

Data Factory



An Introduction to  Azure Control Flows & Data Flows

Paul Andrew | Principal Consultant & Solution Architect



Gold Cloud Platform
Gold Data Analytics
Gold Data Platform
Gold DevOps





<https://github.com/mrpaulandrew>

CommunityEvents

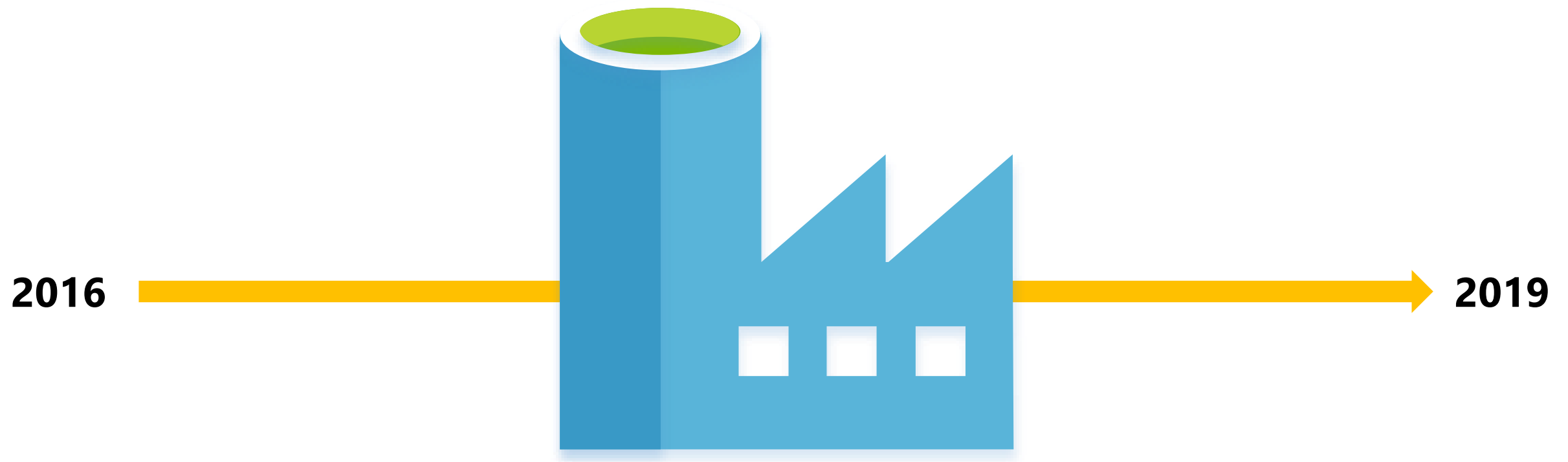
Demo code, content and slides from various community events.

● C++

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

Background

<rant>



Azure Data Factory is SQL Server Integration Services in the cloud.

</rant>

Data Factory



An Introduction to  Azure Control Flows & Data Flows

Data Factory

What is it?

Why use it?

Components/Concepts

Data Factory Extensibility

SSIS, Functions,
Custom Activities

Data Factory in Production

CI/CD
Cost

Data Transformations

Data Flows with
Databricks

Data Factory



An Introduction to  Azure **Control Flows & Data Flows**

Data Factory

What is it?

Why use it?

Components/Concepts

Data Factory Extensibility

SSIS, Functions,
Custom Activities

Data Factory in Production

CI/CD
Cost

Data Transformations

Data Flows with
Databricks

Data Factory



An Introduction to  Azure **Control Flows & Data Flows**

Data Factory

What is it?

Why use it?

Components/Concepts

Data Factory Extensibility

SSIS, Functions,
Custom Activities

Data Factory in Production

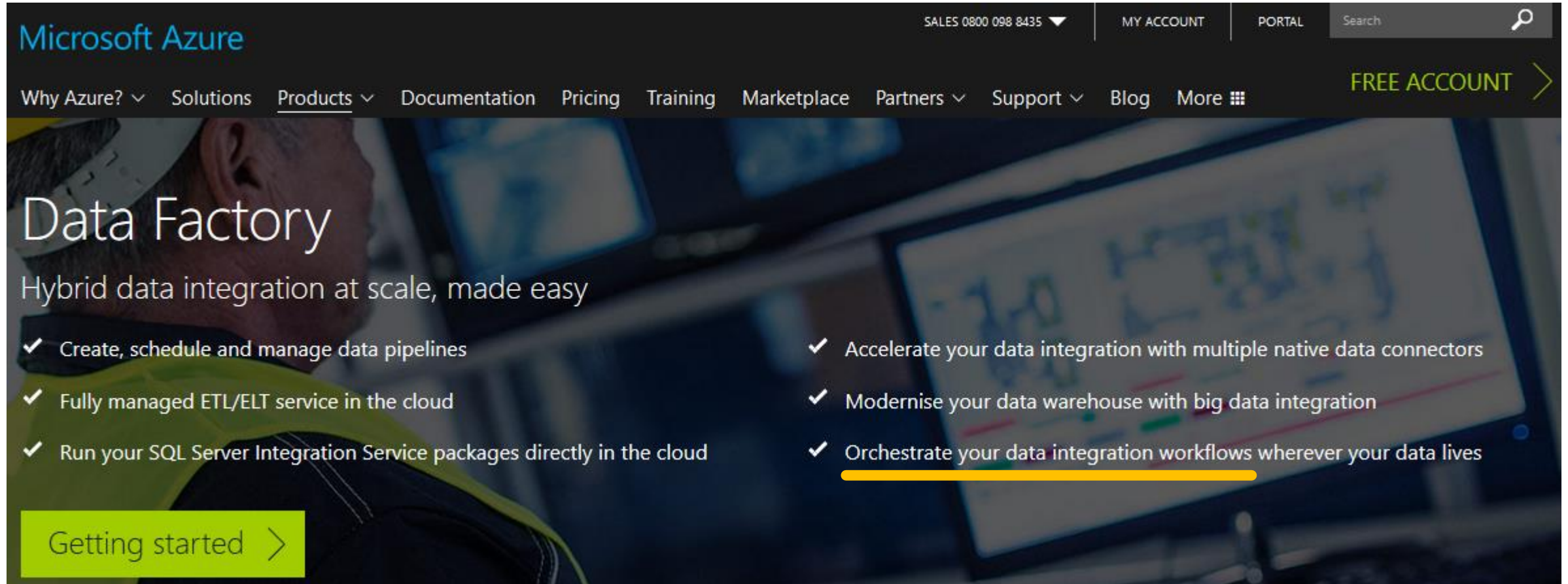
CI/CD
Cost

Data Transformations

Data Flows with
Databricks

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 >

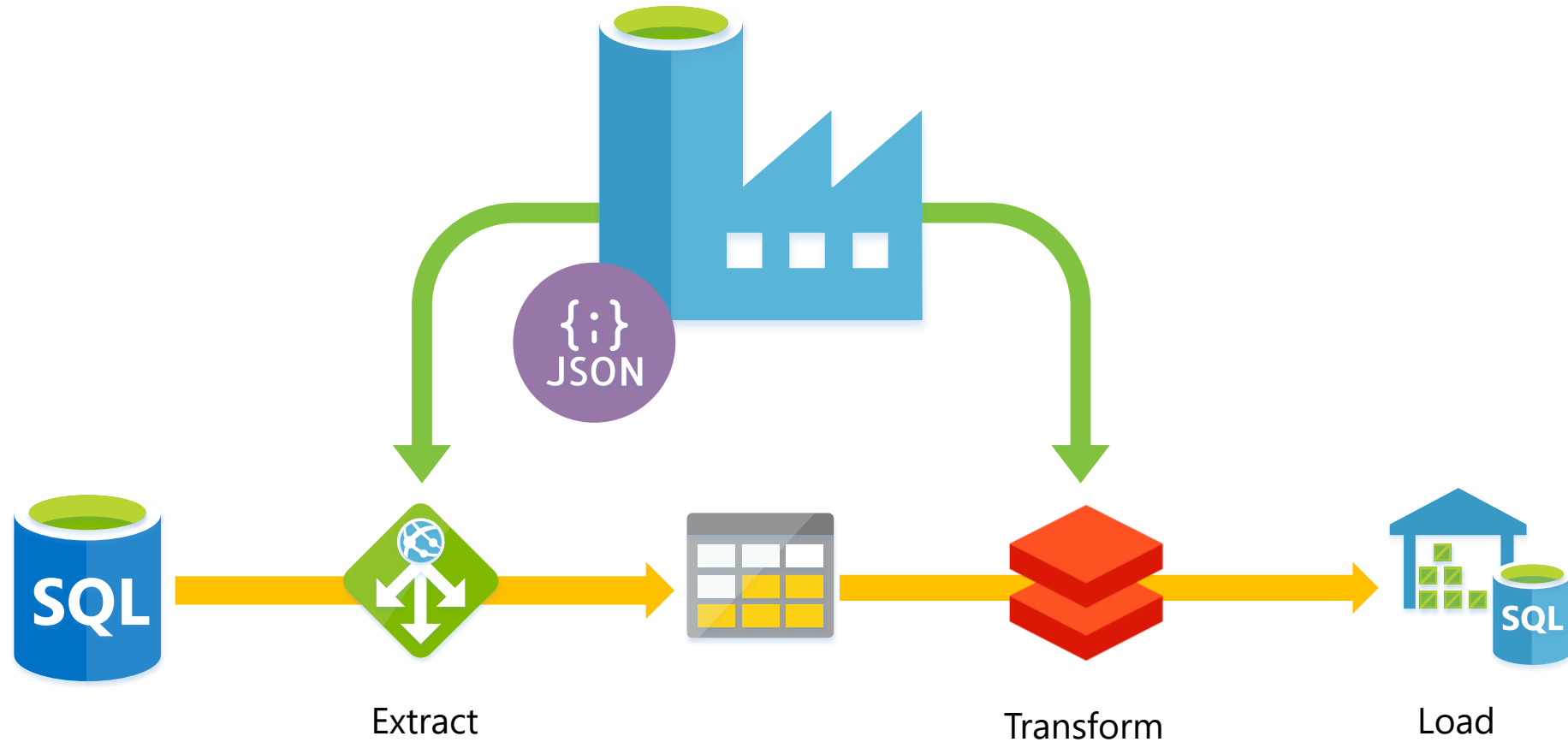
Data Factory

Hybrid data integration at scale, made easy

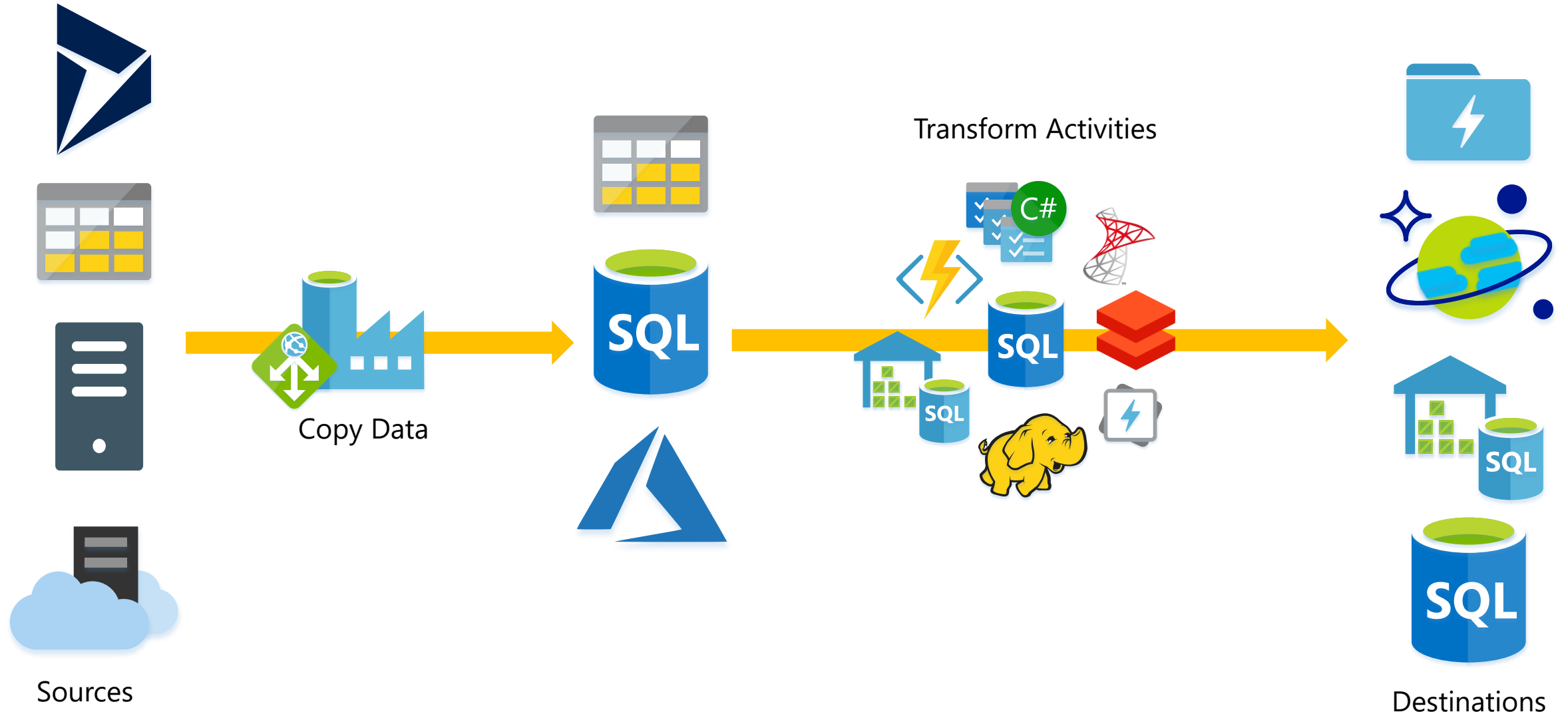
- ✓ Create, schedule and manage data pipelines
- ✓ Fully managed ETL/ELT service in the cloud
- ✓ Run your SQL Server Integration Service packages directly in the cloud
- ✓ Accelerate your data integration with multiple native data connectors
- ✓ Modernise your data warehouse with big data integration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

Getting started >

What is Azure Data Factory?

