Complex Orchestration

With Dynamic Data Factory Pipelines

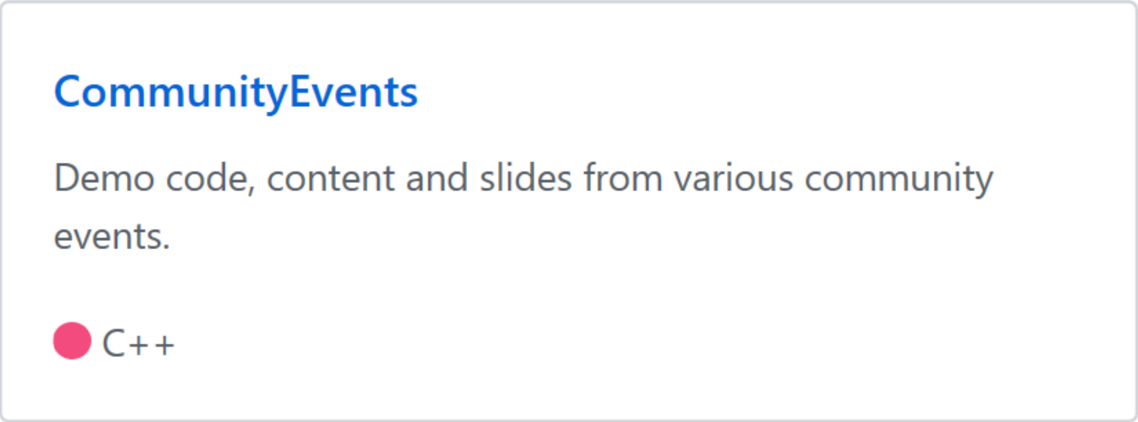
Paul Andrew | Principal Consultant & Solution Architect

Microsoft® Most Valuable Professional

@MrPaulAndrew

In/MrPaulAndrew

https://github.com/mrpaulandrew



CommunityEvents

Demo code, content and slides from various community events.

● C++

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

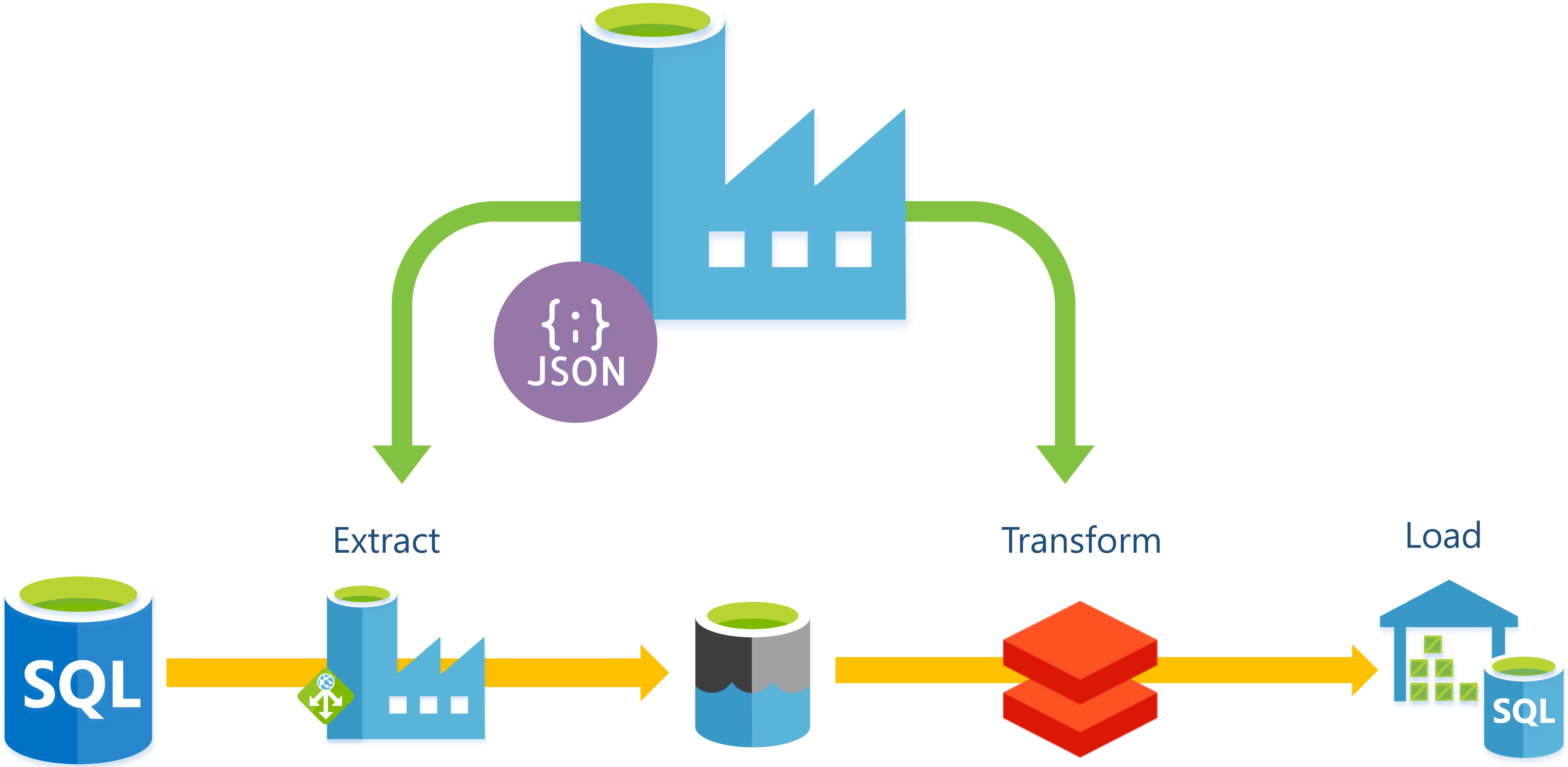
Session Agenda

- Data Factory – A Quick Overview
- Dynamic Pipelines
- Extending Data Factory
  - Web Activities
  - Custom Activities

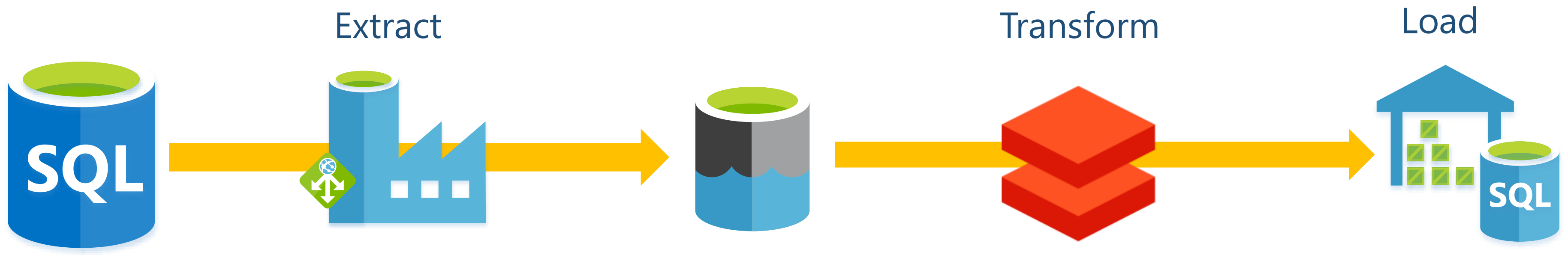
- True Scale Out Execution
  - SSIS Integration Runtime
- Data Factory – In Production
  - Bootstrapping
  - DevOps

Azure Data Factory
A Quick Overview

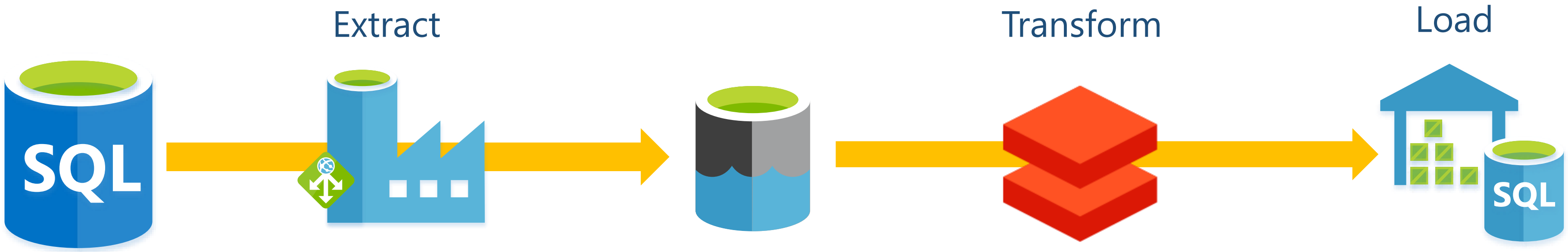
What is Azure Data Factory?



