

The Data Engineering before the Data Science

- made easy with Azure Data Factory



Paul Andrew

Principal Consultant & Solution Architect

altius



@MrPaulAndrew

GitHub



<https://github.com/mrpaulandrew>

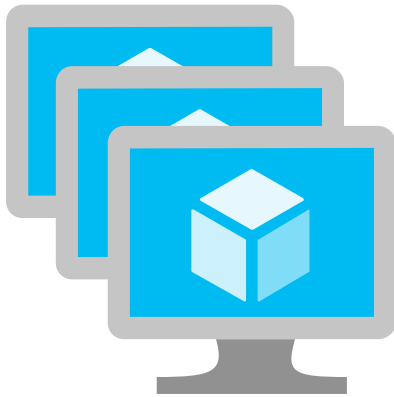
CommunityEvents

Demo code, content and slides from various community events.

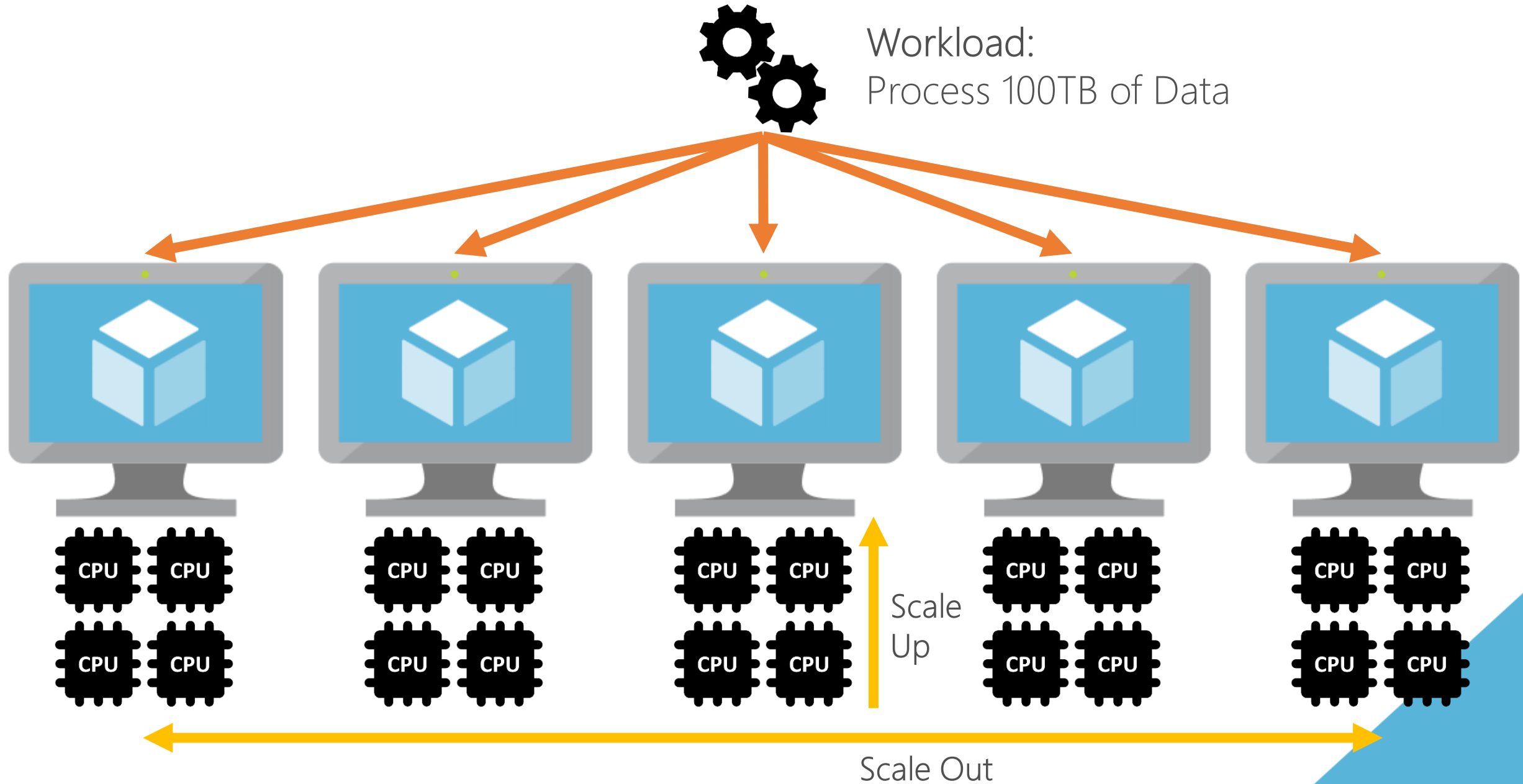
● C++

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

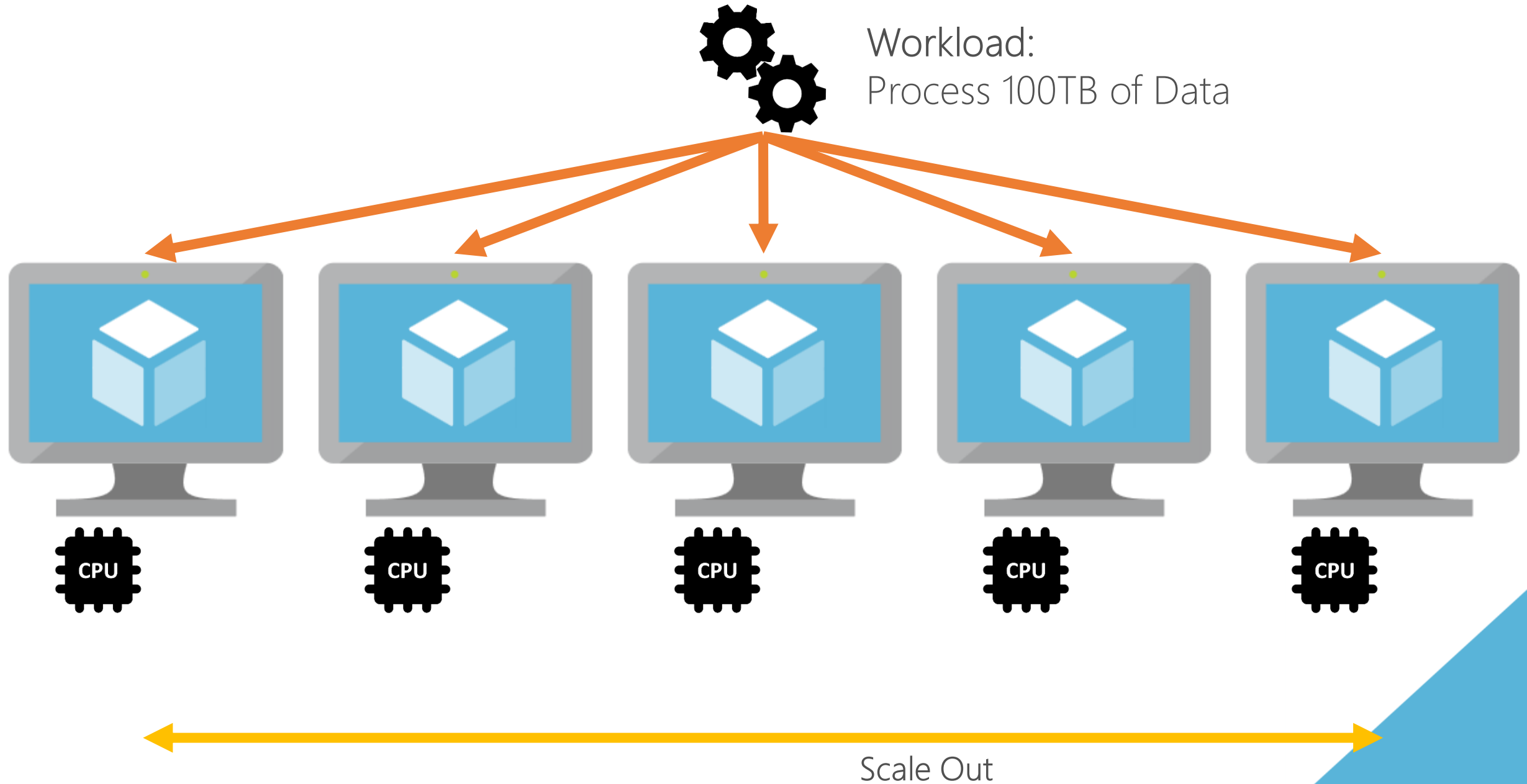
Scale Up vs Scale Out



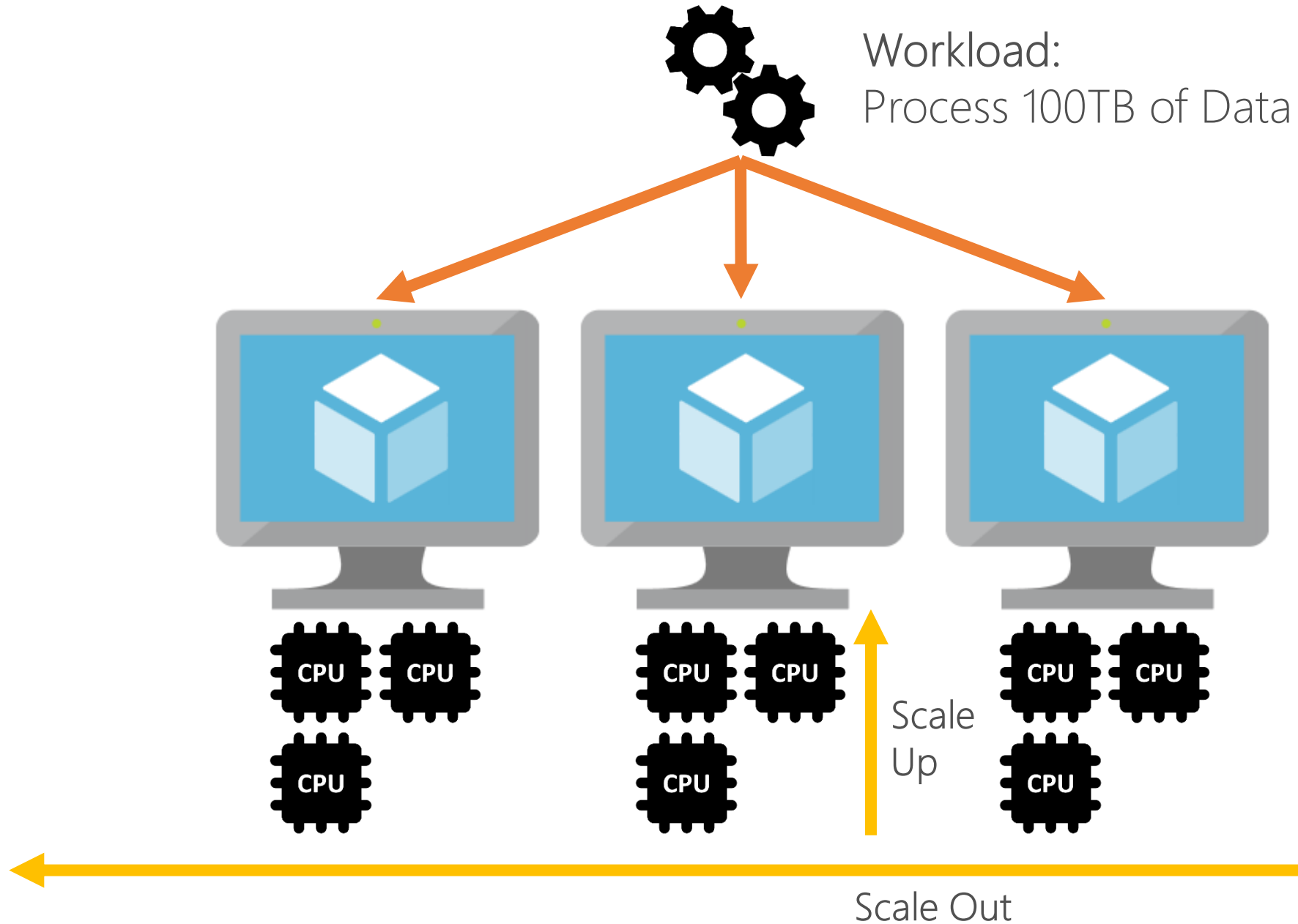
Scale Up and Scale Out



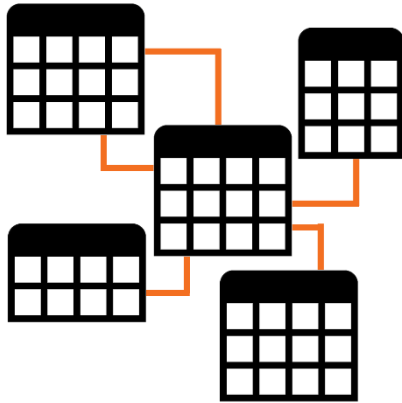
Scale Up and Scale Out



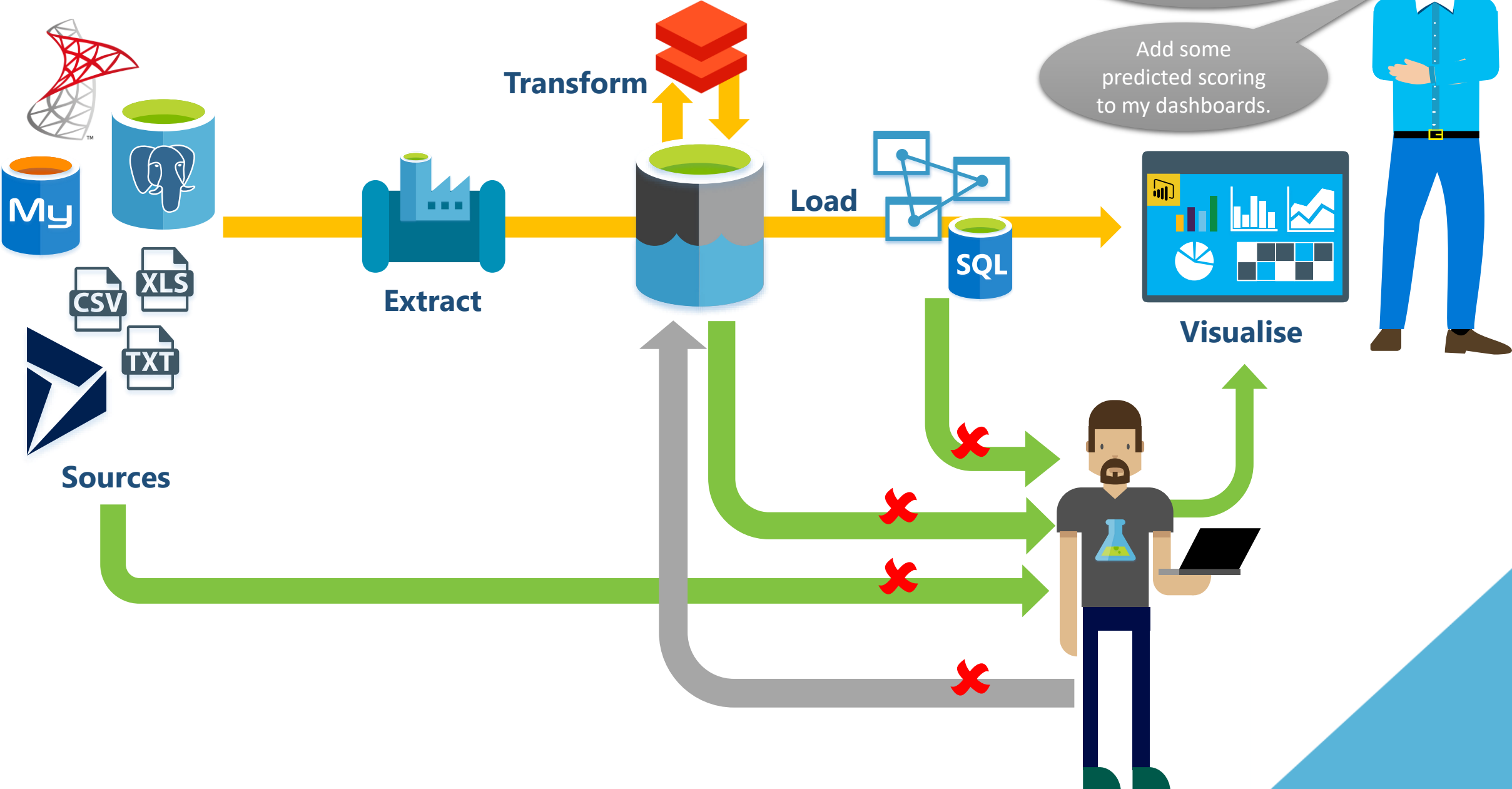
Scale Up and Scale Out



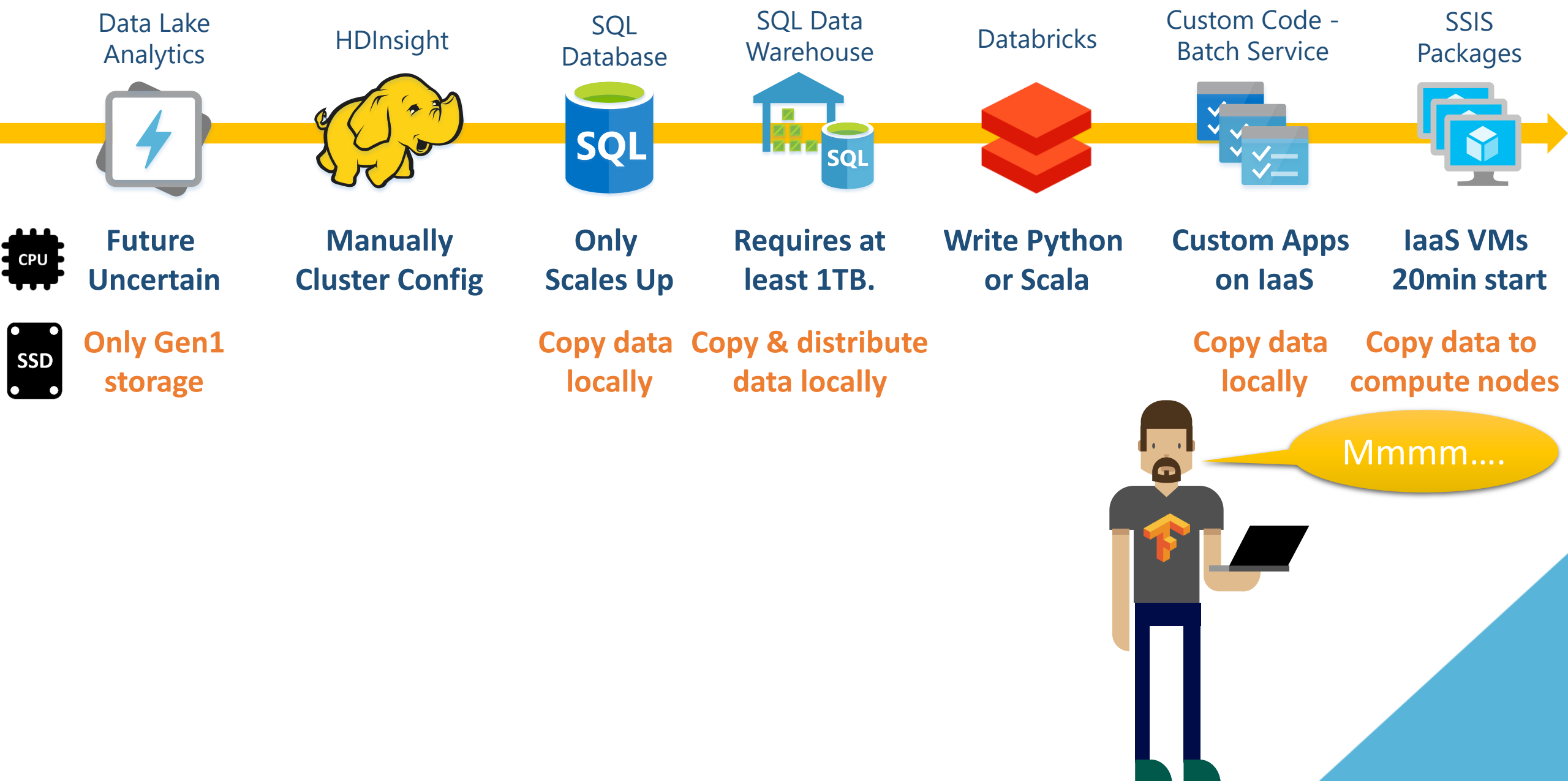
Problem: Getting the Data Ready



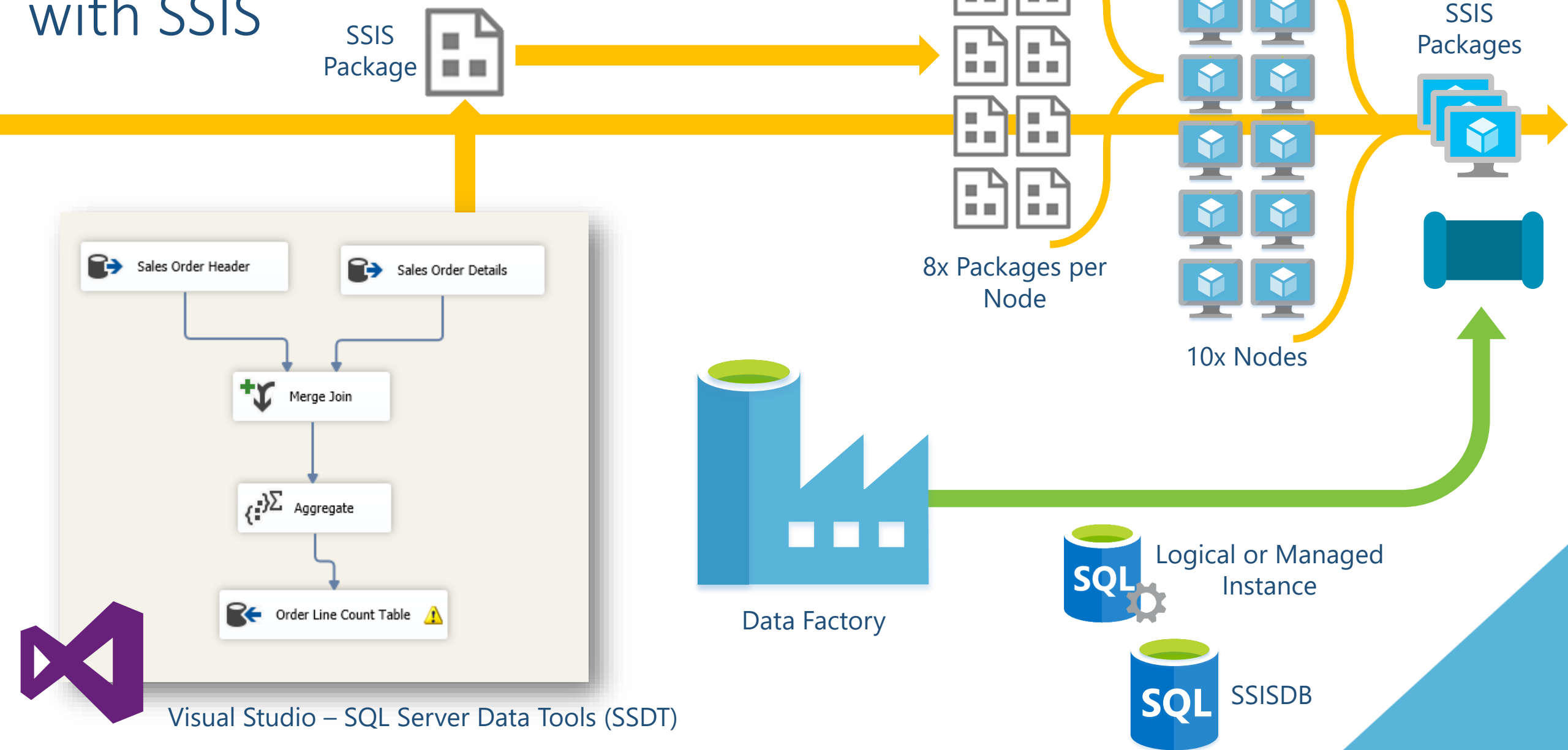
Getting the Data Ready



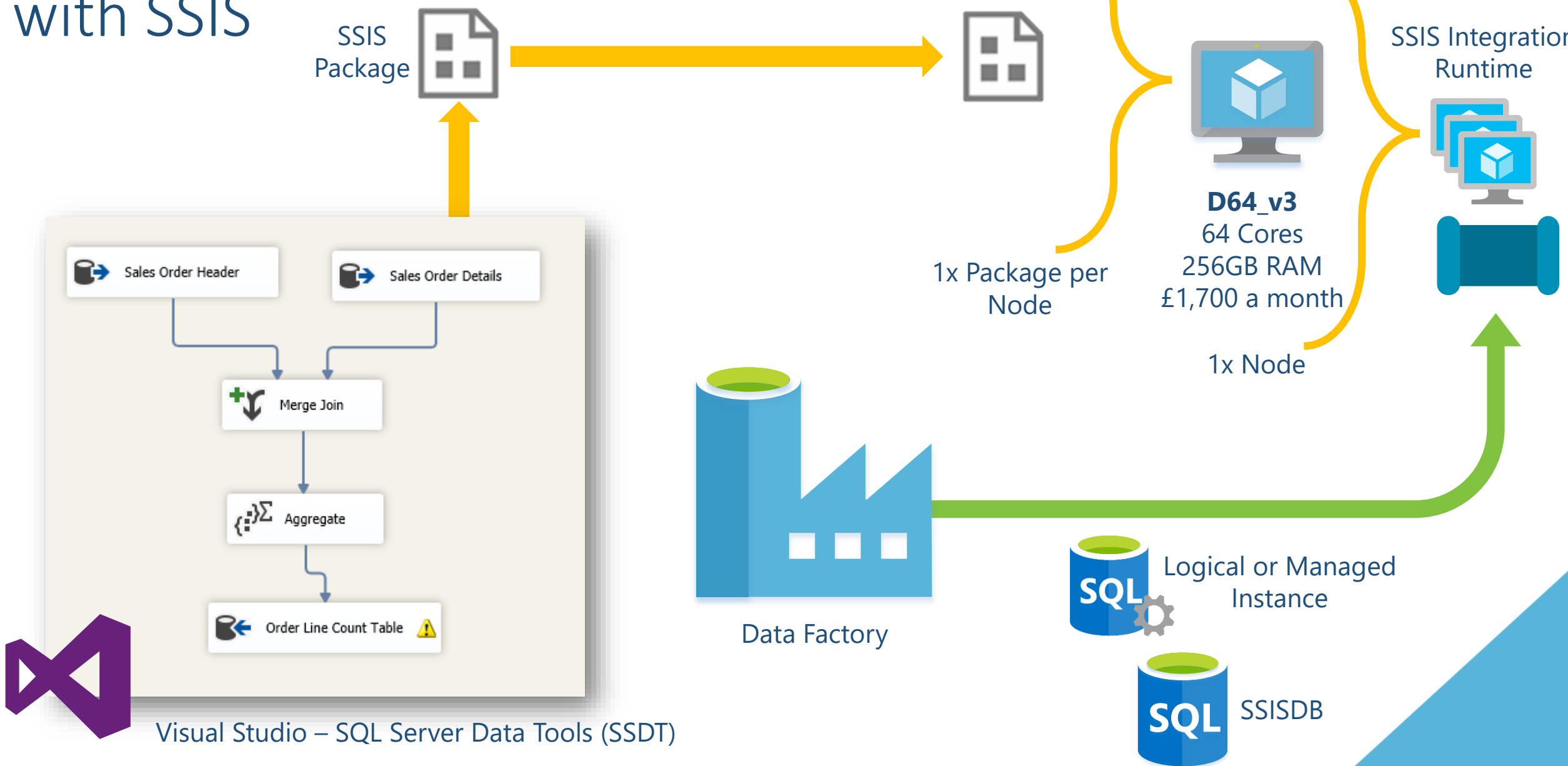
Data Transformation in zure



Data Transformation in zure with SSIS



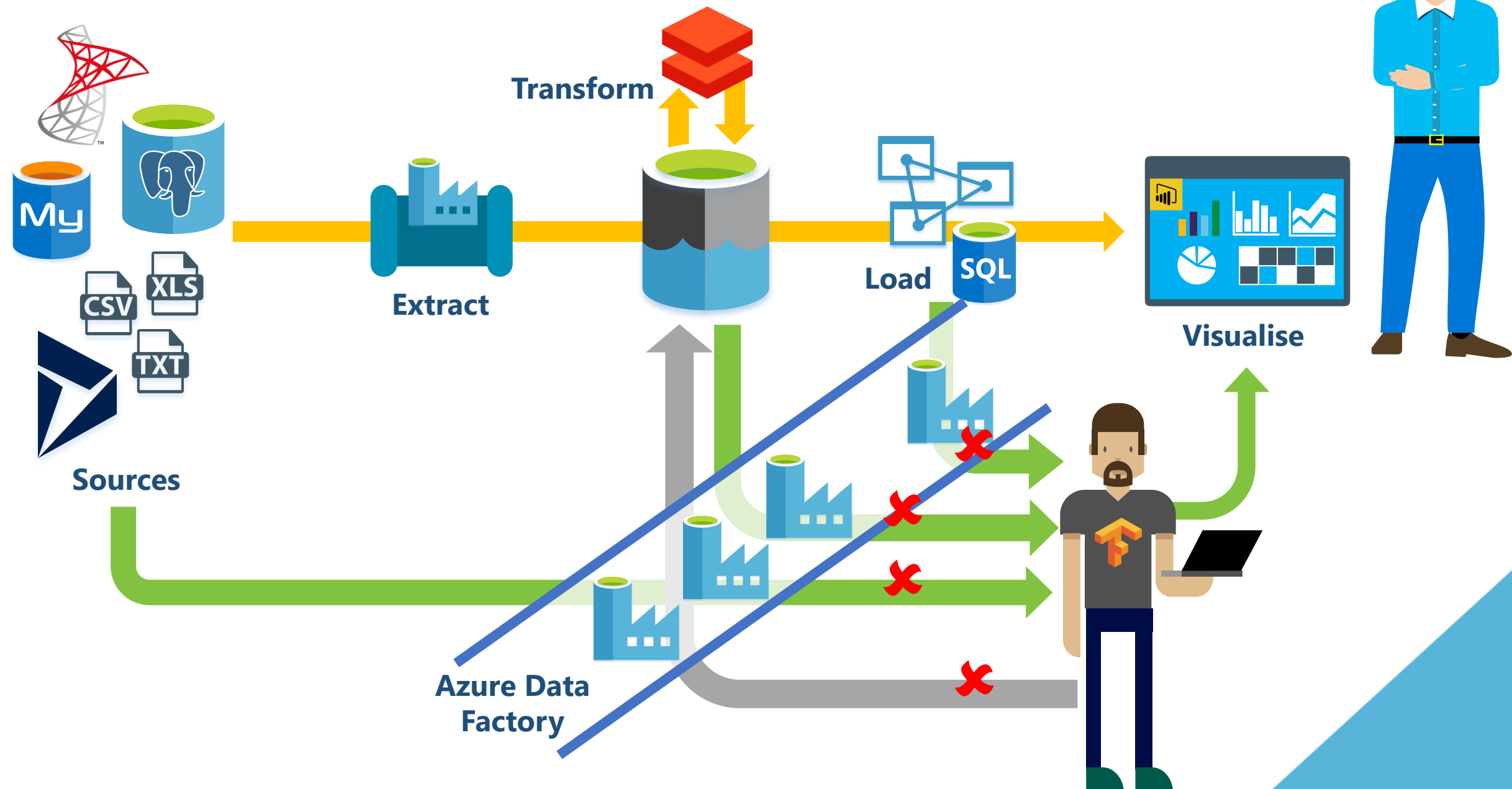
Data Transformation in zure with SSIS



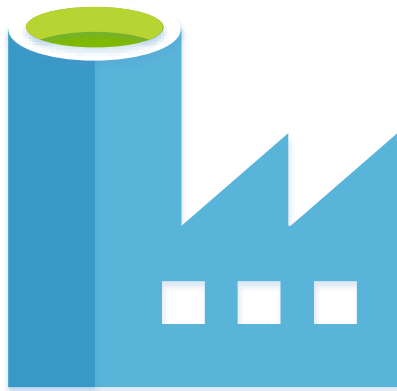
Data Transformation in zure



