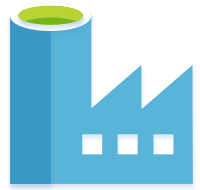
Data Community

SQL Day

# Complex Orchestration



## With Dynamic Data Factory Pipelines


Paul Andrew | Principal Consultant & Solution Architect

Microsoft® Most Valuable Professional

@MrPaulAndrew

Microsoft Partner
Microsoft
Gold Cloud Platform
Gold Data Analytics
Gold Data Platform
Gold DevOps
Gold Datacenter

adatis

Data Community

SQL Day

## PLATINUM SPONSOR

TECHNOLOGY INNOVATION DATA KNOWLEDGE tidk

## STRATEGIC PARTNER

Microsoft

## GOLD SPONSORS

CLOUDS ON MARS

Future Processing

SPYROSOFT

## SILVER SPONSOR

dbWatch DATABASE CONTROL

elitmind think bright

VEEAM

## BRONZE SPONSOR

Największa księgarnia IT w Polsce
NOVATECH
www.novatech.com.pl

# GitHub

https://github.com/mrpaulandrew

**CommunityEvents**

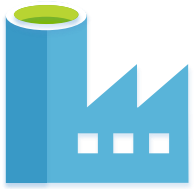
Demo code, content and slides from various community events.

C++

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

# Agenda

# Complex Orchestration

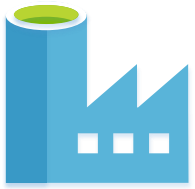


## With Dynamic Data Factory Pipelines


| Azure Data Factory | Extensibility & Parallelism | More Design Patterns |
| --- | --- | --- |
| A very quick overview | Custom Activities<br>SSIS IR & Packages | Bootstrapping<br>Hosted IR vs IaaS<br>Frameworks |

# Agenda

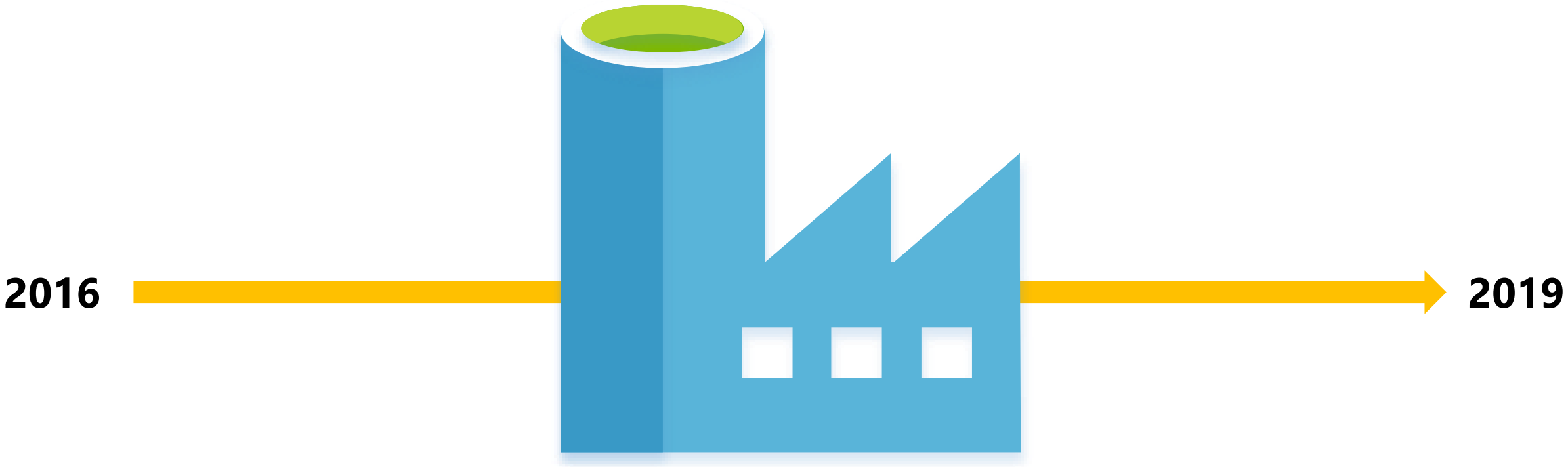
# Complex Orchestration



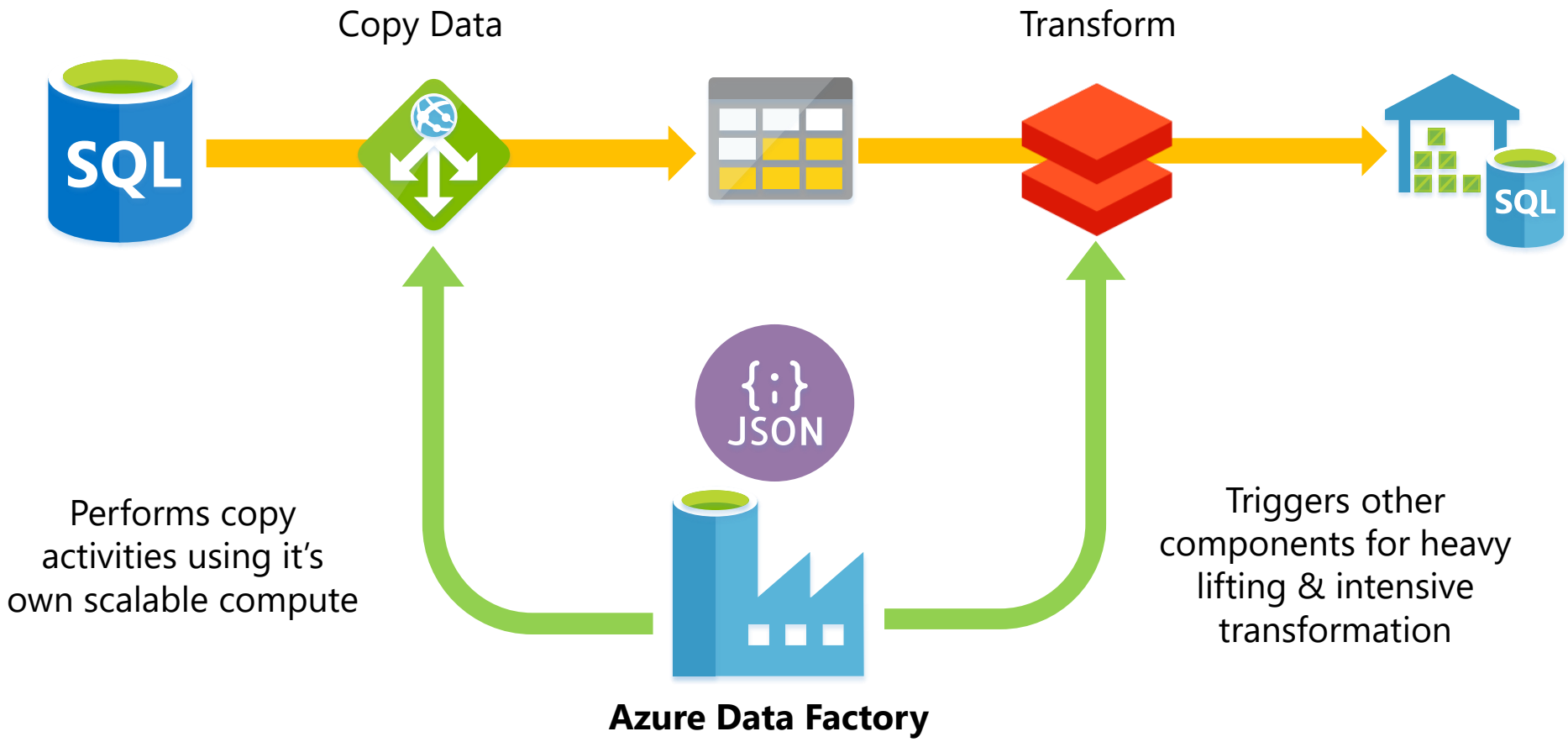
## With Dynamic Data Factory Pipelines


| Azure Data Factory | Extensibility & Parallelism | More Design Patterns |
| --- | --- | --- |
| A very quick overview | Custom Activities<br>SSIS IR & Packages | Bootstrapping<br>Hosted IR vs IaaS<br>Frameworks |

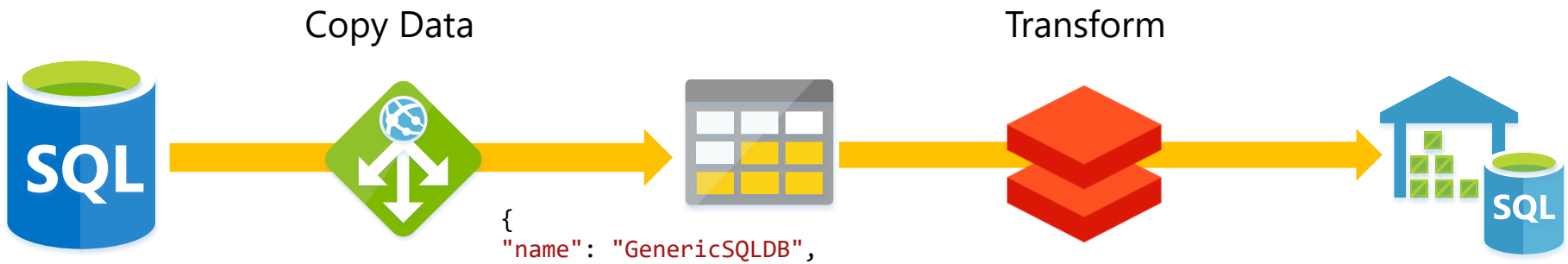
# Azure Data Factory



# What is Azure Data Factory?



# Data Factory Components – Recap




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

JSON

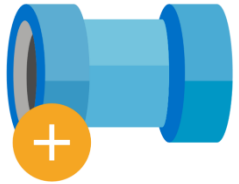
```
{
"name": "GenericSQLDB",
"type": "Microsoft.DataFactory/factories/linkedservices",
"properties": {
    "parameters": {
        "ServerInstance": {
            "type": "String"
        },
        "DatabaseName": {
            "type": "String"
        },
        "SQLUser": {
            "type": "String"
        },
        "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}"
    }
}
```

