ETL in Azure Made Easy

with Data Factory Data Flows

Paul Andrew

Principal Consultant & Solution Architect

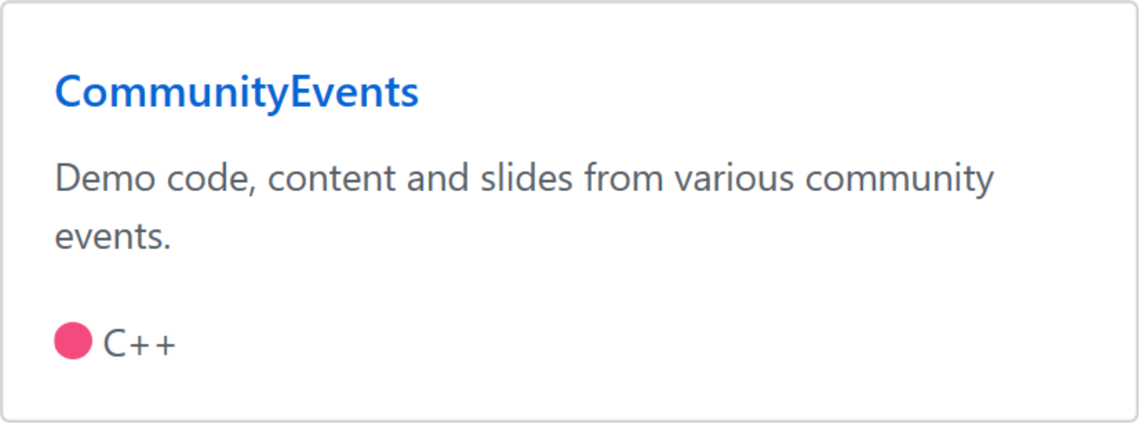
altius PASS

MVP Microsoft® Most Valuable Professional

@MrPaulAndrew

GitHub

https://github.com/mrpaulandrew



CommunityEvents

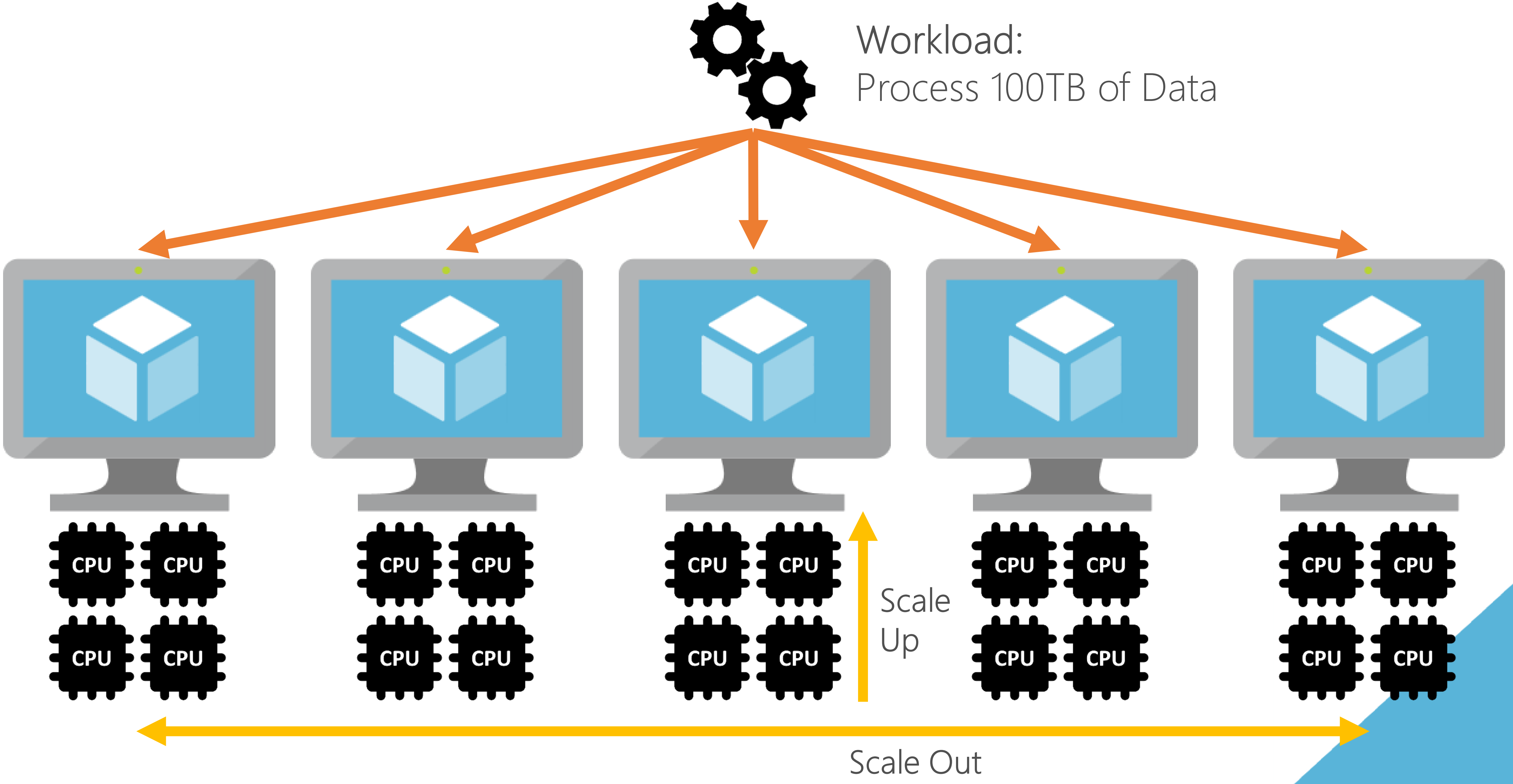
Demo code, content and slides from various community events.

● C++

{Event/Location}-{Month}-{Year}

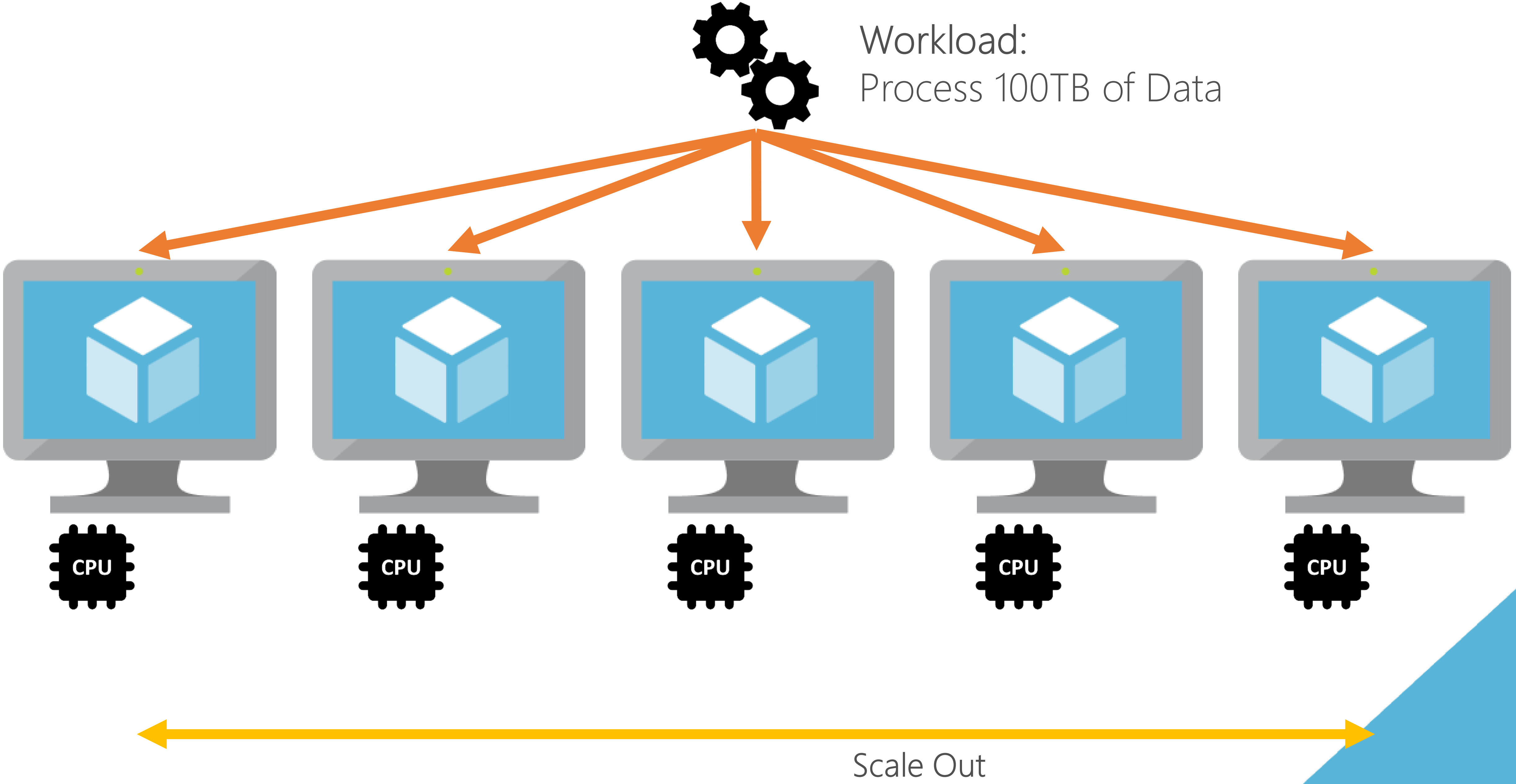
Scale Up vs Scale Out

Scale Up and Scale Out



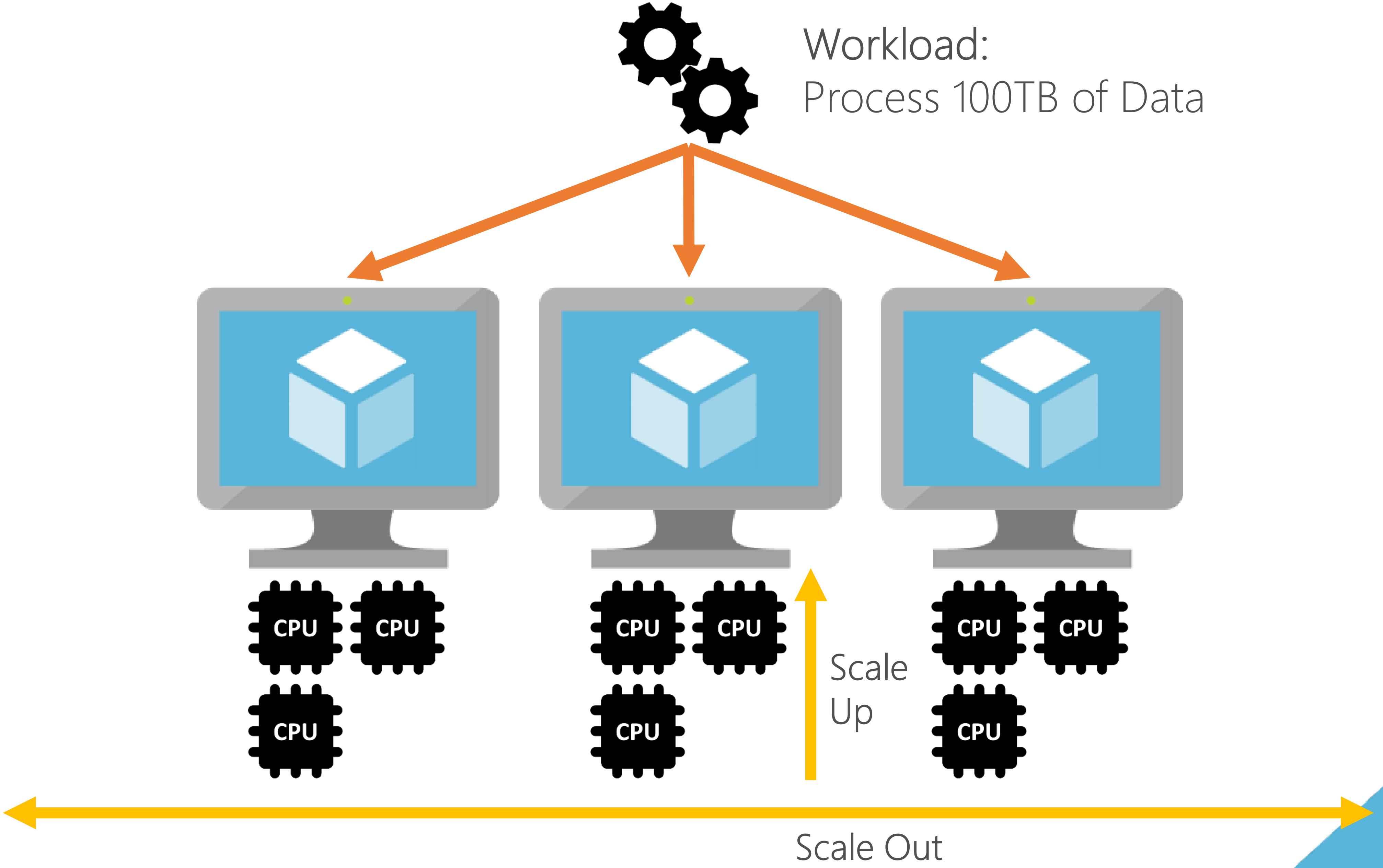
Workload:
Process 100TB of Data

Scale Up and Scale Out


Workload:
Process 100TB of Data

Scale Out

Scale Up and Scale Out



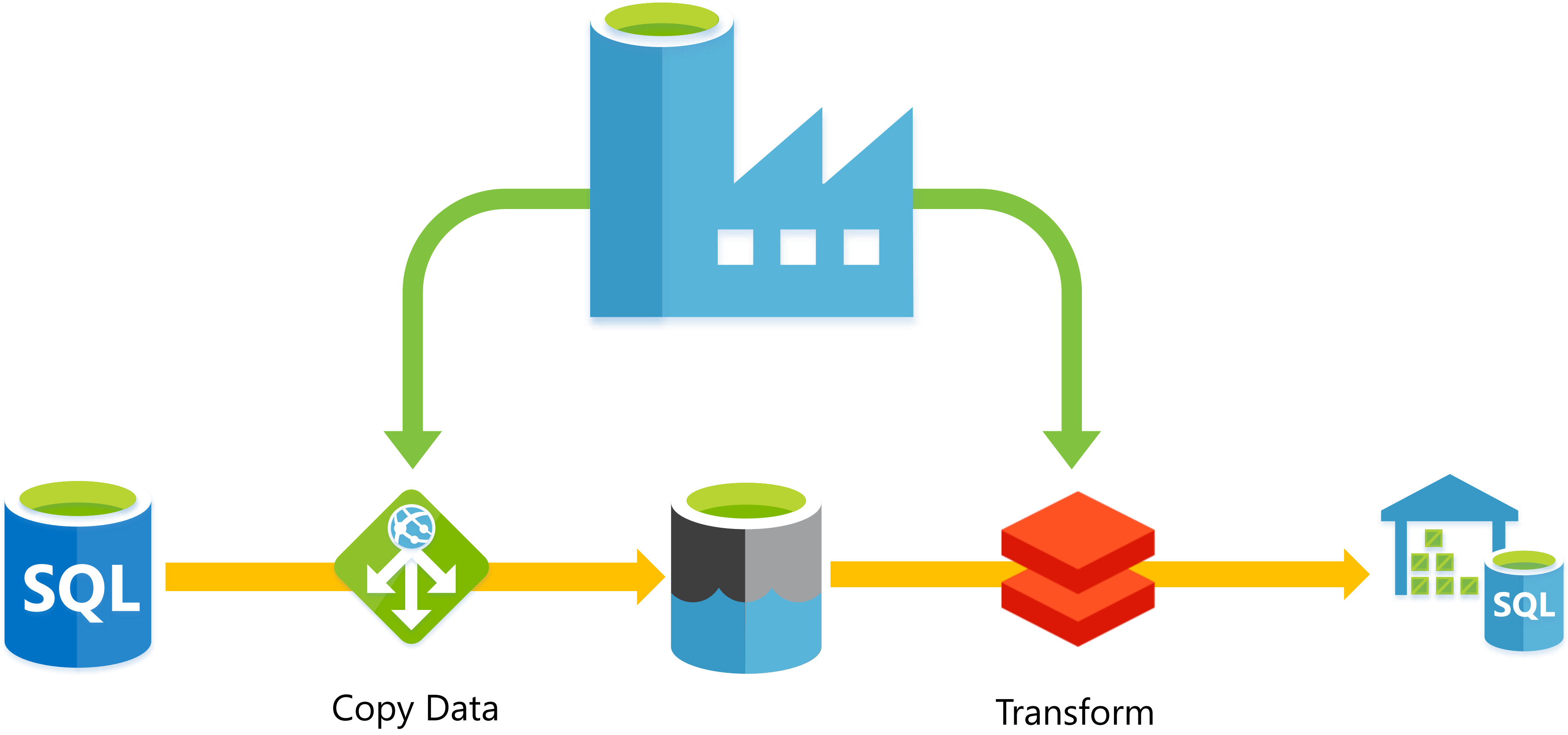
Workload:
Process 100TB of Data

Azure Data Factory

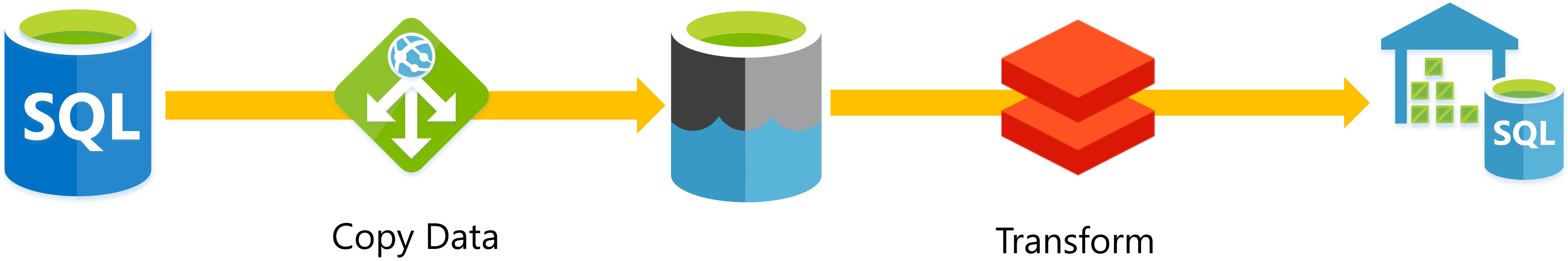
What is Azure Data Factory?

2016
V1
V2
2019

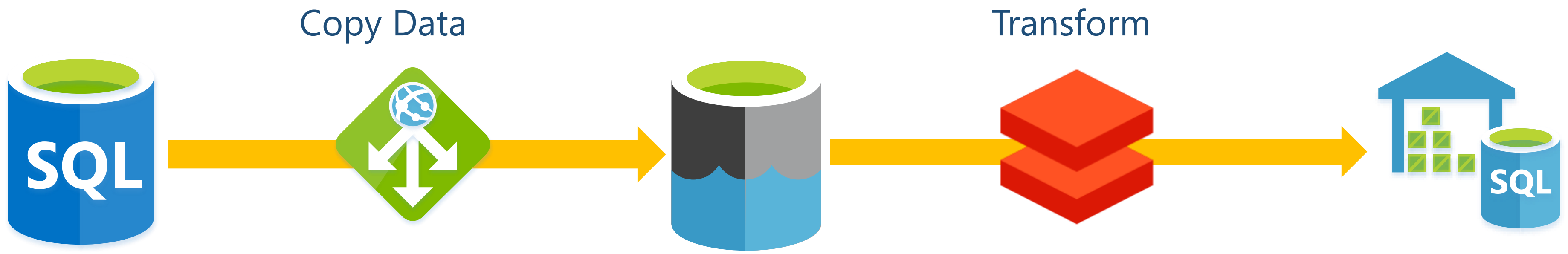
What is Azure Data Factory?



What is Azure Data Factory?



Data Factory Components



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

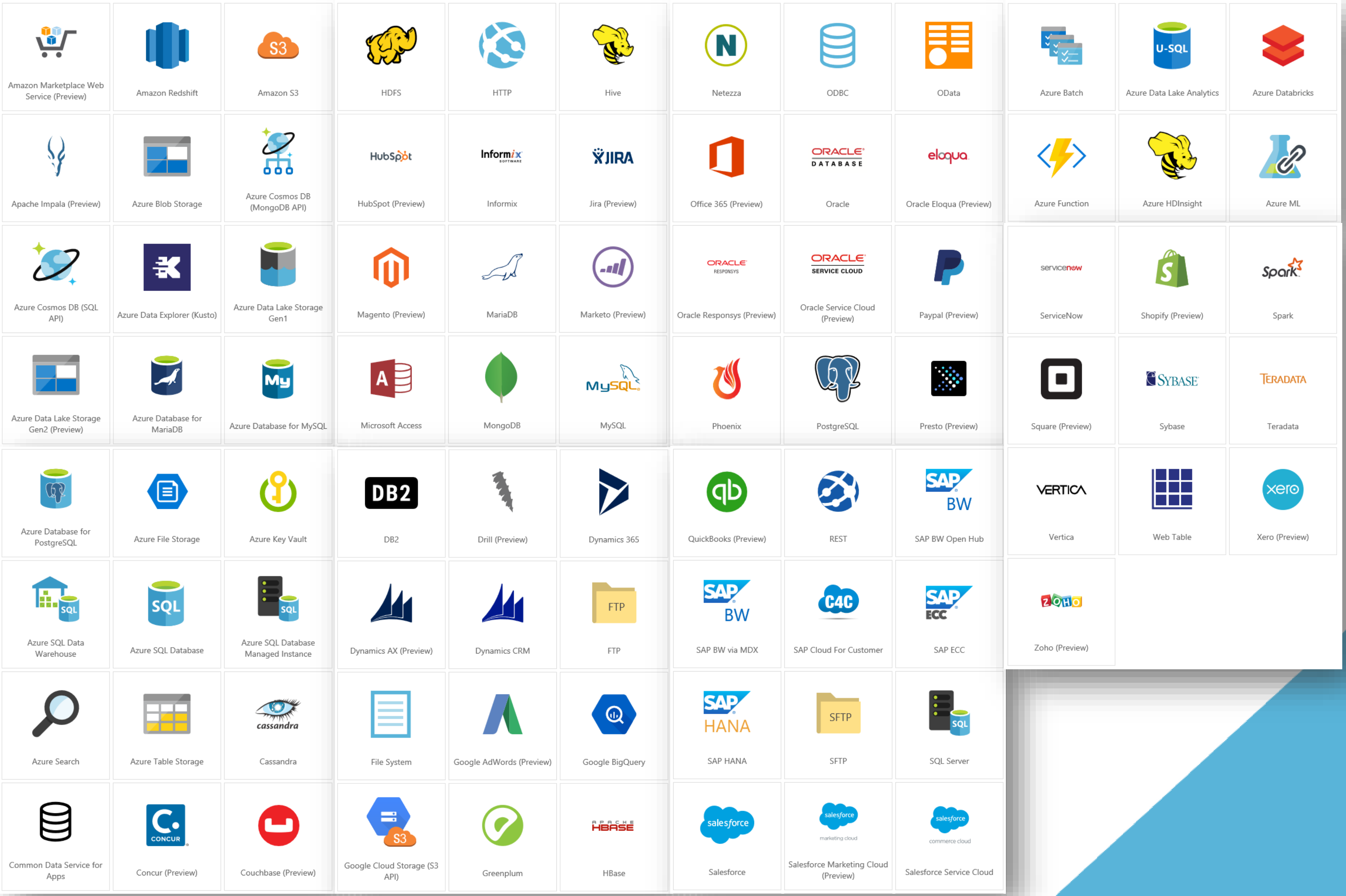
SQL

SQLDBLinkedService

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

Data Factory Components

1 **Linked Services** –



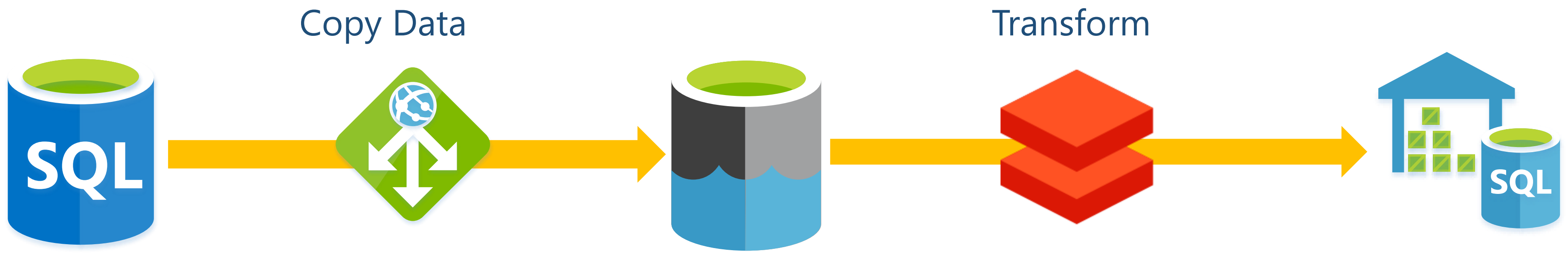
Amazon Marketplace Web Service (Preview)
Amazon Redshift
Amazon S3
Apache Impala (Preview)
Azure Blob Storage
Azure Cosmos DB (MongoDB API)
Azure Cosmos DB (SQL API)
Azure Data Explorer (Kusto)
Azure Data Lake Storage Gen1
Azure Data Lake Storage Gen2 (Preview)
Azure Database for MariaDB
Azure Database for MySQL
Azure Database for PostgreSQL
Azure File Storage
Azure Key Vault
Azure SQL Data Warehouse
Azure SQL Database
Azure SQL Database Managed Instance
Azure Search
Azure Table Storage
Cassandra
Common Data Service for Apps
Concur (Preview)
Couchbase (Preview)

HDFS
HTTP
Hive
HubSpot (Preview)
Informix
Jira (Preview)
Magento (Preview)
MariaDB
Marketo (Preview)
Microsoft Access
MongoDB
MySQL
DB2
Drill (Preview)
Dynamics 365
Dynamics AX (Preview)
Dynamics CRM
FTP
File System
Google AdWords (Preview)
Google BigQuery
Google Cloud Storage (S3 API)
Greenplum
HBase