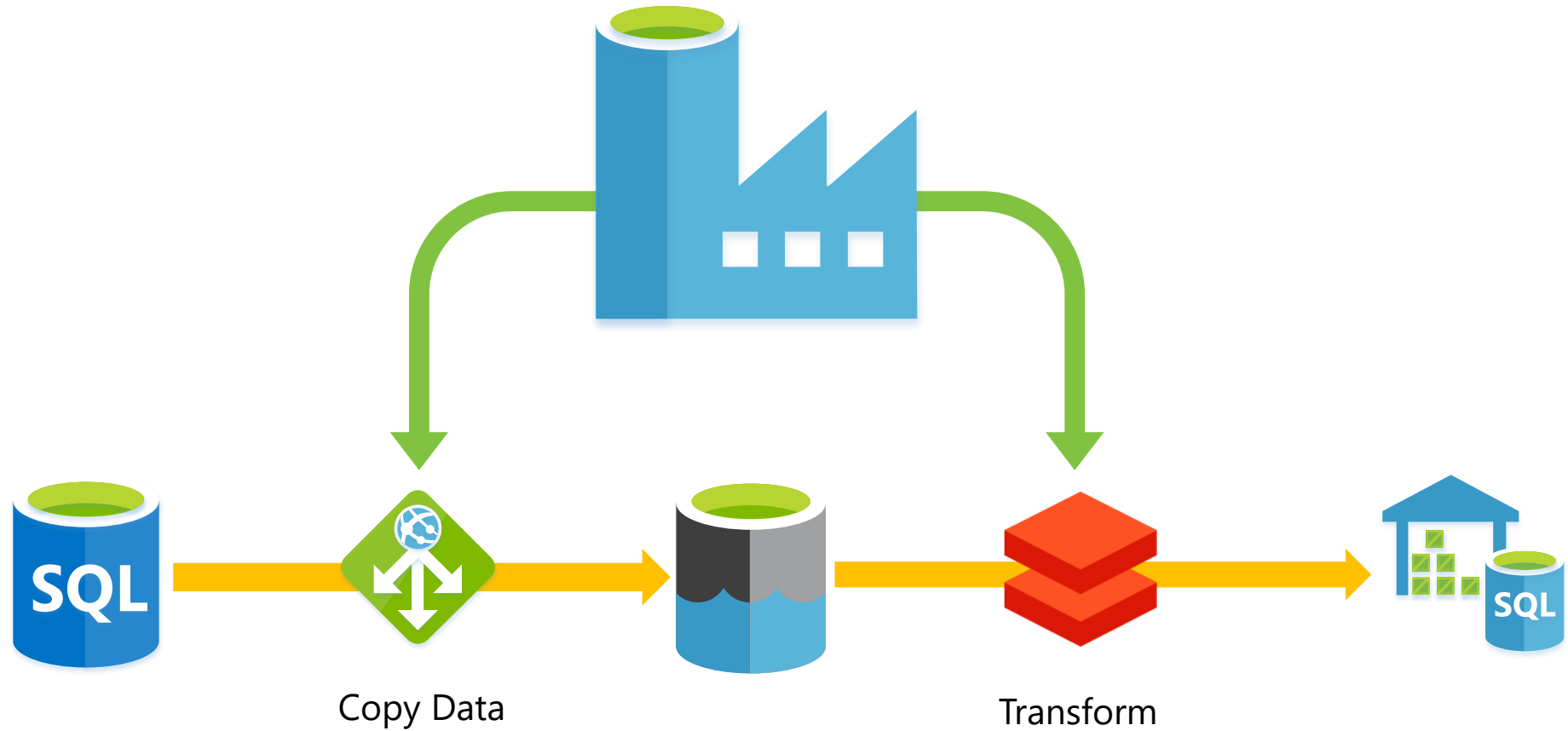
Getting the Data Ready



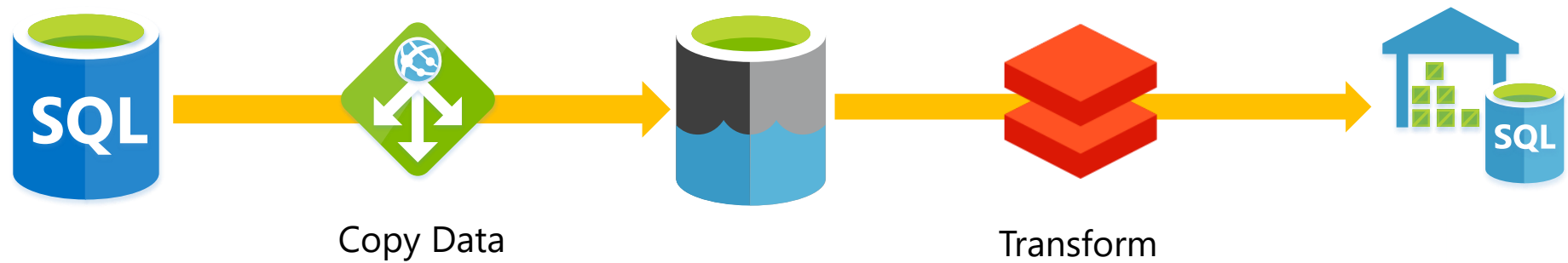
Solution: Azure Data Factory



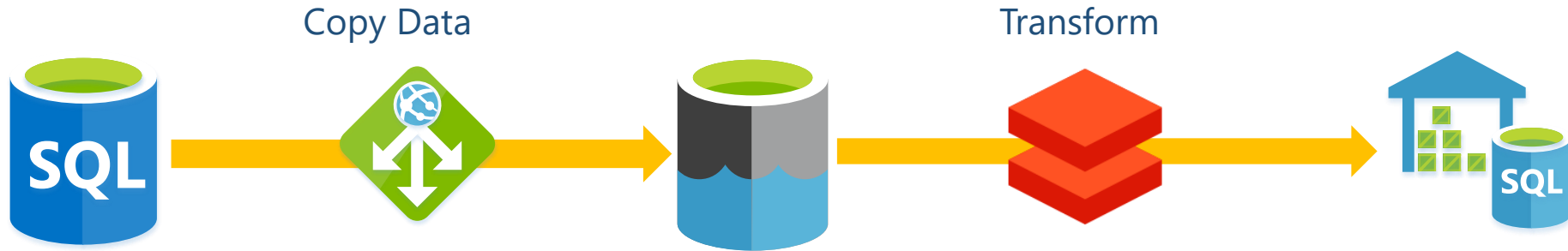
What is Azure Data Factory?



What is Azure Data Factory?



Data Factory Components



1

Linked Services – How and what to connect to. Like the SSIS connection manager.



SQLDBLinkedService

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

Data Factory Components



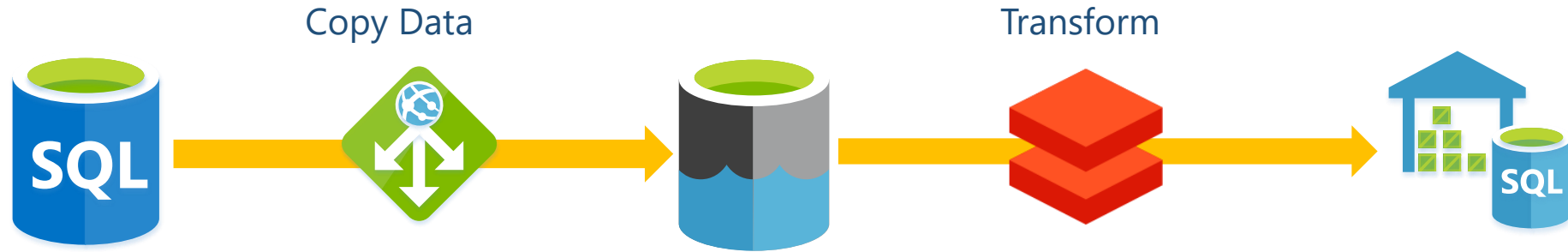
1

Linked Services –



 Amazon Marketplace Web Service (Preview)	 Amazon Redshift	 Amazon S3	 HDFS	 HTTP	 Hive	 Netezza	 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			

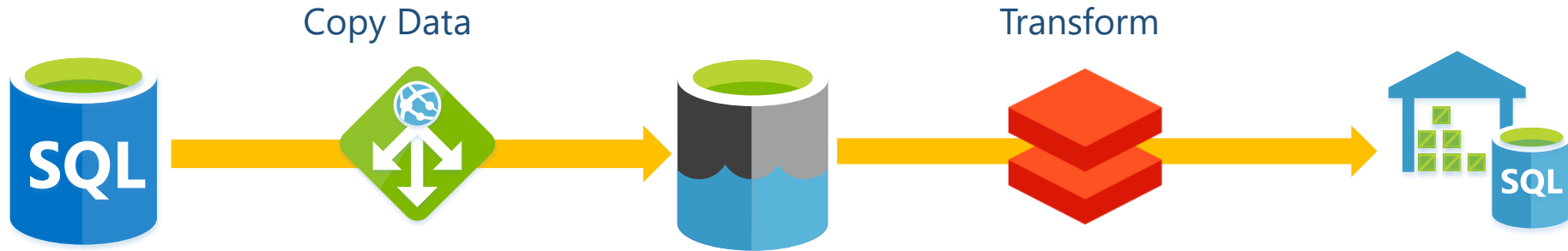
Data Factory Components



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?

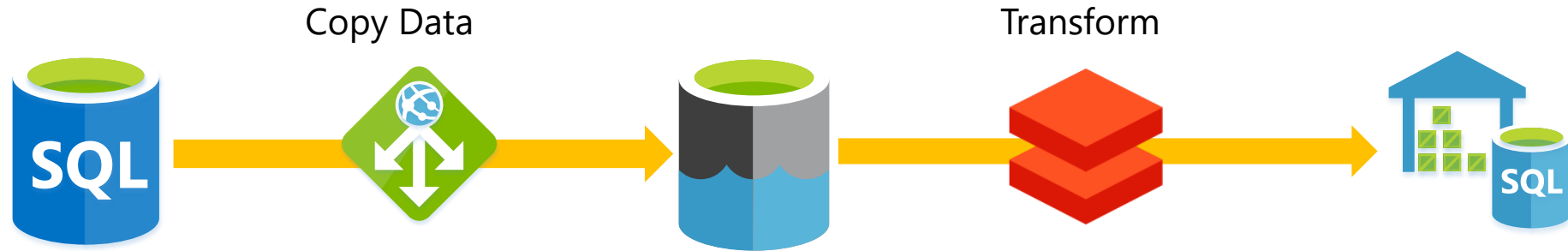


dbo.DimCustomer



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

Data Factory Components



1

Linked Services

2

Data Sets

3

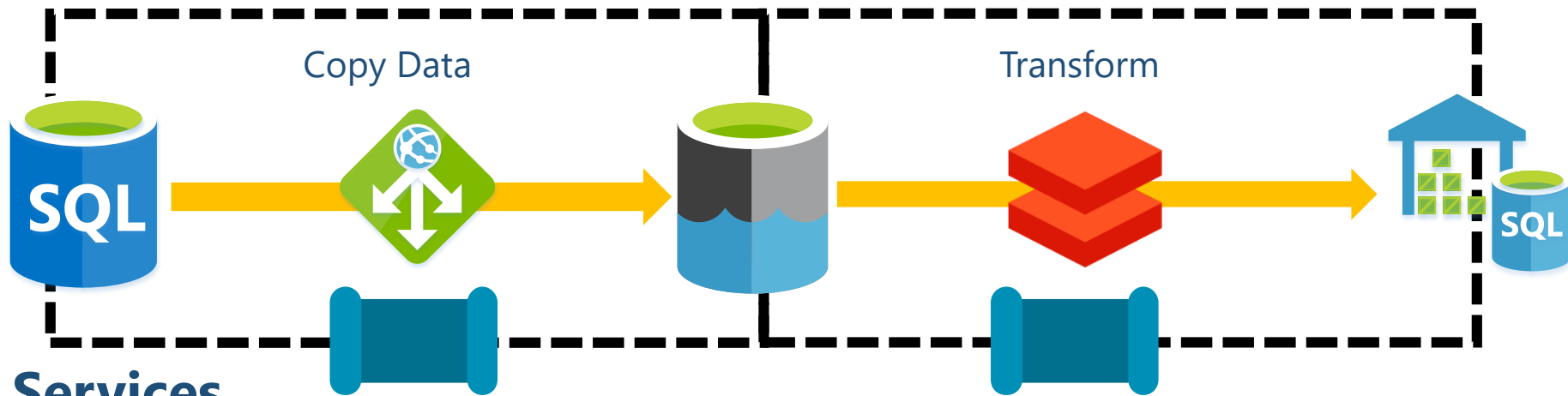
Activities – What do we want to happen?
With what conditions?