# Integration Runtimes

1

**Azure**
Integration Runtime

Movement Hours



Activity
Orchestration


Flexible Region

2

**SSIS**
Integration Runtime

SSIS Package
Execution

Specified Region

3

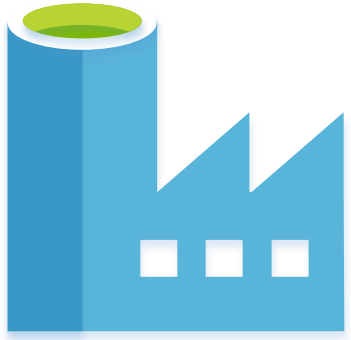
**Self Hosted**
Integration Runtime

Gateway Access

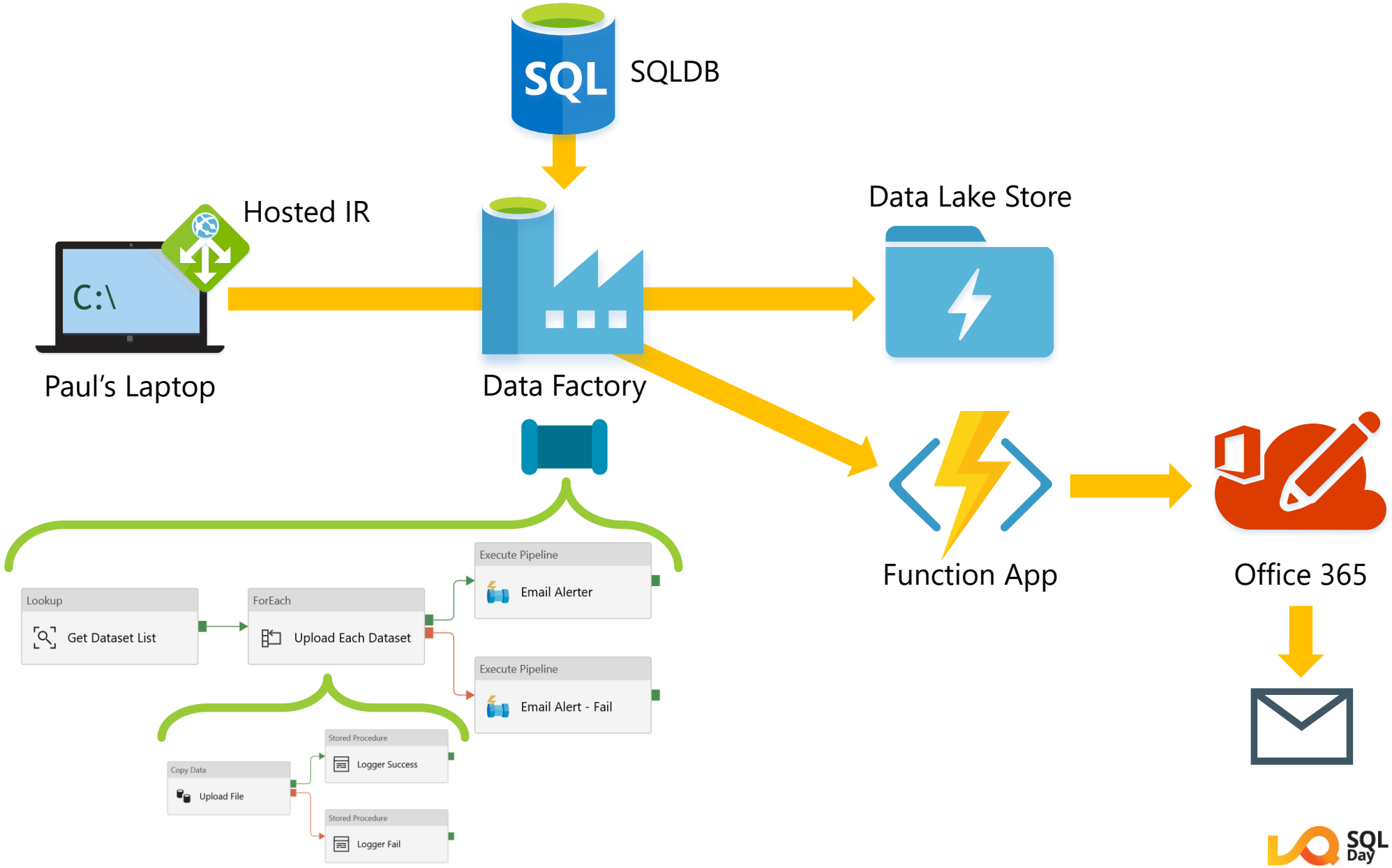
Activity
Orchestration


Virtual Machine

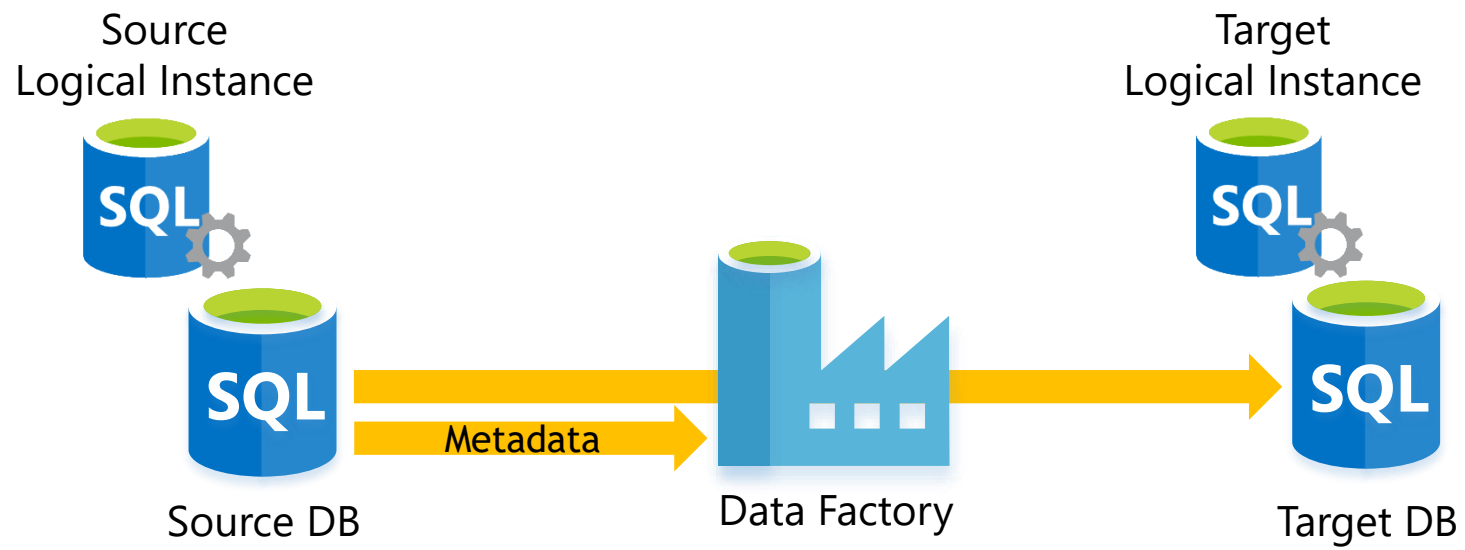
# Demo

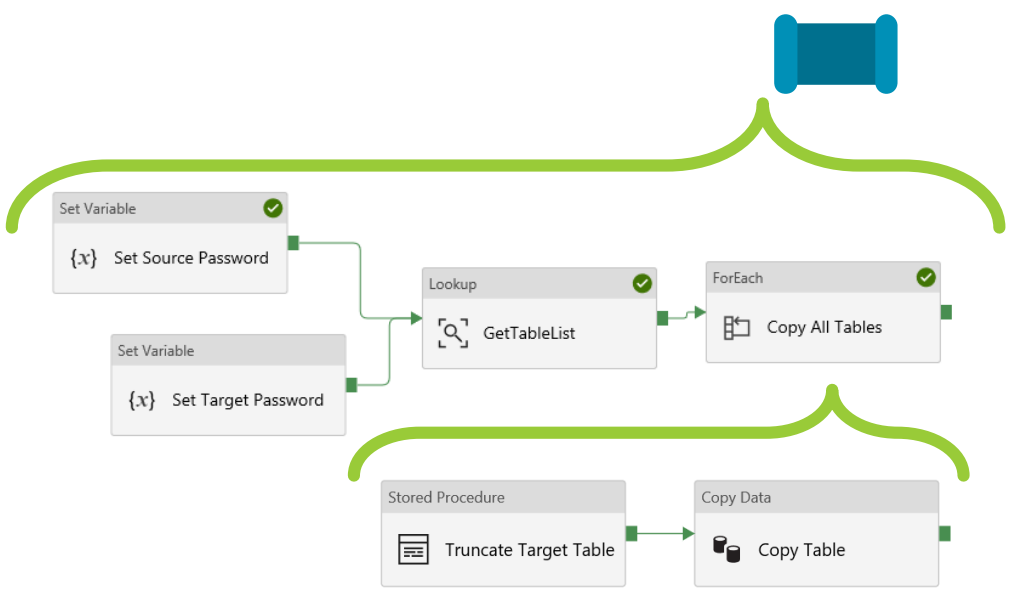


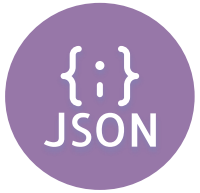
# Demo Architecture 1



# Demo Architecture 2




1x Linked Service

1x Dataset

# Agenda

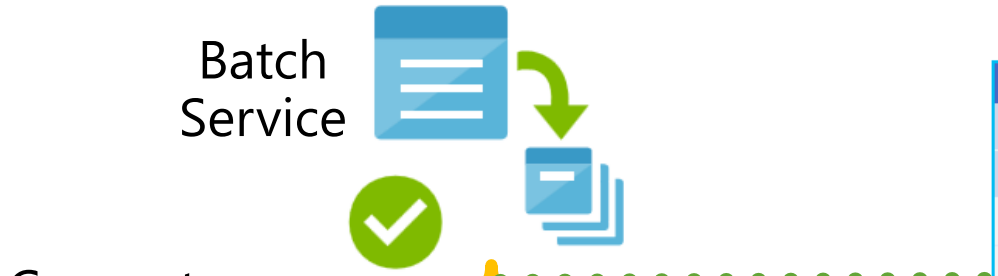
# Complex Orchestration

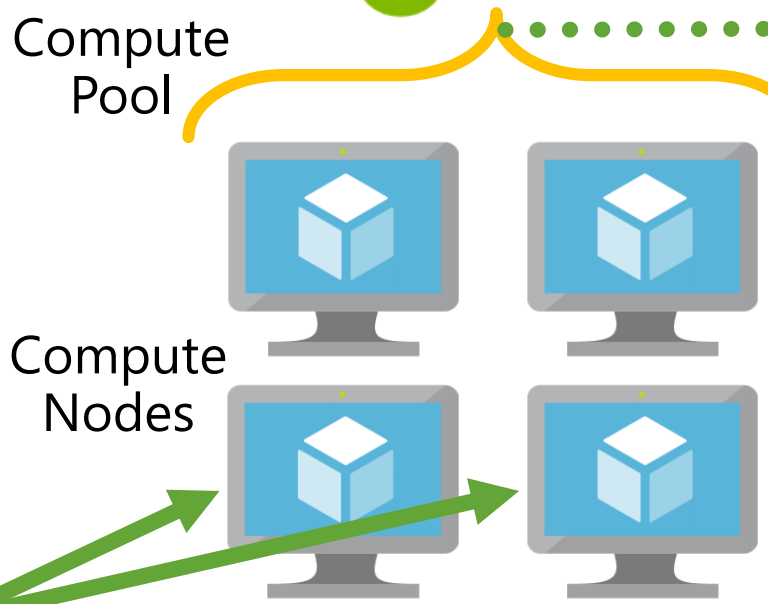
## With Dynamic Data Factory Pipelines


| Azure Data Factory | Extensibility & Parallelism | More Design Patterns |
| --- | --- | --- |
| A very quick overview | Custom Activities<br>SSIS IR & Packages | Bootstrapping<br>Hosted IR vs IaaS<br>Frameworks |

# ADF Extensibility

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





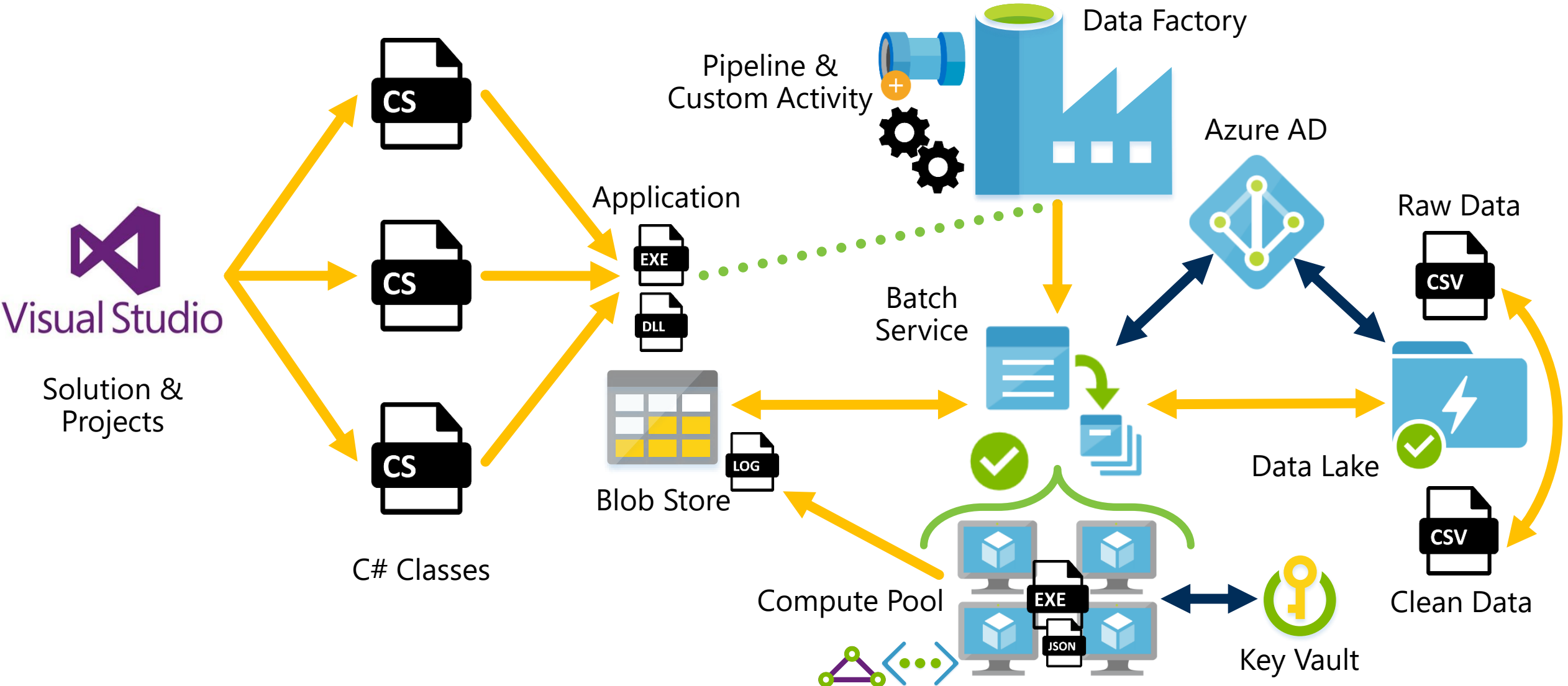
VM node size set per compute pool:

| 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

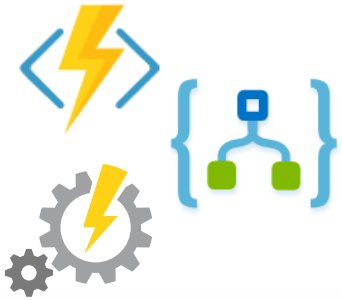
1 **Custom Activities** – A .Net Console App Executed Using Azure Batch Service

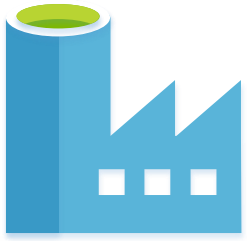


# ADF Extensibility Continued

1. **Custom Activities** – A .Net Console App Executed Using Azure Batch Service
2. **Rest API Calls** – Eg. Web Activities Calling:

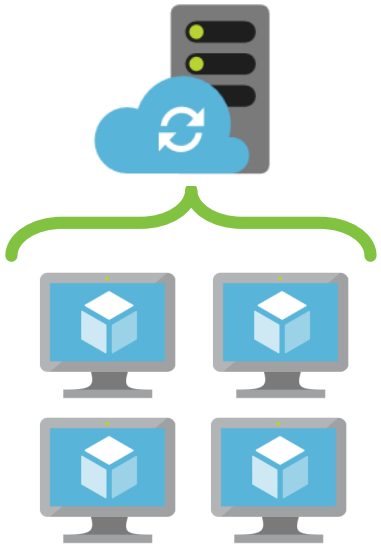
~~Azure Functions~~
Azure Logic Apps
Azure Automation



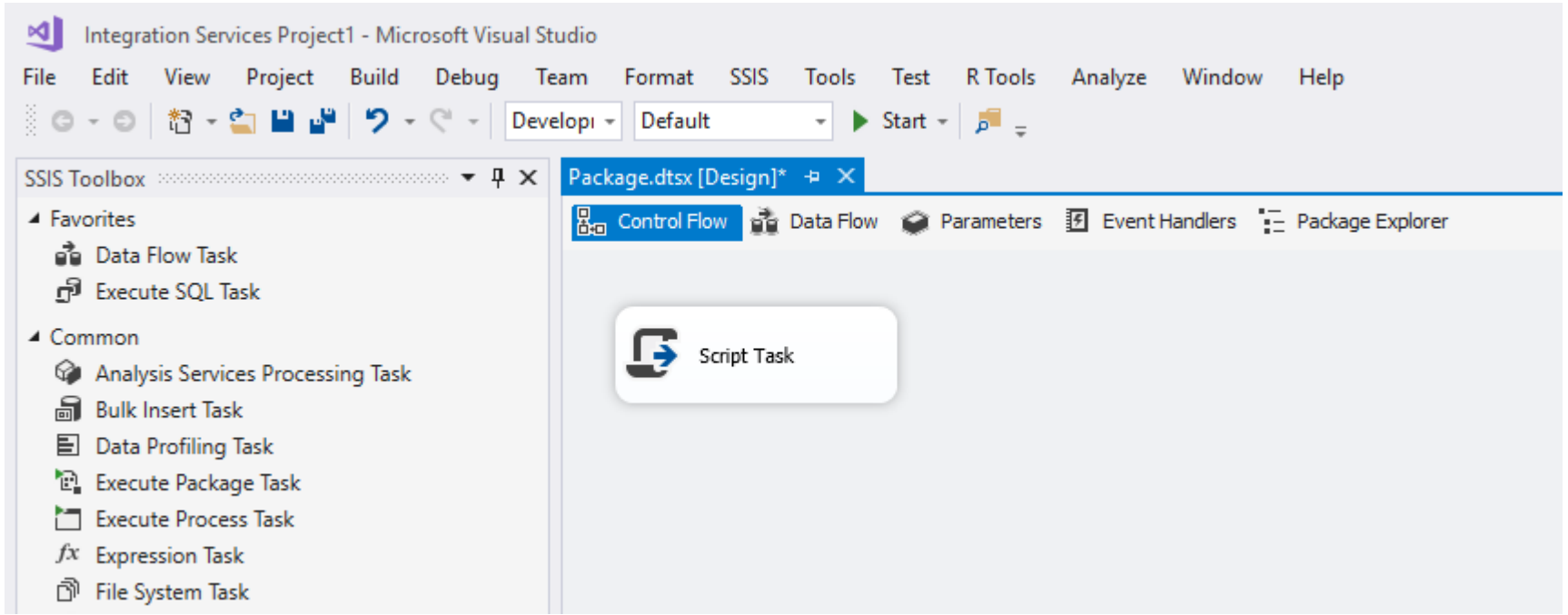


# ADF Extensibility Continued

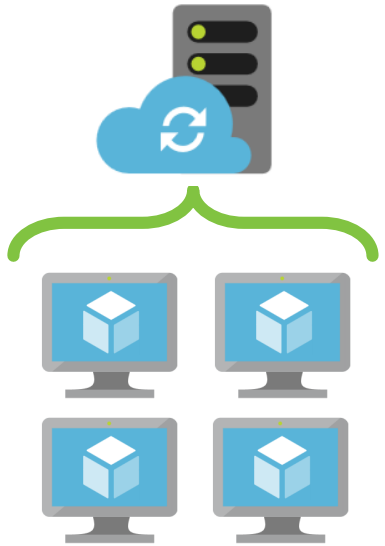
1. **Custom Activities**
2. **Rest API Calls**
3. **SSIS** – Packages with Control Flows and Data Flows