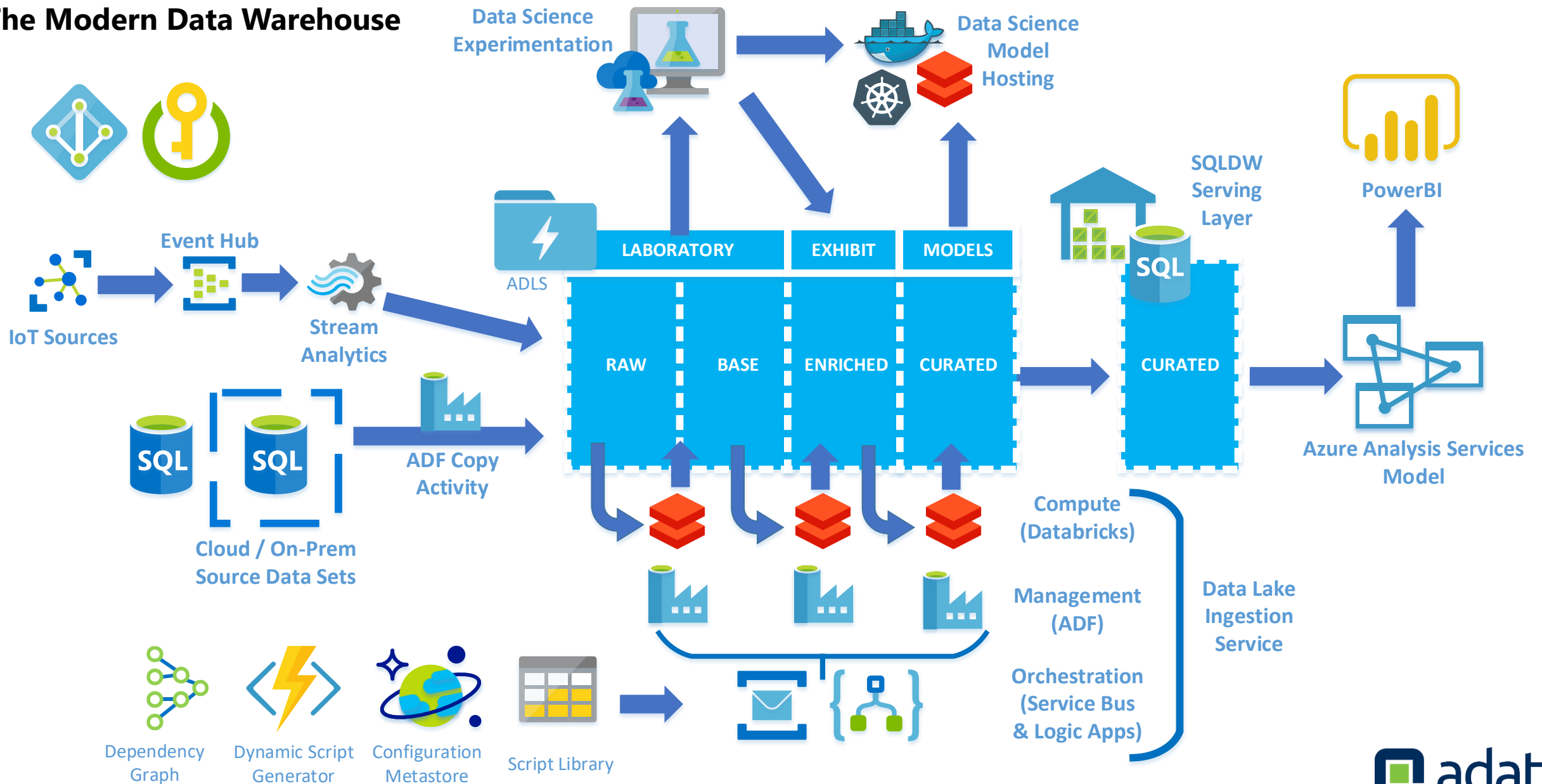


What does Azure Data Factory do?

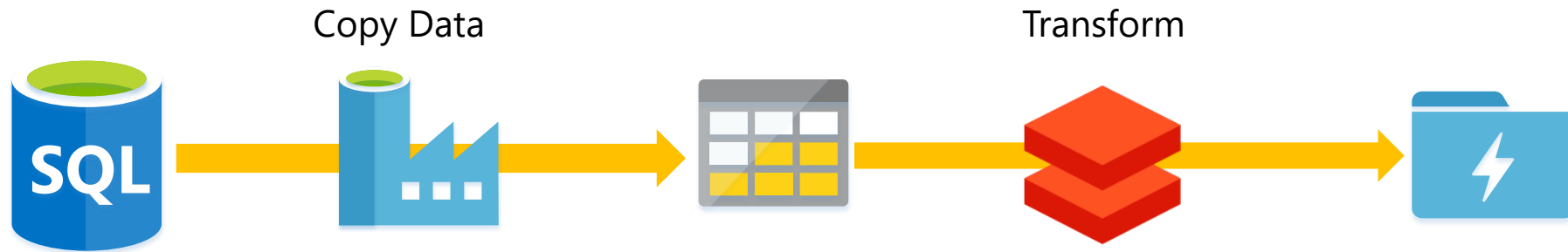


Why use Azure Data Factory?

The Modern Data Warehouse



























































































Data Factory Components



1 Linked Services – How do I connect?

Like the SSIS Connection Manager!

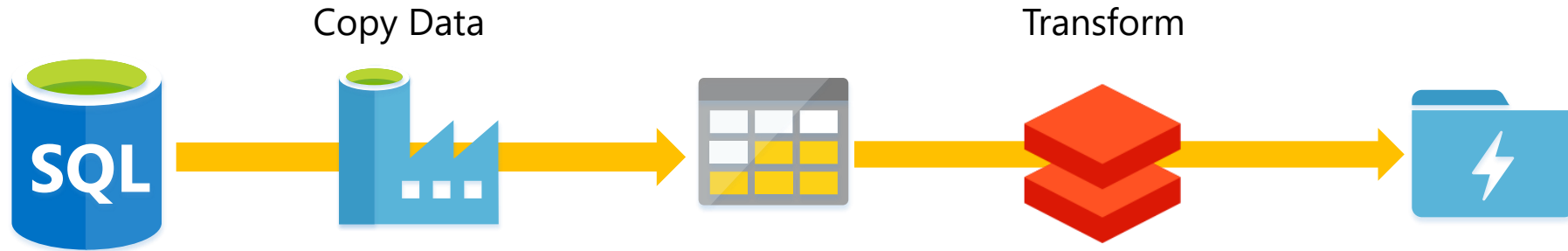


 Amazon Marketplace Web Service (Preview)	 Amazon Redshift	 Amazon S3	 HDFS	 HTTP	 Hive	 Nettezza	 ODBC	 OData	 Azure Batch	 Azure Data Lake Analytics	 Azure Databricks
 Apache Impala (Preview)	 Azure Blob Storage	 Azure Cosmos DB (MongoDB API)	 HubSpot (Preview)	 Informix	 Jira (Preview)	 Office 365 (Preview)	 Oracle	 Oracle Eloqua (Preview)	 Azure Function	 Azure HDInsight	 Azure ML
 Azure Cosmos DB (SQL API)	 Azure Data Explorer (Kusto)	 Azure Data Lake Storage Gen1	 Magento (Preview)	 MariaDB	 Marketo (Preview)	 Oracle Responsys (Preview)	 Oracle Service Cloud (Preview)	 Paypal (Preview)	 ServiceNow	 Shopify (Preview)	 Spark
 Azure Data Lake Storage Gen2 (Preview)	 Azure Database for MariaDB	 Azure Database for MySQL	 Microsoft Access	 MongoDB	 MySQL	 Phoenix	 PostgreSQL	 Presto (Preview)	 Square (Preview)	 Sybase	 Teradata
 Azure Database for PostgreSQL	 Azure File Storage	 Azure Key Vault	 DB2	 Drill (Preview)	 Dynamics 365	 QuickBooks (Preview)	 REST	 SAP BW Open Hub	 Vertica	 Web Table	 Xero (Preview)
 Azure SQL Data Warehouse	 Azure SQL Database	 Azure SQL Database Managed Instance	 Dynamics AX (Preview)	 Dynamics CRM	 FTP	 SAP BW via MDX	 SAP Cloud For Customer	 SAP ECC	 Zoho (Preview)		
 Azure Search	 Azure Table Storage	 Cassandra	 File System	 Google AdWords (Preview)	 Google BigQuery	 SAP HANA	 SFTP	 SQL Server			
 Common Data Service for Apps	 Concur (Preview)	 Couchbase (Preview)	 Google Cloud Storage (S3 API)	 Greenplum	 HBase	 Salesforce	 Salesforce Marketing Cloud (Preview)	 Salesforce Service Cloud			



88x linked services supported as 1st class citizens within Azure Data Factory. As of 5th Feb 2019.

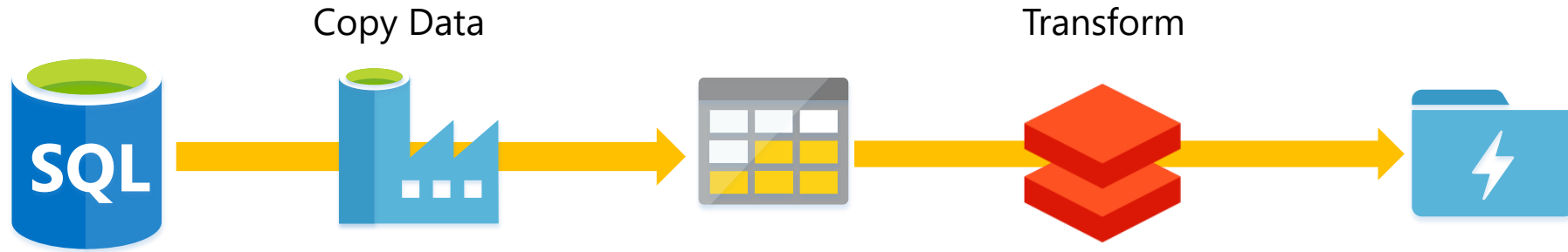
Data Factory Components



1

Linked Services

Data Factory Components



1

Linked Services

2

Data Sets – Where is my data? What format? What file path/table do I need?

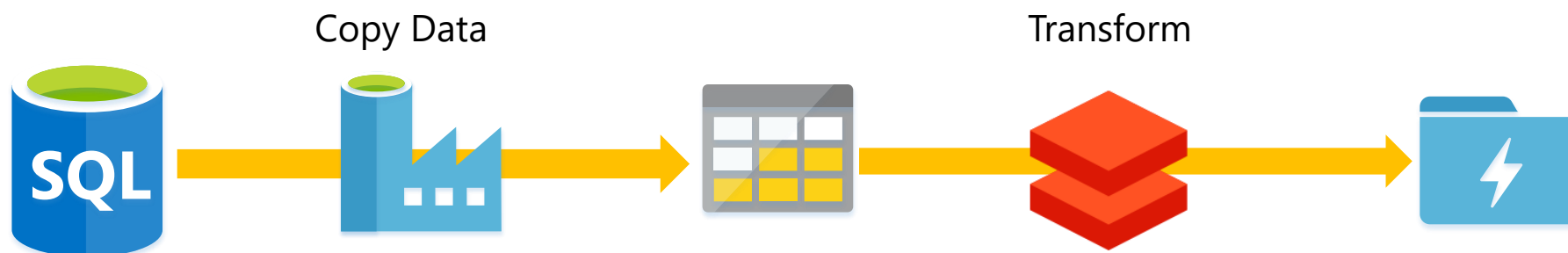


dbo.DimCustomer



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

Data Factory Components



1 Linked Services

2 Data Sets

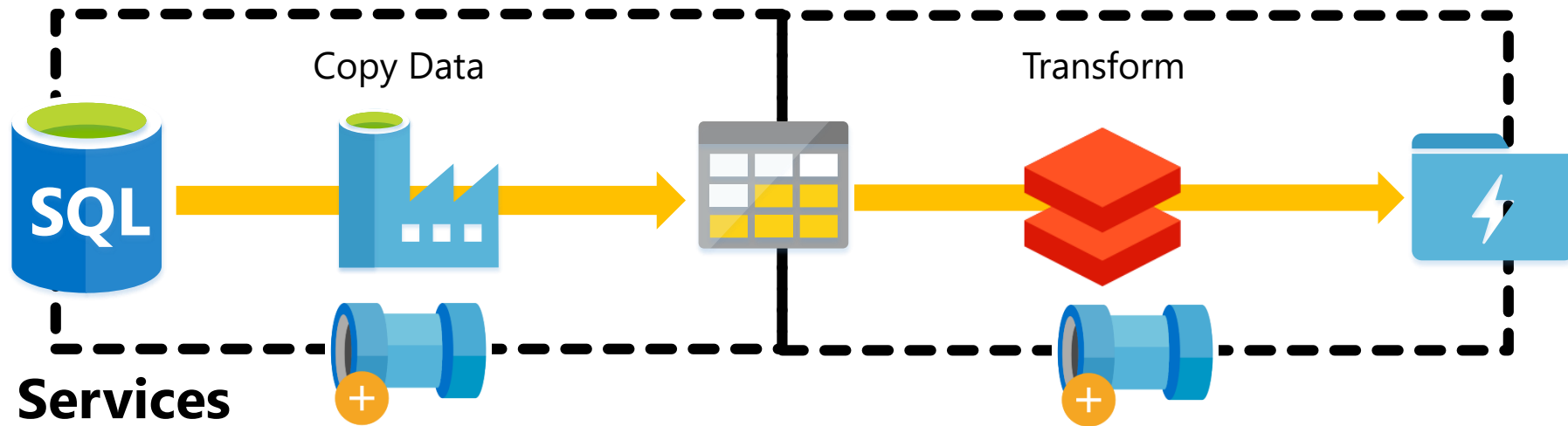
3 Activities – What do we want to happen? With what conditions?



Databricks Notebook Activity

```
notebookPath: /Playground/Playing
baseParameters: Testing
libraries[jar]: dbfs:/lib1.jar
linkedServiceName: BricksOfData01
```