What is Azure Data Factory?



Data Factory Components



1 **Linked Services** ✓

2 **Data Sets** ✓

3 **Activities** ✓

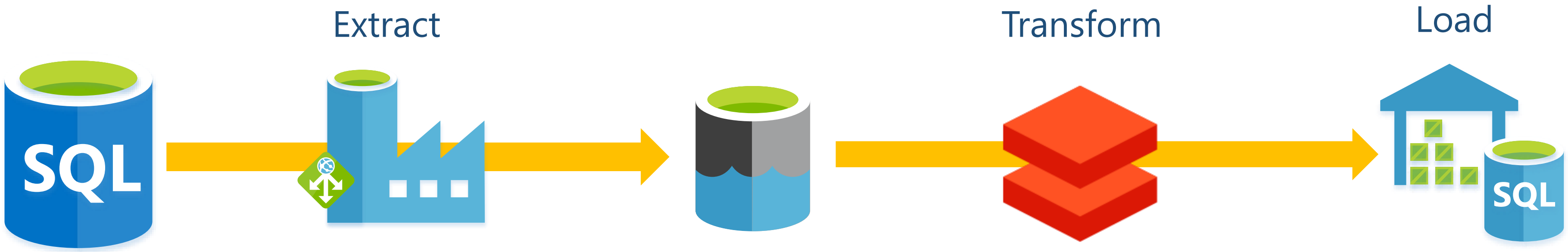
4 **Pipelines** ✓

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}"
    }
}
```

Data Factory Components



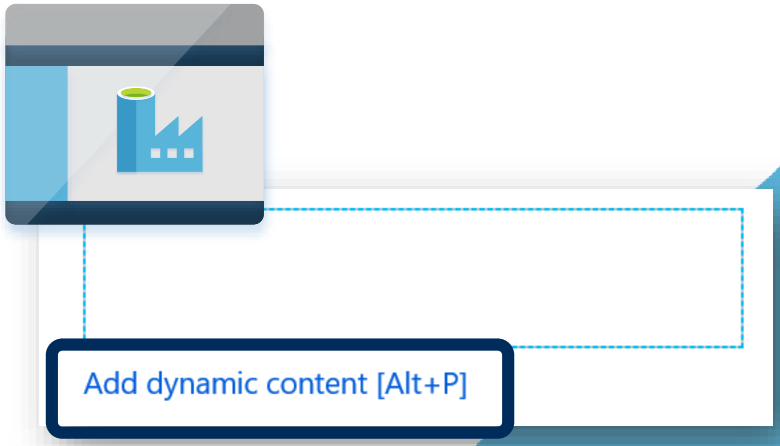
1. **Linked Services** ✓
2. **Data Sets** ✓
3. **Activities** ✓
4. **Pipelines** ✓
5. **Triggers** ✗

{;} JSON

Expression Builder

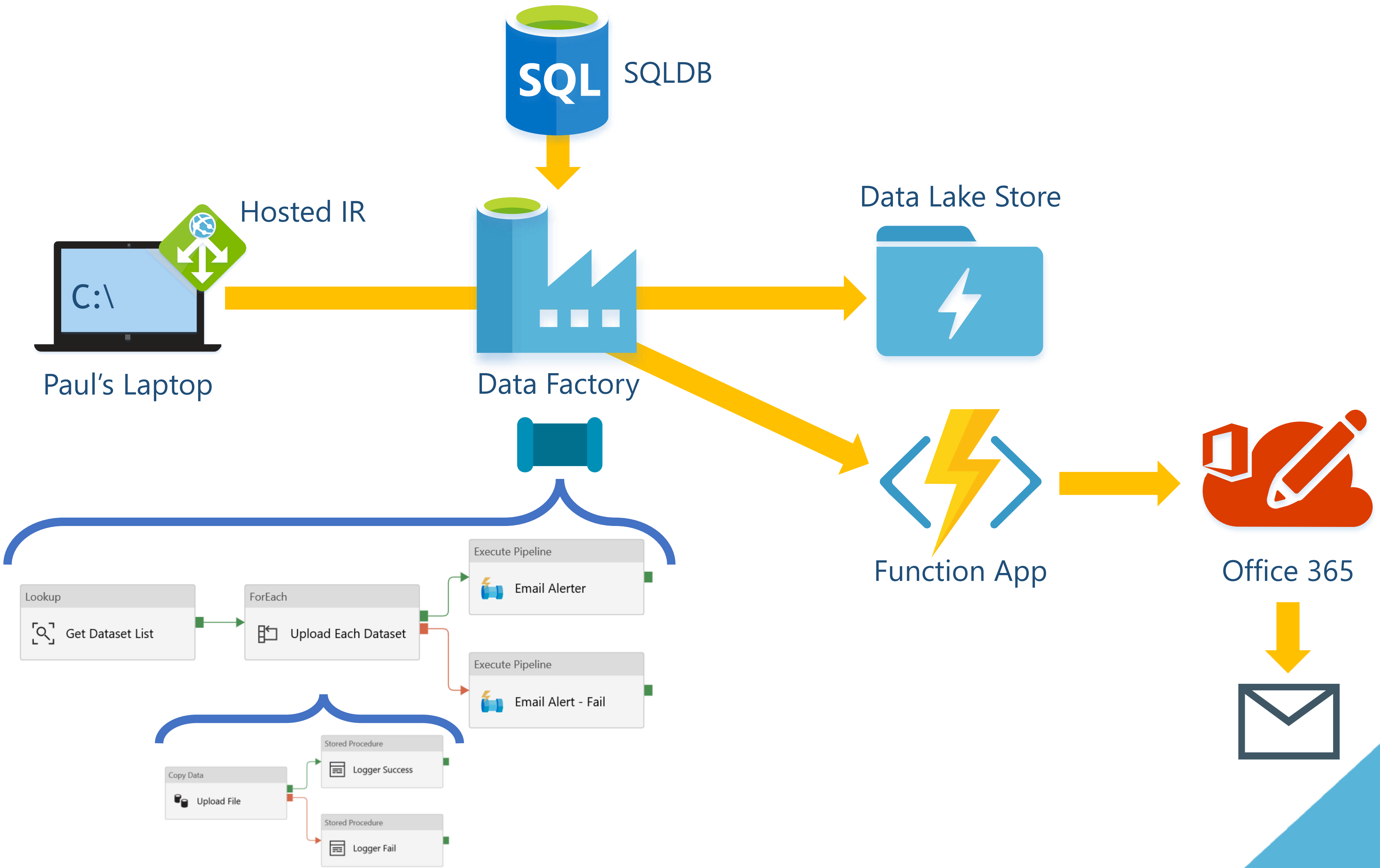
@{.........} ← Parameters
System Variables

- Collection
- Conversation
- Date
- Logical
- Math
- String

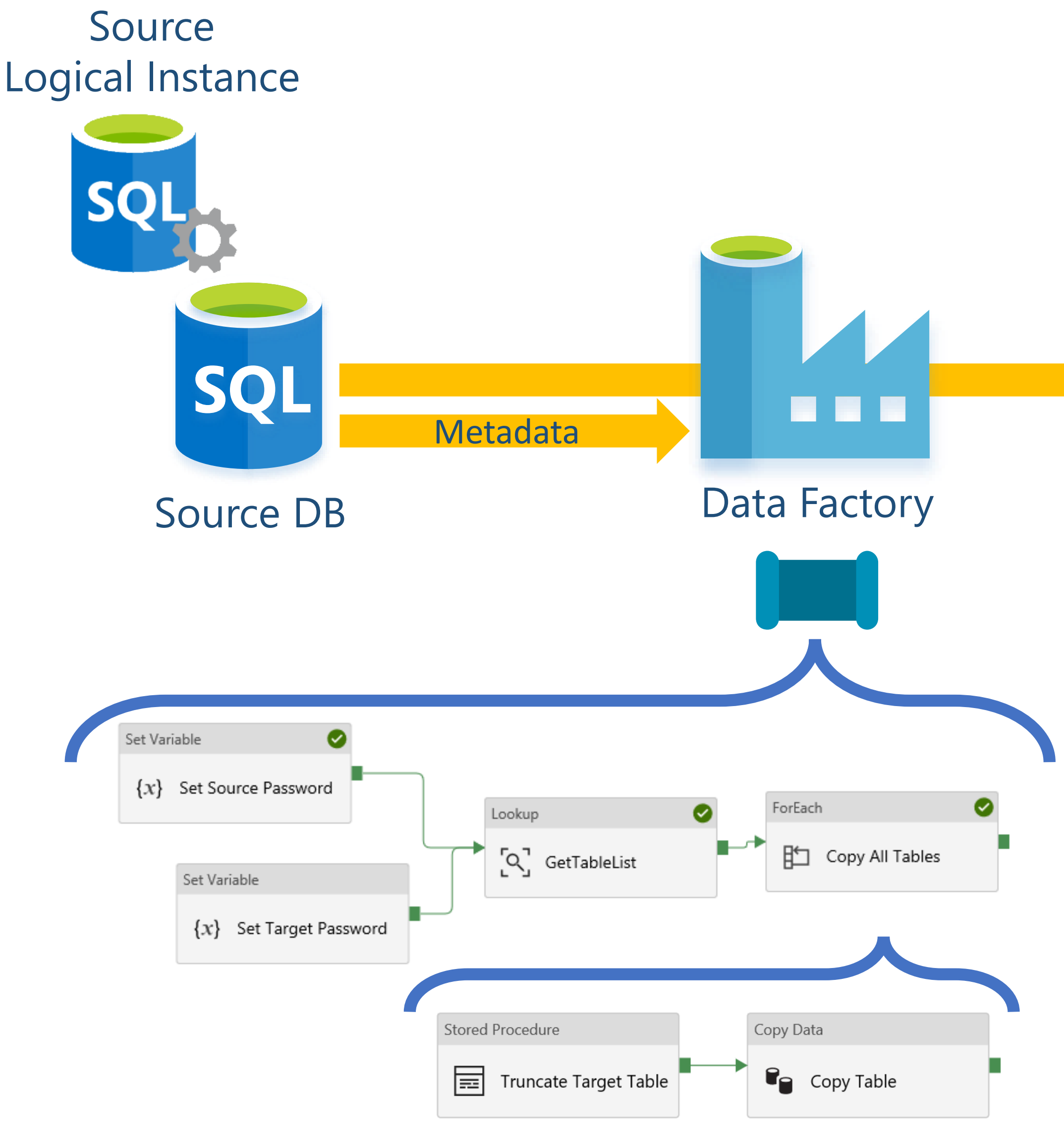

Add dynamic content [Alt+P]

