Building an Azure Business Intelligence Solution End to End

Hands On Workshop

Paul Andrew | Senior Consultant

Terry McCann | Principal Consultant

Simon Whiteley | Cloud Architect

adatis

Microsoft Partner | Gold Data Analytics
Gold Data Platform
Gold Cloud Platform

Microsoft

https://github.com/Adatis

ModernDataWarehouseWorkshop

adatis

Agenda for the Day

Module 1	Module 2	Module 3	Module 4
Microsoft Azure	Storage Uploading Data Data Lake	Real-time Data Streaming Power BI	U-SQL - Data Transformation Basics
Module 5	Module 6	Module 7	Module 8
USQL - Advanced Analytics Cognitive Services	Data Factory Orchestration Dynamic Pipelines	Data Presentation & Consumption Power BI Models	Other Services Q&A

Session Agenda

Data Factory Recap

Concepts
Components

ADFv2

Features Update
The Integration Runtime

Data Factory Extensibility

SSIS, Functions, Custom Activities
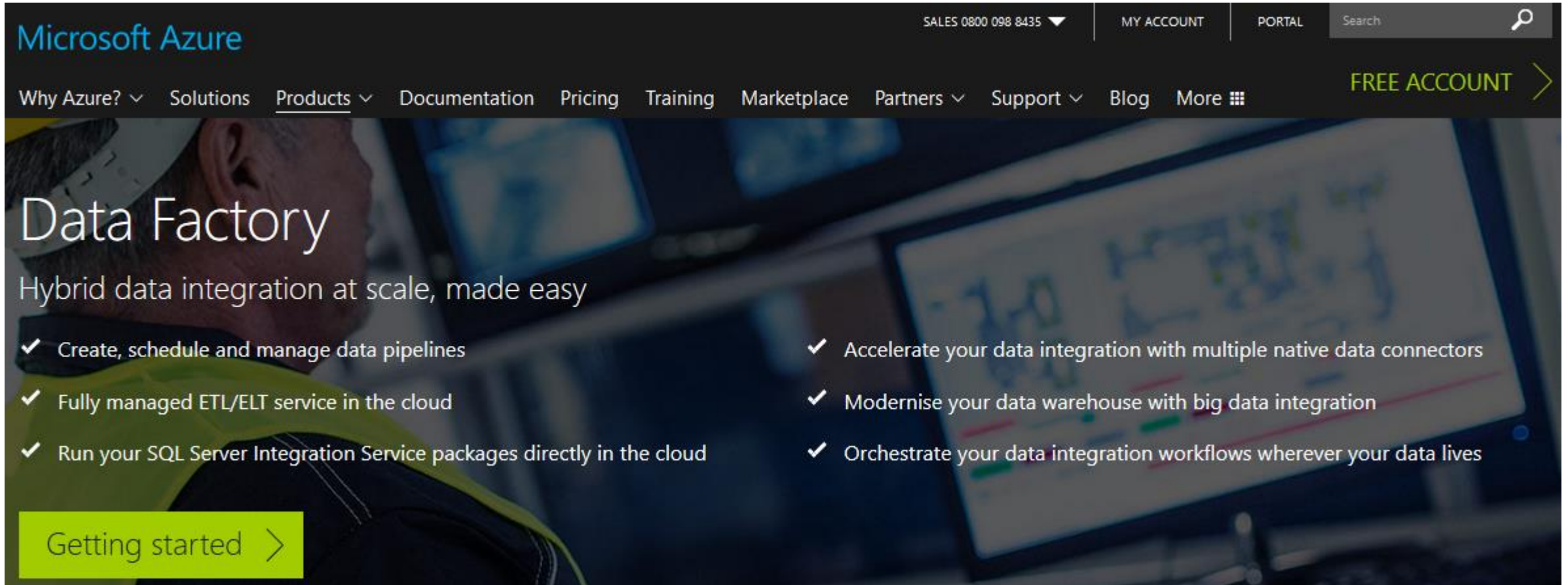
Agenda

Data Factory Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

Data Factory
Extensibility

SSIS, Functions,
Custom Activities

What is Azure Data Factory?

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

What is Azure Data Factory?

Copy Data

Transform

SQL

JSON

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?

Transform Activities

Copy Data

Sources

Destinations

Data Factory Components

Copy Data

Transform
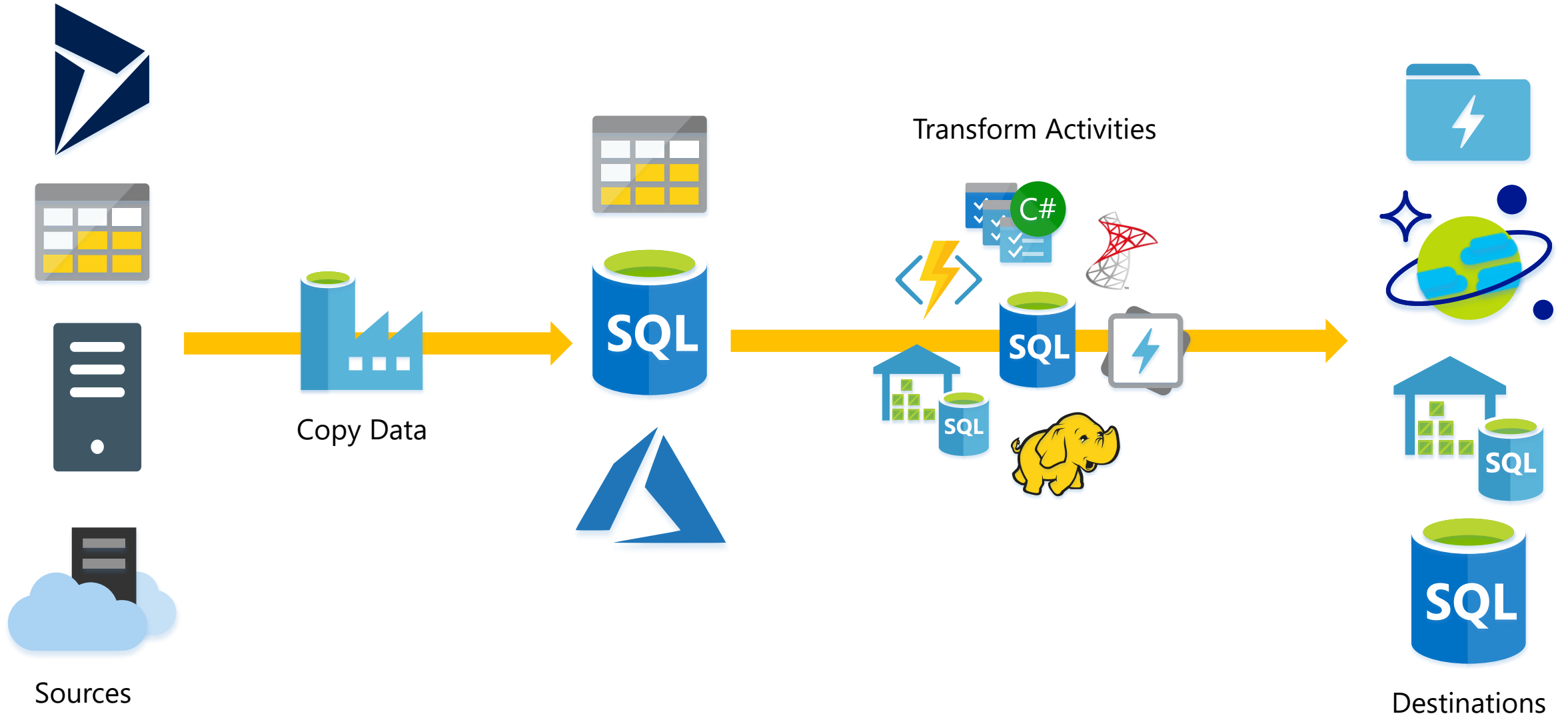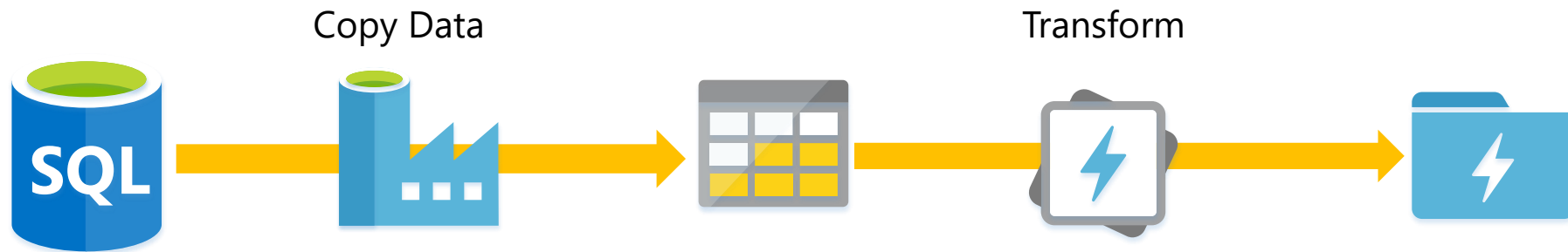
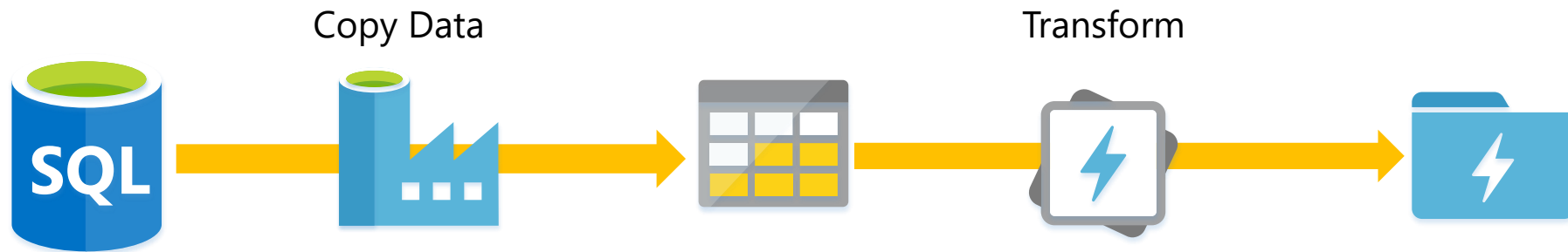**1** **Linked Services** – How do I connect?   Like the SSIS Connection Manager!