Data Factory Components



1

Linked Services

2

Data Sets

3

Activities

4

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



Sequence Container



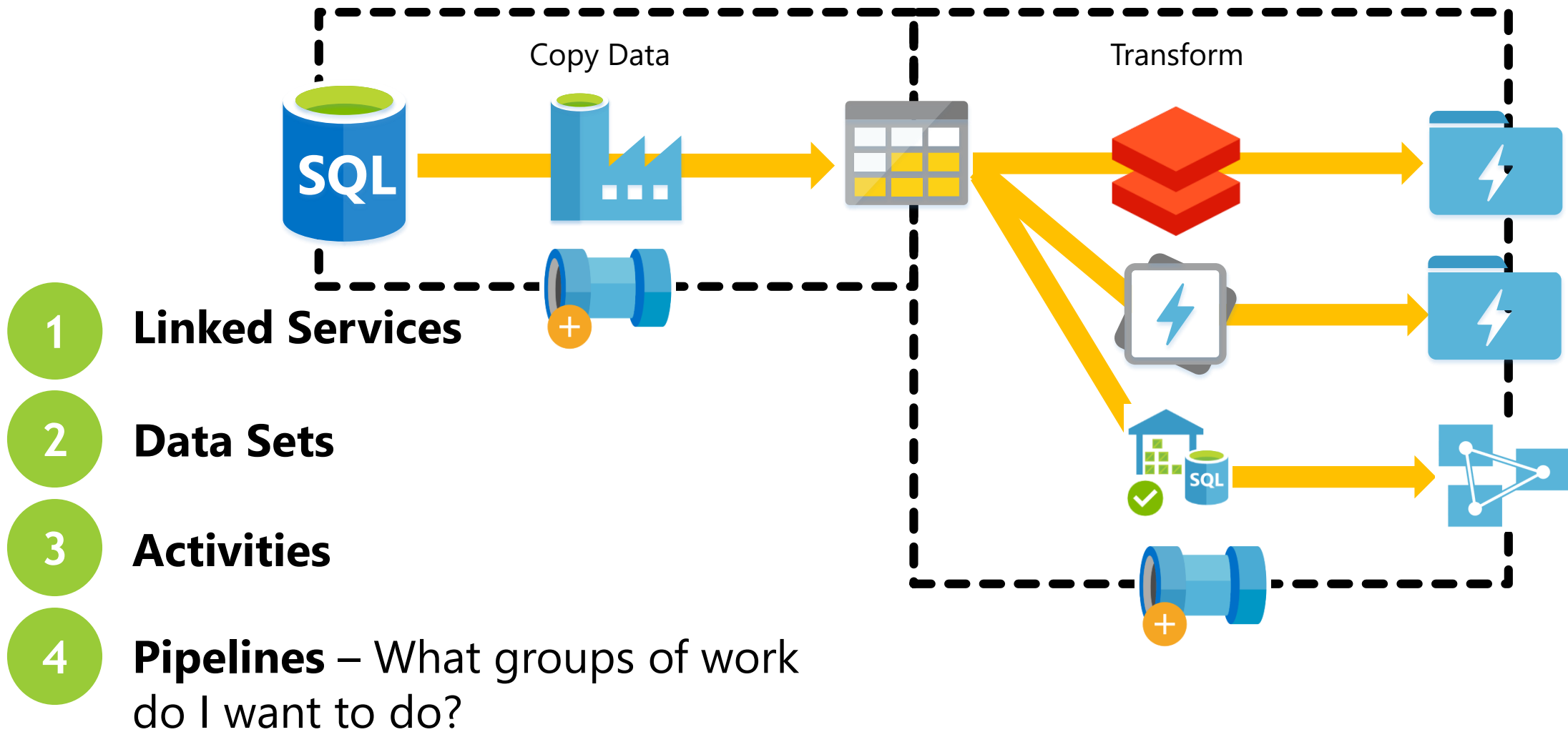
Execute Package Task

Execute Pipeline

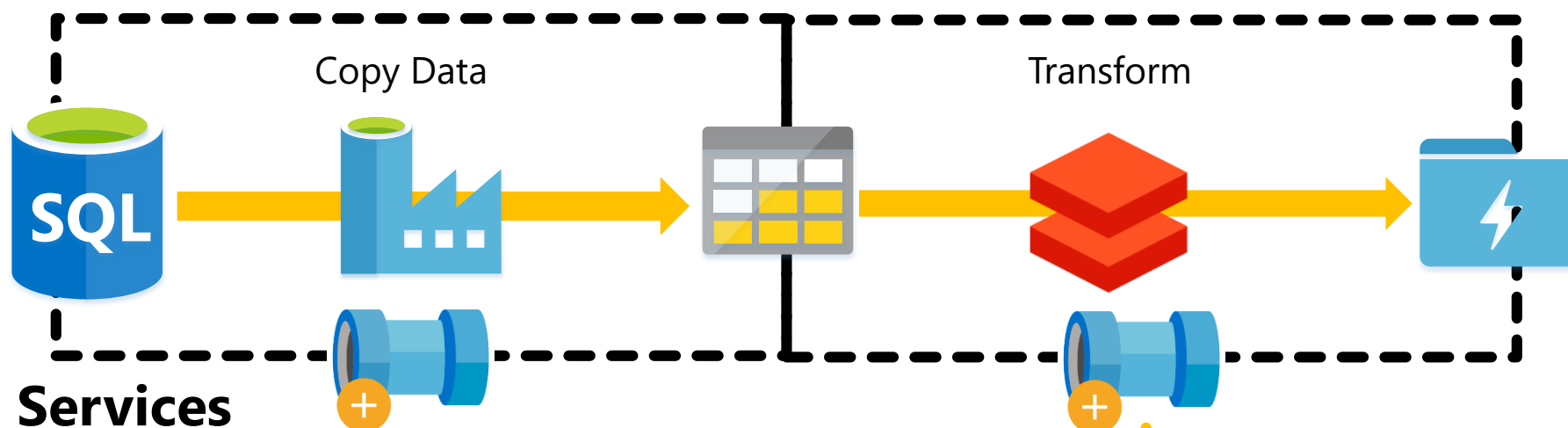


Execute Pipeline1

Data Factory Components



Data Factory Components



1

Linked Services

2

Data Sets

3

Activities

4

Pipelines

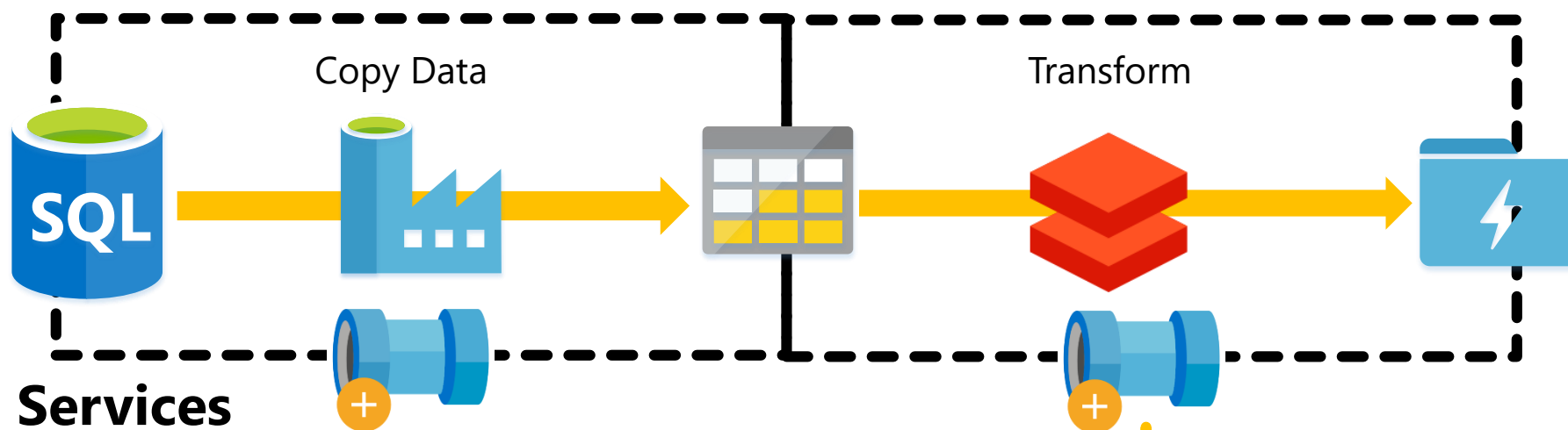
5

Triggers – How are we going to tell our pipeline(s) to execute?



- Manual via UI
- Tumbling Windows
- Scheduled
- Blob File Events
- Logic App Calls

Data Factory Components



1 **Linked Services**

2 **Data Sets**

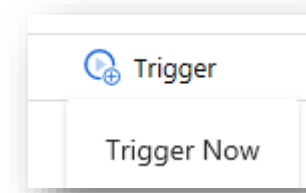
3 **Activities**

4 **Pipelines**

5 **Triggers**

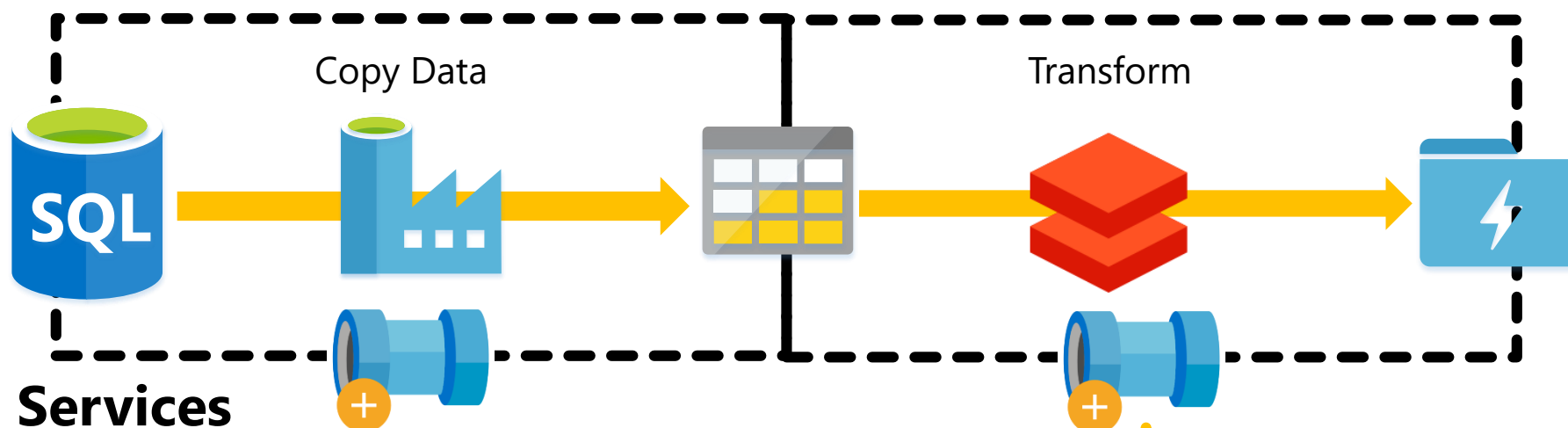


- **Manual**
- Tumbling Windows
- Scheduled
- Blob File Events
- Logic App Calls



```
Invoke-AzureRmDataFactoryV2Pipeline  
-DataFactoryName $dataFactoryName  
-ResourceGroupName $resourceGroupName  
-PipelineName $pipelineName
```

Data Factory Components



1 Linked Services

2 Data Sets

3 Activities

4 Pipelines

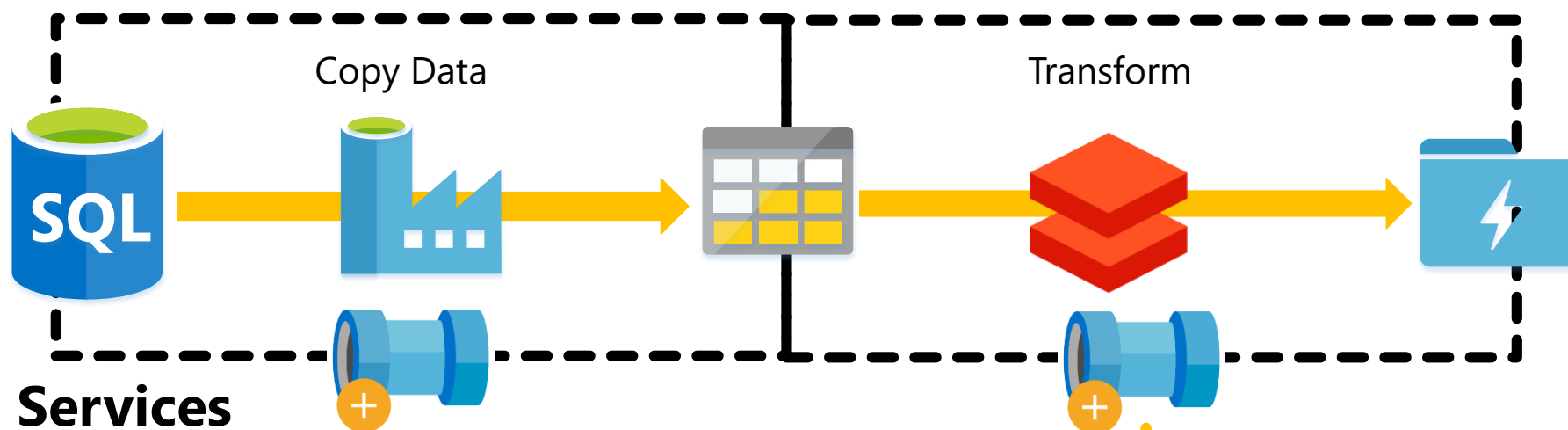
5 Triggers



- Manual via UI
- **Tumbling Windows** - AKA Time Slices
- Scheduled
- Blob File Events
- Logic App Calls



Data Factory Components



1 **Linked Services**

2 **Data Sets**

3 **Activities**

4 **Pipelines**

5 **Triggers**

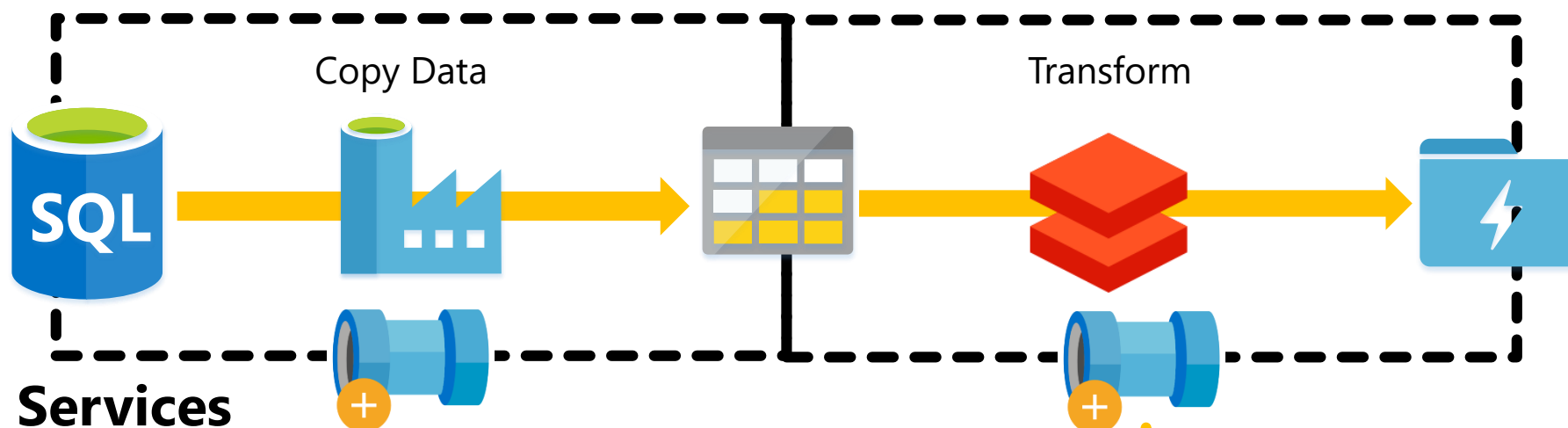


- Manual via UI
- Tumbling Windows
- **Scheduled**
- Blob File Events
- Logic App Calls



- Every 1 minute.
- UTC

Data Factory Components



1 Linked Services

2 Data Sets

3 Activities

4 Pipelines

5 Triggers

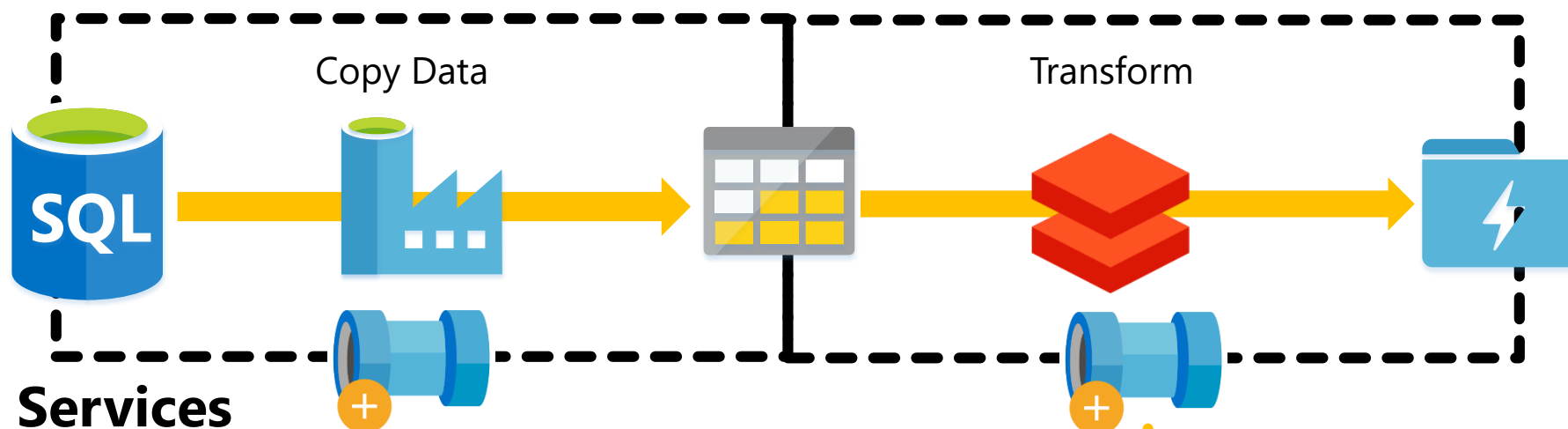


- Manual via UI
- Tumbling Windows
- Scheduled
- **Blob File Events**
- Logic App Calls