Netezza
ODBC
OData
Office 365 (Preview)
Oracle
Oracle Eloqua (Preview)
Oracle Responsys (Preview)
Oracle Service Cloud (Preview)
Paypal (Preview)
Phoenix
PostgreSQL
Presto (Preview)
QuickBooks (Preview)
REST
SAP BW Open Hub
SAP BW via MDX
SAP Cloud For Customer
SAP ECC
SAP HANA
SFTP
SQL Server
Salesforce
Salesforce Marketing Cloud (Preview)
Salesforce Service Cloud

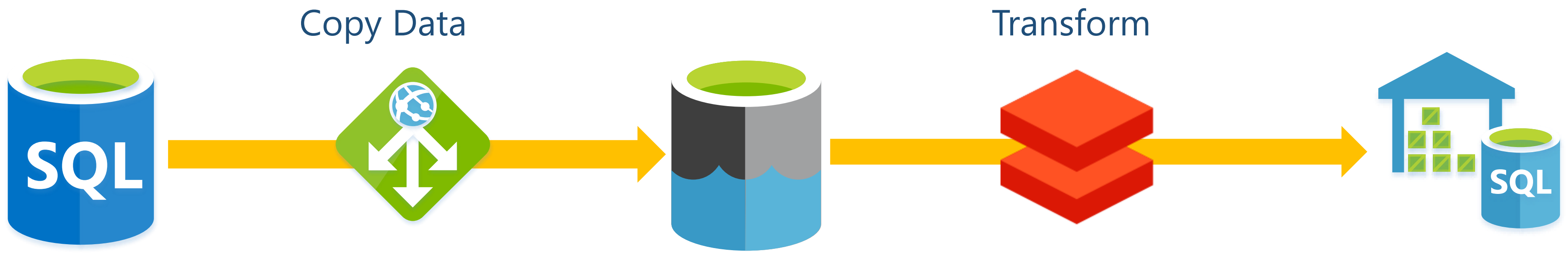
Azure Batch
Azure Data Lake Analytics
Azure Databricks
Azure Function
Azure HDInsight
Azure ML
ServiceNow
Shopify (Preview)
Spark
Square (Preview)
Sybase
Teradata
Vertica
Web Table
Xero (Preview)
Zoho (Preview)

Data Factory Components



1 **Linked Services**

Data Factory Components



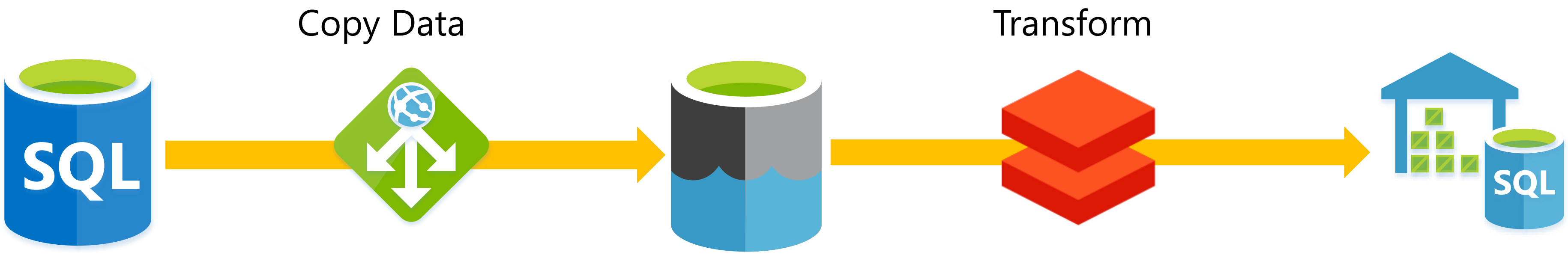
1 **Linked Services**

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**

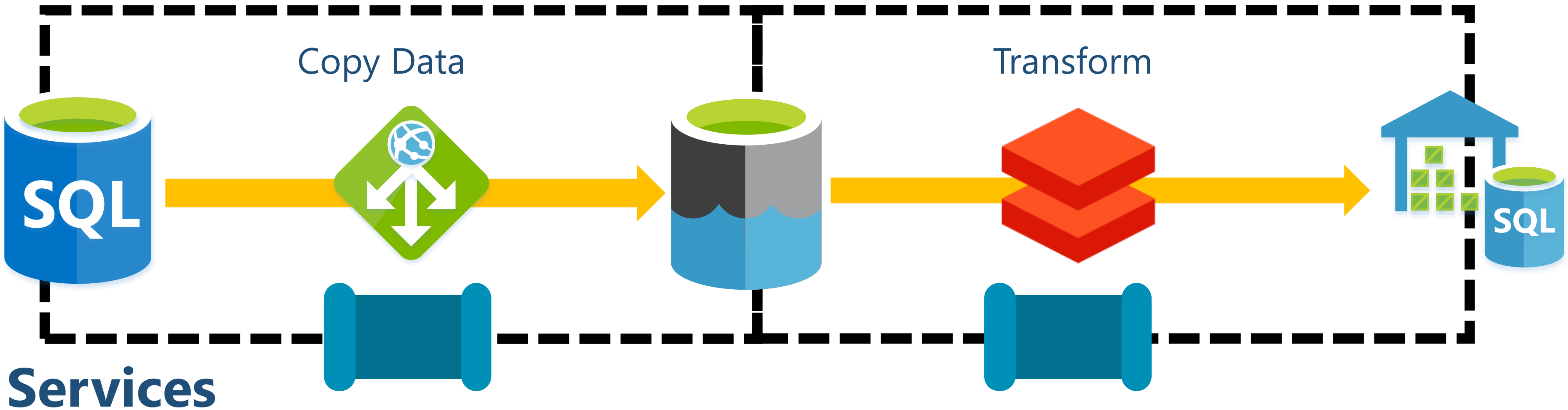
2 **Data Sets**

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

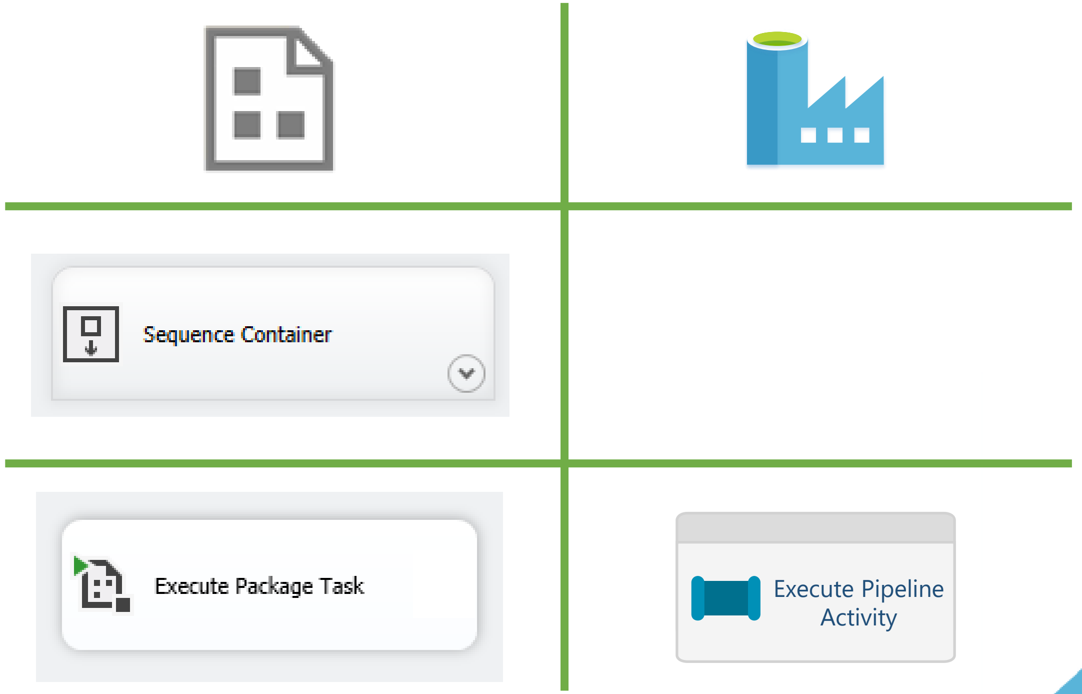
Databricks Notebook Activity

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

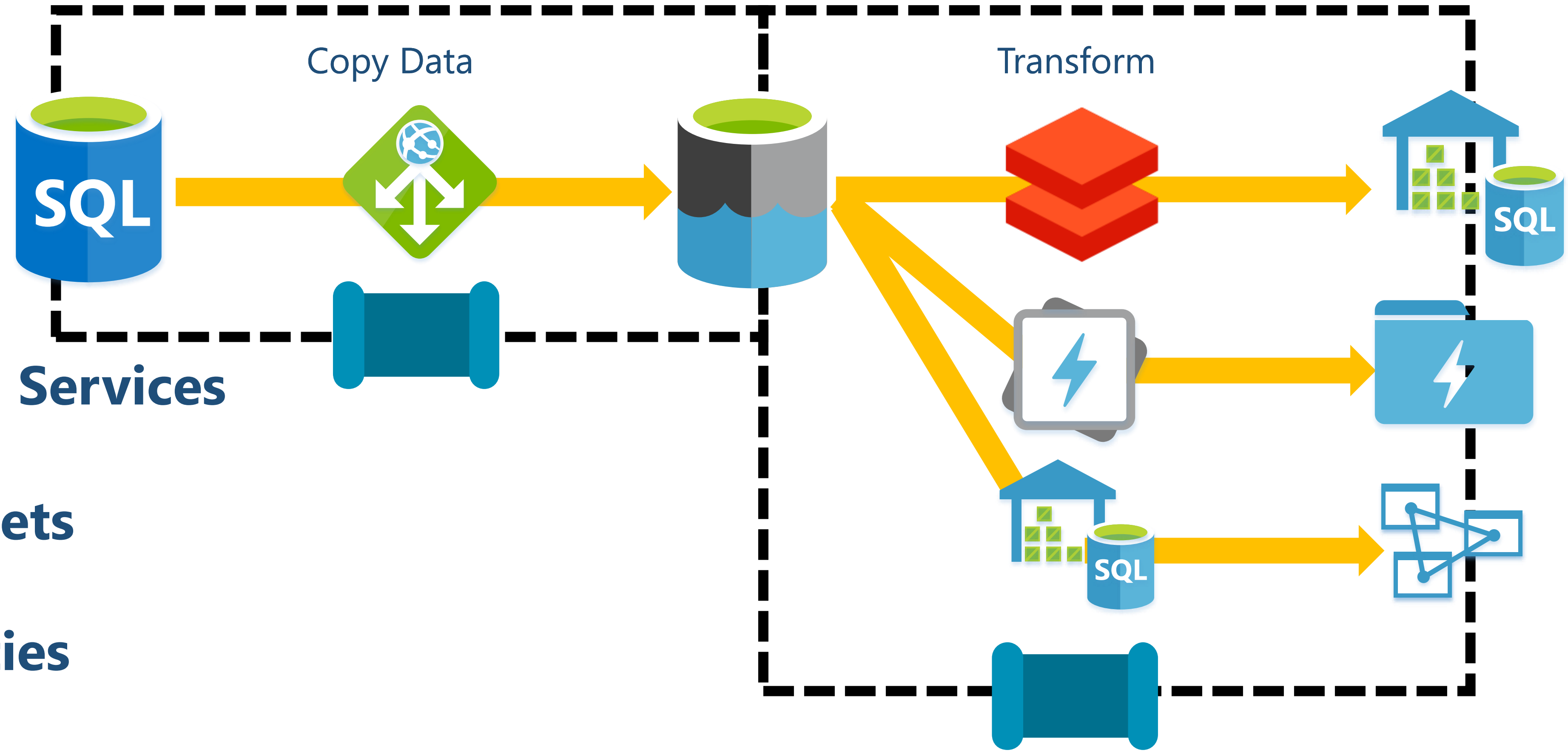
Data Factory Components



1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines** – What groups of work do I want to do?



Data Factory Components



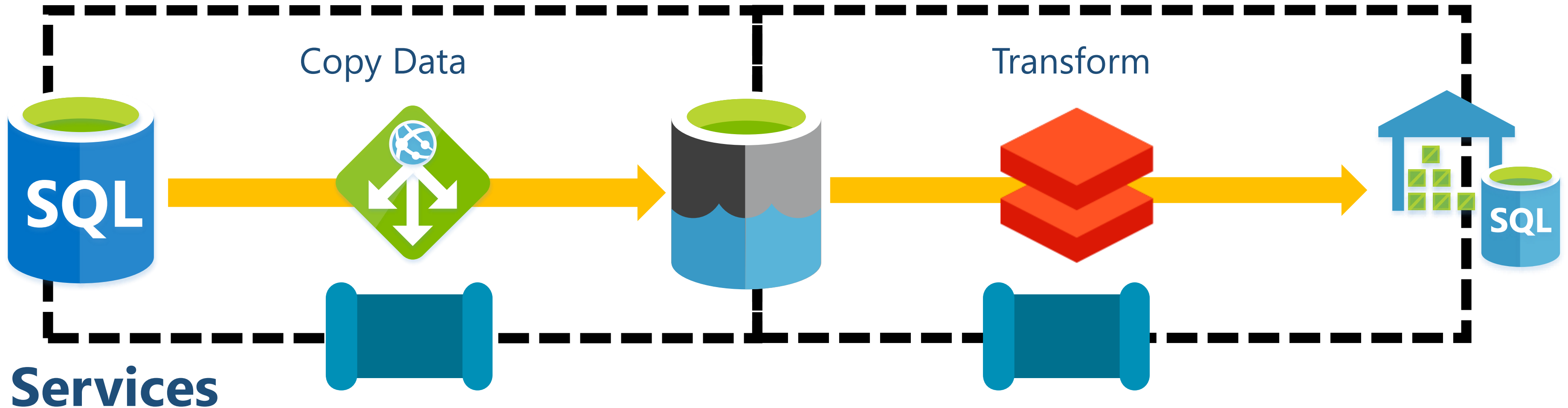
1 **Linked Services**

2 **Data Sets**

3 **Activities**

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

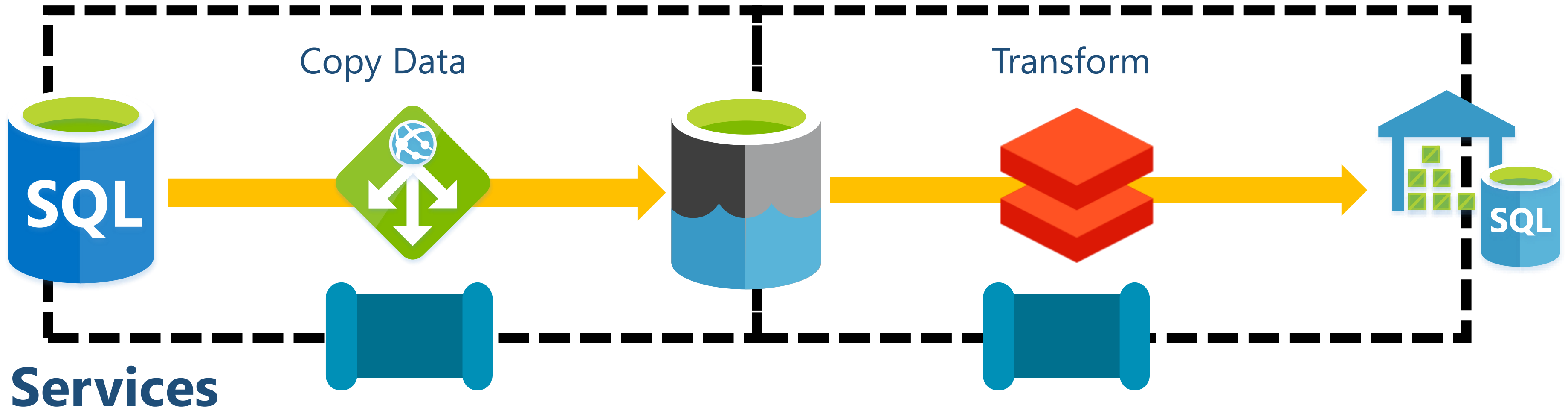
Data Factory Components



1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
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

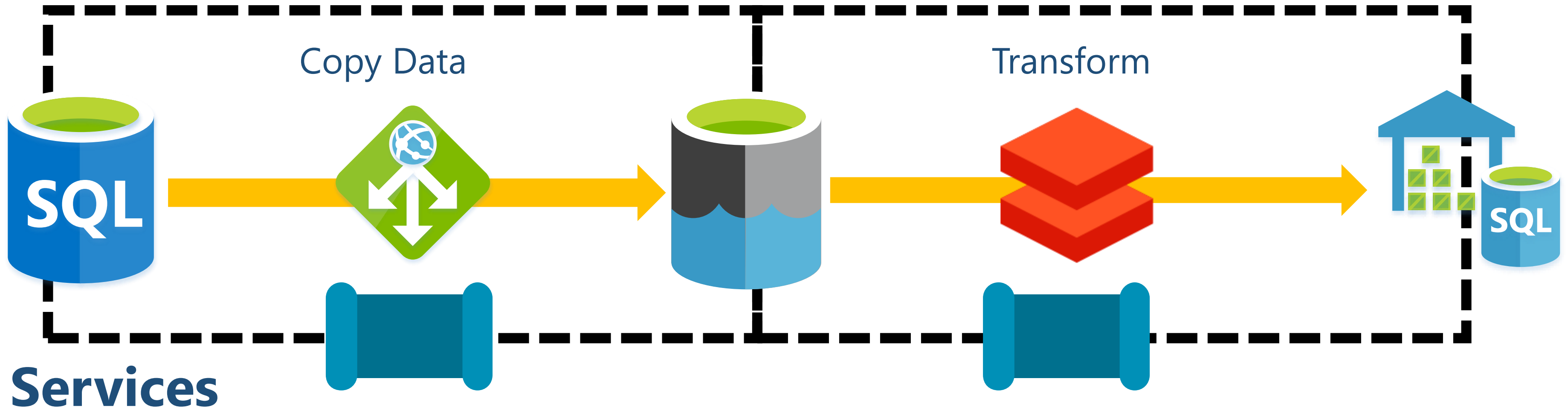


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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

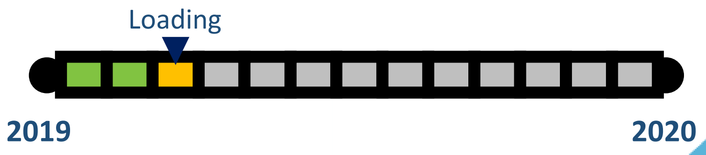
```
Invoke-AzureRmDataFactoryV2Pipeline
-DataFactoryName $dataFactoryName
-ResourceGroupName $resourceGroupName
-PipelineName $pipelineName
```

Data Factory Components

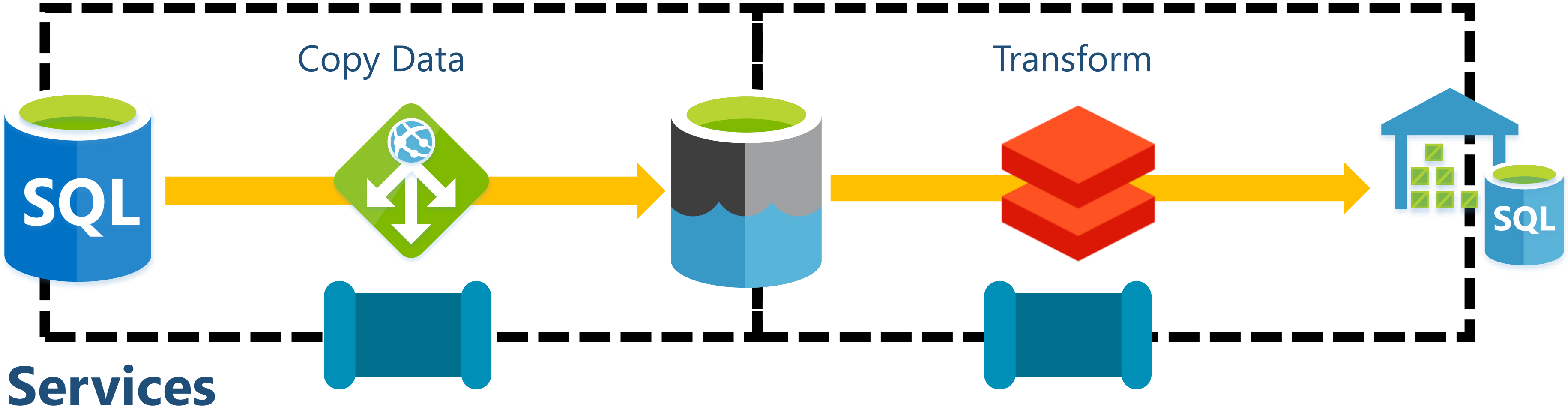


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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



Data Factory Components

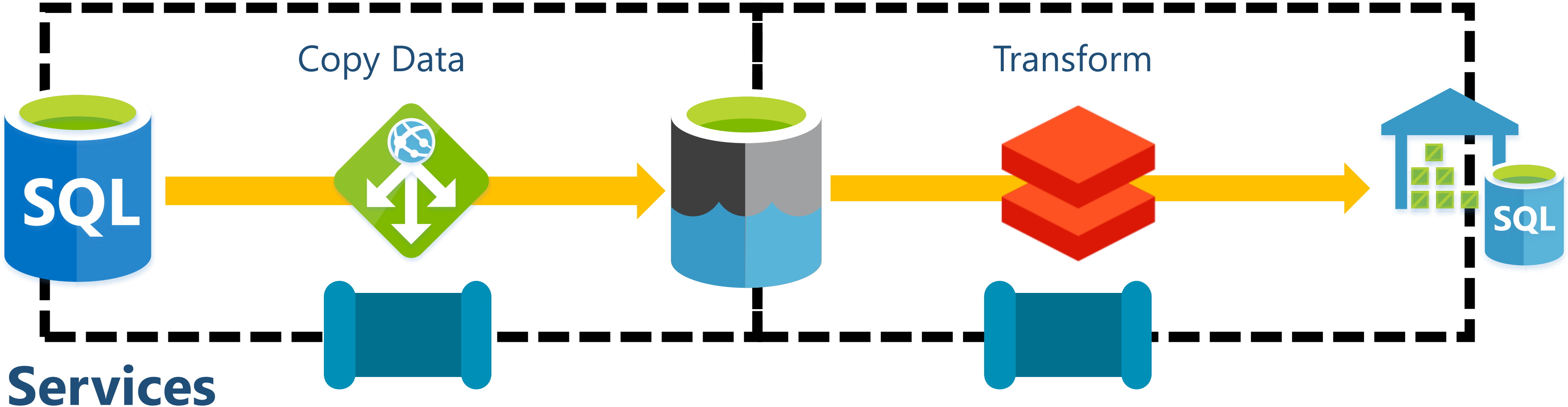


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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

- Every 1 minute.
- UTC

Data Factory Components

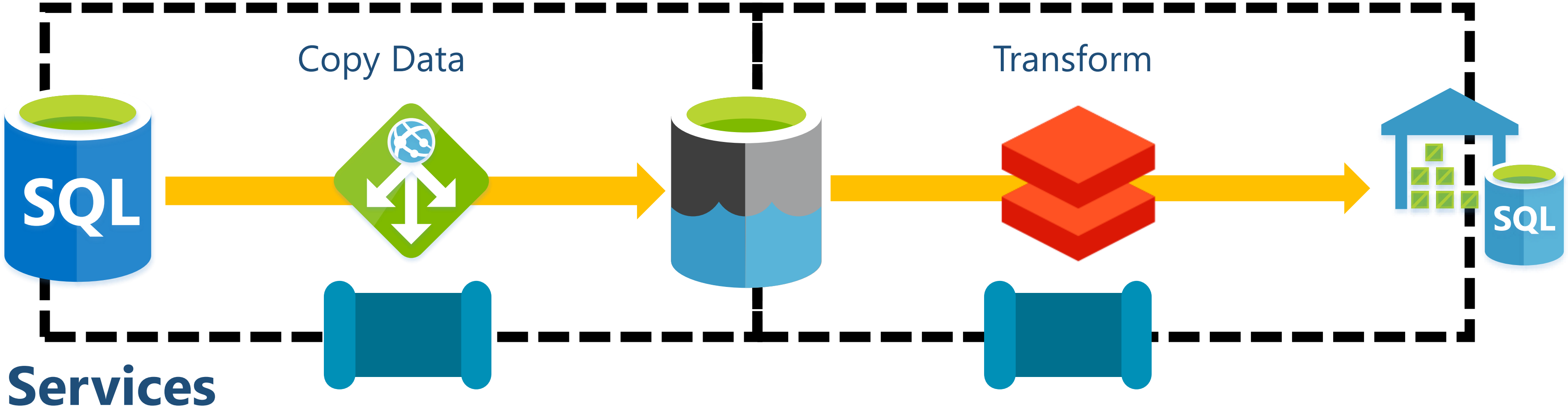


1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

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

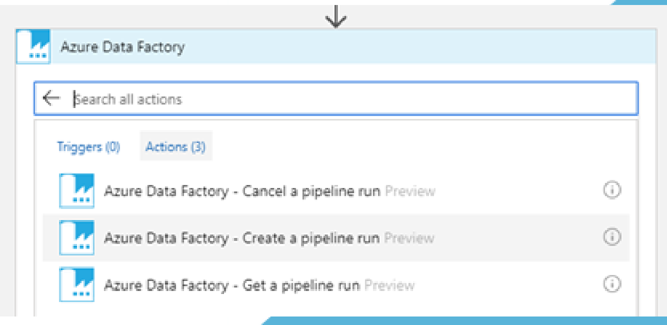
{Path} Created
{Path} Deleted

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**



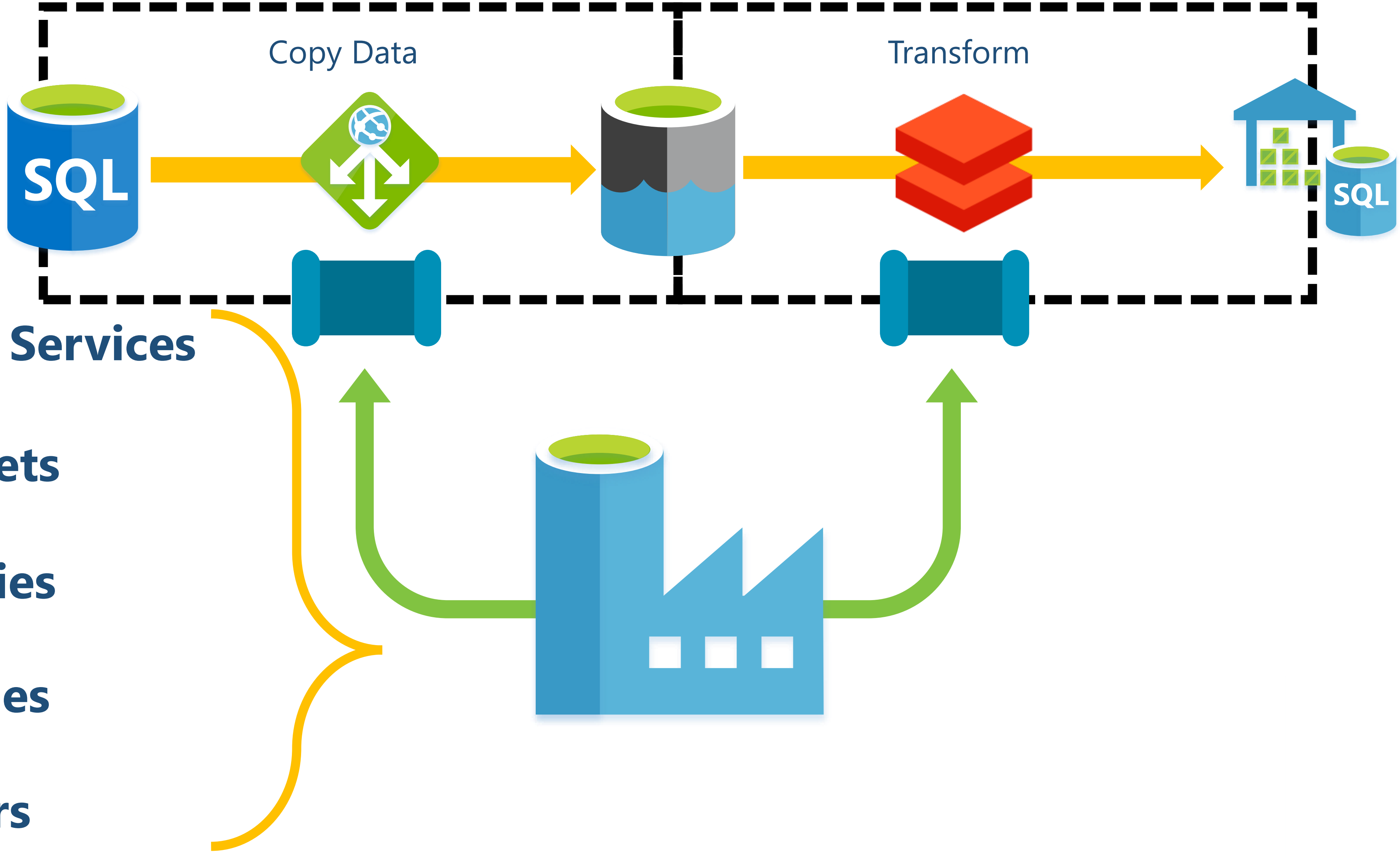
Azure Data Factory

Search all actions

Triggers (0) Actions (3)

