

Complex Orchestration

With Dynamic Data Factory Pipelines



Paul Andrew | Principal Consultant & Solution Architect

PLATINUM SPONSOR

STRATEGIC PARTNER

TECHNOLOGY
INNOVATION
DATA
KNOWLEDGE

GOLD SPONSORS



CLOUDS ON MARS



SILVER SPONSOR



BRONZE SPONSOR





<https://github.com/mrpaulandrew>

CommunityEvents

Demo code, content and slides from various community events.

● C++

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

Complex Orchestration

With Dynamic Data Factory Pipelines



Azure Data
Factory

A very quick
overview

Extensibility &
Parallelism

Custom Activities
SSIS IR & Packages

More Design
Patterns

Bootstrapping
Hosted IR vs IaaS
Frameworks

Complex Orchestration

With Dynamic Data Factory Pipelines



Azure Data
Factory

A very quick
overview

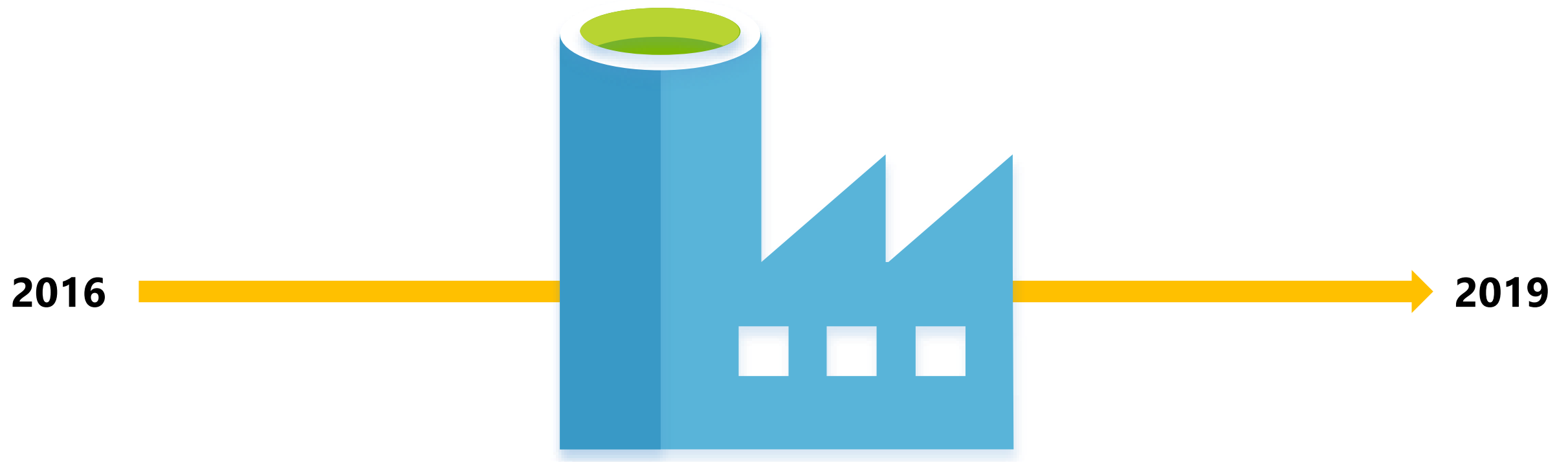
Extensibility &
Parallelism

Custom Activities
SSIS IR & Packages

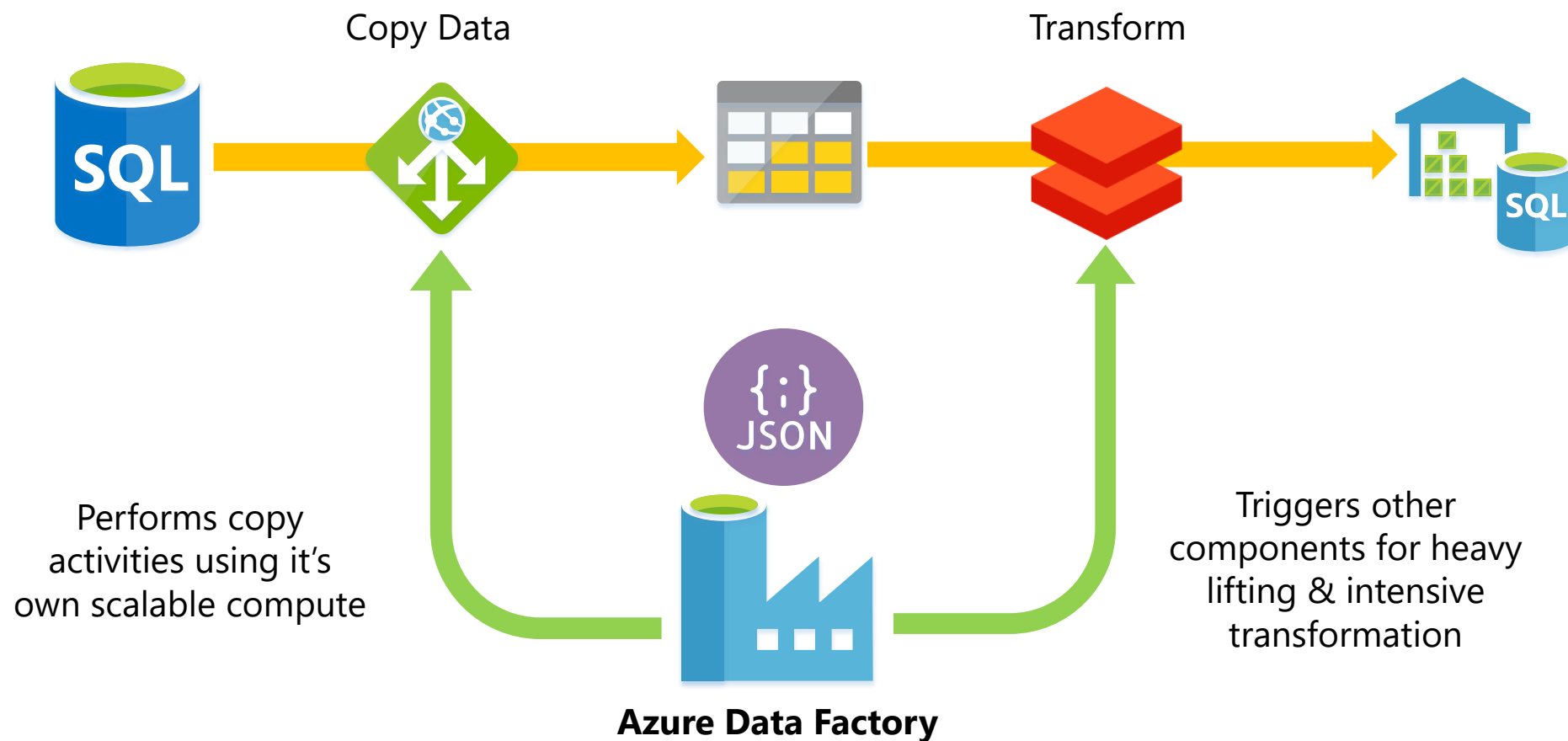
More Design
Patterns

Bootstrapping
Hosted IR vs IaaS
Frameworks

Azure Data Factory



What is Azure Data Factory?



Data Factory Components – Recap



```
{
  "name": "GenericSQLDB",
  "type": "Microsoft.DataFactory/factories/linkedservices",
  "properties": {
    "parameters": {
      "ServerInstance": {
        "type": "String"
      },
      "DatabaseName": {
        "type": "String"
      },
      "SQLUser": {
        "type": "String"
      },
      "SQLPassword": {
        "type": "String"
      }
    },
    "type": "AzureSqlDatabase",
    "typeProperties": {
      "connectionString": "Integrated Security=False;Encrypt=True;ConnectionTimeout=30;
Data Source=@{linkedService().ServerInstance};
InitialCatalog=@{linkedService().DatabaseName};
UserID=@{linkedService().SQLUser};
Password=@{linkedService().SQLPassword}"
    }
  }
}
```