Databricks Notebook Activity

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

Data Factory Components



1

Linked Services

2

Data Sets

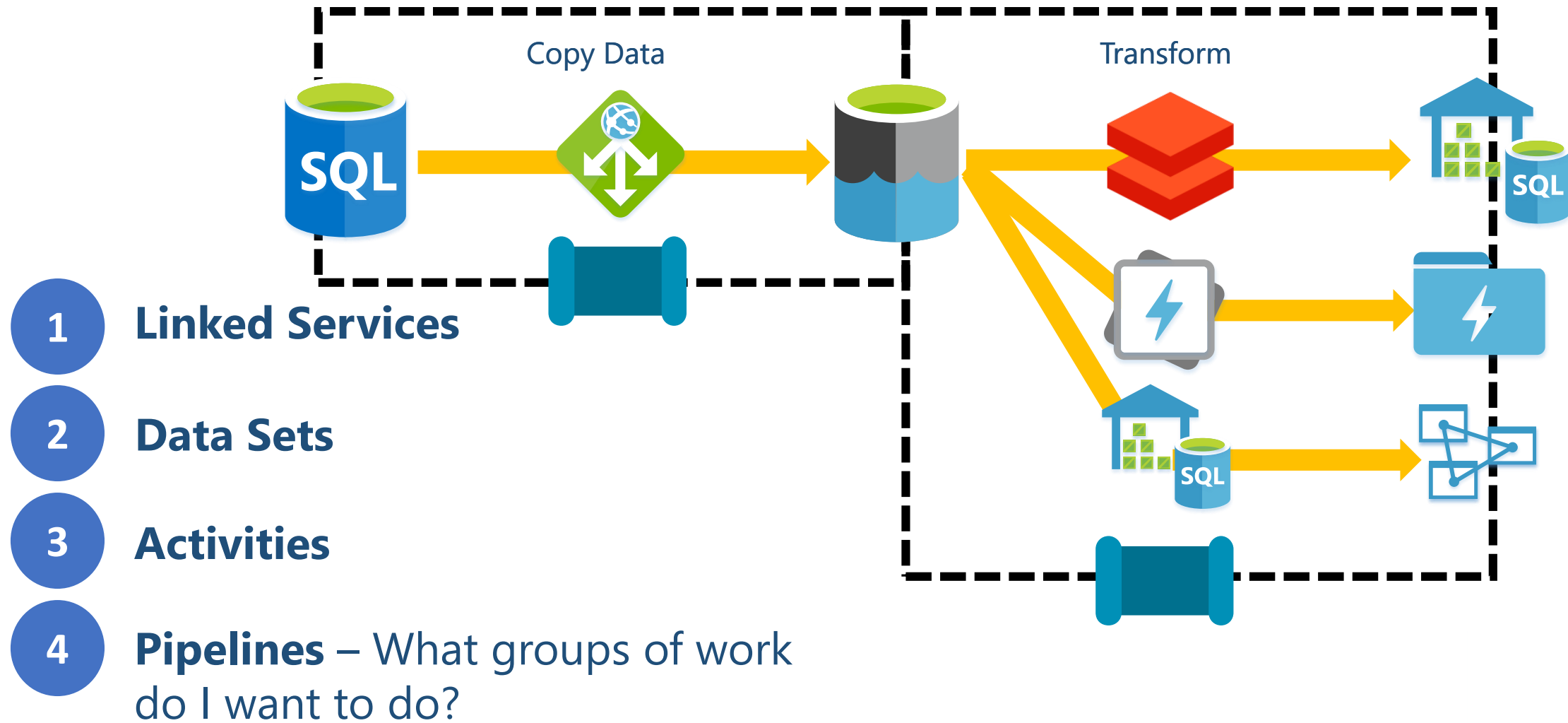
3

Activities

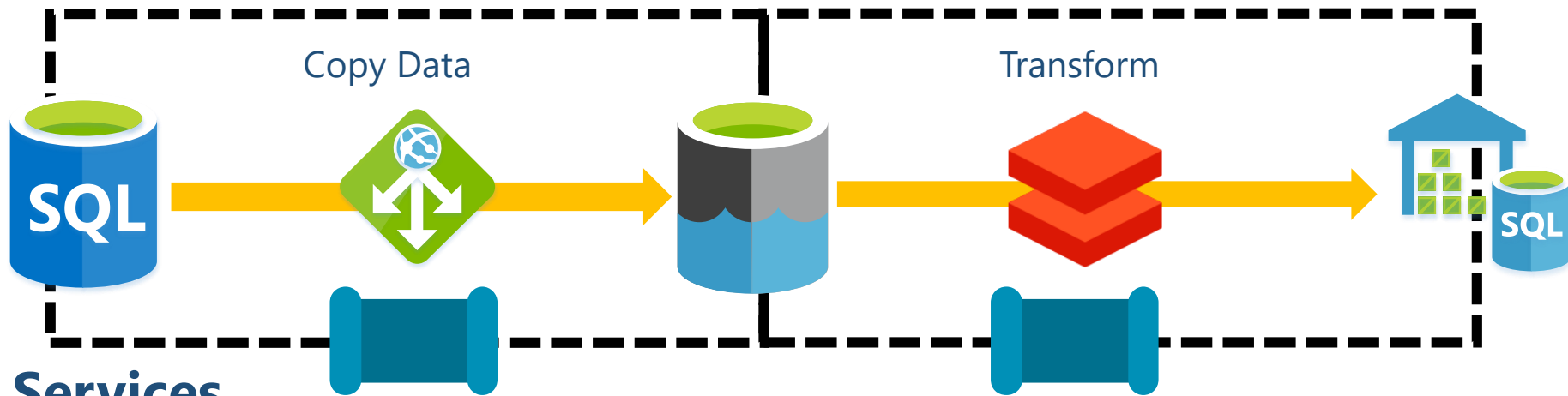
4

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

Data Factory Components



Data Factory Components



1

Linked Services

2

Data Sets

3

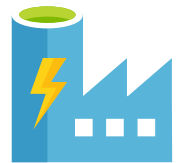
Activities

4

Pipelines

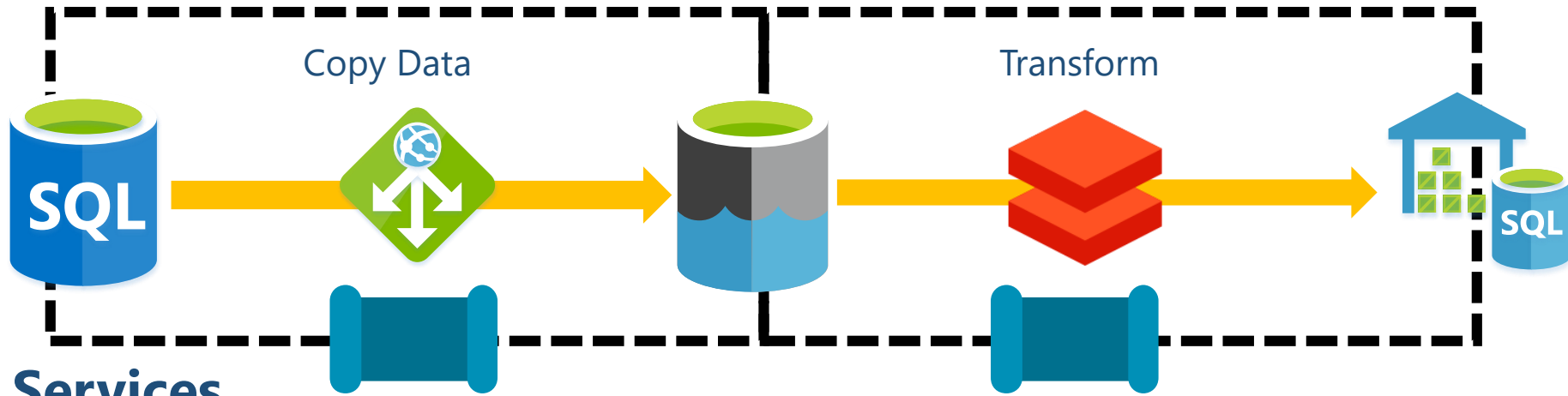
5

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

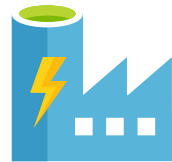
Activities

4

Pipelines

5

Triggers

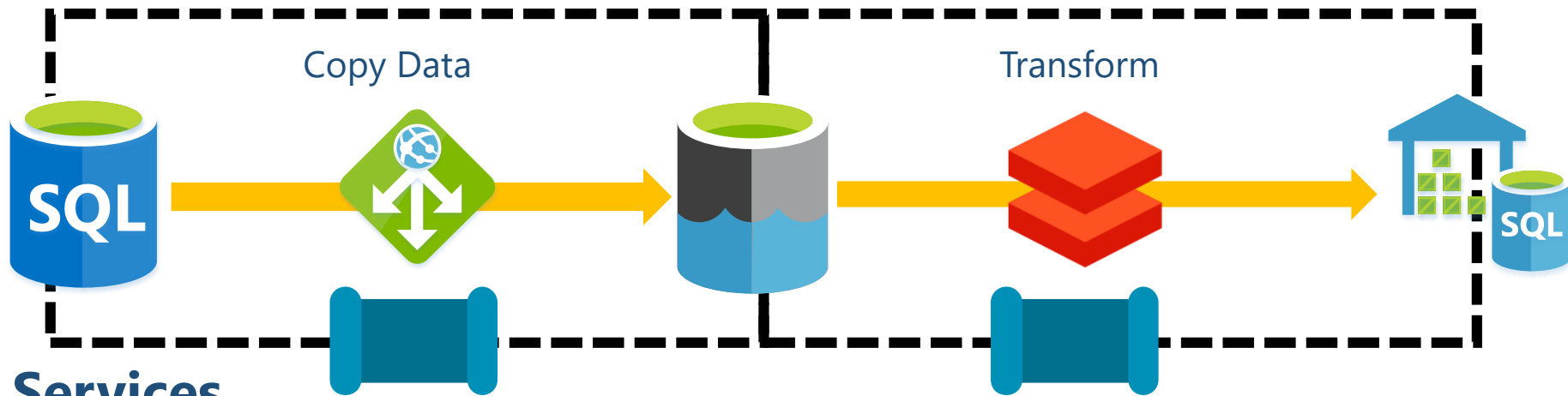


- **Manual**
- Tumbling Windows
- Scheduled
- Blob File Events
- Logic App Calls



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

Data Factory Components



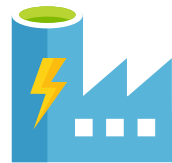
1 Linked Services

2 Data Sets

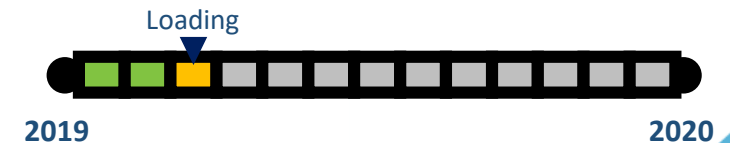
3 Activities

4 Pipelines

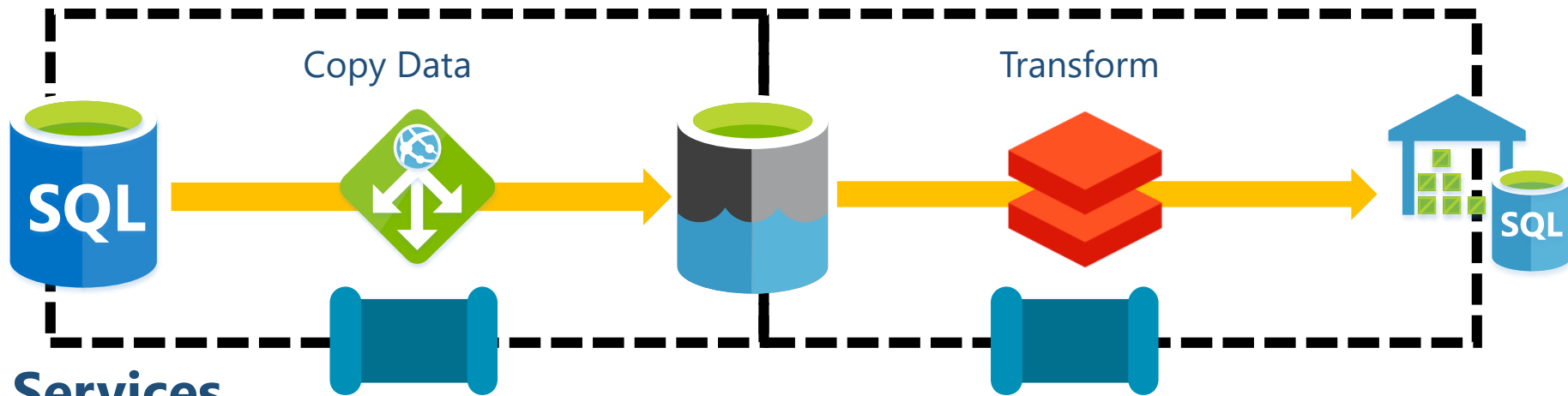
5 Triggers



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



Data Factory Components



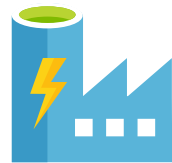
1 Linked Services

2 Data Sets

3 Activities

4 Pipelines

5 Triggers

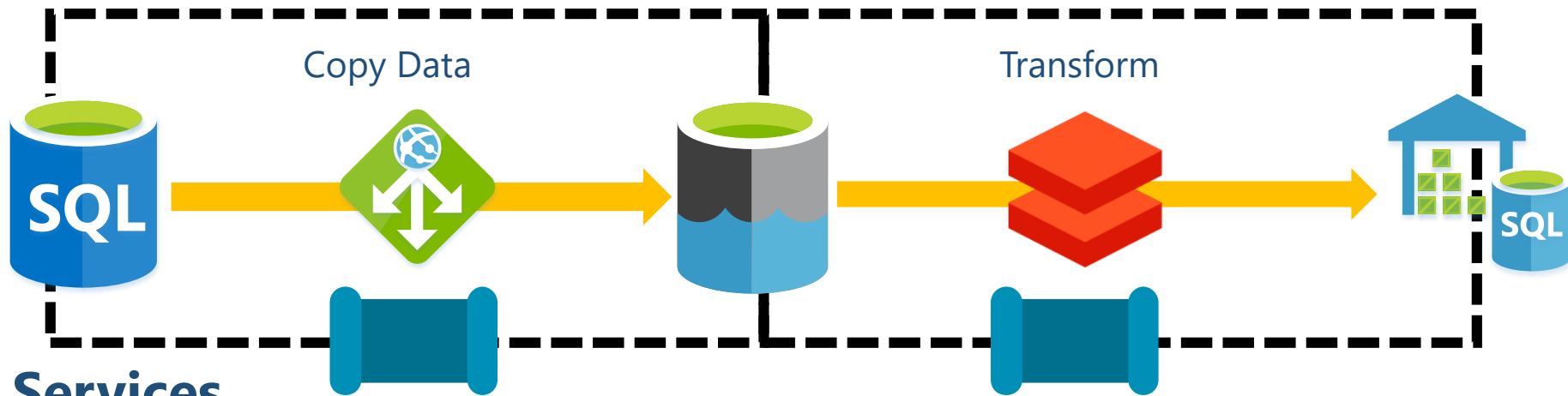


- Manual via UI
- Tumbling Windows
- **Scheduled**
- Blob File Events
- Logic App Calls



- Every 1 minute.
- UTC

Data Factory Components



1

Linked Services

2

Data Sets

3

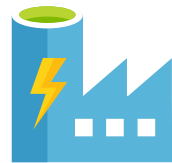
Activities

4

Pipelines

5

Triggers

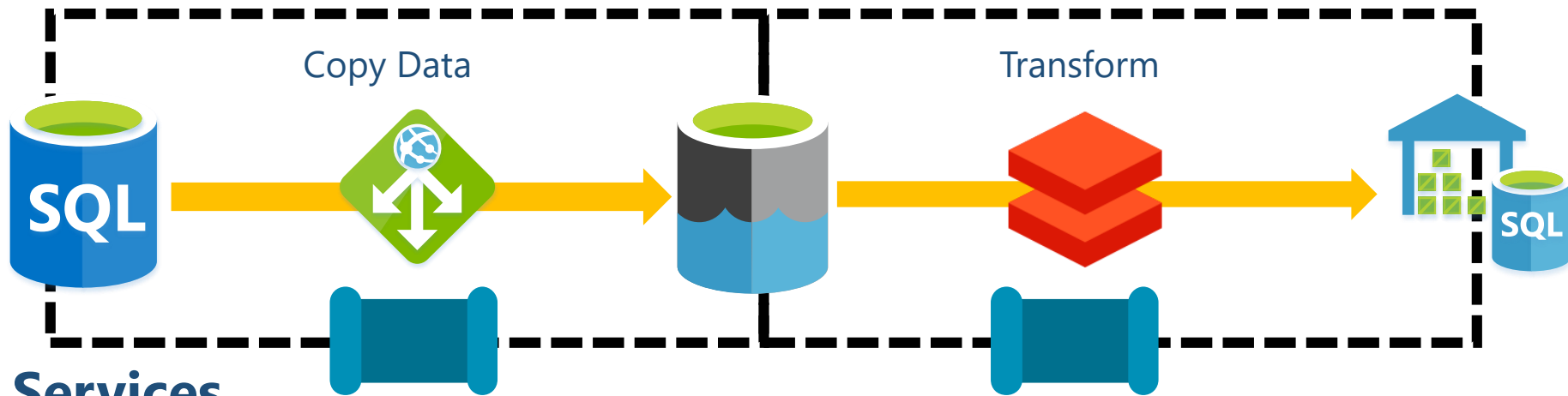


- Manual via UI
- Tumbling Windows
- Scheduled
- **Blob File Events**
- Logic App Calls



{Path} Created
{Path} Deleted

Data Factory Components



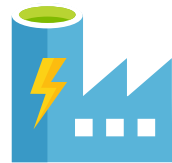
1 Linked Services

2 Data Sets

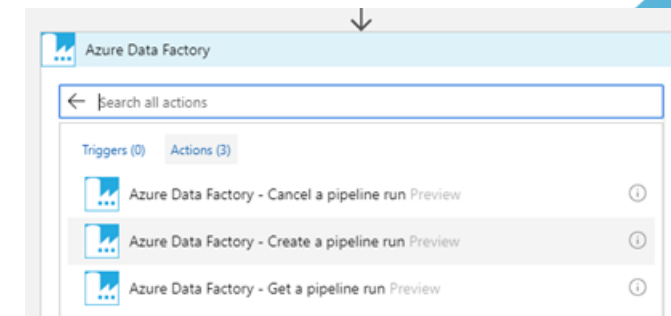
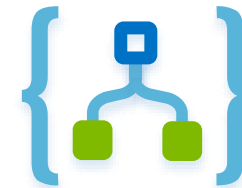
3 Activities