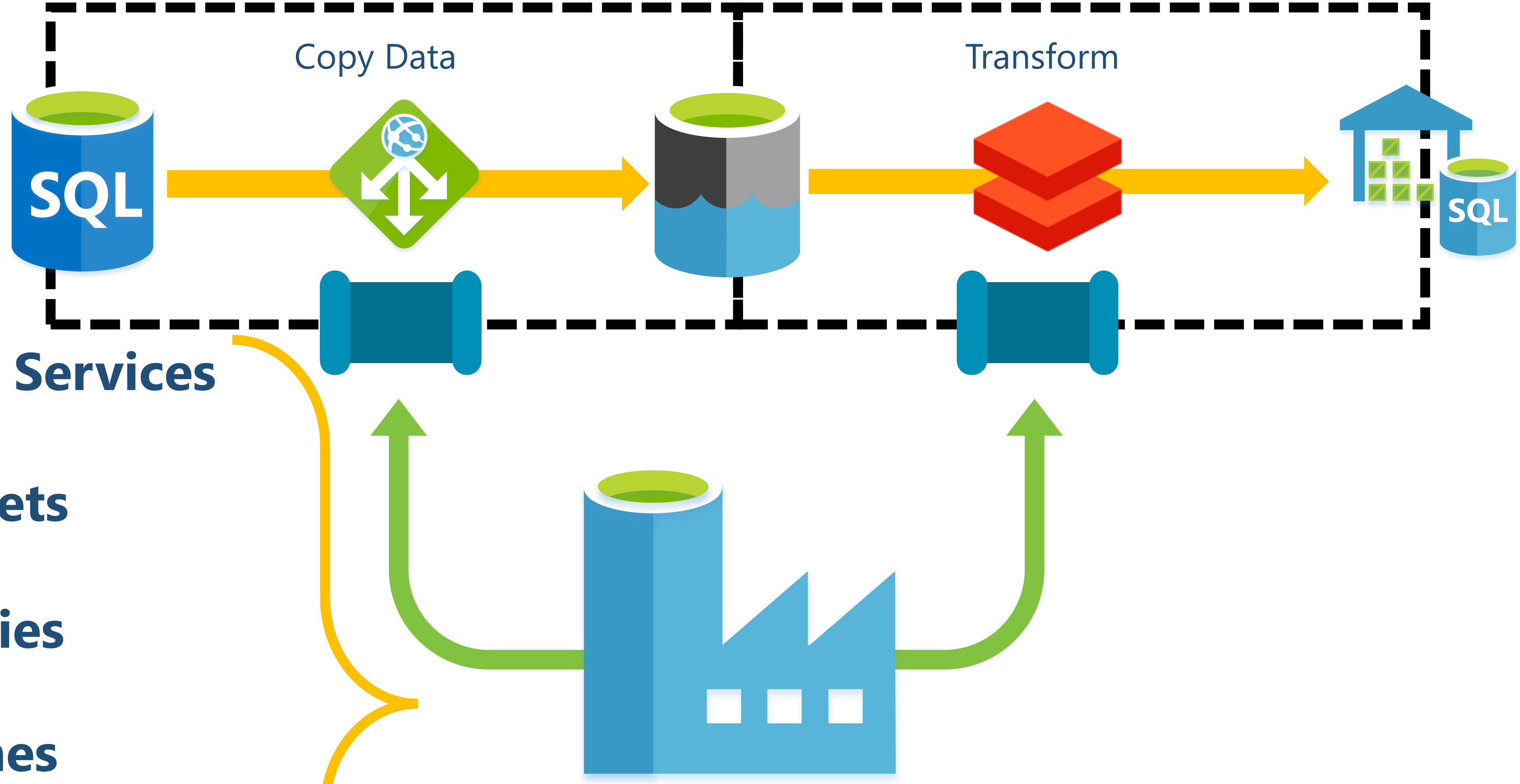
- Azure Data Factory - Cancel a pipeline run Preview
- Azure Data Factory - Create a pipeline run Preview
- Azure Data Factory - Get a pipeline run Preview

Data Factory Components



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

Data Factory Control Flow Components



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

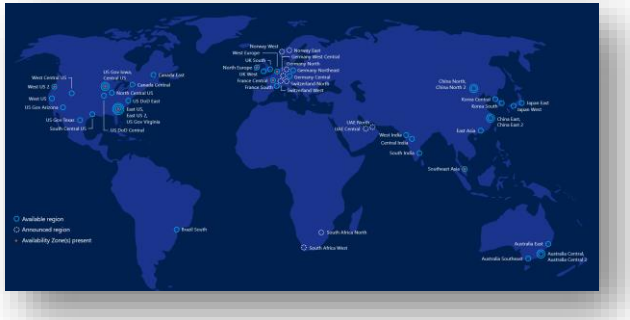
Integration Runtimes

1 **Azure** Integration Runtime

Movement Hours

Activity Orchestration

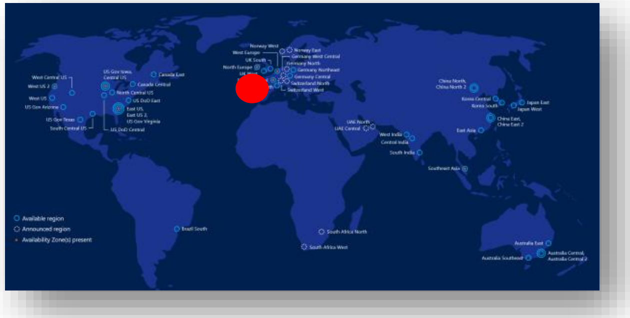
Flexible Region



2 **SSIS** Integration Runtime

SSIS Package Execution

Specified Region



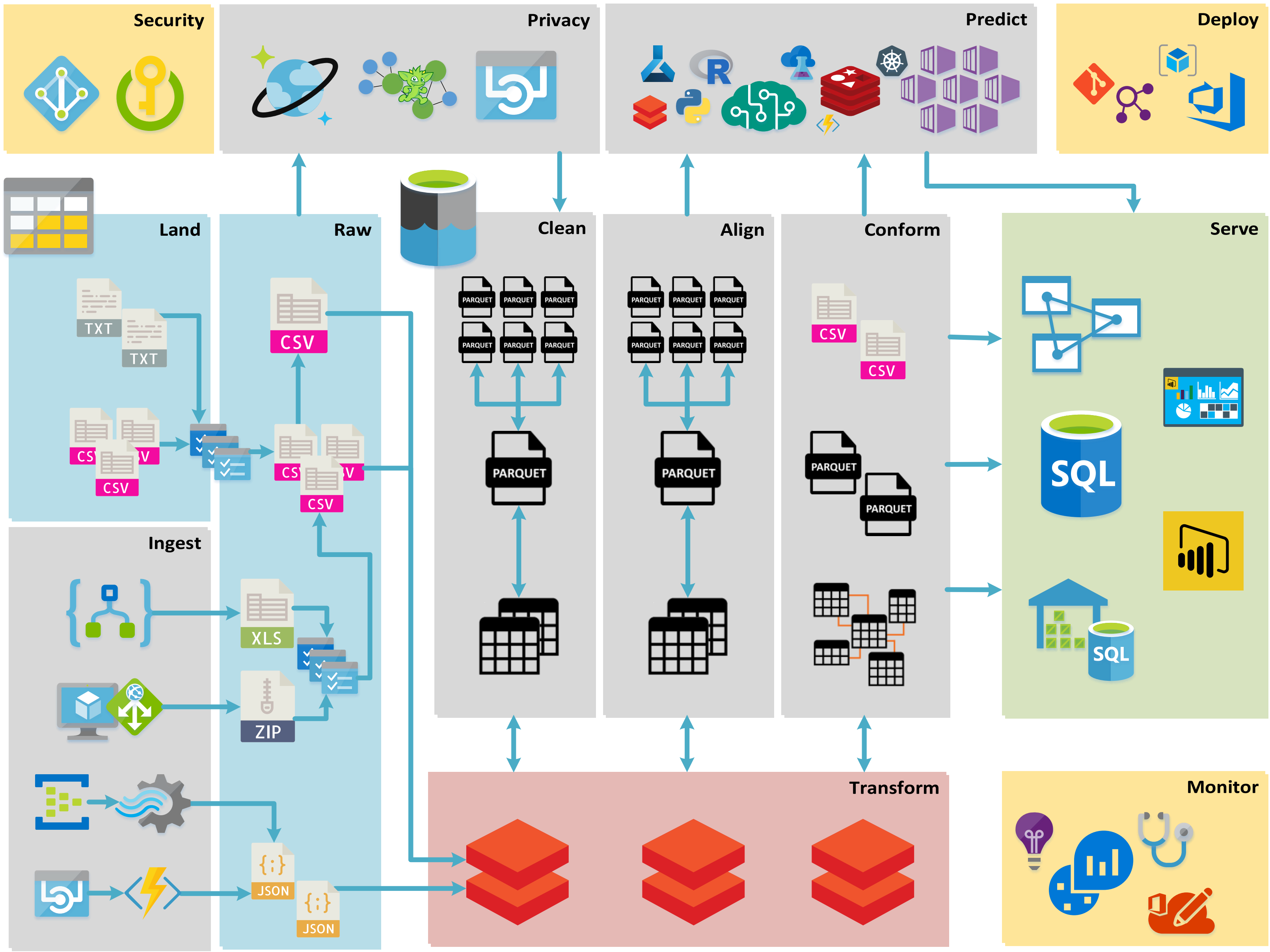
3 **Self Hosted** Integration Runtime

Gateway Access

Activity Orchestration

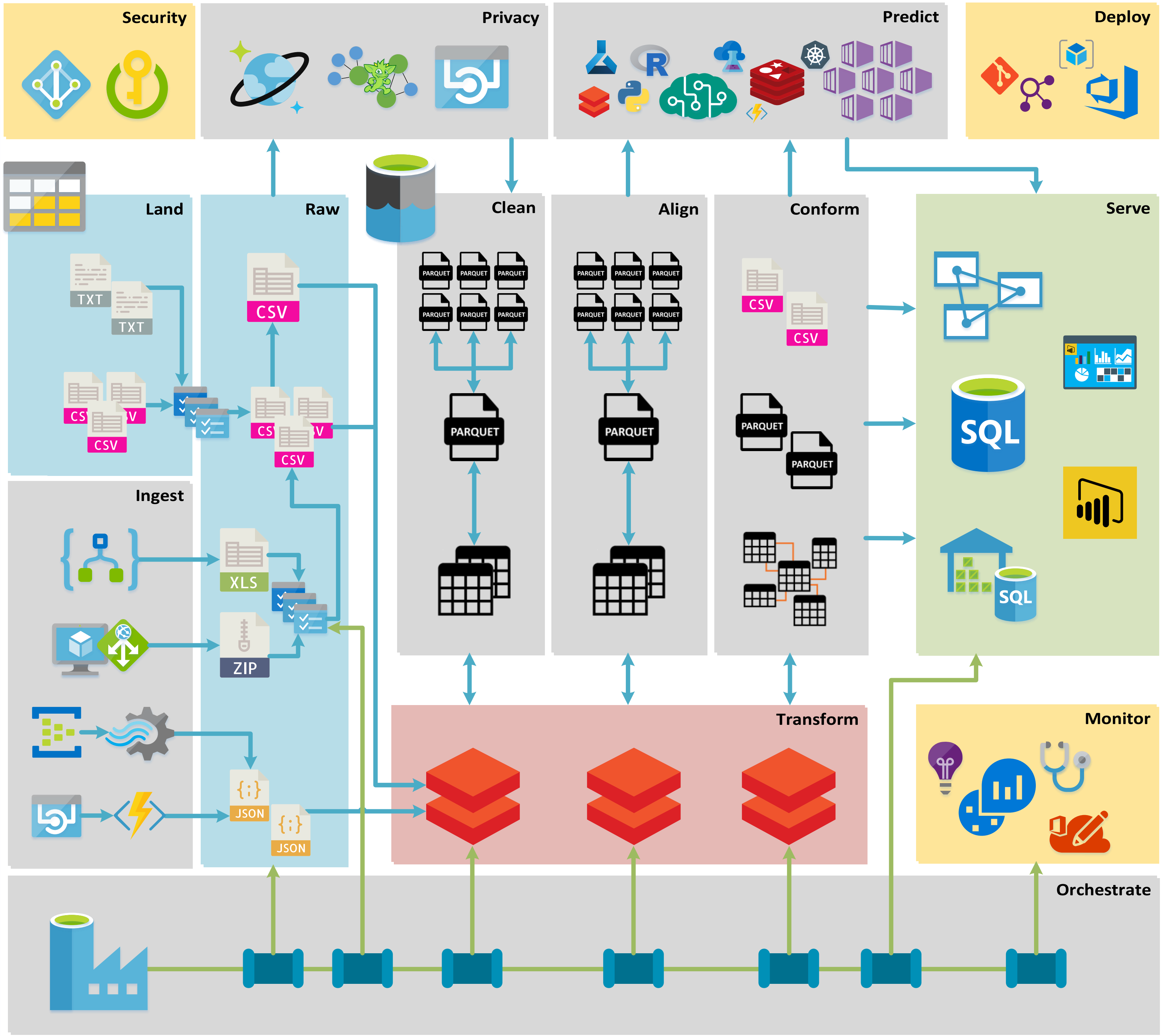
Virtual Machine

Why use Azure Data Factory?



Azure PaaS Data Warehouse & Analytics Platform

Why use Azure Data Factory?



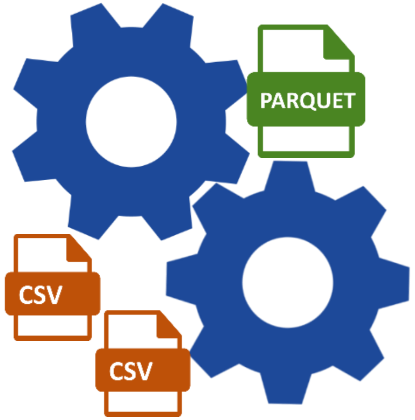
Azure PaaS Data Warehouse & Analytics Platform

Data Factory What & Why - Recap

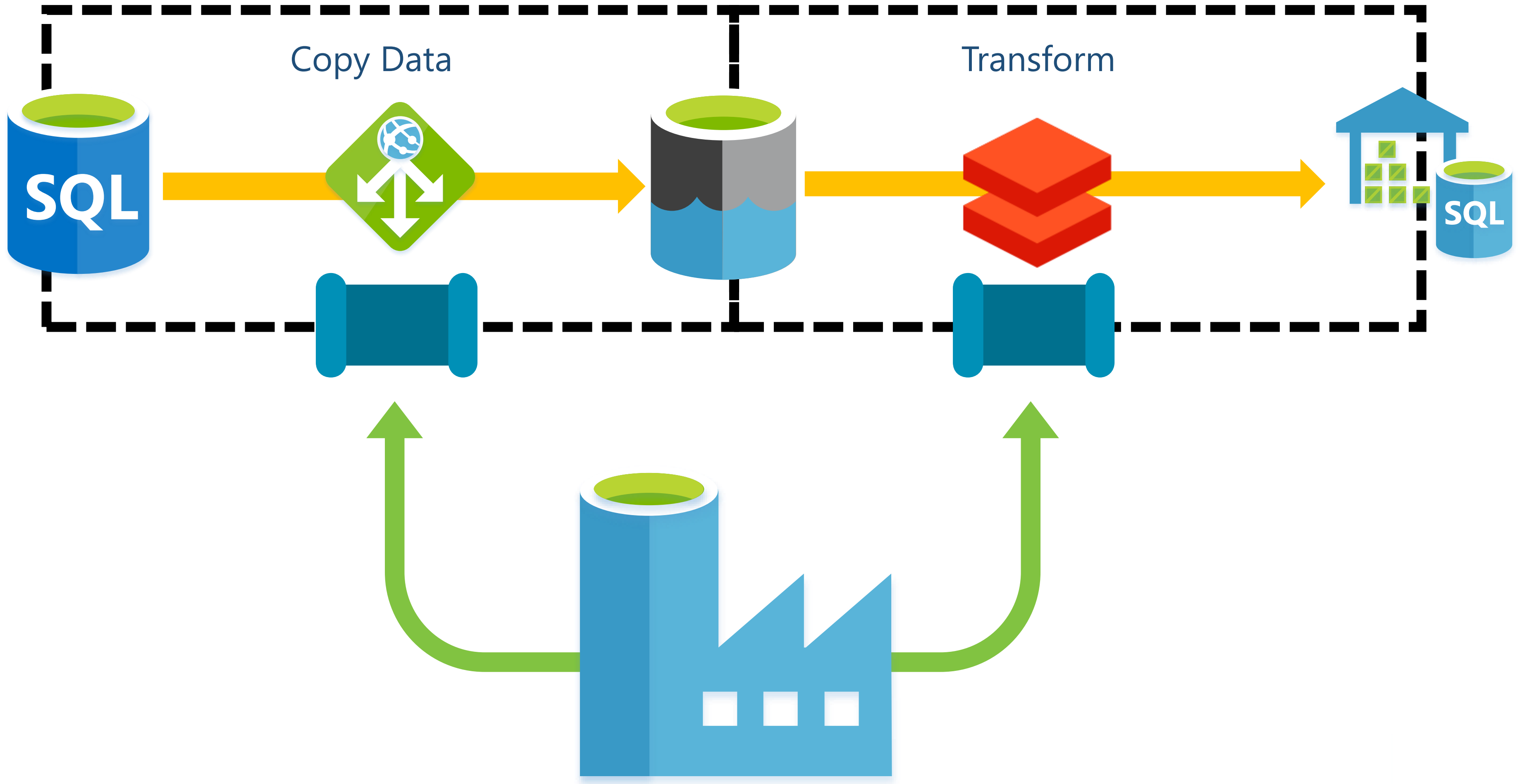
1. **Linked Services**
2. **Data Sets**
3. **Activities**
4. **Pipelines**
5. **Triggers**

1. **Azure**
   Integration Runtime
2. **SSIS**
   Integration Runtime
3. **Self Hosted**
   Integration Runtime