Dynamic Data Factory Pipelines

Demo Architecture 1



Demo Architecture 2



Source
Logical Instance

SQL

SQL

Metadata

Source DB

Data Factory

Target
Logical Instance

SQL

SQL

Target DB

Set Variable
{x} Set Source Password

Set Variable
{x} Set Target Password

Lookup
GetTableList

ForEach
Copy All Tables

Stored Procedure
Truncate Target Table

Copy Data
Copy Table

{;}
JSON

1x Linked Service

1x Dataset

Web Activities vs Web Hook Activities

Web Hook vs Web Activity

Web	Web Hook
PUT POST GET DELETE	POST
1 Minute Timeout	Configurable Timeout
Retry Capabilities	No Retry
Linked Services Datasets	No Artifact Support
One Way Call	Call Back URL

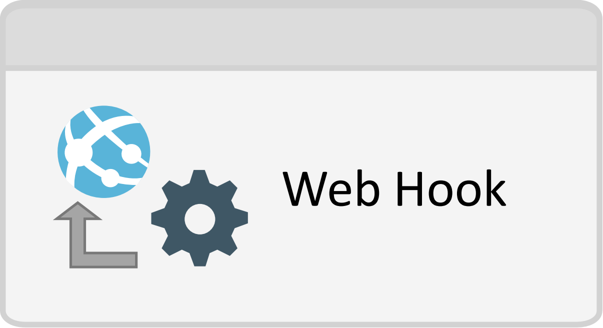