1 **Linked Services** ✓

2 **Data Sets** ✓

3 **Activities** ✓

4 **Pipelines** ✓

5 **Triggers** ✗

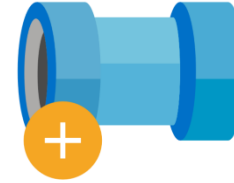
{:}
JSON

Integration Runtimes

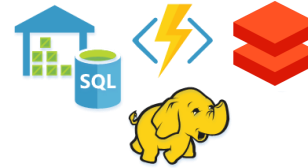
1

Azure
Integration Runtime

Movement Hours



Activity
Orchestration



Flexible Region



2

SSIS
Integration Runtime

SSIS Package
Execution



Specified Region



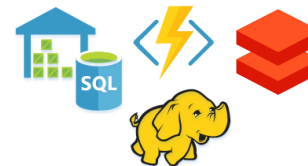
3

Self Hosted
Integration Runtime

Gateway Access



Activity
Orchestration



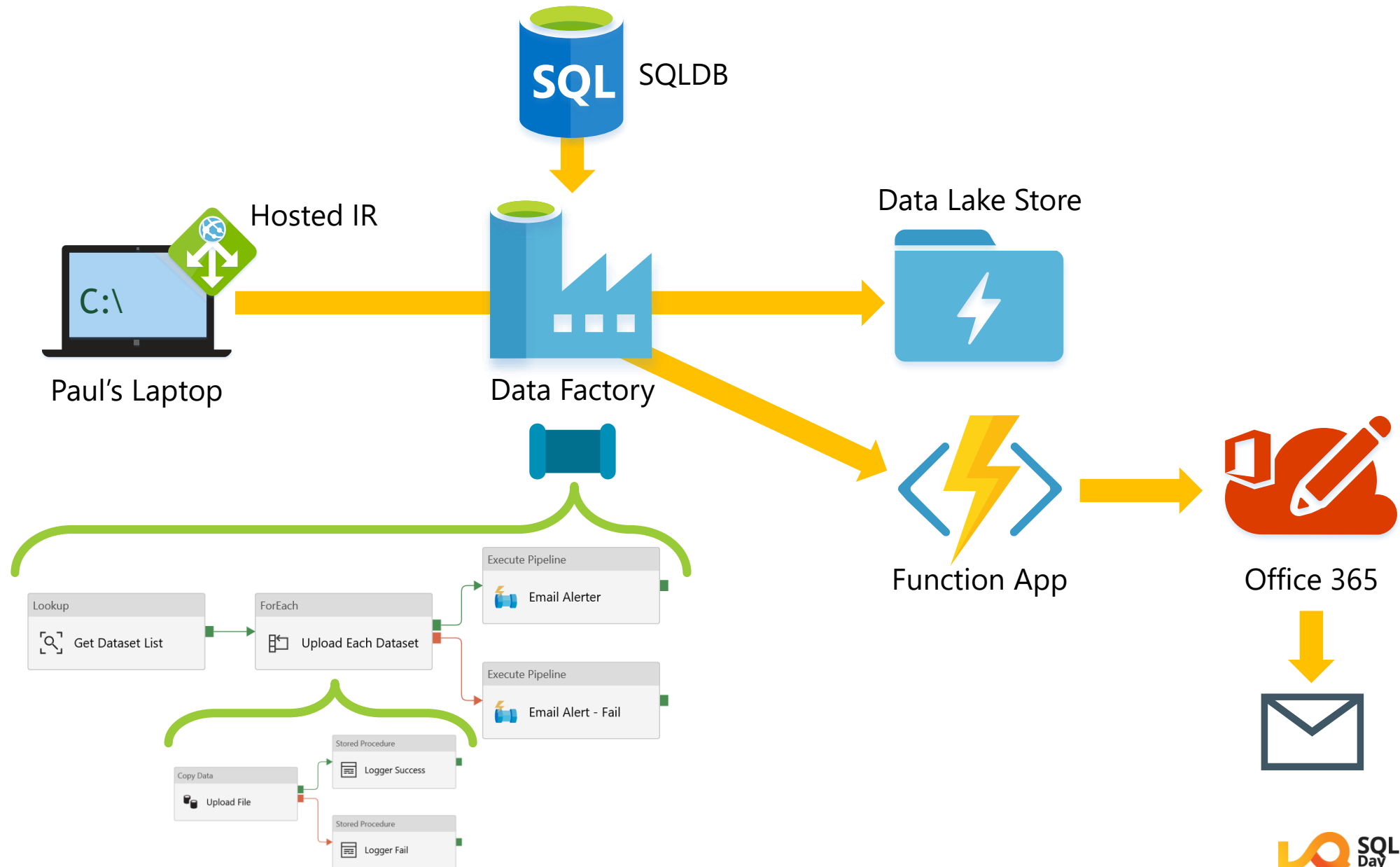
Virtual Machine



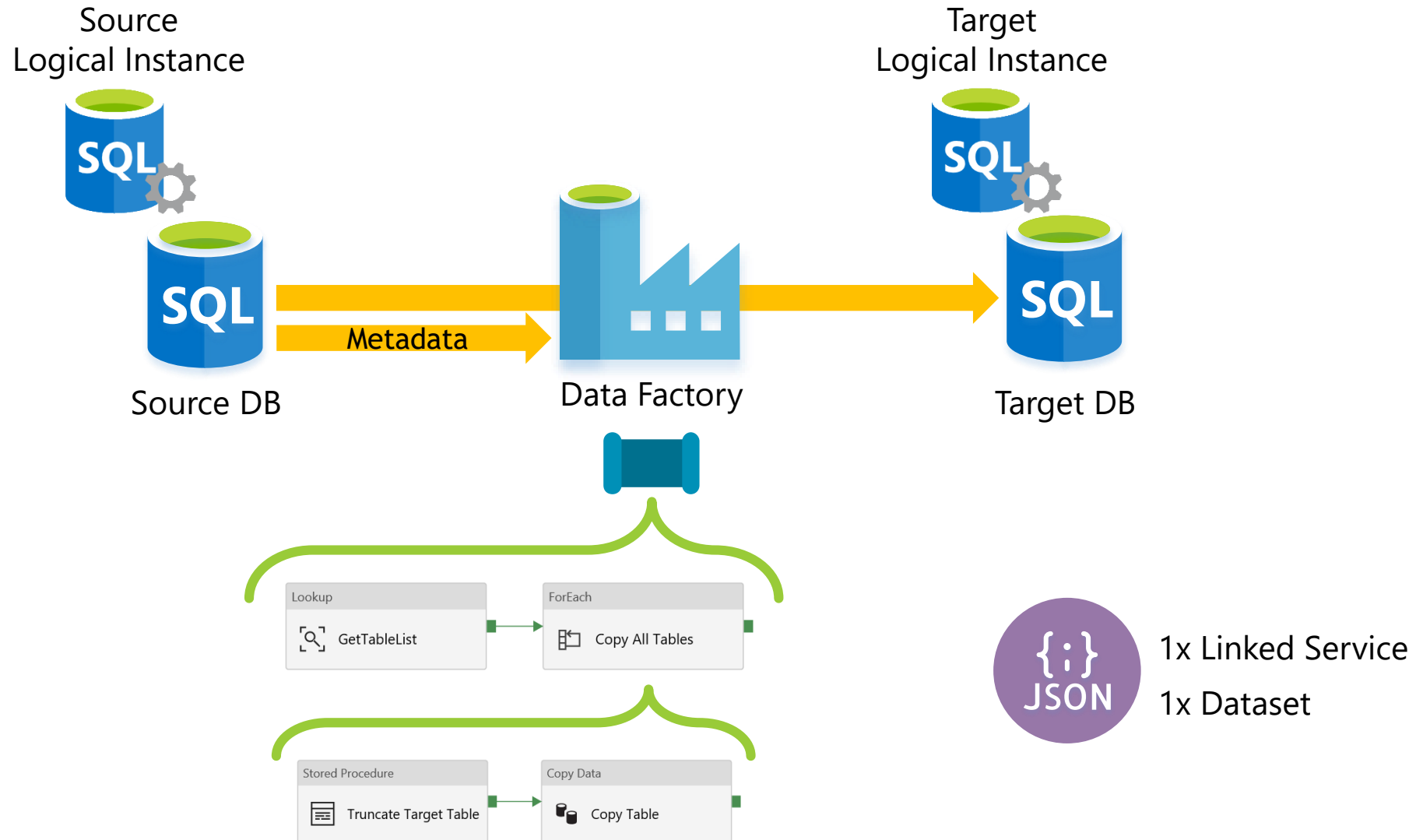
Demo



Demo Architecture 1



Demo Architecture 2



Complex Orchestration

With Dynamic Data Factory Pipelines



Azure Data
Factory

A very quick
overview

Extensibility &
Parallelism

Custom Activities
SSIS IR & Packages

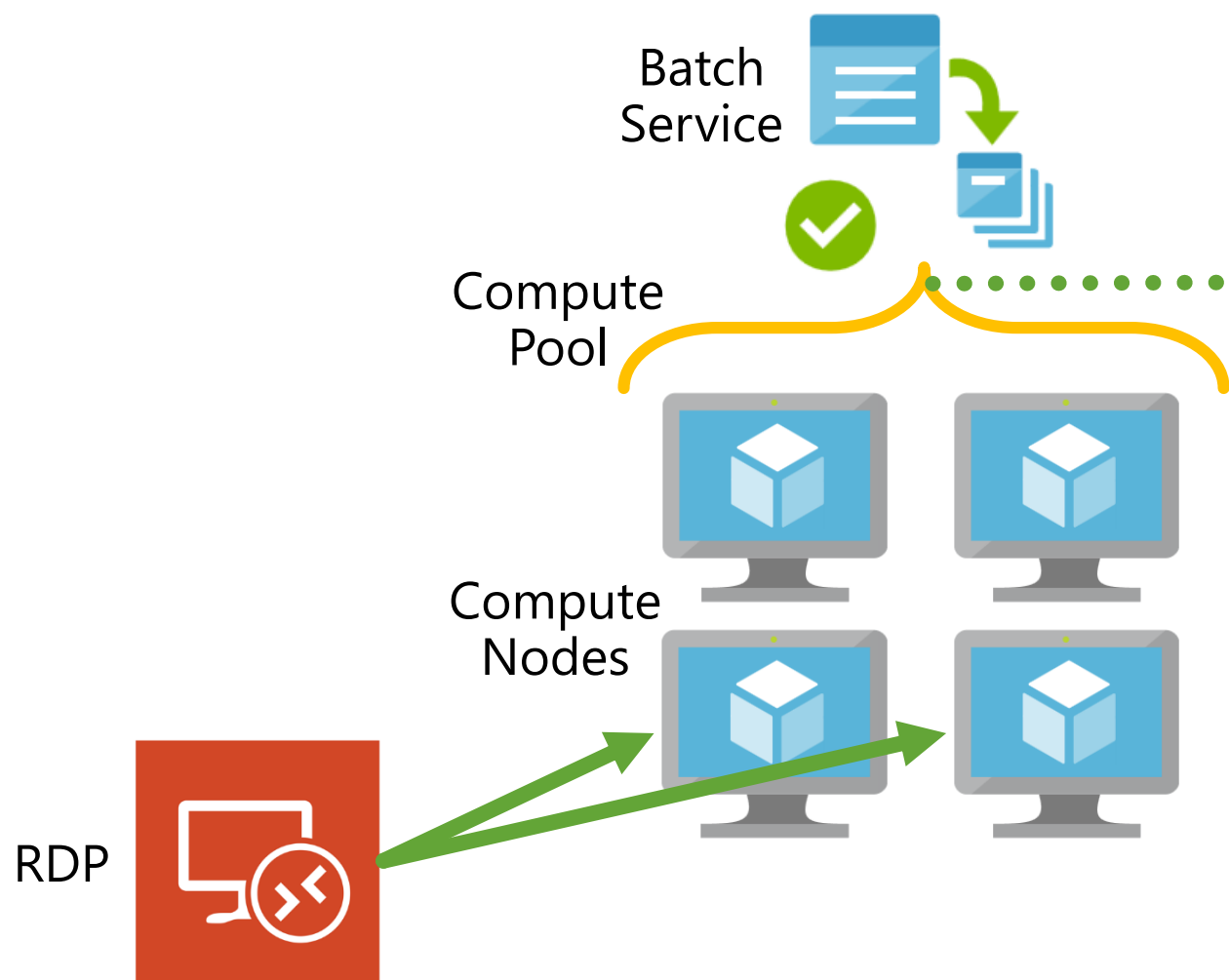
More Design
Patterns

Bootstrapping
Hosted IR vs IaaS
Frameworks










ADF Extensibility

1

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



VM node size set per compute pool:

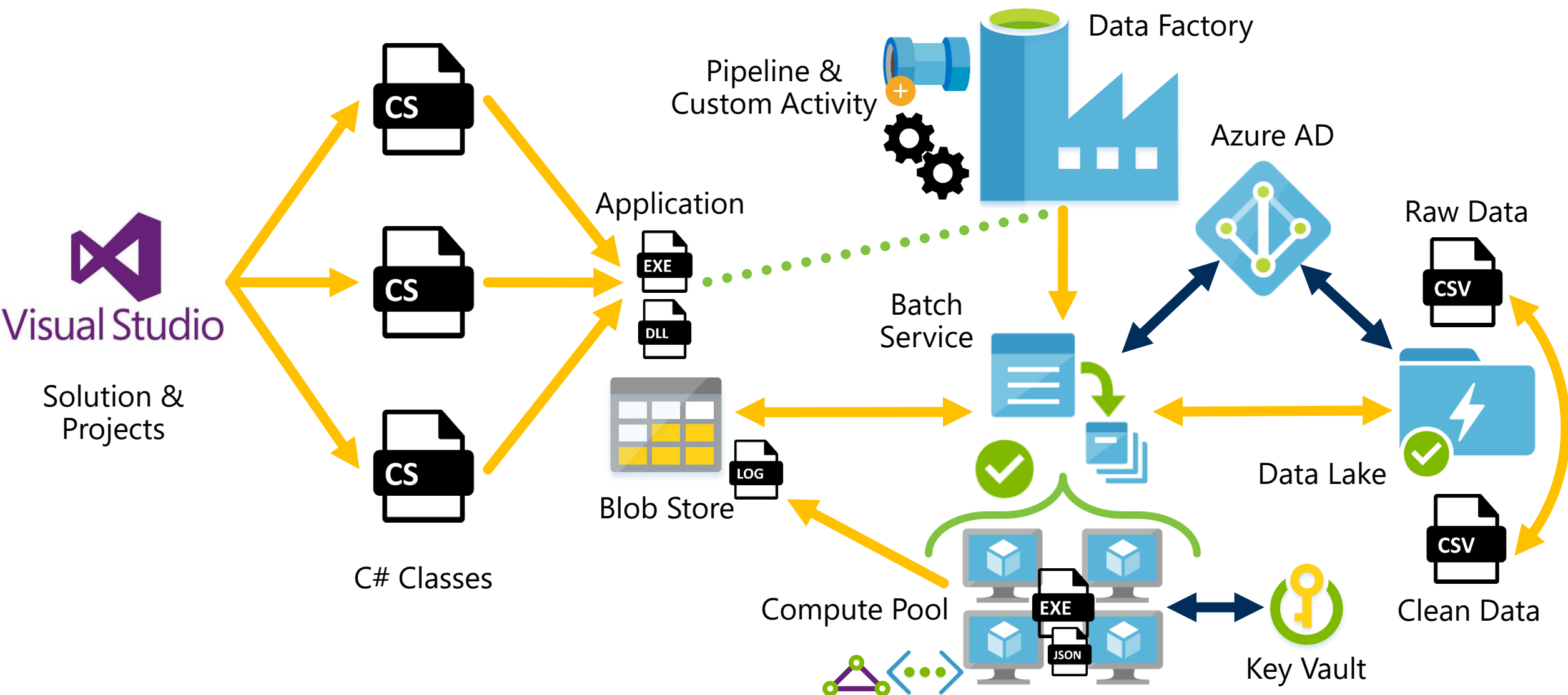
A1	Standard	★	A2	Standard	★	A3	Standard	★
1	Cores		2	Cores		4	Cores	
1.8	GB		3.5	GB		7	GB	
	1 TB	OS disk size		1 TB	OS disk size		1 TB	OS disk size
	70 GB	Resource disk size		135 GB	Resource disk size		285 GB	Resource disk size
	2	Max data disk		4	Max data disk		8	Max data disk
Unable to display pricing			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

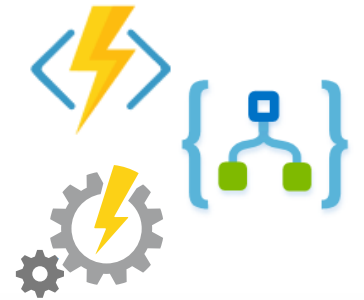


ADF Extensibility Continued

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

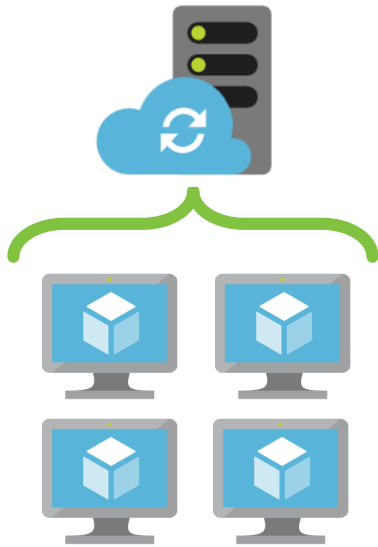
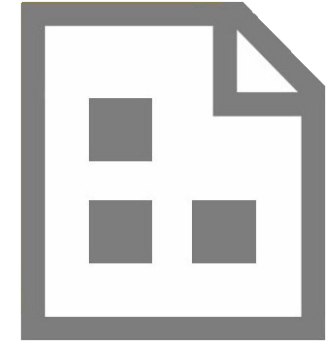
Use [expressions, functions](#) or refer to [system variables](#) in the 'value' column.

Parameterizable properties ⓘ

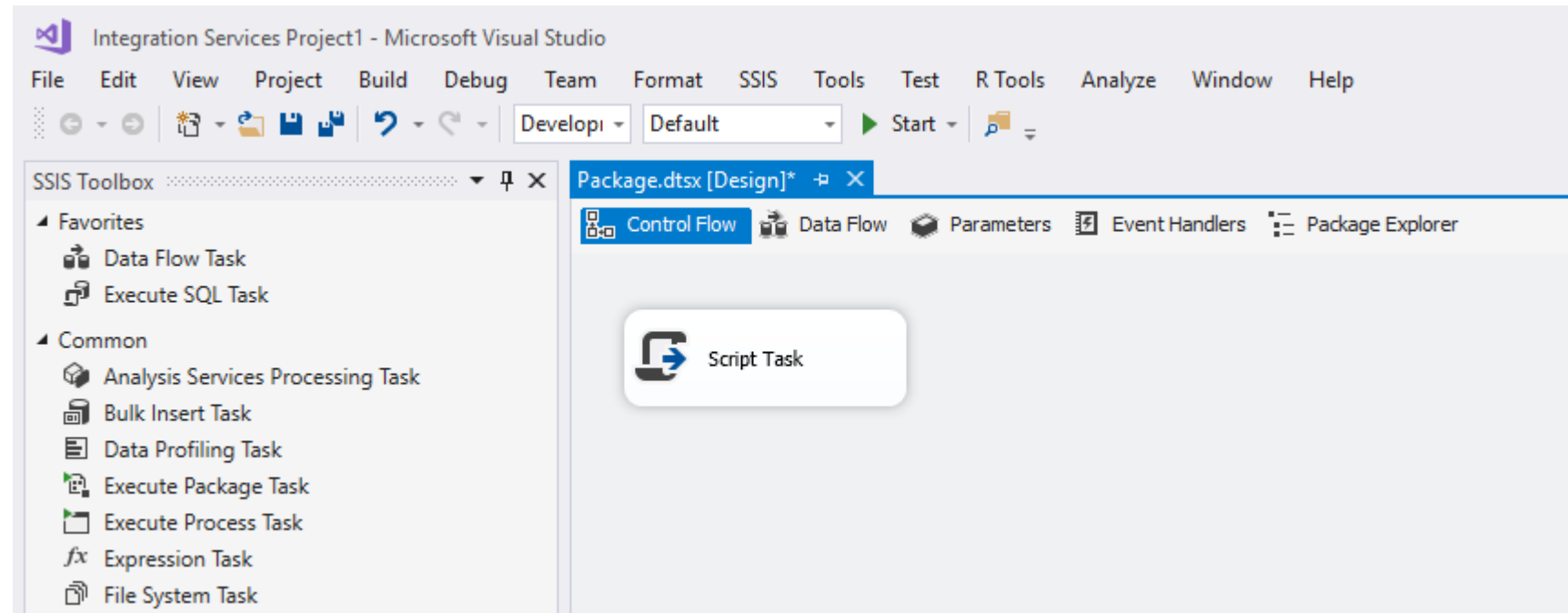
NAME	VALUE
url	<input type="text" value="Value"/>
body	<input type="text" value="Value"/>
Timeout	<input type="text" value="Value"/>
Retry	<input type="text" value="Value"/>

ADF Extensibility Continued

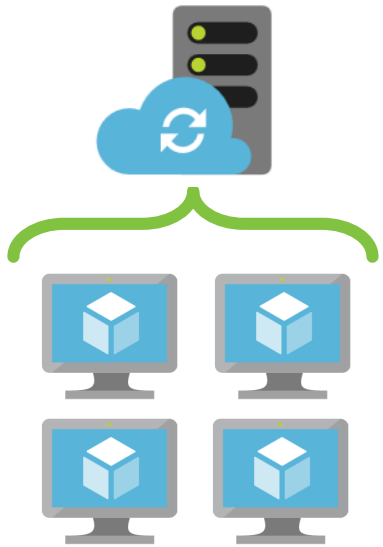
- 1 Custom Activities
- 2 Rest API Calls
- 3 **SSIS** – Packages with Control Flows and Data Flows