SQLDBLinkedService

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

Data Factory Components

Copy Data

Transform

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

Copy Data

Transform
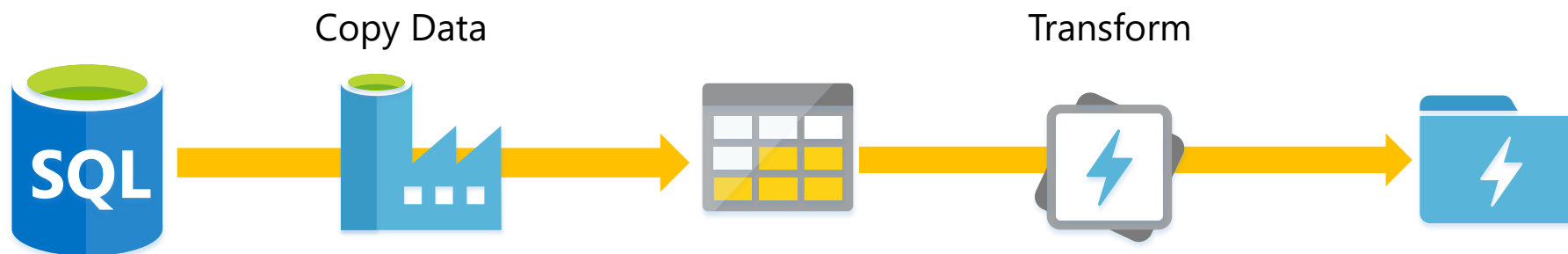
1. **Linked Services**

2. **Data Sets**

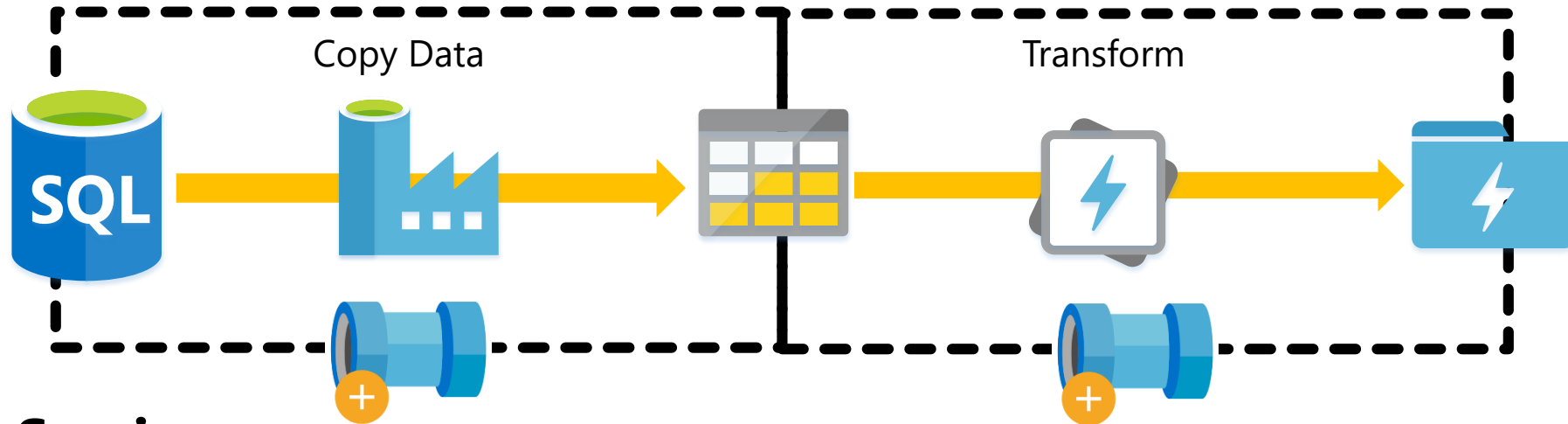3. **Activities** – What do we want to happen? With what conditions?