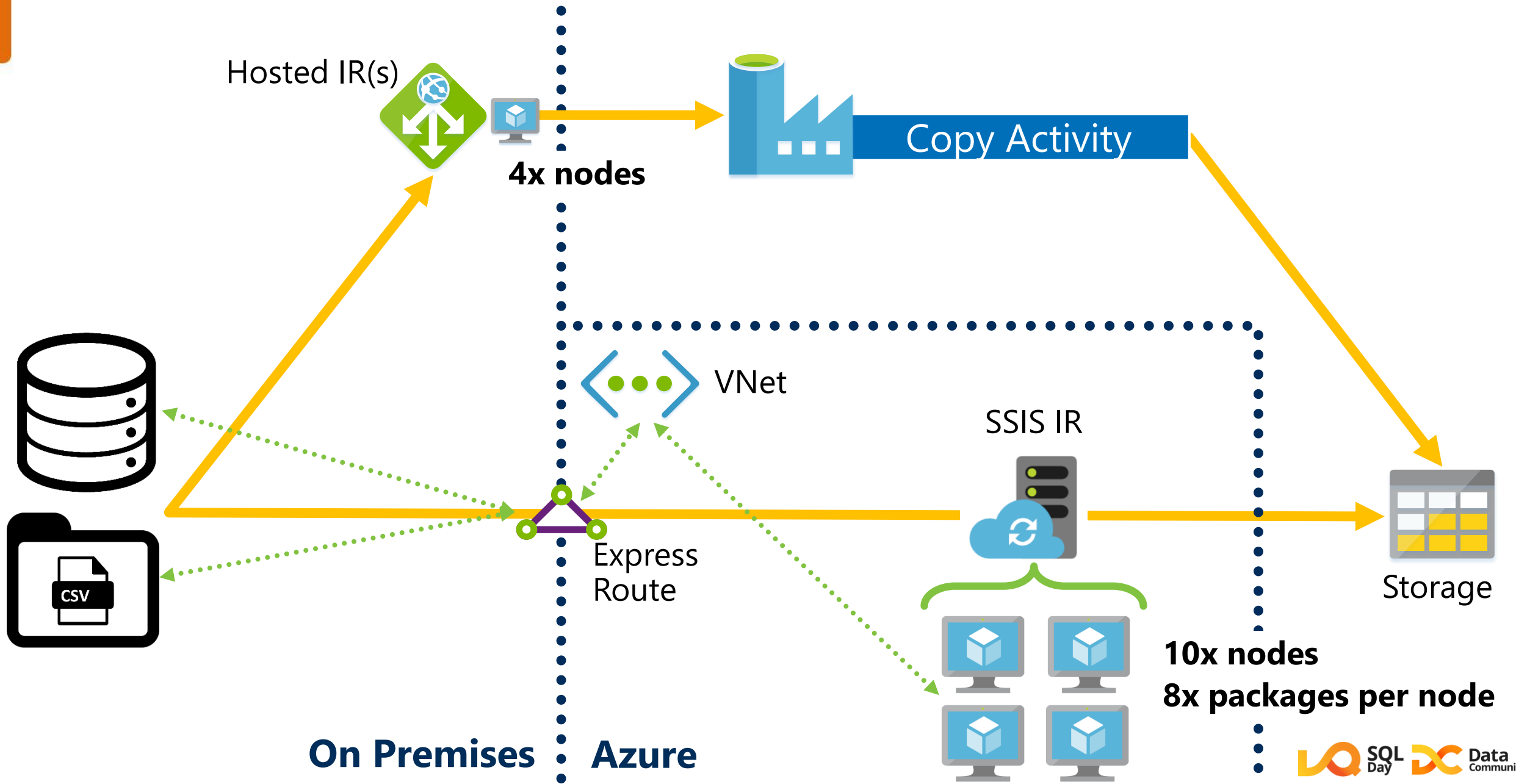
**ADF SSIS IR**



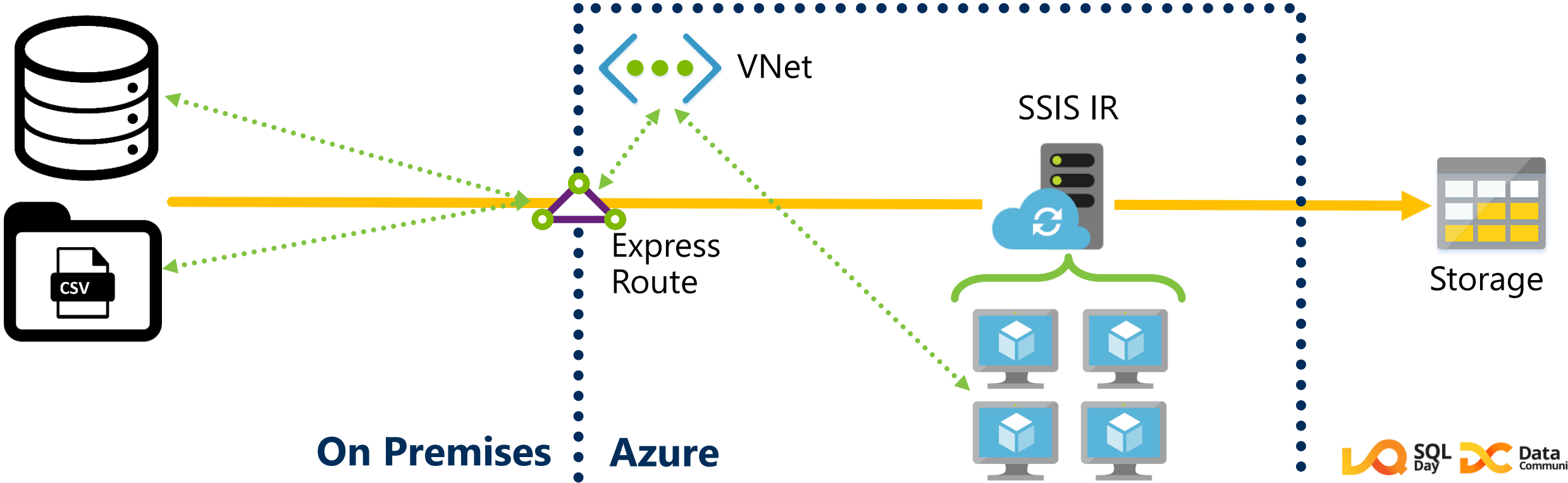
# ADF Extensibility Continued



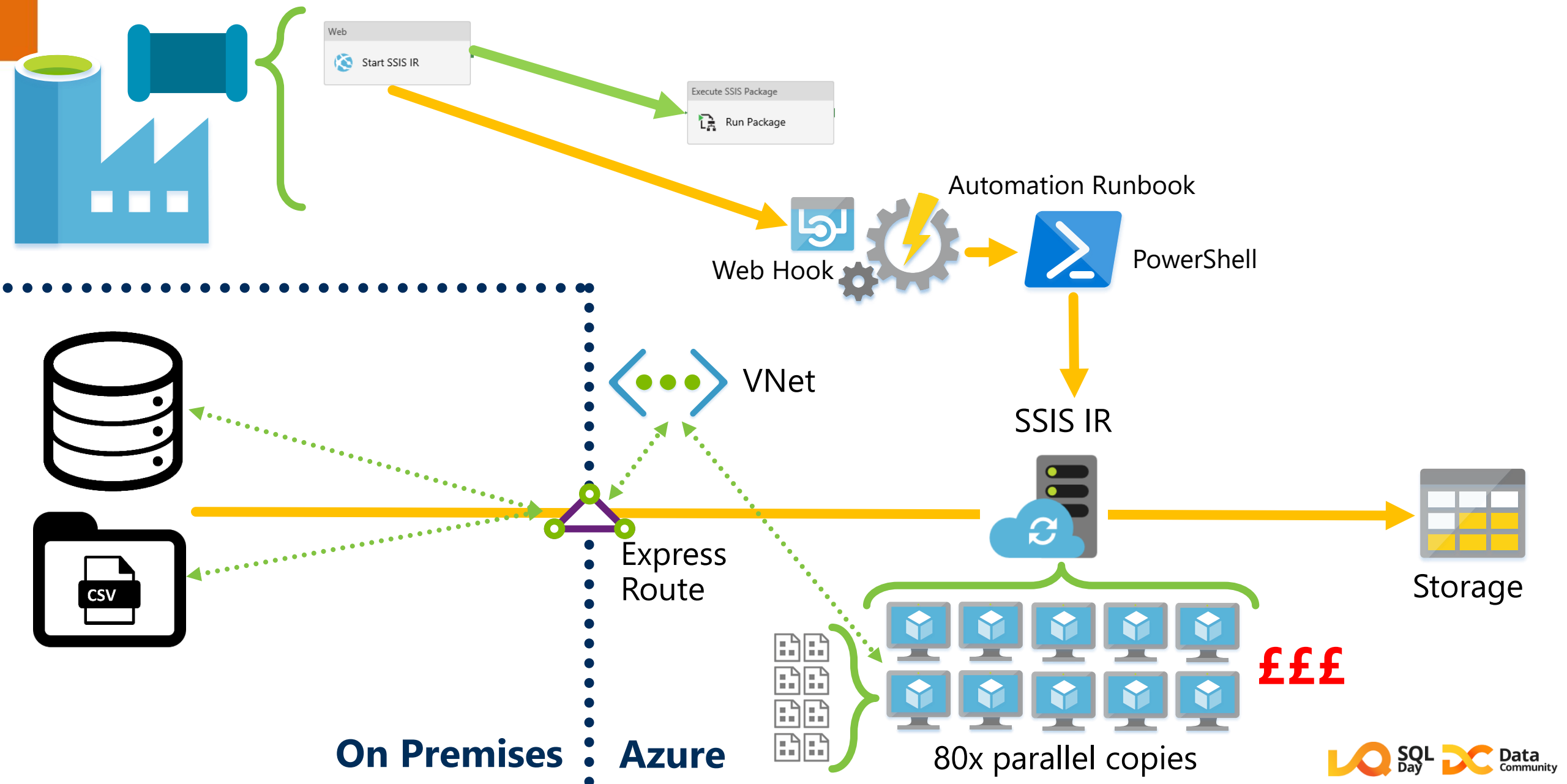
# The SSIS IR vs Hosted IR with Express Route