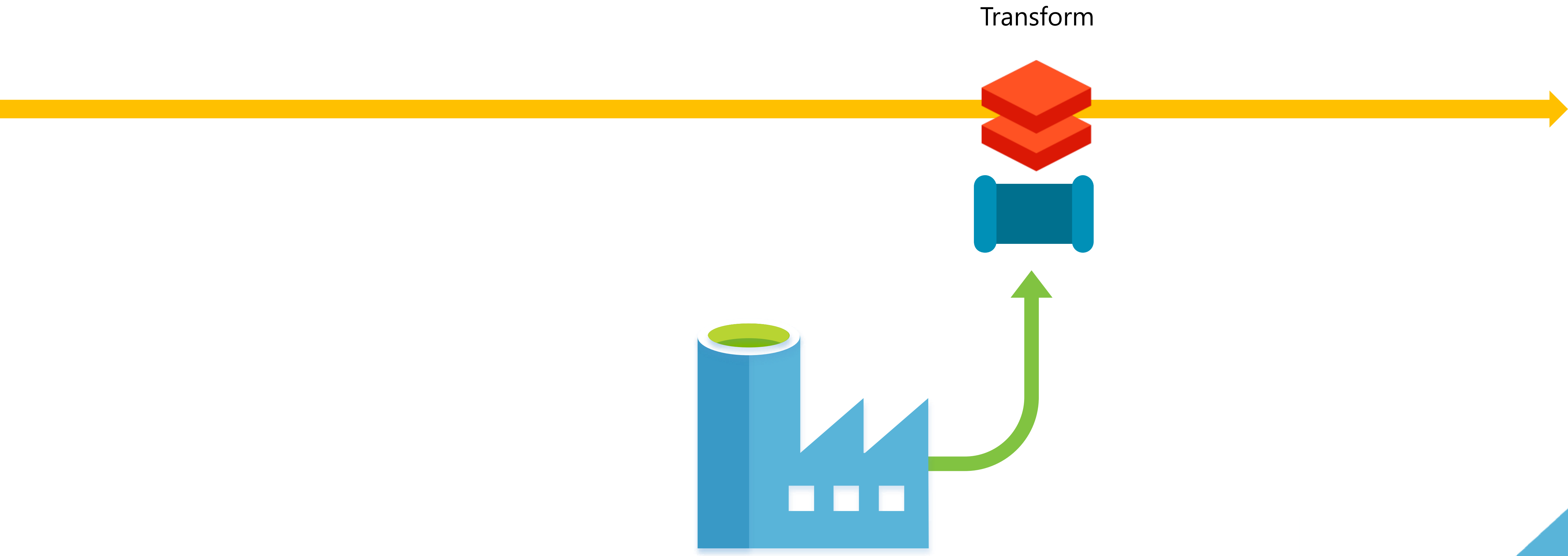
Data Transformation in Azure



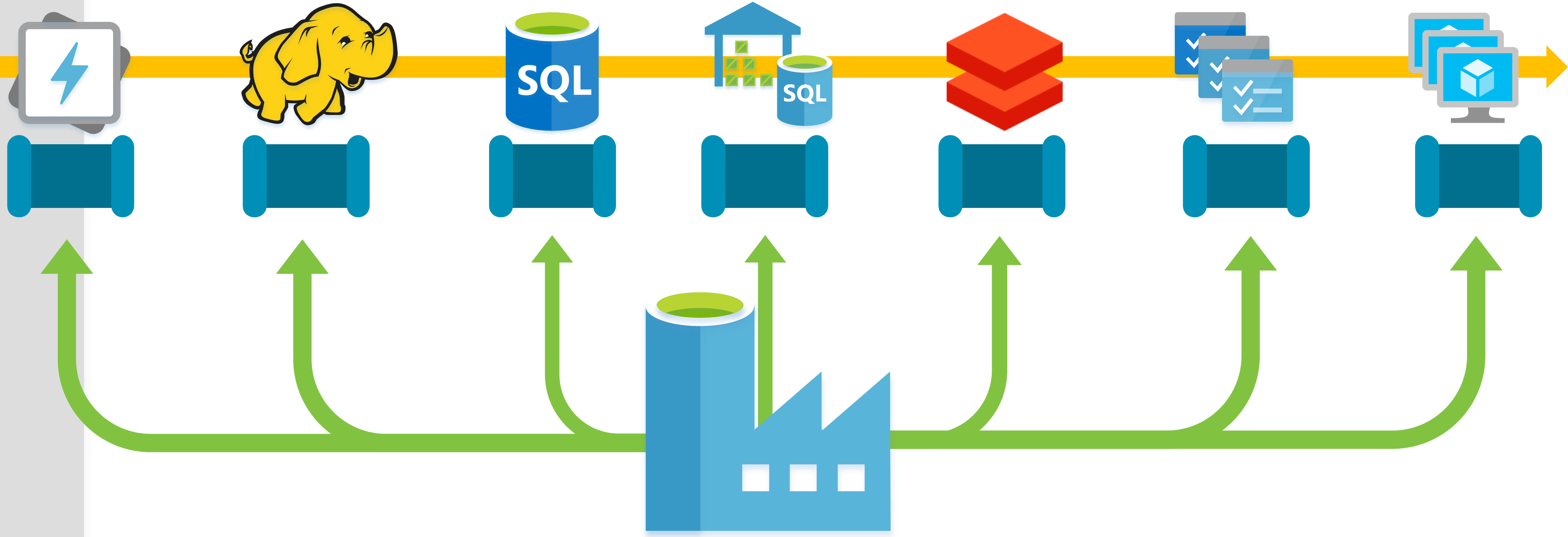
Data Factory Control Flow Components



Data Transformation in Azure



Data Transformation in Azure



Data Lake Analytics

HDInsight

SQL Database

SQL Data Warehouse

Databricks

Custom Activity Batch Service

SSIS Integration Runtime

Future Uncertain

Expensive Clusters

Only Scales Up

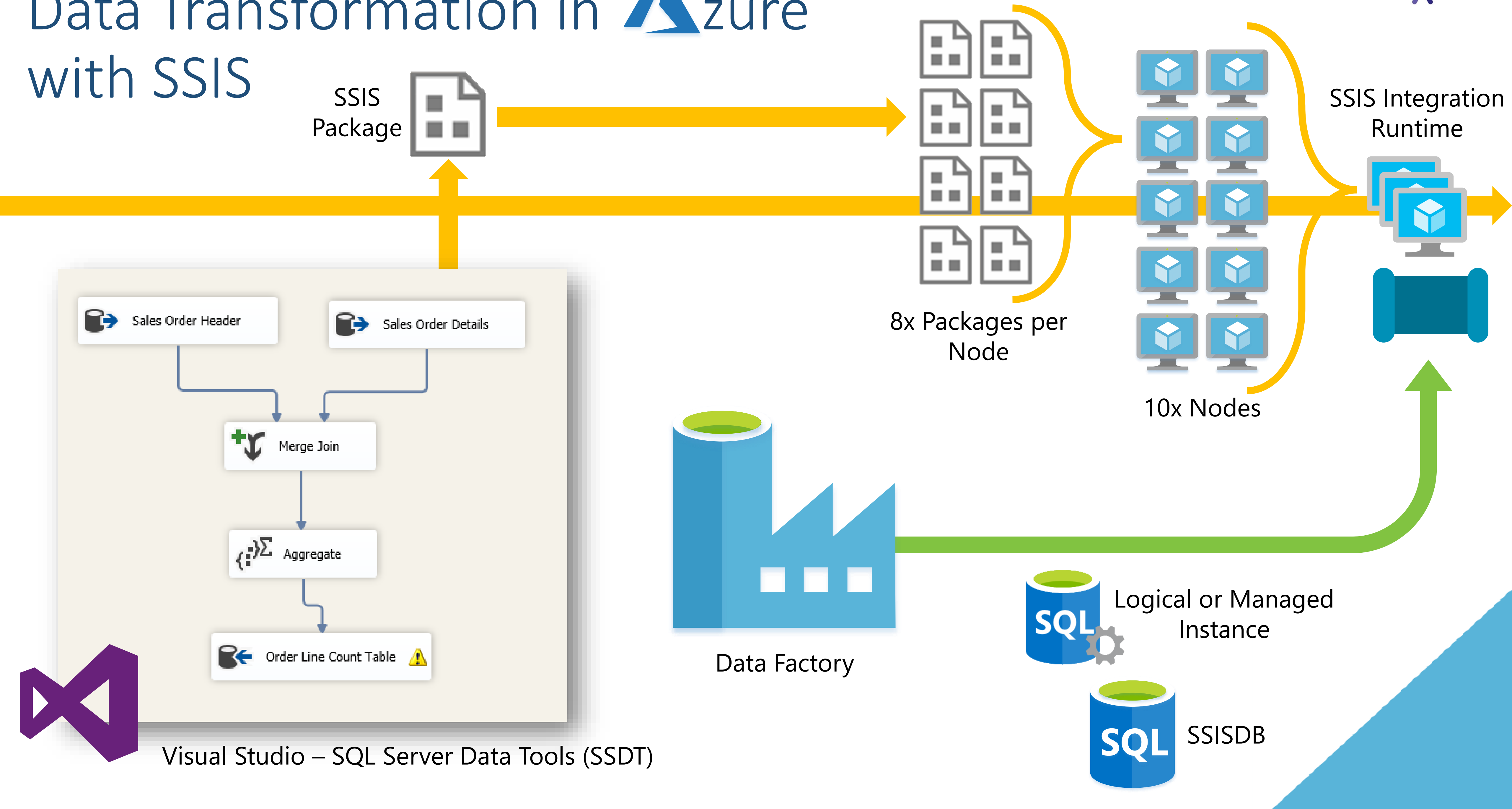
Requires at least 1TB

Learn Spark, Python/Scala

Custom Apps on IaaS

IaaS VMs 20min start

Data Transformation in Azure with SSIS



SSIS Package

8x Packages per Node

10x Nodes

SSIS Integration Runtime

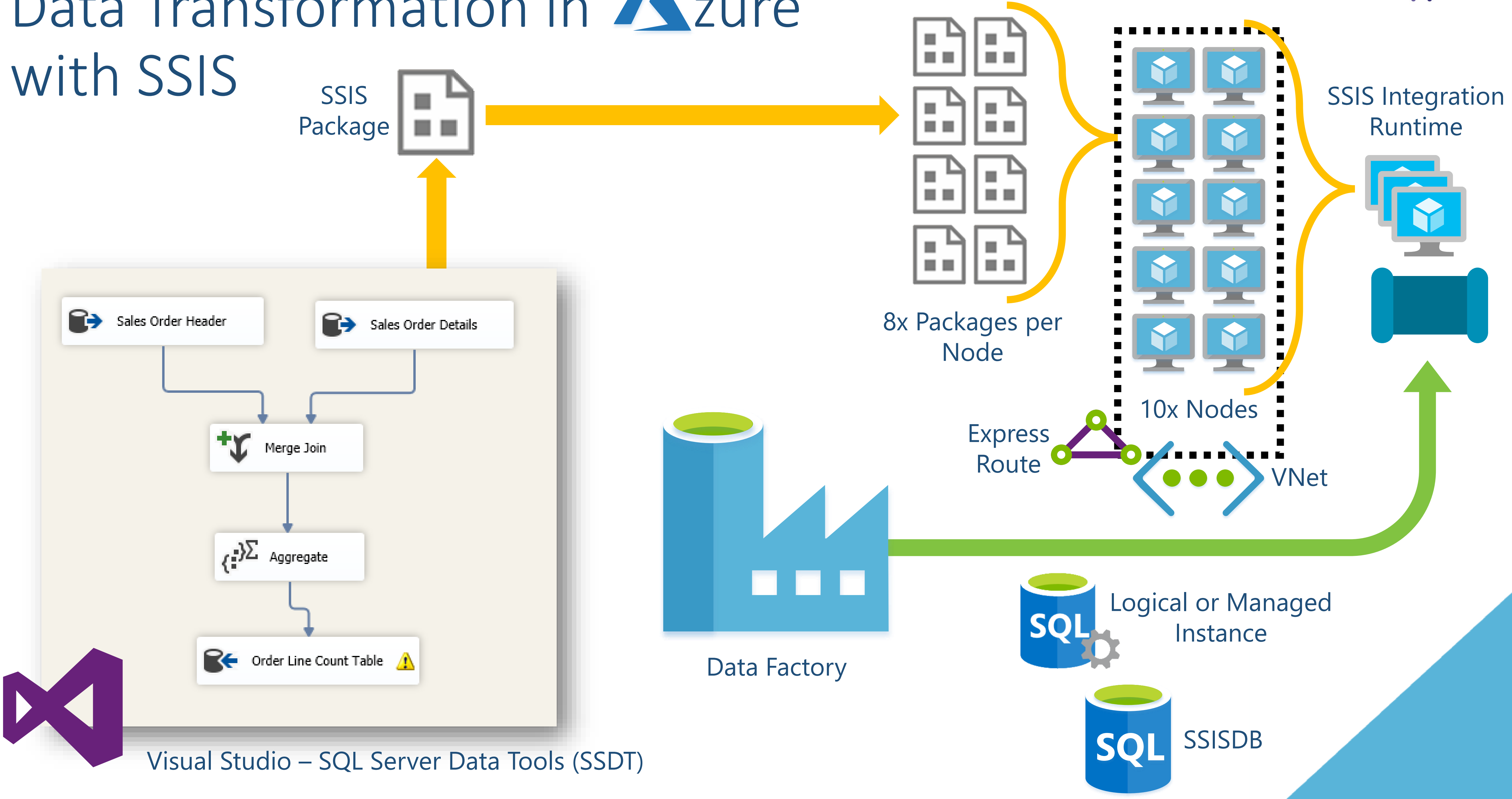
Data Factory

Logical or Managed Instance

SSISDB

Visual Studio – SQL Server Data Tools (SSDT)

Data Transformation in Azure with SSIS



SSIS Package

8x Packages per Node

10x Nodes

SSIS Integration Runtime

Express Route

VNet

Data Factory

Logical or Managed Instance

SSISDB

Sales Order Header

Sales Order Details

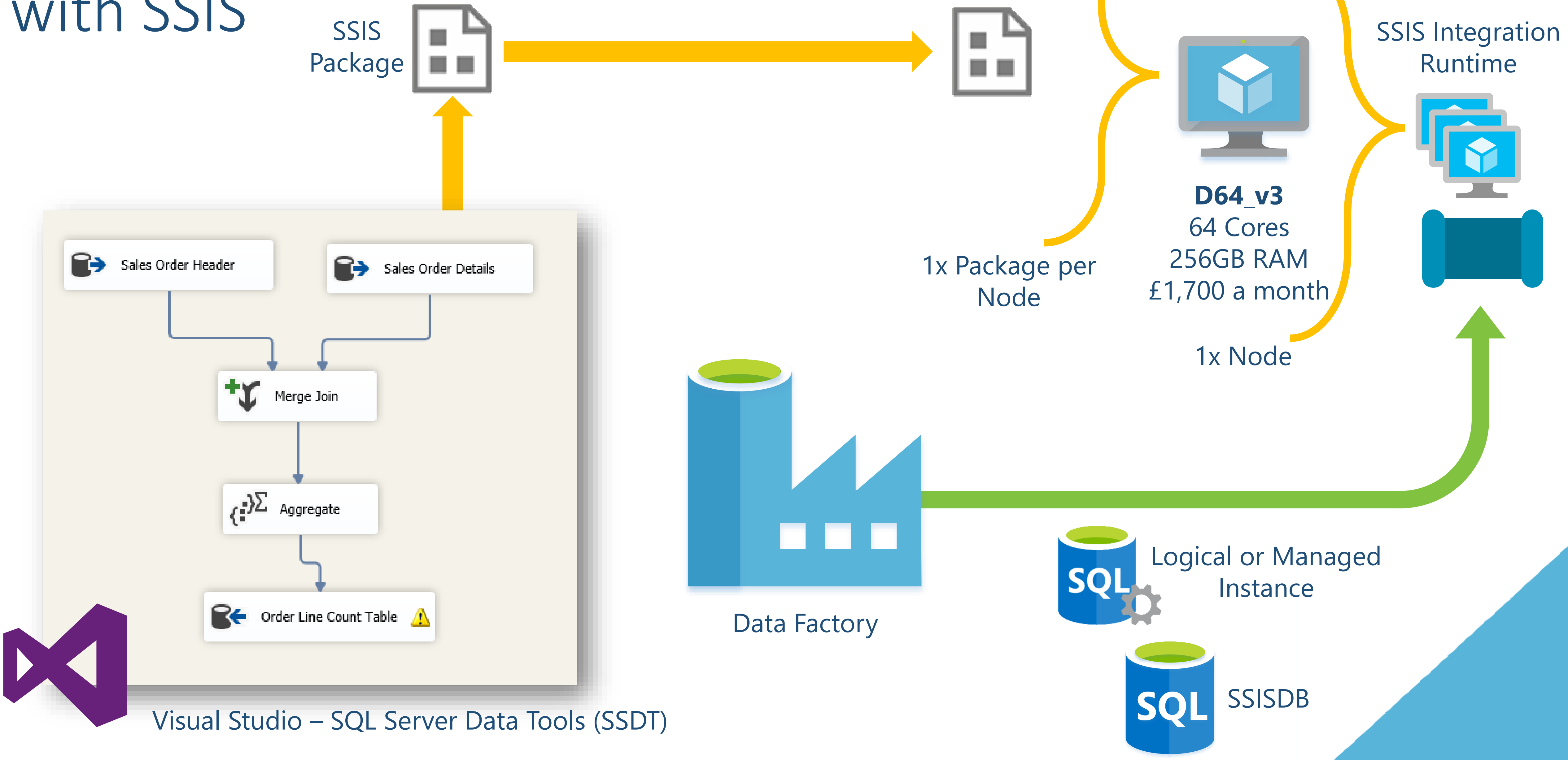
Merge Join

Aggregate

Order Line Count Table

Visual Studio – SQL Server Data Tools (SSDT)

Data Transformation in Azure with SSIS



SSIS Package

1x Package per Node

D64_v3
64 Cores
256GB RAM
£1,700 a month

1x Node

SSIS Integration Runtime

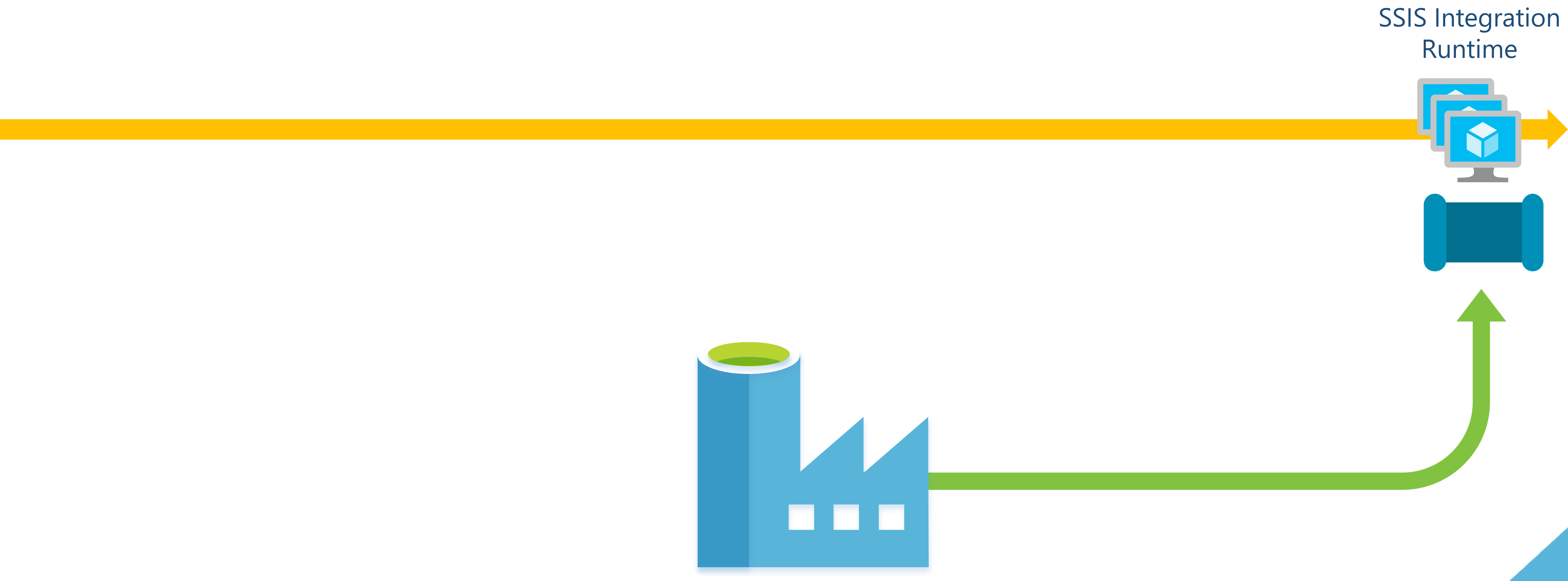
Data Factory

Logical or Managed Instance

SSISDB

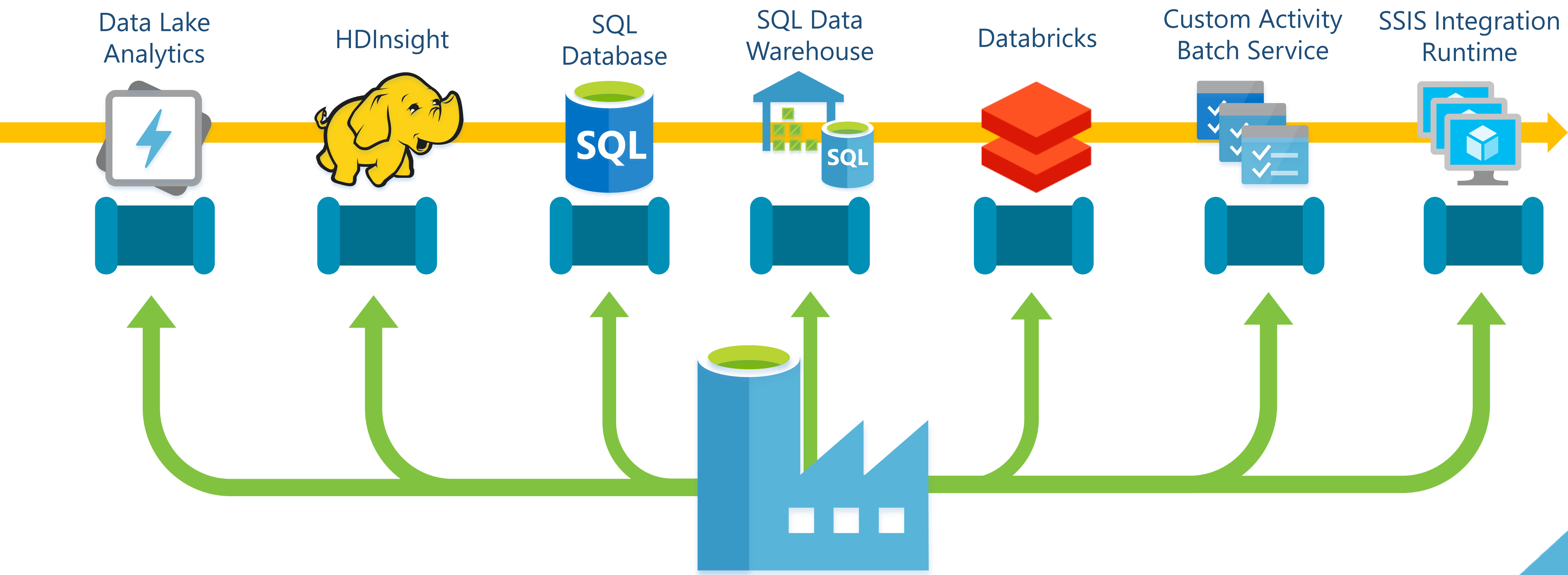
Visual Studio – SQL Server Data Tools (SSDT)