U-SQL Activity

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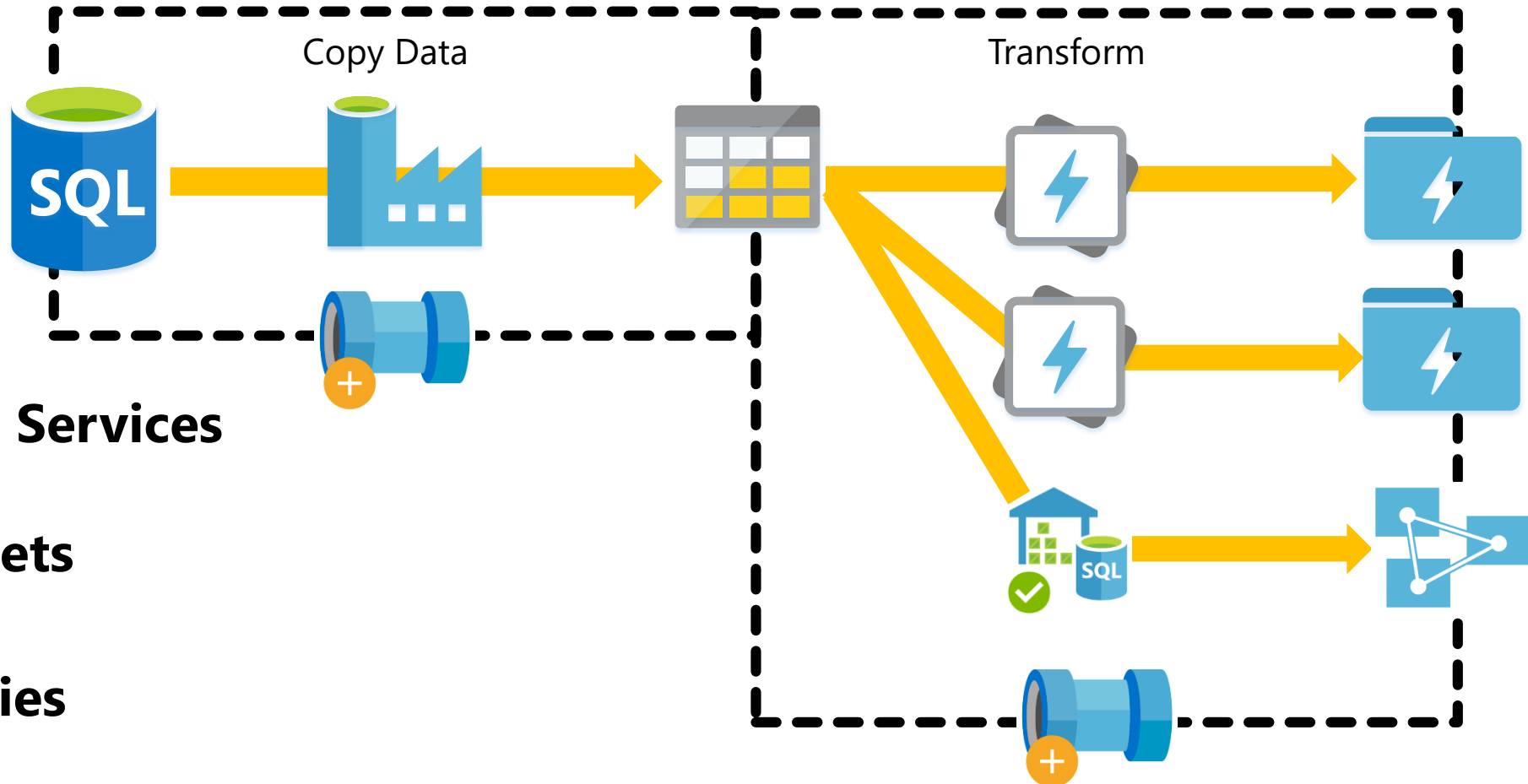
Data Factory Components