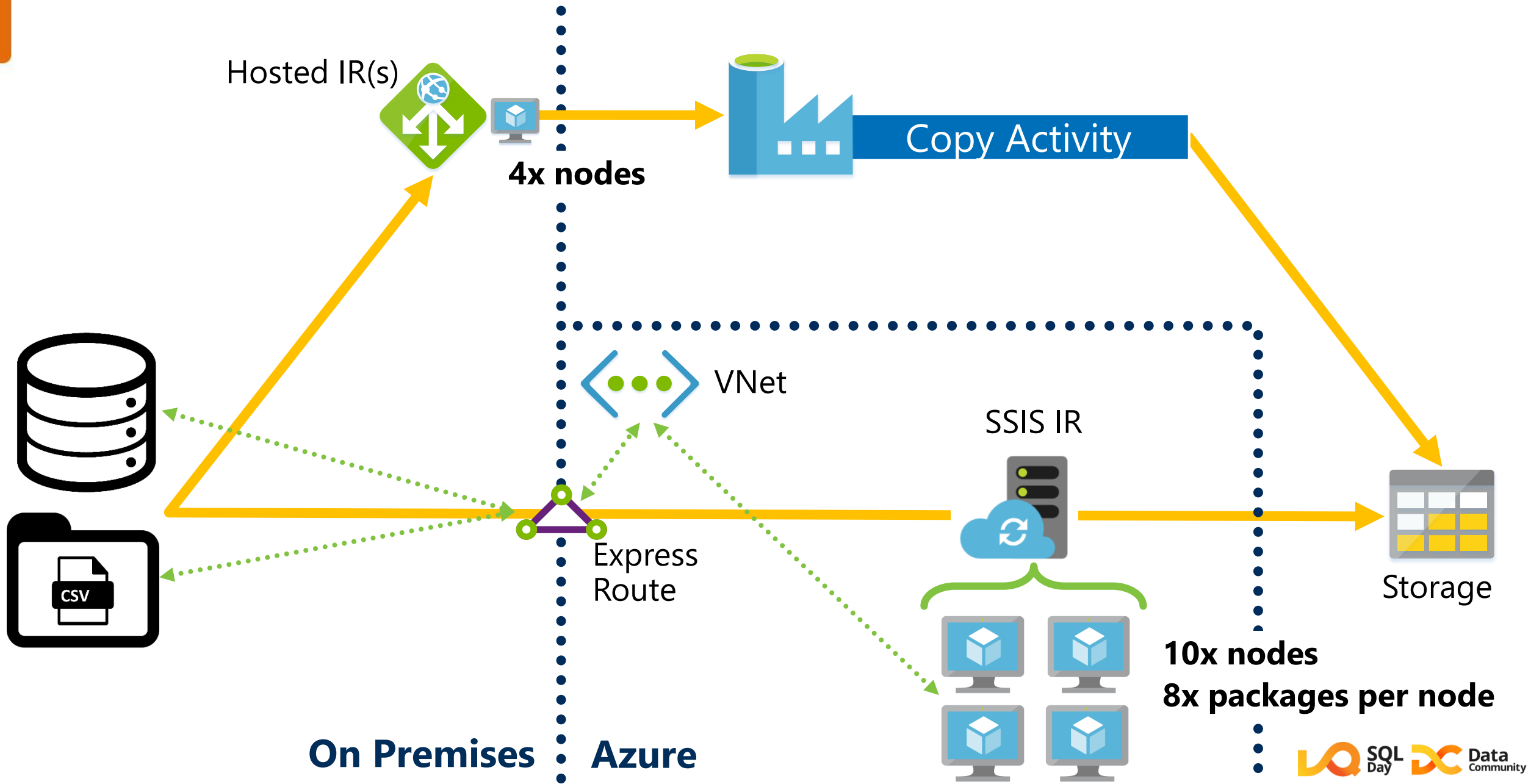
ADF SSIS IR



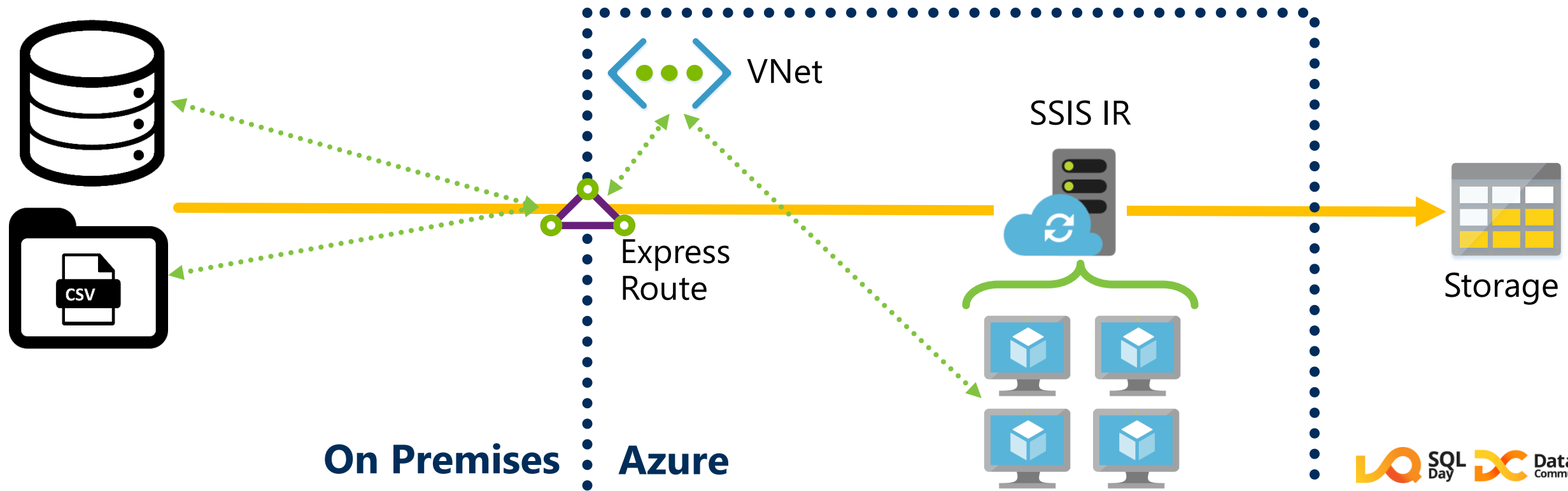
ADF Extensibility Continued



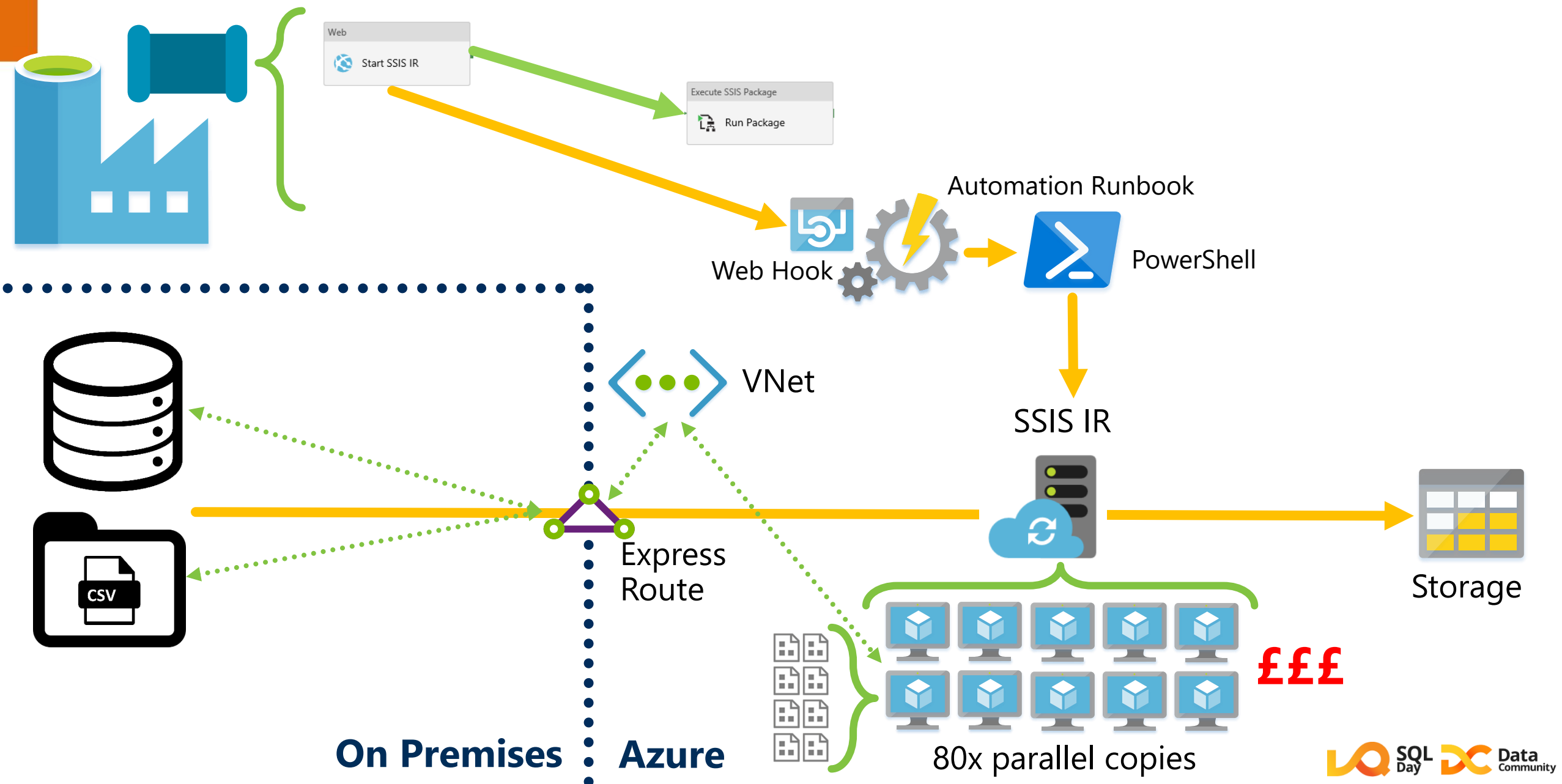
The SSIS IR vs Hosted IR with Express Route