Data Transformation in Azure

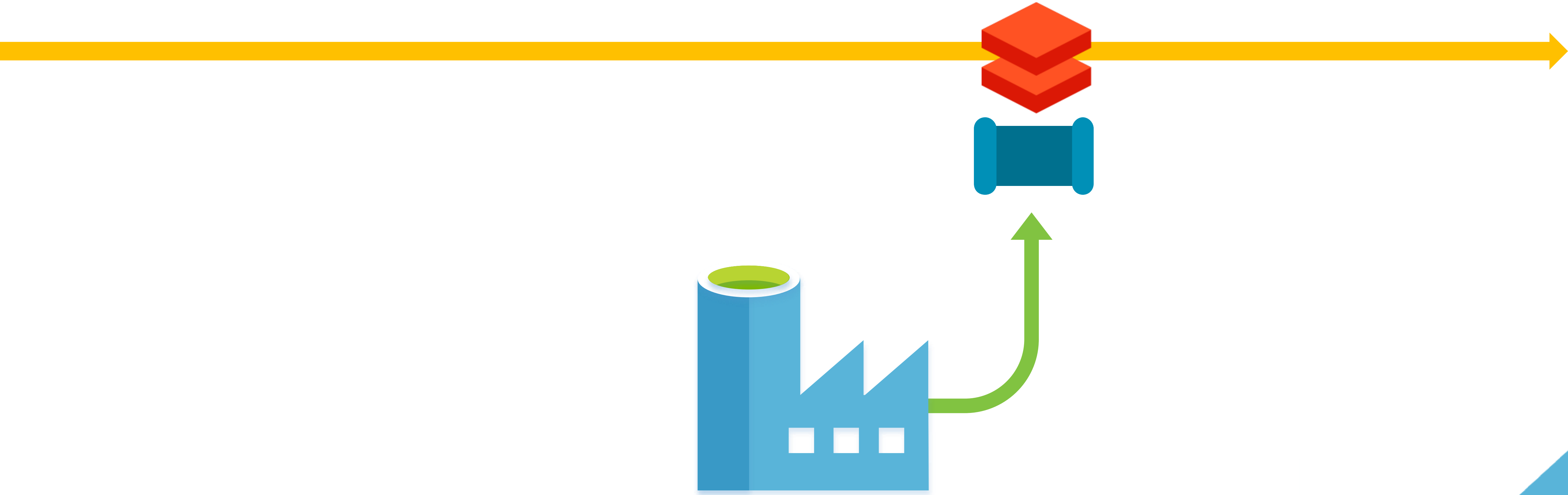


SSIS Integration Runtime

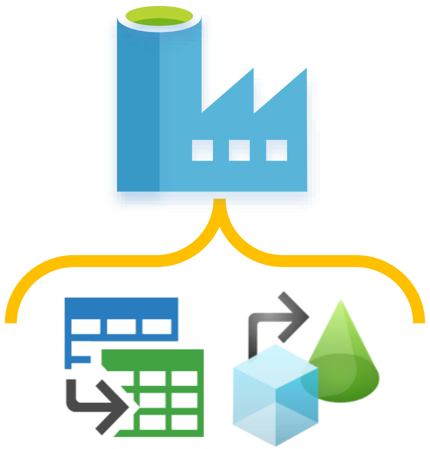
Data Transformation in Azure



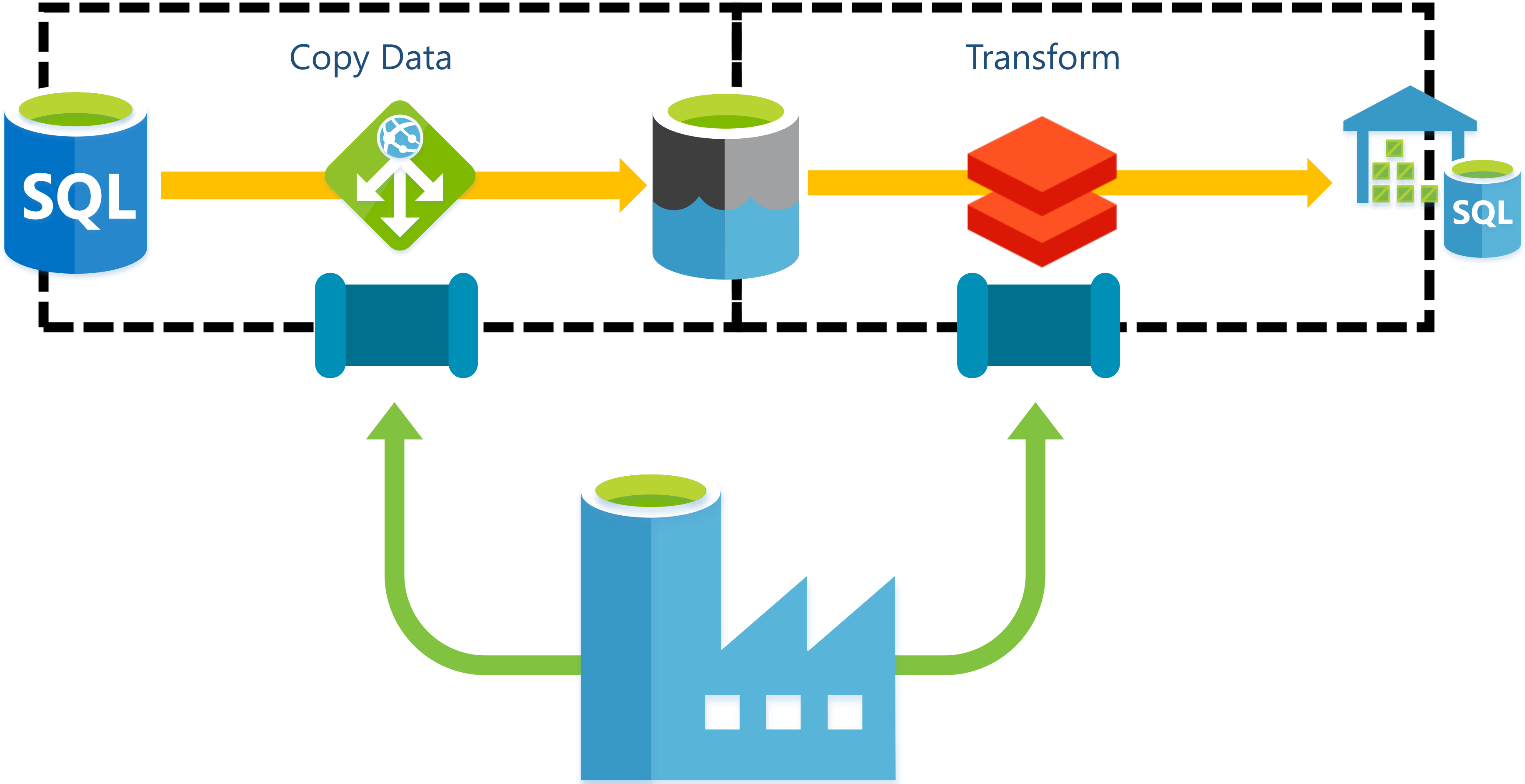
Data Transformation in Azure



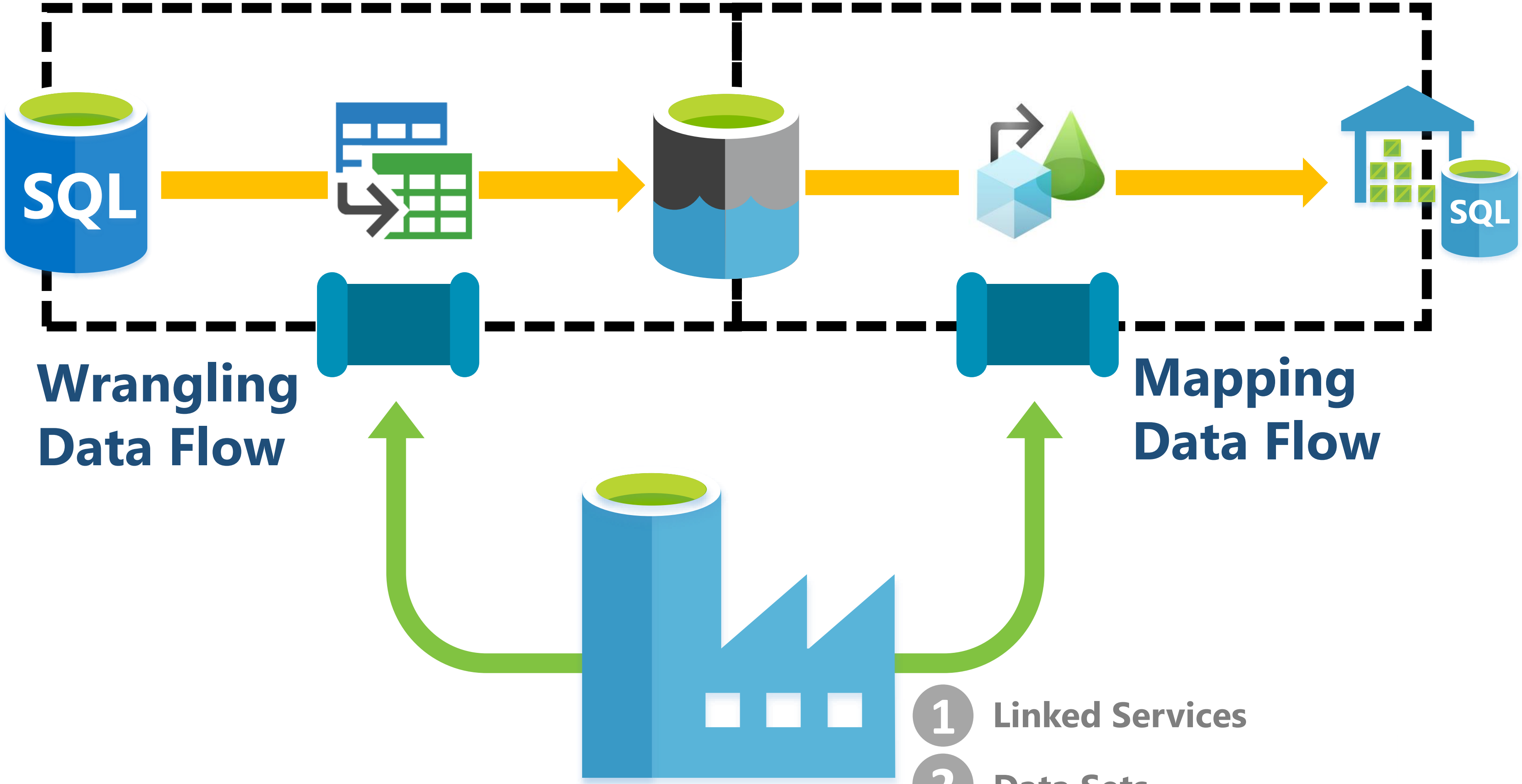
Data Factory Data Flows



Data Factory Control Flow Components



Data Factory Data Flows



1 Linked Services

2 Data Sets

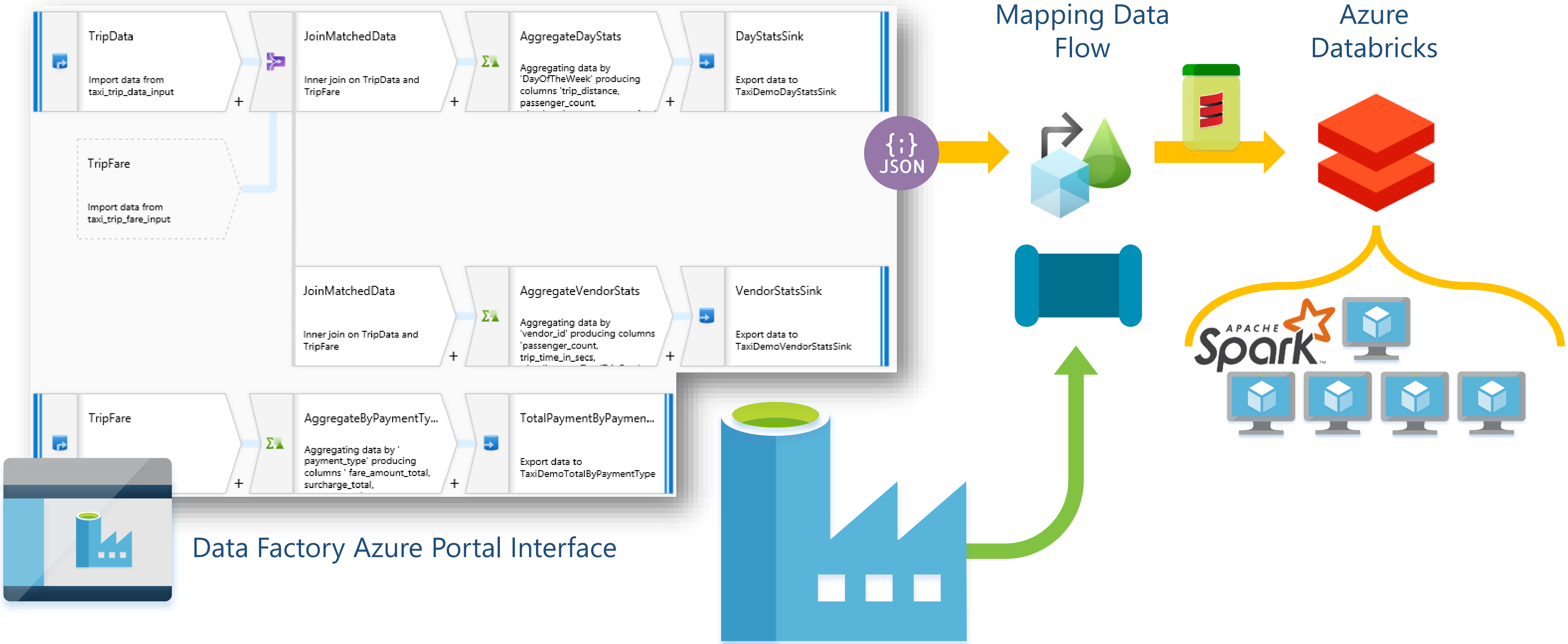
3 **Activities**

4 Pipelines

5 Triggers

Mapping Data Flows

What is a Mapping Data Flow?



TripData

Import data from taxi_trip_data_input

JoinMatchedData

Inner join on TripData and TripFare

AggregateDayStats

Aggregating data by 'DayOfTheWeek' producing columns 'trip_distance, passenger_count,

DayStatsSink

Export data to TaxiDemoDayStatsSink

TripFare

Import data from taxi_trip_fare_input

JoinMatchedData

Inner join on TripData and TripFare

AggregateVendorStats

Aggregating data by 'vendor_id' producing columns 'passenger_count, trip_time_in_secs,

VendorStatsSink

Export data to TaxiDemoVendorStatsSink

TripFare

AggregateByPaymentTy...

Aggregating data by ' payment_type' producing columns ' fare_amount_total, surcharge_total,

TotalPaymentByPaymen...

Export data to TaxiDemoTotalByPaymentType

Mapping Data Flow

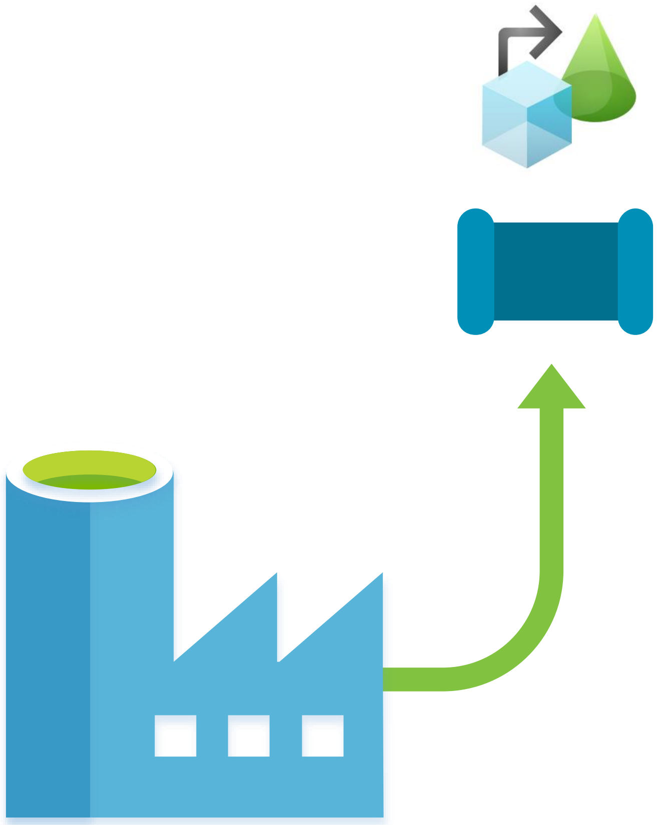
Azure Databricks

Data Factory Azure Portal Interface

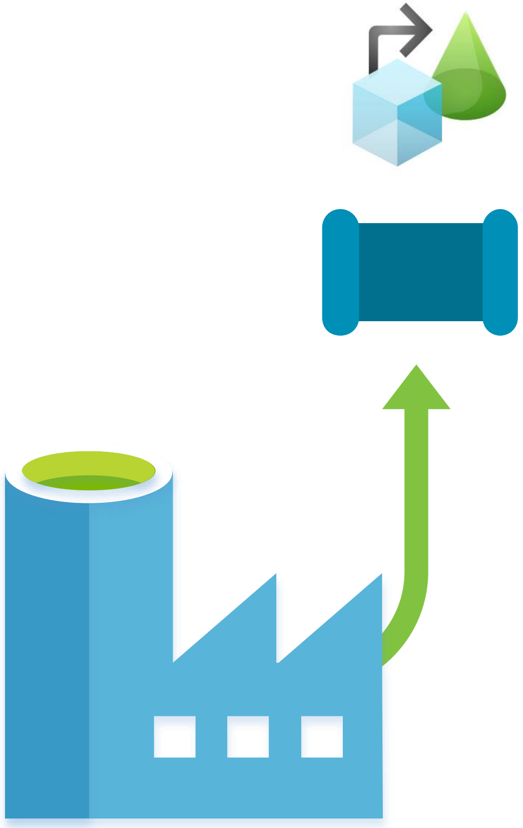
Data Factory

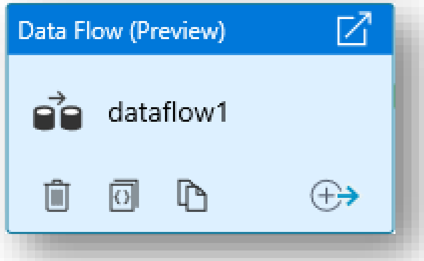
https://mrpaulandrew.com/2018/10/22/what-is-azure-data-factory-data-flow/

Mapping Data Flows



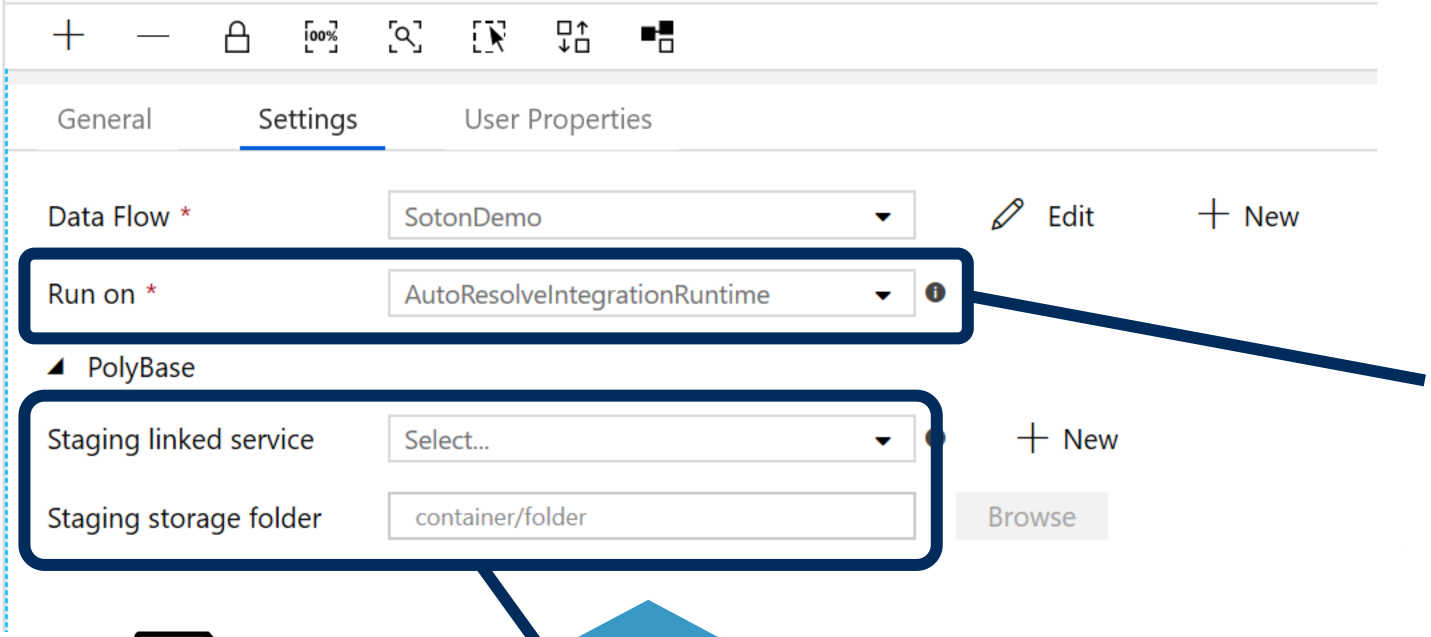
Mapping Data Flows – Settings & Concepts




Data Flow (Preview)

dataflow1

New Control Flow Activity



Data Flow * SotonDemo

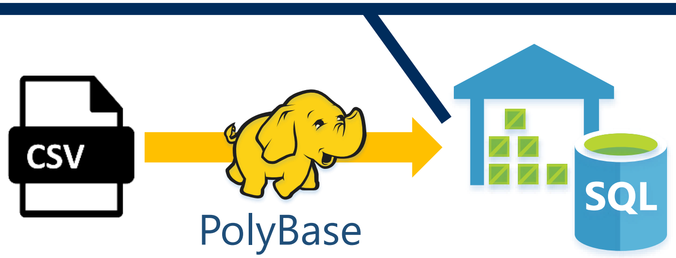
Run on * AutoResolveIntegrationRuntime

◢ PolyBase

Staging linked service Select...

Staging storage folder container/folder

Azure Databricks



PolyBase

Azure SQLDW

Integration Runtimes

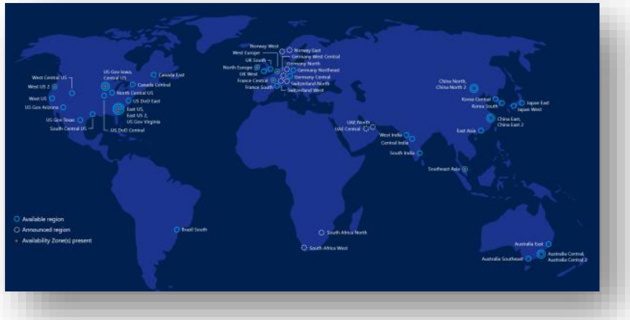
1

Azure
Integration Runtime

Movement Hours

Activity Orchestration

Flexible Region

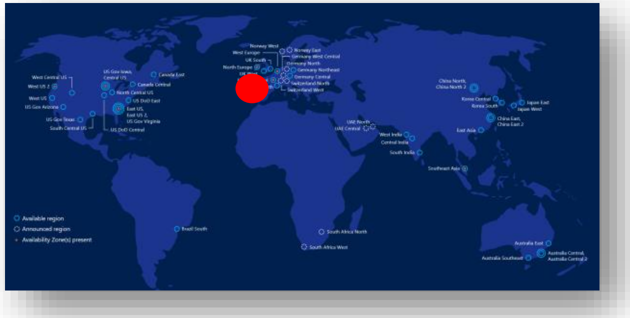


2

SSIS
Integration Runtime

SSIS Package Execution

Specified Region



3

Self Hosted
Integration Runtime

Gateway Access

Activity Orchestration

Virtual Machine

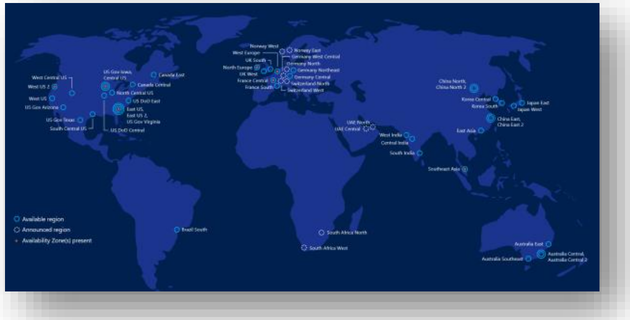
Integration Runtimes – Mapping Data Flow Cluster

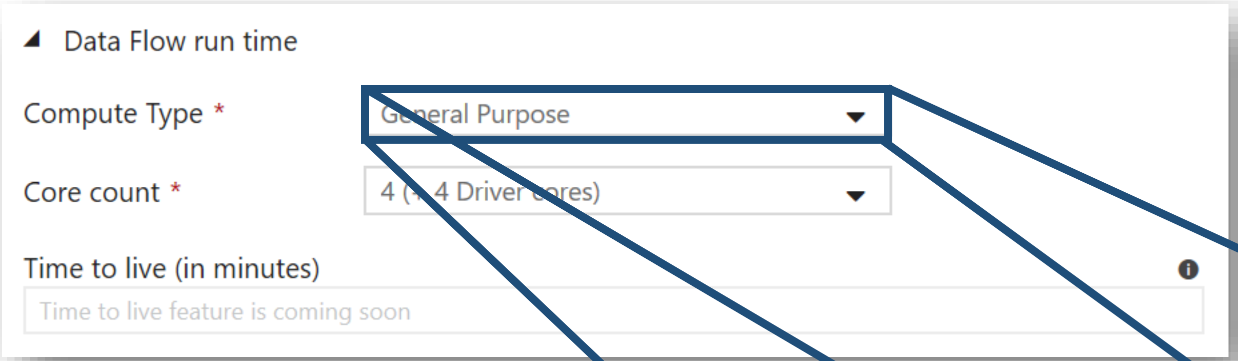
1 **Azure**
Integration Runtime

Movement Hours

Activity Orchestration

Flexible Region





Data Flow run time

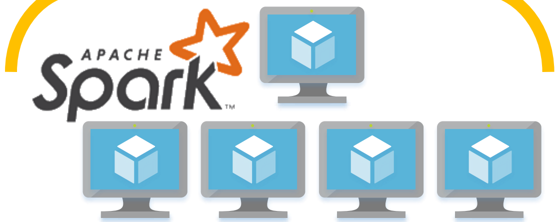
Compute Type * General Purpose

Core count * 4 (+ 4 Driver cores)

Time to live (in minutes)

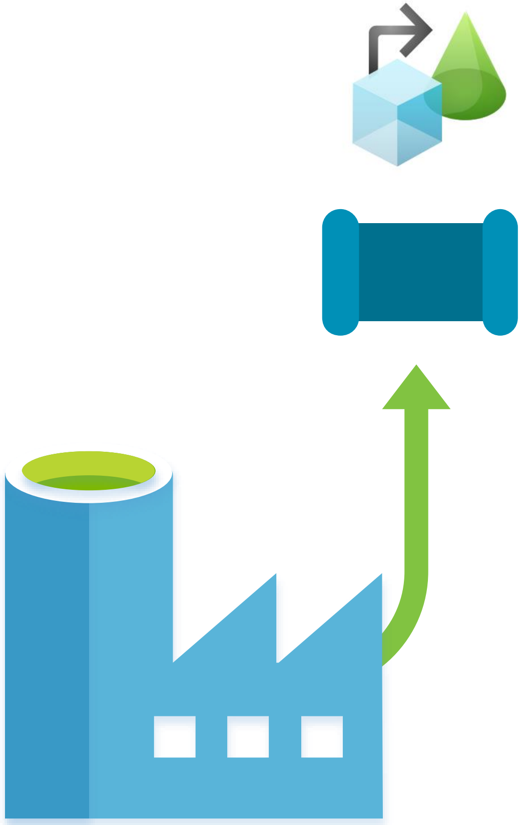
Time to live feature is coming soon

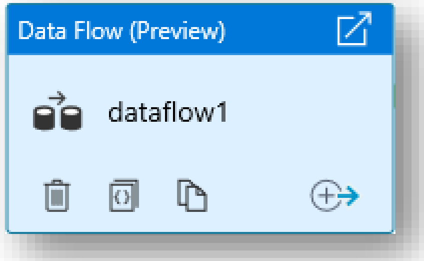
=



- General Purpose
- Memory Optimised
- Compute Optimised

Mapping Data Flows – Settings & Concepts

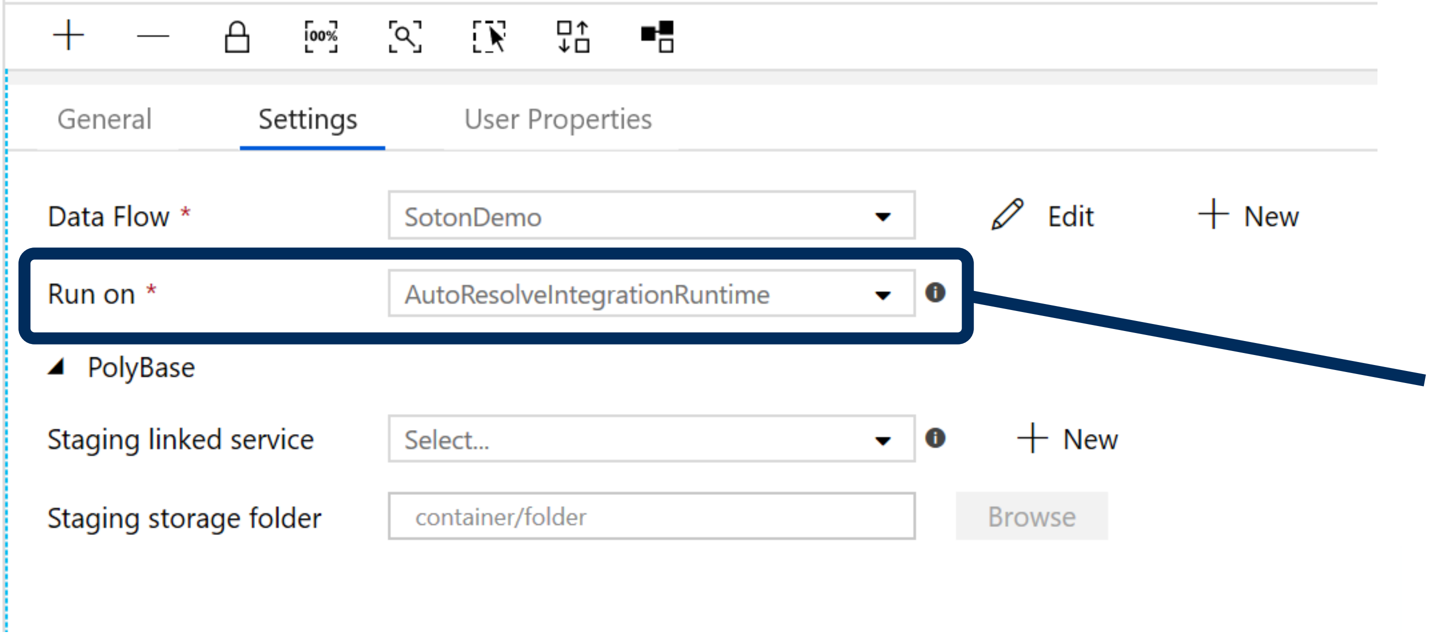




Data Flow (Preview)

dataflow1

New Control Flow Activity



Data Flow * SotonDemo

Run on * AutoResolveIntegrationRuntime

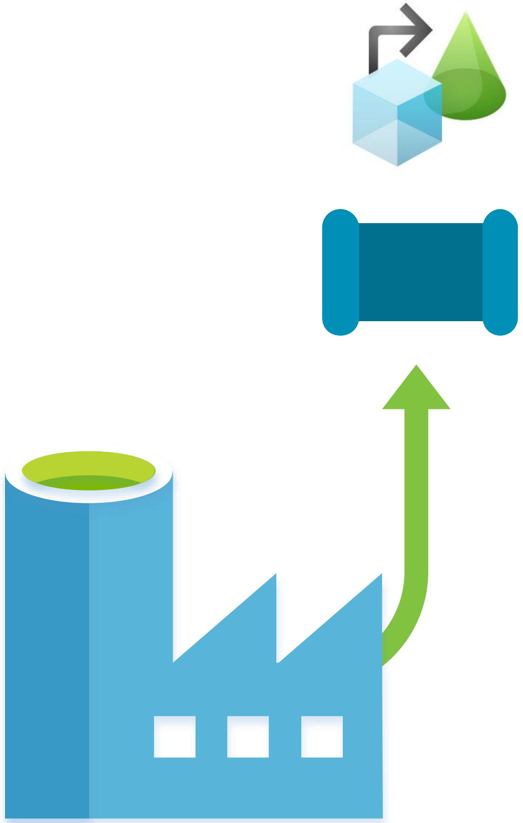
◢ PolyBase

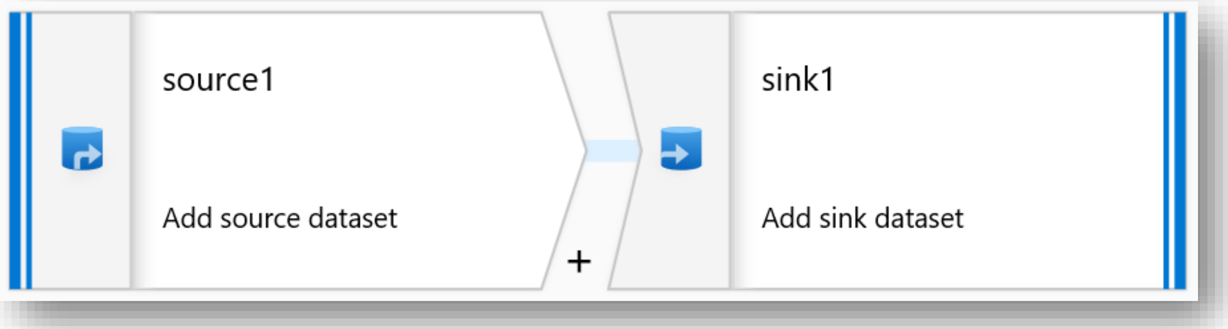
Staging linked service Select...