Copy Data   Transform

SQL

1. **Linked Services**

2. **Data Sets**

3. **Activities**

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

Data Factory Components

Copy Data

Transform

SQL

1 **Linked Services**
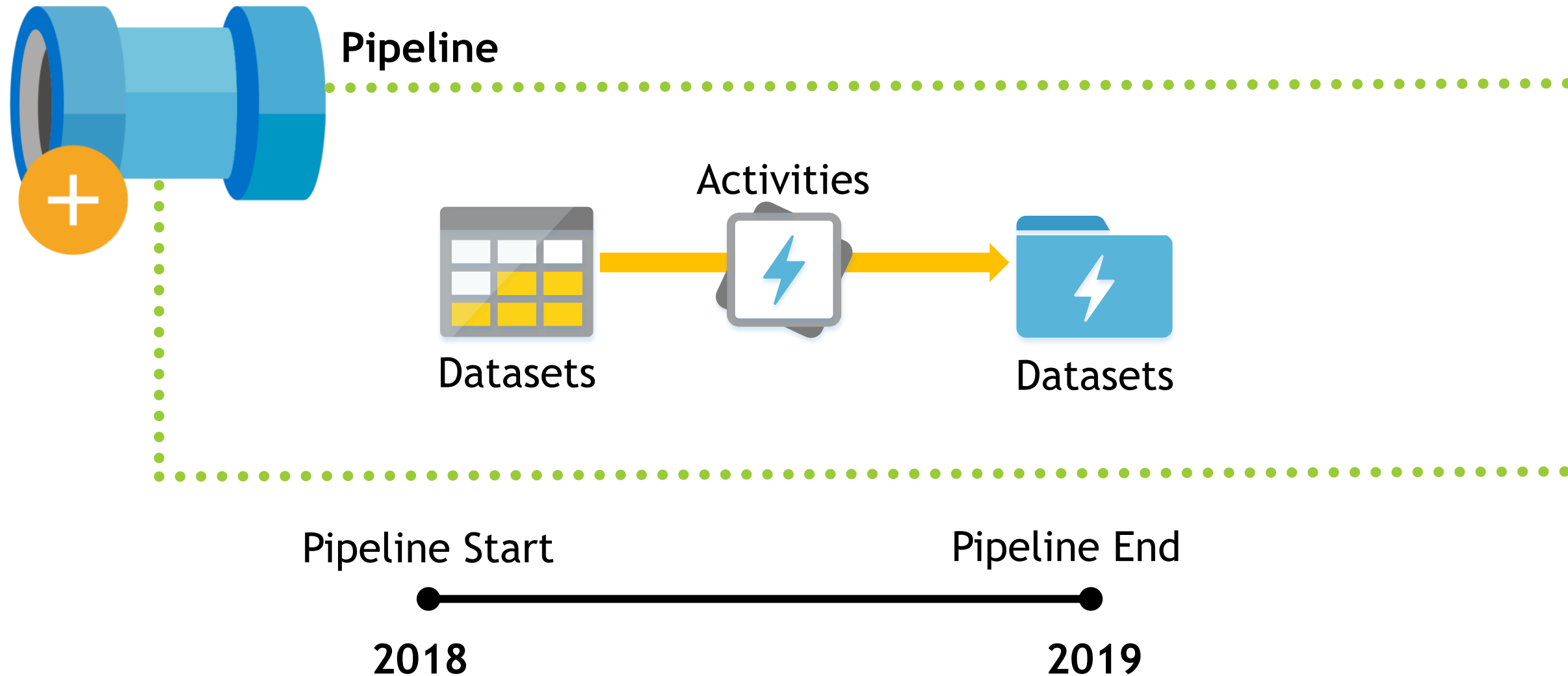
2 **Data Sets**

3 **Activities**

4 **Pipelines** – What groups of work do I want to do?
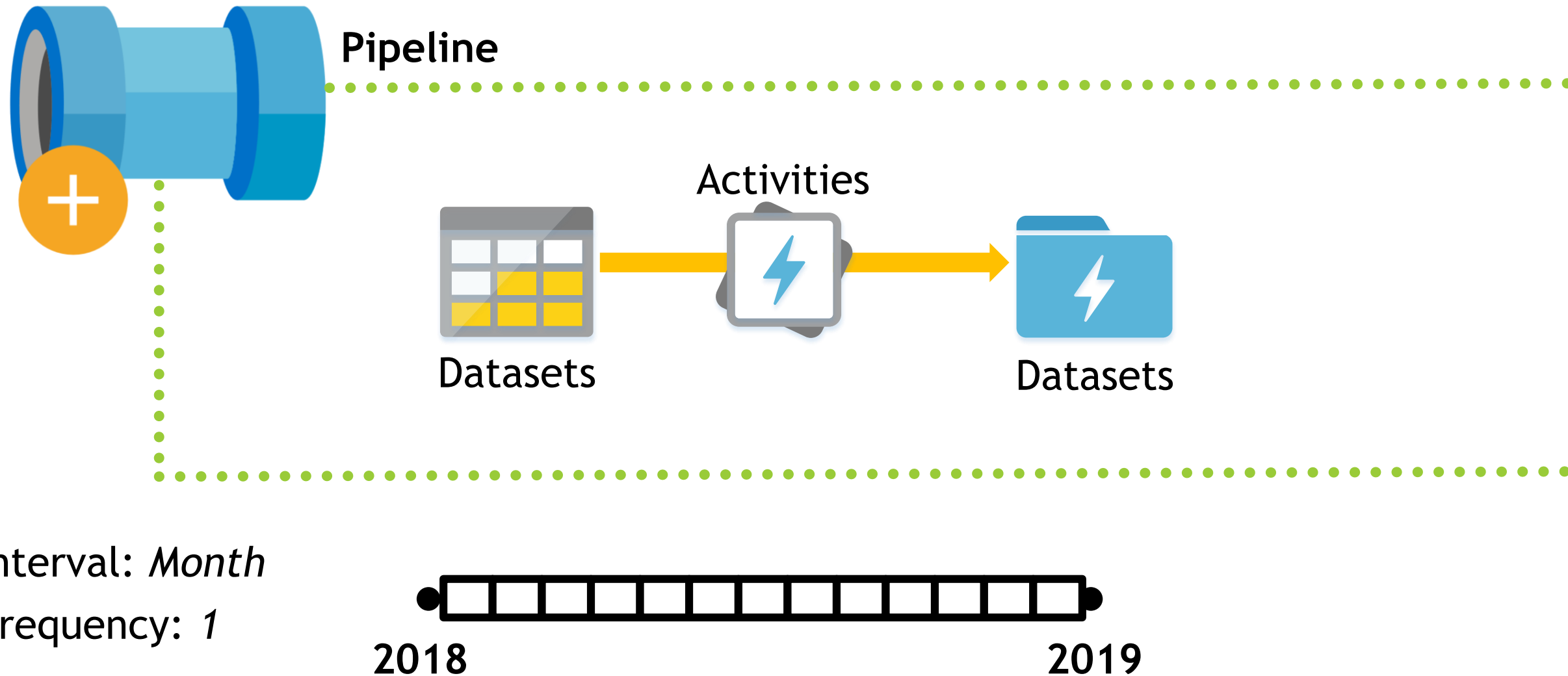
Azure Data Factory Concepts

Time Slices – triggering an activity execution.

Pipeline

Activities

Datasets
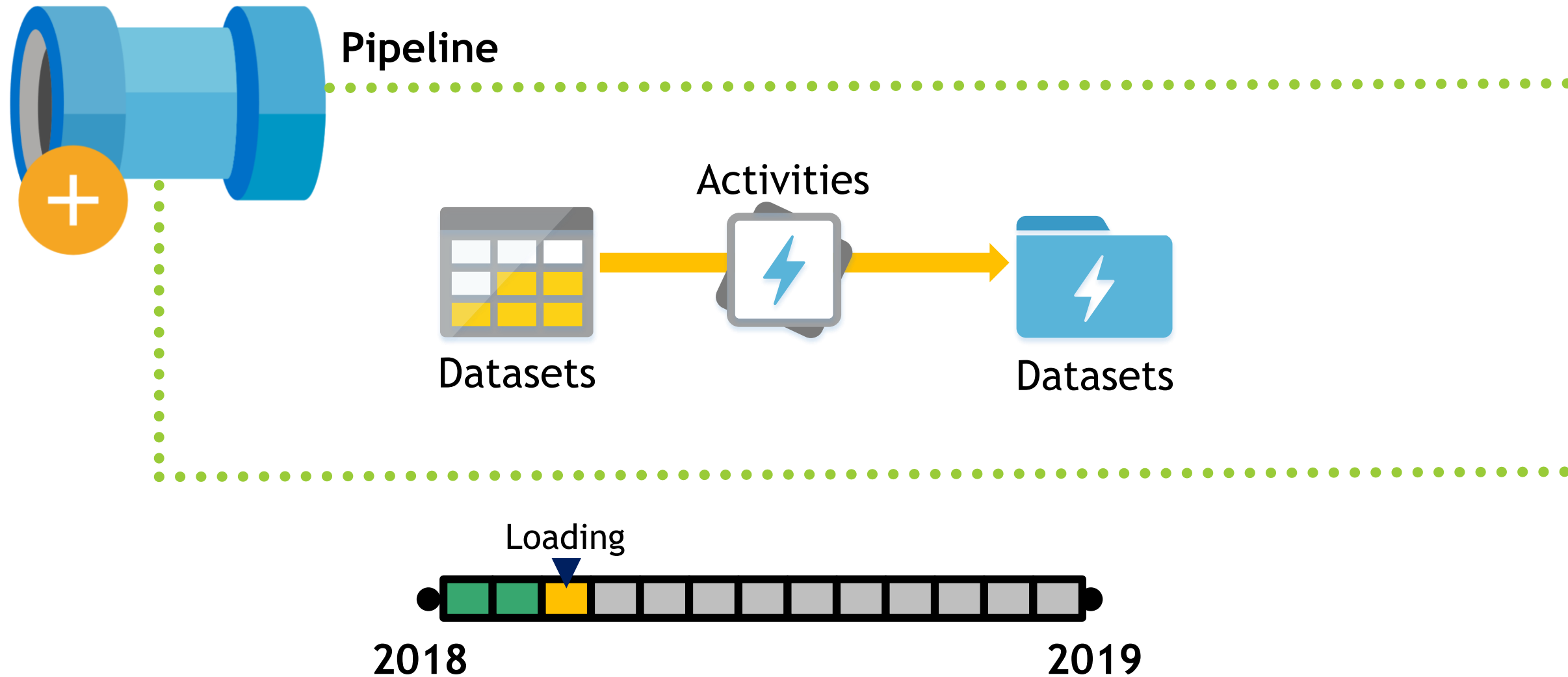
Datasets

Pipeline Start

Pipeline End

2018

2019

Azure Data Factory Concepts Continued

Time Slices – triggering an activity execution.

Pipeline

Activities

Datasets

Datasets

Interval: *Month*

Frequency: *1*

2018

2019

Azure Data Factory Concepts Continued

Time Slices – triggering an activity execution.

Pipeline

Activities

Datasets

Datasets

Loading

2018   2019

Time Slice Problems...