4 Pipelines

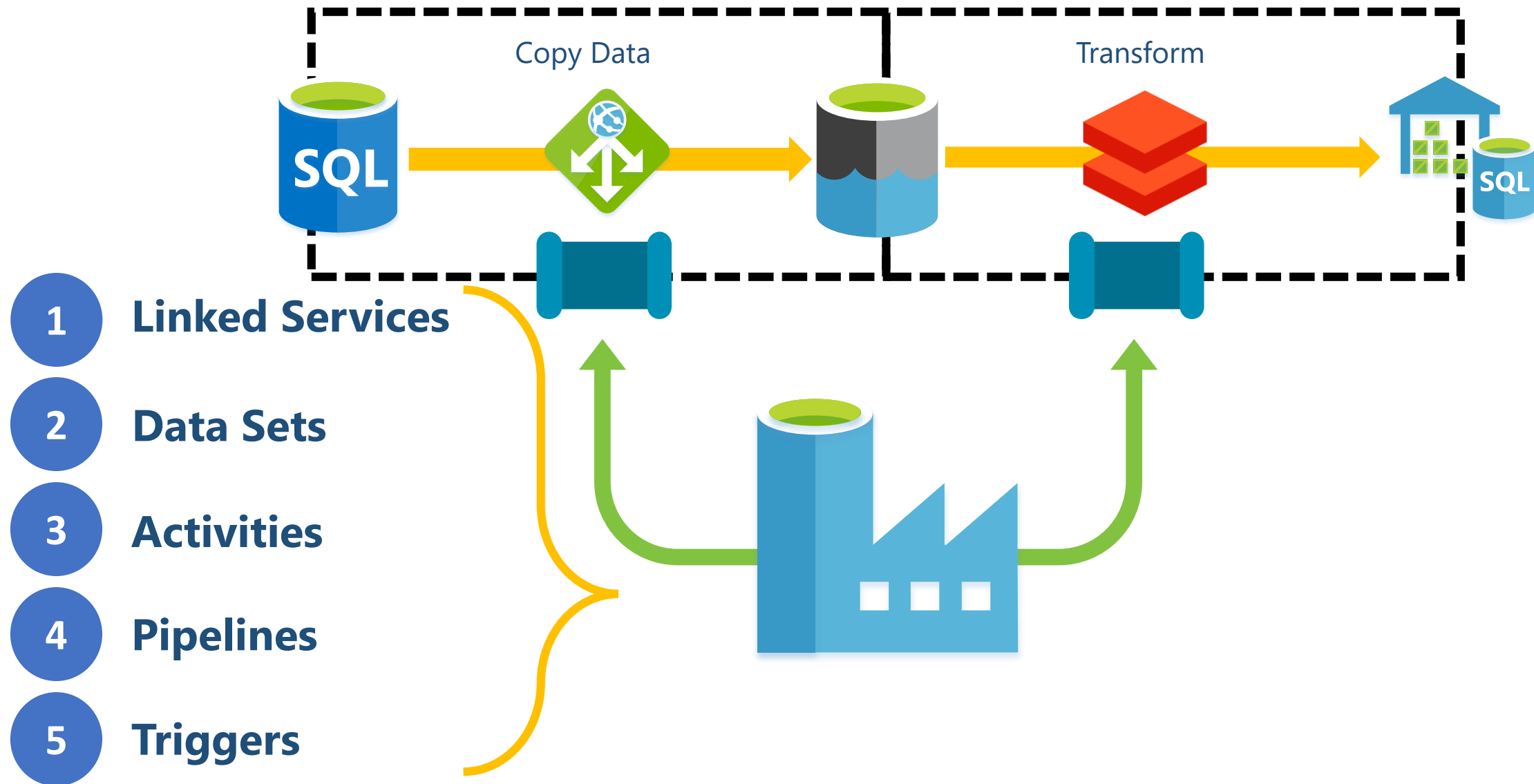
5 Triggers



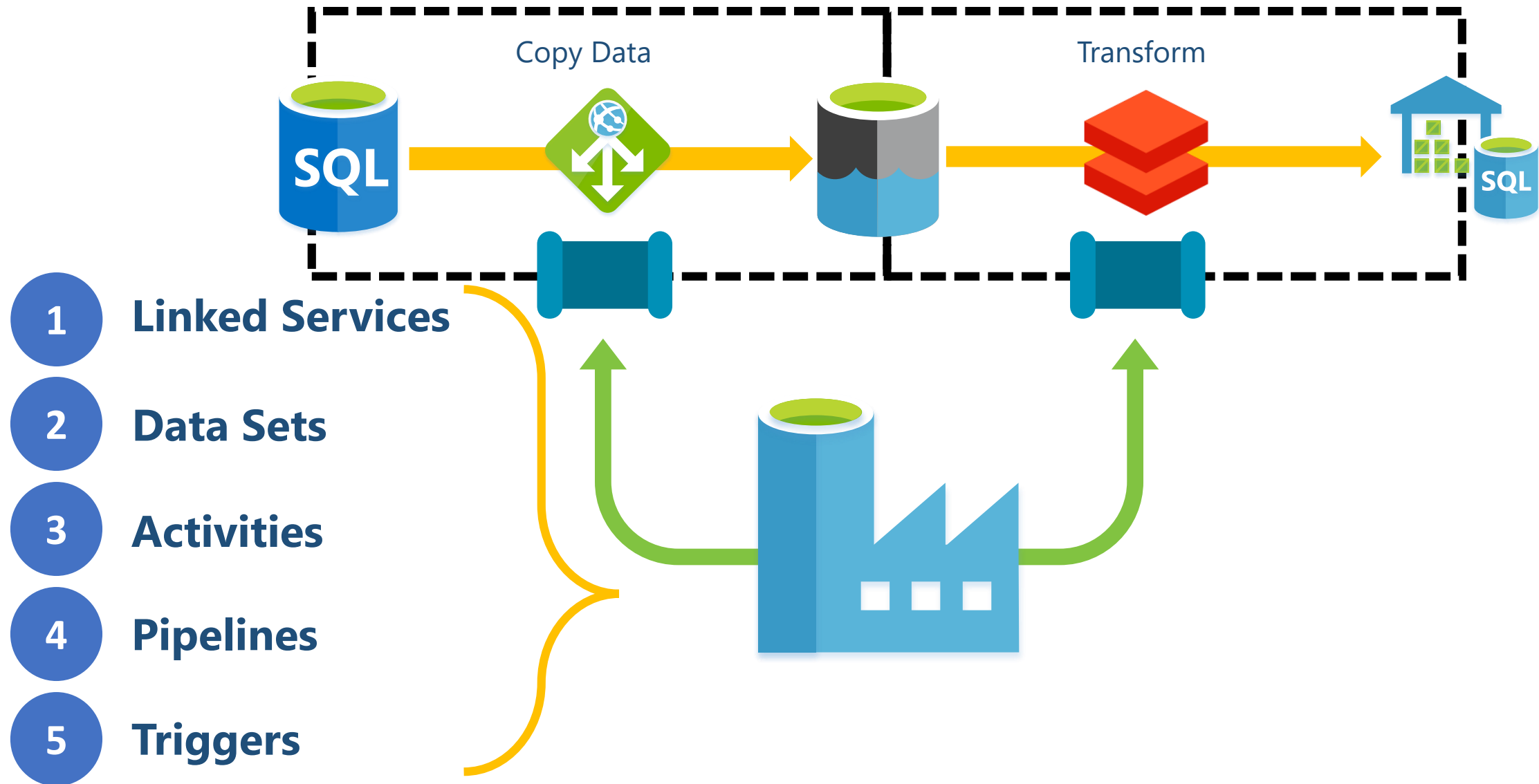
- Manual via UI
- Tumbling Windows
- Scheduled
- Blob File Events
- **Logic App Calls**