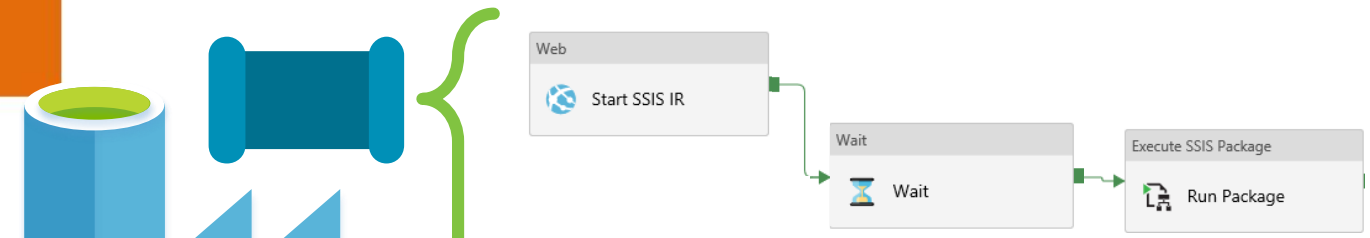
# The SSIS IR Start/Stop

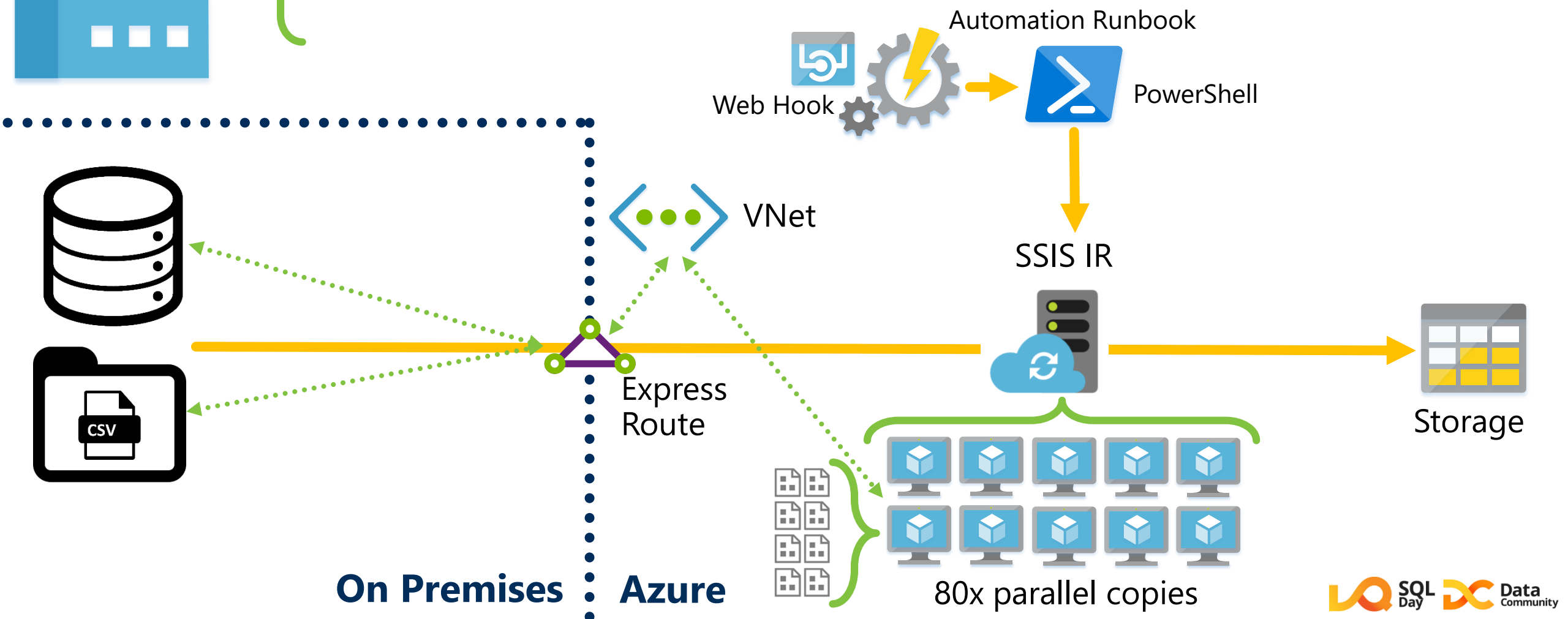


# The SSIS IR Start/Stop

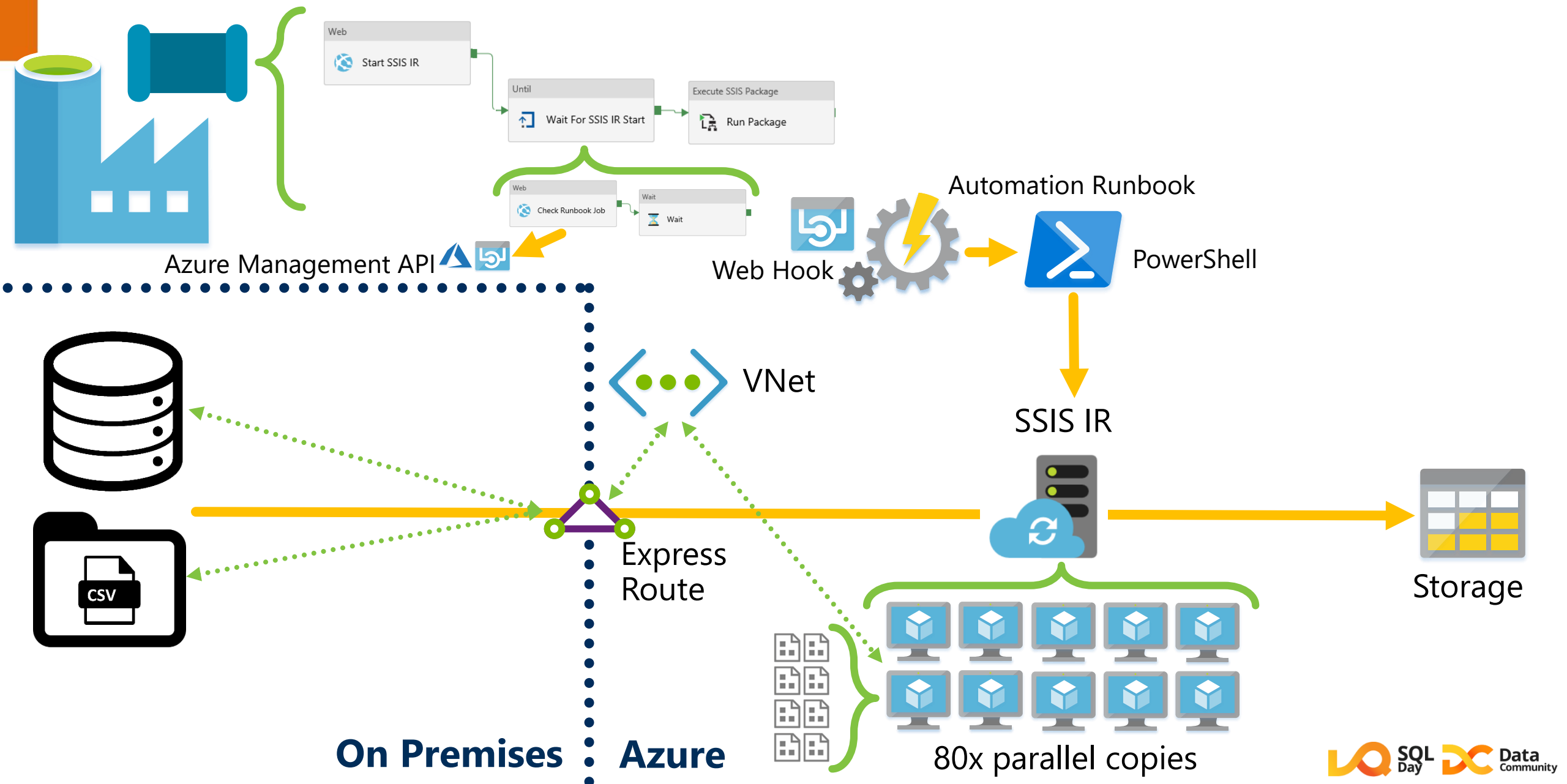


# The SSIS IR Start/Stop

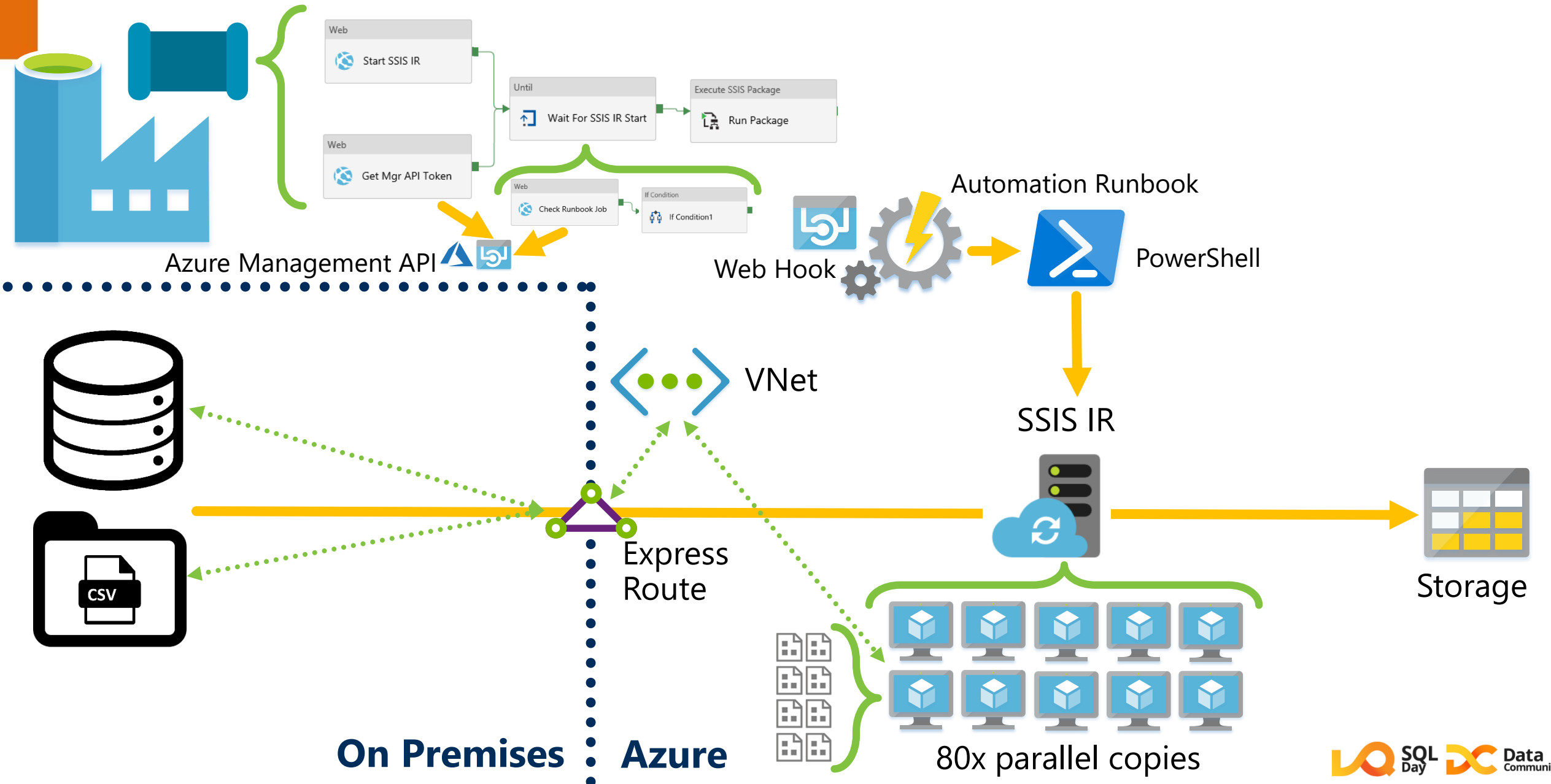


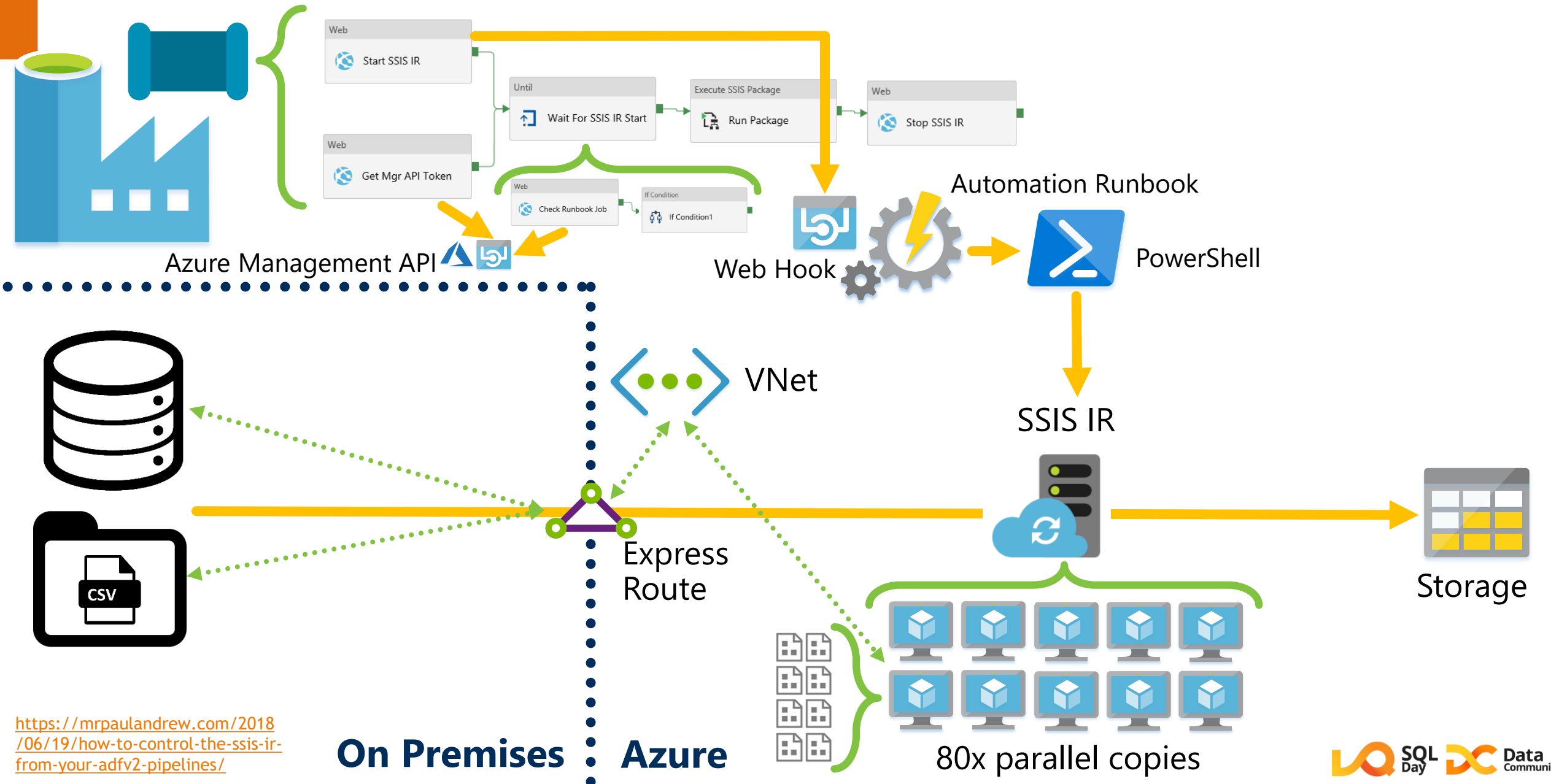
# The SSIS IR Start/Stop



# The SSIS IR Start/Stop

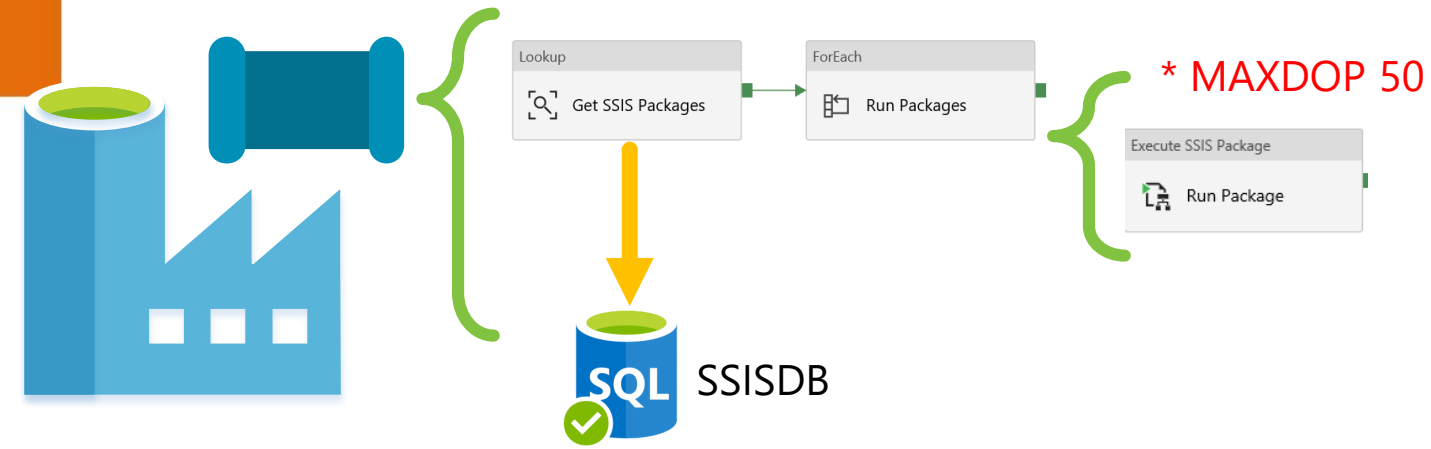


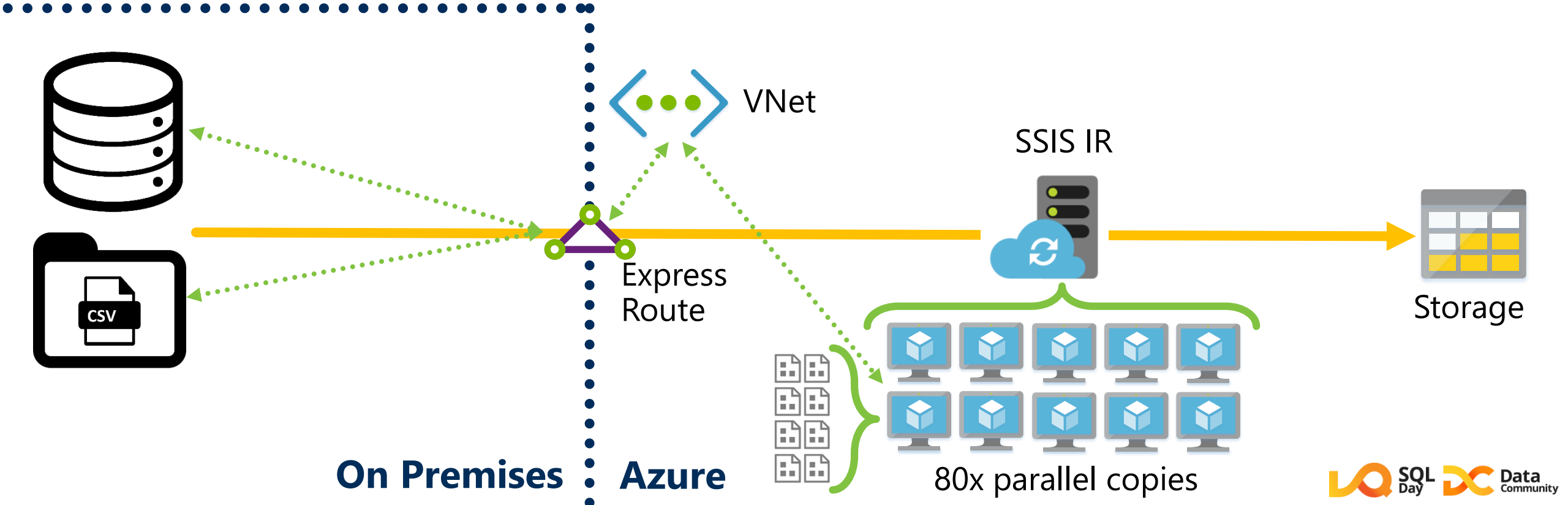
# The SSIS IR Start/Stop



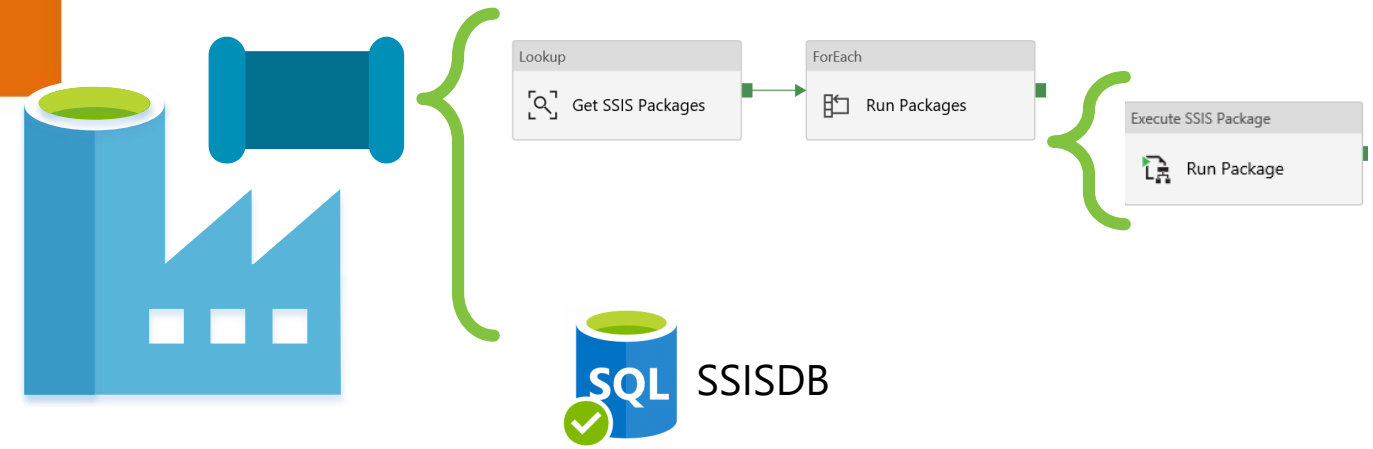

https://mrpaulandrew.com/2018/06/19/how-to-control-the-ssis-ir-from-your-adfv2-pipelines/