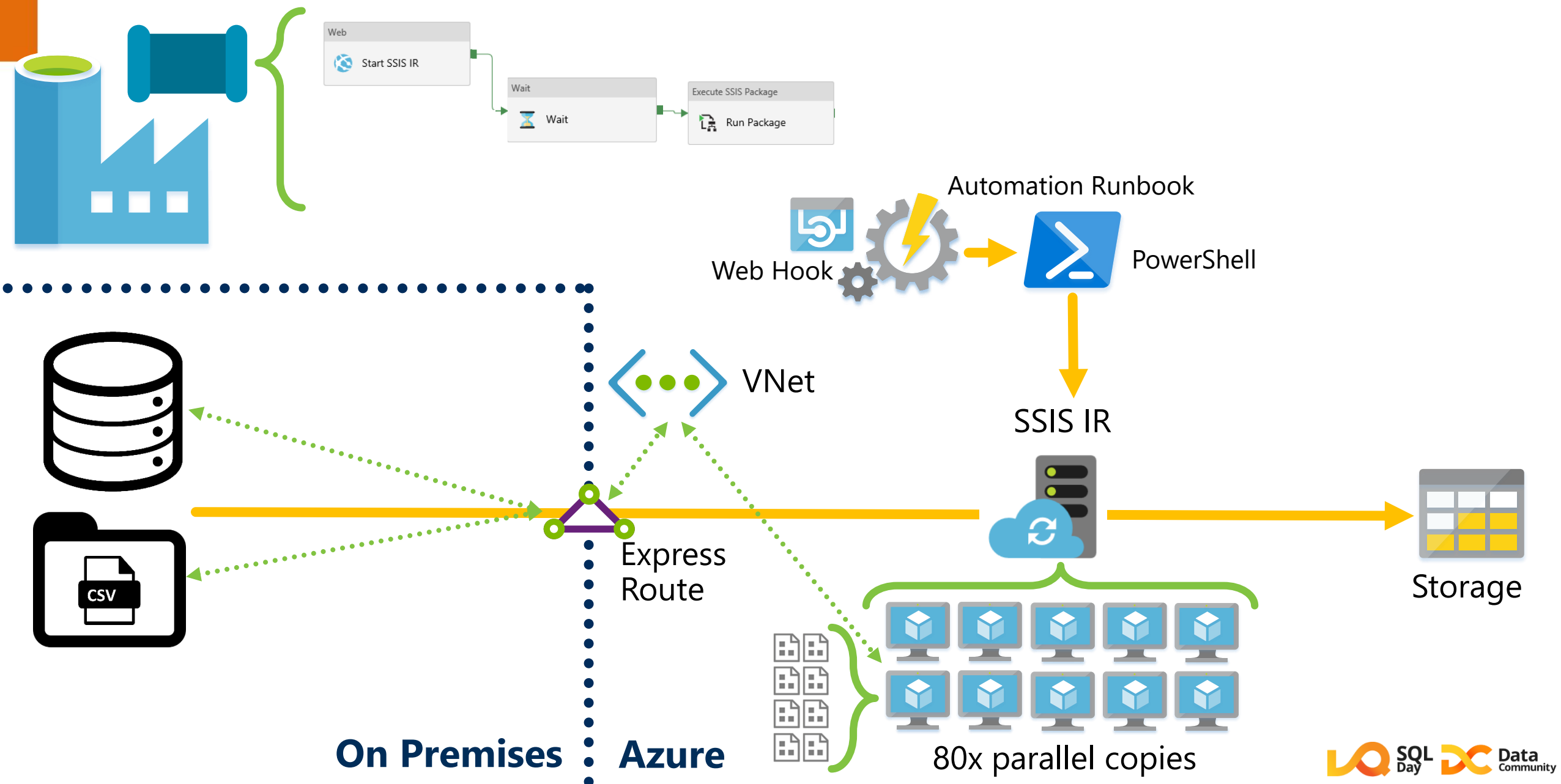
The SSIS IR Start/Stop



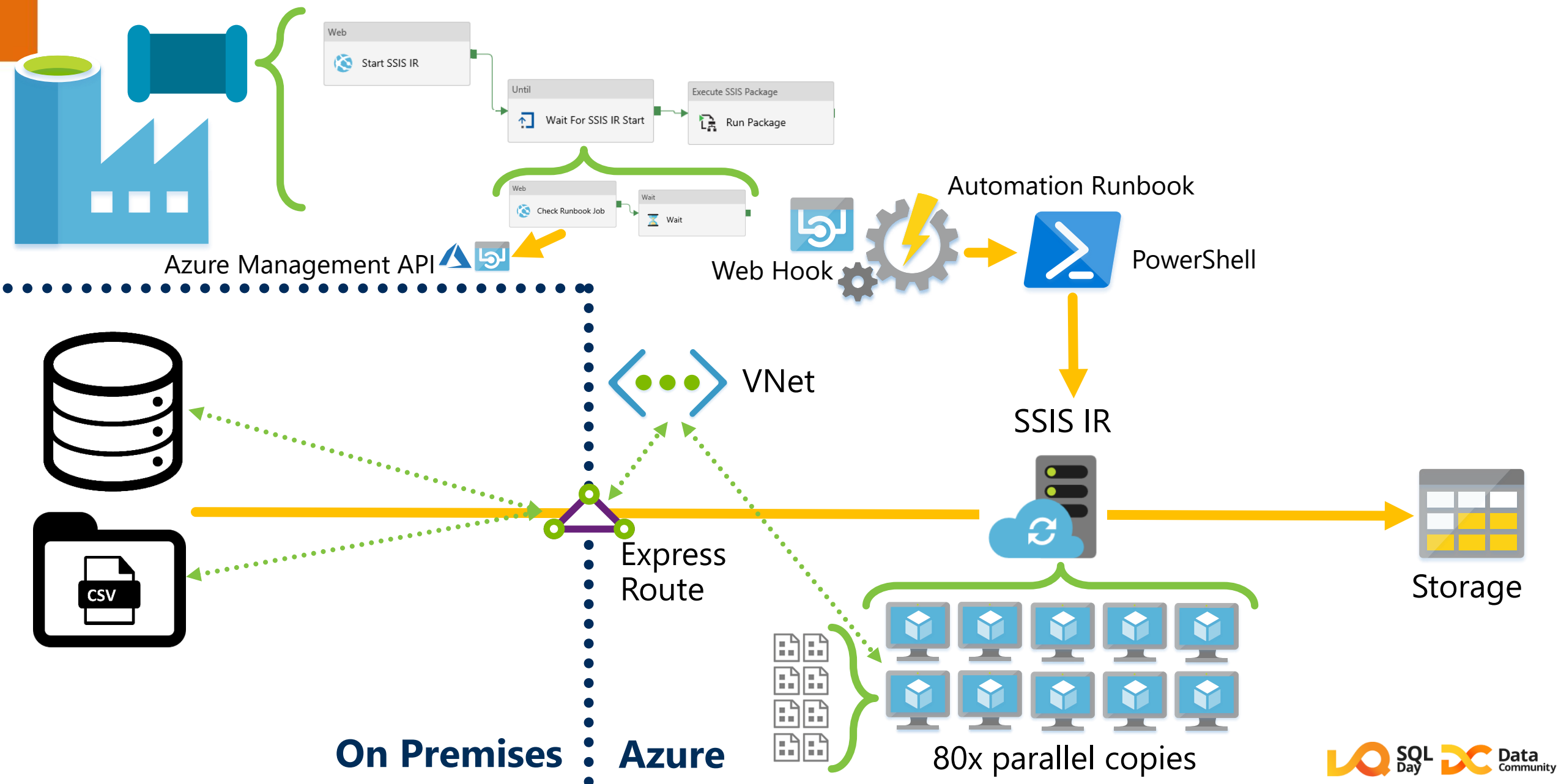
The SSIS IR Start/Stop



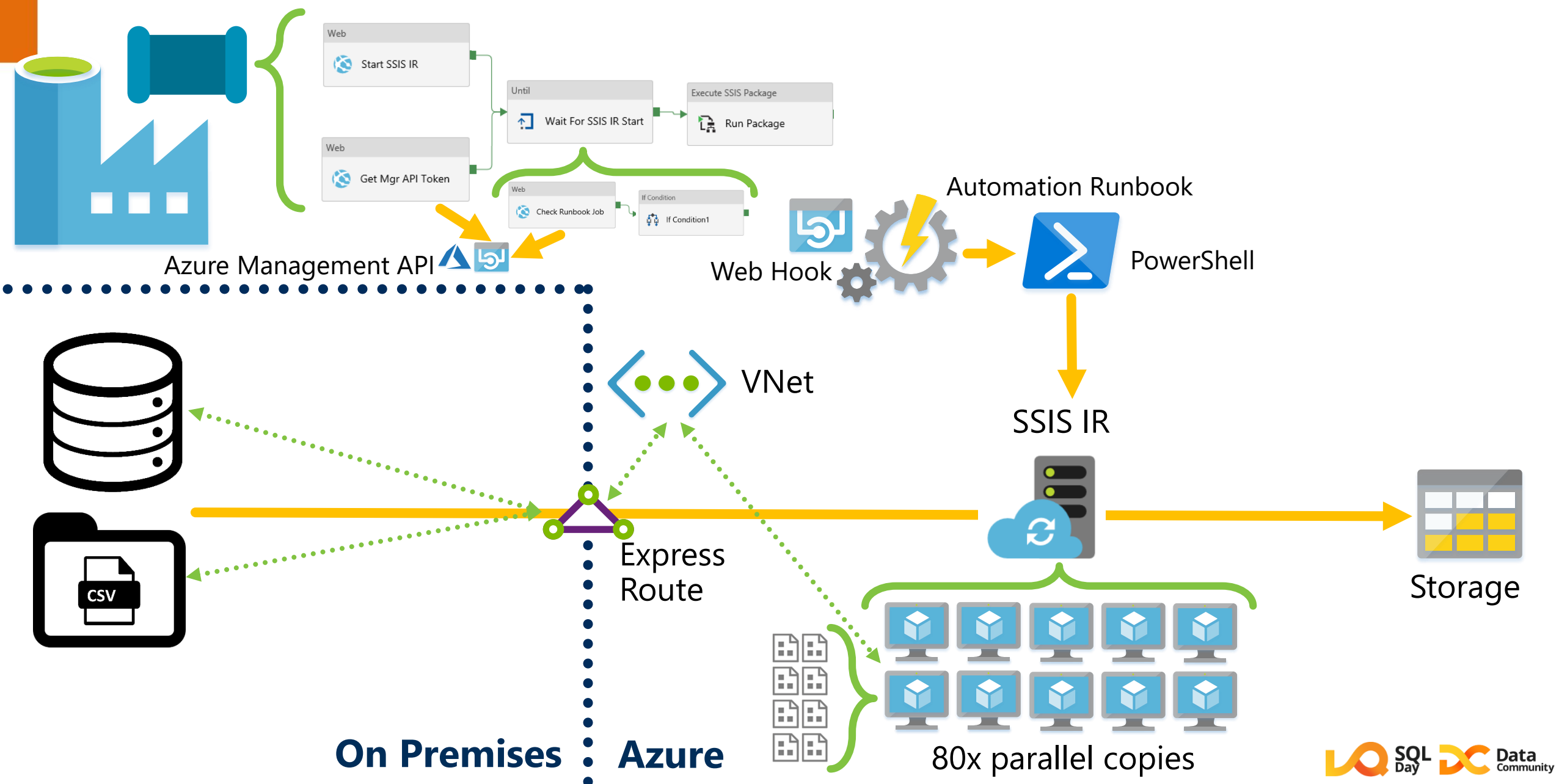
The SSIS IR Start/Stop



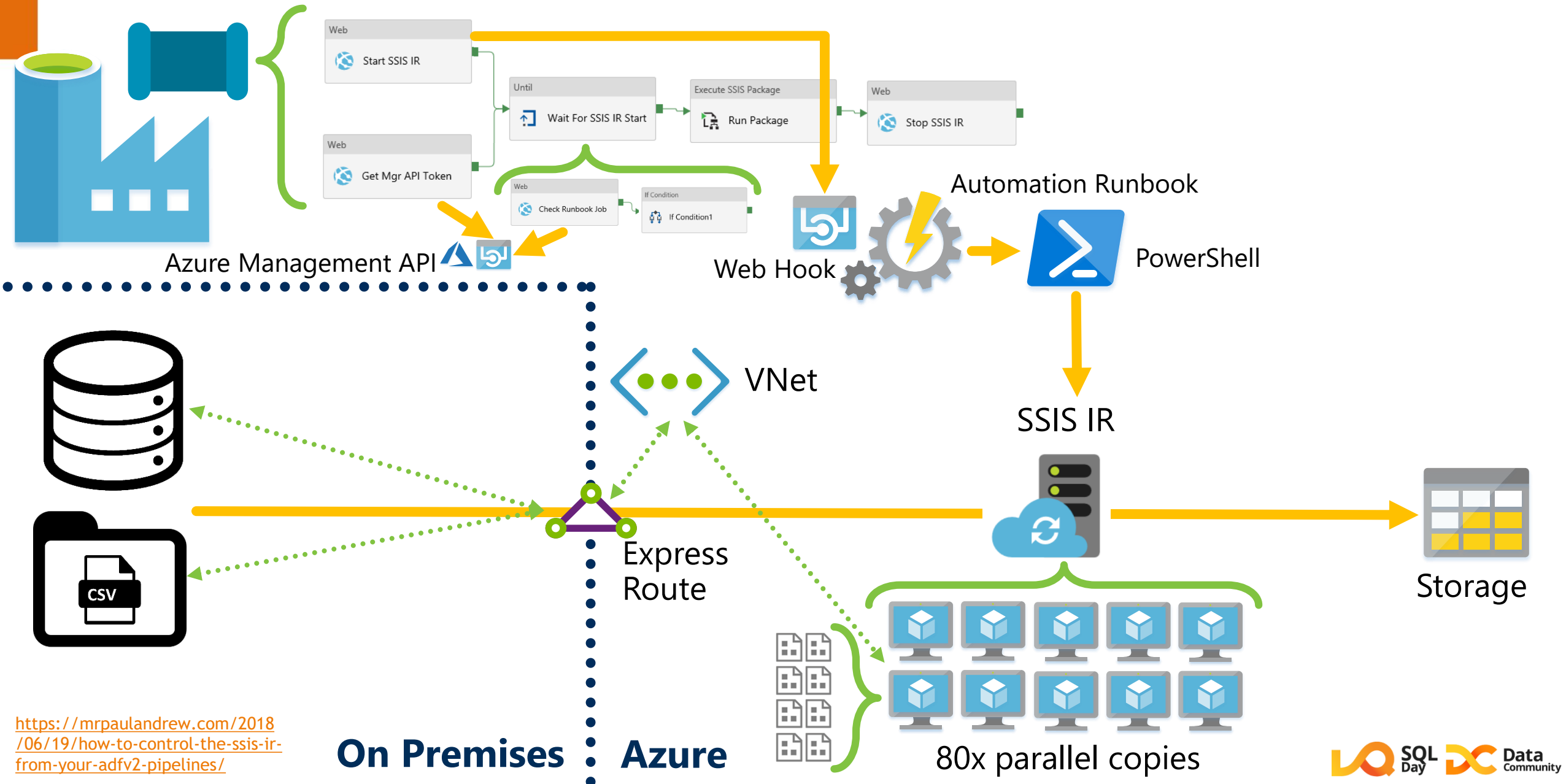
The SSIS IR Start/Stop



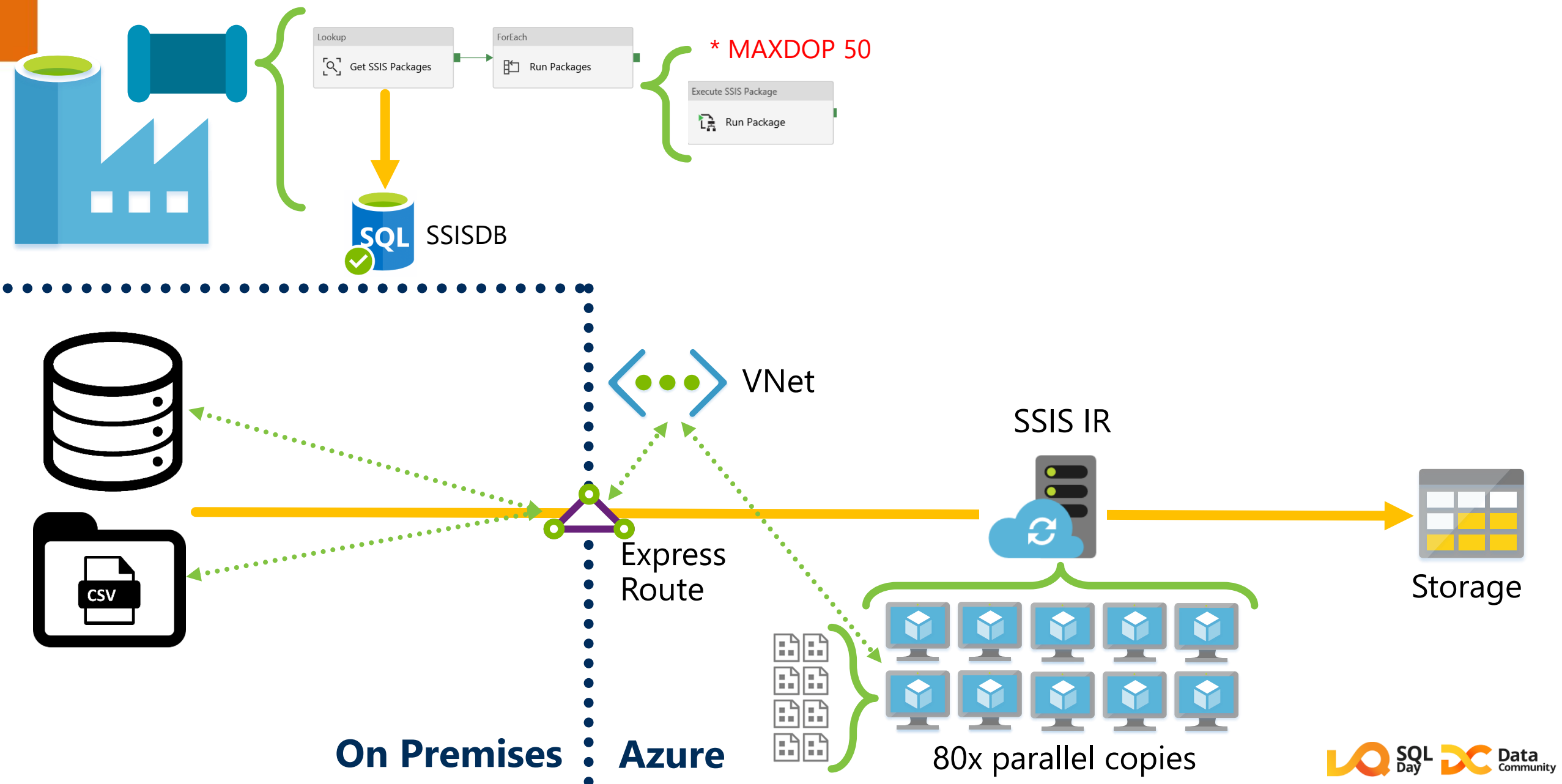
The SSIS IR Start/Stop



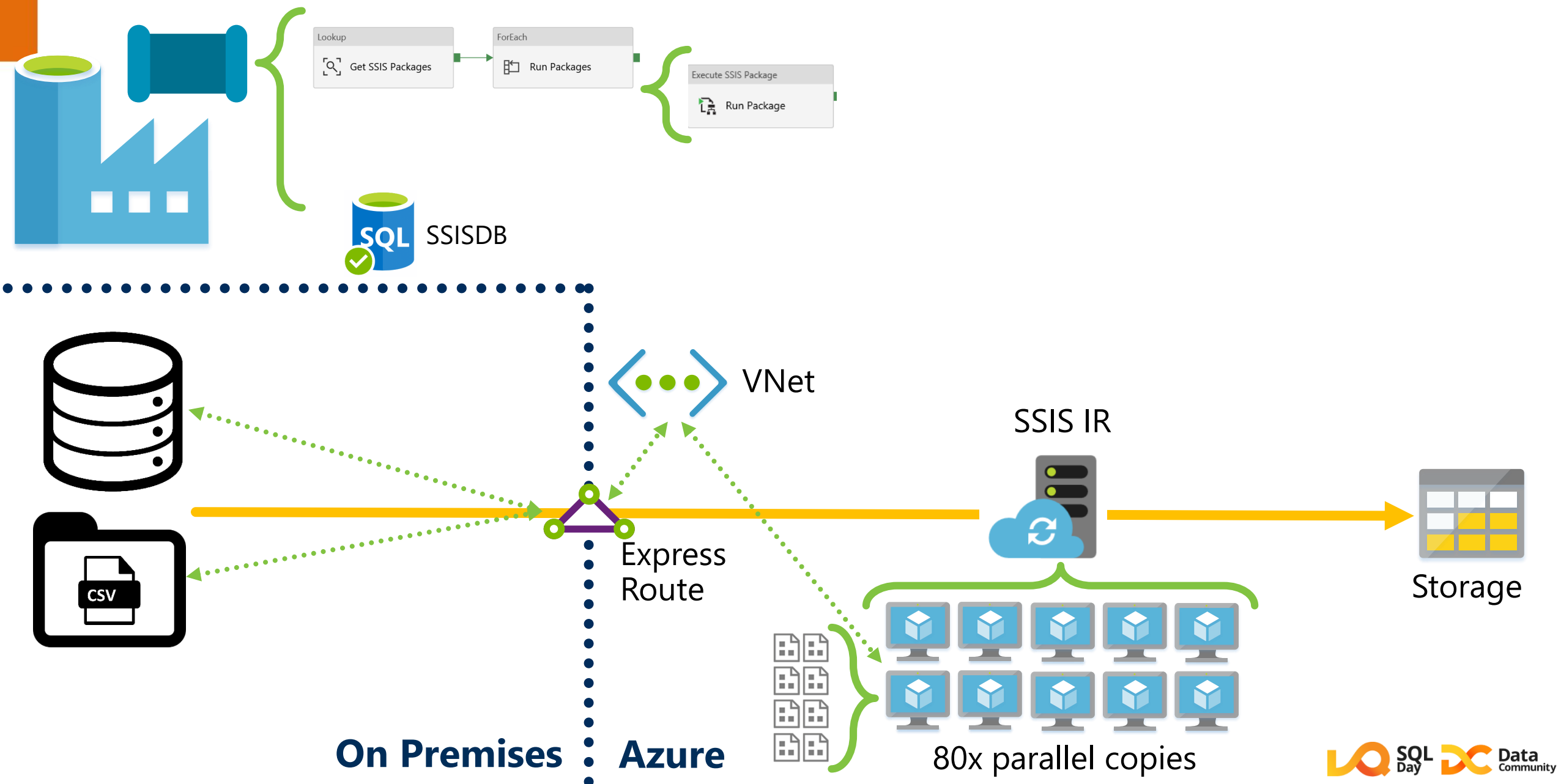
The SSIS IR Start/Stop



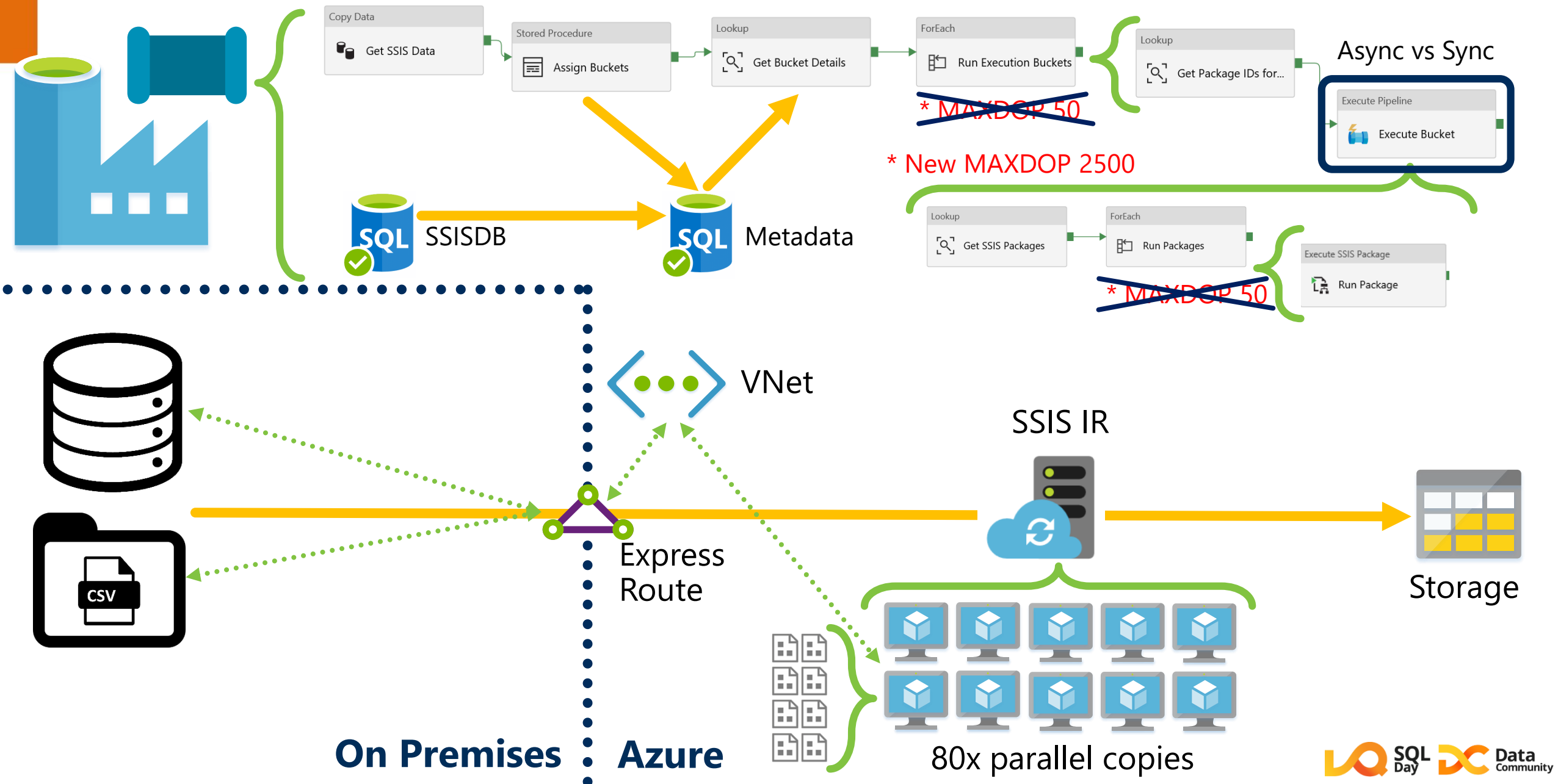
The SSIS IR Parallelism



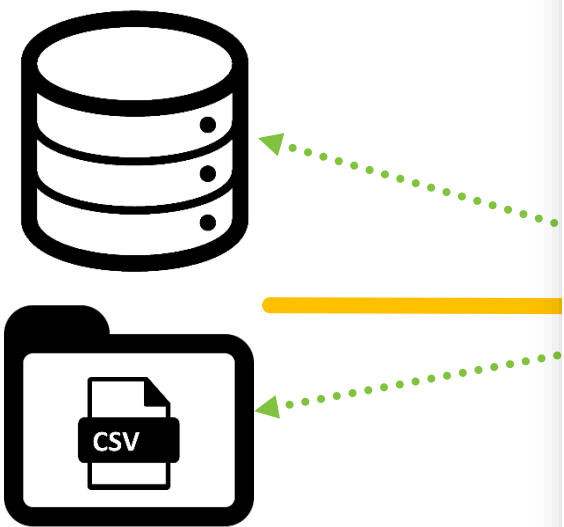
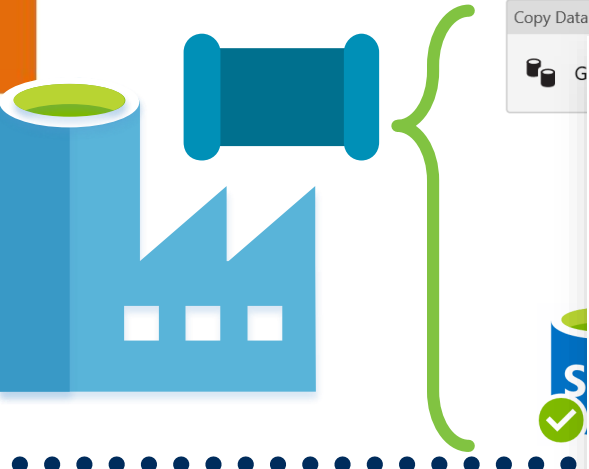
The SSIS IR Parallelism