Data Factory Components



Data Factory Control Flow Components



Integration Runtimes



1

Azure
Integration Runtime

Movement Hours

Activity
Orchestration







2

SSIS
Integration Runtime

SSIS Package
Execution







3

Self Hosted
Integration Runtime

Gateway Access

Activity
Orchestration





Data Factory What & Why - Recap

1 Linked Services

2 Data Sets

3 Activities

4 Pipelines

5 Triggers

1

Azure
Integration Runtime

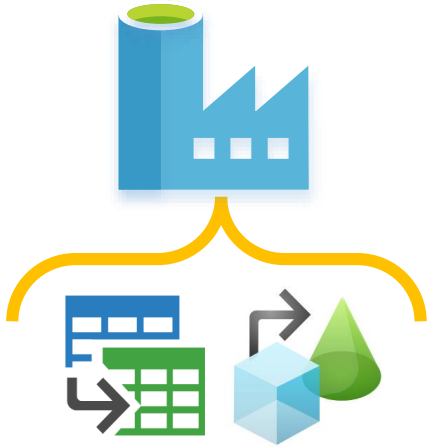
2

SSIS
Integration Runtime

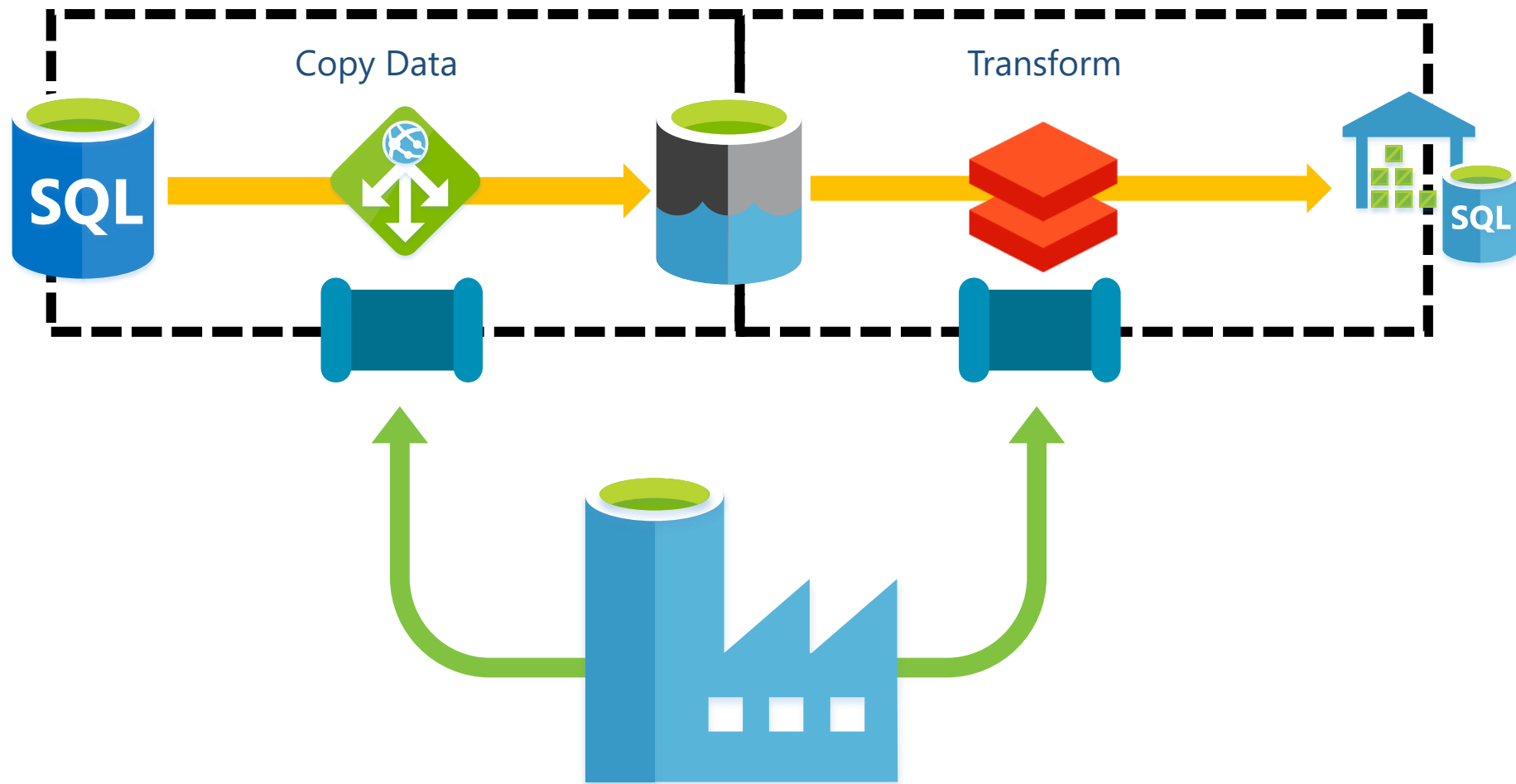
3

Self Hosted
Integration Runtime

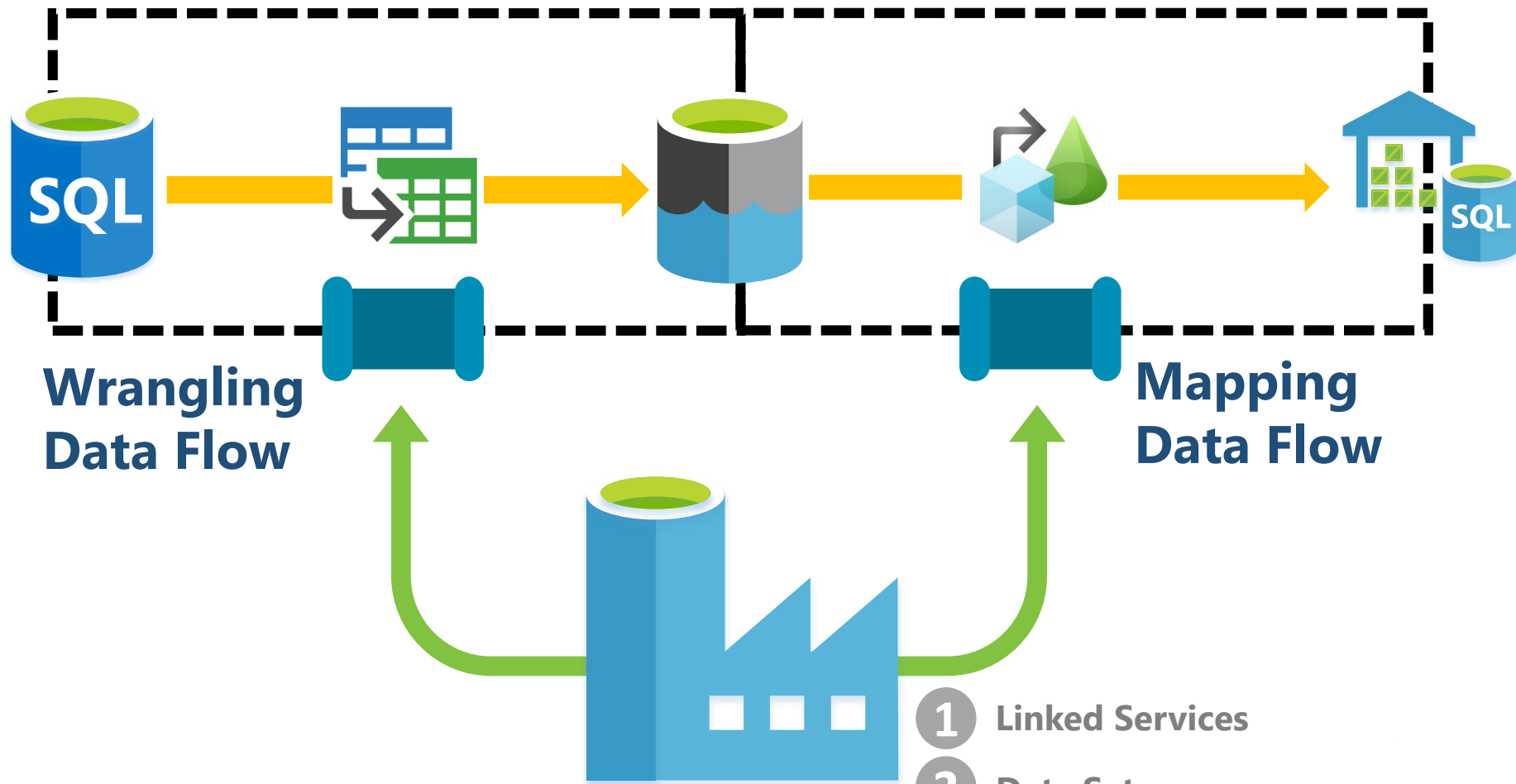
Data Factory Data Flows



Data Factory Control Flow Components



Data Factory Data Flows

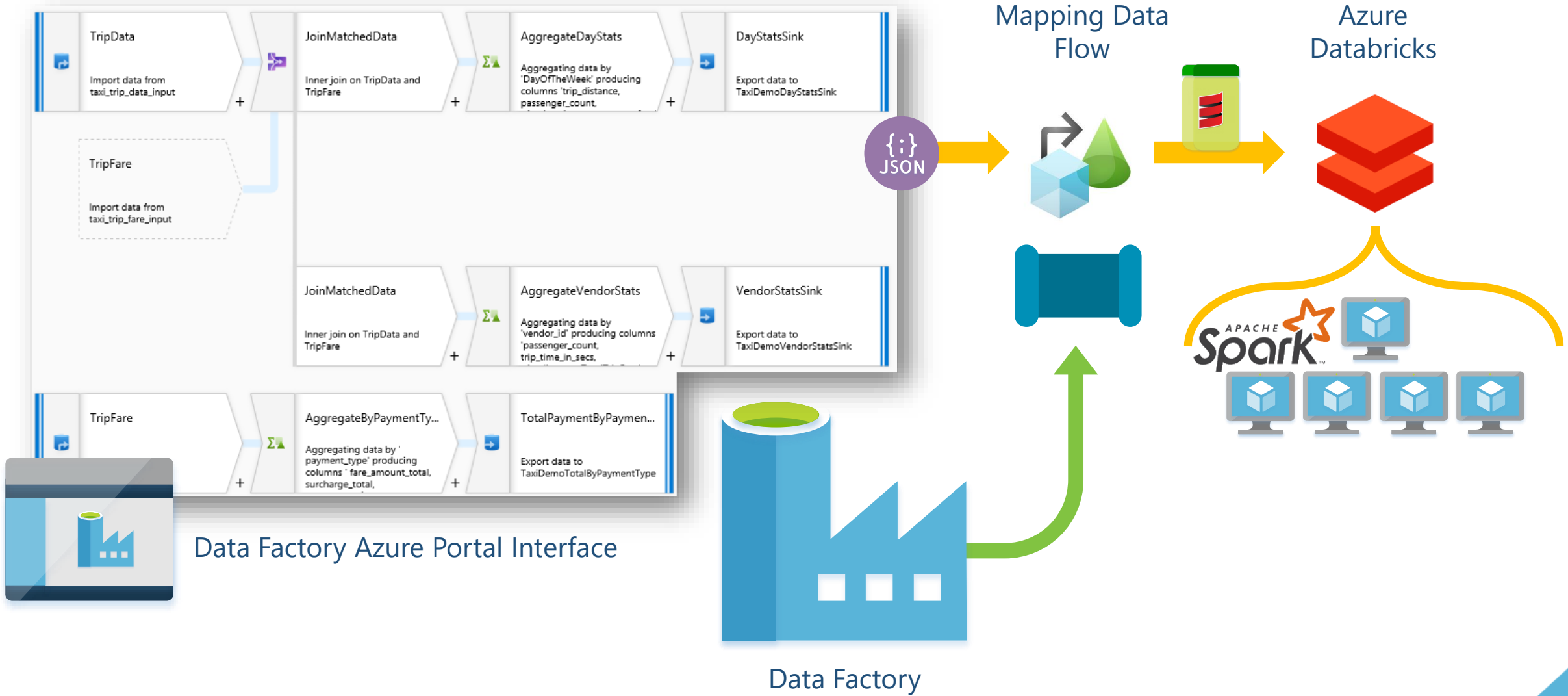


- 1 Linked Services
- 2 Data Sets
- 3 Activities
- 4 Pipelines
- 5 Triggers

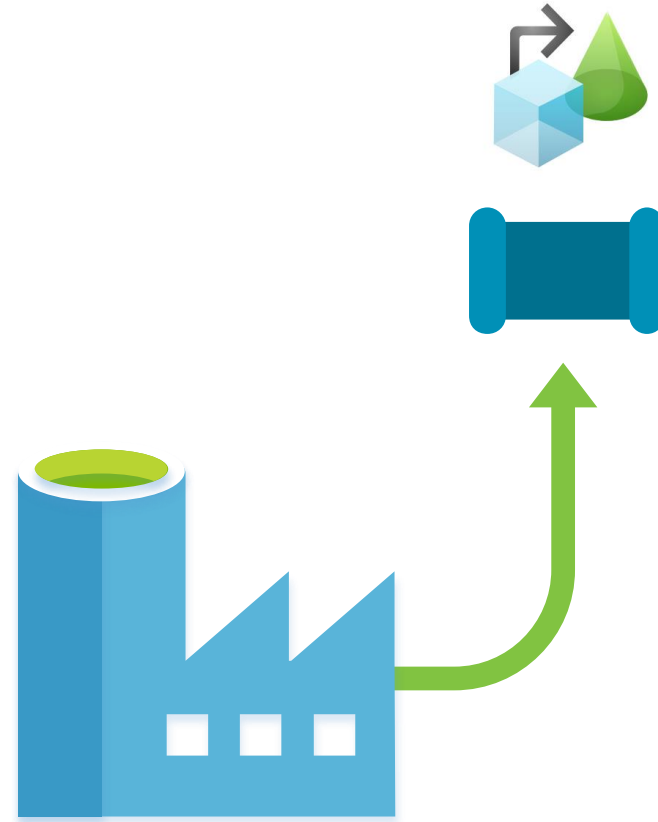
Mapping Data Flows



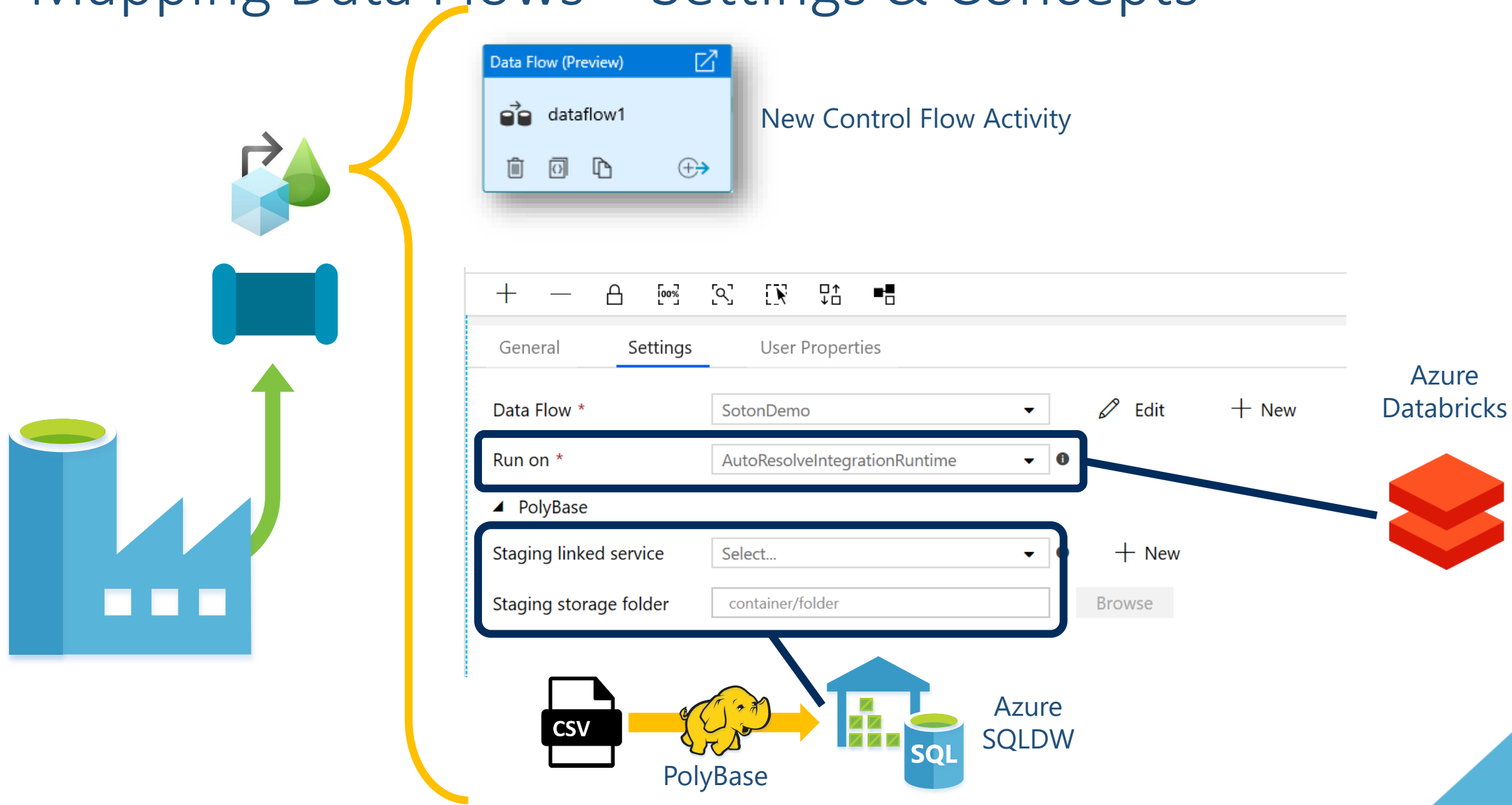
What is a Mapping Data Flow?



Mapping Data Flows



Mapping Data Flows – Settings & Concepts



Integration Runtimes



1

Azure
Integration Runtime

Movement Hours



Activity
Orchestration







2

SSIS
Integration Runtime

SSIS Package
Execution









3

Self Hosted
Integration Runtime

Gateway Access

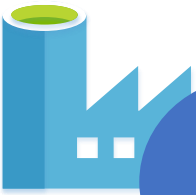


Activity
Orchestration





Integration Runtimes – Mapping Data Flow Cluster



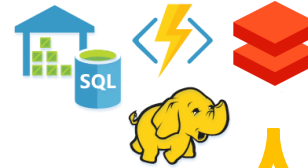
1

Azure Integration Runtime

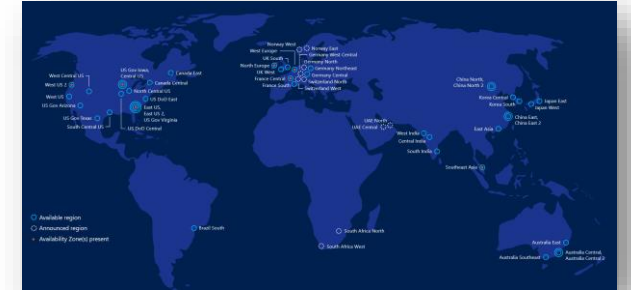
Movement Hours



Activity
Orchestration



Flexible Region



▲ Data Flow run time

Compute Type *