https://mrpaulandrew.com/2019/06/18/azure-data-factory-web-hook-vs-web-activity/

Web Hook vs Web Activity

Asynchronous Web	**Synchronous** Web Hook
PUT POST GET DELETE	POST
1 Minute Timeout	Configurable Timeout
Retry Capabilities	No Retry
Linked Services Datasets	No Artifact Support
One Way Call	Call Back URL

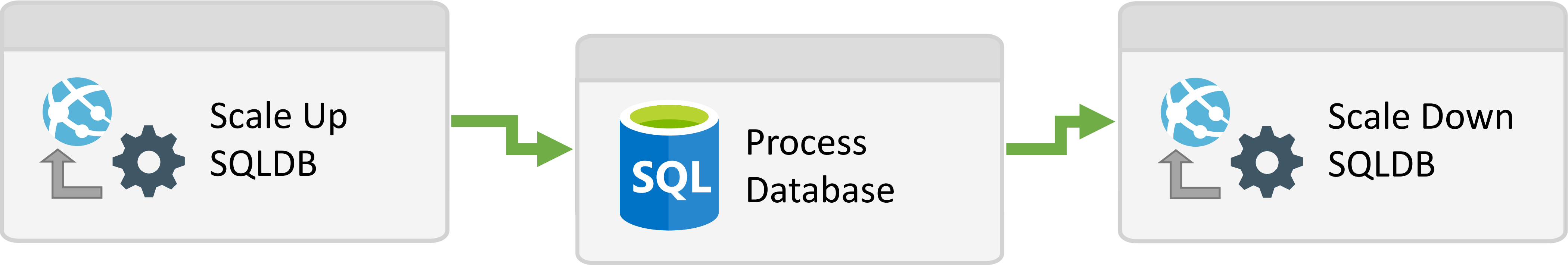
https://mrpaulandrew.com/2019/06/18/azure-data-factory-web-hook-vs-web-activity/

Web Hook vs Web Activity



Web Hook

Web Hook vs Web Activity

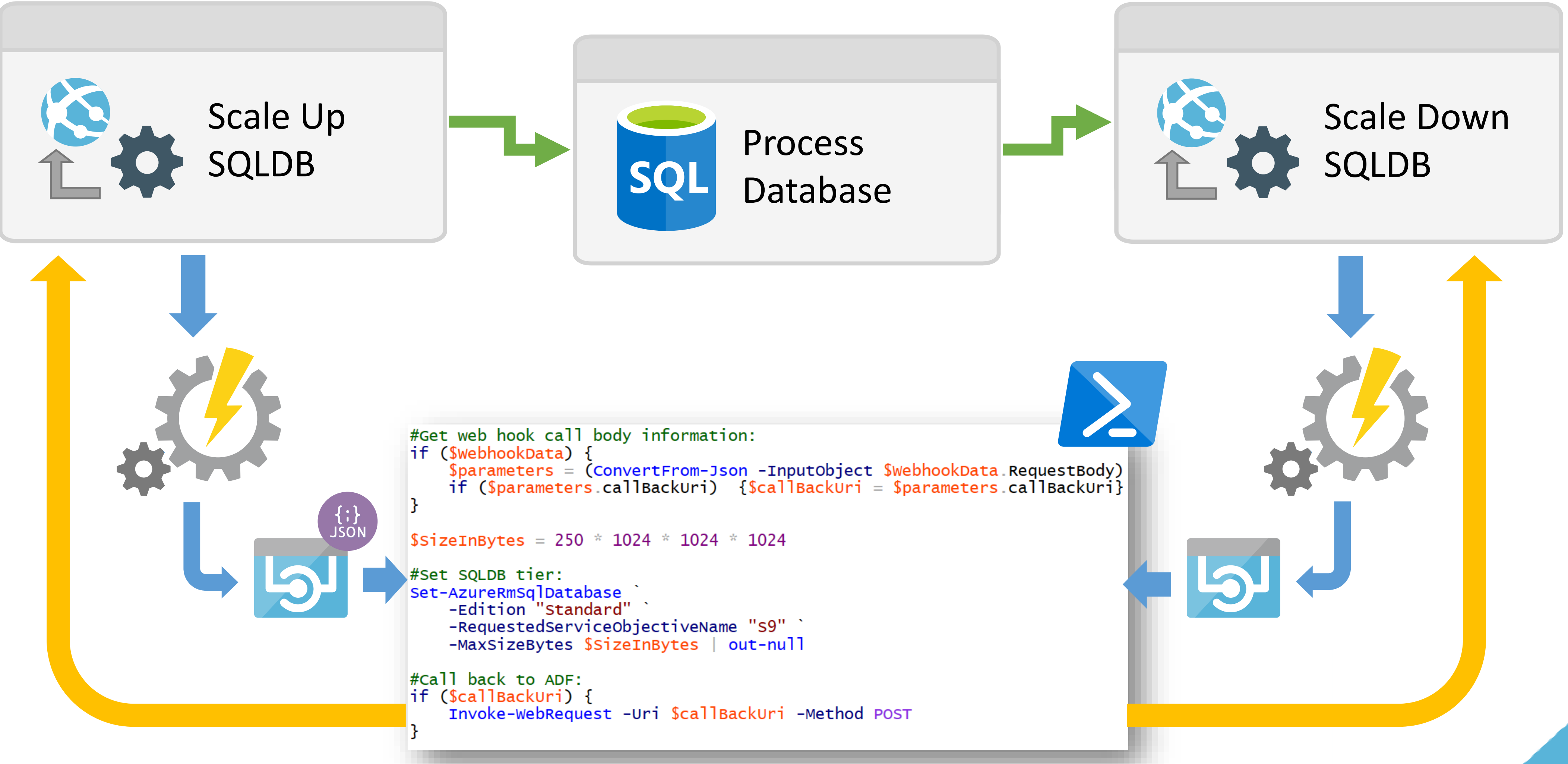


Scale Up SQLDB

Process Database

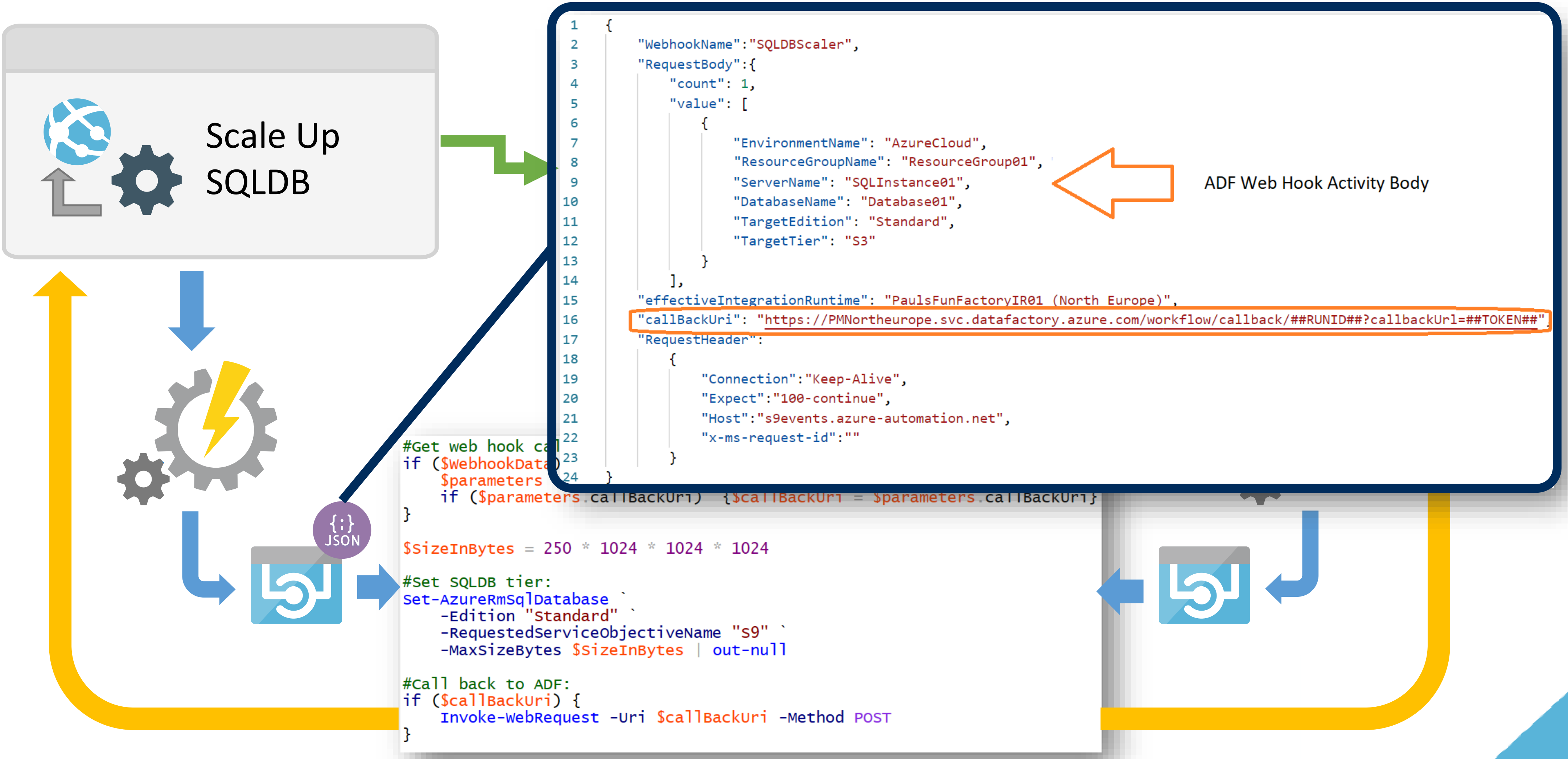
Scale Down SQLDB

Web Hook vs Web Activity