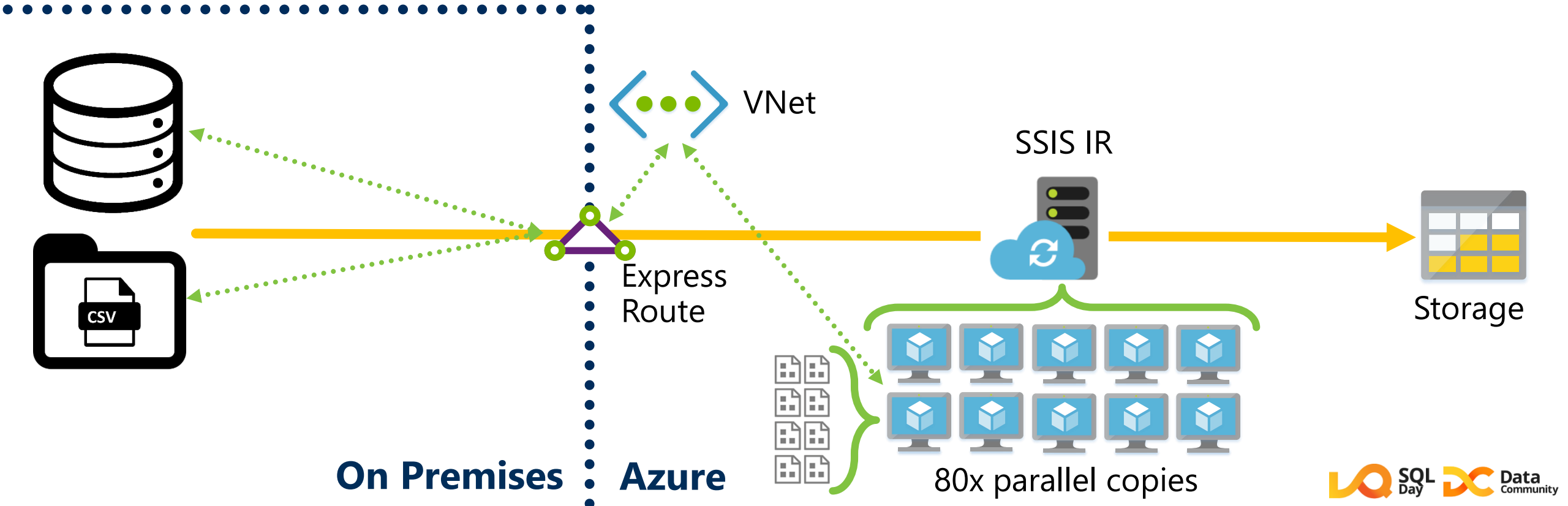
# The SSIS IR Parallelism



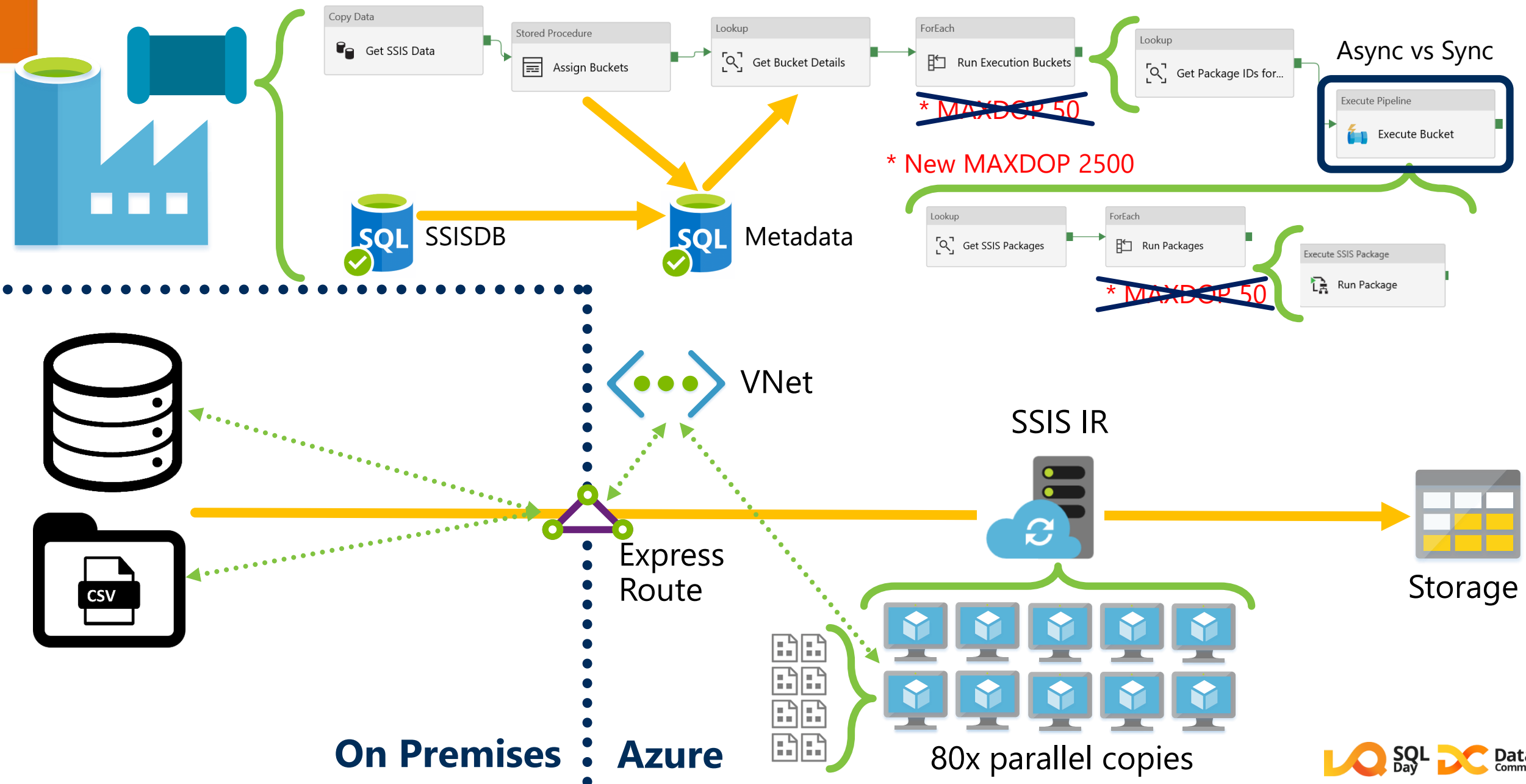



# The SSIS IR Parallelism

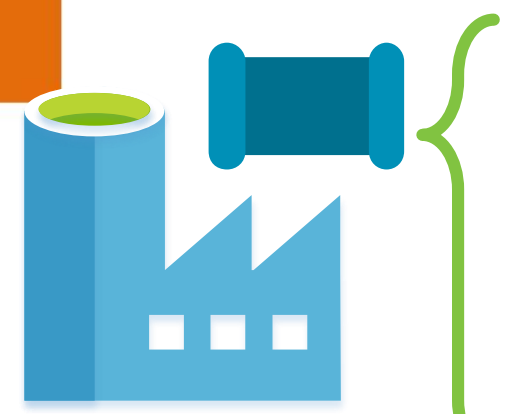


# The SSIS IR Parallelism



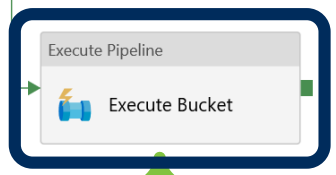
# The SSIS IR Parallelism

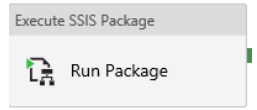



| Resource | Default Limit | Maximum Limit |
|---|---|---|
| Data factories in an Azure subscription | 50 | Contact support |
| Total number of entities (Pipeline, Datasets, Triggers, Linked Services, Integration runtimes) within a data factory | 5000 | Contact support |
| Total CPU cores for Azure-SSIS 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

Async vs Sync



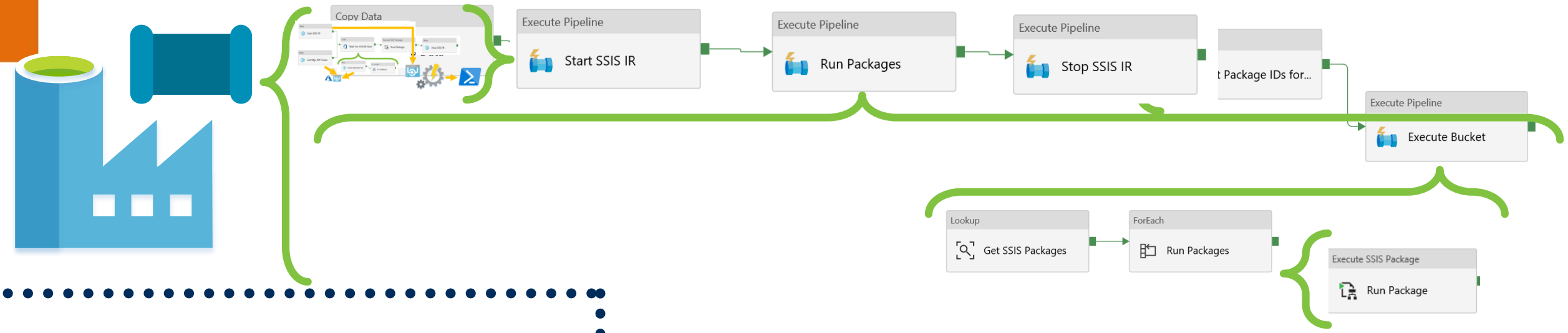



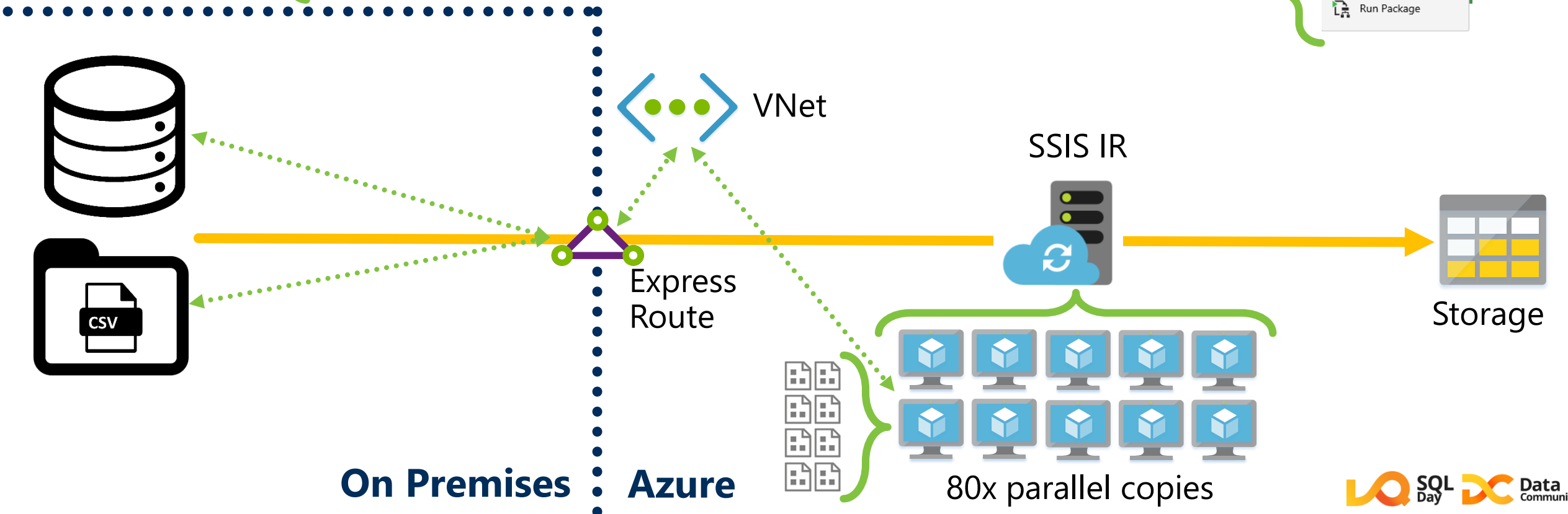

Storage

On Premises : Azure

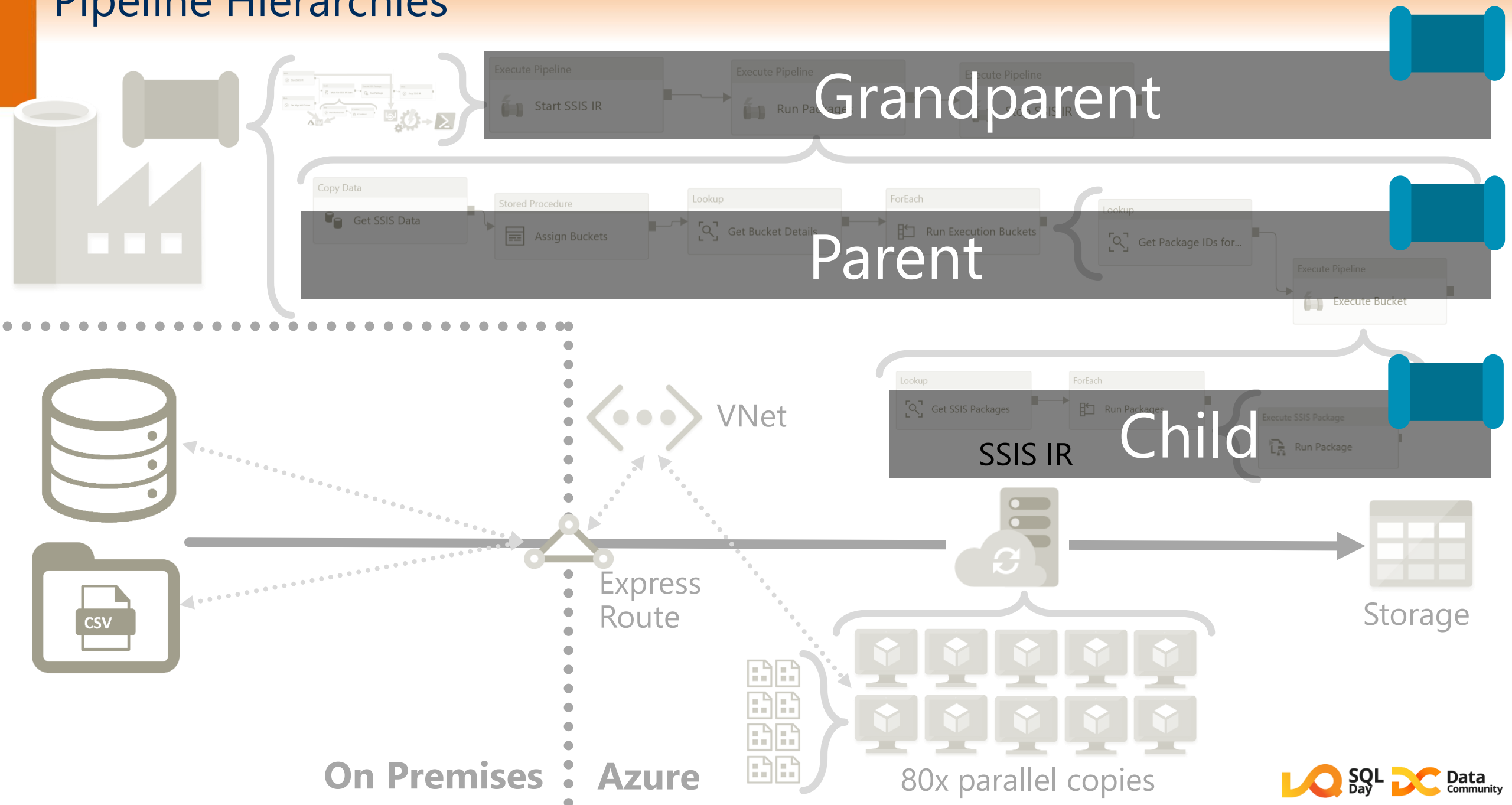
80x parallel copies

# SSIS IR & Package Complete Orchestration Solution



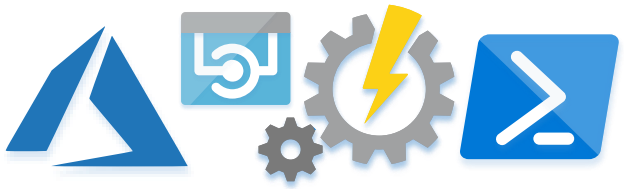



# Pipeline Hierarchies



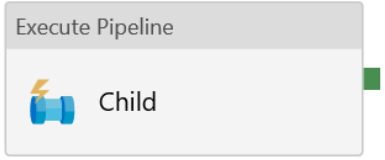
# Pattern Summary


**High Level Control Flow and Pipeline Triggers**

**Platform Component Control**

**Manage Parallel Activities**

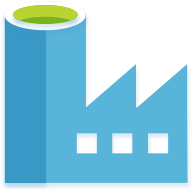




**Service Level Executions**

# Agenda

# Complex Orchestration



## With Dynamic Data Factory Pipelines


| Azure Data Factory | Extensibility & Parallelism | More Design Patterns |
| --- | --- | --- |
| A very quick overview | Custom Activities<br>SSIS IR & Packages | Bootstrapping<br>Hosted IR vs IaaS<br>Frameworks |

# Bootstrapping

Data Factory

Azure Function

Azure Function1

Functions

```
Console.WriteLine("Creating pipeline run...");
Dictionary<string, object> parameters = new Dictionary<string,
object> {
                    {"inputPath", inputBlobPath }, {
"outputPath", outputBlobPath }
                    };
CreateRunResponse runResponse =
client.Pipelines.CreateRunWithHttpMessagesAsync(resourceGroup,
dataFactoryName, pipelineName, parameters:
parameters).Result.Body;
Console.WriteLine("Pipeline run ID: " + runResponse.RunId);
```

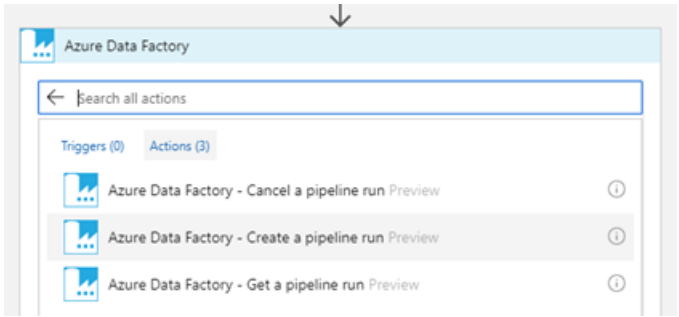
Web

Web1

Automation Runbooks

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

Logic Apps



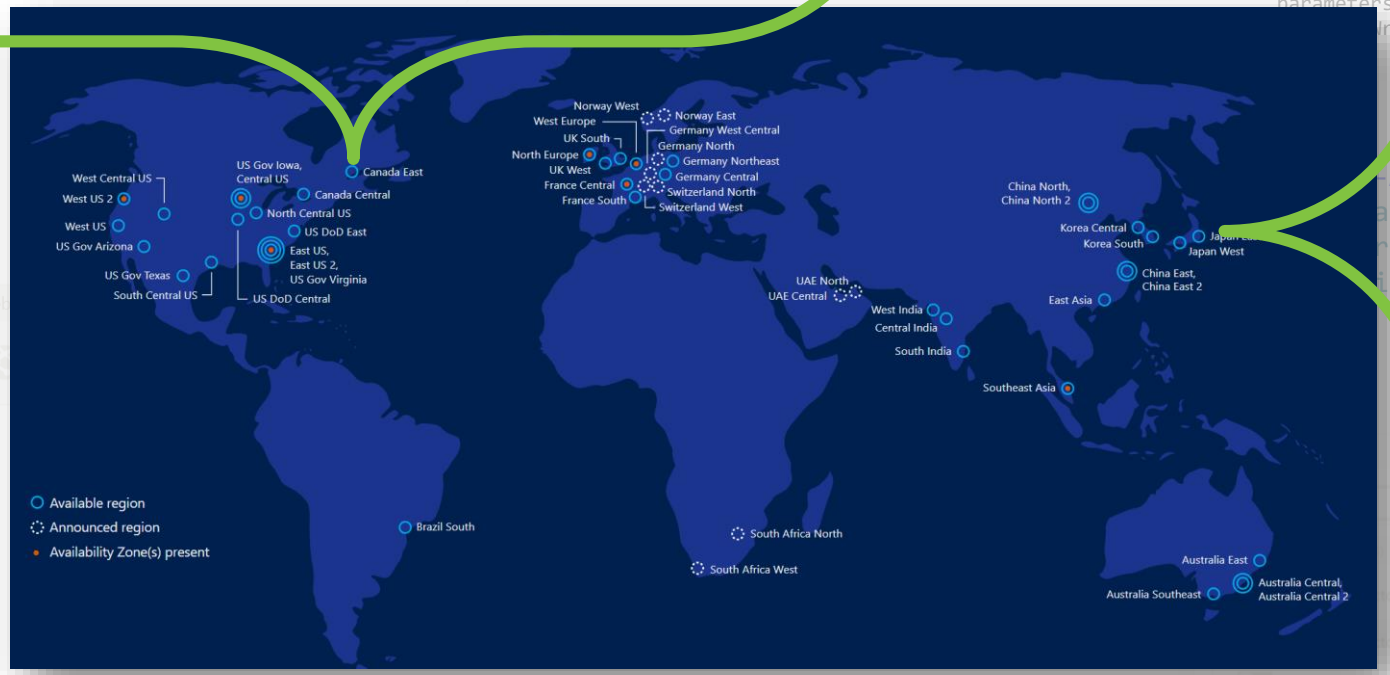
# Bootstrapping – Why?