General Purpose

Core count *

4 (4 Driver Cores)

Time to live (in minutes)

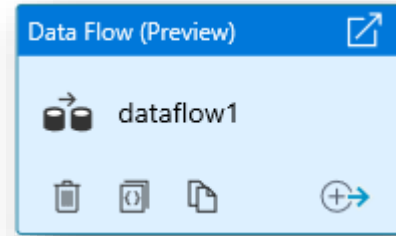
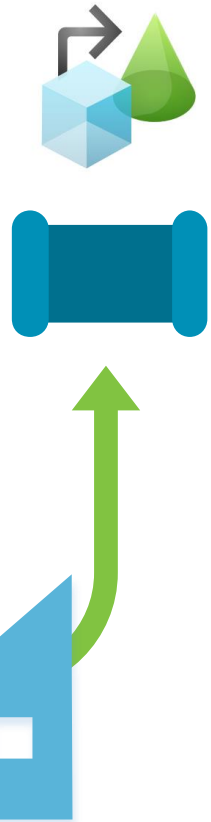
Time to live feature is coming soon

- General Purpose
- Memory Optimised
- Compute Optimised

=



Mapping Data Flows – Settings & Concepts



New Control Flow Activity

Settings panel for Data Flow (Preview) showing the 'Settings' tab.

General Settings:

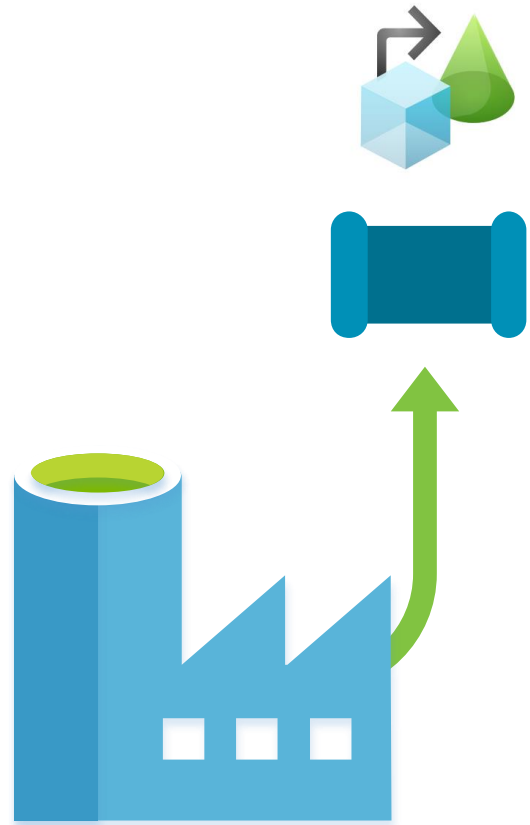
- Data Flow *: SotonDemo [Edit] [New]
- Run on *: AutoResolveIntegrationRuntime [Info]
- Staging linked service: Select... [Info] [New]
- Staging storage folder: container/folder [Browse]

Additional settings include PolyBase, Staging linked service, and Staging storage folder.

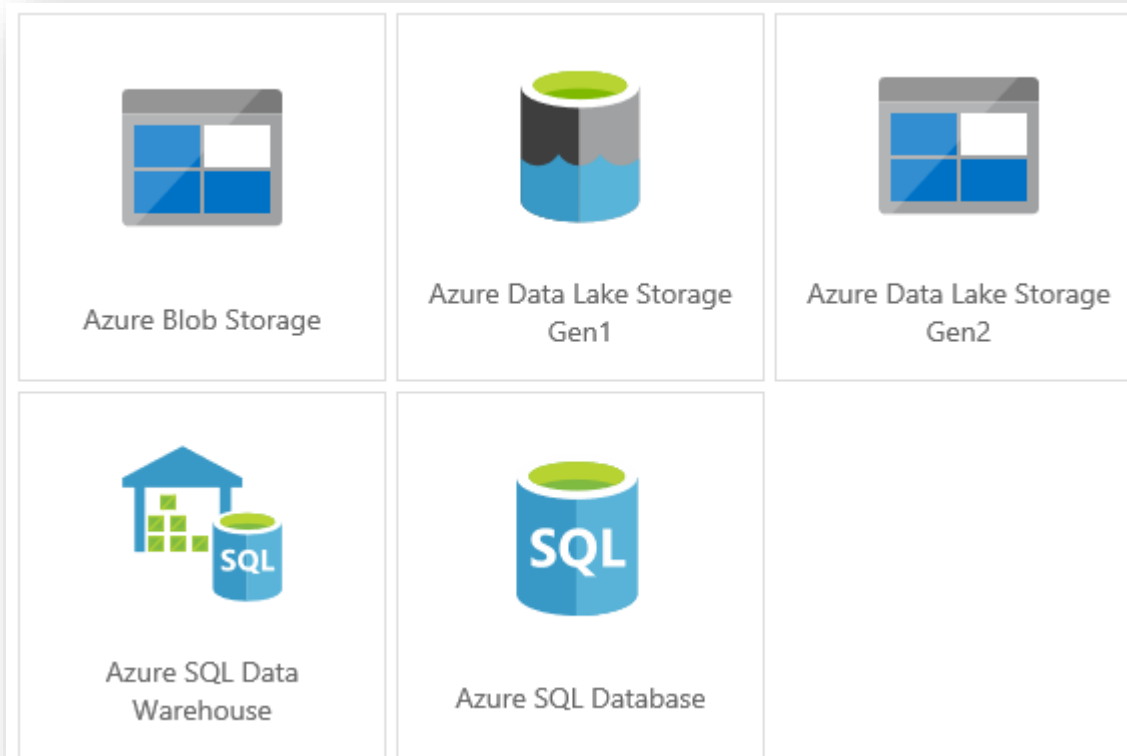
Azure Databricks



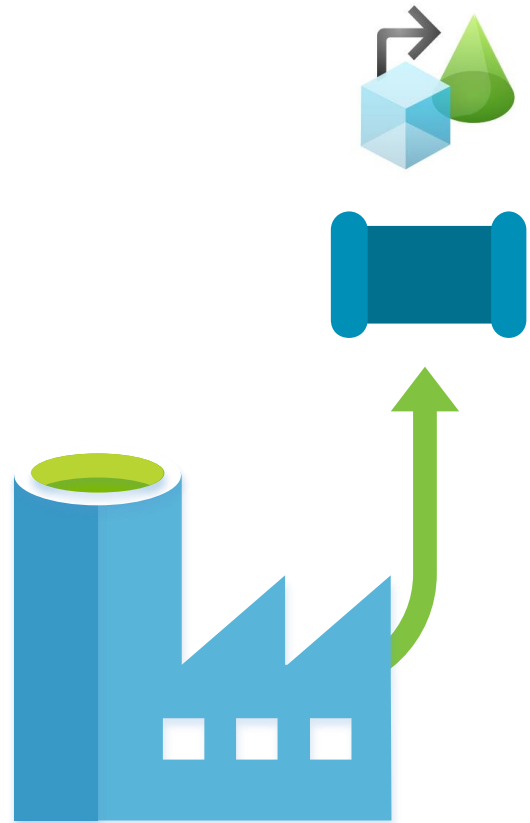
Mapping Data Flows – Settings & Concepts



Currently Available:



Mapping Data Flows – Settings & Concepts



source1

Add source dataset

+

Source Settings

Output stream name * Table1

Source dataset * GenericSQLTable Edit + New

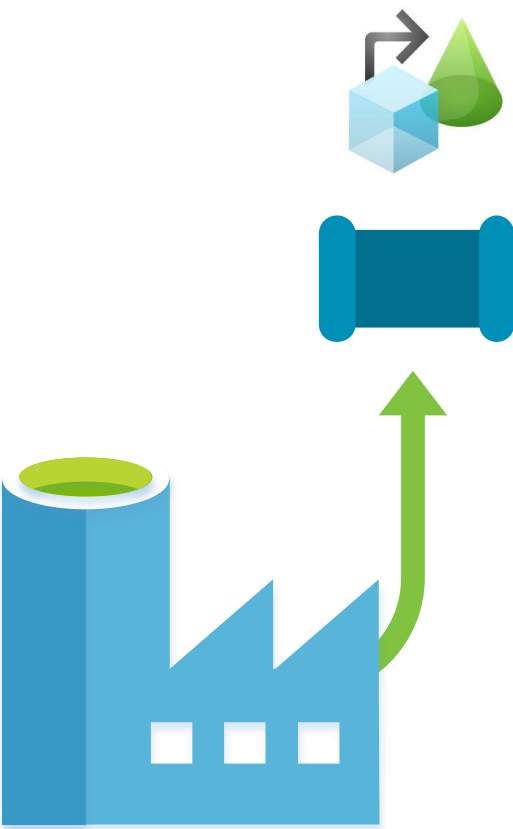
Options

- ☒ Allow schema drift ⓘ
- ☒ Validate schema ⓘ

Sampling * ☒ Enable ☐ Disable ⓘ

Rows limit 100

Mapping Data Flows – Settings & Concepts



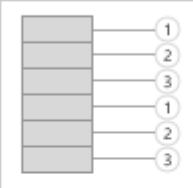
sink1

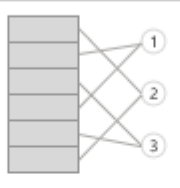
Add sink dataset

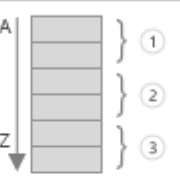
SinkSettingsMappingOptimizeInspectData Preview

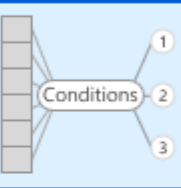
Partition option *
☐ Use current partitioning ☐ Single partition ☒ Set Partitioning