Staging storage folder container/folder

Azure Databricks

Mapping Data Flows – Settings & Concepts





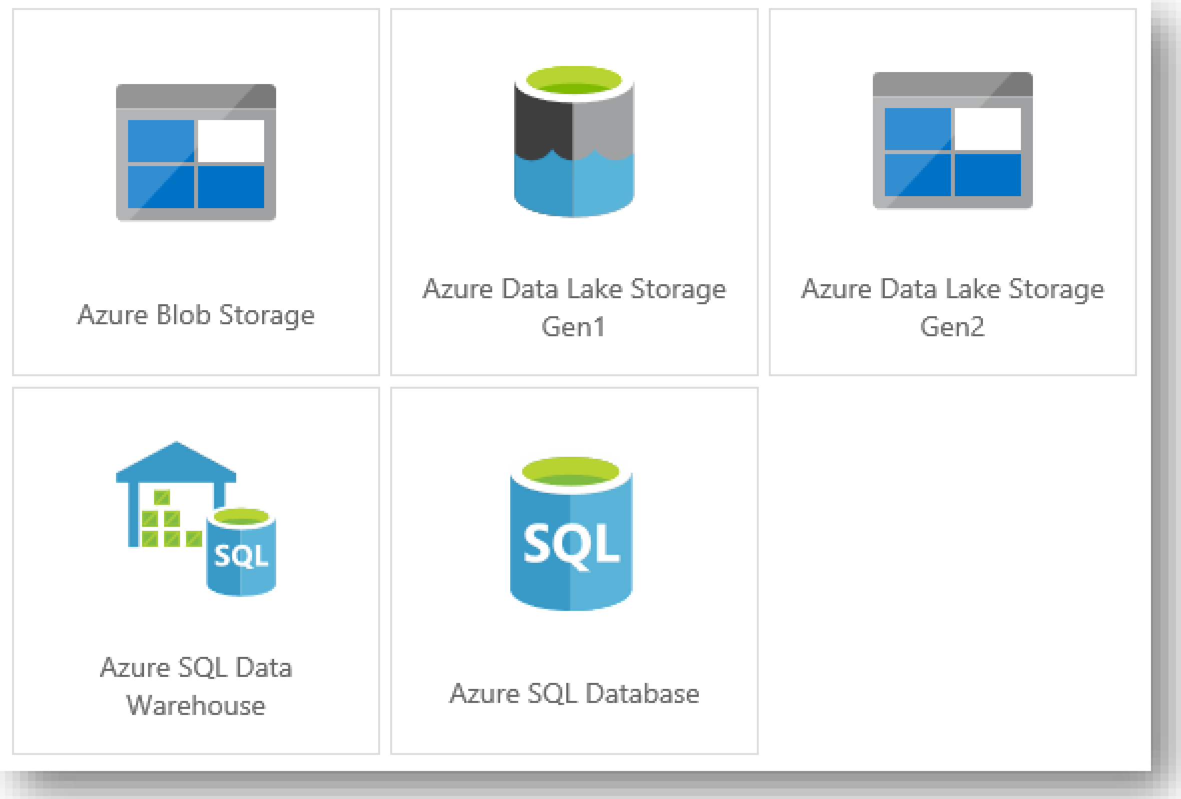
source1

Add source dataset

sink1

Add sink dataset

Currently Available:



Azure Blob Storage

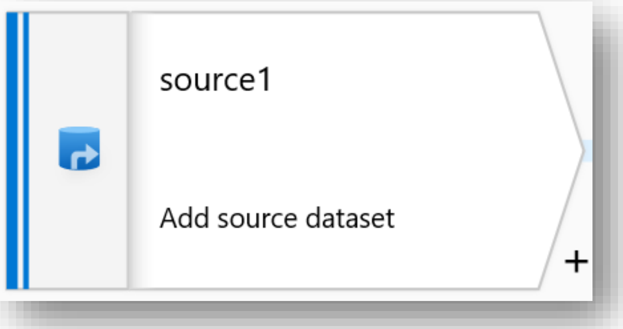
Azure Data Lake Storage Gen1

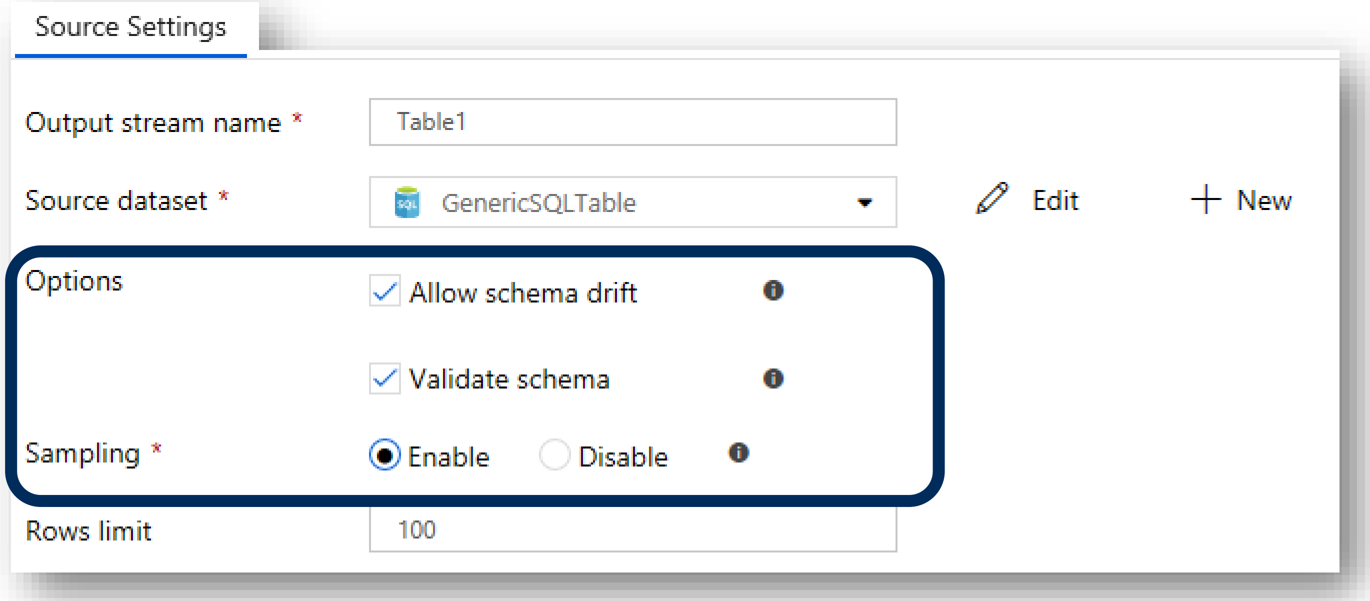
Azure Data Lake Storage Gen2

Azure SQL Data Warehouse

Azure SQL Database

Mapping Data Flows – Settings & Concepts

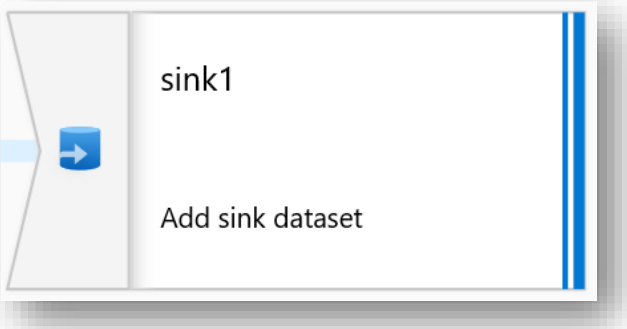




Source Settings

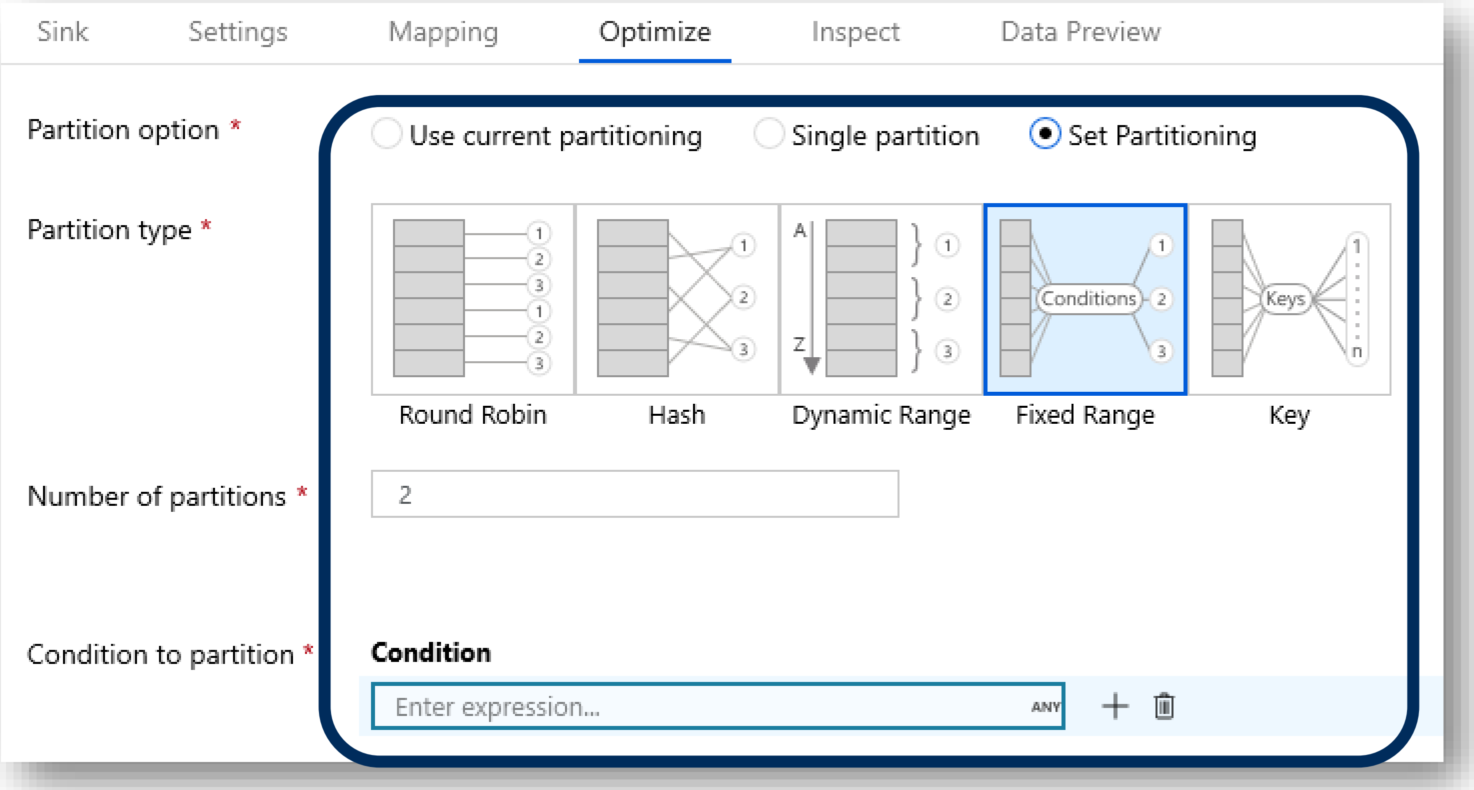
Output stream name *	Table1		
Source dataset *	GenericSQLTable	Edit	New
Options	☑ Allow schema drift		
	☑ Validate schema		
Sampling *	● Enable ○ Disable		
Rows limit	100		

Mapping Data Flows – Settings & Concepts



sink1

Add sink dataset



Sink | Settings | Mapping | Optimize | Inspect | Data Preview

Partition option * ○ Use current partitioning ○ Single partition ● Set Partitioning

Partition type * Round Robin | Hash | Dynamic Range | Fixed Range | Key

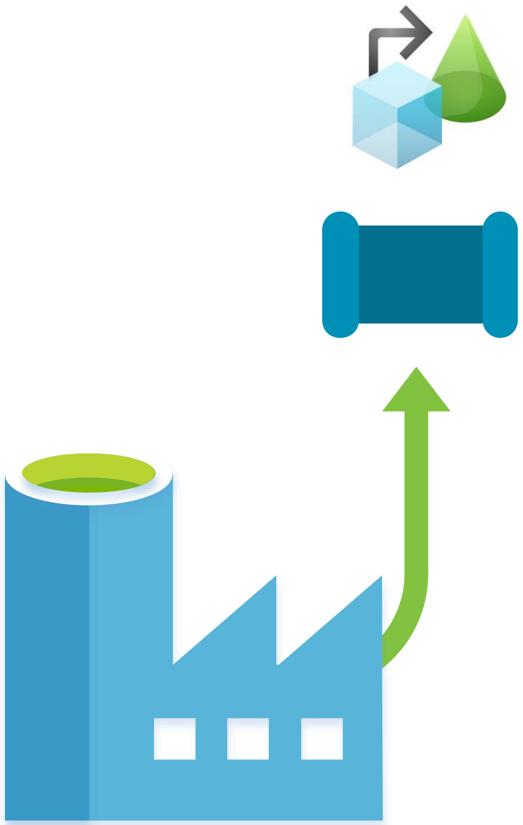
Number of partitions * 2

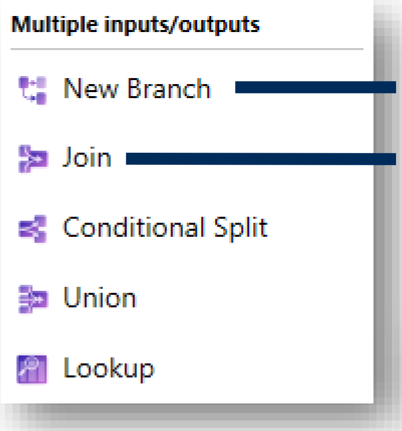
Condition to partition *

Condition

Enter expression... ANY + 🗑

Mapping Data Flows – Transformations



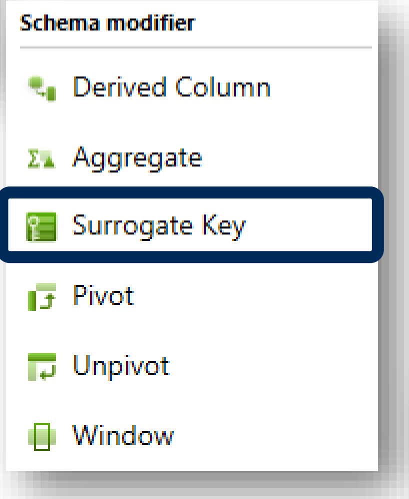


Multiple inputs/outputs

- New Branch
- Join
- Conditional Split
- Union
- Lookup

Multicast

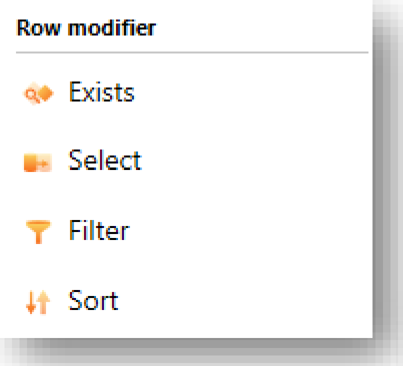
Merge Join



Schema modifier

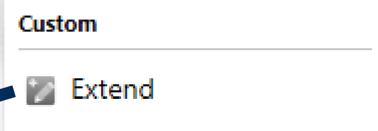
- Derived Column
- Aggregate
- Surrogate Key
- Pivot
- Unpivot
- Window

https://docs.microsoft.com/en-gb/azure/data-factory/data-flow-surrogate-key



Row modifier

- Exists
- Select
- Filter
- Sort

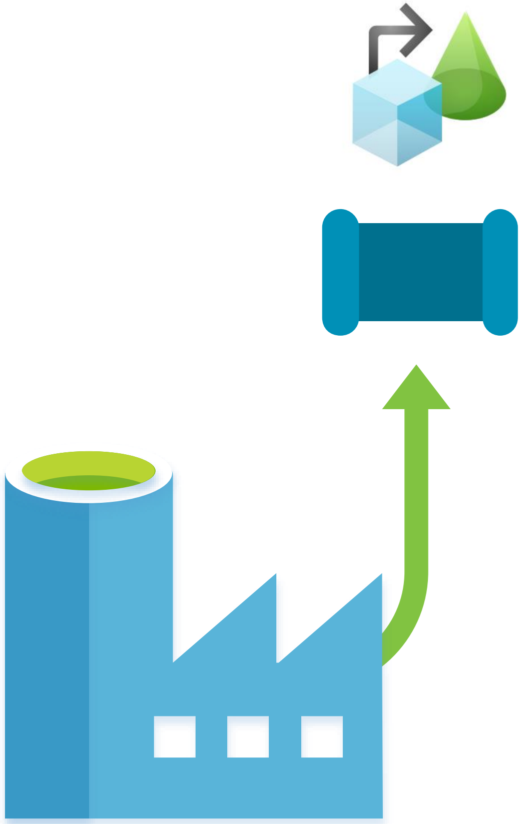


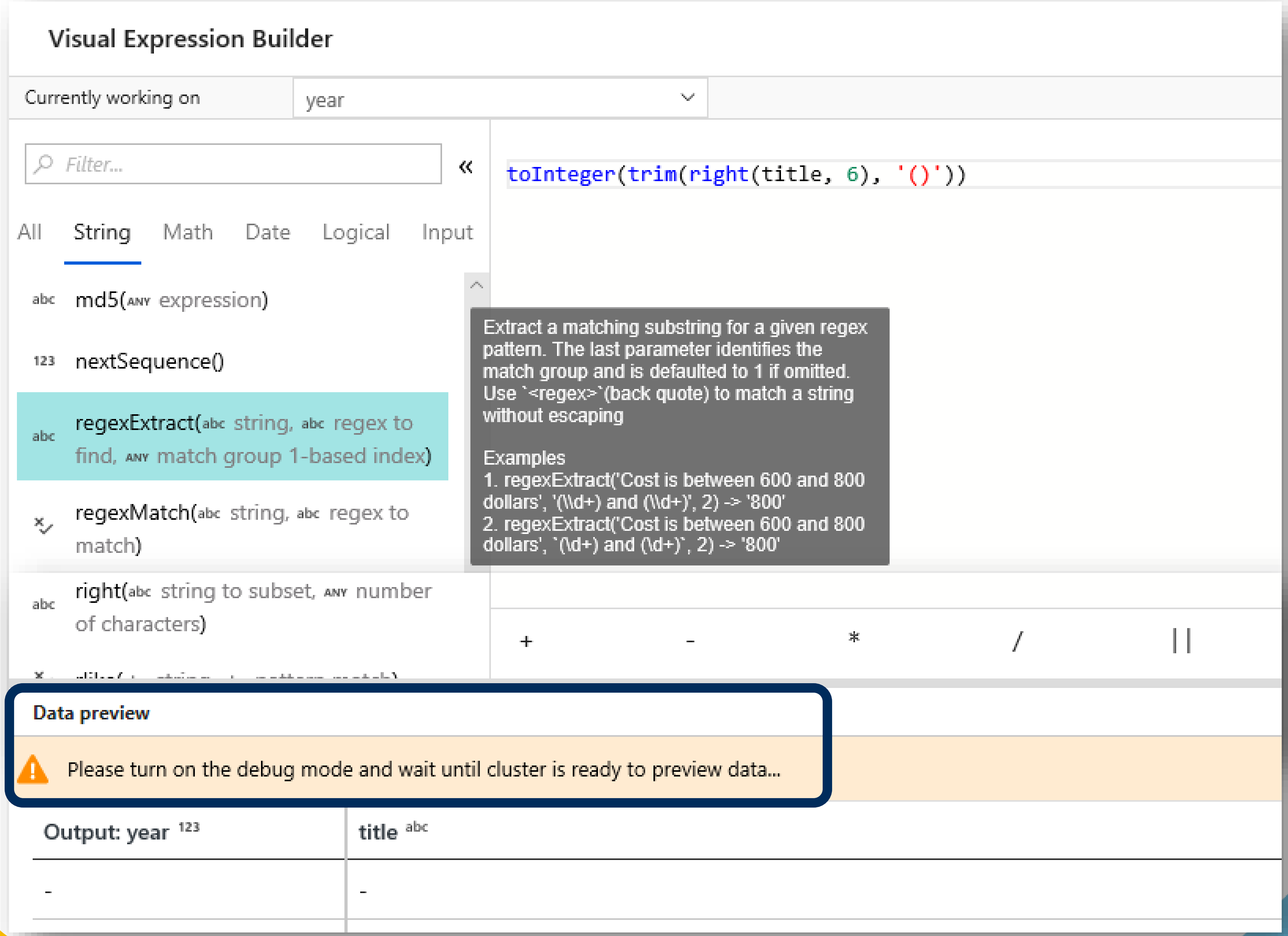
Custom

- Extend

Script Component

Mapping Data Flows – Expression Builder





Visual Expression Builder

Currently working on: year

Filter...

All | 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)
- rlike(abc string, abc pattern match)

```
toInteger(trim(right(title, 6), '()'))
```

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

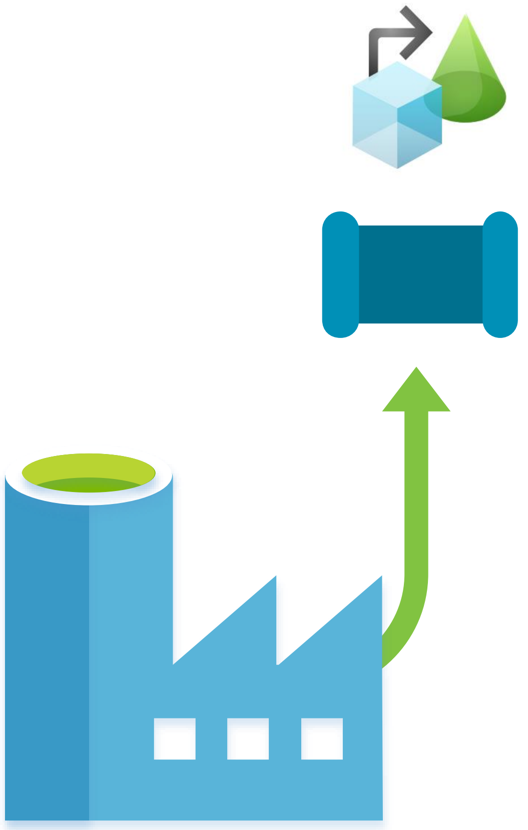
Examples
1. regexExtract('Cost is between 600 and 800 dollars', '(\\d+) and (\\d+)', 2) -> '800'
2. regexExtract('Cost is between 600 and 800 dollars', `(\d+) and (\d+)`, 2) -> '800'