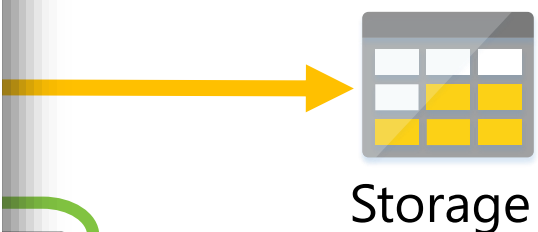
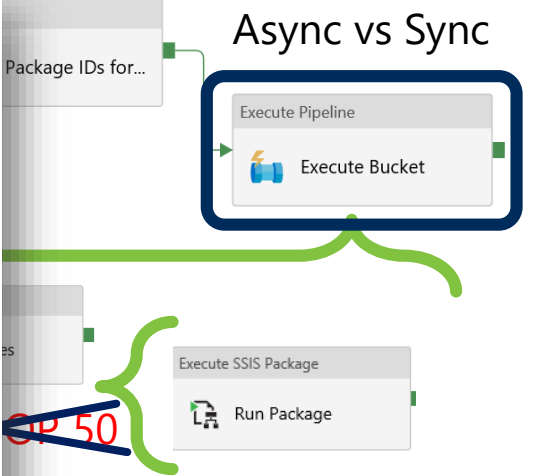
The SSIS IR Parallelism



The SSIS IR Parallelism



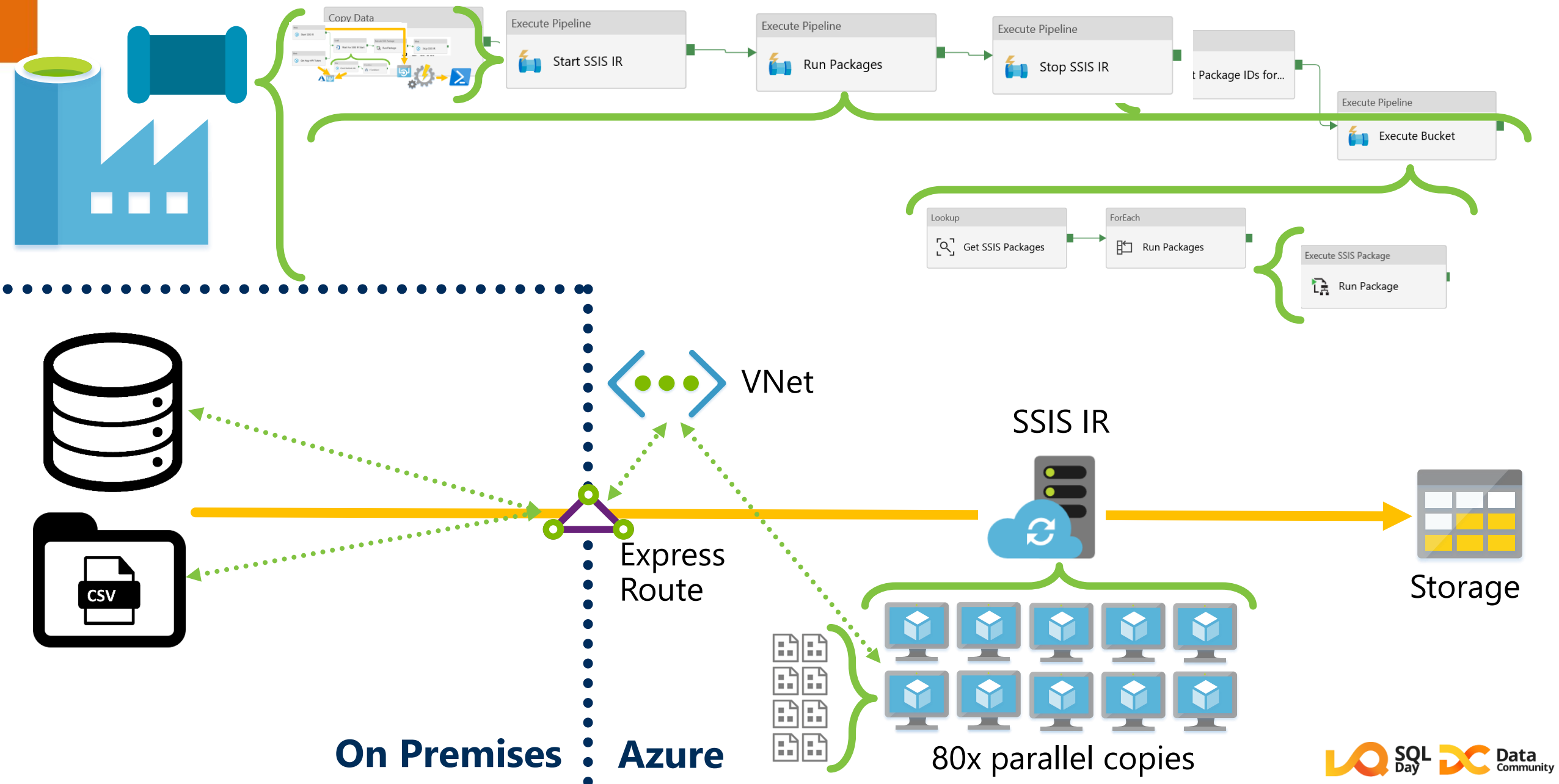
Resource	Default Limit	Maximum Limit
Data factories in an Azure subscription	50	Contact support
Total number of entities (Pipeline, Datasets, Triggers, Linked Services, Integration runtimes) within a data factory	5000	Contact support
Total CPU cores for Azure-SSIS Integration Runtime(s) under one subscription	256	Contact support
Concurrent pipeline runs per data factory (shared among all pipelines in the factory)	10,000	Contact support
Max activities per pipeline (includes inner activities for containers)	40	40
Max number of Linked Integration Runtime that can be created against a single Self-hosted Integration Runtime	20	Contact support
Max parameters per pipeline	50	50
ForEach items	100,000	100,000
ForEach parallelism	20	50
Characters per expression	8,192	8,192
Minimum Tumbling Window Trigger interval	15 min	15 min
Max Timeout for pipeline activity runs	7 days	7 days
Bytes per object for pipeline objects ¹	200 KB	200 KB
Bytes per object for dataset and linked service objects ¹	100 KB	2000 KB
Data integration units per copy activity run ³	256	Contact support



<https://github.com/MicrosoftDocs/azure-docs/blob/master/includes/azure-data-factory-limits.md>