```
#Get web hook call body information:
if ($WebhookData) {
    $parameters = (ConvertFrom-Json -InputObject $WebhookData.RequestBody)
    if ($parameters.callBackUri)  {$callBackUri = $parameters.callBackUri}
}

$SizeInBytes = 250 * 1024 * 1024 * 1024

#Set SQLDB tier:
Set-AzureRmSqlDatabase `
    -Edition "Standard" `
    -RequestedServiceObjectiveName "S9" `
    -MaxSizeBytes $SizeInBytes | out-null

#Call back to ADF:
if ($callBackUri) {
    Invoke-WebRequest -Uri $callBackUri -Method POST
}
```

Web Hook vs Web Activity



```
{
    "WebhookName":"SQLDBScaler",
    "RequestBody":{
        "count": 1,
        "value": [
            {
                "EnvironmentName": "AzureCloud",
                "ResourceGroupName": "ResourceGroup01",
                "ServerName": "SQLInstance01",
                "DatabaseName": "Database01",
                "TargetEdition": "Standard",
                "TargetTier": "S3"
            }
        ],
    "effectiveIntegrationRuntime": "PaulsFunFactoryIR01 (North Europe)",
    "callBackUri": "https://PMNortheurope.svc.datafactory.azure.com/workflow/callback/##RUNID##?callbackUrl=##TOKEN##"
    "RequestHeader":
        {
            "Connection":"Keep-Alive",
            "Expect":"100-continue",
            "Host":"s9events.azure-automation.net",
            "x-ms-request-id":""
        }
}
```

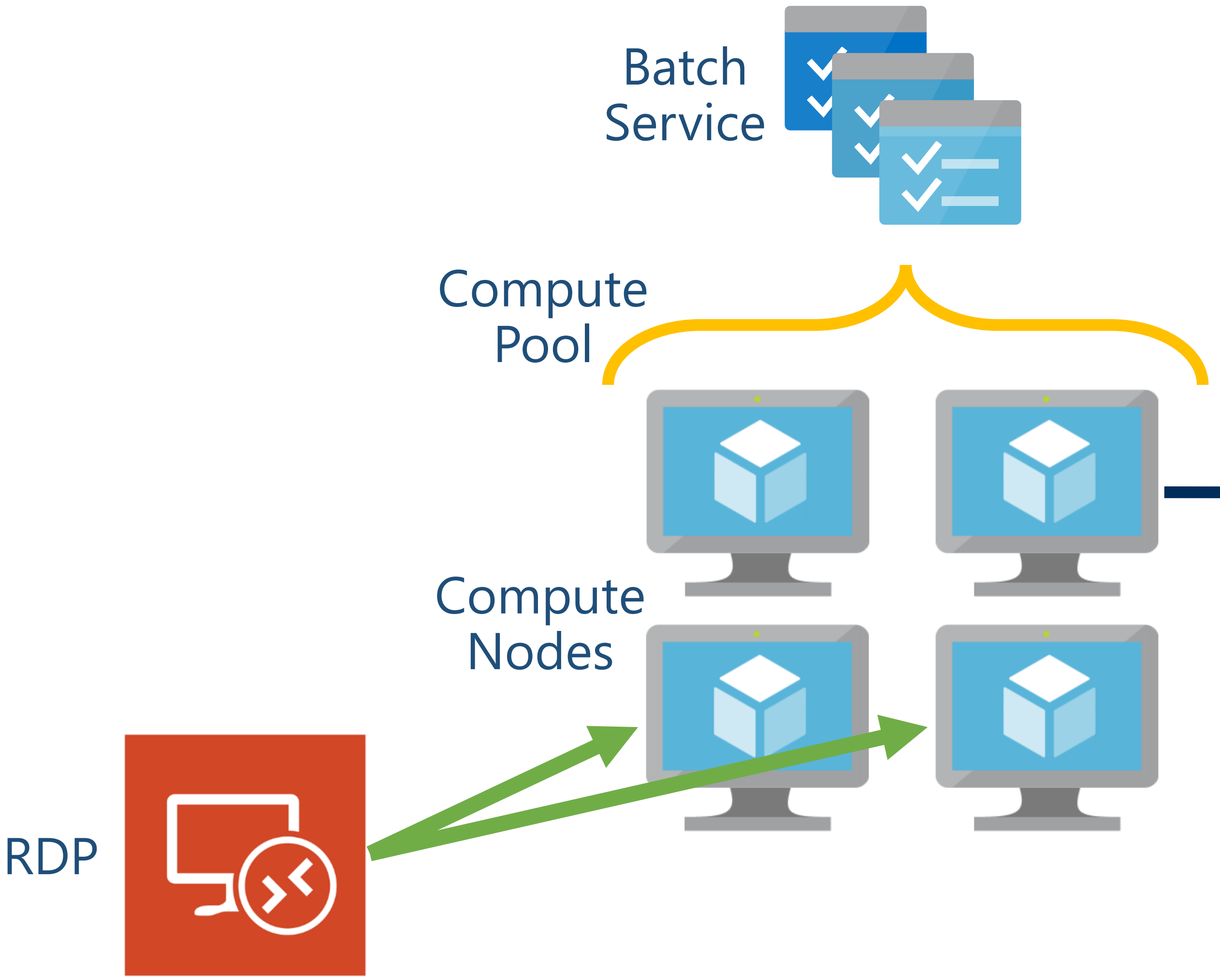
ADF Web Hook Activity Body

```
#Get web hook ca
if ($WebhookData
    $parameters
    if ($parameters.callBackUri) {$callBackUri = $parameters.callBackUri}
}

$SizeInBytes = 250 * 1024 * 1024 * 1024

#Set SQLDB tier:
Set-AzureRmSqlDatabase `
    -Edition "Standard" `
    -RequestedServiceObjectiveName "S9" `
    -MaxSizeBytes $SizeInBytes | out-null