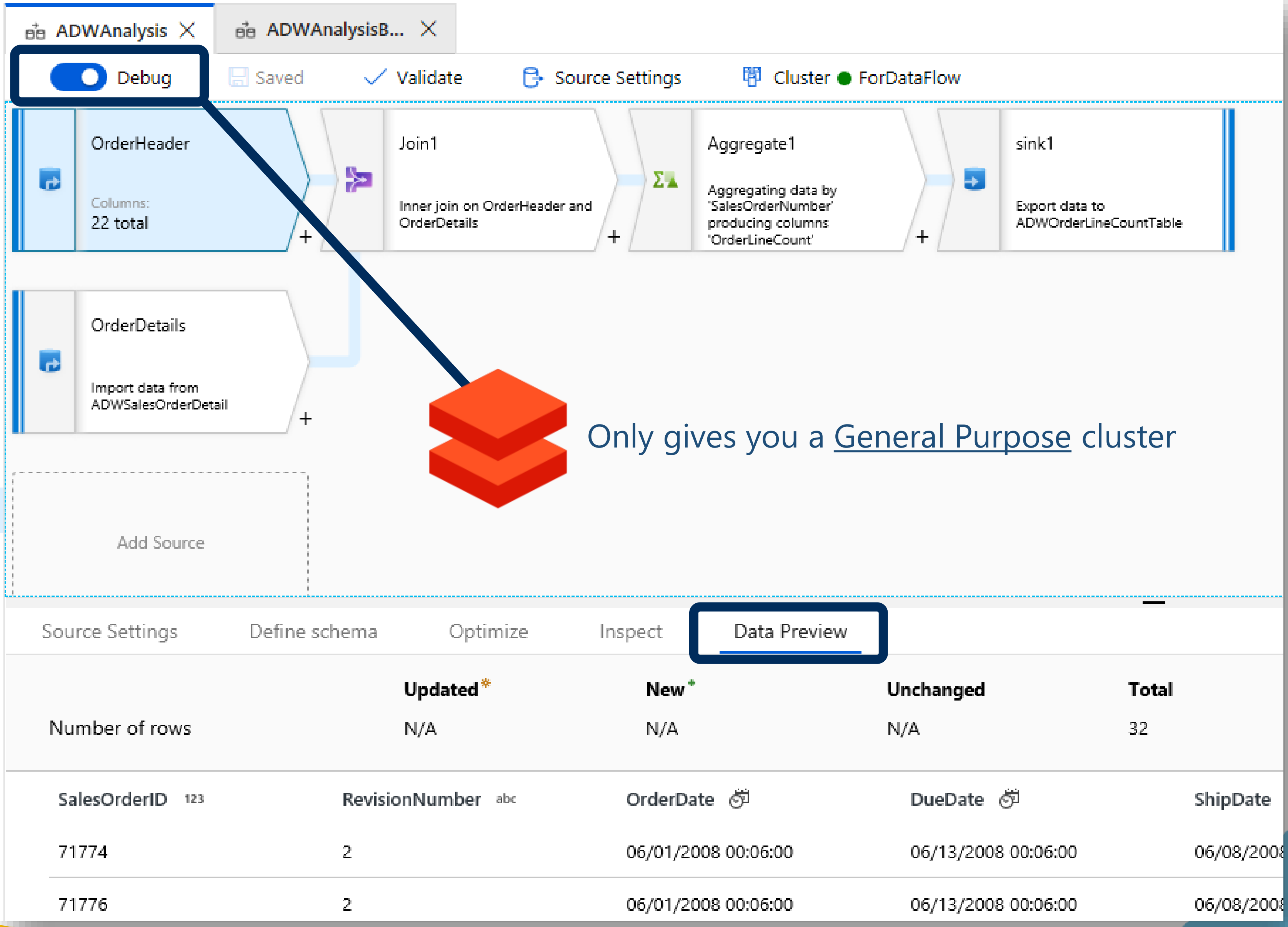
Data preview

⚠ 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



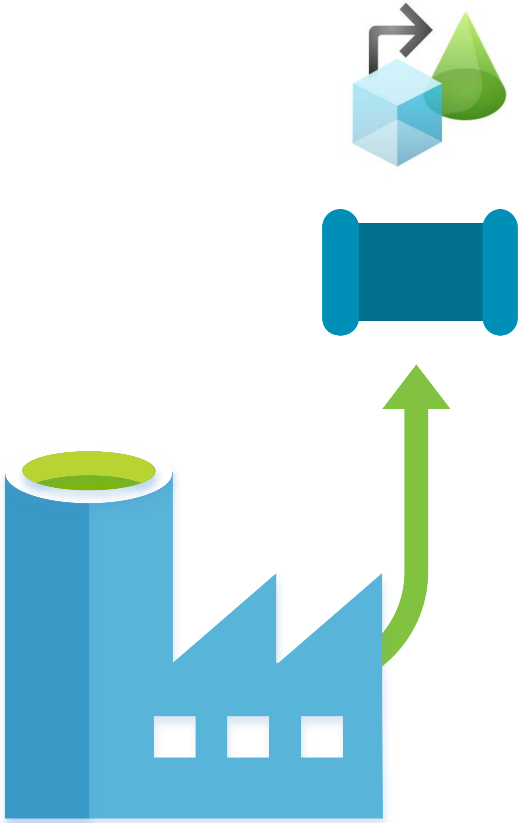


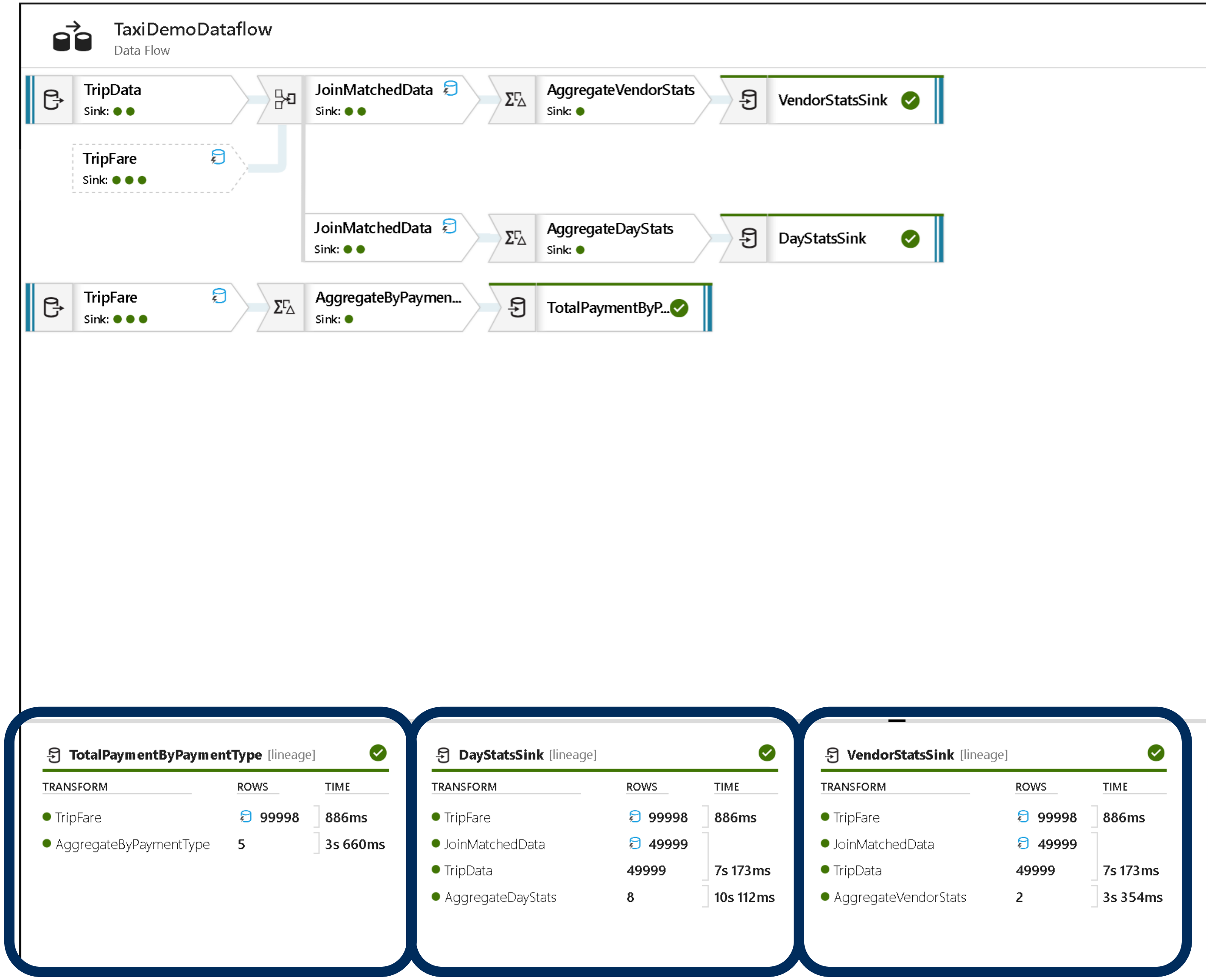
Only gives you a General Purpose cluster

	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





TotalPaymentByPaymentType [lineage] ✓

TRANSFORM	ROWS	TIME
● TripFare	99998	886ms
● AggregateByPaymentType	5	3s 660ms

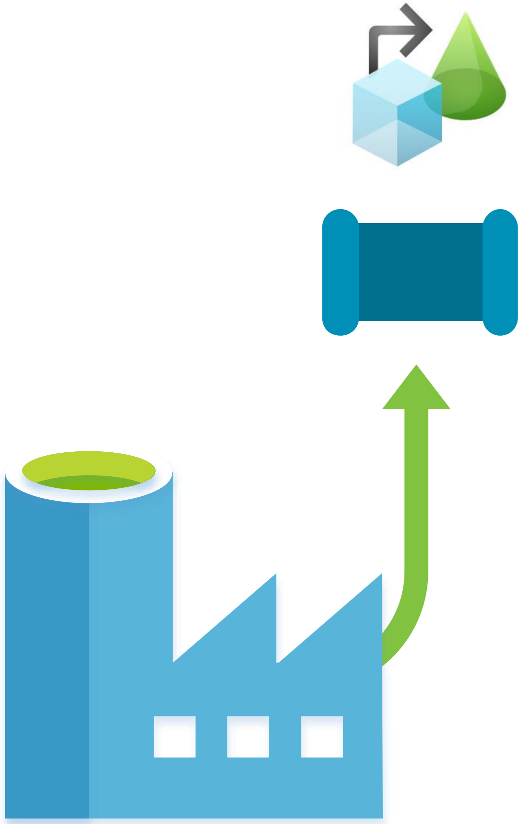
DayStatsSink [lineage] ✓

TRANSFORM	ROWS	TIME
● TripFare	99998	886ms
● JoinMatchedData	49999	
● TripData	49999	7s 173ms
● AggregateDayStats	8	10s 112ms

VendorStatsSink [lineage] ✓

TRANSFORM	ROWS	TIME
● TripFare	99998	886ms
● JoinMatchedData	49999	
● TripData	49999	7s 173ms
● AggregateVendorStats	2	3s 354ms

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

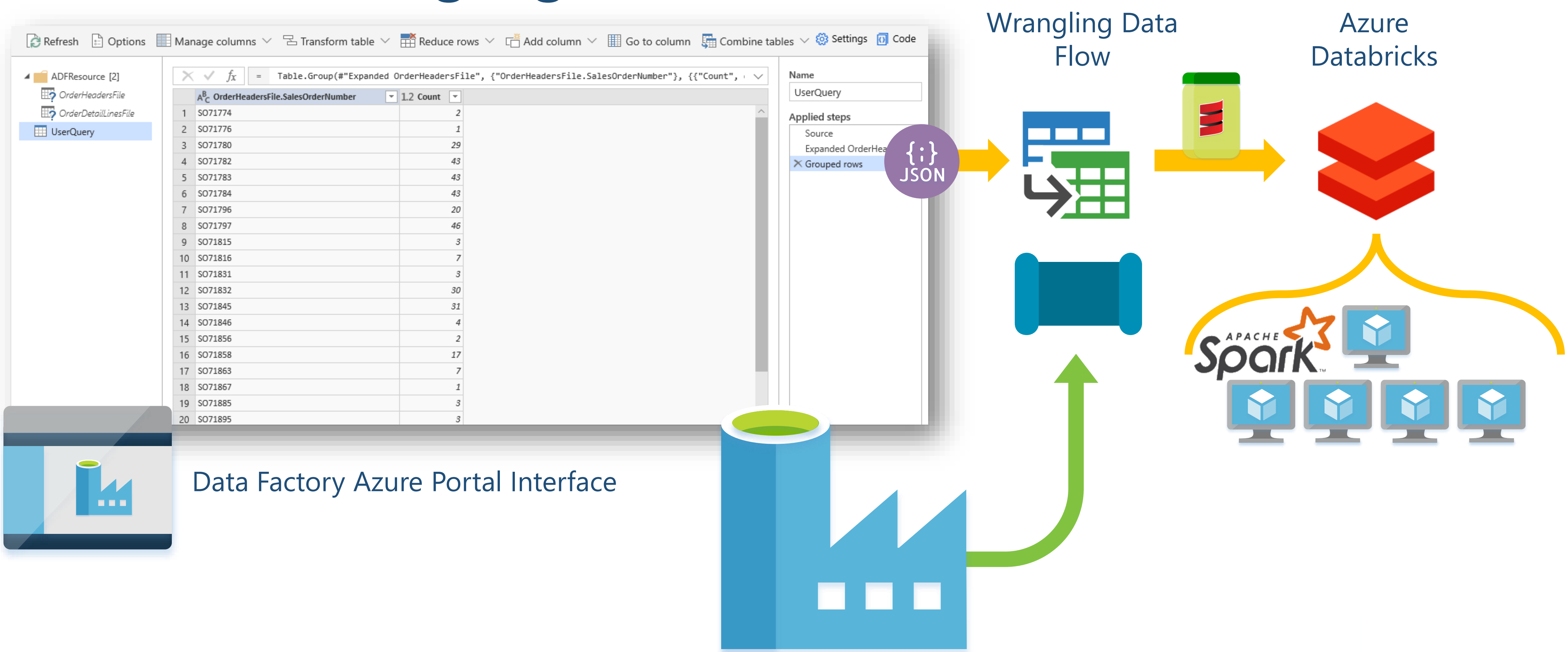
5 **Debug Mode**
https://docs.microsoft.com/en-gb/azure/data-factory/concepts-data-flow-debug-mode

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?



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

	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 OrderHea
- Grouped rows

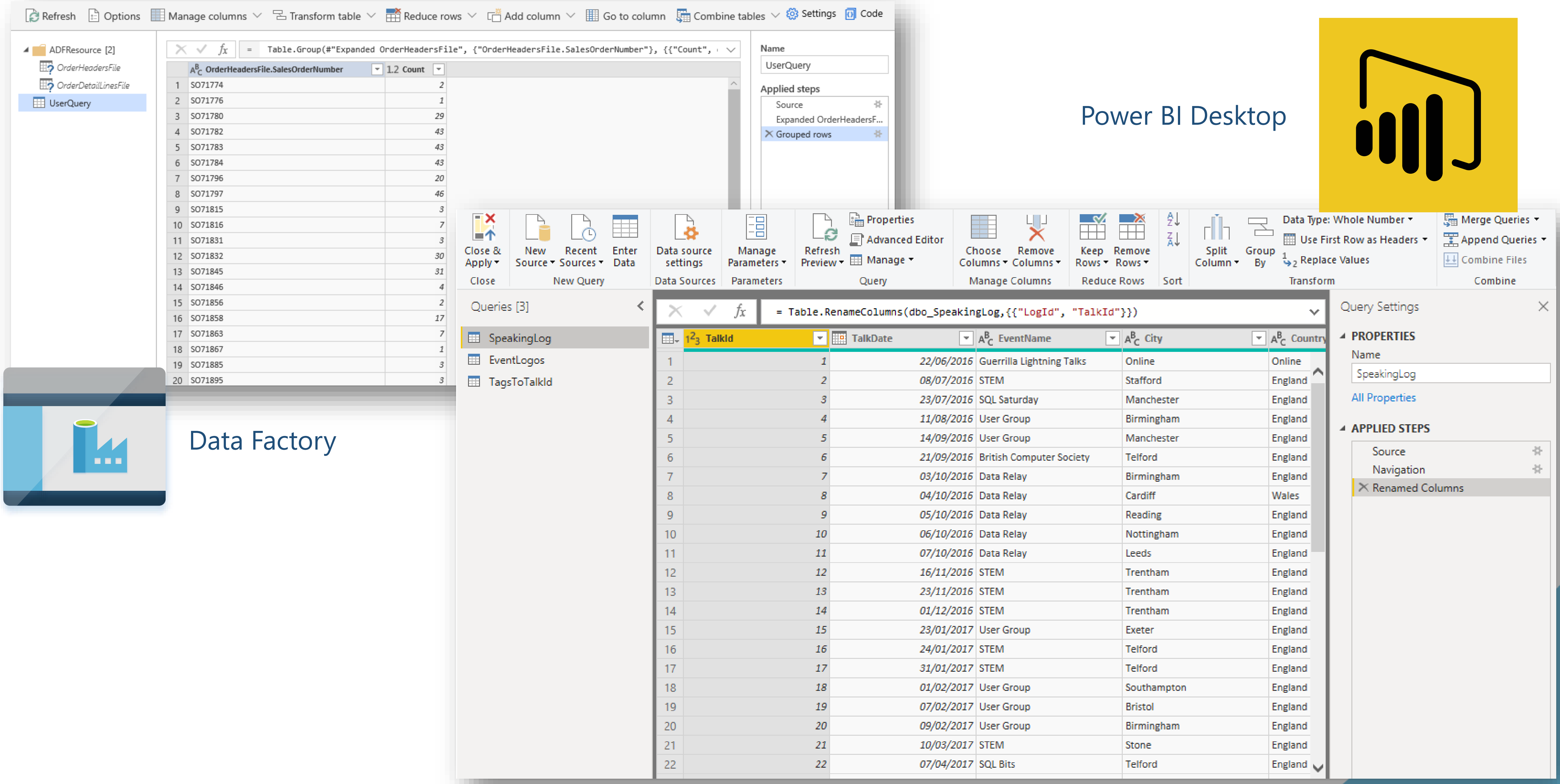
Wrangling Data Flow

Azure Databricks

Data Factory Azure Portal Interface

Data Factory

What is a Wrangling Data Flow?



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

	OrderHeadersFile.SalesOrderNumber	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

Data Factory

Power BI Desktop

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

	TalkId	TalkDate	EventName	City	Countr
1	1	22/06/2016	Guerrilla Lightning Talks	Online	Online
2	2	08/07/2016	STEM	Stafford	England
3	3	23/07/2016	SQL Saturday	Manchester	England
4	4	11/08/2016	User Group	Birmingham	England
5	5	14/09/2016	User Group	Manchester	England
6	6	21/09/2016	British Computer Society	Telford	England
7	7	03/10/2016	Data Relay	Birmingham	England
8	8	04/10/2016	Data Relay	Cardiff	Wales
9	9	05/10/2016	Data Relay	Reading	England
10	10	06/10/2016	Data Relay	Nottingham	England
11	11	07/10/2016	Data Relay	Leeds	England
12	12	16/11/2016	STEM	Trentham	England
13	13	23/11/2016	STEM	Trentham	England
14	14	01/12/2016	STEM	Trentham	England
15	15	23/01/2017	User Group	Exeter	England
16	16	24/01/2017	STEM	Telford	England
17	17	31/01/2017	STEM	Telford	England
18	18	01/02/2017	User Group	Southampton	England
19	19	07/02/2017	User Group	Bristol	England
20	20	09/02/2017	User Group	Birmingham	England
21	21	10/03/2017	STEM	Stone	England
22	22	07/04/2017	SQL Bits	Telford	England

Query Settings

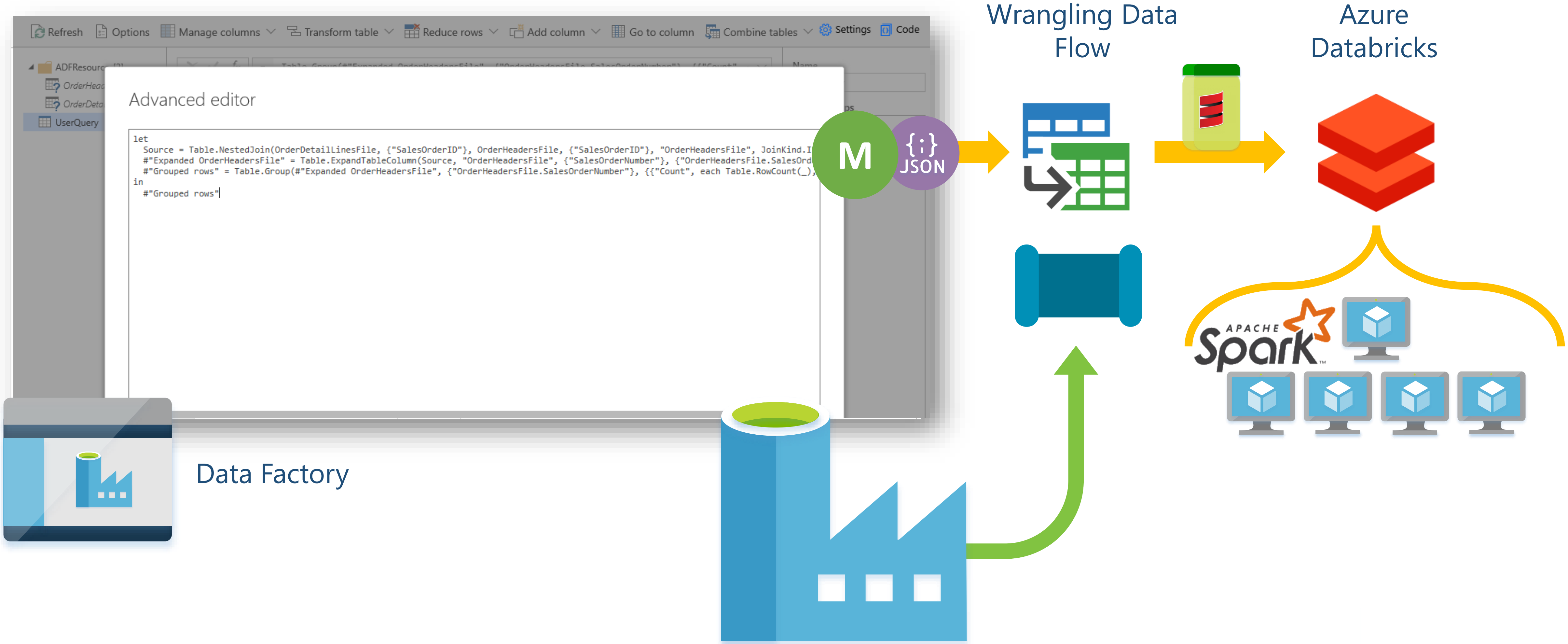
PROPERTIES

Name: SpeakingLog

All Properties

APPLIED STEPS
- Source
- Navigation
- Renamed Columns

What is a Wrangling Data Flow?



```
let
  Source = Table.NestedJoin(OrderDetailLinesFile, {"SalesOrderID"}, OrderHeadersFile, {"SalesOrderID"}, "OrderHeadersFile", JoinKind.I
  #"Expanded OrderHeadersFile" = Table.ExpandTableColumn(Source, "OrderHeadersFile", {"SalesOrderNumber"}, {"OrderHeadersFile.SalesOrd
  #"Grouped rows" = Table.Group(#"Expanded OrderHeadersFile", {"OrderHeadersFile.SalesOrderNumber"}, {{"Count", each Table.RowCount(_),
in
  #"Grouped rows"
```

Wrangling Data Flow

Azure Databricks

Data Factory

Data Factory

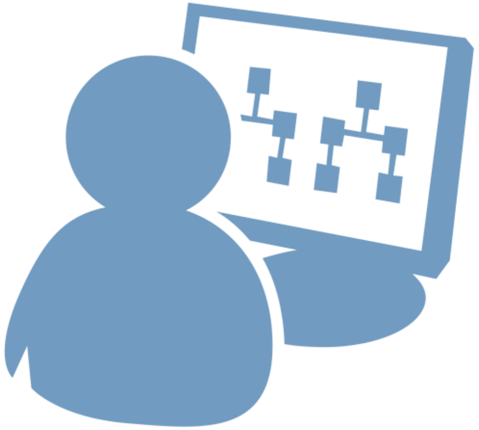
DEMO

Note to self - start clusters!

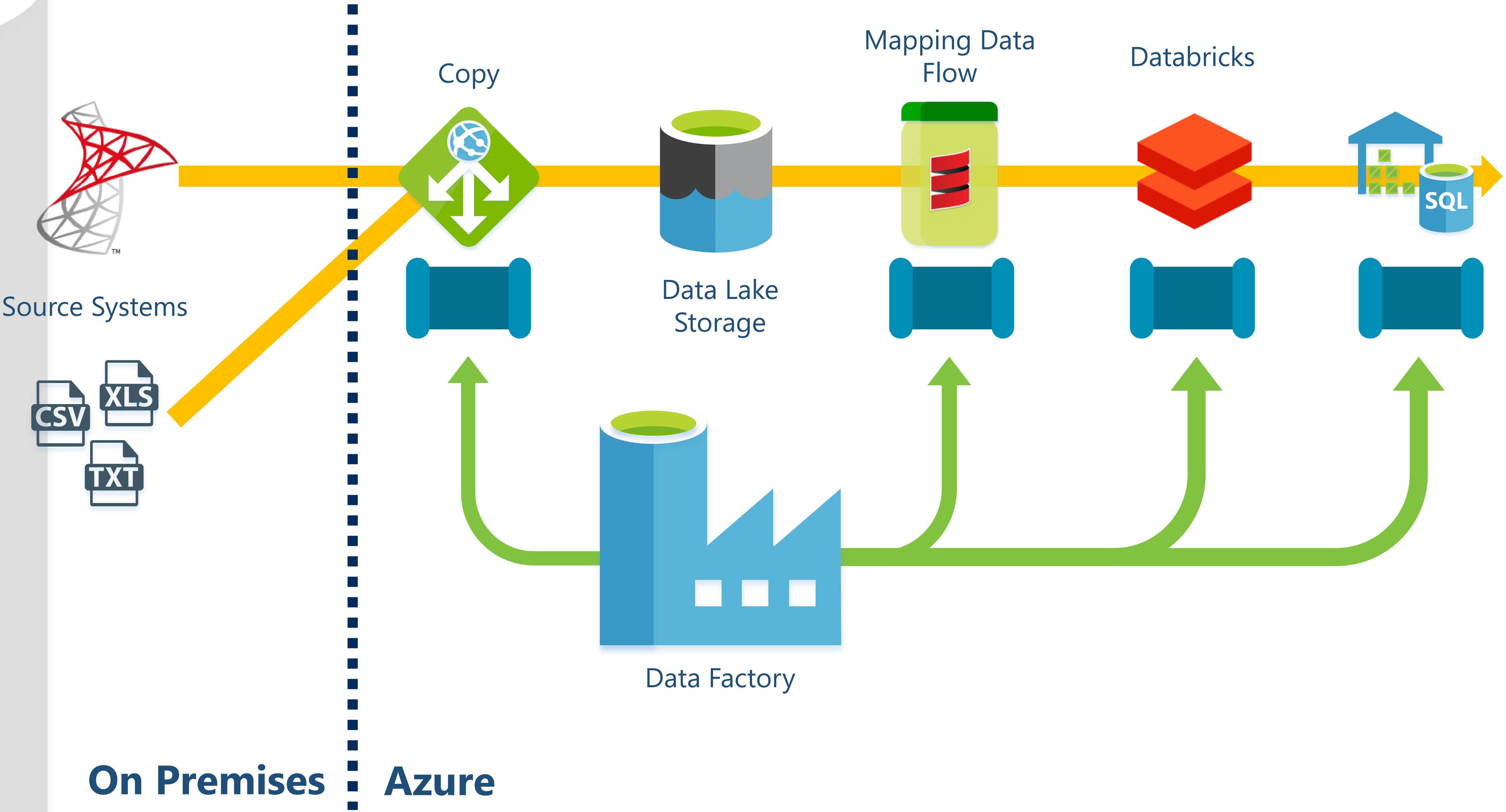
Demo Summary

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

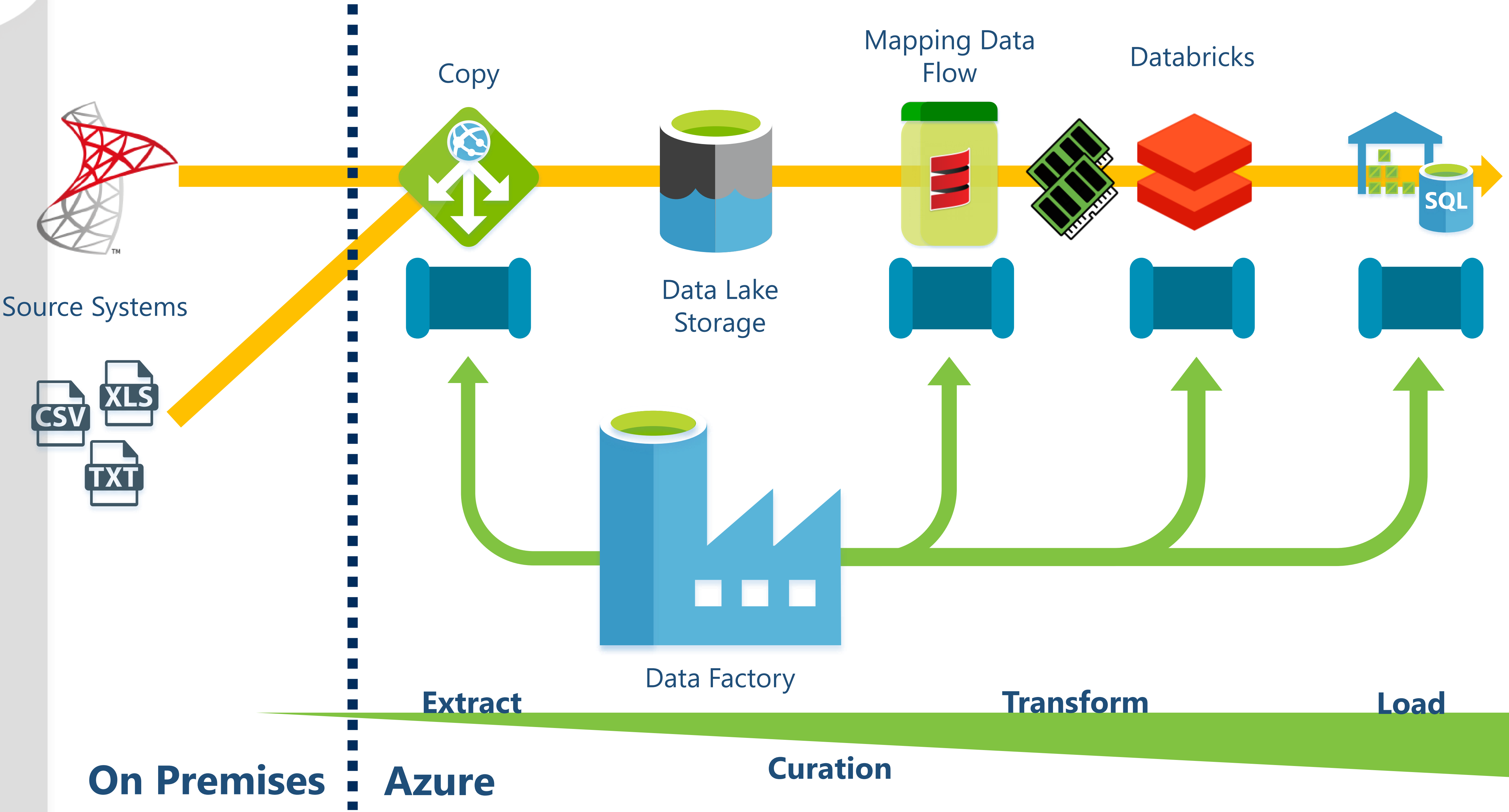
Design Patterns



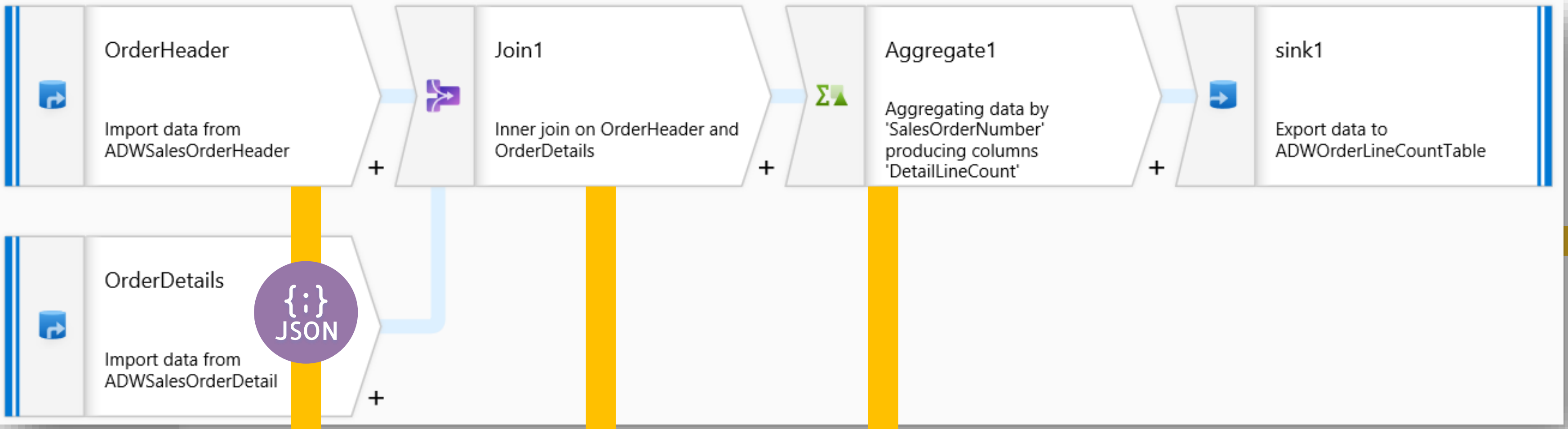
Mapping Data Flow Future Design Patterns ???