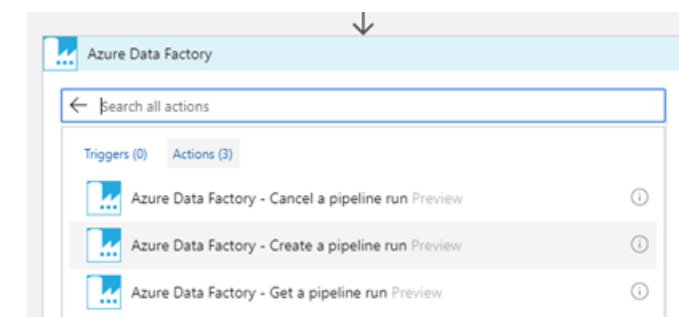
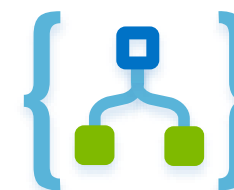


- {Path} Created
- {Path} Deleted

Data Factory Components

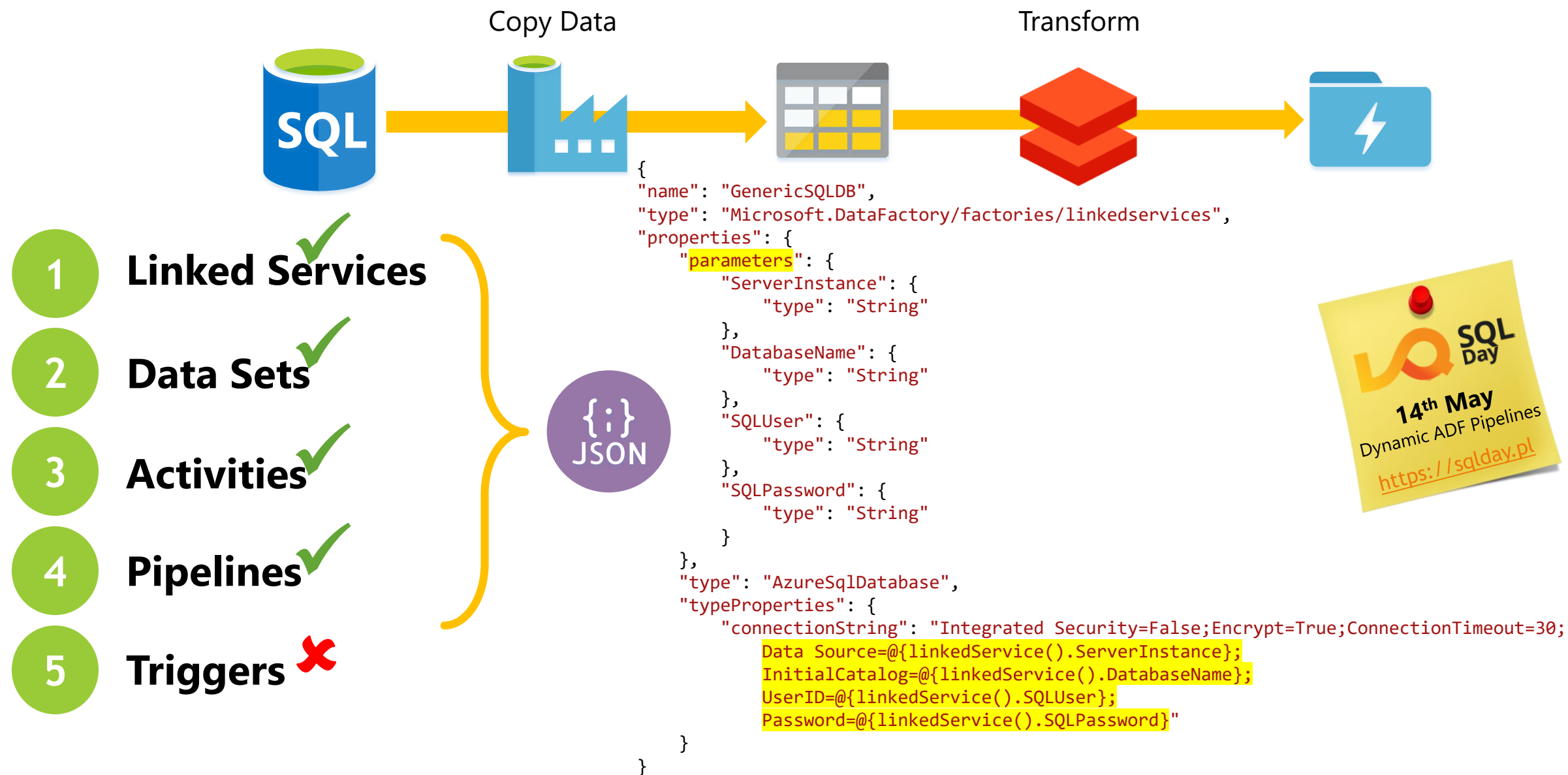


- Manual via UI
- Tumbling Windows
- Scheduled
- Blob File Events
- **Logic App Calls**



<http://blogs.adatis.co.uk/matthow/post/Using-ADF-V2-Activities-in-Logic-Apps>

Data Factory Components – Recap



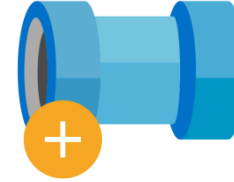
Integration Runtimes



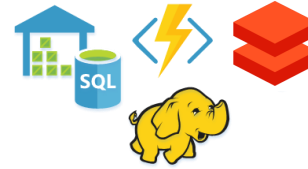
1

Azure
Integration Runtime

Movement Hours



Activity
Orchestration



Flexible Region



2

SSIS
Integration Runtime

SSIS Package
Execution



Specified Region



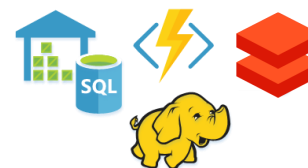
3

Self Hosted
Integration Runtime

Gateway Access



Activity
Orchestration



Virtual Machine



Integration Runtimes



Gateway Access

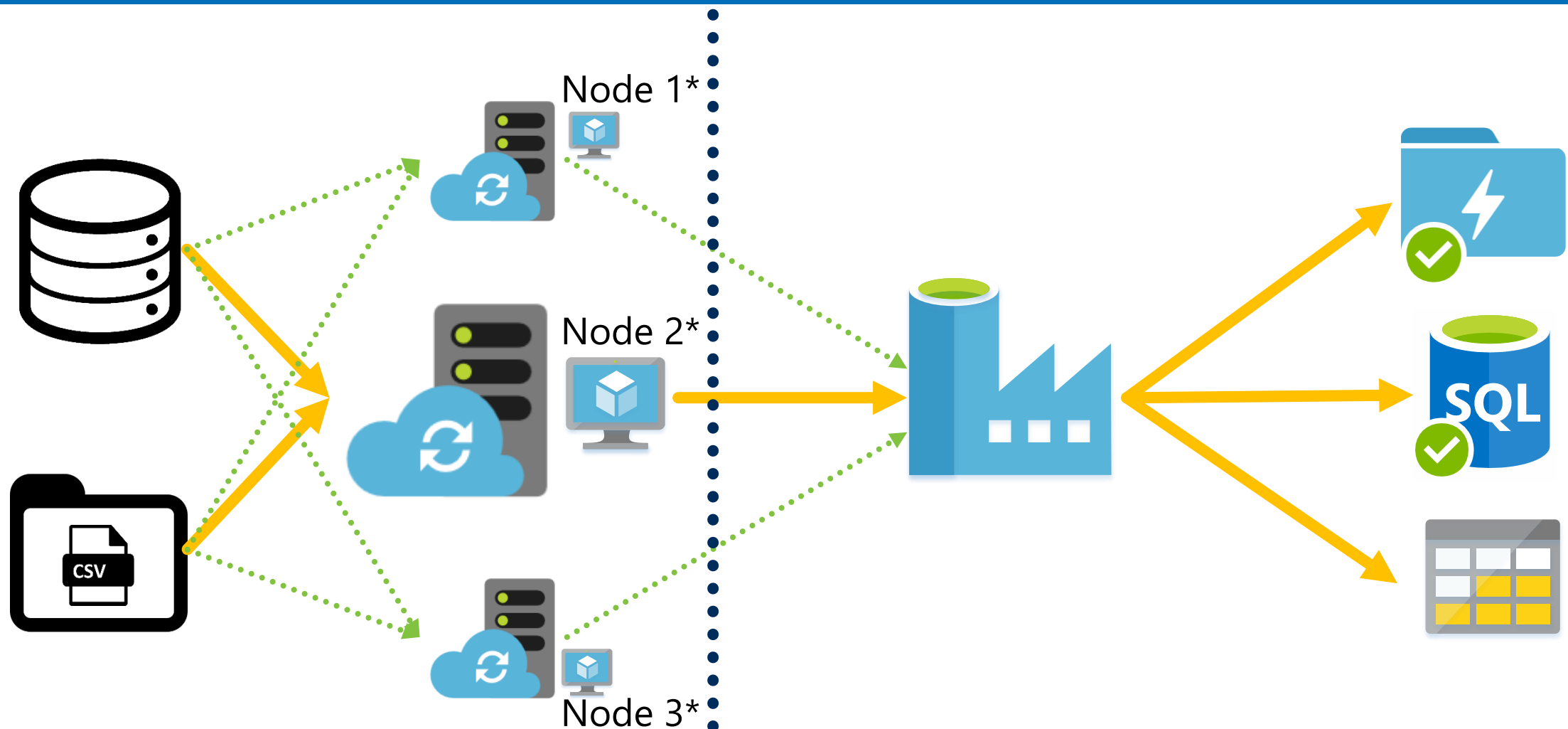
Activity
Orchestration



Virtual Machine



The Hosted Integration Runtime

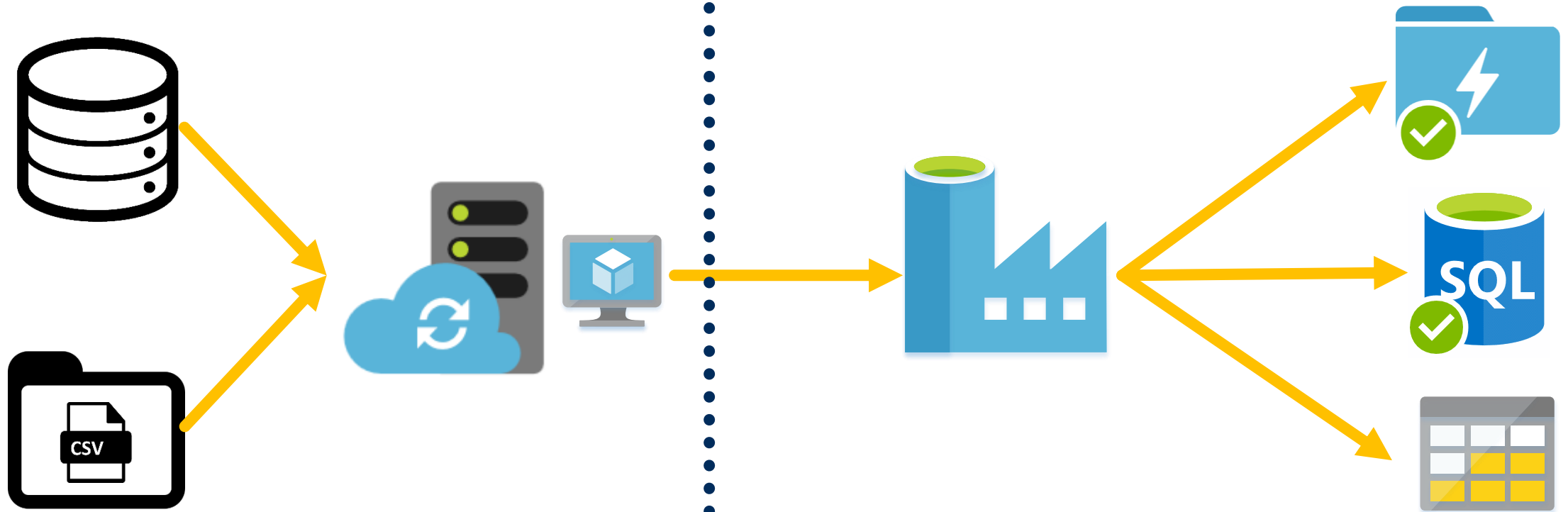


On Premises

Azure

*Failover & Load Balancing

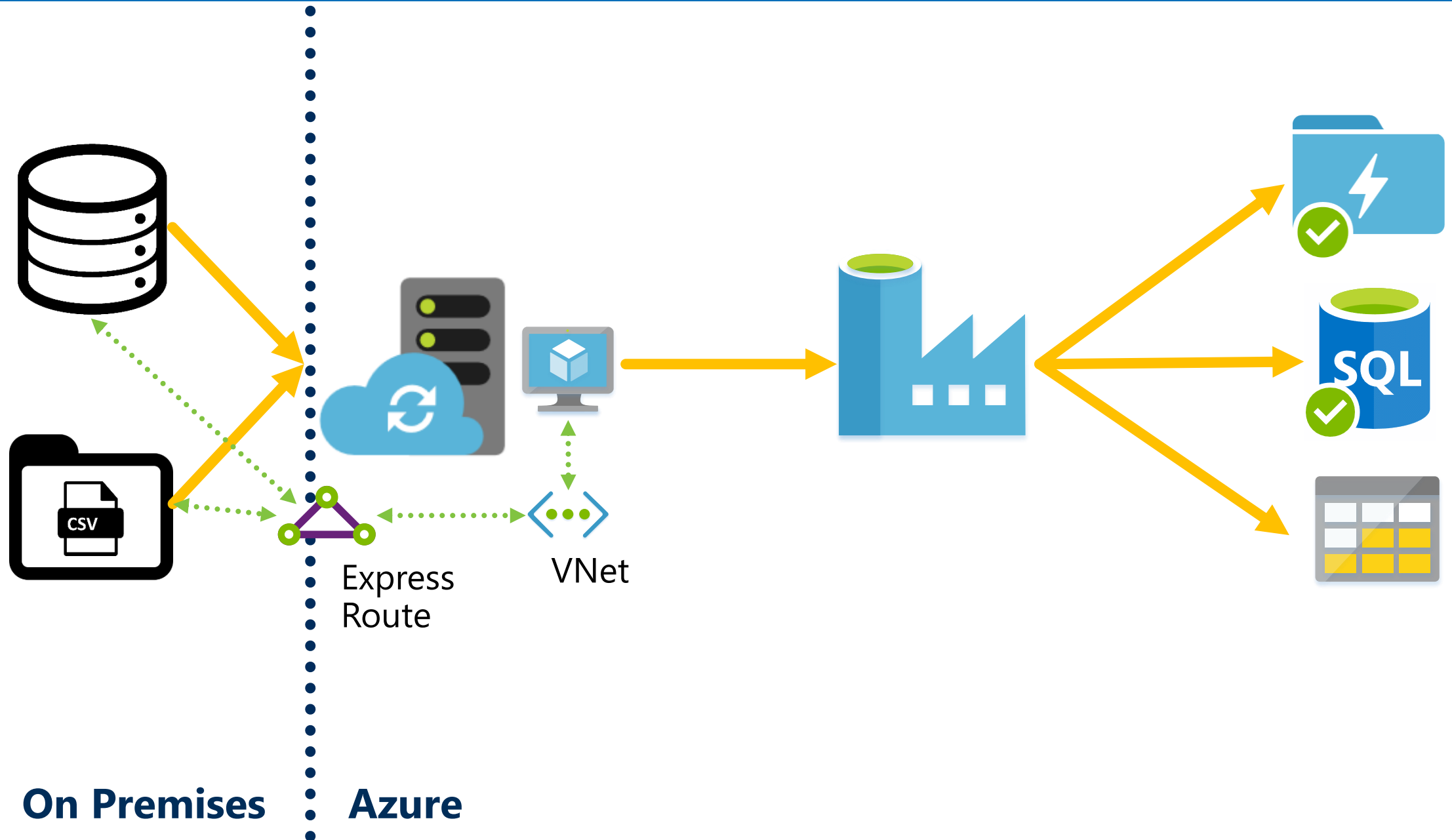
The Hosted Integration Runtime



On Premises

Azure

The Hosted Integration Runtime with Express Route



Data Factory



An Introduction to  Azure **Control Flows & Data Flows**

Data Factory

What is it?