#Call back to ADF:
if ($callBackUri) {
    Invoke-WebRequest -Uri $callBackUri -Method POST
}
```

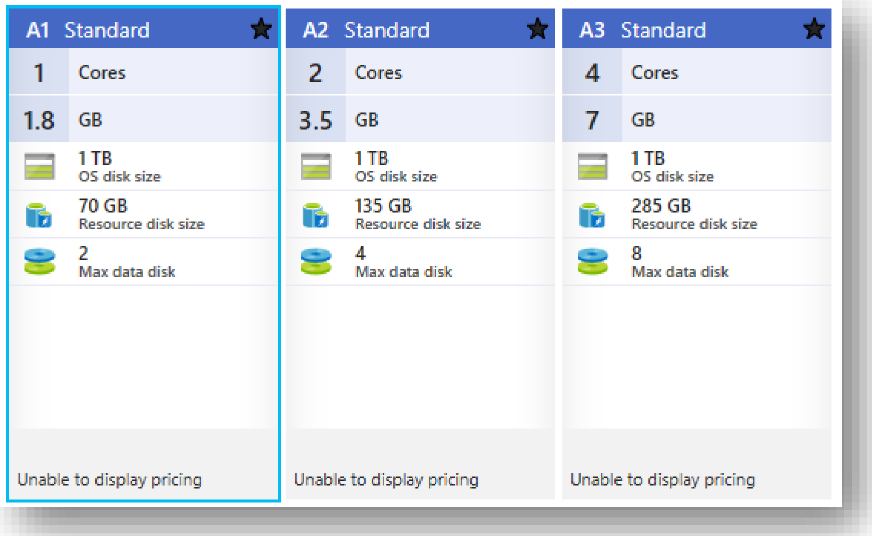
Custom Activities

Azure Batch Service

Scale out compute delivered using PaaS technology with IaaS underneath.



VM node size set per compute pool:

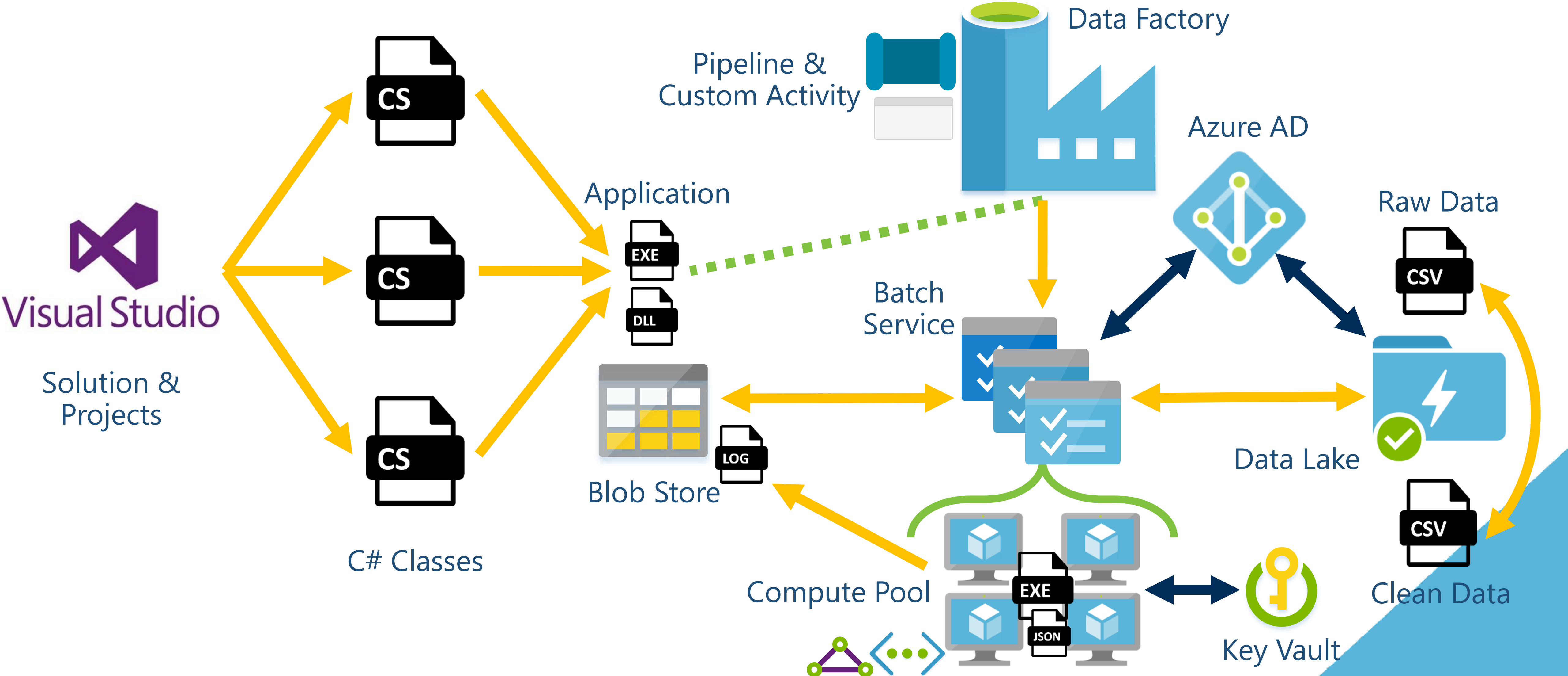


A1 Standard ★	A2 Standard ★	A3 Standard ★
1 Cores	2 Cores	4 Cores
1.8 GB	3.5 GB	7 GB
1 TB OS disk size	1 TB OS disk size	1 TB OS disk size
70 GB Resource disk size	135 GB Resource disk size	285 GB Resource disk size