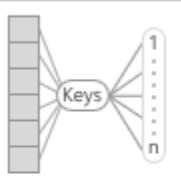
Partition type *


Round Robin


Hash


Dynamic Range


Fixed Range

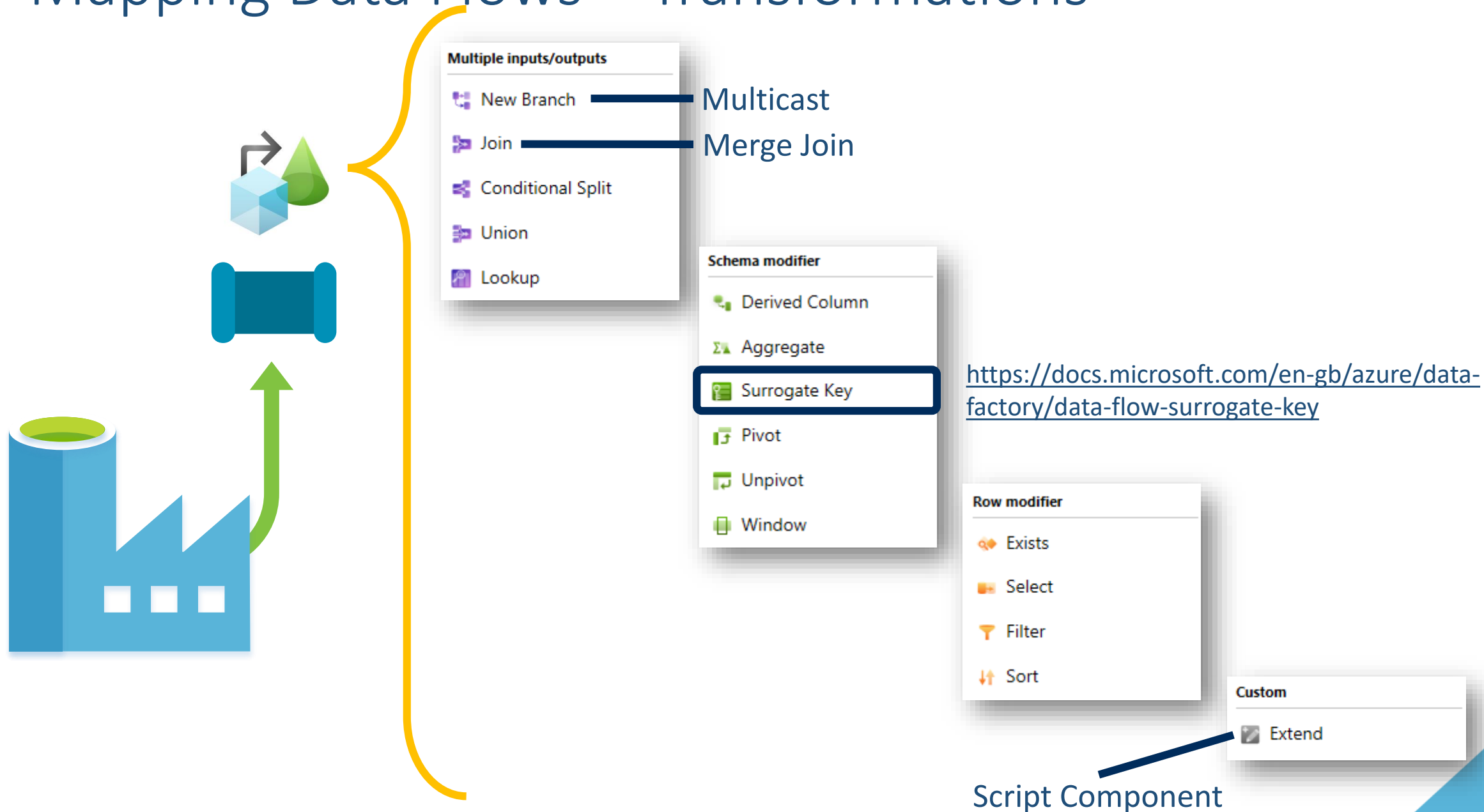

Key

Number of partitions *

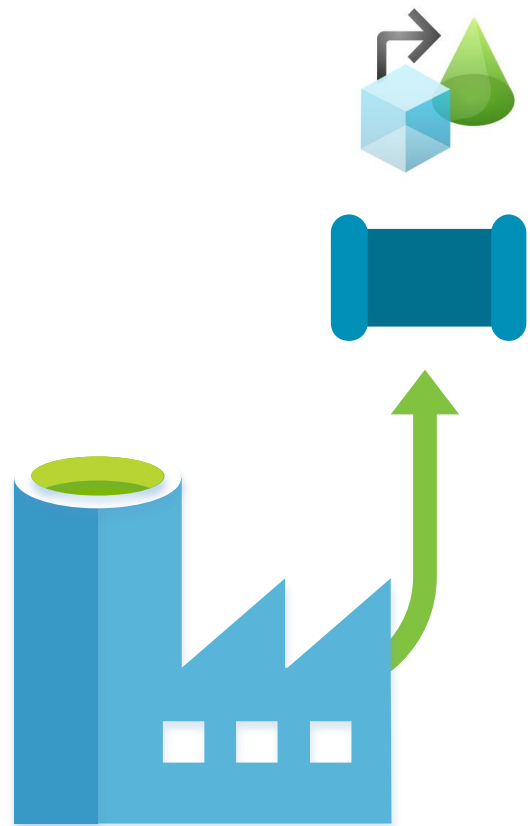
Condition to partition *

Condition
 ANY + -

Mapping Data Flows – Transformations



Mapping Data Flows – Expression Builder



Visual Expression Builder

Currently working on: year

Filter...

String Math Date Logical Input

abc md5(ANY expression)

123 nextSequence()

abc **regexExtract(abc string, abc regex to find, ANY match group 1-based index)**

✕ regexMatch(abc string, abc regex to match)

abc right(abc string to subset, ANY number of characters)

✕

Extract a matching substring for a given regex pattern. The last parameter identifies the match group and is defaulted to 1 if omitted. Use `<regex>` (back quote) to match a string without escaping

Examples

1. regexExtract('Cost is between 600 and 800 dollars', '(\d+) and (\d+)', 2) -> '800'
2. regexExtract('Cost is between 600 and 800 dollars', '(\d+) and (\d+)', 2) -> '800'

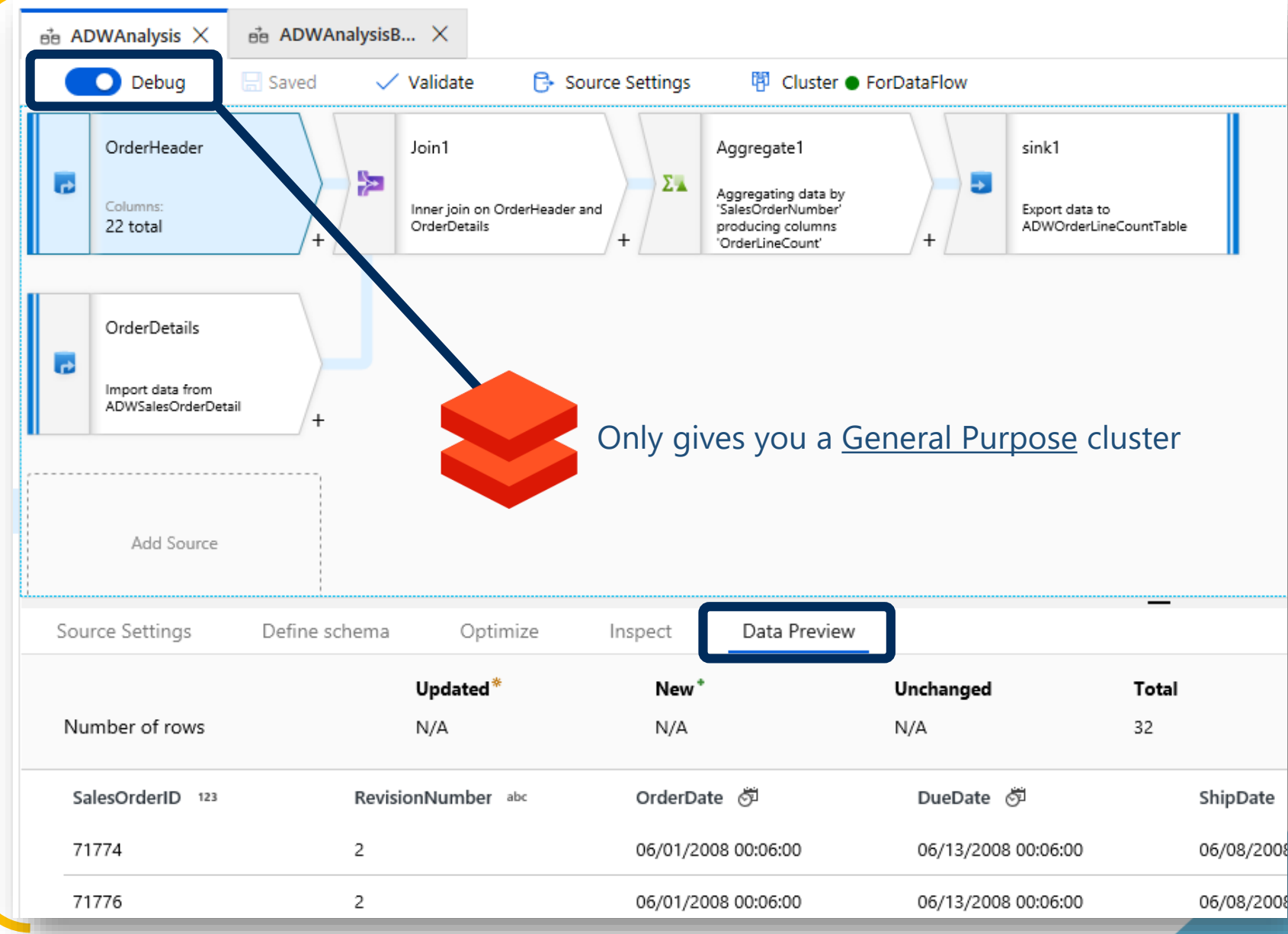

+ - * / ||

Data preview

⚠ Please turn on the debug mode and wait until cluster is ready to preview data...

Output: year 123	title abc
-	-

Mapping Data Flows – Debug Mode



Only gives you a General Purpose cluster

ADWAnalysis X ADWAnalysisB... X

☒ Debug Saved Validate Source Settings Cluster ForDataFlow

OrderHeader
Columns: 22 total

Join1
Inner join on OrderHeader and OrderDetails

Aggregate1
Aggregating data by 'SalesOrderNumber' producing columns 'OrderLineCount'

sink1
Export data to ADWOrderLineCountTable

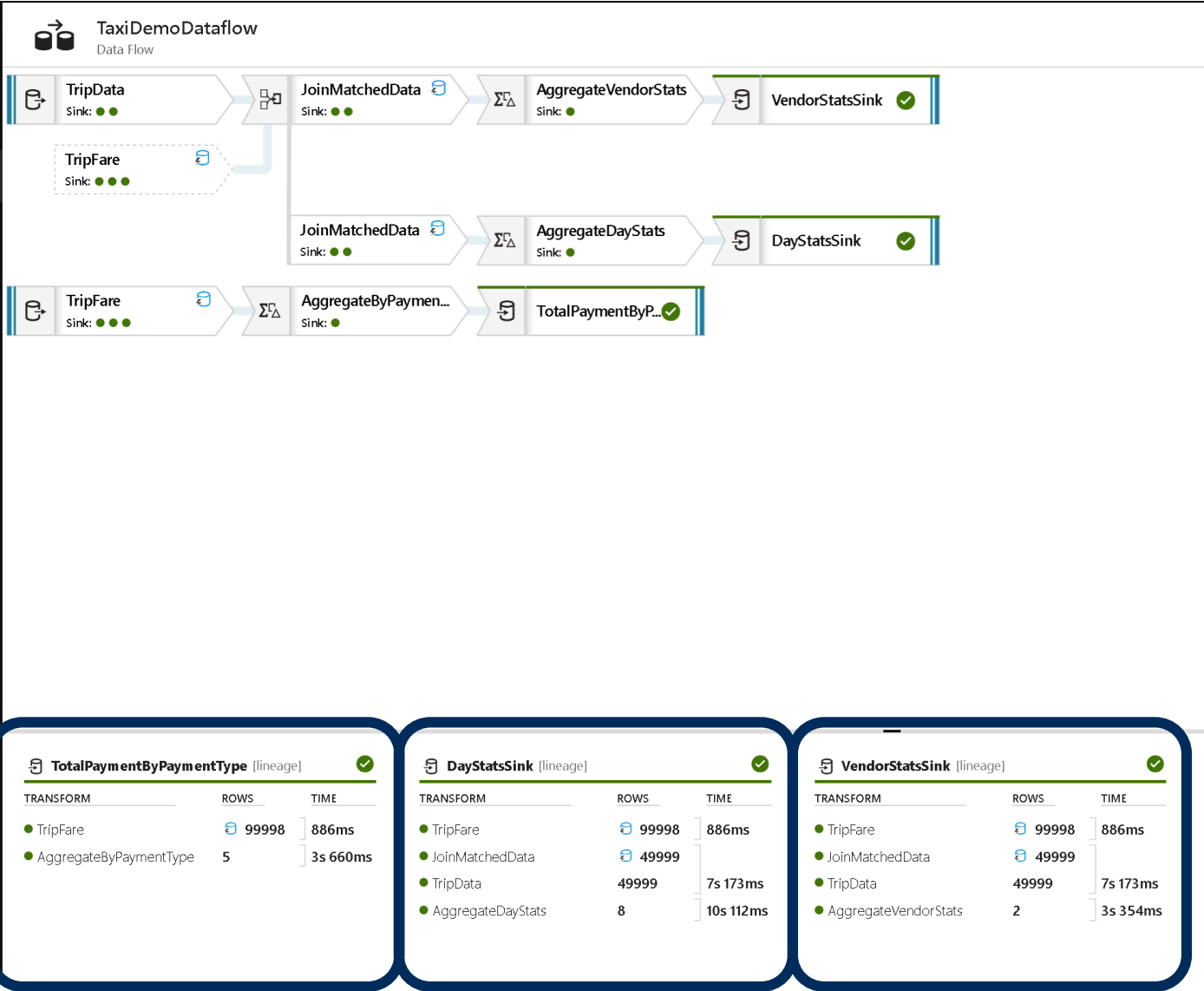
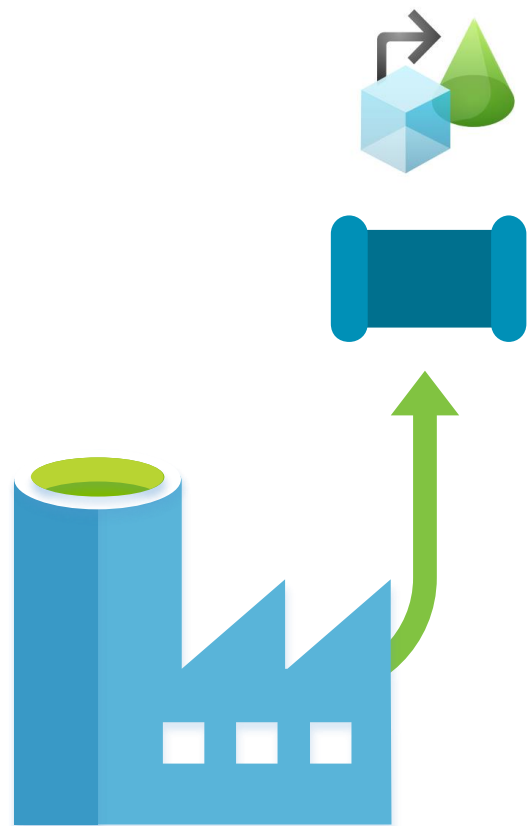
OrderDetails
Import data from ADWSalesOrderDetail

Add Source

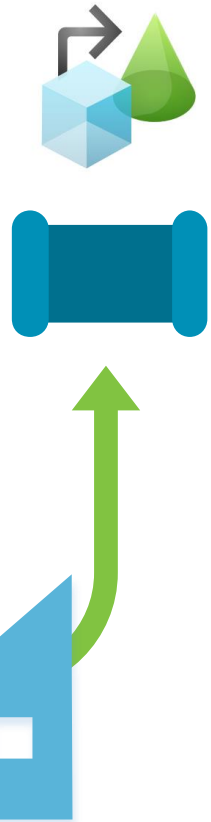
Source Settings Define schema Optimize Inspect **Data Preview**

	Updated*	New*	Unchanged	Total
Number of rows	N/A	N/A	N/A	32
SalesOrderID 123	RevisionNumber abc	OrderDate 🕒	DueDate 🕒	ShipDate
71774	2	06/01/2008 00:06:00	06/13/2008 00:06:00	06/08/2008
71776	2	06/01/2008 00:06:00	06/13/2008 00:06:00	06/08/2008

Mapping Data Flows – Monitoring



Mapping Data Flows



1

Activity

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-overview>

2

Source & Sink

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-schema-drift>

3

Transformations

<https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-aggregate>

4

Expression Builder

<https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-expression-functions>

5

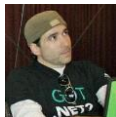
Debug Mode

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-debug-mode>

6

Monitoring

<https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-monitoring>



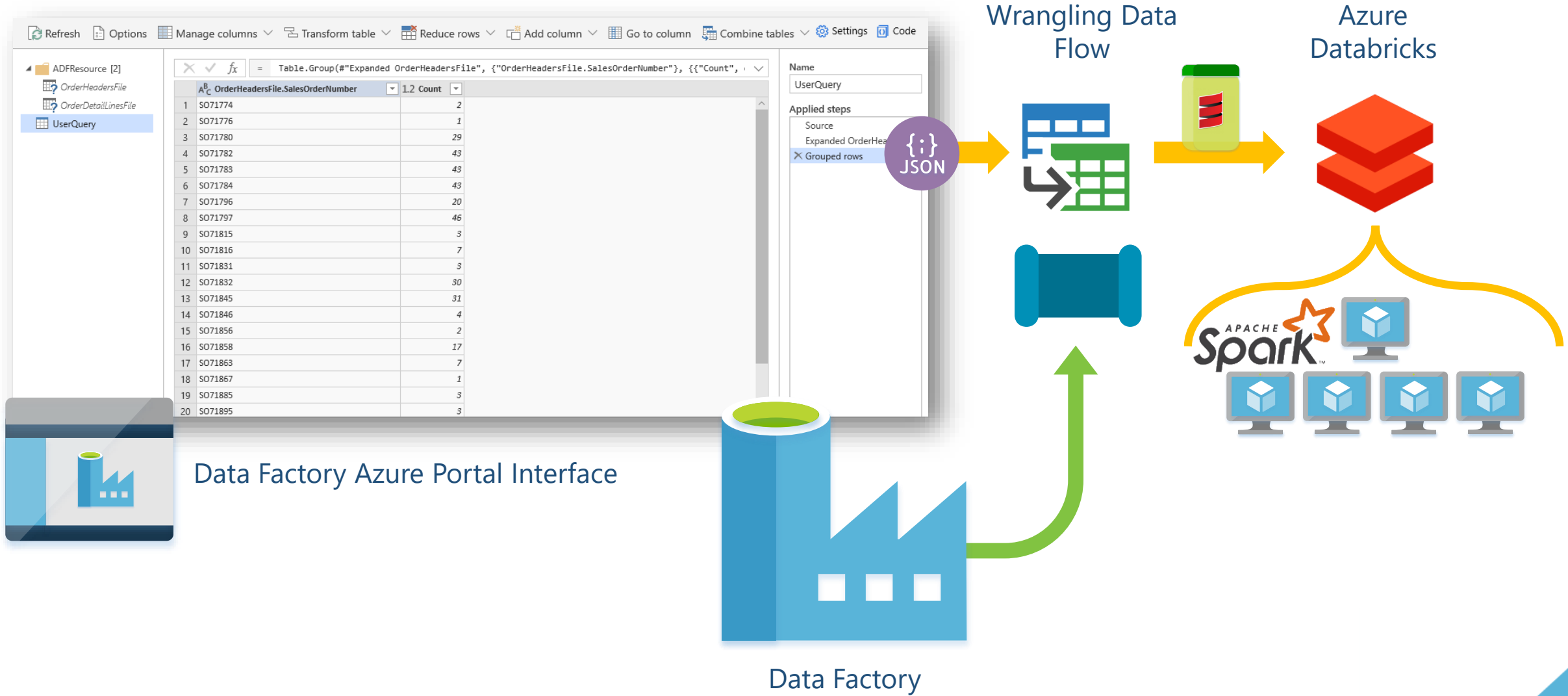
Mark Kromer

<https://github.com/kromerm/adfdataflowdocs>

Wrangling Data Flows



What is a Wrangling Data Flow?