Data Factory

Tenant 1

Subscription 1

Functions




Tenant 2

Subscription 2

```
Console.WriteLine("Creating pipeline run...");
Dictionary<string, object> parameters = new Dictionary<string,
object> {
                    {"inputPath", inputBlobPath }, {
"outputPath", outputBlobPath }
                    };
CreateRunResponse runResponse =
client.Pipelines.CreateRunWithHttpMessagesAsync(resourceGroup,
dataFactoryName, pipelineName, parameters:
parameters).Result.Body;
WriteLine("Pipeline run ID: " + runResponse.RunId);

AzureRmDataFactoryV2Pipeline
actoryName $dataFactoryName
rceGroupName $resourceGroupName
ineName $pipelineName
```

# Bootstrapping – Why?

Data Factory

Azure Function

Azure Function1

Functions

```
Console.WriteLine("Creating pipeline run...");
Dictionary<string, object> parameters = new Dictionary<string,
object> {
                    {"inputPath", inputBlobPath }, {
"outputPath", outputBlobPath }
                    };
CreateRunResponse runResponse =
client.Pipelines.CreateRunWithHttpMessagesAsync(resourceGroup,
dataFactoryName, pipelineName, parameters:
parameters).Result.Body;
Console.WriteLine("Pipeline run ID: " + runResponse.RunId);
```

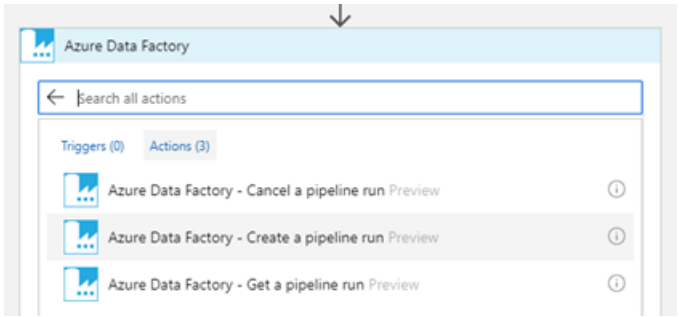
Web

Web1

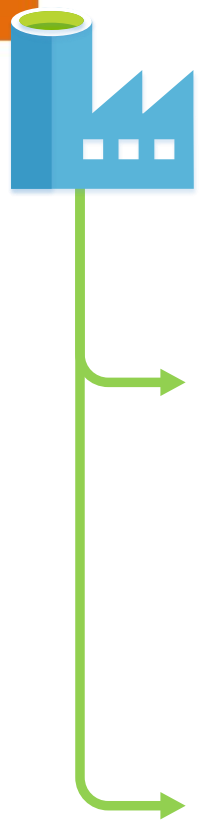
Automation Runbooks

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

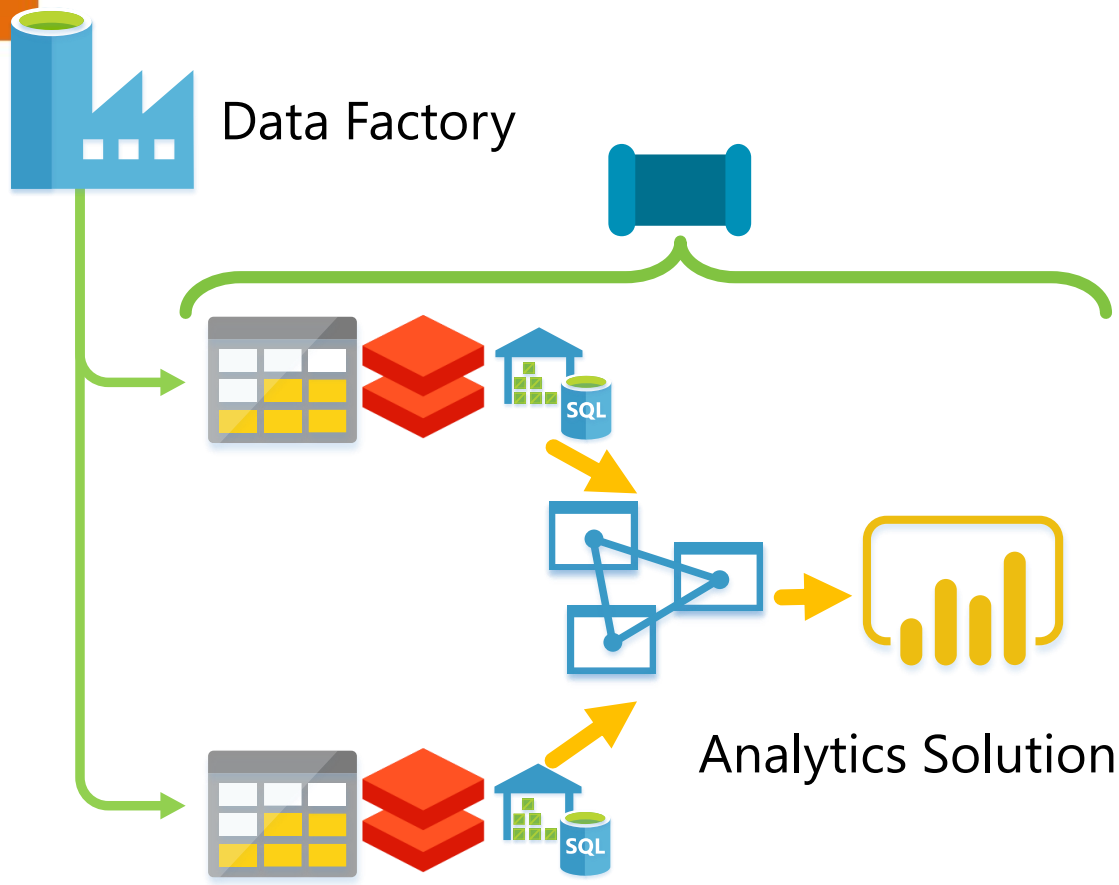
Logic Apps



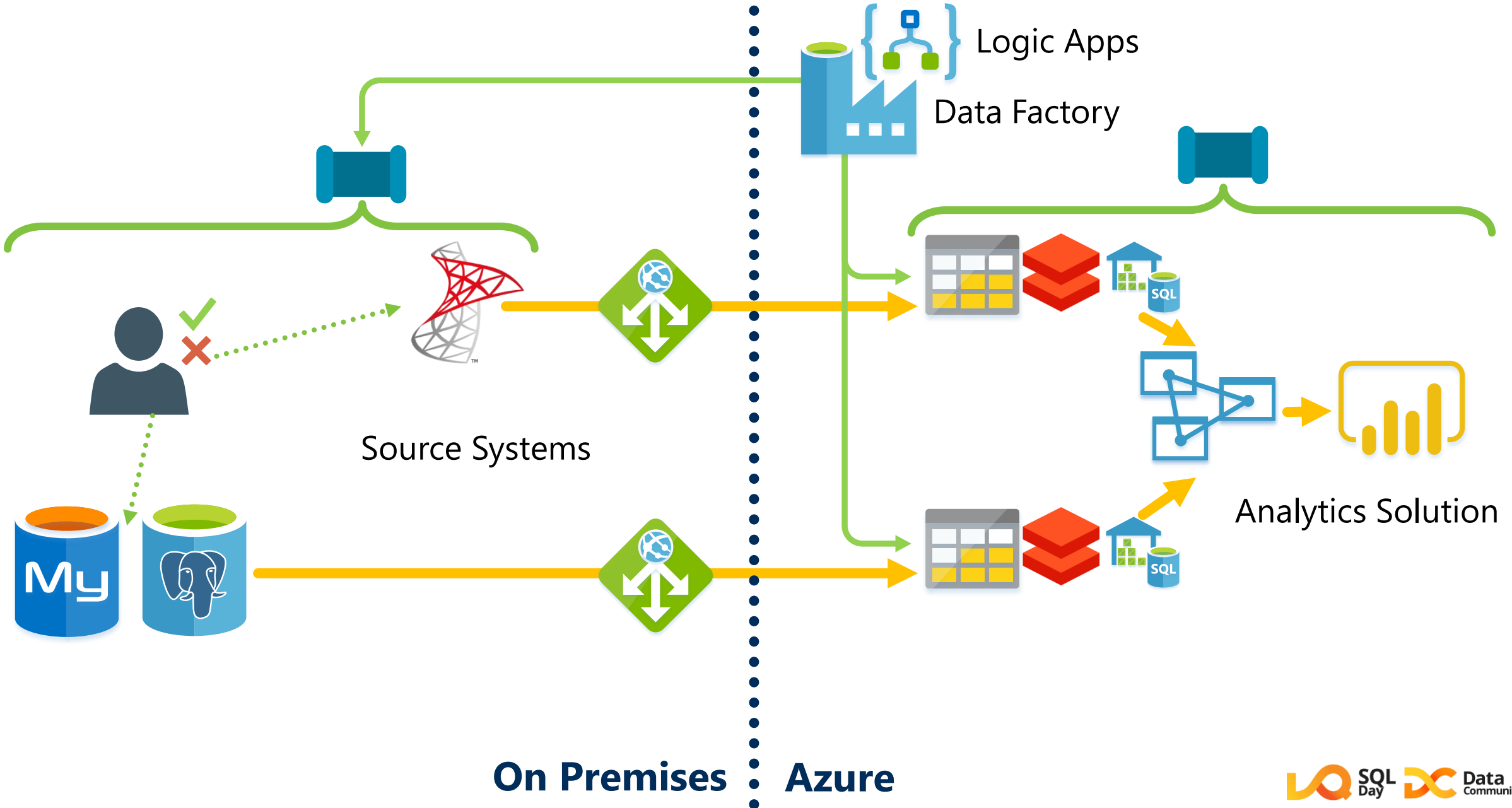
# Bootstrapping – Why?



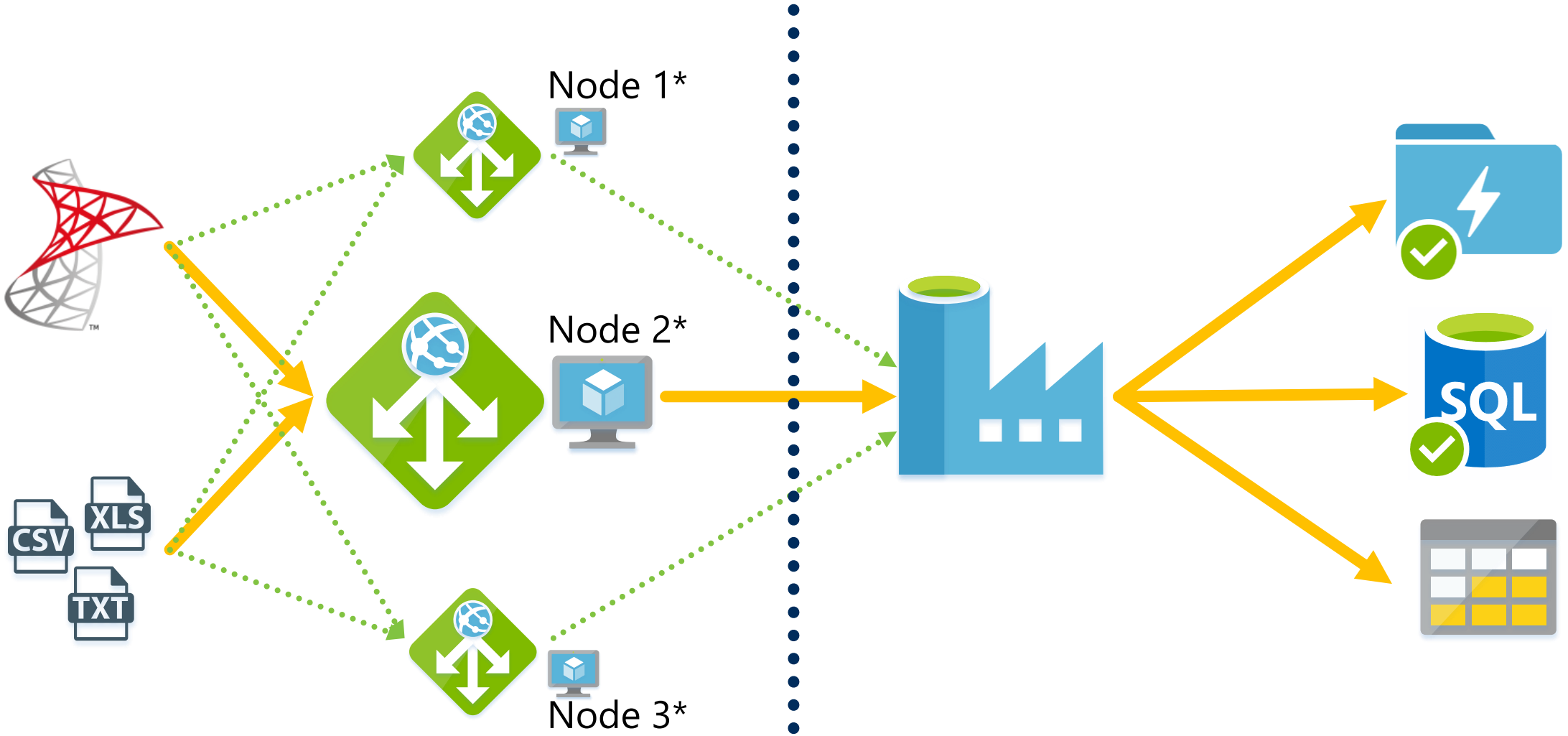
# Bootstrapping – Why?