2 Max data disk	4 Max data disk	8 Max data disk
Unable to display pricing	Unable to display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

Building a Custom Activity

A .Net Console App Executed Using Azure Batch Service.



Extensibility Conclusions

C#

Do Stuff

10 minutes execution

C#

Custom
Activity

Auto scale out compute &
Scale up per compute node

REST

Web

Asynchronous

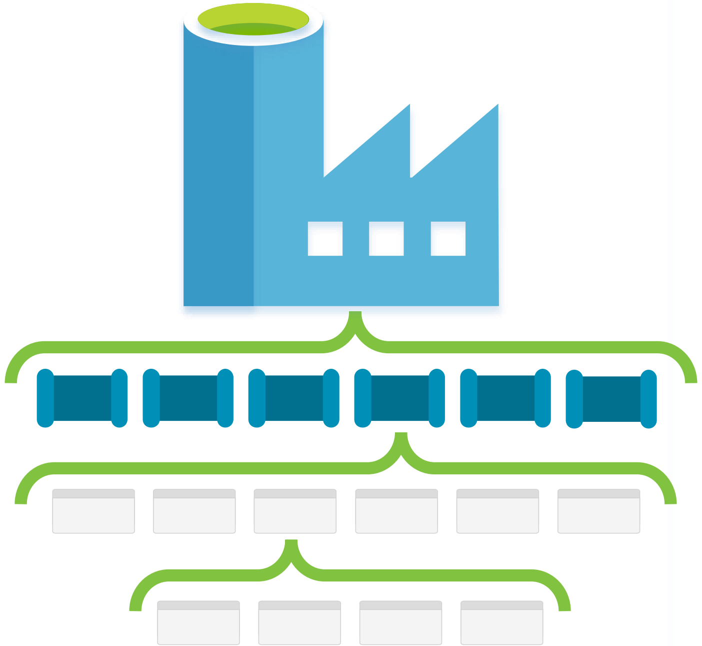
REST

Web Hook

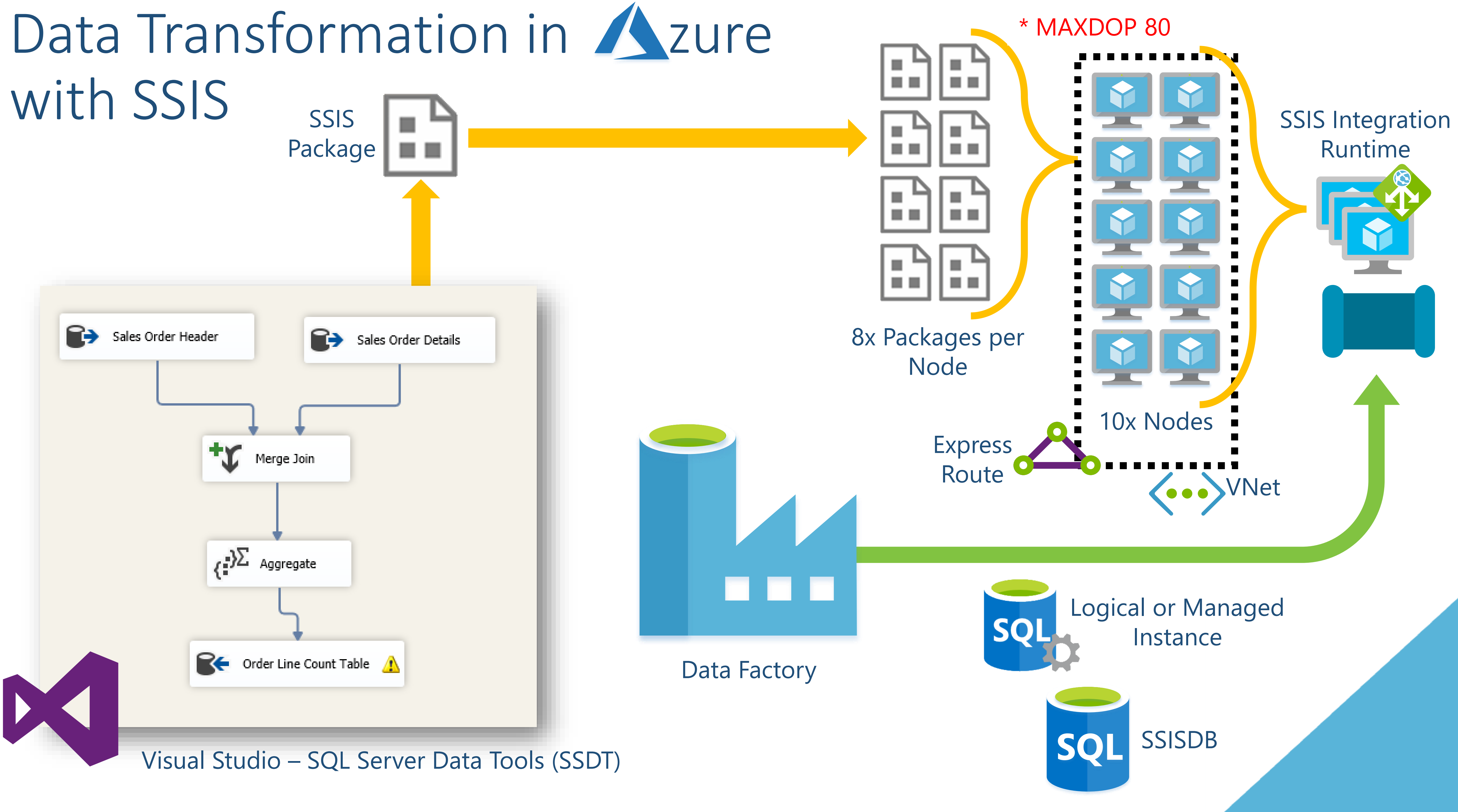
Synchronous
(Call back)

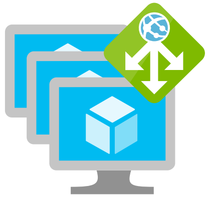