Dataset A

12:00am   12:00pm

Dataset B

Monday   Sunday

Dataset D

Dataset C

Monday

12:20am   12:20pm

Sunday

?
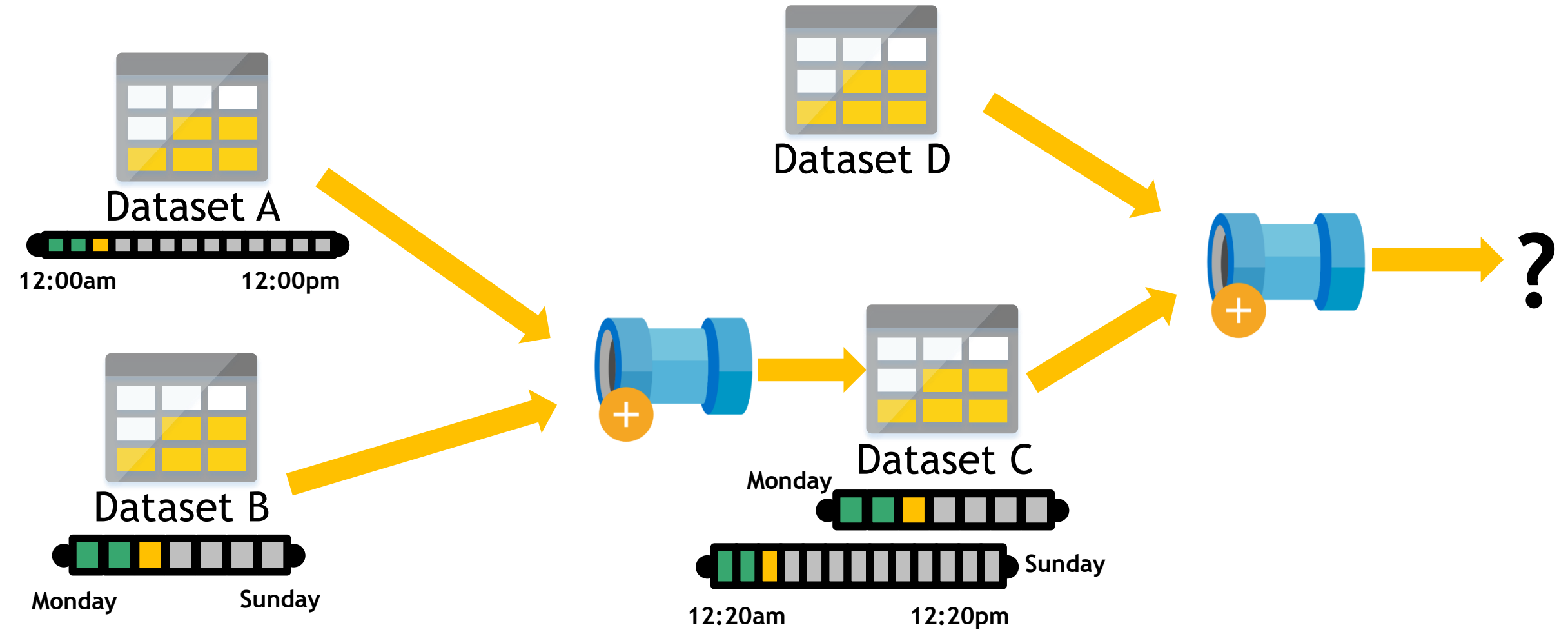
Agenda

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?

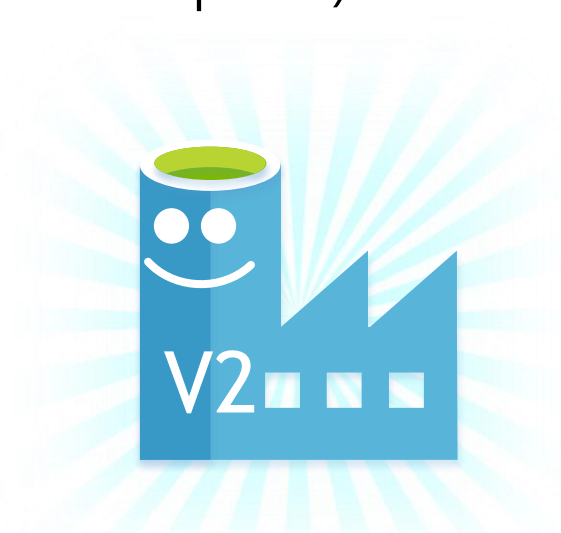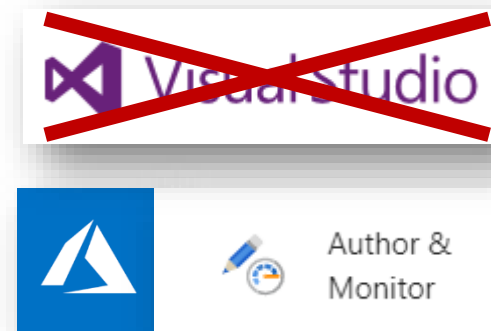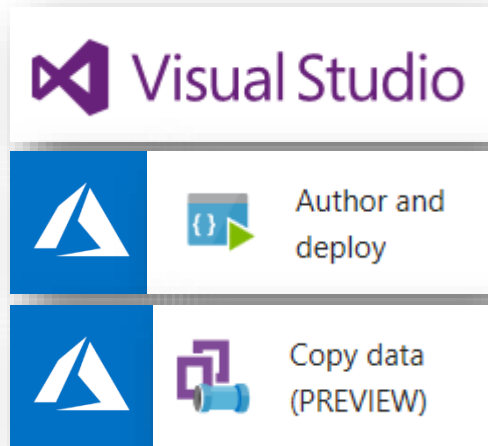| V1		V2
✔	Data Movement (Copy)	✔
✔	Activities	✔
✔	Pipelines	✔
✔	Datasets	✔
✔	Time Slices/Tumbling Windows	✔
✘	Event Triggers	✔
✘	Recurring Schedules	✔
✘	Parameters	✔
✘	Expressions	✔
✘	Conditional Logic	✔
✘	Use of SSIS Packages	✔
✘	Graphic Developer Canvas	✔
✘	Drilldown Monitoring	✔

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)

ADFPipeline.json - Untitled (Workspace) - Visual Studio Code

File   Edit   Selection   View   Go   Debug   Tasks   Help

{} ADFPipeline.json ✕

```
1  {
2      "PipelineName": "Why???"
3  }
```

V1

Visual Studio

Author and deploy

Copy data (PREVIEW)

