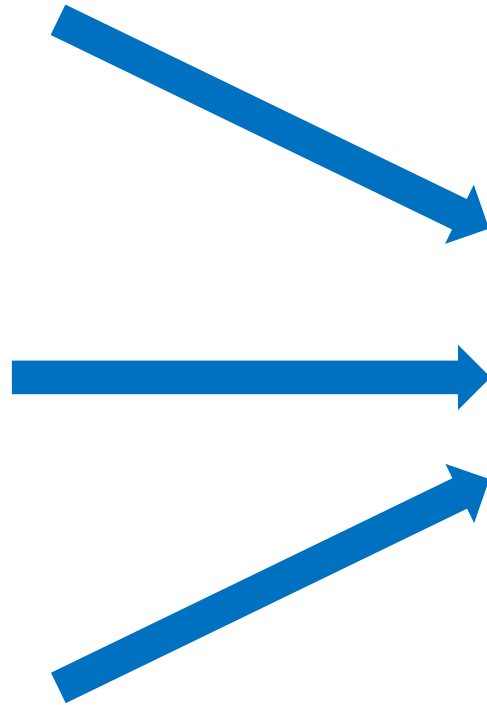
ADF DATA FLOWS

EXECUTE PIPELINE ACTIVITY



ADF DATA FLOWS

PIPELINE TRIGGERS



Pipeline 1



ADF DATA FLOWS



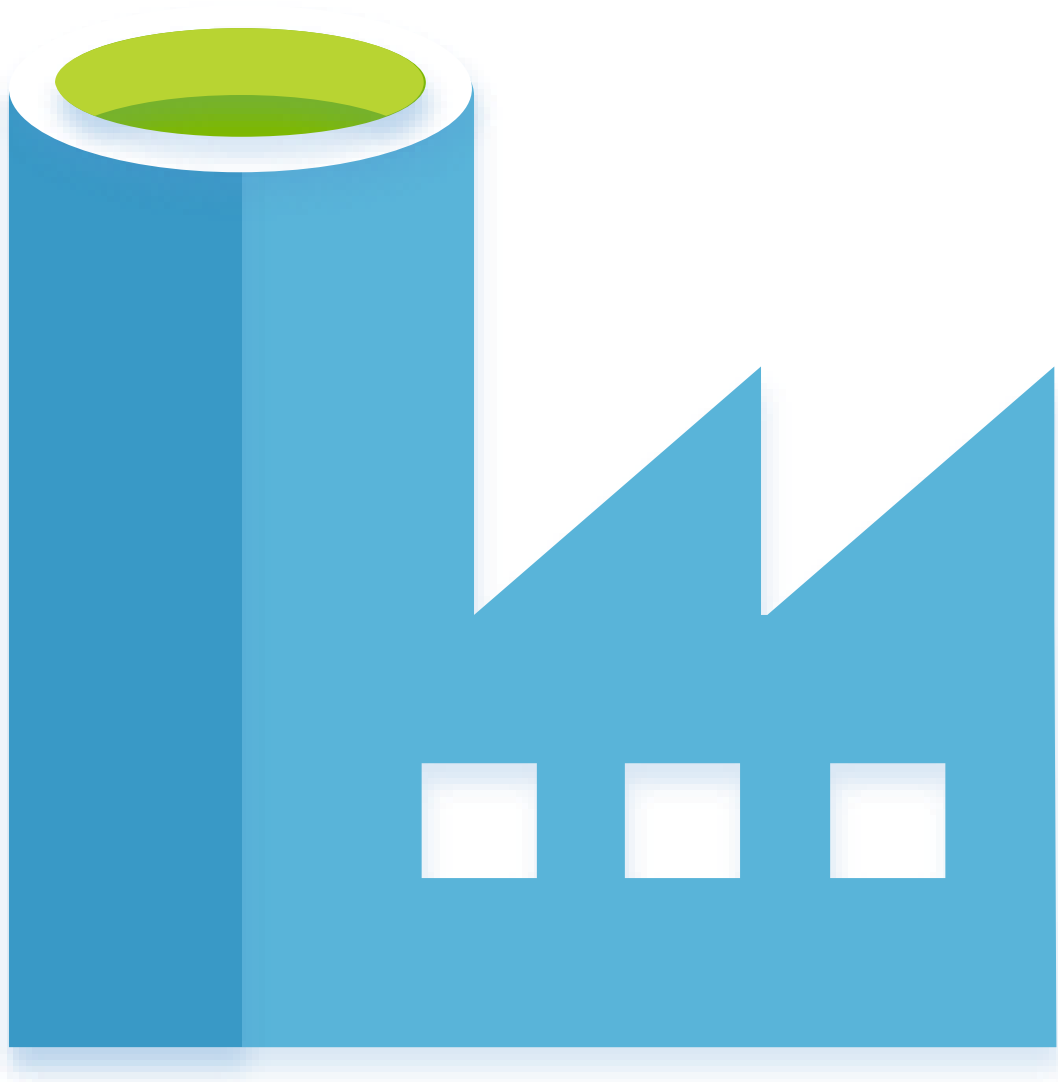
DEMO:

DYNAMIC PIPELINES

- Creating a Generic Pipeline
- Metadata Driven Workflows

But what if I don't
want to write any
code?