# Bootstrapping vs Data Ingestion



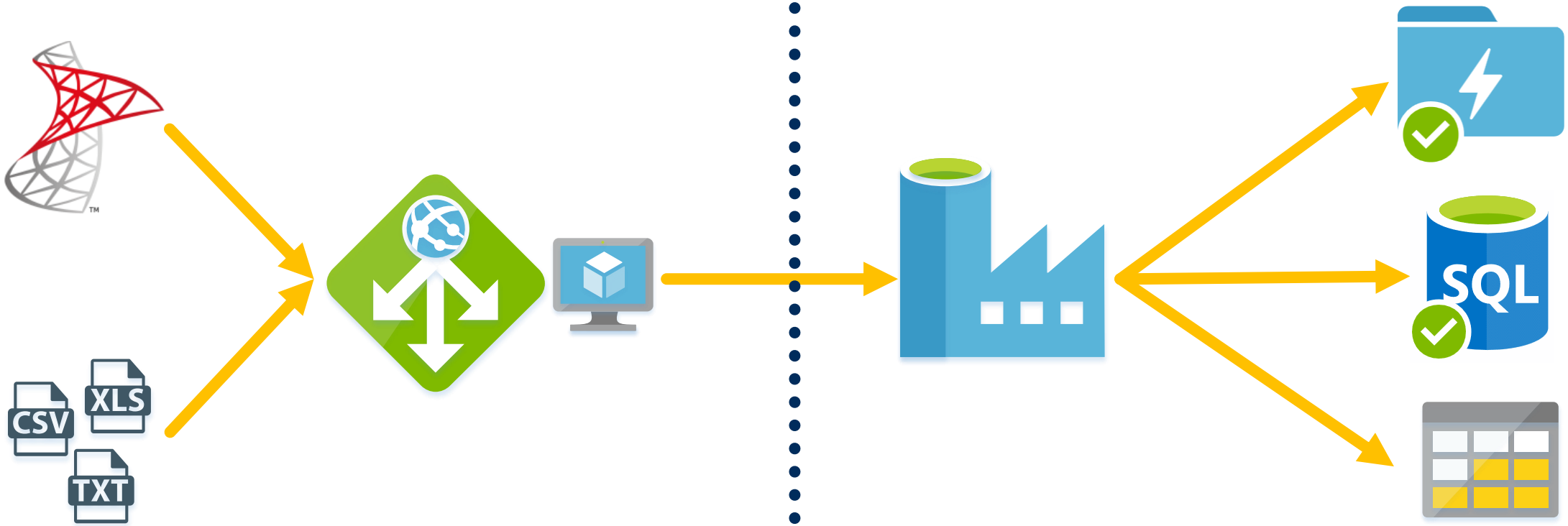
# The Hosted Integration Runtime




*Failover & Load Balancing

**On Premises** **Azure**

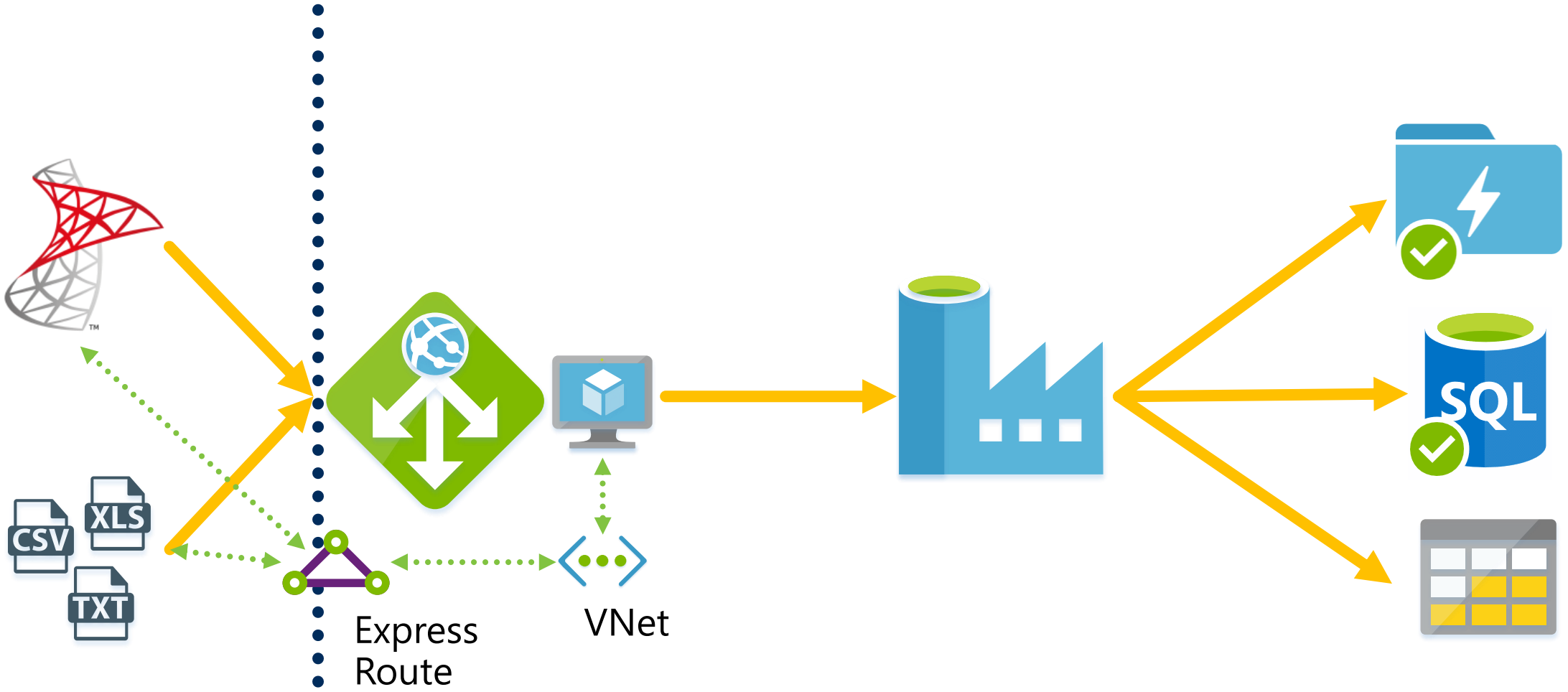
# The Hosted Integration Runtime




**On Premises** **Azure**

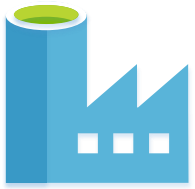
*Failover & Load Balancing

# The Hosted Integration Runtime with Express Route



# Agenda

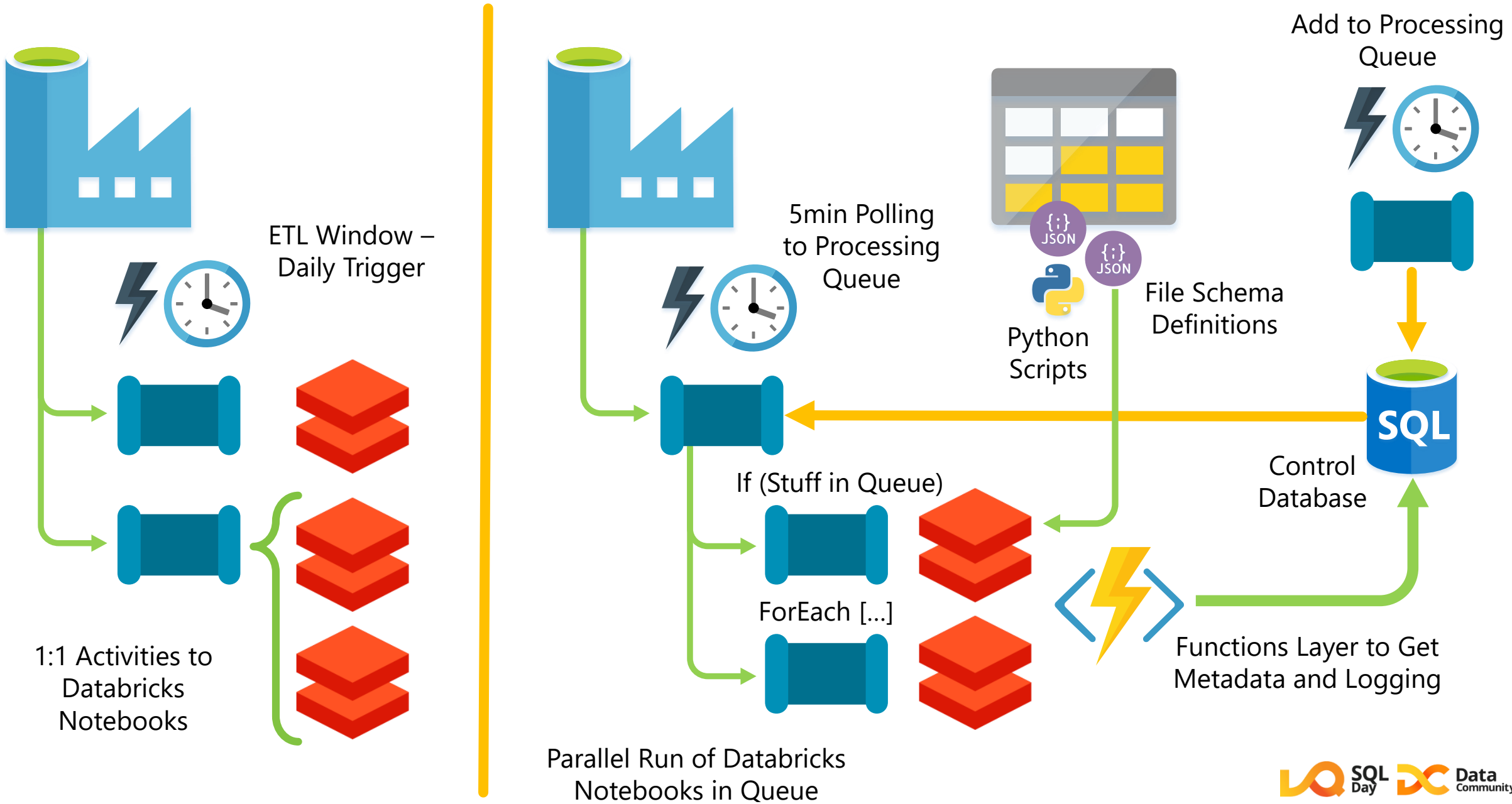
# Complex Orchestration



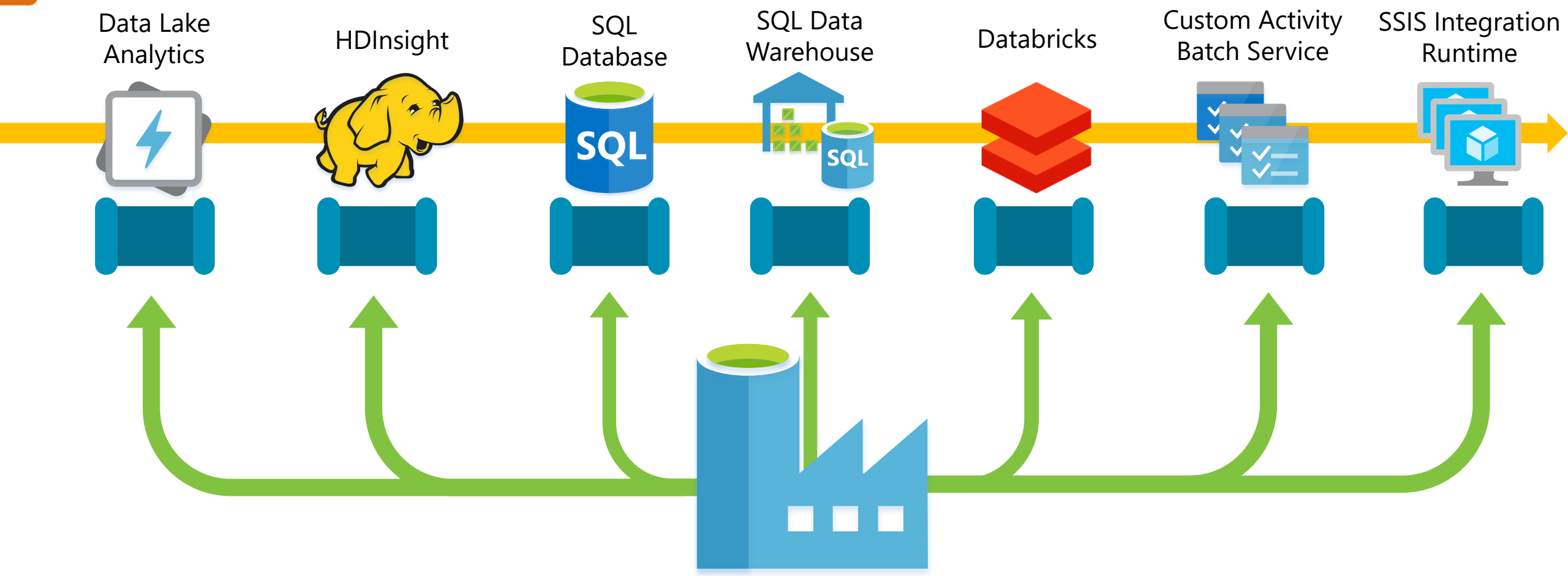
## With Dynamic Data Factory Pipelines


| Azure Data Factory | Extensibility & Parallelism | More Design Patterns |
| --- | --- | --- |
| A very quick overview | Custom Activities<br>SSIS IR & Packages | Bootstrapping<br>Hosted IR vs IaaS<br>Frameworks |