What is a Wrangling Data Flow?

Refresh

Options

Manage columns

Transform table

Reduce rows

Add column

Go to column

Combine tables

Settings

Code

ADFRsource [2]

OrderHeadersFile

OrderDetailLinesFile

UserQuery

fx

= Table.Group(#"Expanded OrderHeadersFile", {"OrderHeadersFile.SalesOrderNumber"}, {"Count",

AB OrderHeadersFile.SalesOrderNumber

1.2 Count

1 SO71774 2

2 SO71776 1

3 SO71780 29

4 SO71782 43

5 SO71783 43

6 SO71784 43

7 SO71796 20

8 SO71797 46

9 SO71815 3

10 SO71816 7

11 SO71831 3

12 SO71832 30

13 SO71845 31

14 SO71846 4

15 SO71856 2

16 SO71858 17

17 SO71863 7

18 SO71867 1

19 SO71885 3

20 SO71895 3

Name

UserQuery

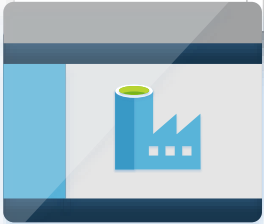
Applied steps

Source

Expanded OrderHeadersF...

Grouped rows

Power BI Desktop



Data Factory

Close & Apply

New Source

Recent Sources

Enter Data

Data source settings

Manage Parameters

Refresh Preview

Properties

Advanced Editor

Manage

Choose Columns

Remove Columns

Keep Rows

Remove Rows

Sort

Split Column

Group By

Data Type: Whole Number

Use First Row as Headers

Replace Values

Merge Queries

Append Queries

Combine Files

Combine

Queries [3]

SpeakingLog

EventLogos

TagsToTalkId

fx

= Table.RenameColumns(dbo_SpeakingLog,{{"LogId", "TalkId"}})

1 2 3 TalkId

TalkDate

AB EventName

AB City

AB Country

1 22/06/2016 Guerrilla Lightning Talks Online Online

2 08/07/2016 STEM Stafford England

3 23/07/2016 SQL Saturday Manchester England

4 11/08/2016 User Group Birmingham England

5 14/09/2016 User Group Manchester England

6 21/09/2016 British Computer Society Telford England

7 03/10/2016 Data Relay Birmingham England

8 04/10/2016 Data Relay Cardiff Wales

9 05/10/2016 Data Relay Reading England

10 06/10/2016 Data Relay Nottingham England

11 07/10/2016 Data Relay Leeds England

12 16/11/2016 STEM Trentham England

13 23/11/2016 STEM Trentham England

14 01/12/2016 STEM Trentham England

15 23/01/2017 User Group Exeter England

16 24/01/2017 STEM Telford England

17 31/01/2017 STEM Telford England

18 01/02/2017 User Group Southampton England

19 07/02/2017 User Group Bristol England

20 09/02/2017 User Group Birmingham England

21 10/03/2017 STEM Stone England

22 07/04/2017 SQL Bits Telford England

Query Settings

PROPERTIES

Name

SpeakingLog

All Properties

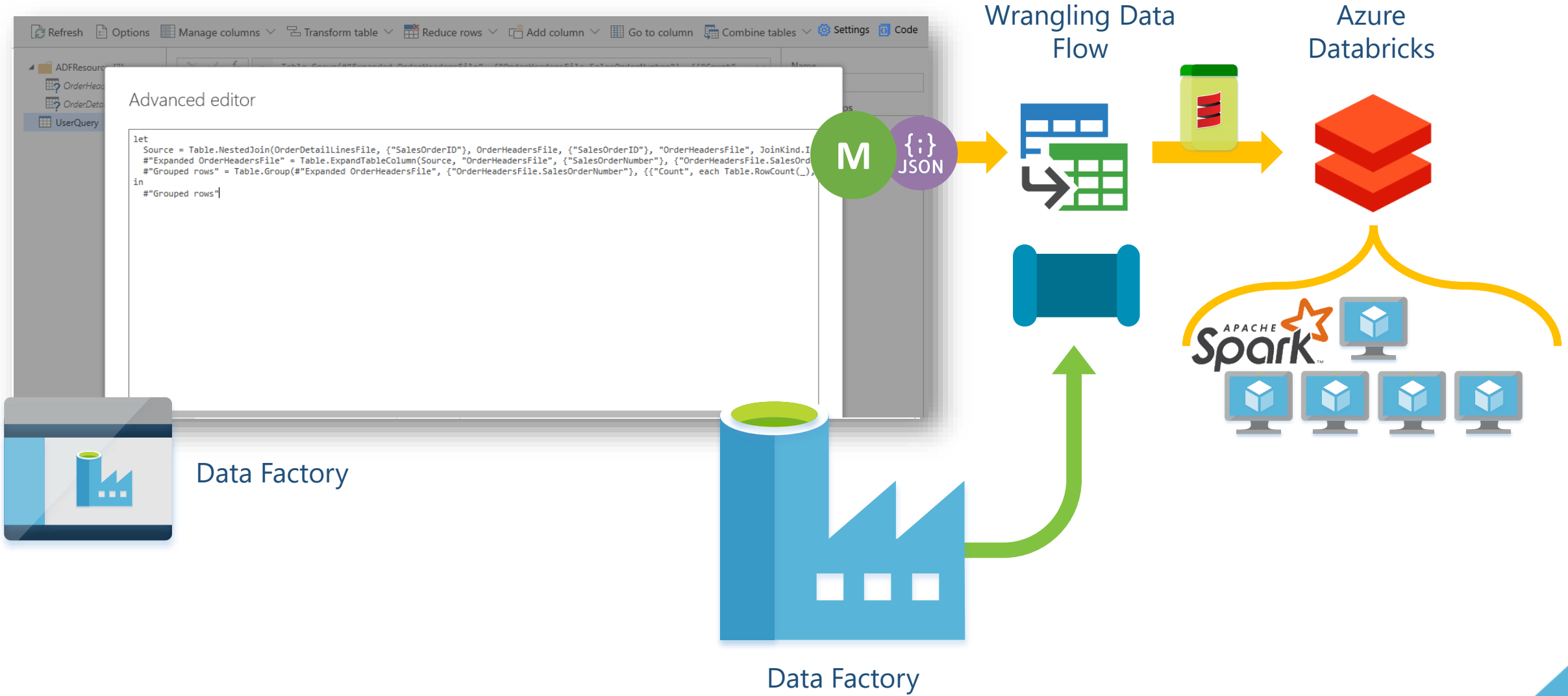
APPLIED STEPS

Source

Navigation

Renamed Columns

What is a Wrangling Data Flow?







A 3D-style blue cylinder with a green circular top. The word "DEMO" is written in white capital letters on the green top. The cylinder has a slight shadow and a white rim at the top.

DEMO

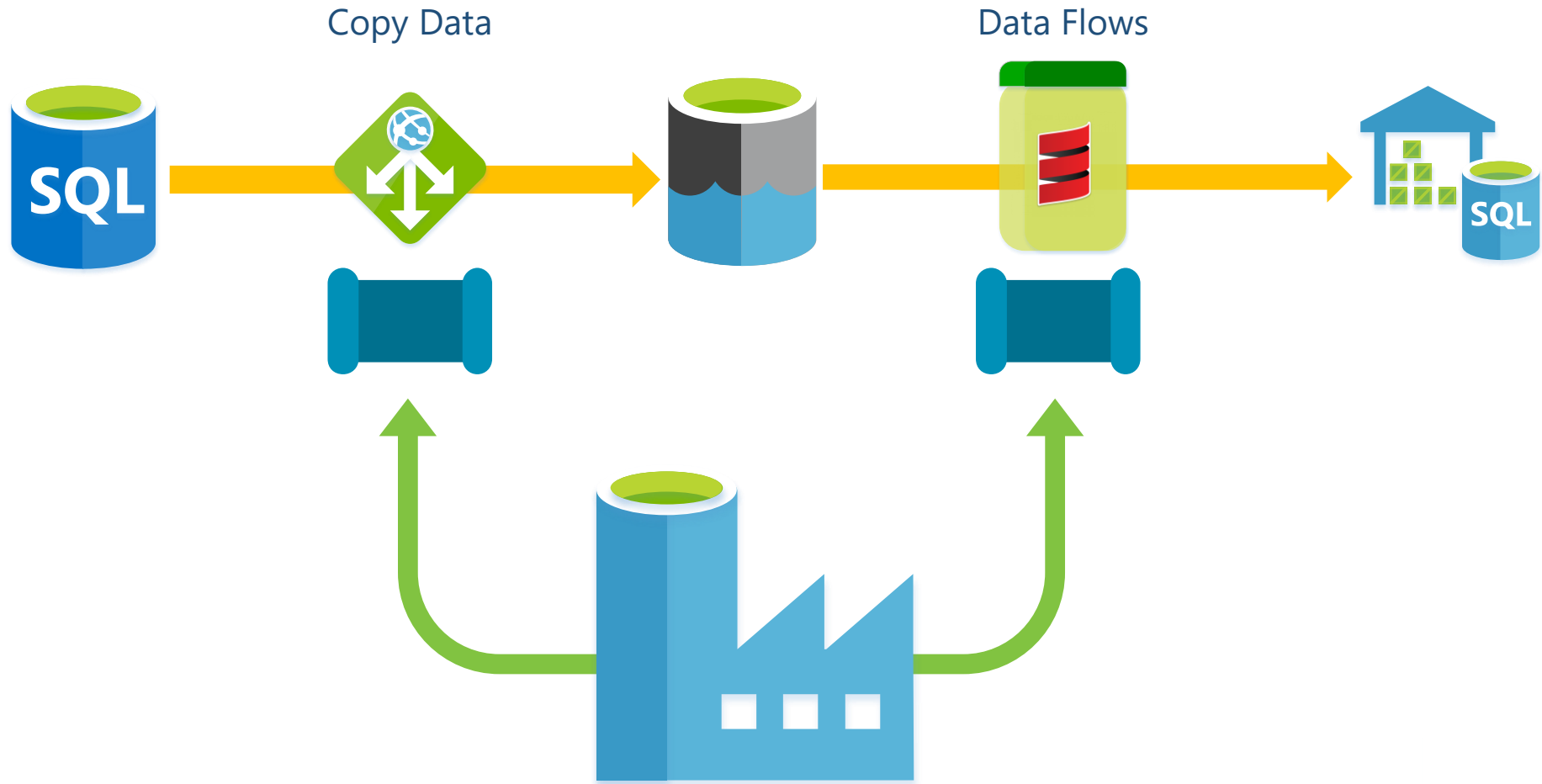
Conclusions



Transformations in Azure Summary

Transformation Method		Graphical UI	Scales Out	Scales Up	Cloud Native Tech
	T-SQL (SQLDB)	✗	✗	✓	✗
	SSIS	✓	✗	✓	✗
	Scala (Databricks)	✗	✓	✓	✓
	Data Factory Data Flows	✓	✓	✓	✓

What is Azure Data Factory?

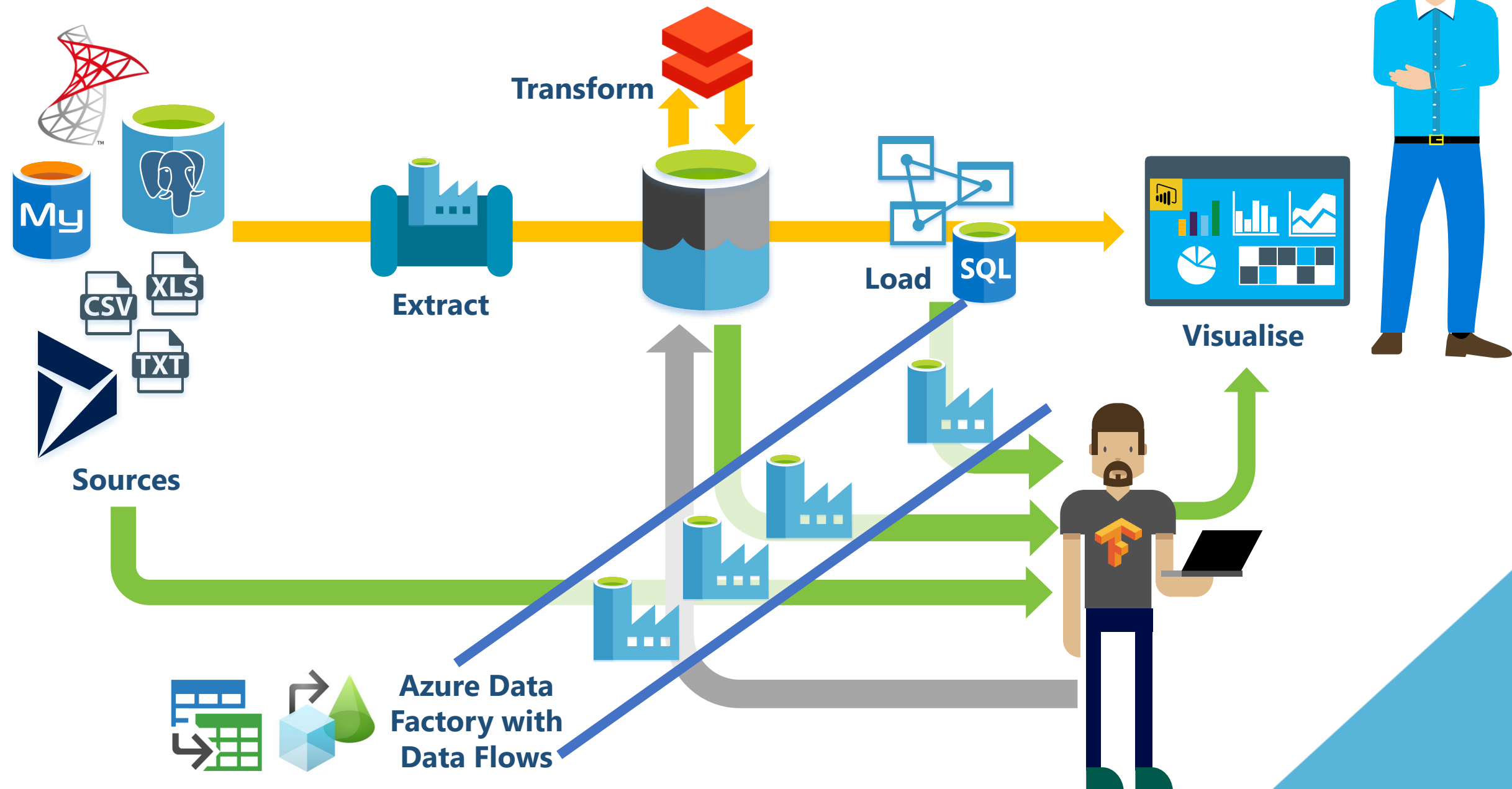


Orchestrator of our solution Control Flow operations.

Orchestrator of our solution Data Flow transformations.

... using cloud native technology in  Azure and now with a user interface for both.

Data Factory for the Data Scientist



Thanks for Listening

Paul Andrew

 @MrPaulAndrew



altius

Email: paul@mrpaulandrew.com

Blog: mrpaulandrew.com

GitHub: github.com/mrpaulandrew