On Premises : Azure 80x parallel copies

SSIS IR & Package Complete Orchestration Solution



Pattern Summary

Execute Pipeline



Grandparent

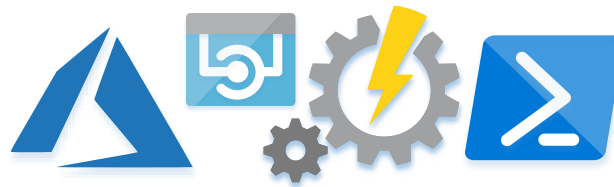


High Level Control Flow and Pipeline Triggers

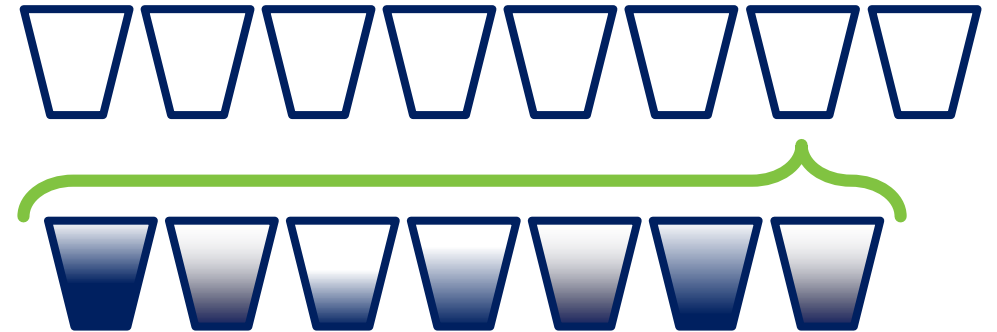
Execute Pipeline



Parent



Platform Component Control



Manage Parallel Streams

Execute Pipeline



Child



Service Level Executions

Complex Orchestration

With Dynamic Data Factory Pipelines



Azure Data
Factory

A very quick
overview

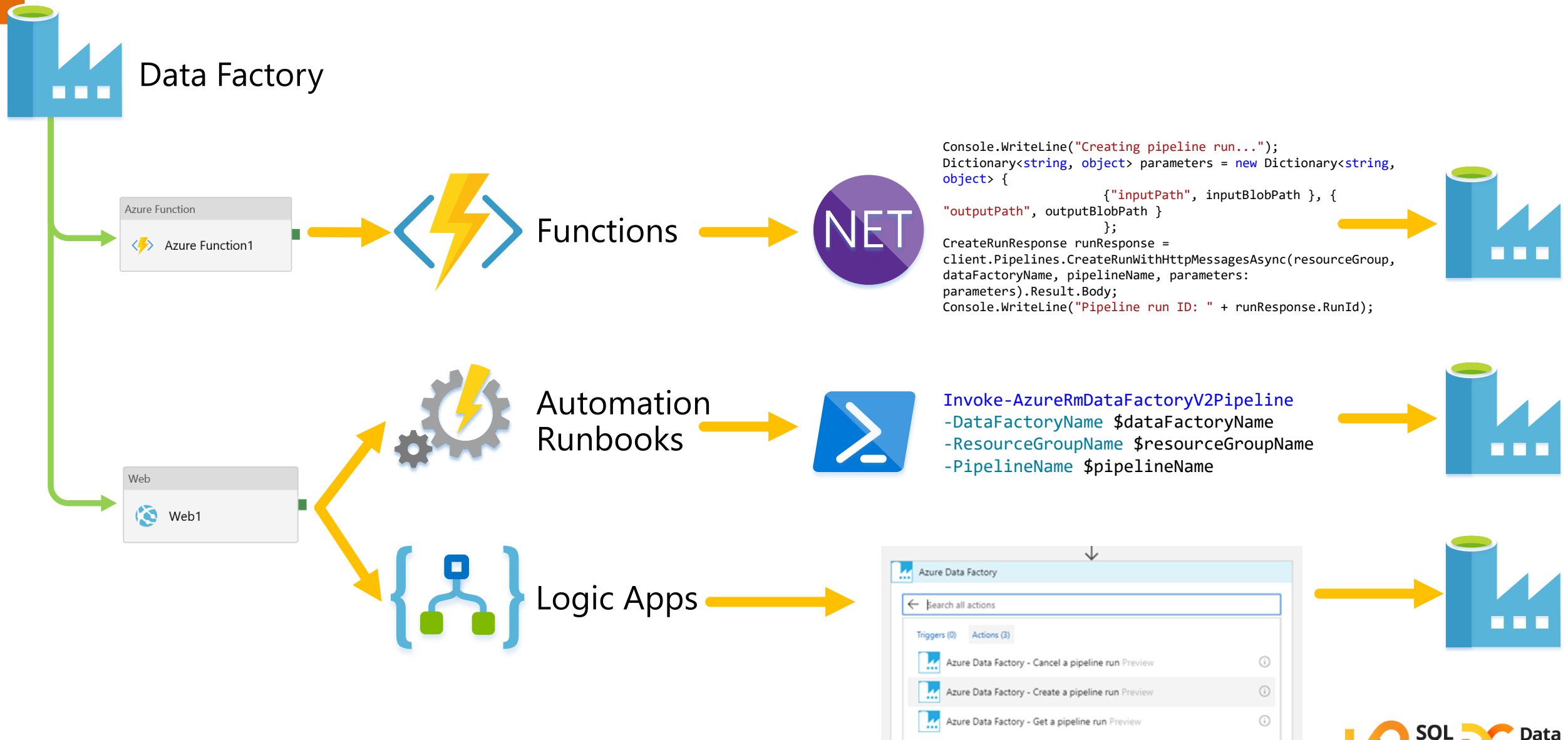
Extensibility &
Parallelism

Custom Activities
SSIS IR & Packages

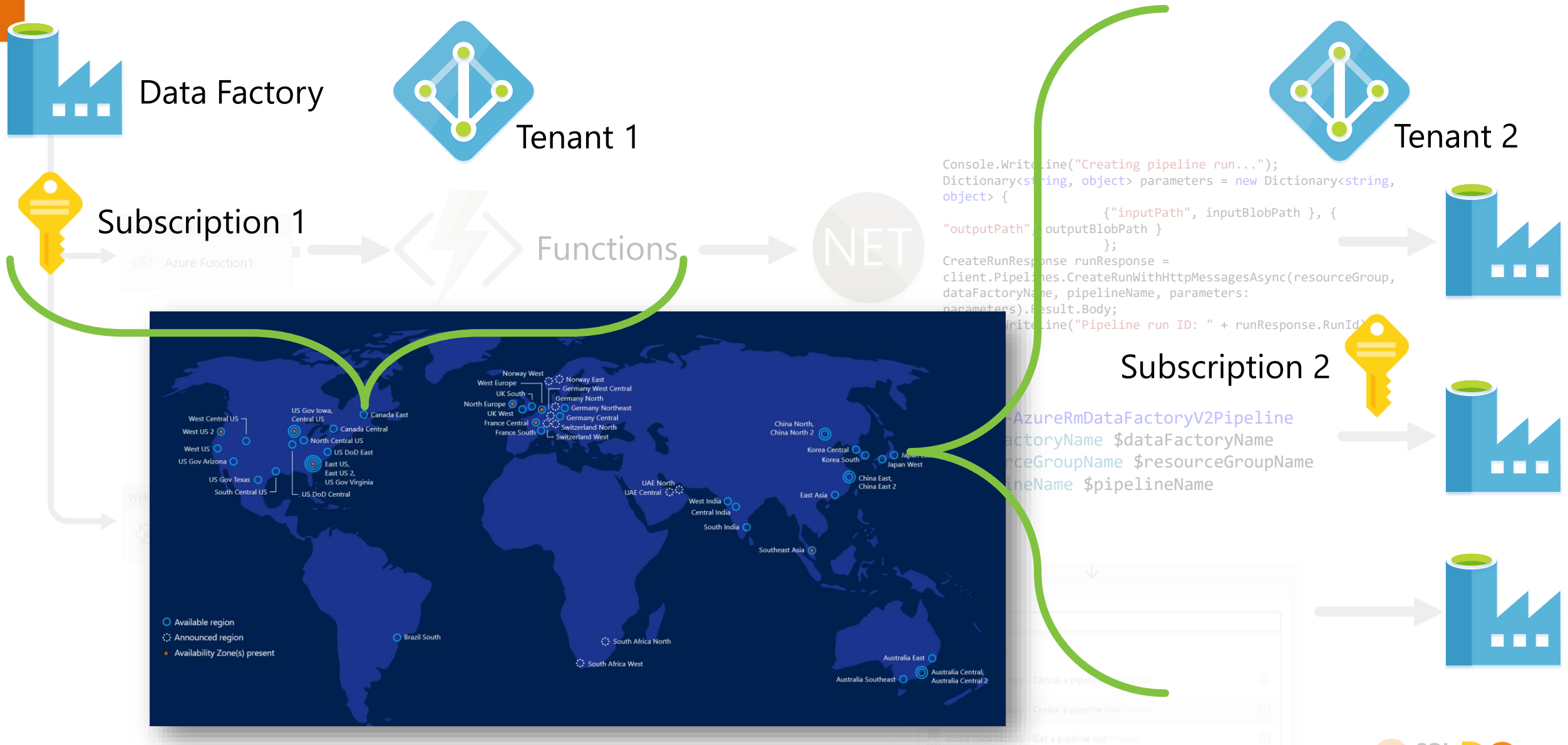
More Design
Patterns

Bootstrapping
Hosted IR vs IaaS
Frameworks

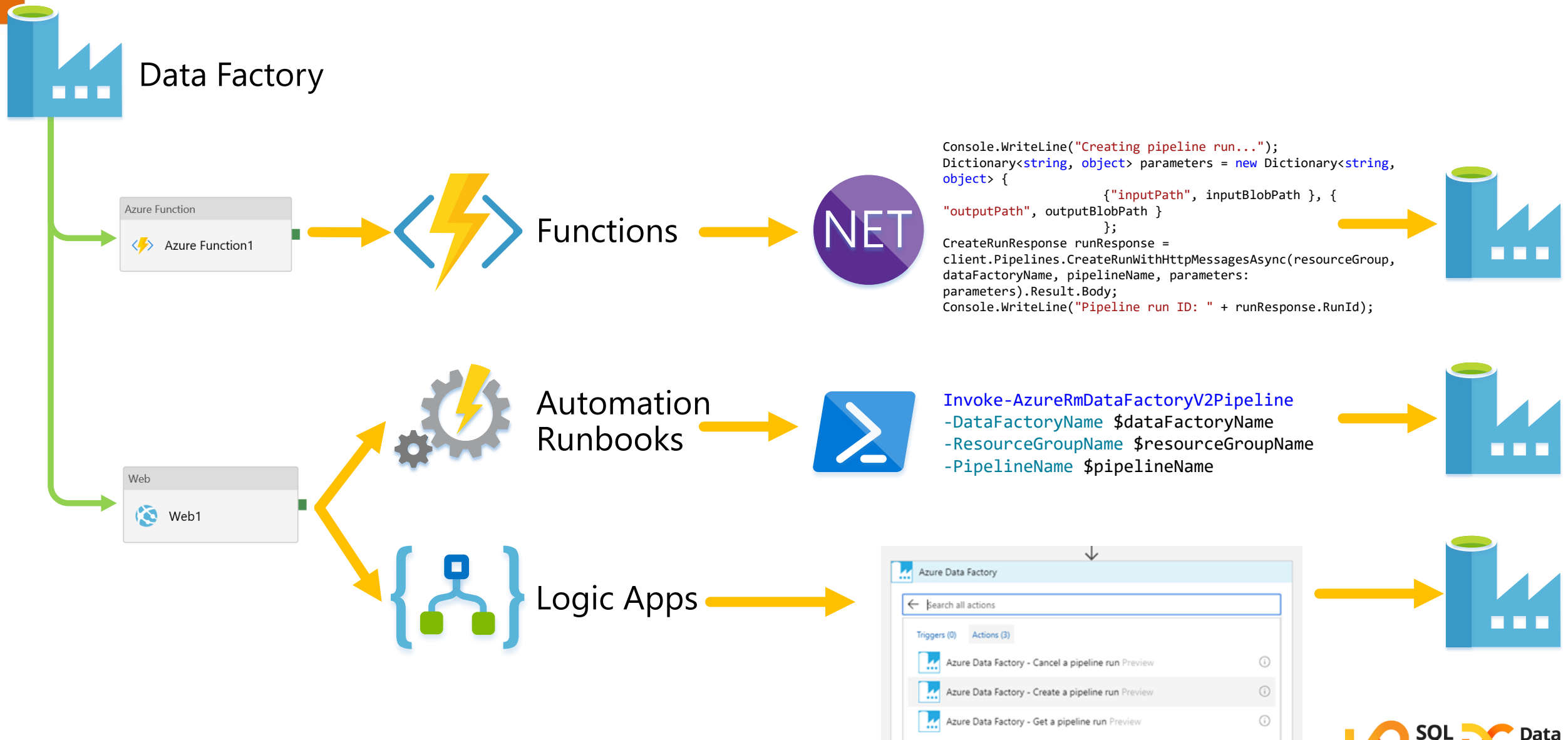
Bootstrapping



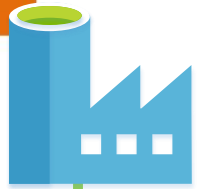
Bootstrapping – Why?



Bootstrapping – Why?



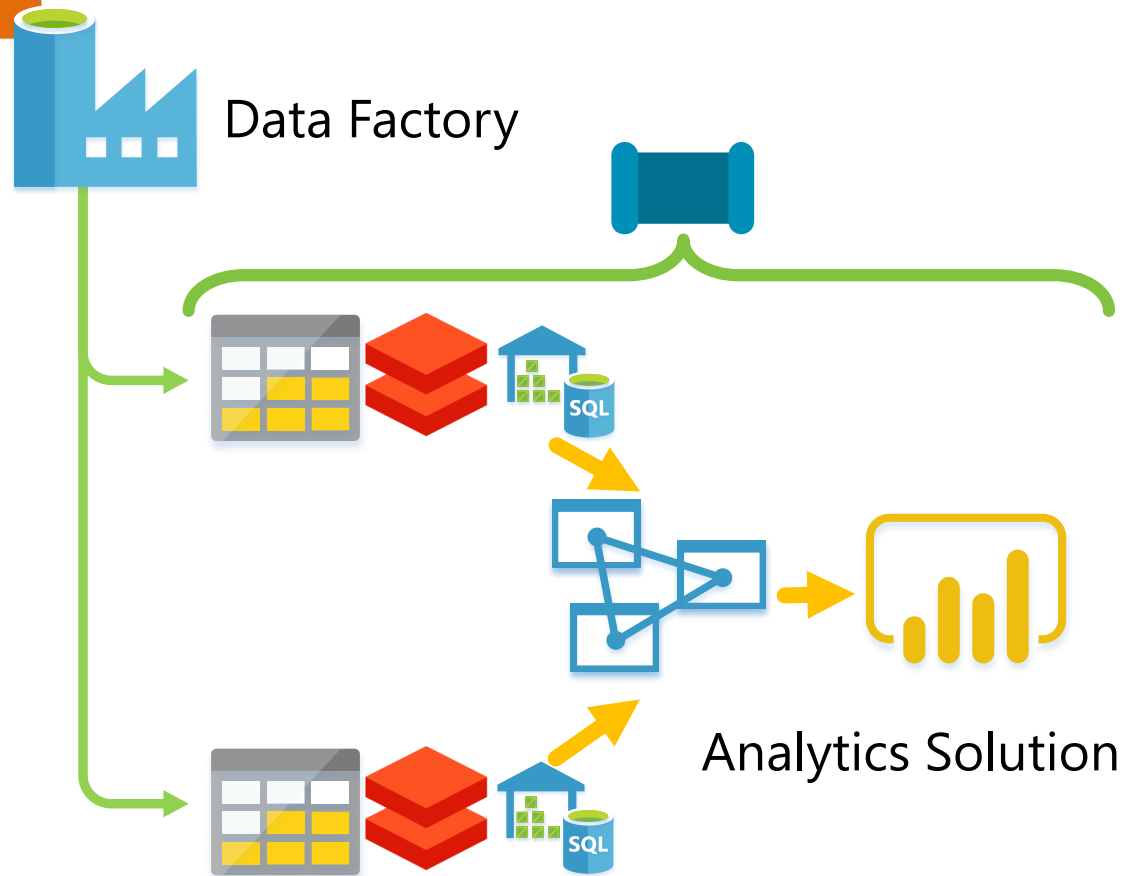
Bootstrapping – Why?



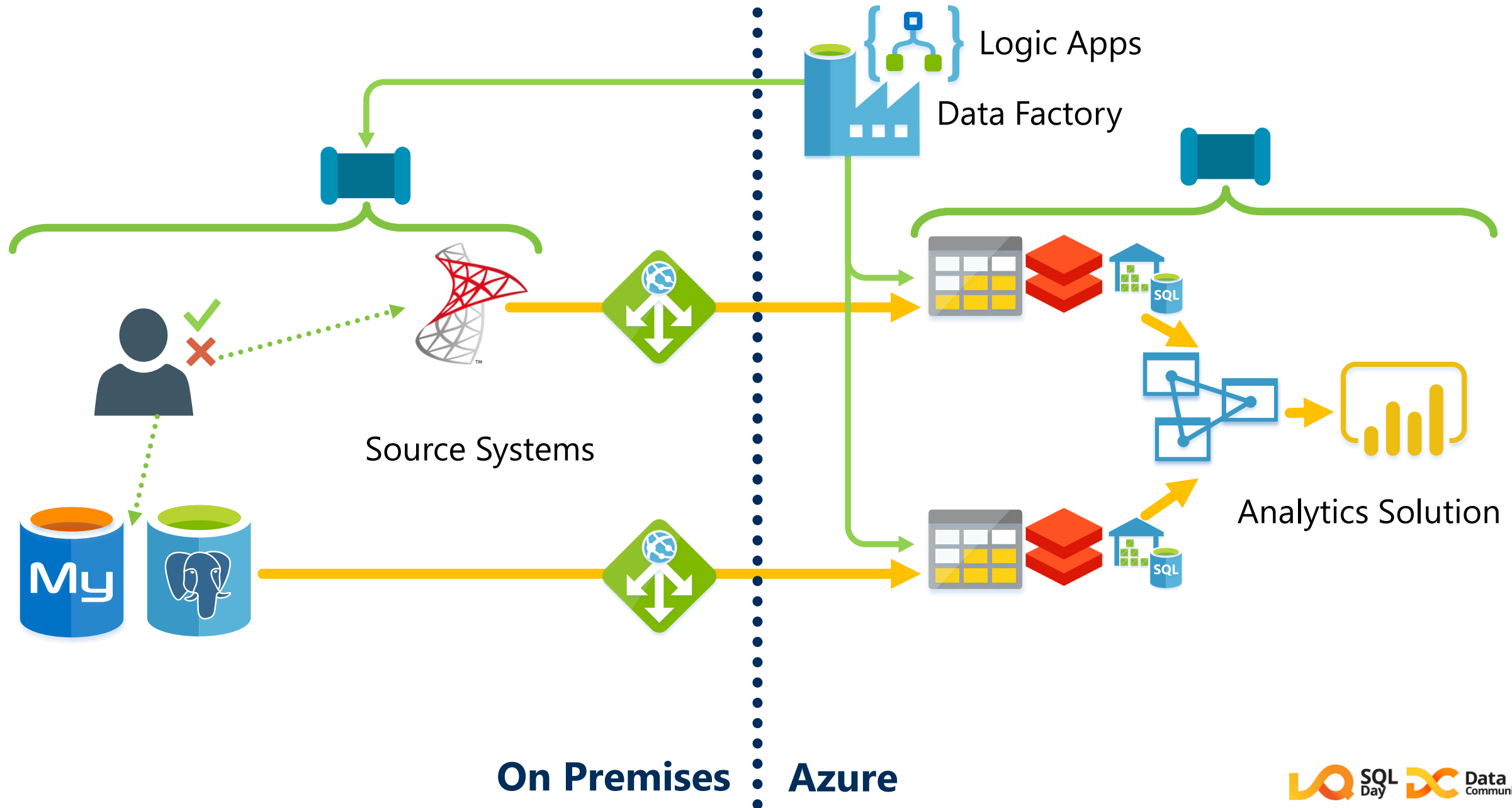
Data Factory



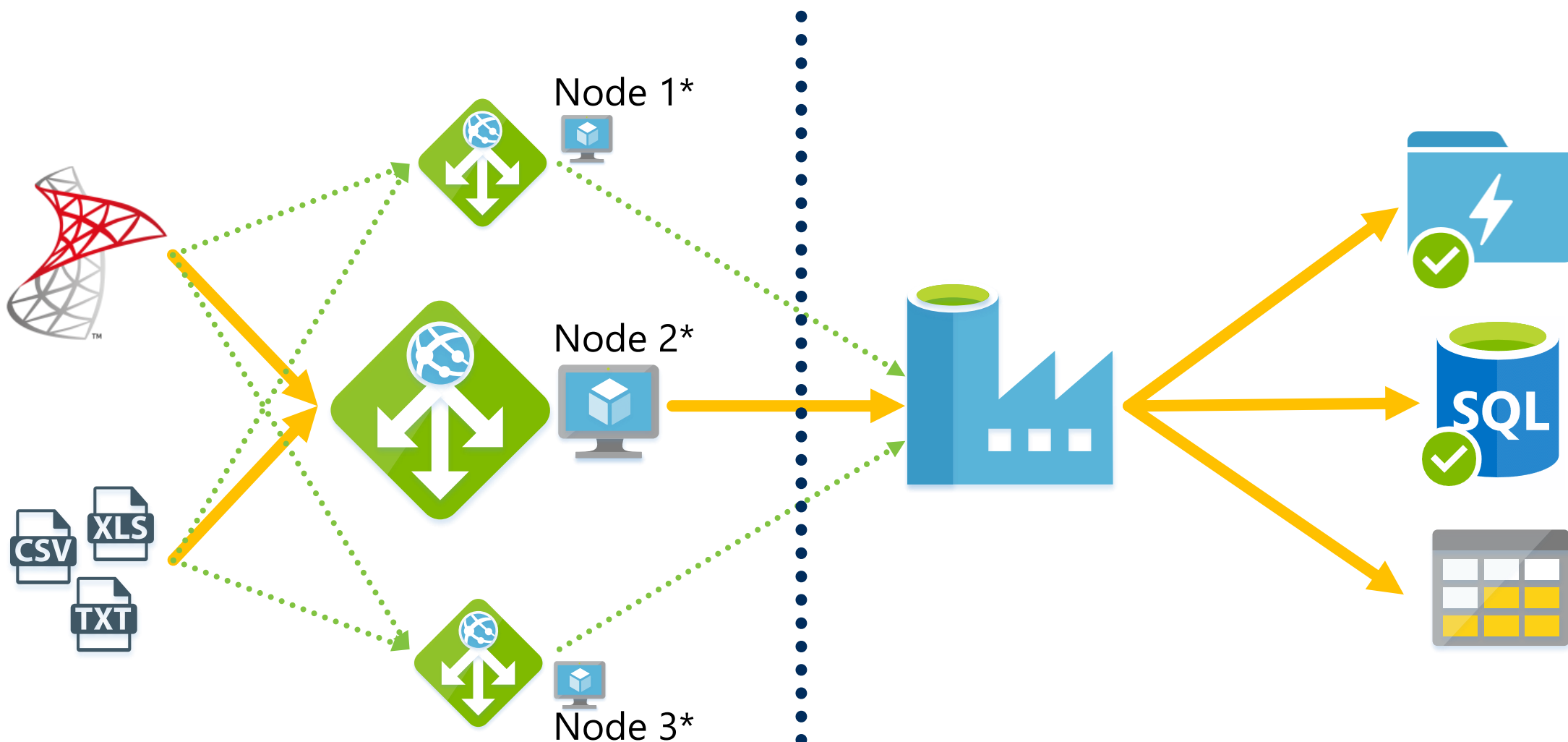
Bootstrapping – Why?