Why use it?

Components/Concepts

Data Factory Extensibility

SSIS, Functions,
Custom Activities

Data Factory in Production

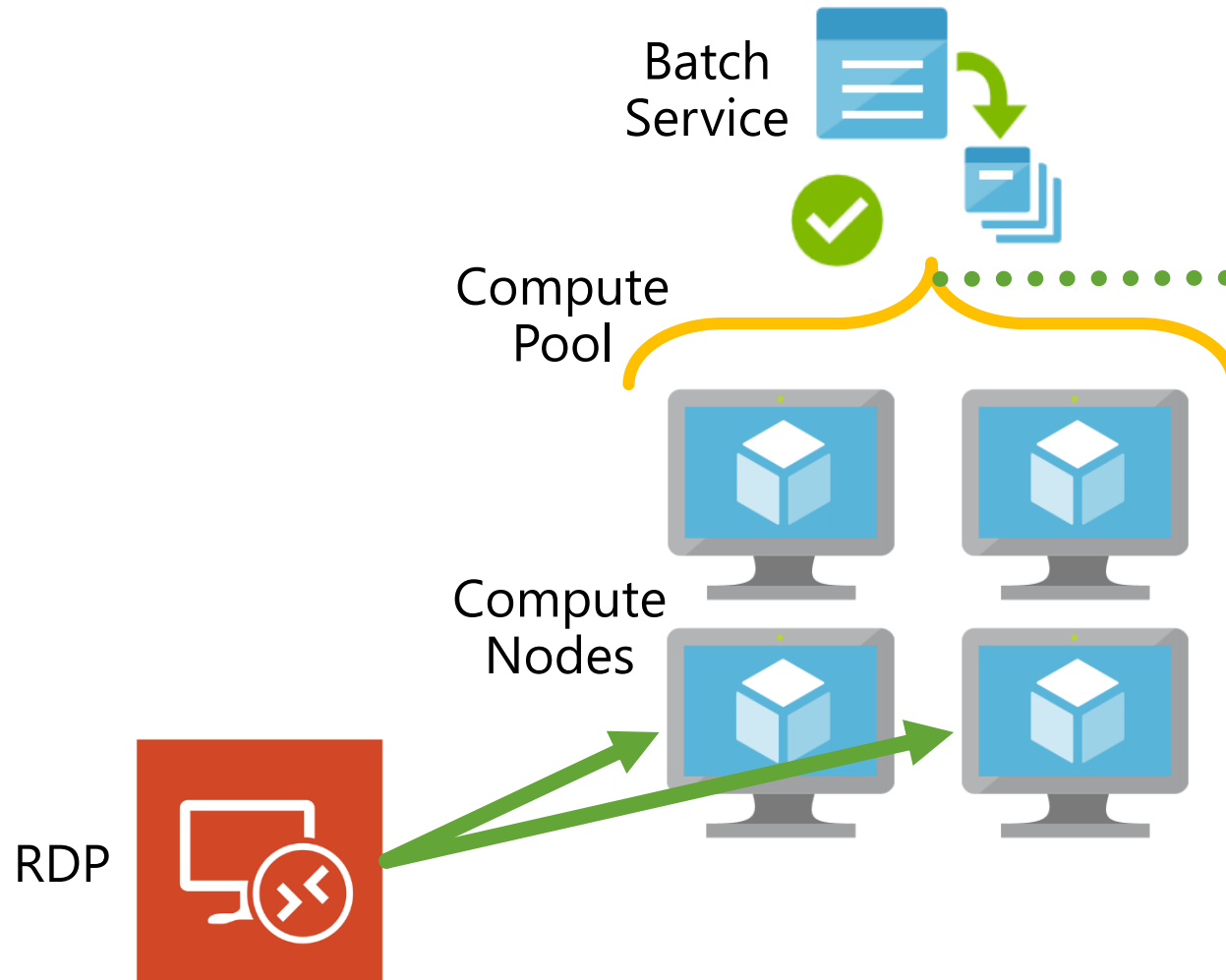
CI/CD
Cost

Coming Soon!

Data Flows with
Data Bricks

1

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



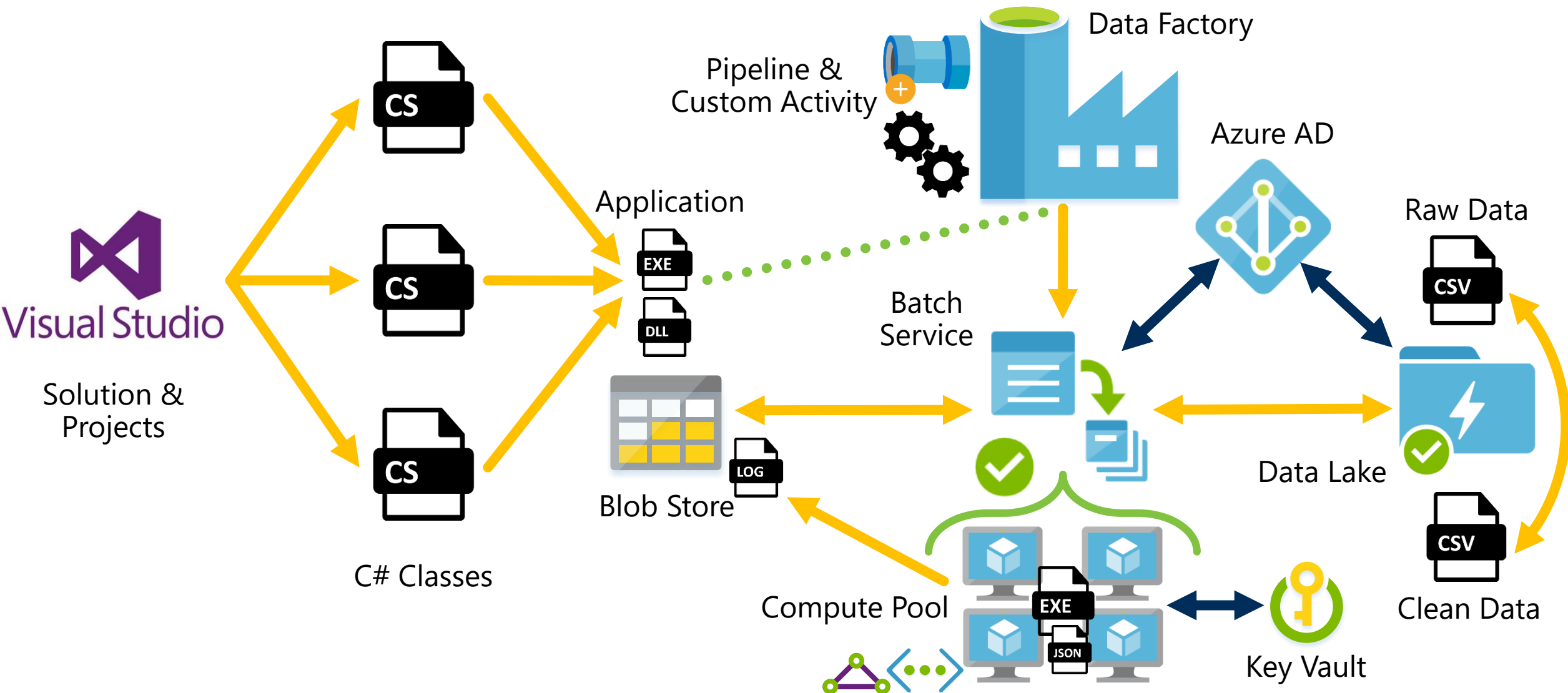
VM node size set per compute pool:

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

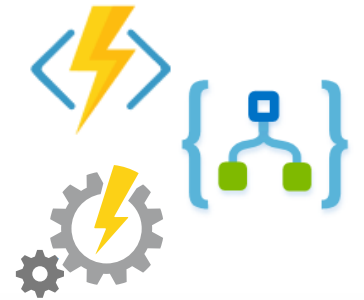


ADF Extensibility Continued

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

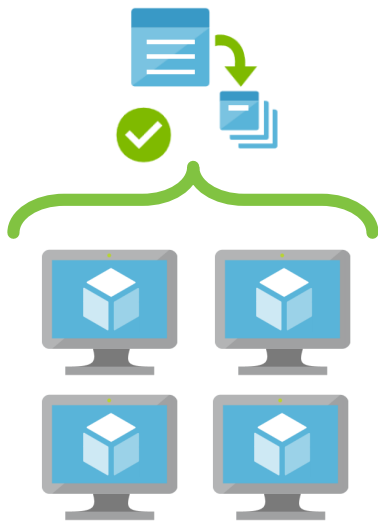
Use [expressions](#), [functions](#) or refer to [system variables](#) in the 'value' column.

Parameterizable properties ⓘ

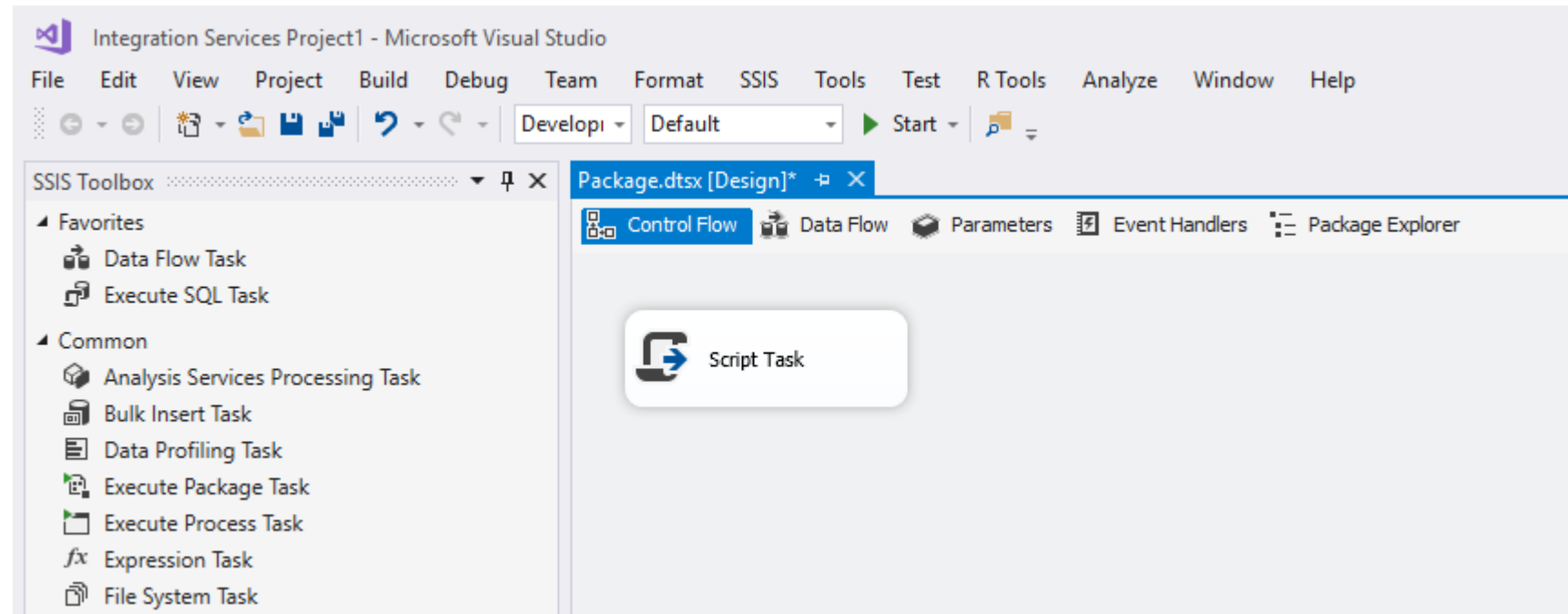
NAME	VALUE
url	<input type="text" value="Value"/>
body	<input type="text" value="Value"/>
Timeout	<input type="text" value="Value"/>
Retry	<input type="text" value="Value"/>

ADF Extensibility Continued

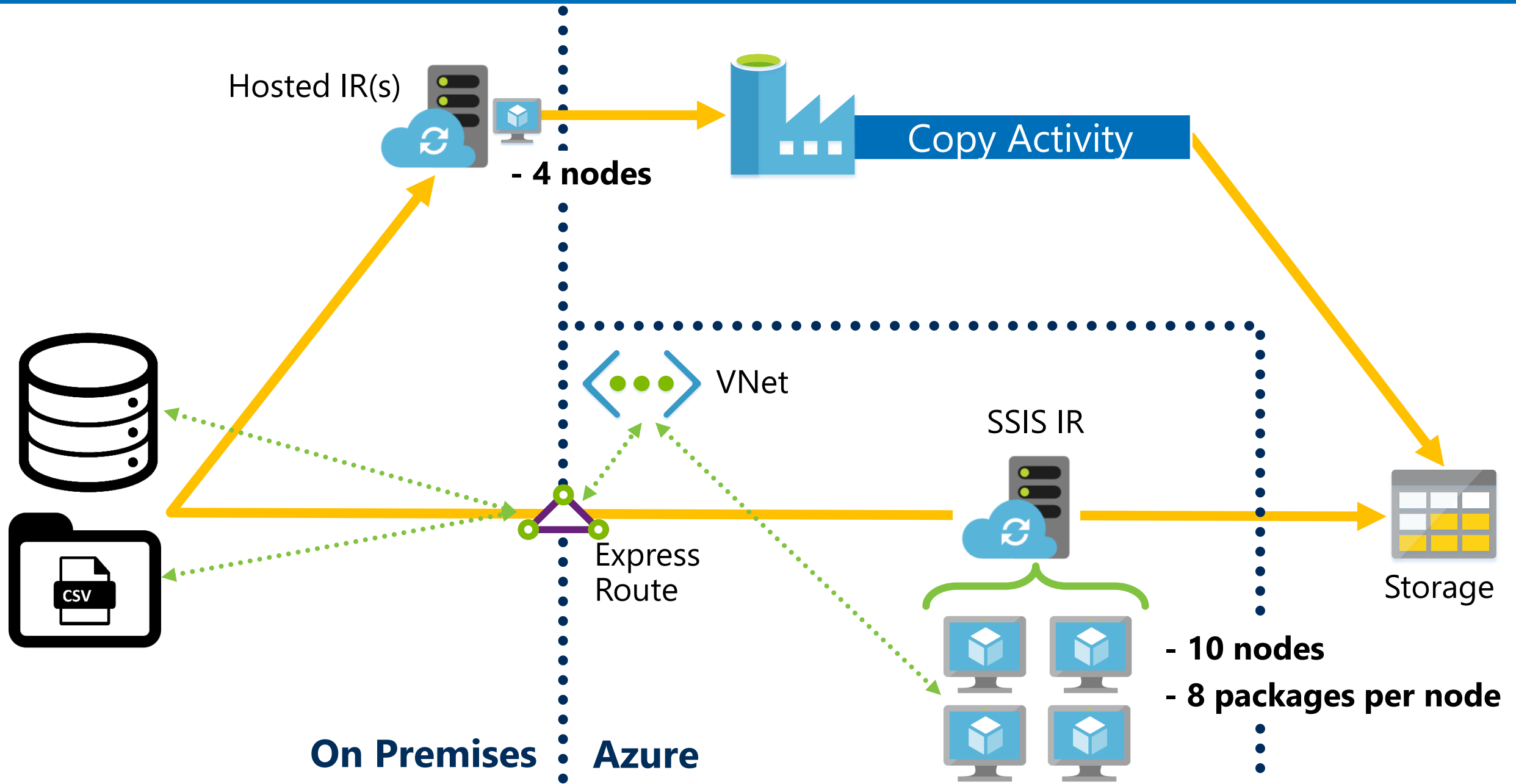
- 1 **Custom Activities**
- 2 **Rest API Calls**
- 3 **SSIS** – Packages with Control Flows and Data Flows



ADF SSIS IR

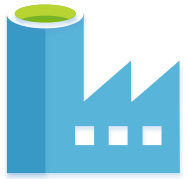


The SSIS IR vs Hosted IR with Express Route



How do we schedule an SSIS Package in Azure?