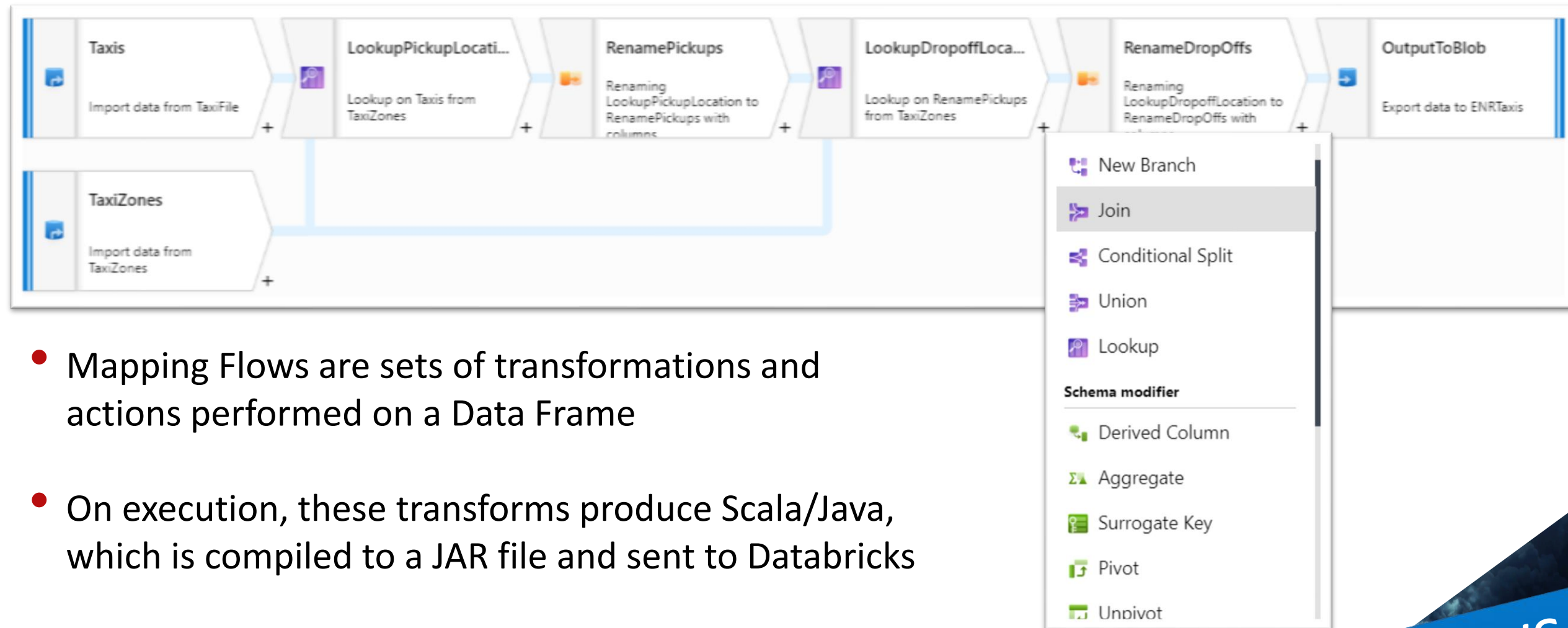


Azure Data Factory

Data Flows

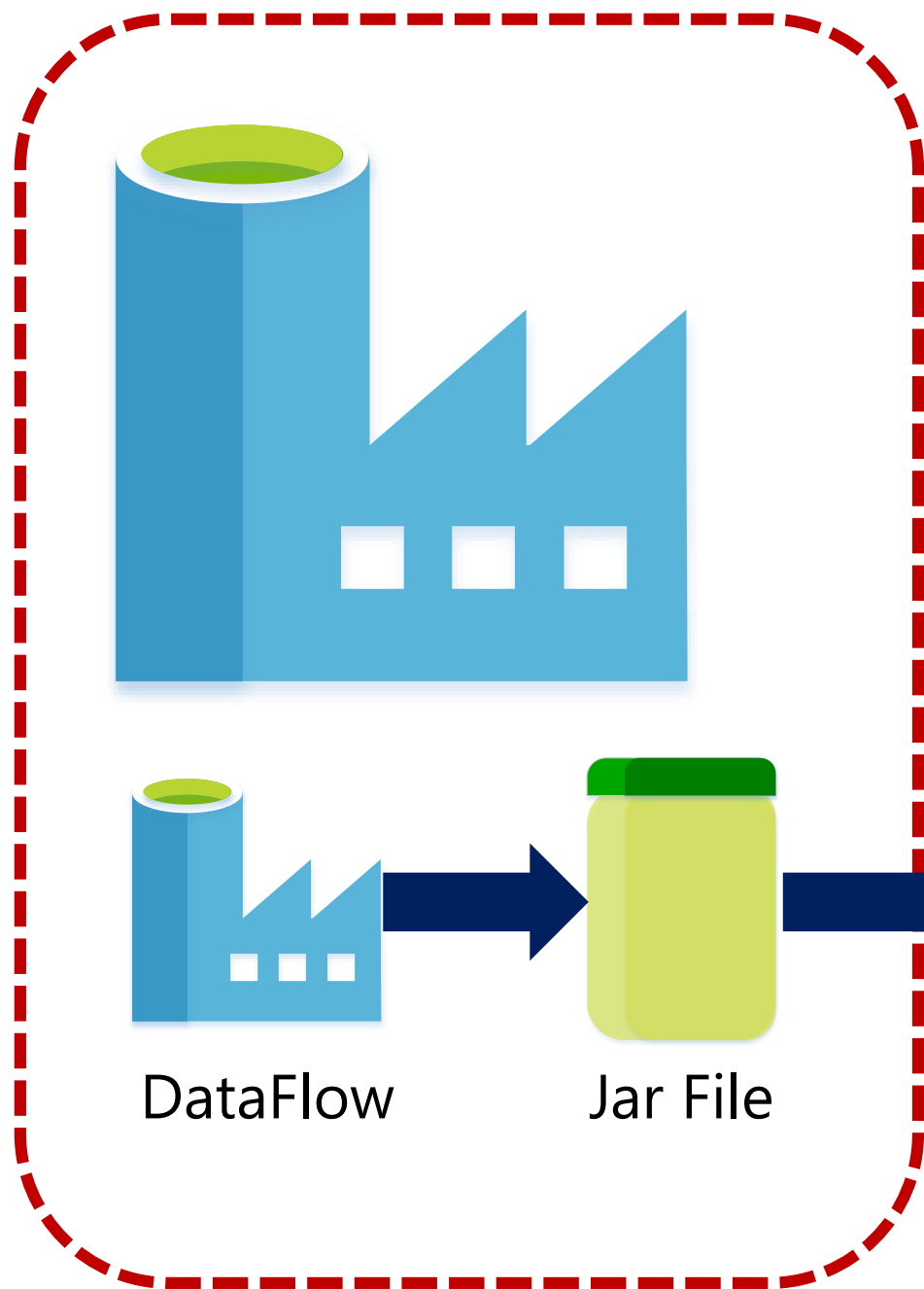
ADF DATA FLOWS

ADF MAPPING DATA FLOWS



- Mapping Flows are sets of transformations and actions performed on a Data Frame
- On execution, these transforms produce Scala/Java, which is compiled to a JAR file and sent to Databricks
- ADF uses it's own built-in Databricks cluster for this, but should be able to use your own in future

ADF DATA FLOWS

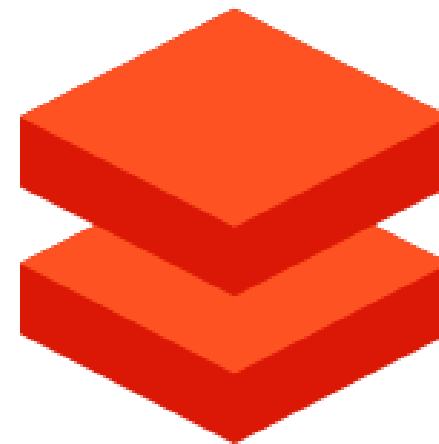


DataFlow

Jar File

Dataflows will compile down to a JAR file which will be sent to the Databricks cluster for execution

This means
it uses
Scala!



ADF DATA FLOWS



DEMO:

ADF MAPPING DATA FLOW

- Databricks Transformation Notebooks
- Creating an ADF Mapping Data Flow
- Running & Monitoring ADF Mapping Data Flows