Bootstrapping vs Data Ingestion



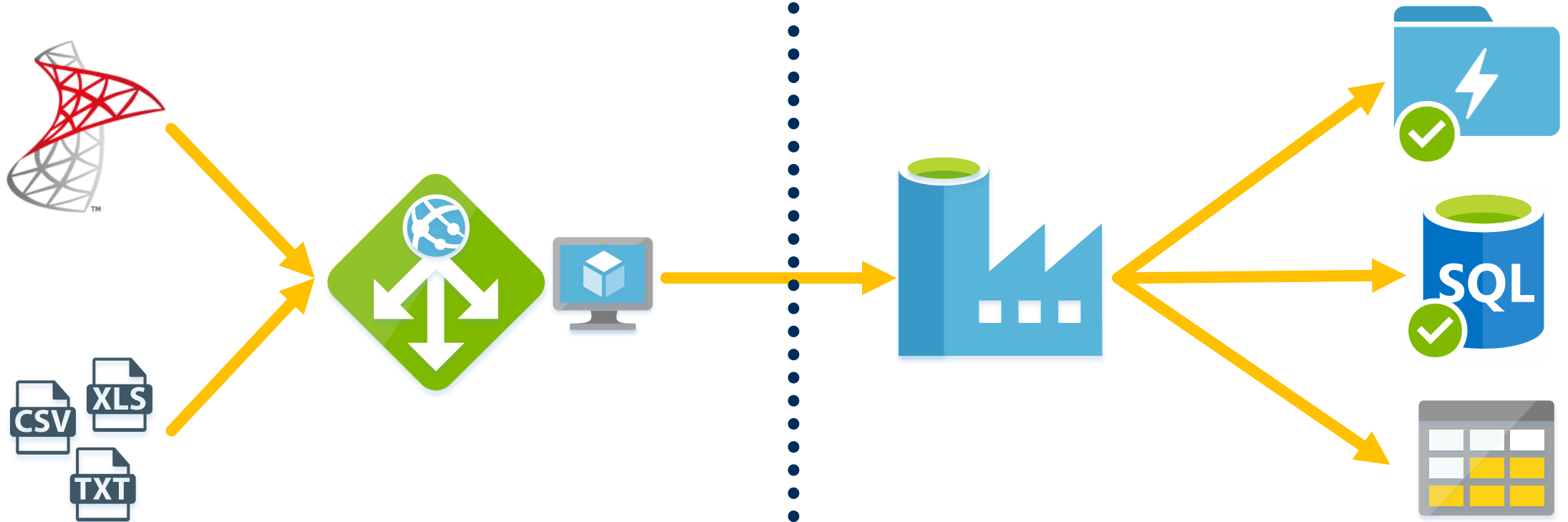
The Hosted Integration Runtime



*Failover & Load Balancing

On Premises • **Azure**

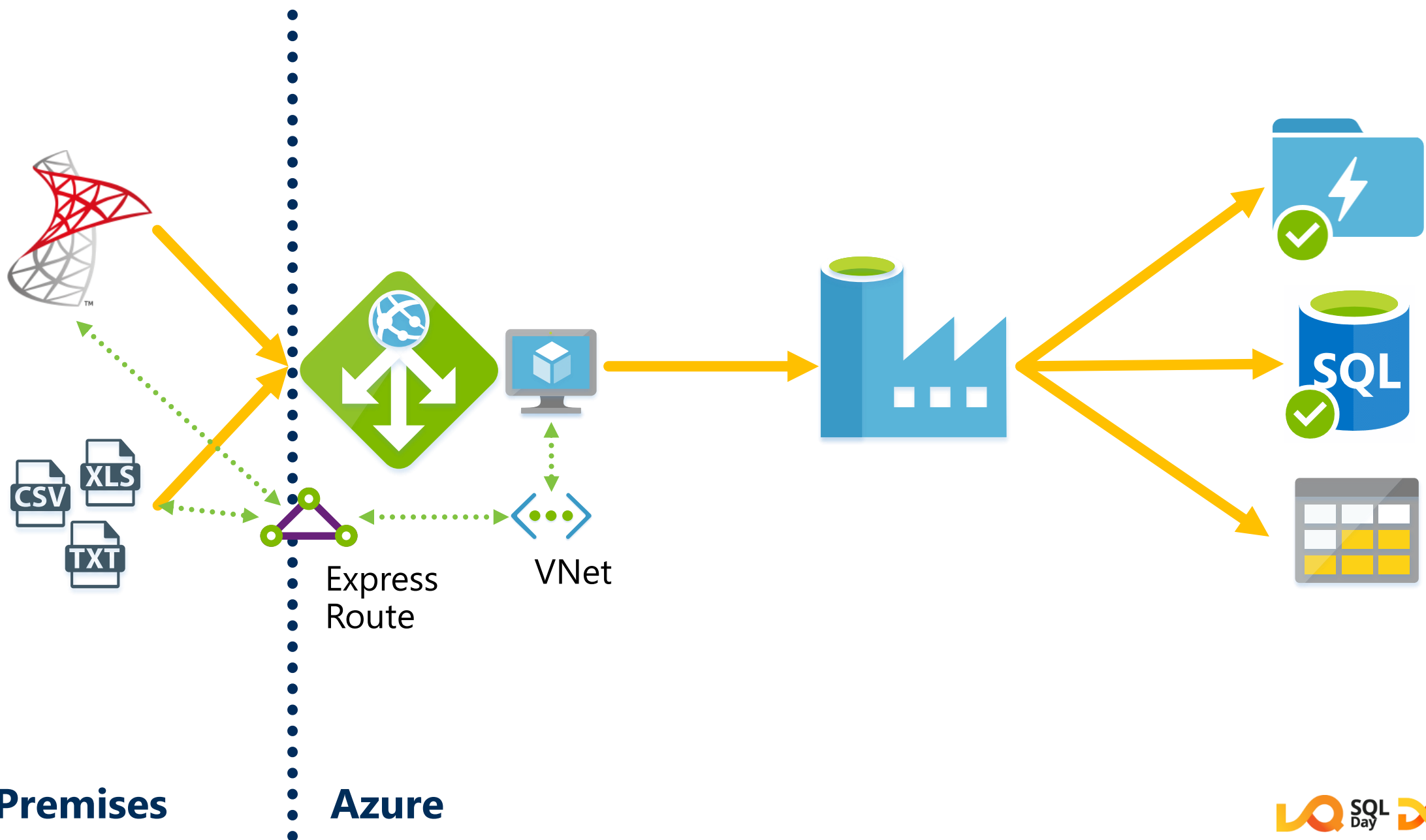
The Hosted Integration Runtime



On Premises : **Azure**

*Failover & Load Balancing

The Hosted Integration Runtime with Express Route



On Premises

Azure

Complex Orchestration

With Dynamic Data Factory Pipelines



Azure Data
Factory

A very quick
overview

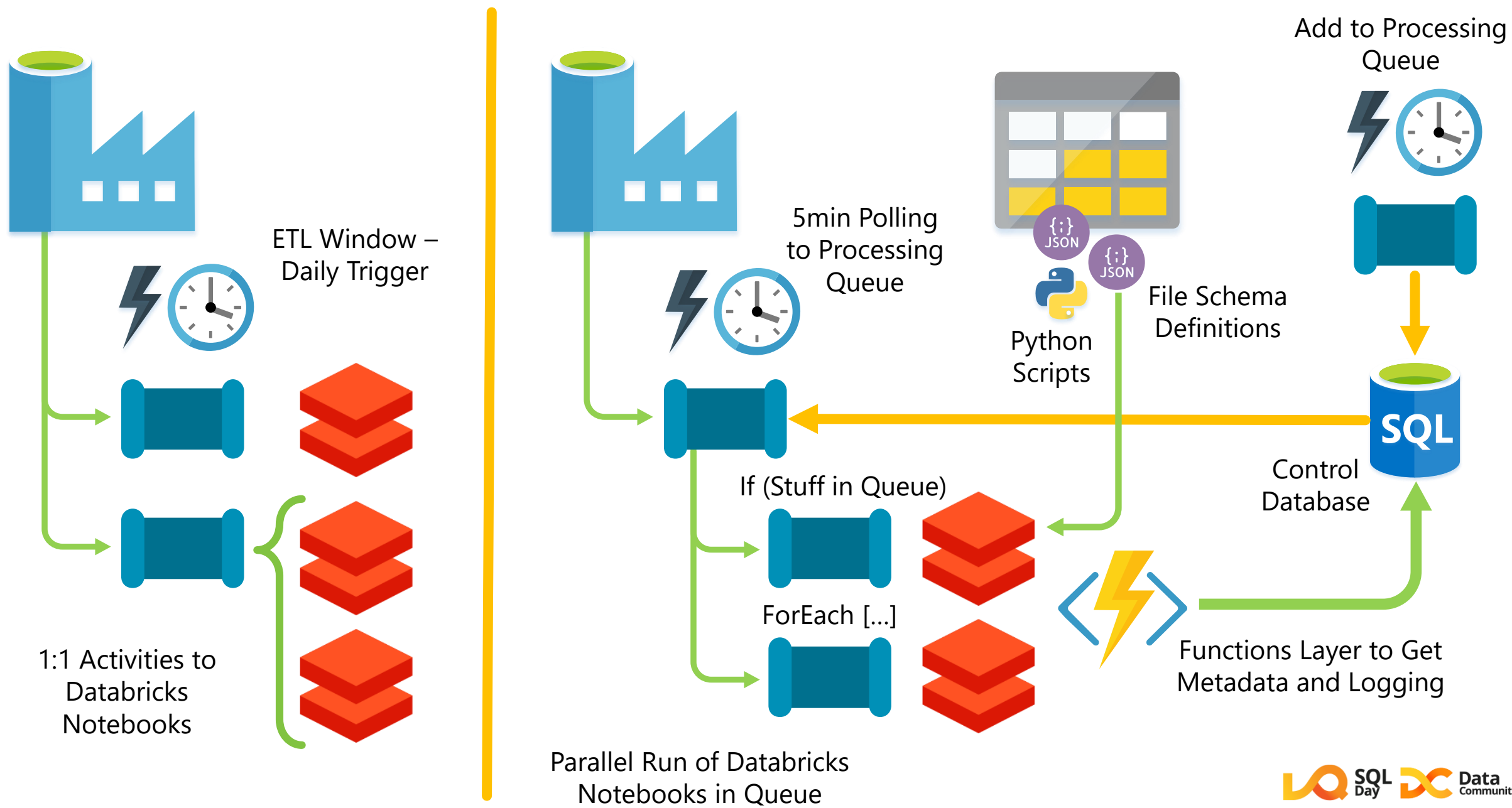
Extensibility &
Parallelism

Custom Activities
SSIS IR & Packages

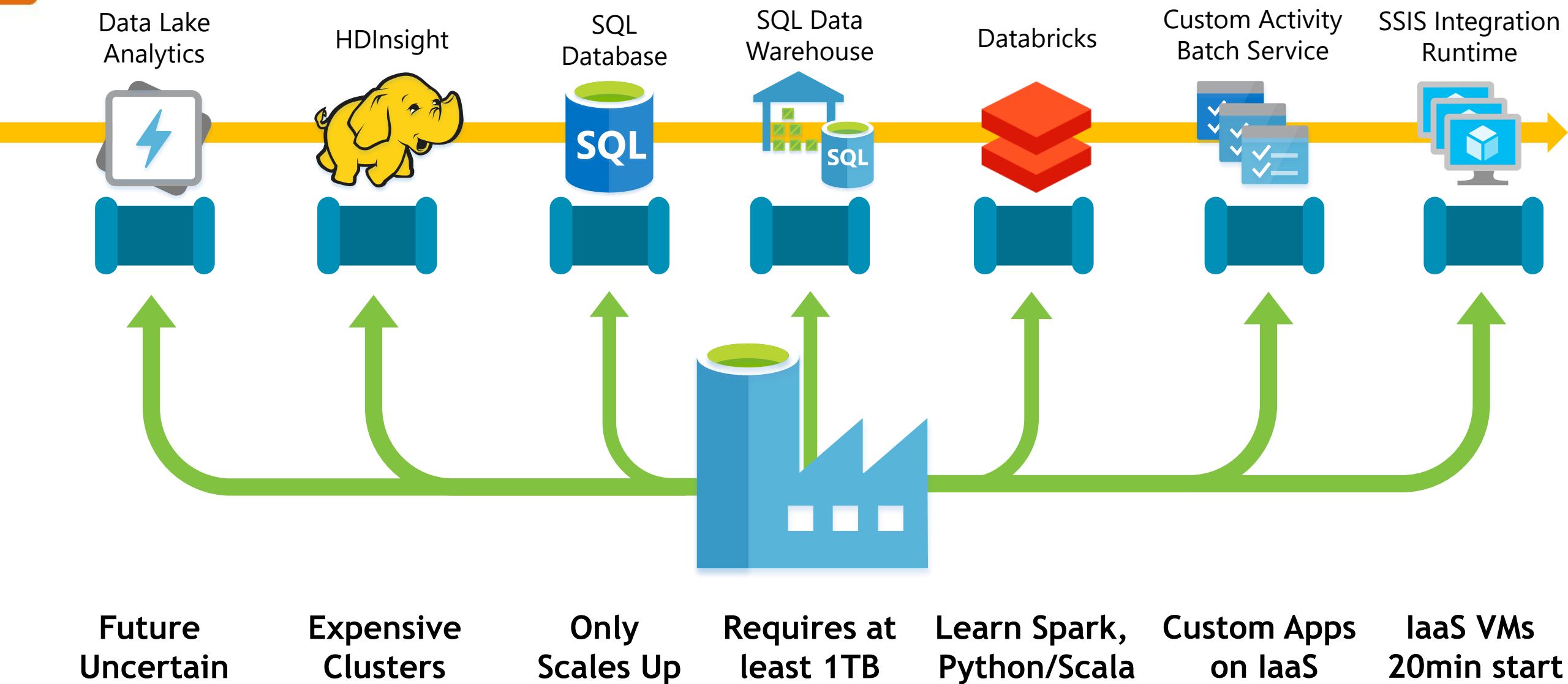
More Design
Patterns

Bootstrapping
Hosted IR vs IaaS
Frameworks

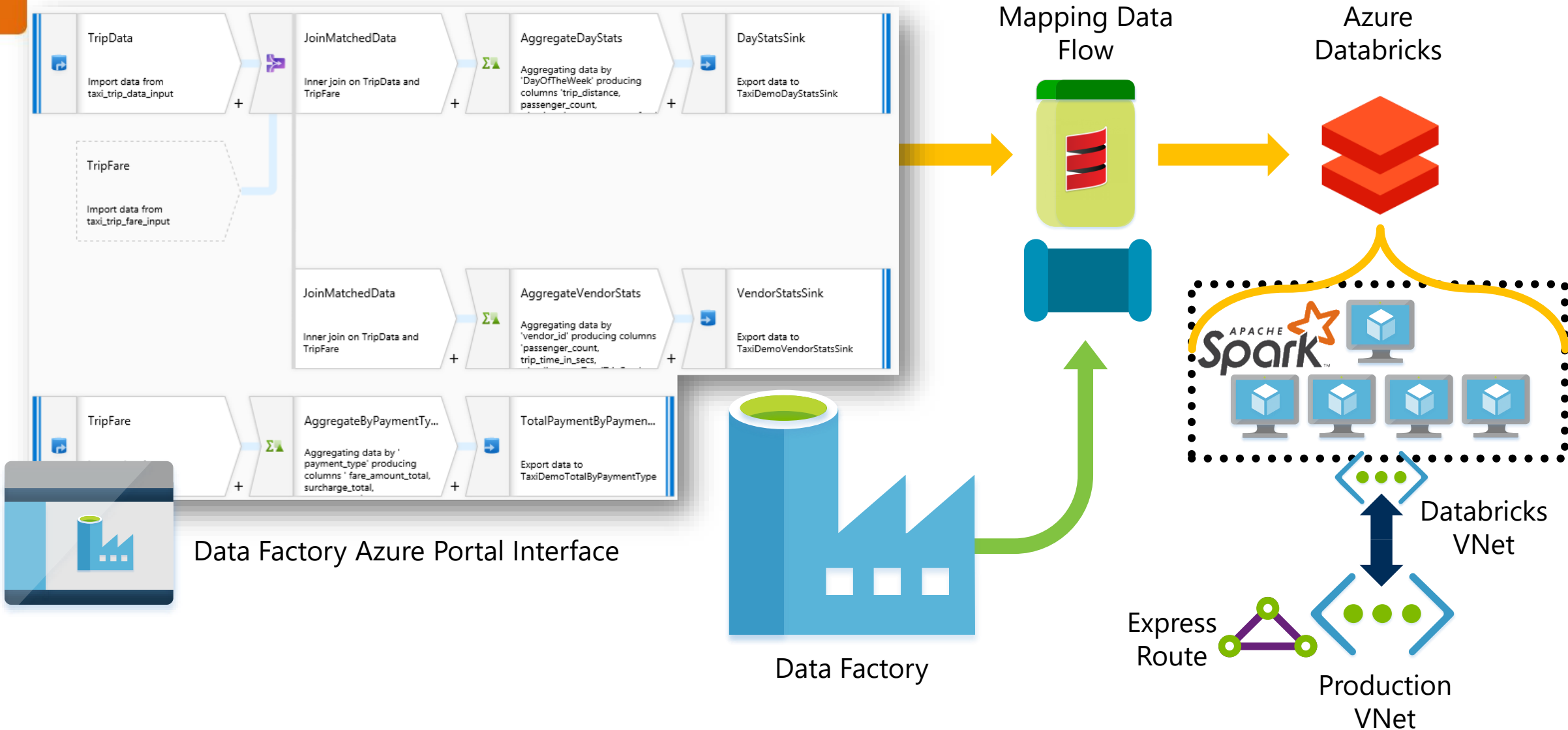
Framework Processing with Dynamic Pipelines