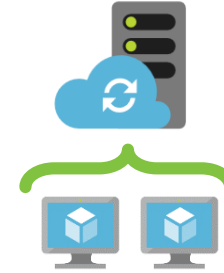
Azure Data Factory



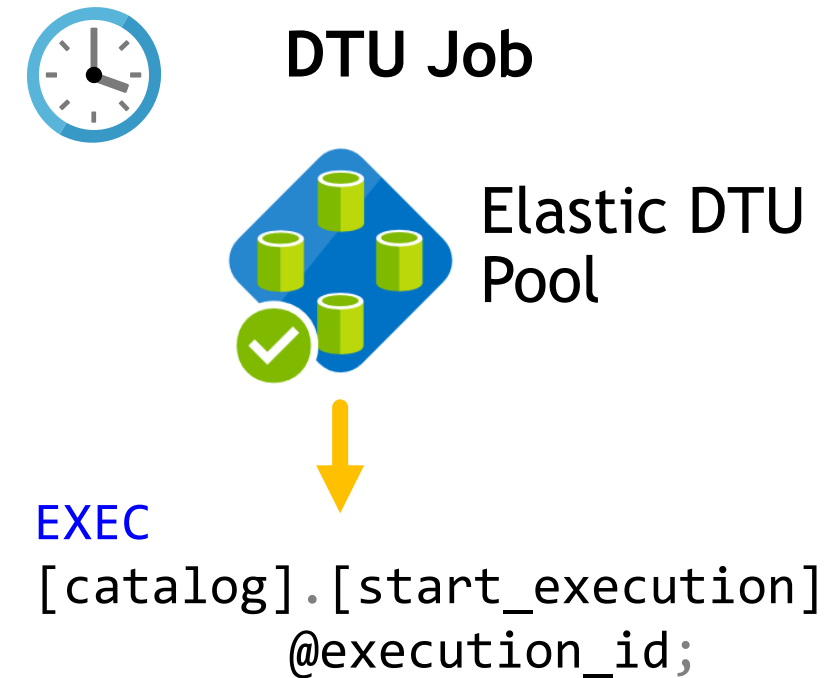
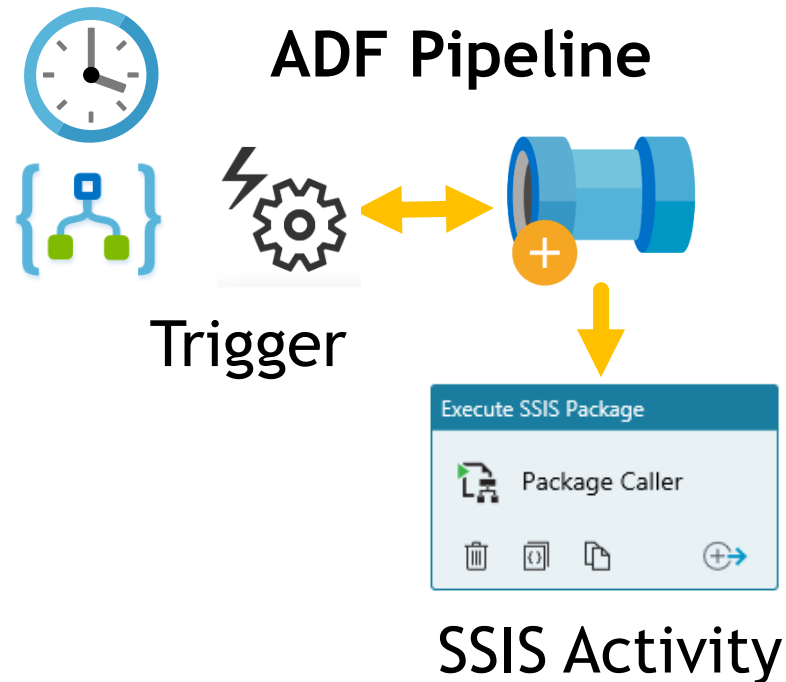
Azure Logical or MI SQL Server Instance



SSIS IR



Azure SQLDB (SSISDB)



Data Factory



An Introduction to  Azure Control Flows & Data Flows

Data Factory

What is it?

Why use it?

Components/Concepts

Data Factory Extensibility

SSIS, Functions,
Custom Activities

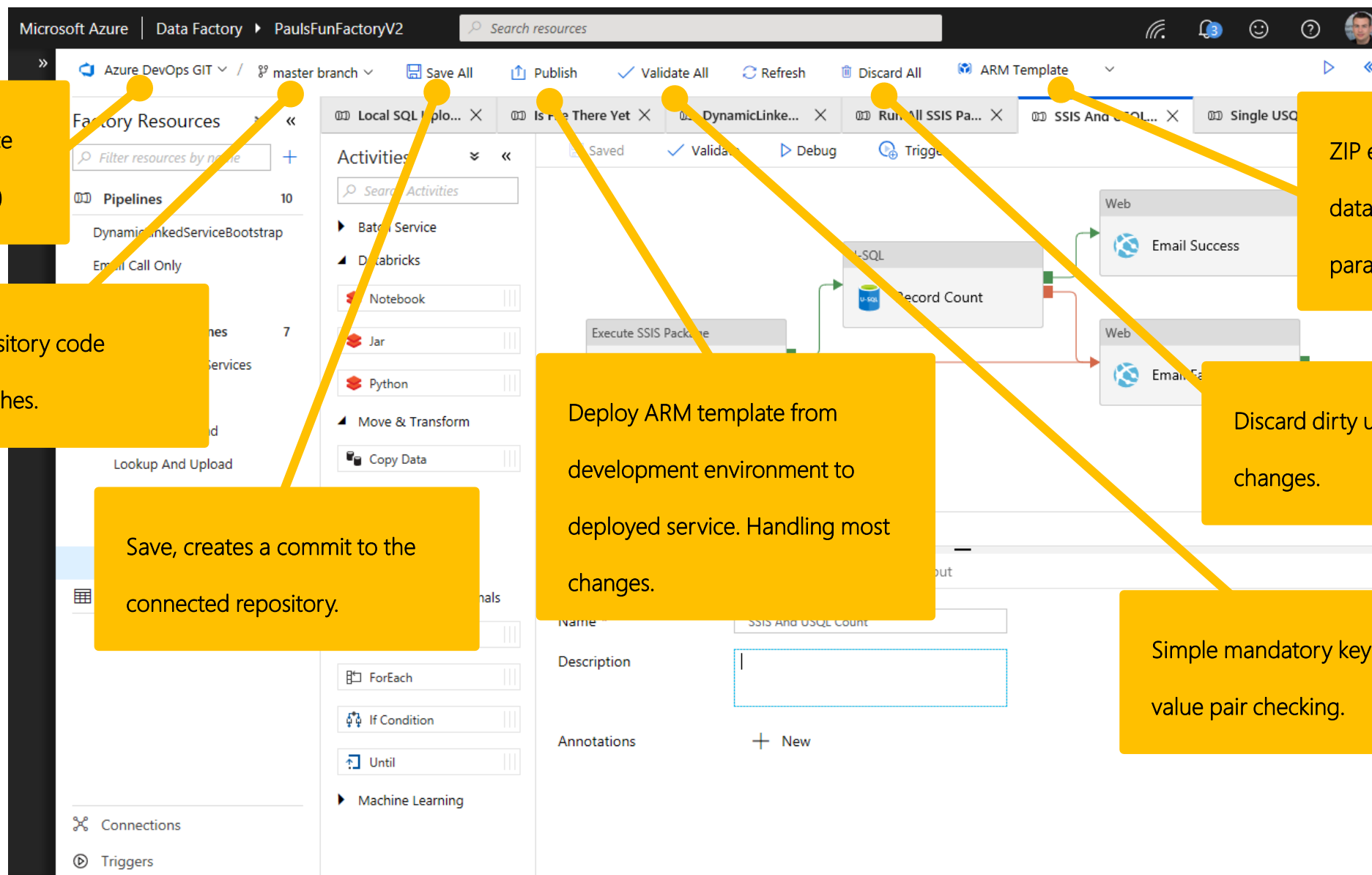
Data Factory in Production

CI/CD
Cost

Data Transformations

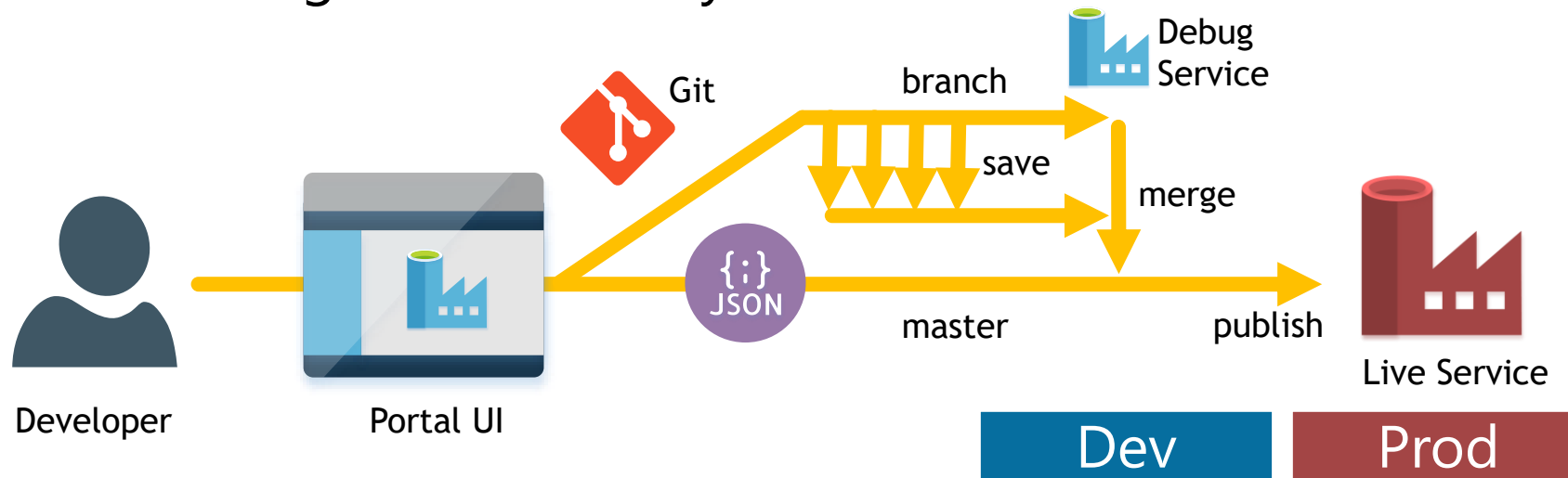
Data Flows with
Data Bricks

Data Factory CI/CD

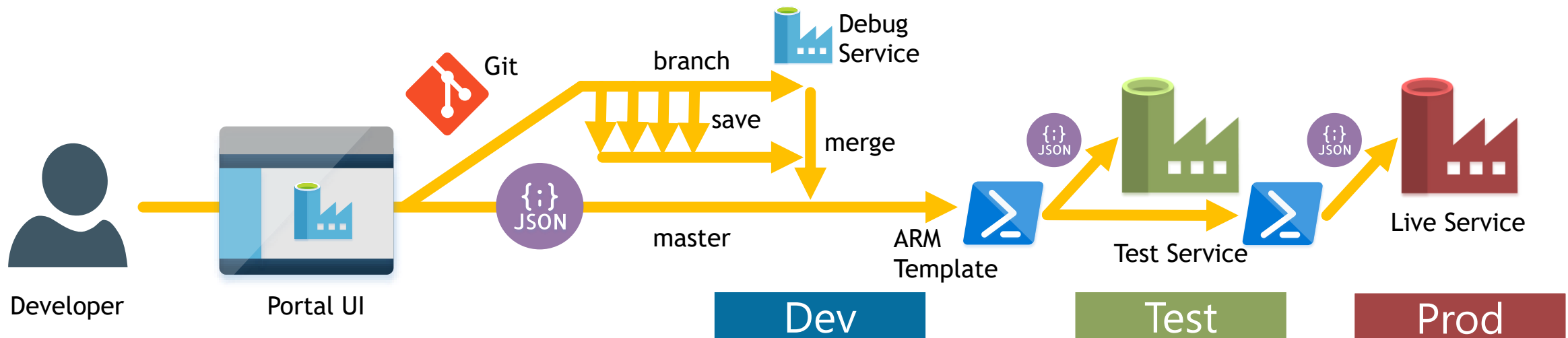


Data Factory CI/CD

Option 1 – Single Data Factory Service

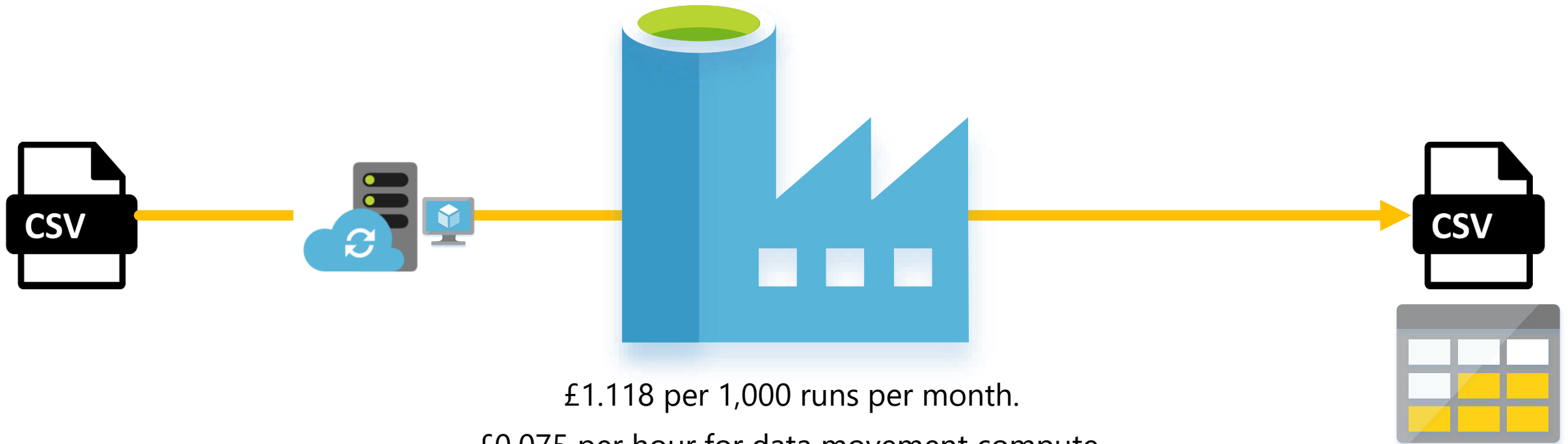


Option 2 – ARM Templates for Multiple Data Factory Services



How much does Azure Data Factory cost?