~~Visual Studio~~

Author & Monitor

Microsoft® .NET

PowerShell

V2

Only Visual Studio 2015

http://bit.ly/2tsyD90

Monitoring

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

Node 1

Node 2

Node 3

On Premises

Azure
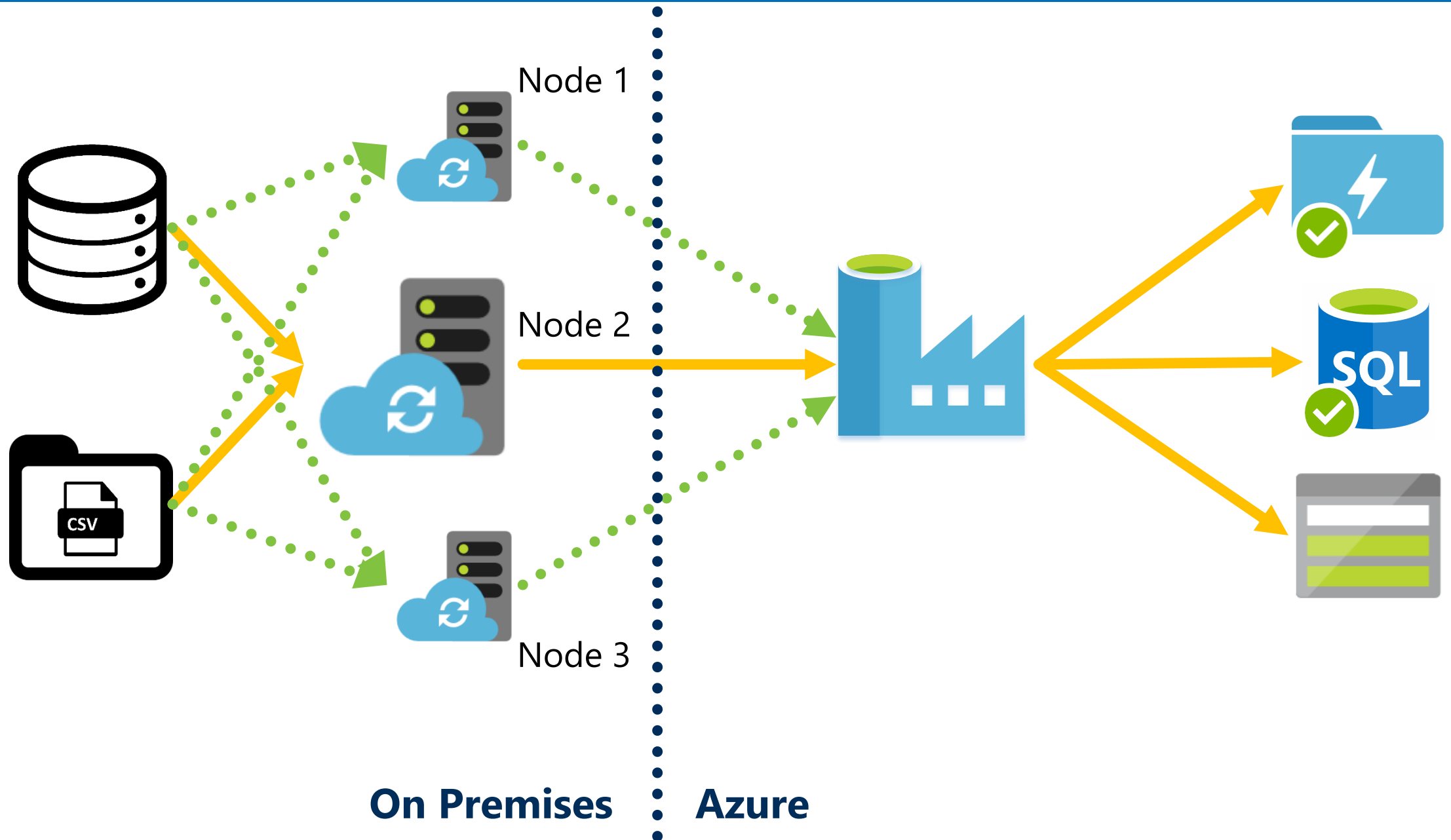
Integration Runtimes

1 Default Integration Runtime

Movement Hours

Activity Orchestration
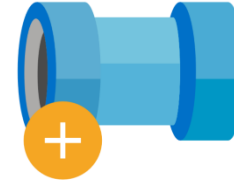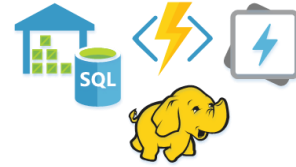
Flexible Region

2 SSIS Integration Runtime

SSIS Package Execution

Specified Region

3 Self Hosted Integration Runtime

Local Compute

Activity Orchestration

On-Prem Server

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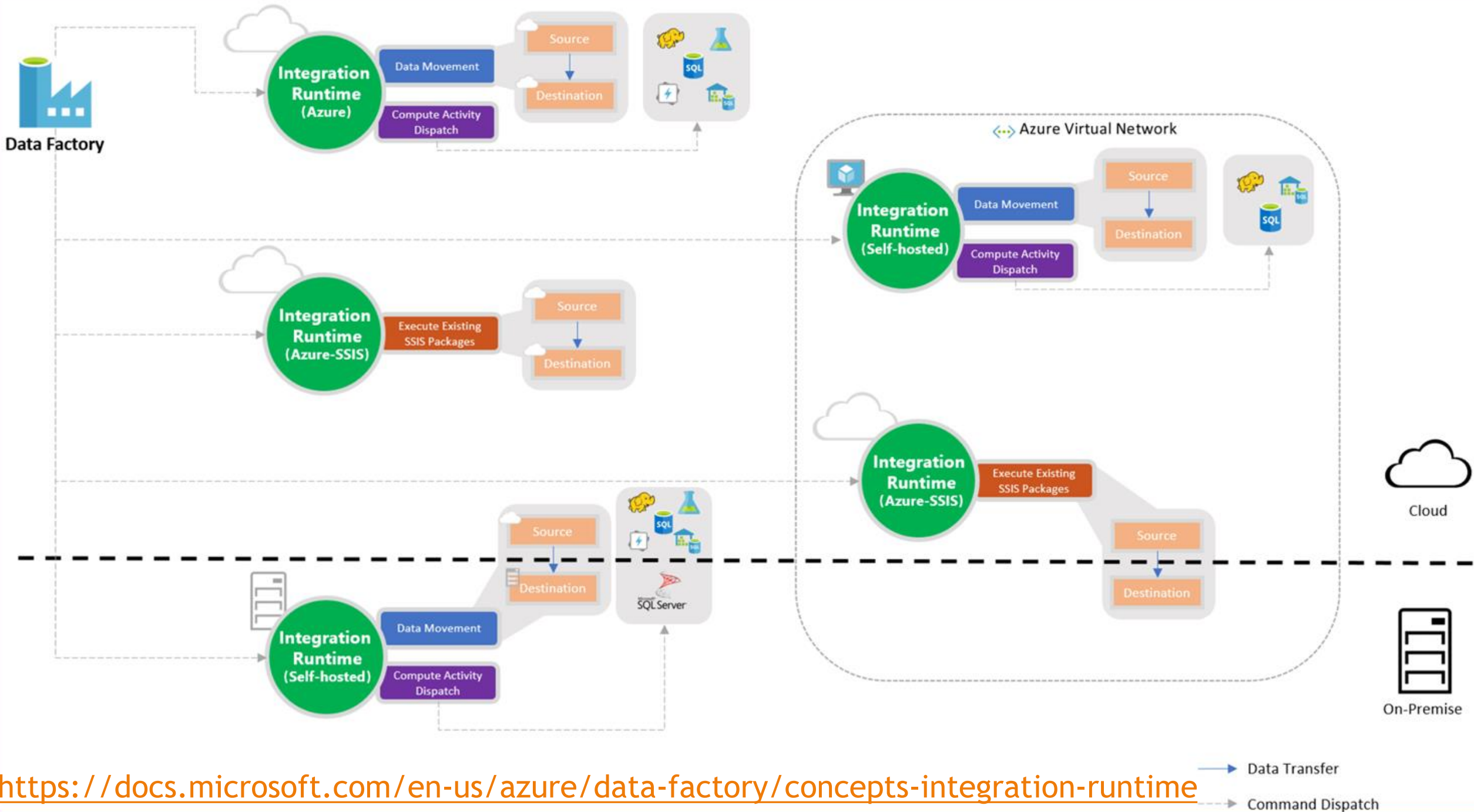
Agenda