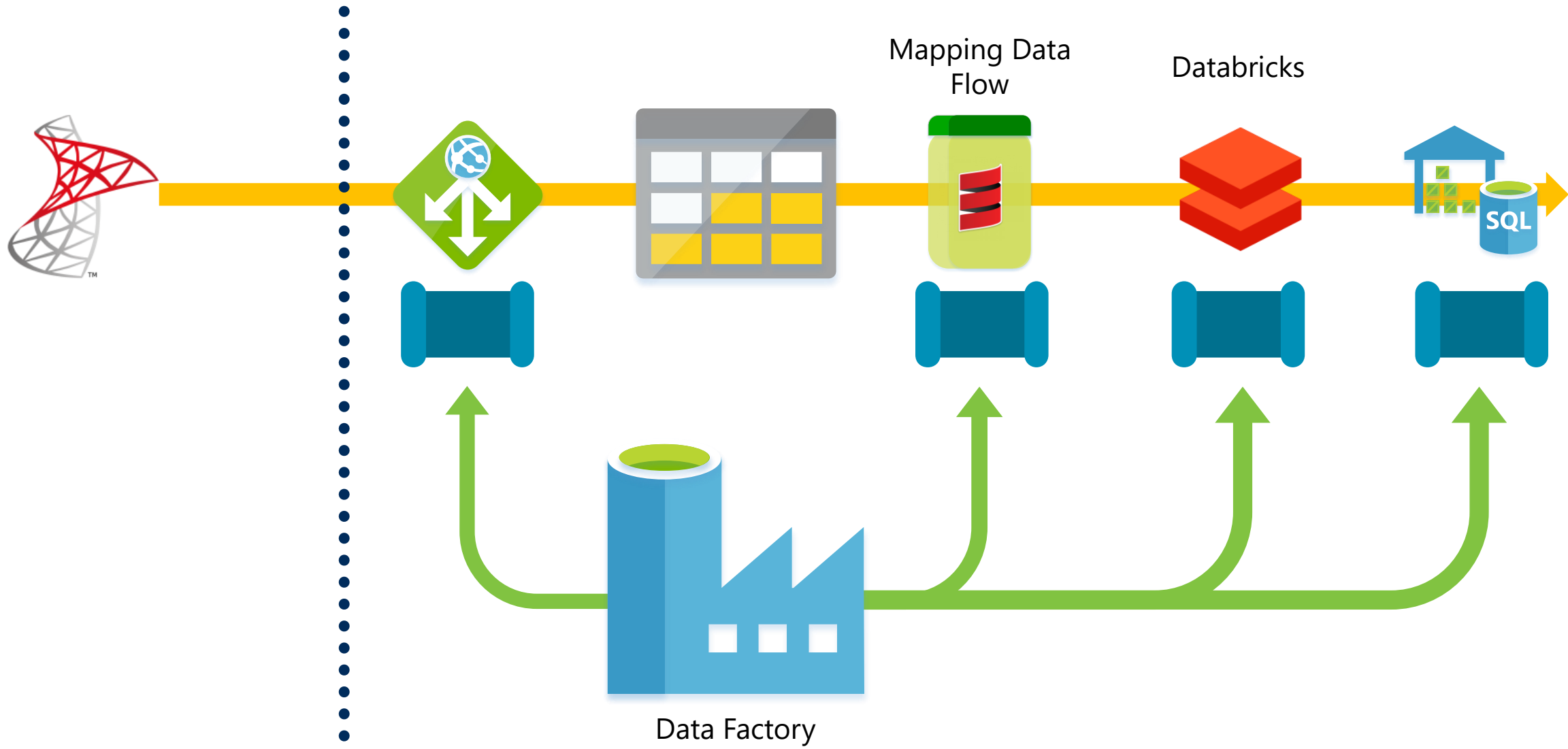
Data Transformation in Azure



What is a Mapping Data Flow?



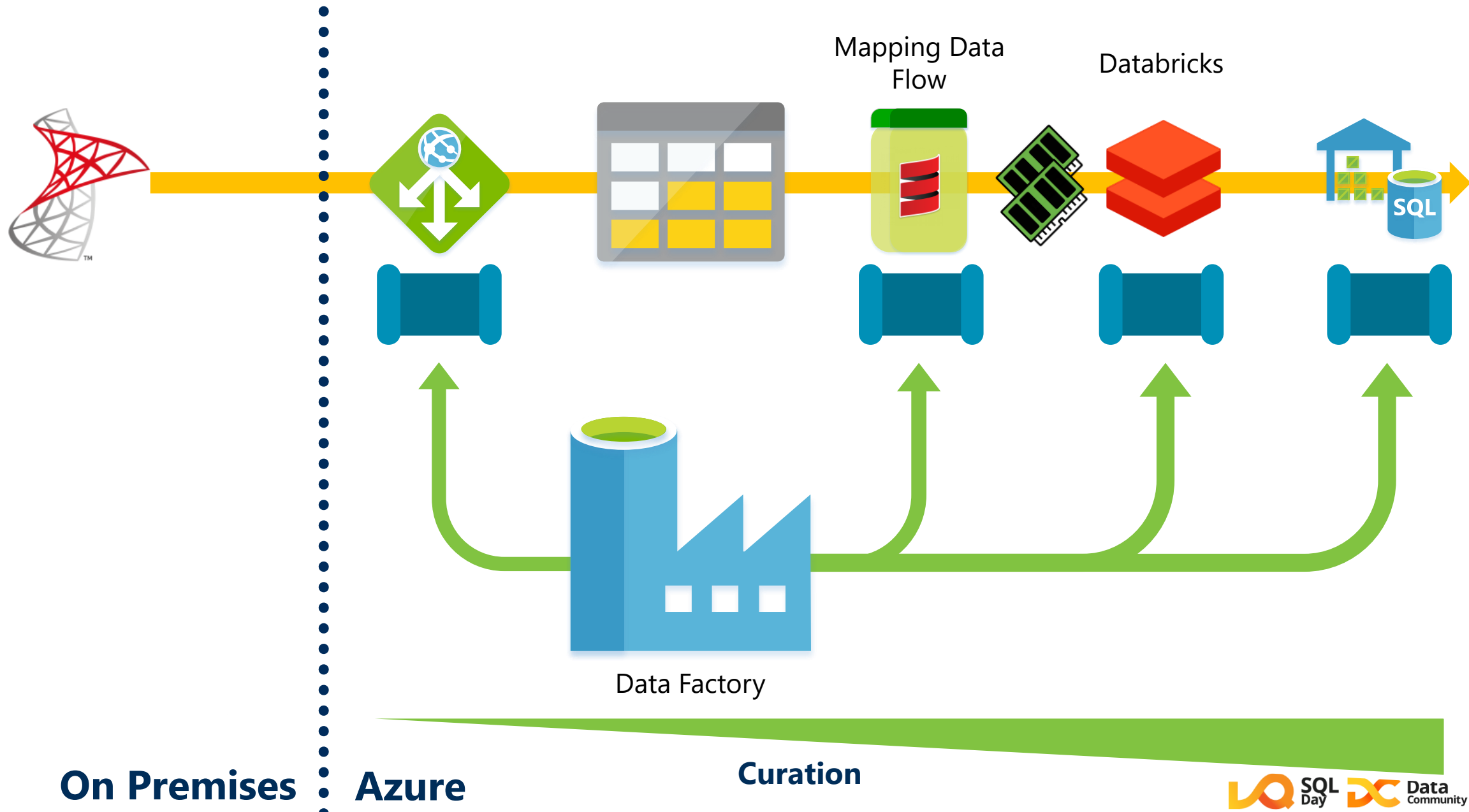
Mapping Data Flow Future Design Patterns ???



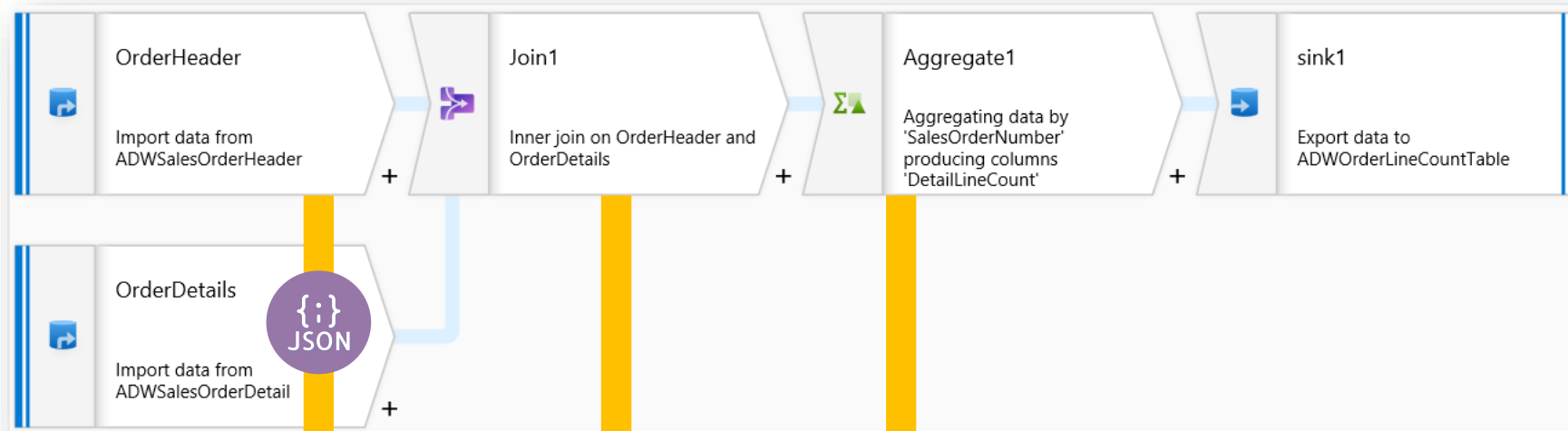
On Premises

Azure

Mapping Data Flow Future Design Patterns ???



Mapping Data Flow Future Design Patterns ???



```

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

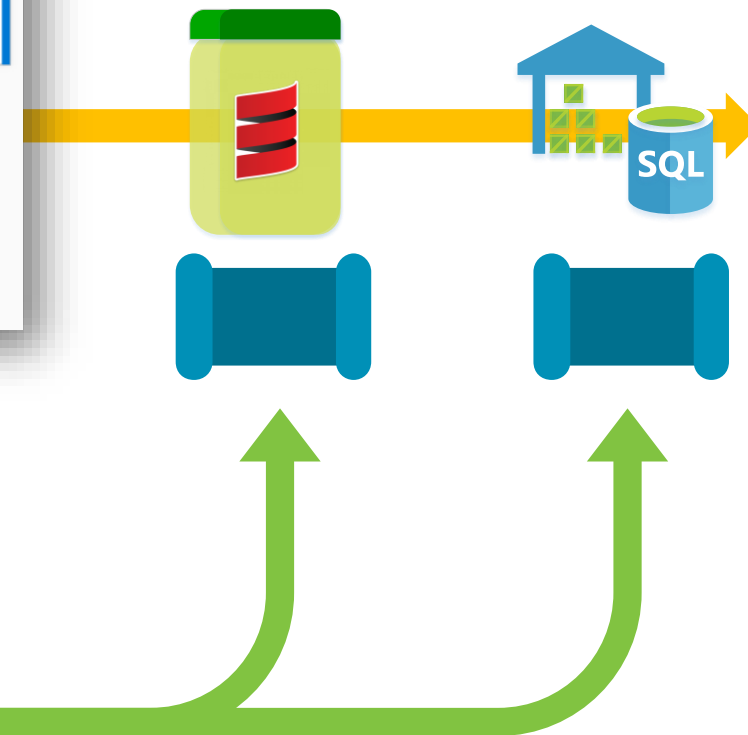
```

```

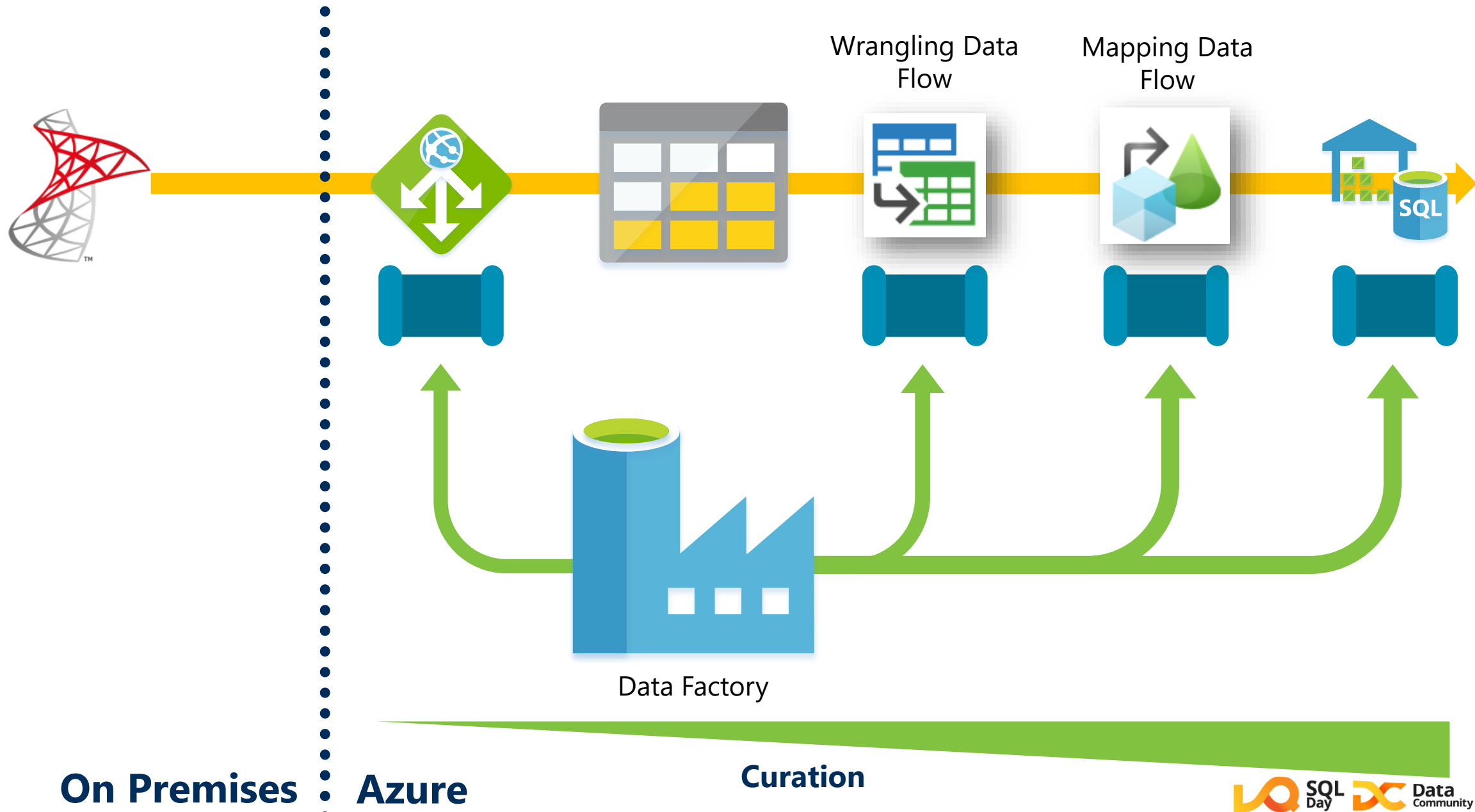
"transformations": [
  {
    "name": "Join1",
    "script": "OrderHeader, OrderDetail join(OrderHeader@SalesOrderID == OrderDetail@SalesOrderID, \n\tjoinType:'inner', \n\tbroadcast: 'none') ~> Join1"
  },
  {
    "name": "Aggregate1",
    "script": "Join1 aggregate(groupBy(SalesOrderNumber), \n\tDetailLineCount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

```

Mapping Data Flow



Future Design Patterns ???



Complex Orchestration

With Dynamic Data Factory Pipelines



Azure Data Factory

A very quick overview

Extensibility & Parallelism

Custom Activities
SSIS IR & Packages

More Design Patterns

Bootstrapping
Hosted IR vs IaaS Frameworks

Thanks for Listening

Paul Andrew

 @MrPaulAndrew



Blog: mrpaulandrew.com

Email: paul@mrpaulandrew.com

GitHub: github.com/mrpaulandrew