1x file copied every day to Azure from on premises for 1 year using a hosted IR taking an hour or less to copy.



$$(\pounds 1.118 \times 12) + (\pounds 0.075 \times 365)$$

$$\pounds 13.42 + \pounds 27.38 = \underline{\pounds 40.80}$$

Data Factory



An Introduction to  Azure **Control Flows & Data Flows**

Data Factory

What is it?

Why use it?

Components/Concepts

Data Factory Extensibility

SSIS, Functions,
Custom Activities

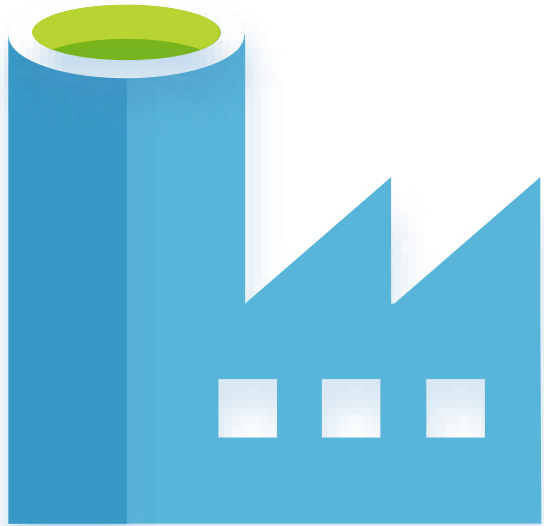
Data Factory in Production

CI/CD
Cost

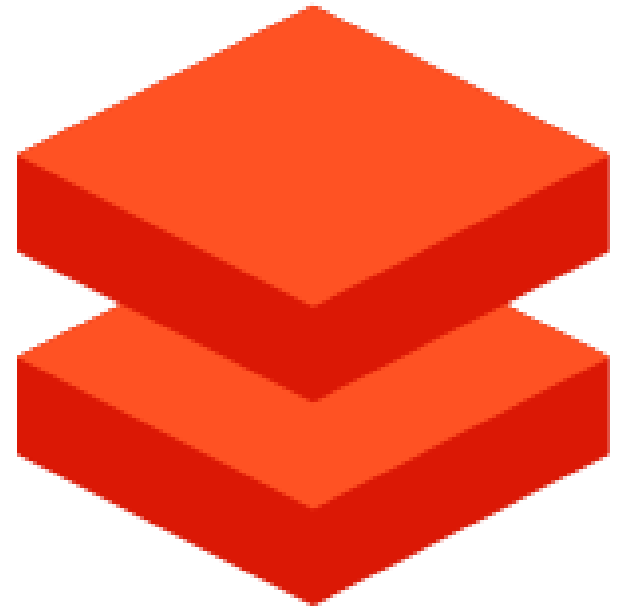
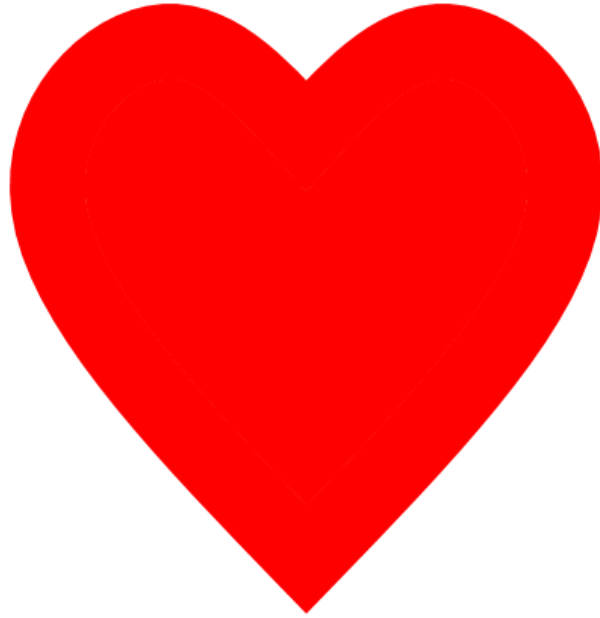
Data Transformations

Data Flows with
Databricks

Data Factory meets Databricks

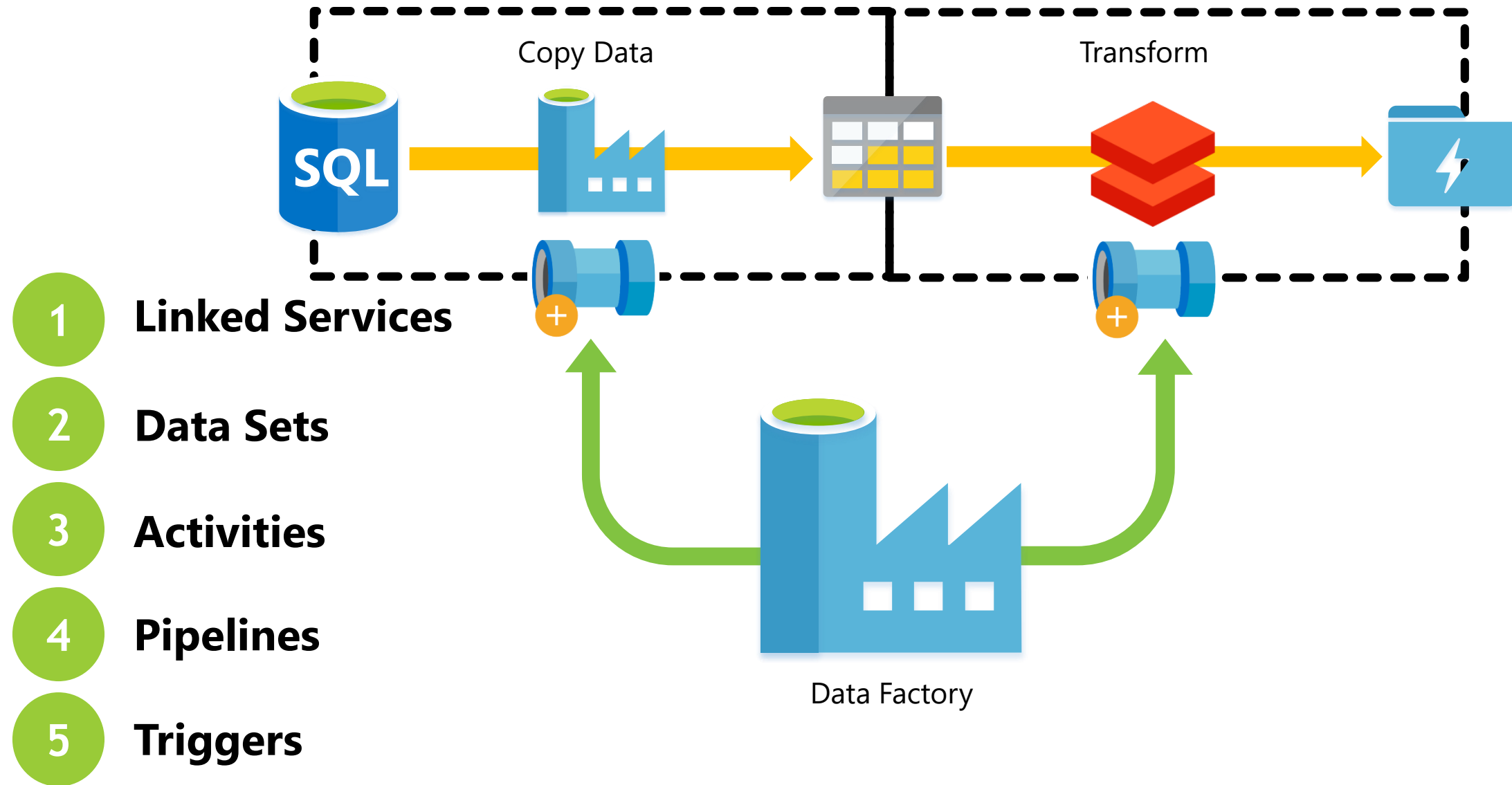


Data Factory

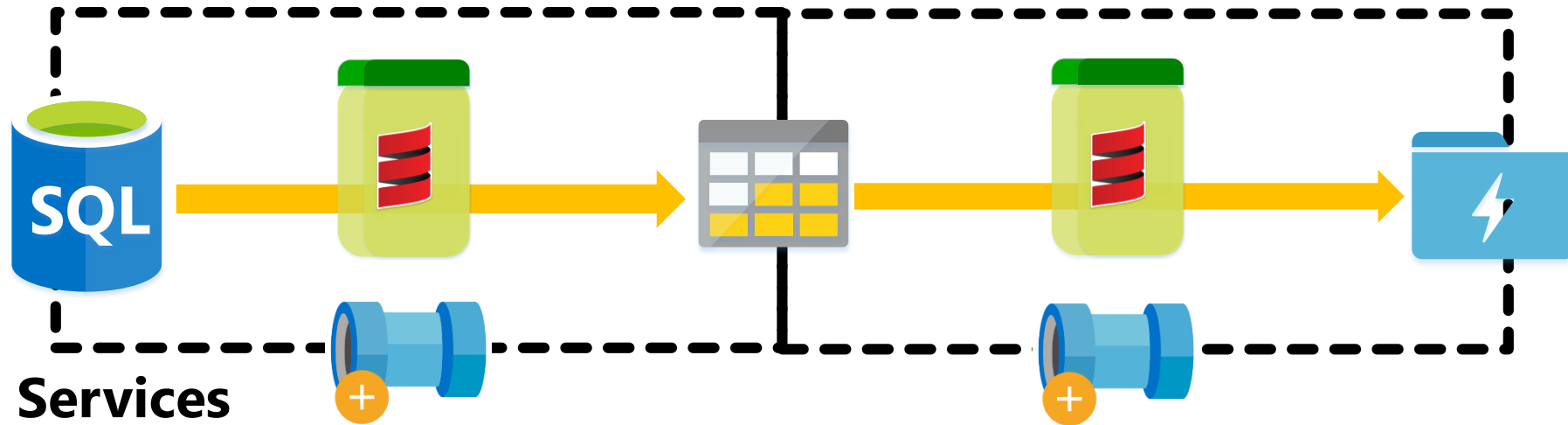


Databricks

Data Factory Control Flow Components



Data Factory Control Flow Components



1

Linked Services

2

Data Sets

3

Activities – Mapping Data Flow

4

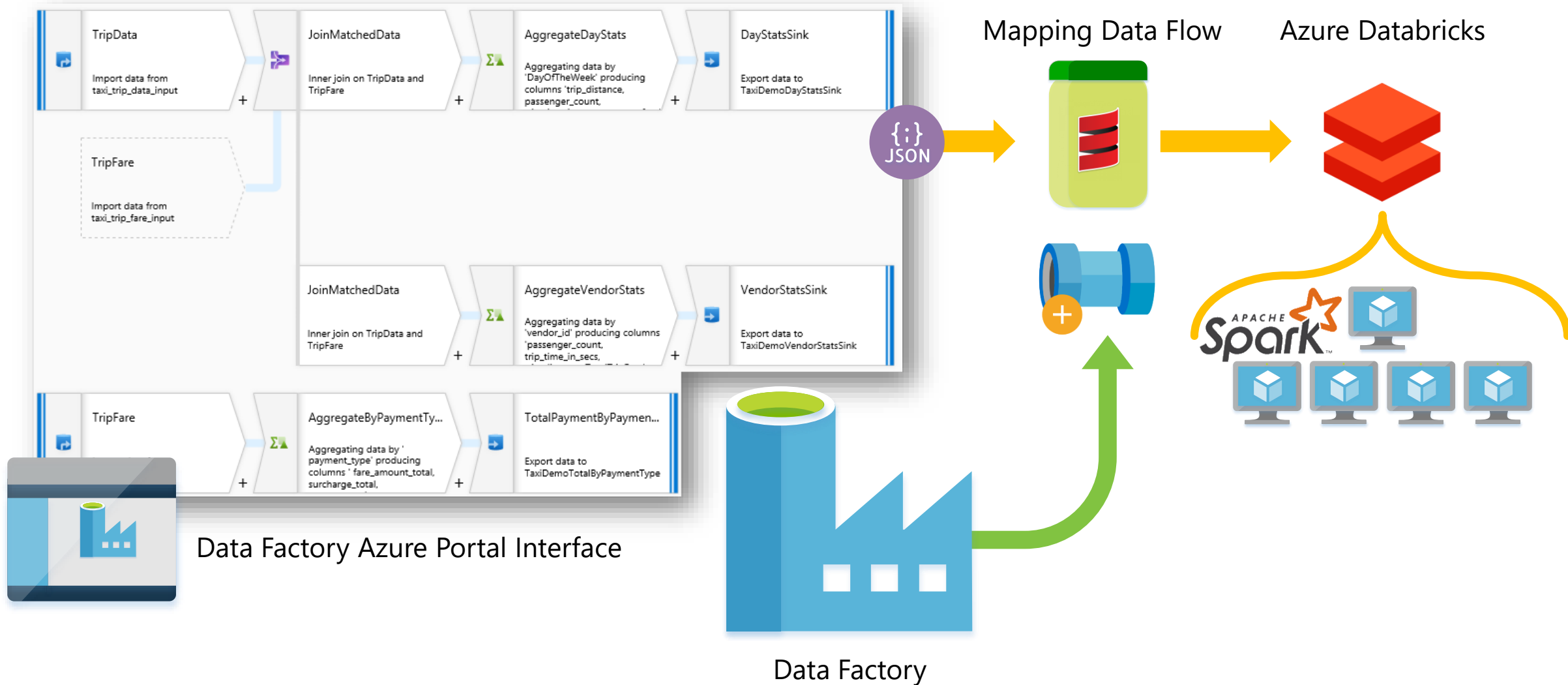
Pipelines

5

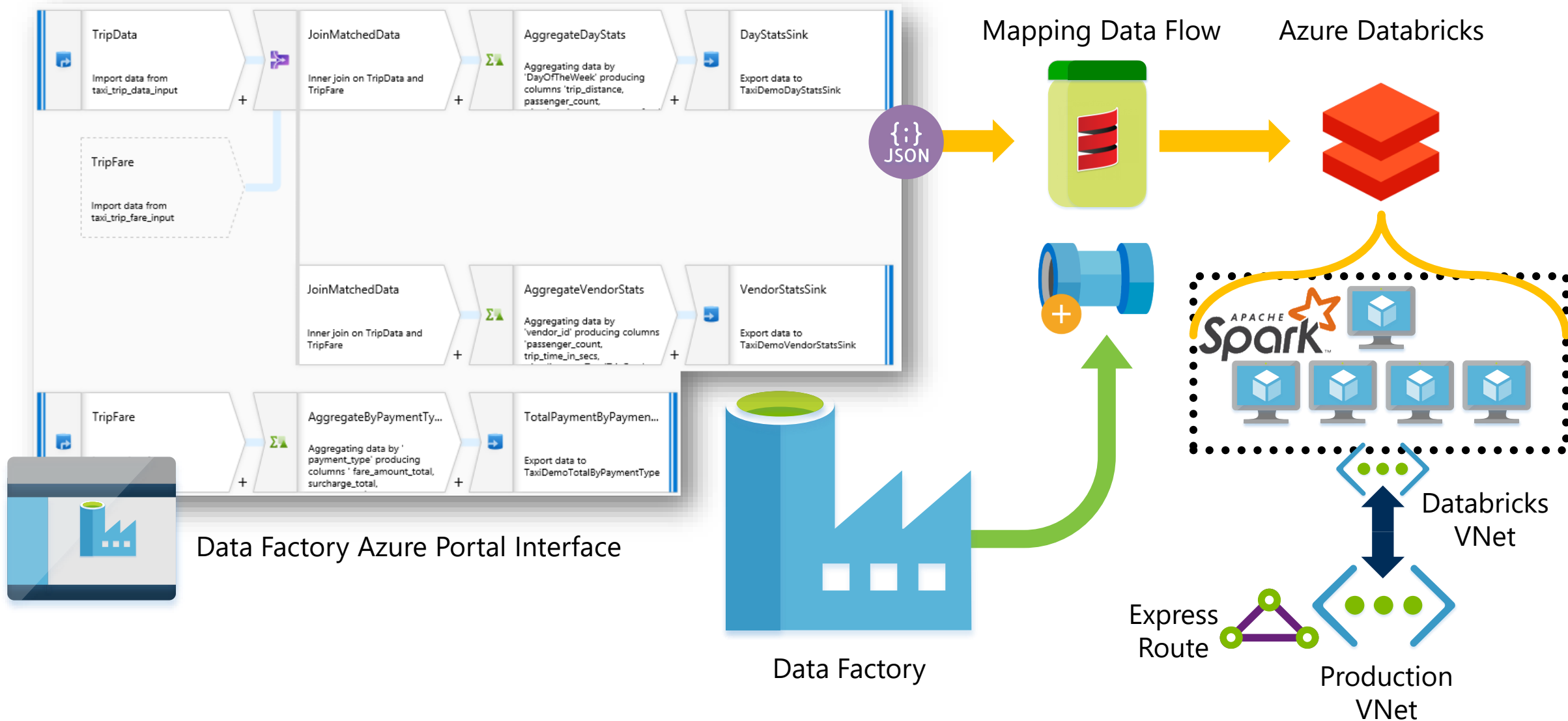
Triggers

Data Factory

What is a Mapping Data Flow?