Data Factory Recap

Concepts
Components

ADFv2

Features Update
The Integration
Runtime

Data Factory
Extensibility

SSIS, Functions,
Custom Activities

ADF Extensibility

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

Batch Service

Compute Pool

Compute Nodes

RDP
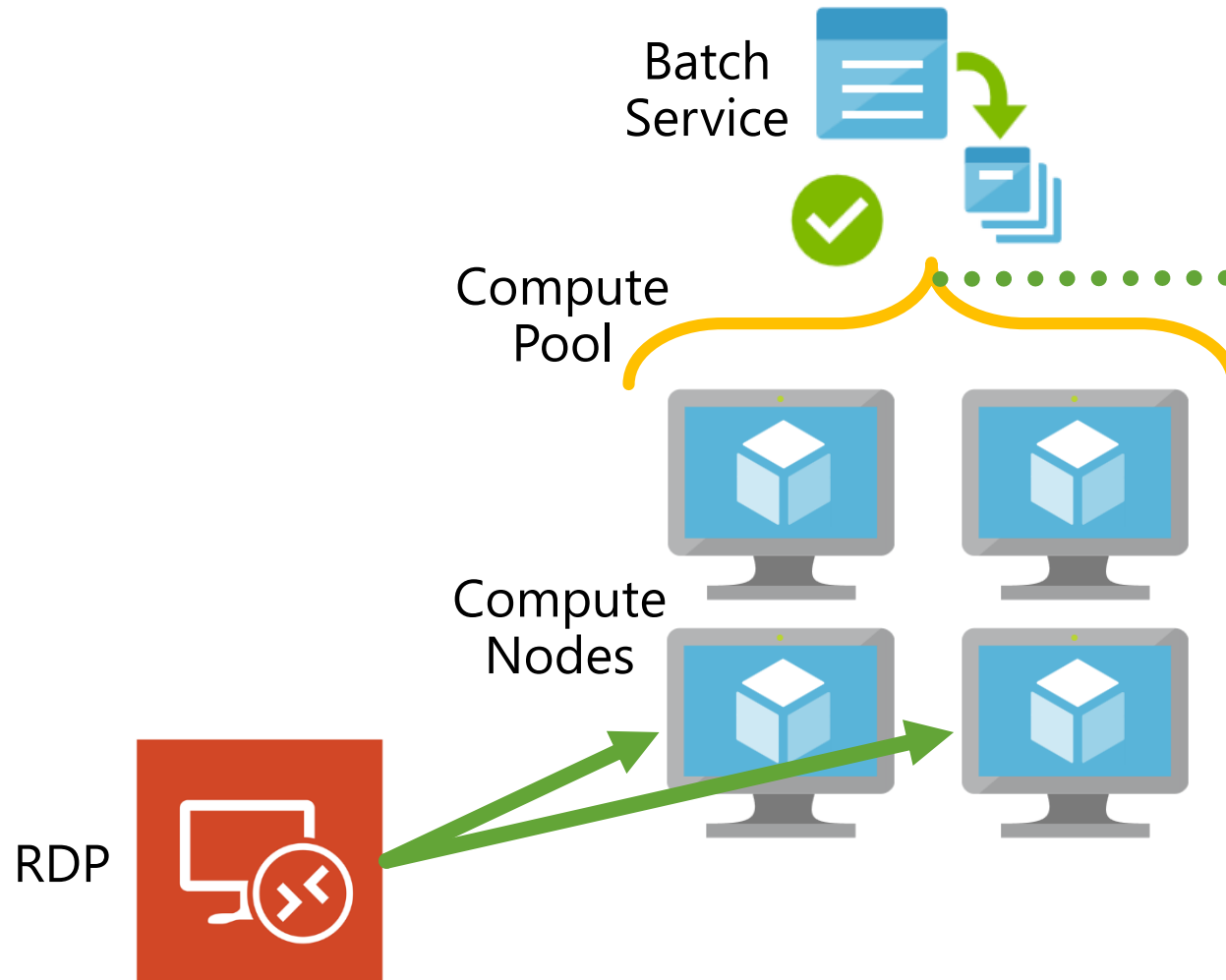
VM node size set per compute pool:

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
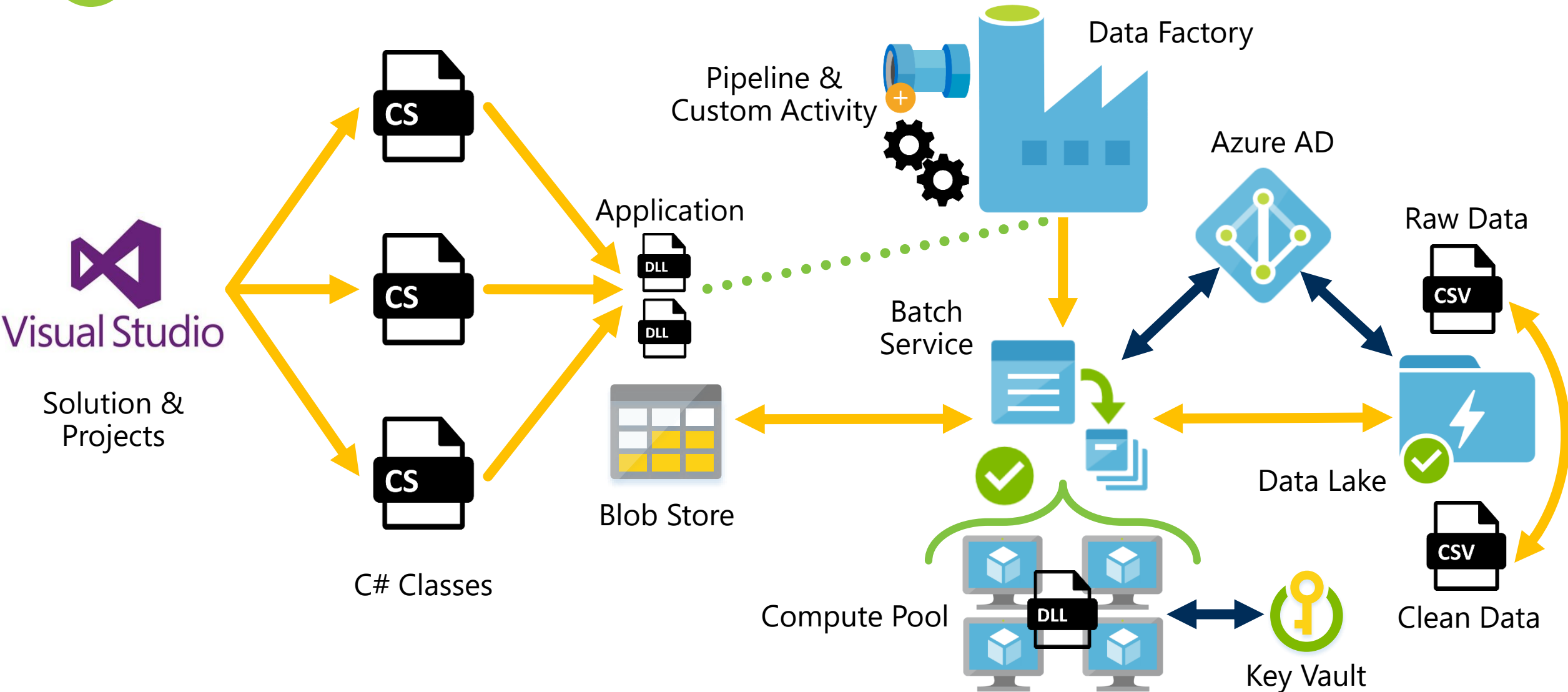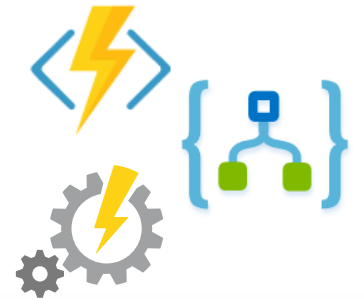
ADF Extensibility Continued

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	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