# Framework Processing with Dynamic Pipelines



# Data Transformation in Azure




**Future Uncertain**

**Expensive Clusters**

**Only Scales Up**

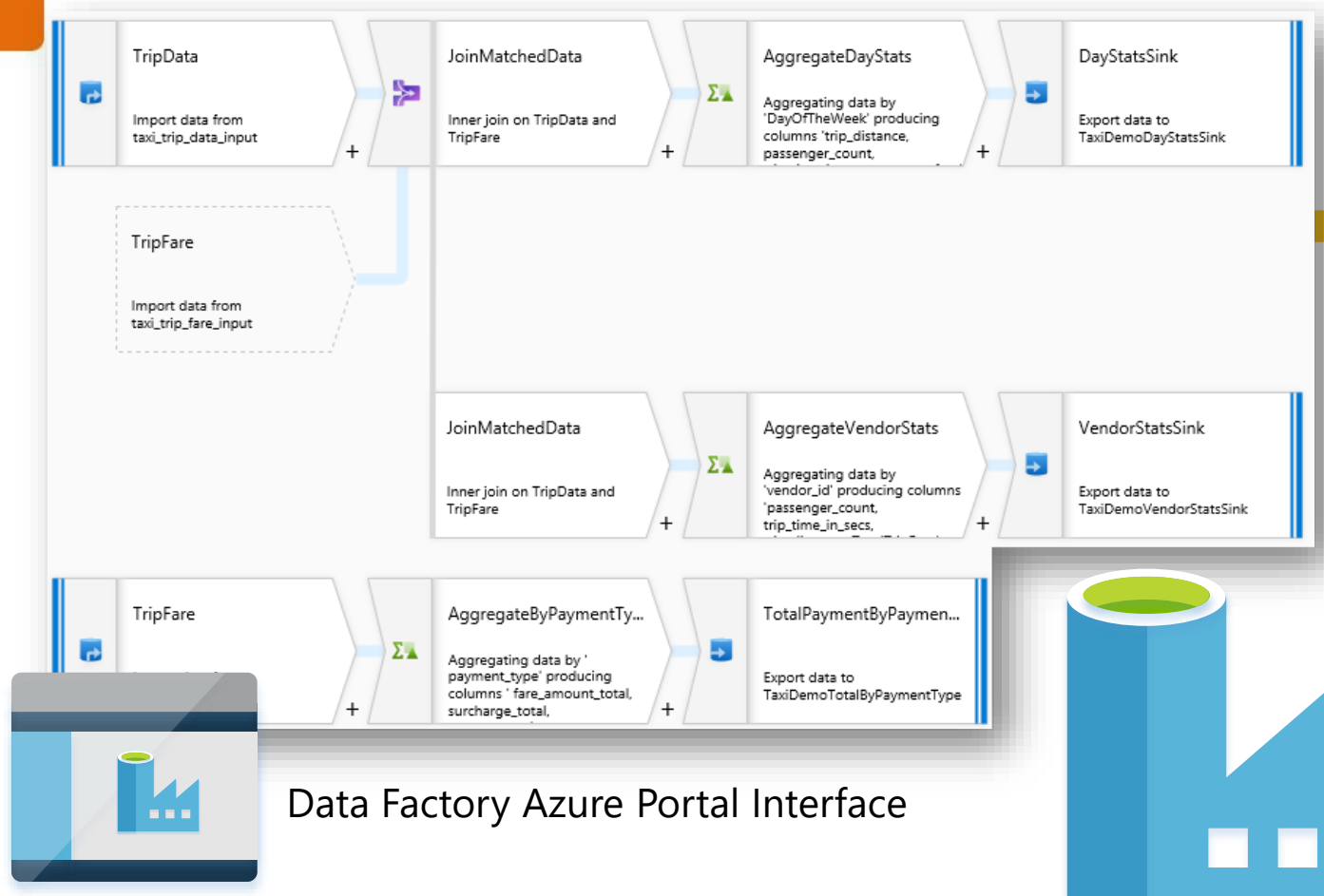
**Requires at least 1TB**

**Learn Spark, Python/Scala**

**Custom Apps on IaaS**

**IaaS VMs 20min start**

# What is a Mapping Data Flow?




Data Factory Azure Portal Interface

Mapping Data Flow

Azure Databricks

Data Factory

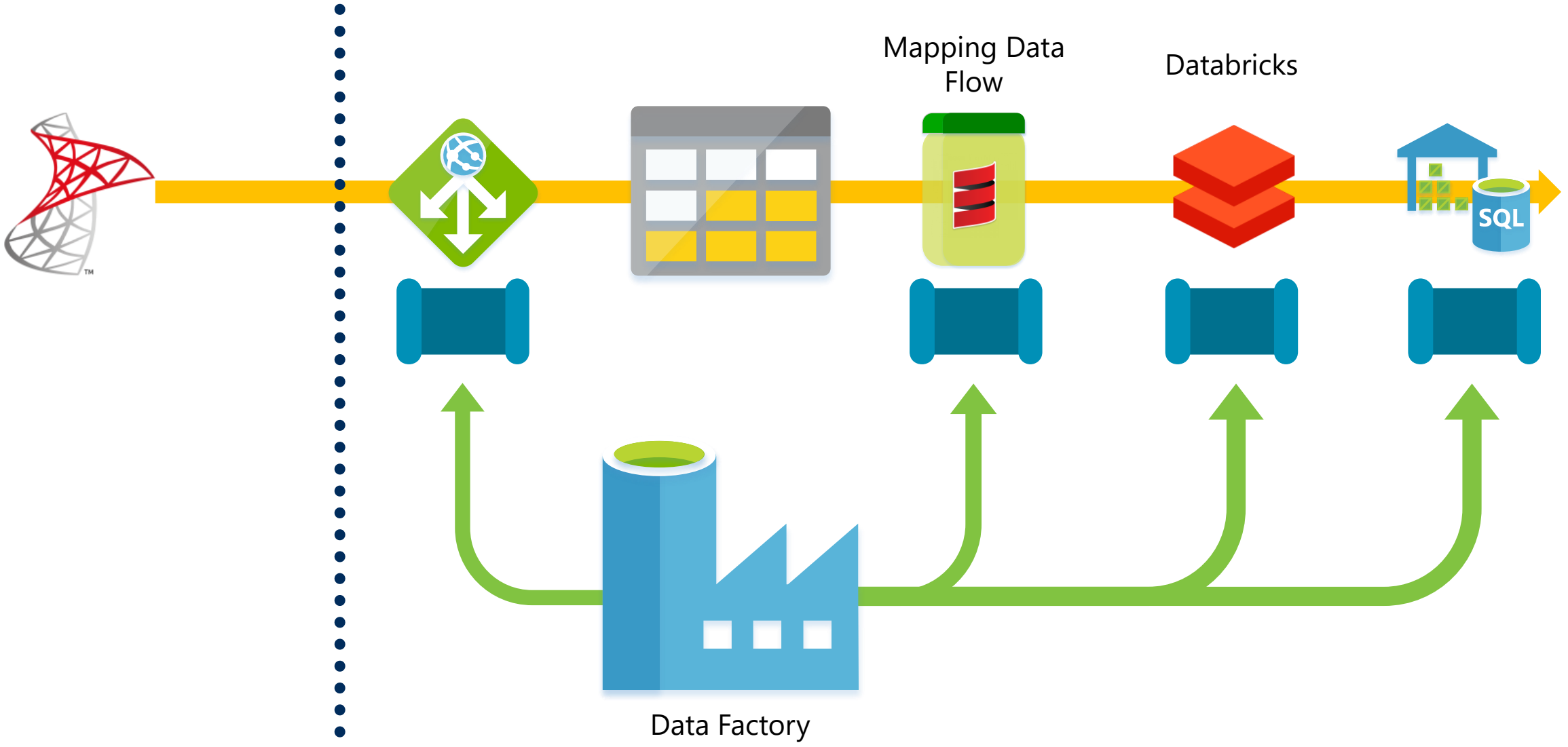
Databricks VNet

Express Route

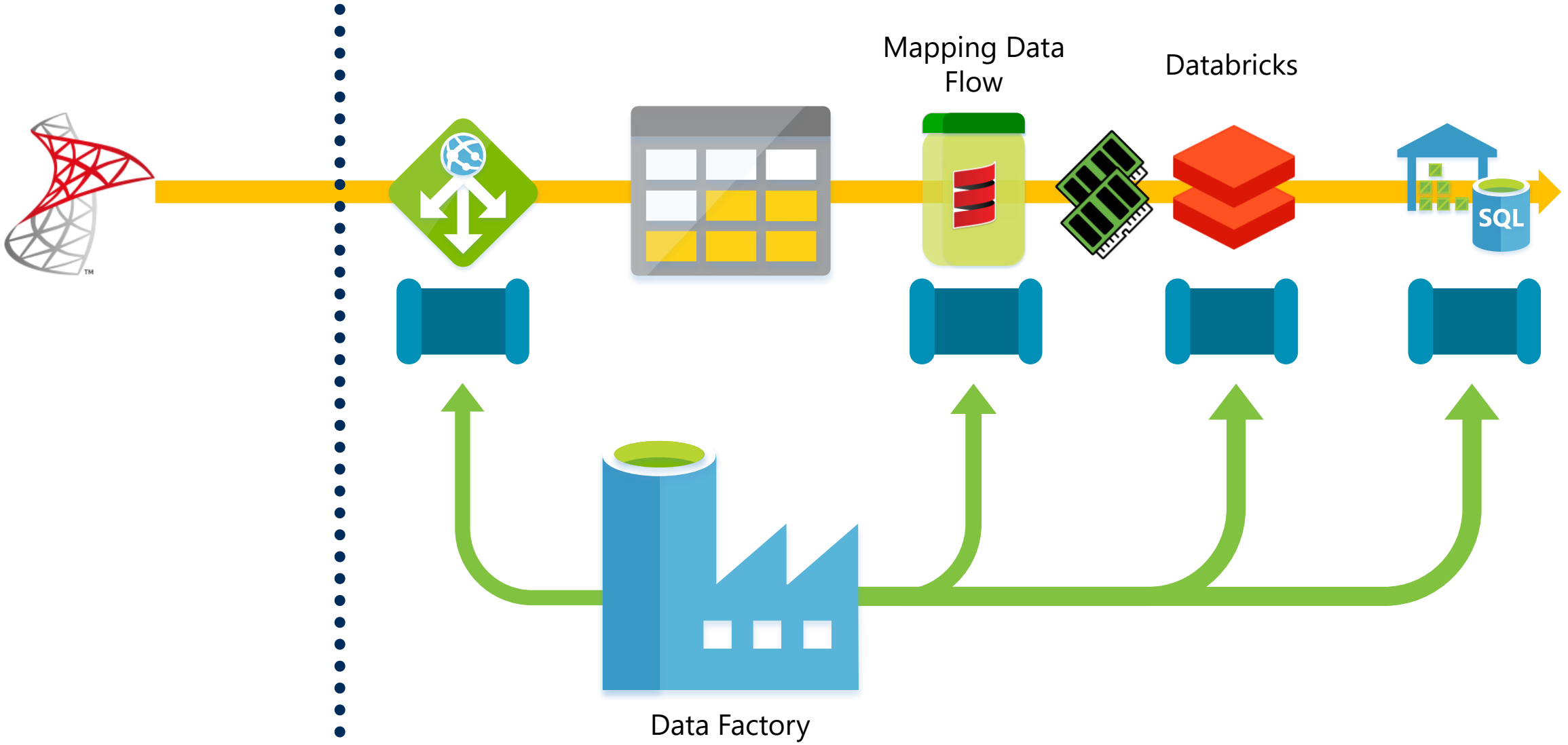
Production VNet

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

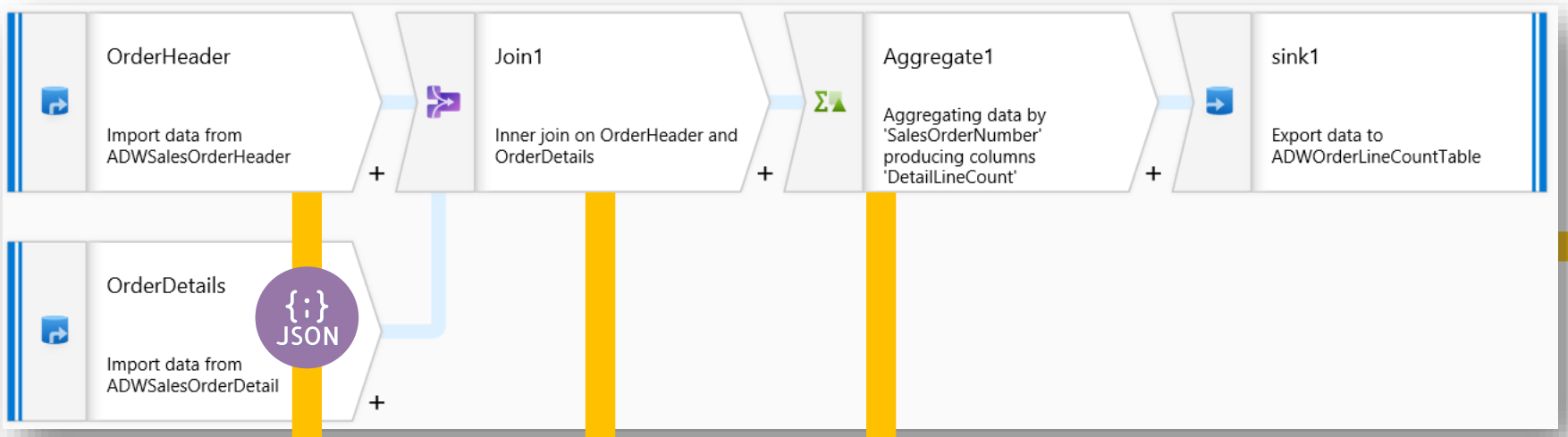
# Mapping Data Flow Future Design Patterns ???



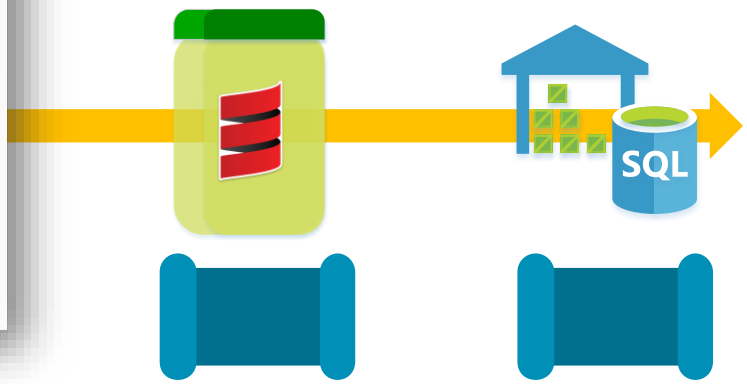
# Mapping Data Flow Future Design Patterns ???

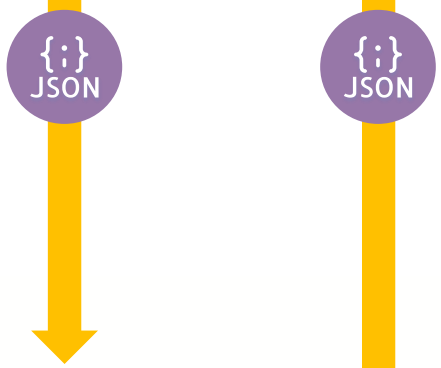


# Mapping Data Flow Future Design Patterns ???



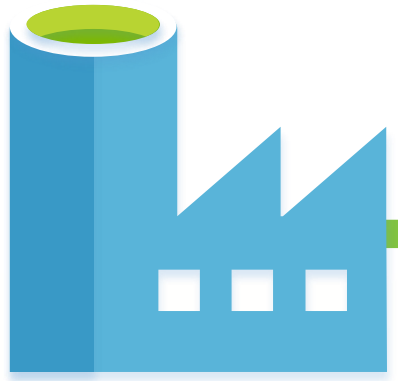

Mapping Data Flow





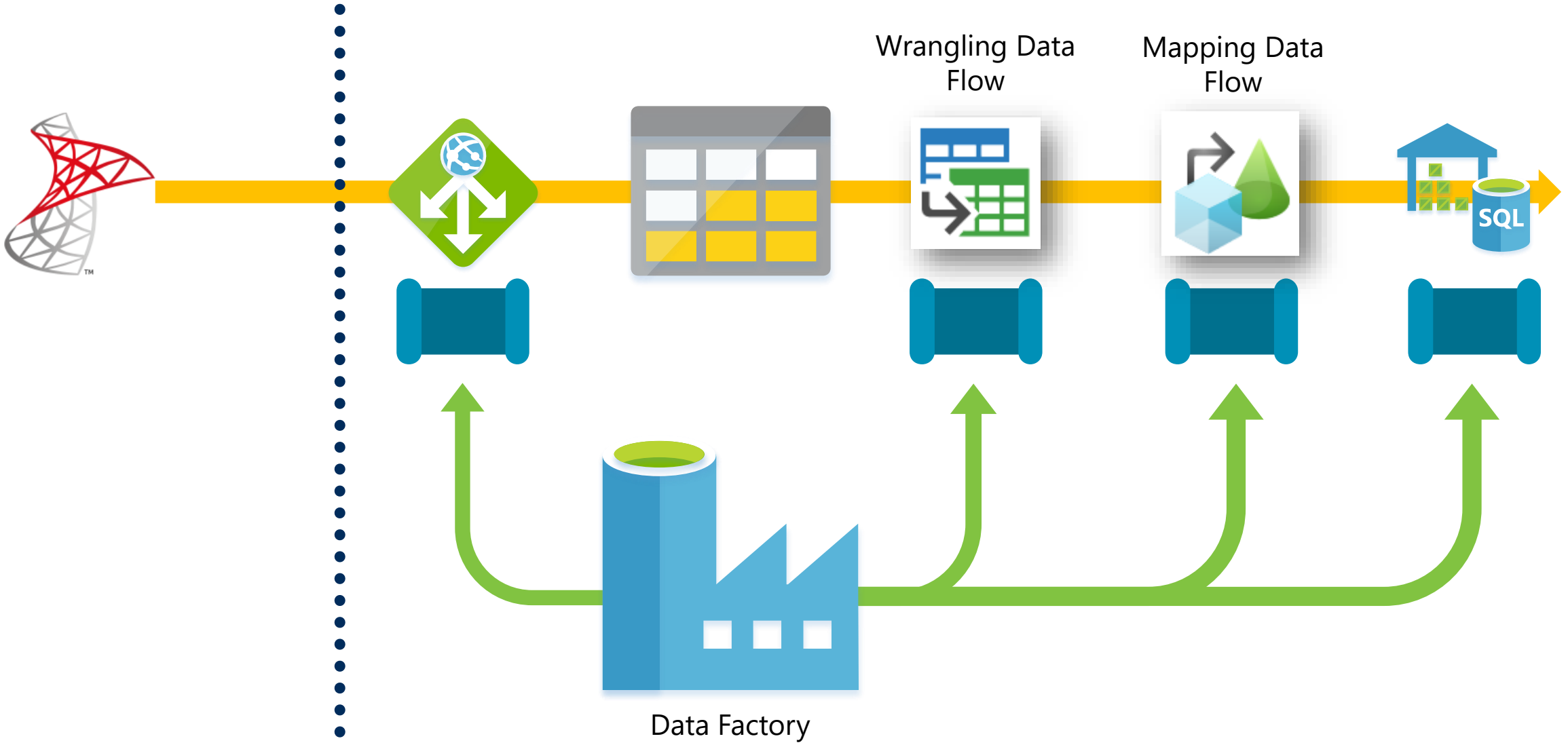




```
"fileName": {
    "value": "@dataset().FileName",
    "type": "Expression"
},
"folderPath": {
    "value": "@dataset().SourceDIR",
    "type": "Expression"
}
```

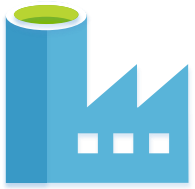
```
"transformations": [
    {
        "name": "Join1",
        "script": "OrderHeader, OrderDetail join(OrderHeader@SalesOrderID == Orde etail@SalesOrderID,\n\tjoinType:'inner',\n\tbroadcast: 'none')~> Join1"
    },
    {
        "name": "Aggregate1",
        "script": "Join1 aggregate(groupBy(SalesOrderNumber),\n\tDetailLineCount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

# Future Design Patterns ???



# Agenda

# Complex Orchestration



## With Dynamic Data Factory Pipelines


| Azure Data Factory | Extensibility & Parallelism | More Design Patterns |
| --- | --- | --- |
| A very quick overview | Custom Activities<br>SSIS IR & Packages | Bootstrapping<br>Hosted IR vs IaaS<br>Frameworks |

# Thanks for Listening

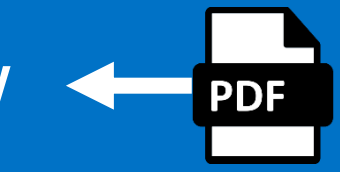

## **Paul Andrew**

@MrPaulAndrew

Microsoft® Most Valuable Professional

adatis

**Blog:** mrpaulandrew.com
**Email:** paul@mrpaulandrew.com
**GitHub:** github.com/mrpaulandrew


Slides