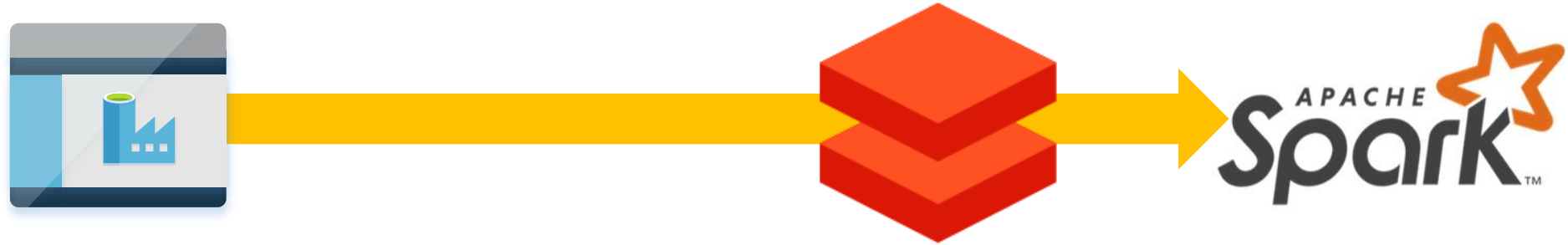


What is a Mapping Data Flow?

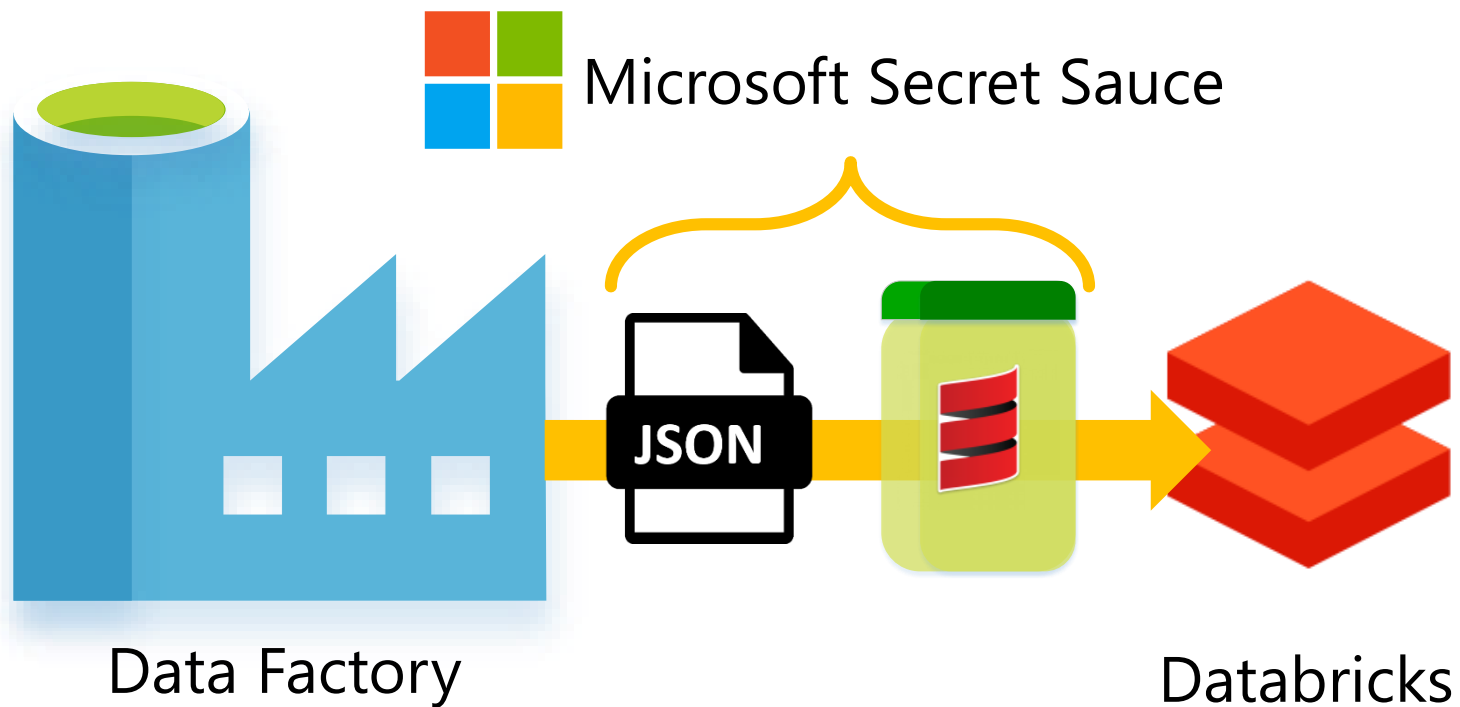


Azure Data Factory with Azure Databricks

<rant>

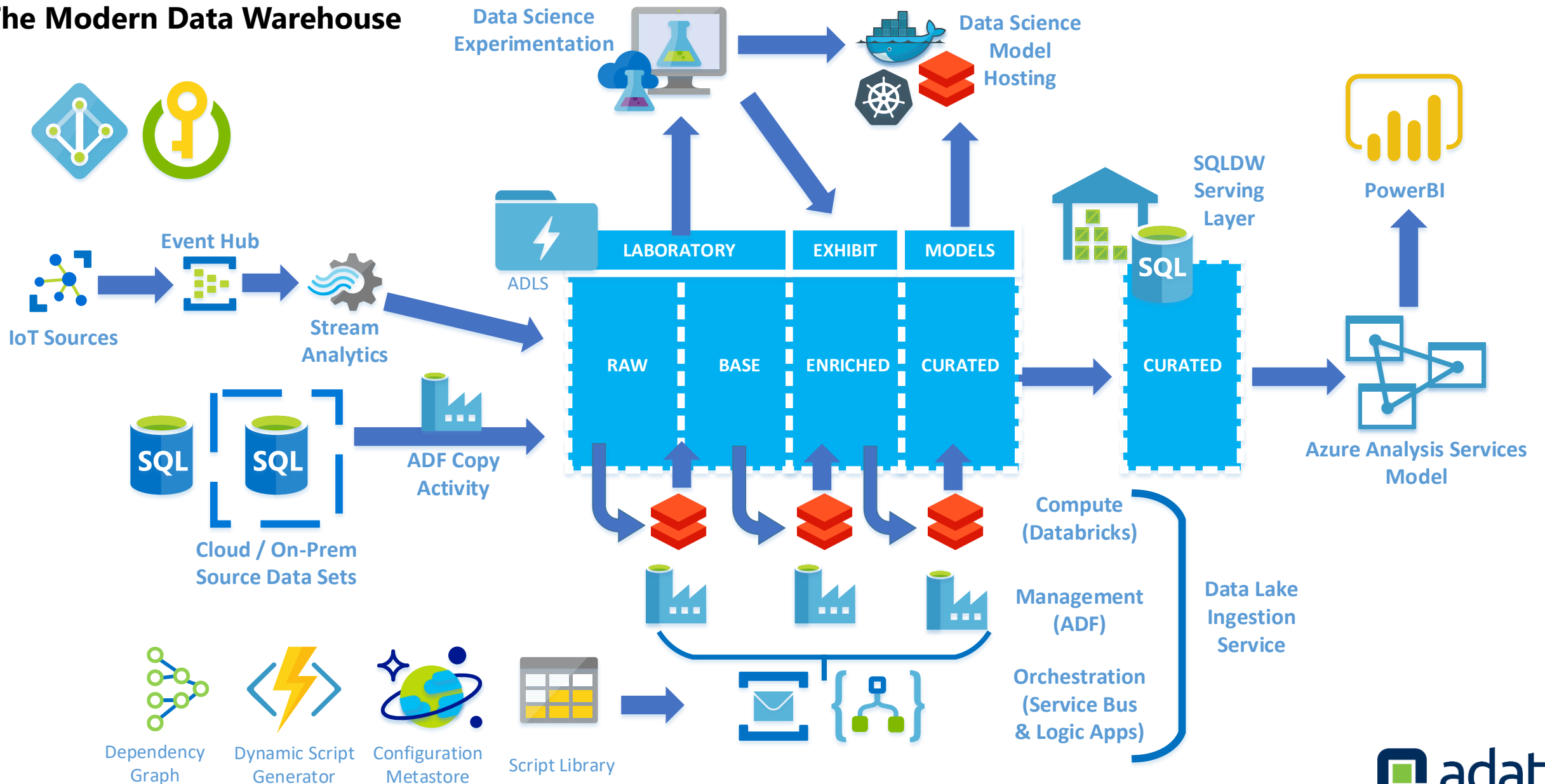


</rant>

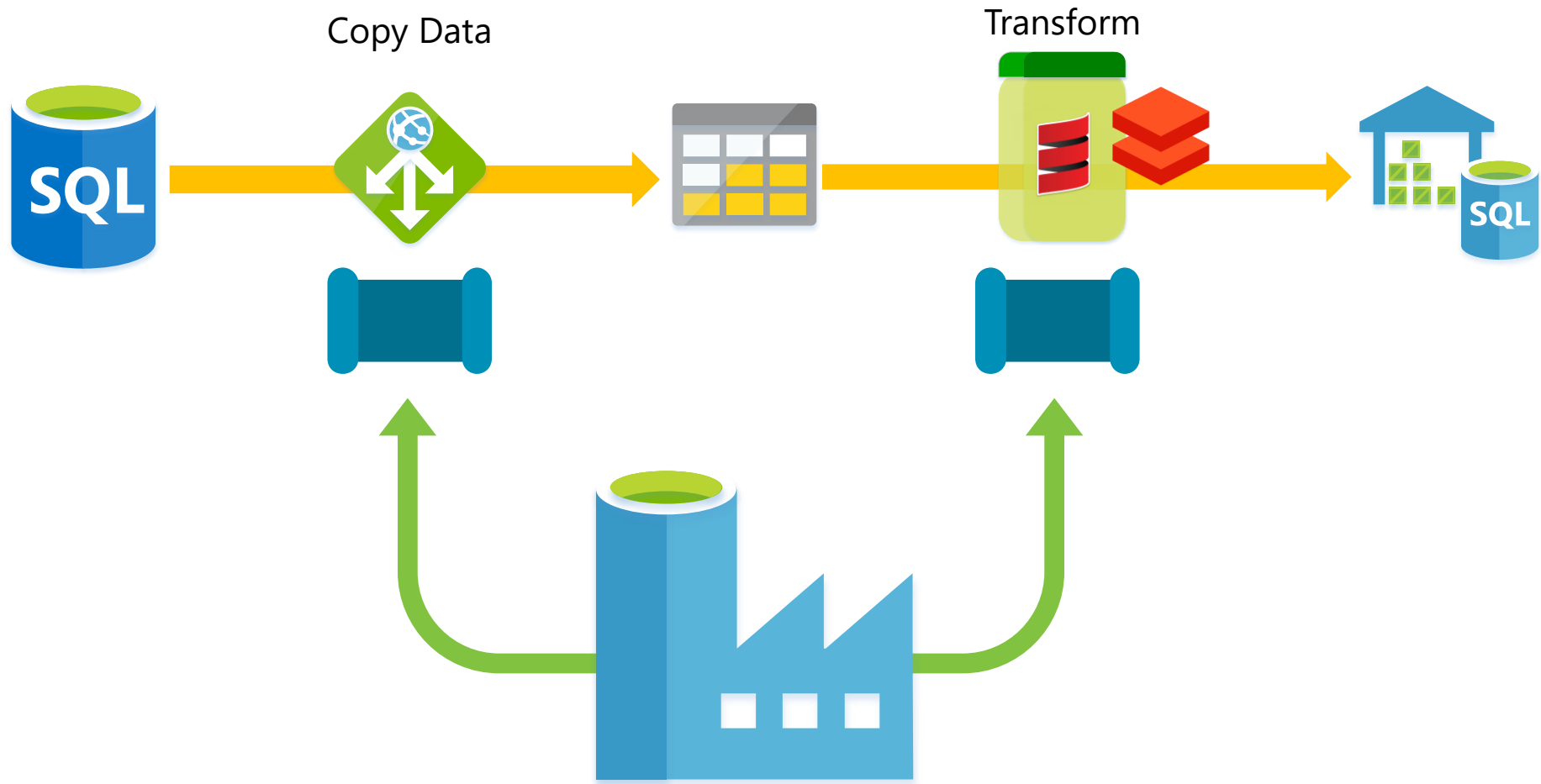


Why use Azure Data Factory?

The Modern Data Warehouse



What is Azure Data Factory?



Orchestrator of our solution Control Flow operations.

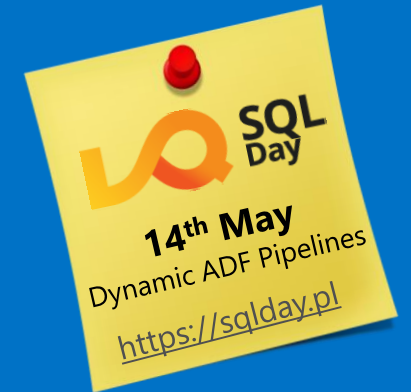
Orchestrator of our solution Data Flow transformations.

... using cloud native technology in  zure and now with an easy developer interface for both.

Thanks for Listening

Paul Andrew

 @MrPaulAndrew



Blog: mrpaulandrew.com

Email: paul@mrpaulandrew.com

GitHub: github.com/mrpaulandrew