Select API method...
GET
POST
PUT

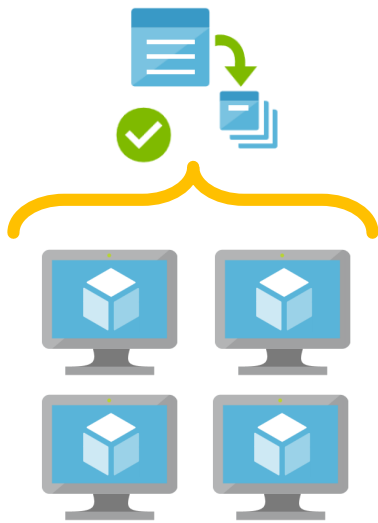| General	Settings ²	Parameters	Advanced

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

◢ Parameterizable properties ⓘ

NAME	VALUE
url	Value
body	Value
Timeout	Value
Retry	Value

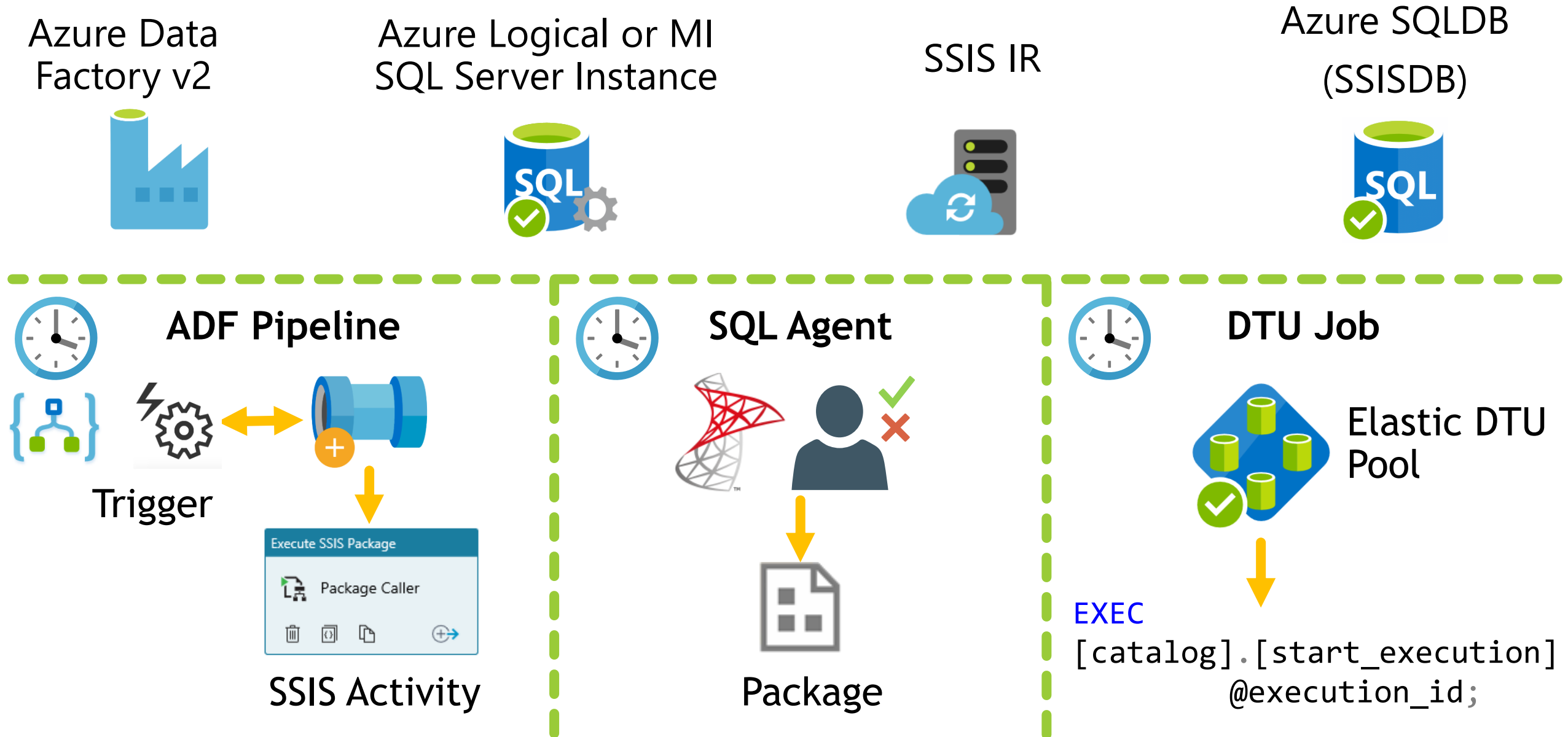# ADF Extensibility Continued

1 **Custom Activities**

2 **Rest API Calls**

3 **SSIS** – Packages with Control Flows and Data Flows

ADF SSIS IR
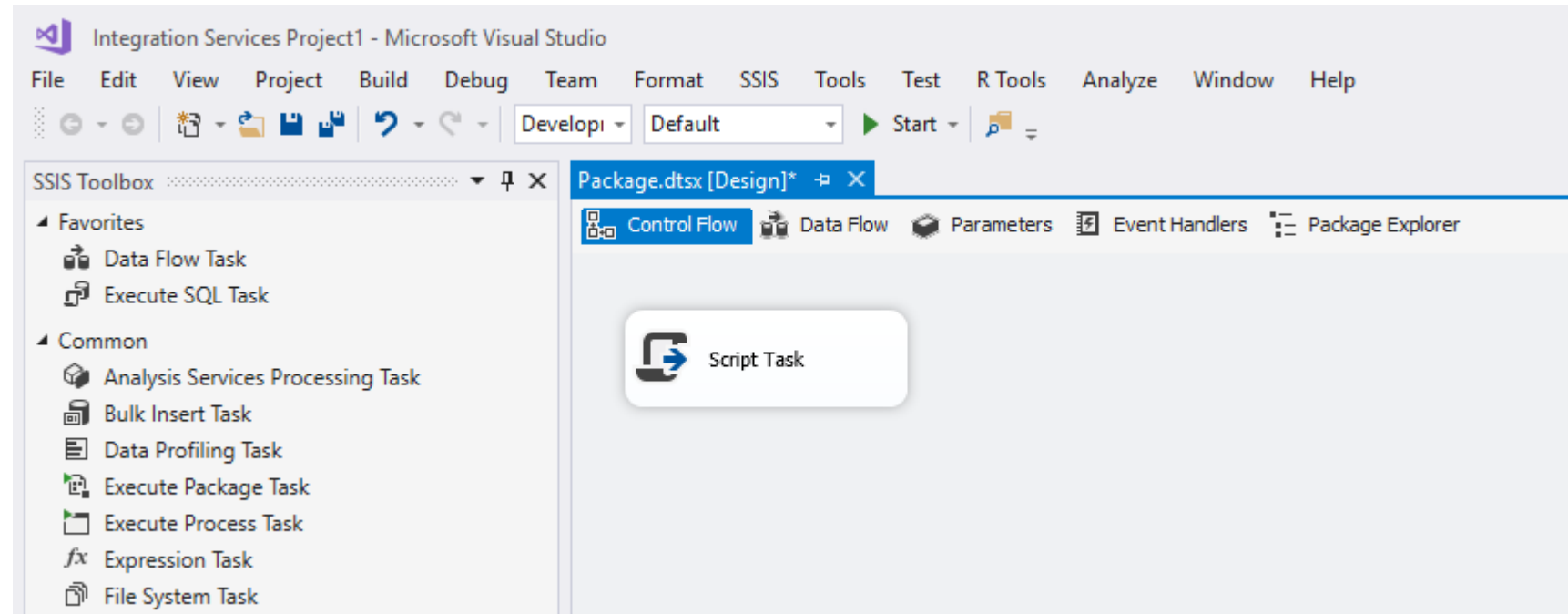
How do we schedule an SSIS Package in Azure?

Azure Data Factory v2

Azure Logical or MI SQL Server Instance

SSIS IR

Azure SQLDB (SSISDB)

ADF Pipeline

Trigger

Execute SSIS Package

Package Caller

SSIS Activity

SQL Agent

Package

DTU Job

Elastic DTU Pool

EXEC
[catalog].[start_execution]
@execution_id;

Agenda for the Day

Module 1

Microsoft Azure

Module 2

Storage
Uploading Data
Data Lake

Module 3

Real-time Data
Streaming
Power BI

Module 4

U-SQL - Data
Transformation
Basics

Module 5

USQL - Advanced
Analytics
Cognitive Services

Module 6

Data Factory
Orchestration
Dynamic Pipelines

Module 7

Data Presentation
& Consumption
Power BI Models

Module 8

Other Services
Q&A