Mapping Data Flow Future Design Patterns ???



Mapping Data Flow Future Design Patterns ???



OrderHeader

Import data from ADWSalesOrderHeader

Join1

Inner join on OrderHeader and OrderDetails

Aggregate1

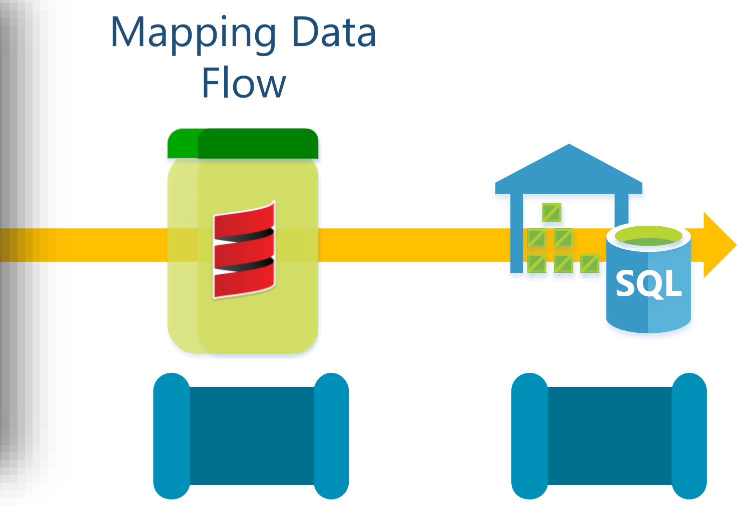
Aggregating data by 'SalesOrderNumber' producing columns 'DetailLineCount'

sink1

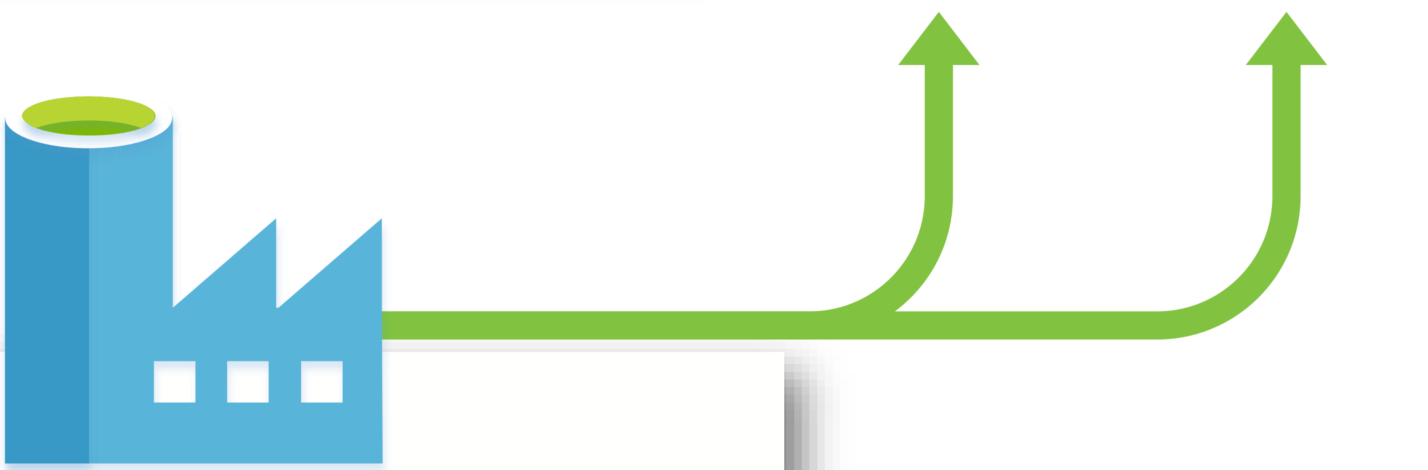
Export data to ADWOrderLineCountTable

OrderDetails

Import data from ADWSalesOrderDetail

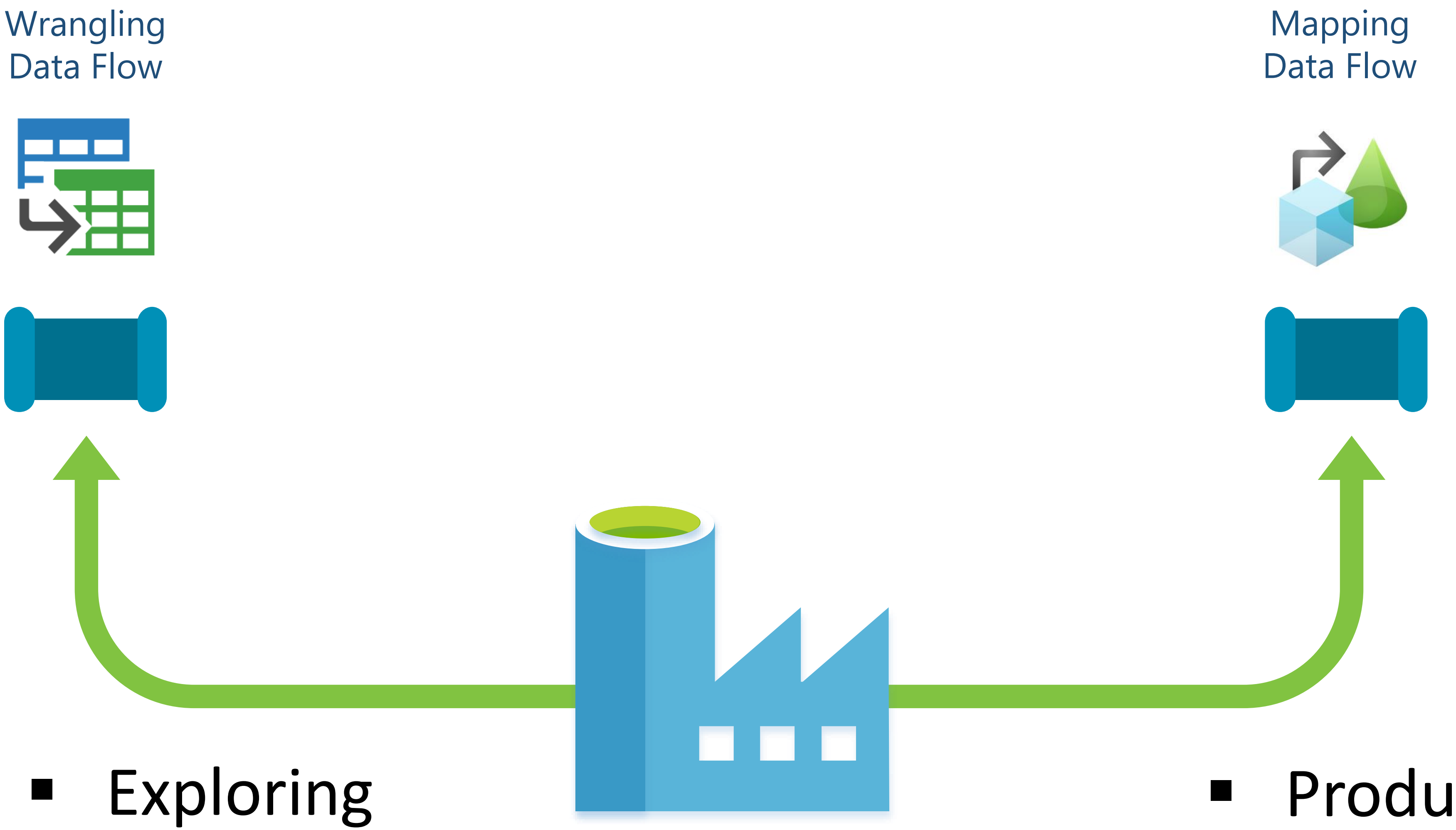


```
"fileName": {
    "value": "@dataset().FileName",
    "type": "Expression"
},
"folderPath": {
    "value": "@dataset().SourceDIR",
    "type": "Expression"
}
```



```
"transformations": [
    {
        "name": "Join1",
        "script": "OrderHeader, OrderDetail join(OrderHeader@SalesOrderID == Orde etail@SalesOrderID,\n\tjoinType:'inner',\n\tbroadcast: 'none')~> Join1"
    },
    {
        "name": "Aggregate1",
        "script": "Join1 aggregate(groupBy(SalesOrderNumber),\n\tDetailLineCount = count(SalesOrderDetailID)) ~> Aggregate1"
    }
]
```

What else?



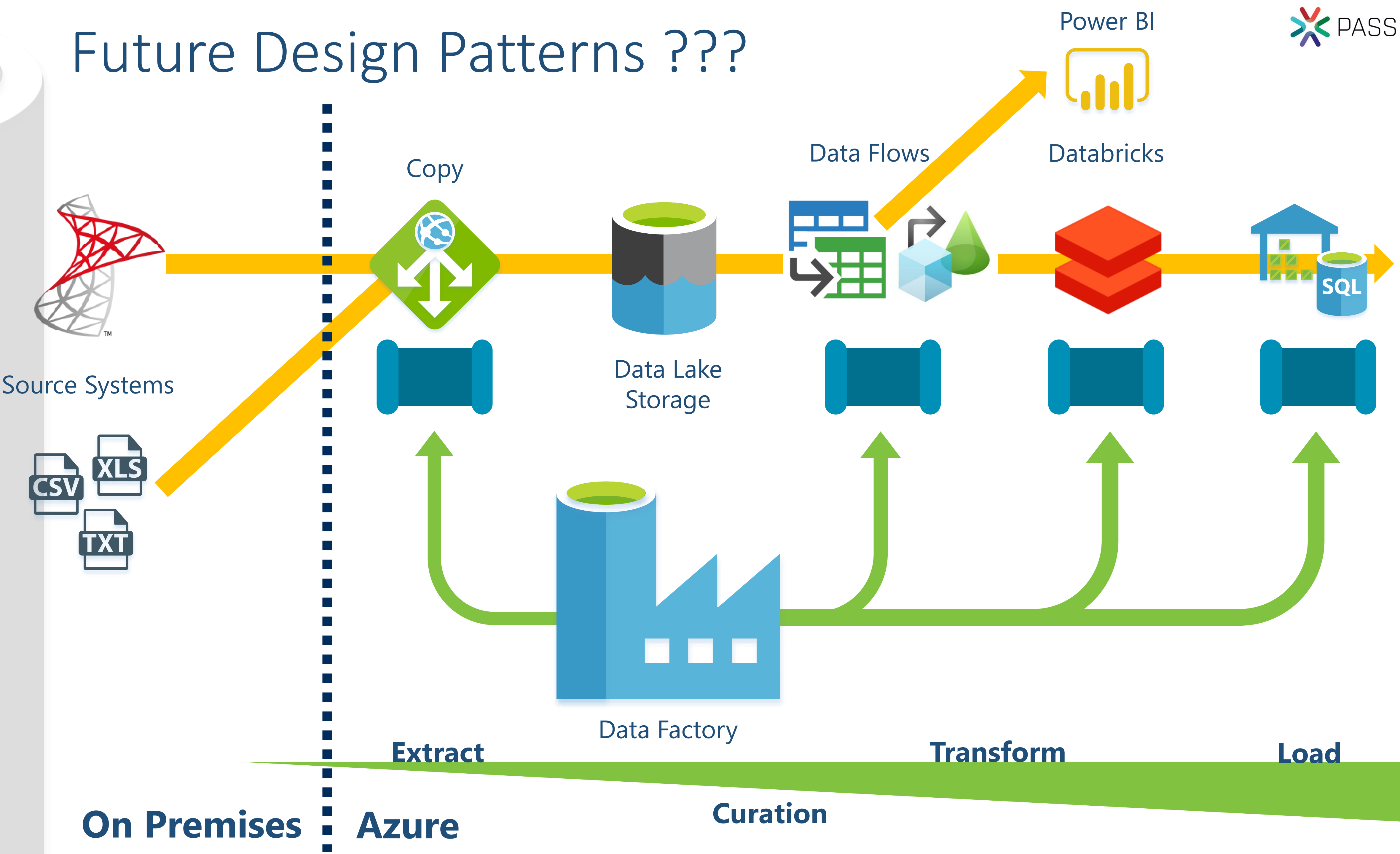
Wrangling Data Flow

Mapping Data Flow

- Exploring
- Experimenting
- Analysing

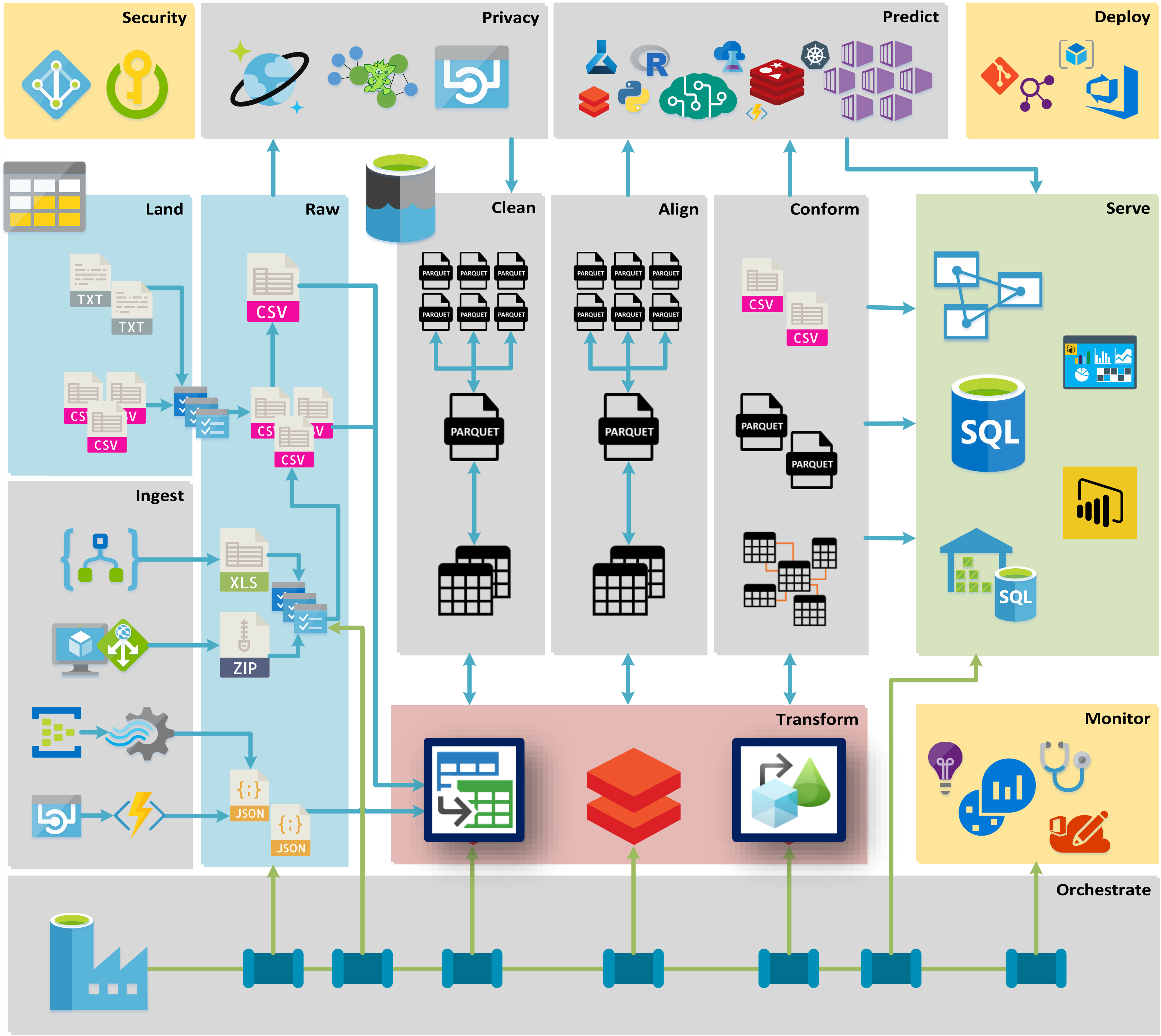
- Productionising
- Engineering
- Warehousing

Future Design Patterns ???



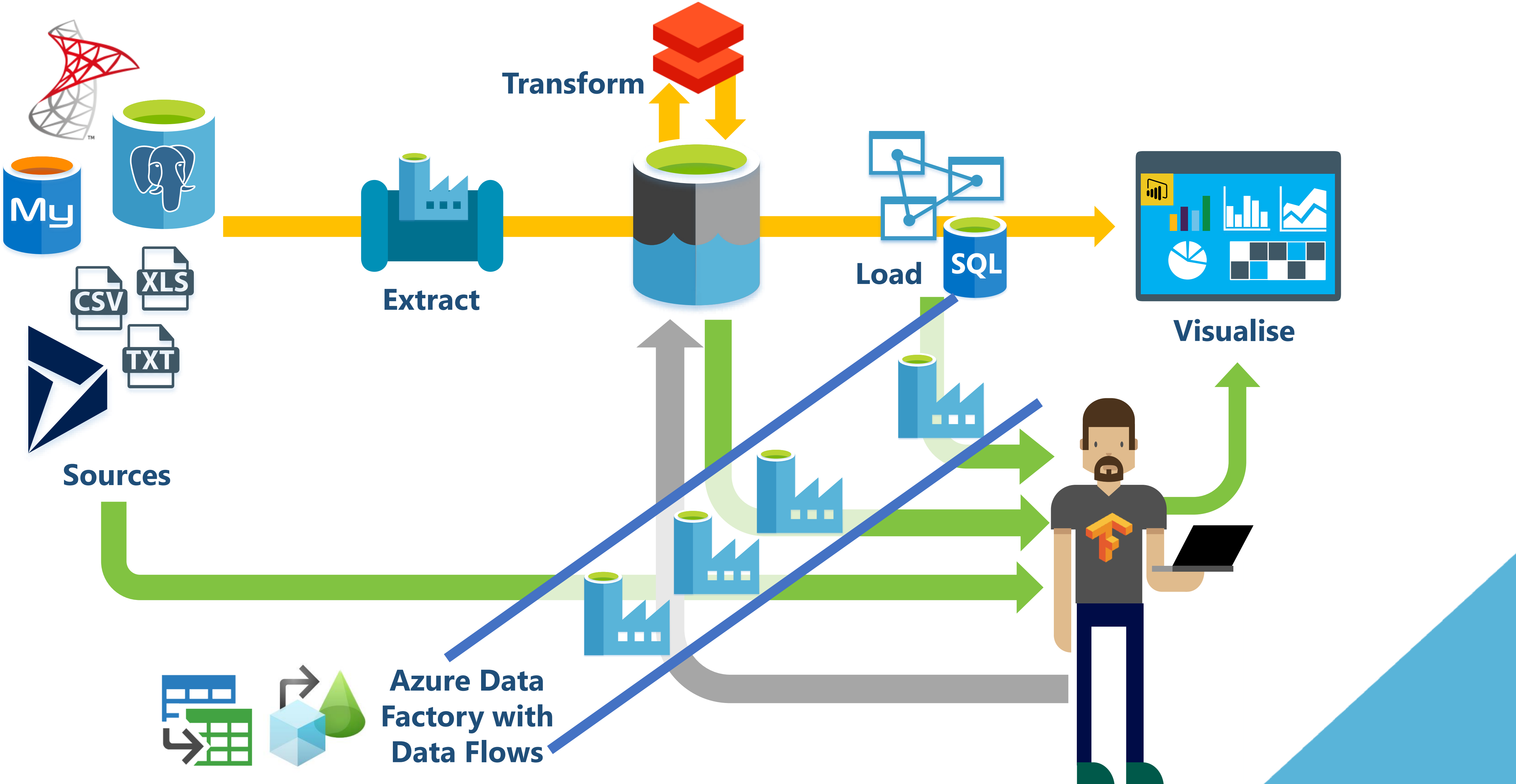
Conclusions

Why use Azure Data Factory?

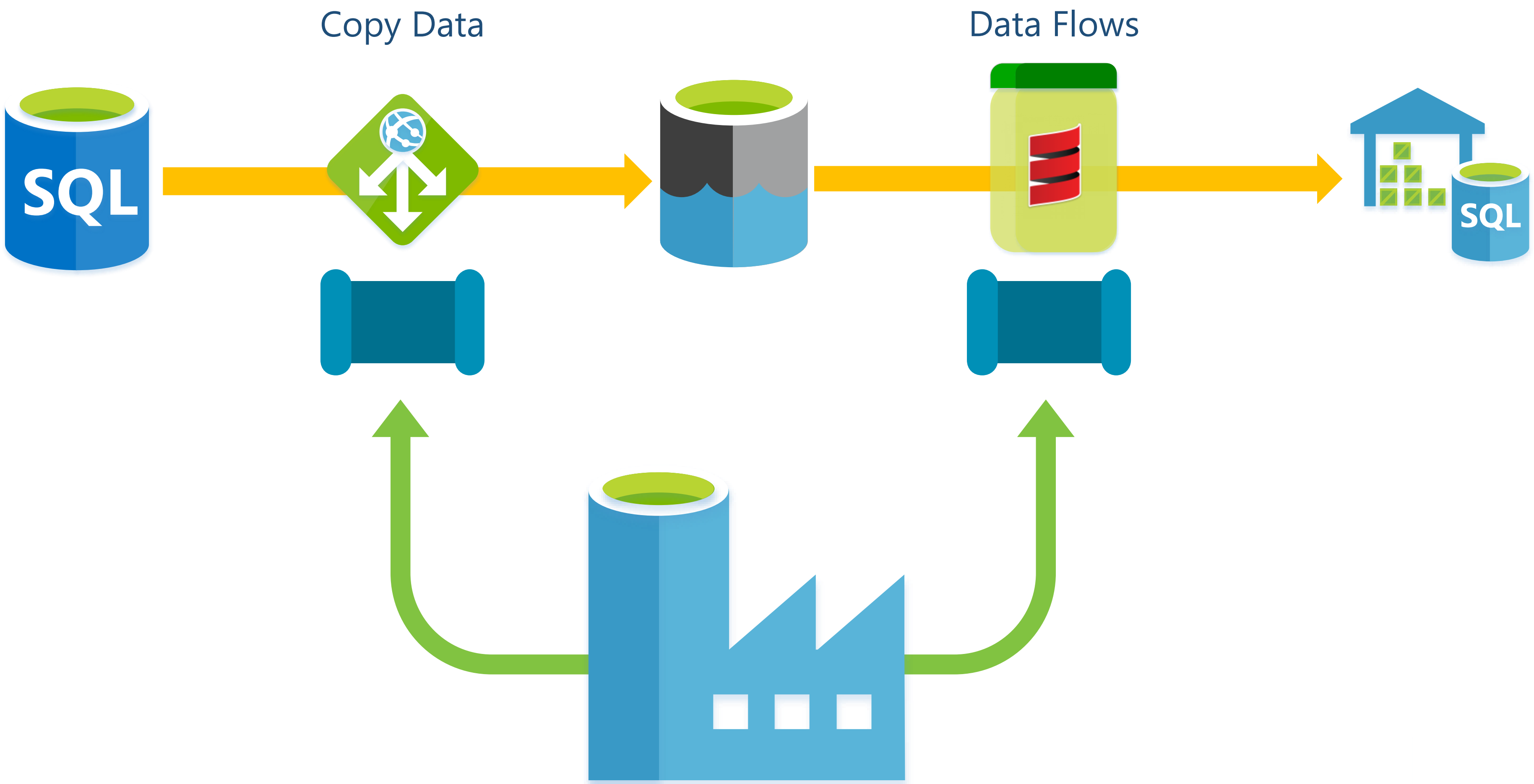


Azure PaaS Data Warehouse & Analytics Platform

Data Factory for the Data Scientist



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.

Thanks for Listening

Paul Andrew

@MrPaulAndrew

Microsoft® MVP Most Valuable Professional

altius

Email: paul@mrpaulandrew.com
Blog: mrpaulandrew.com
GitHub: github.com/mrpaulandrew ← PDF Slides