Everything!

Scale Out ~~Execution~~

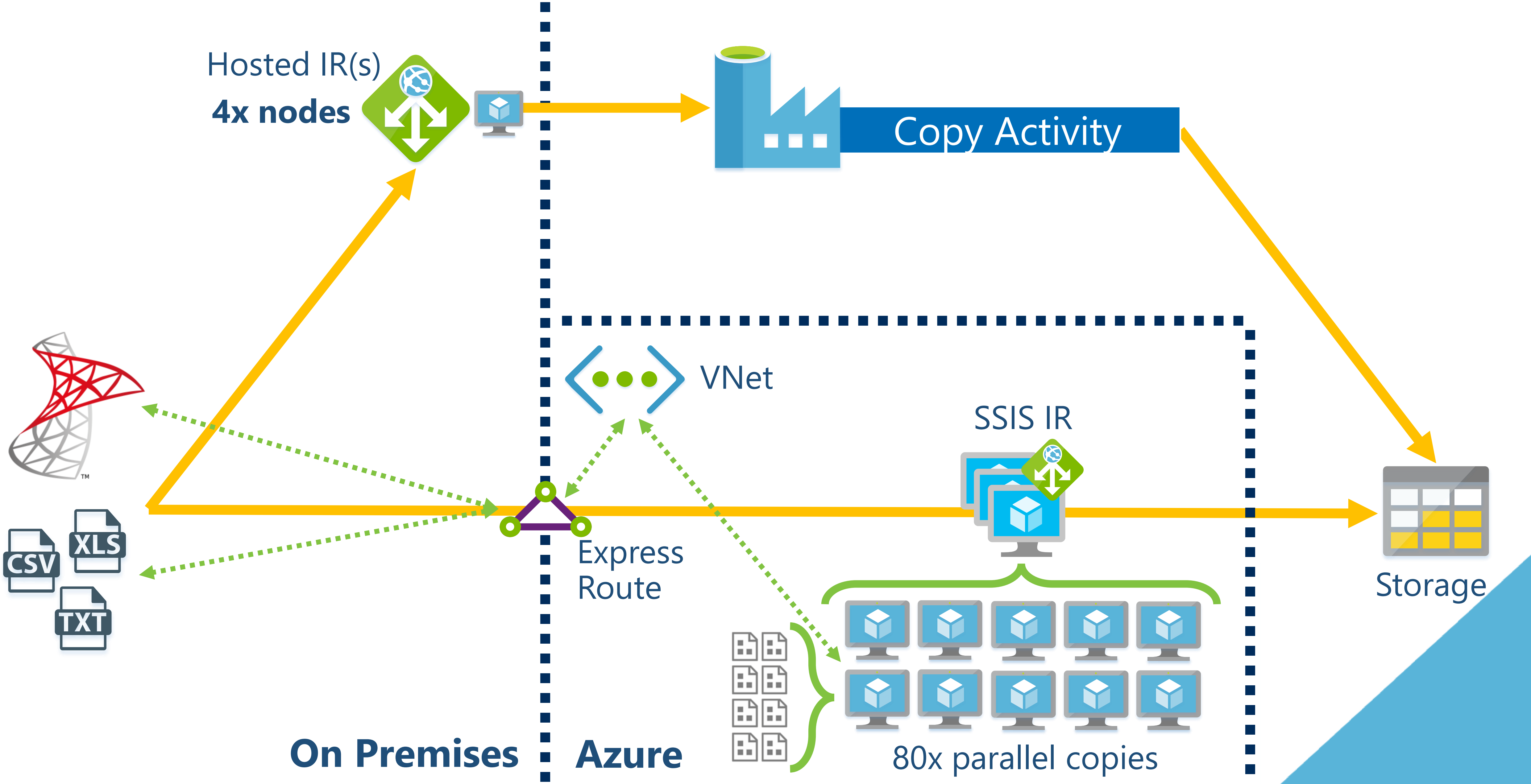


Data Transformation in Azure with SSIS

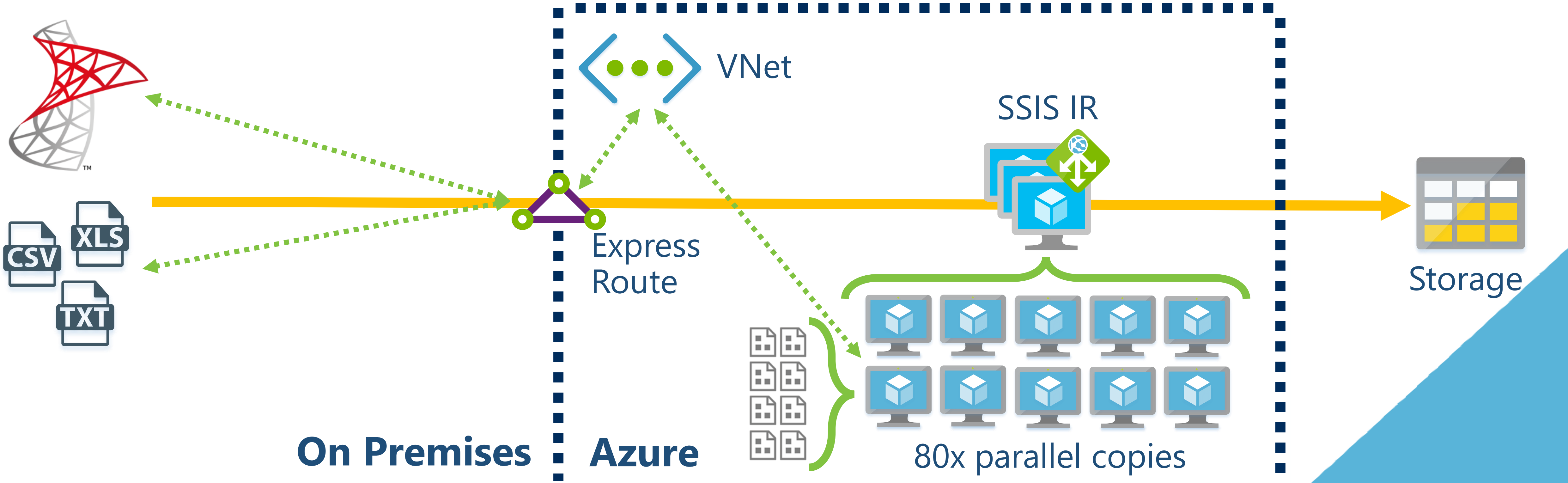




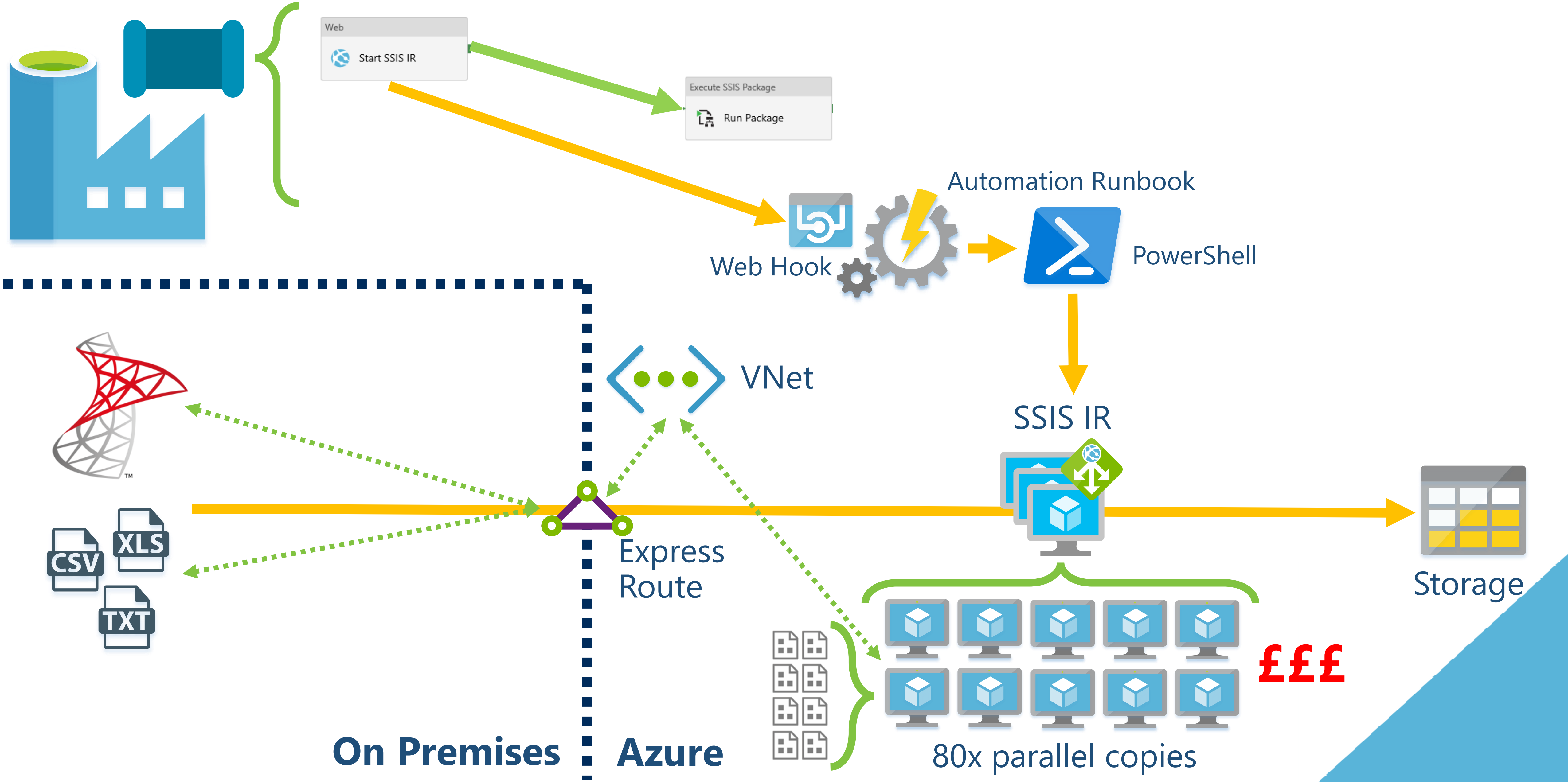
The SSIS IR vs Hosted IR with Express Route



The SSIS IR Start/Stop



The SSIS IR Start/Stop



Web

Start SSIS IR

Execute SSIS Package

Run Package

Automation Runbook

Web Hook

PowerShell

VNet

SSIS IR

Express Route

Storage

CSV

XLS

TXT

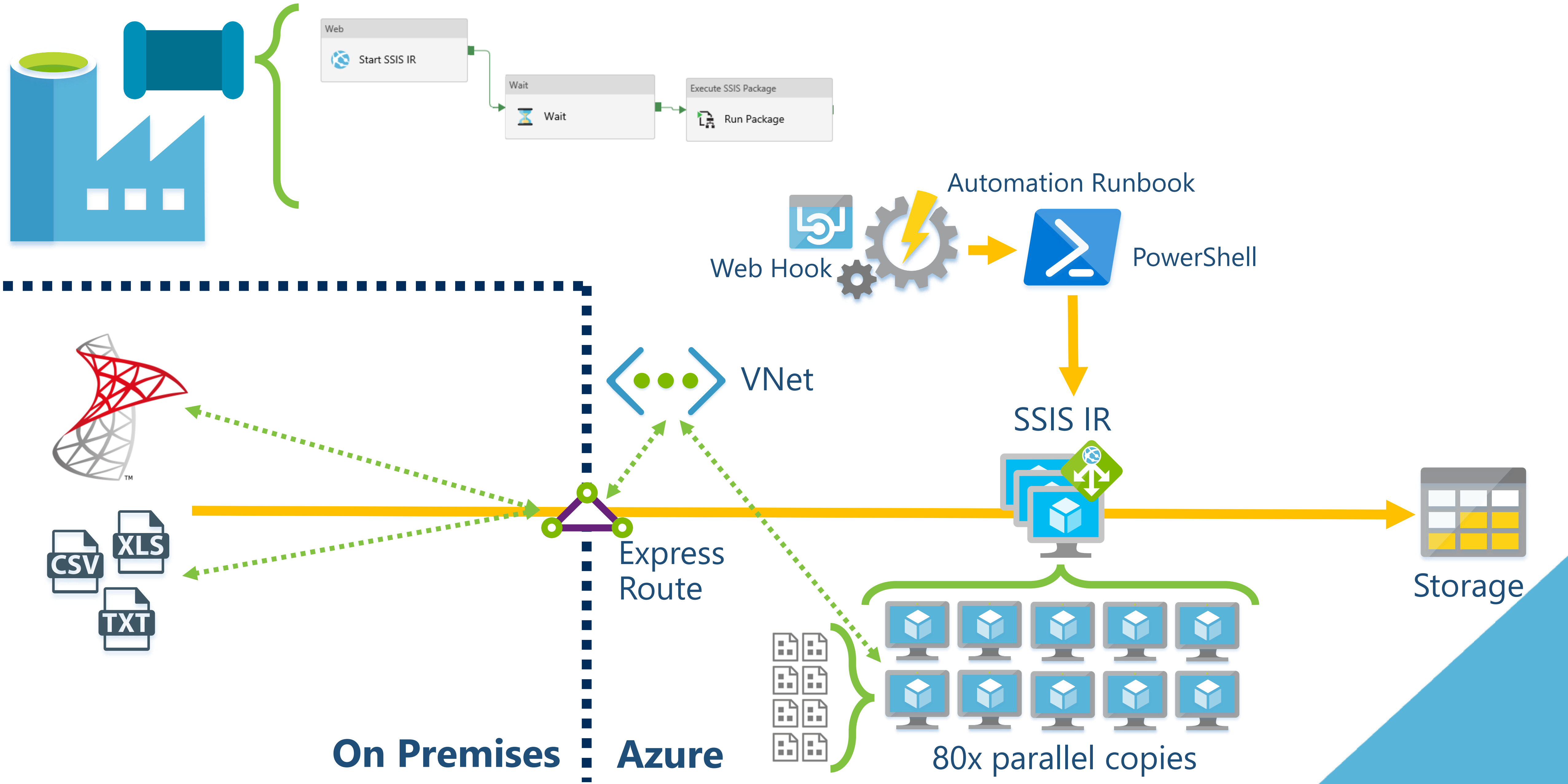
£££

80x parallel copies

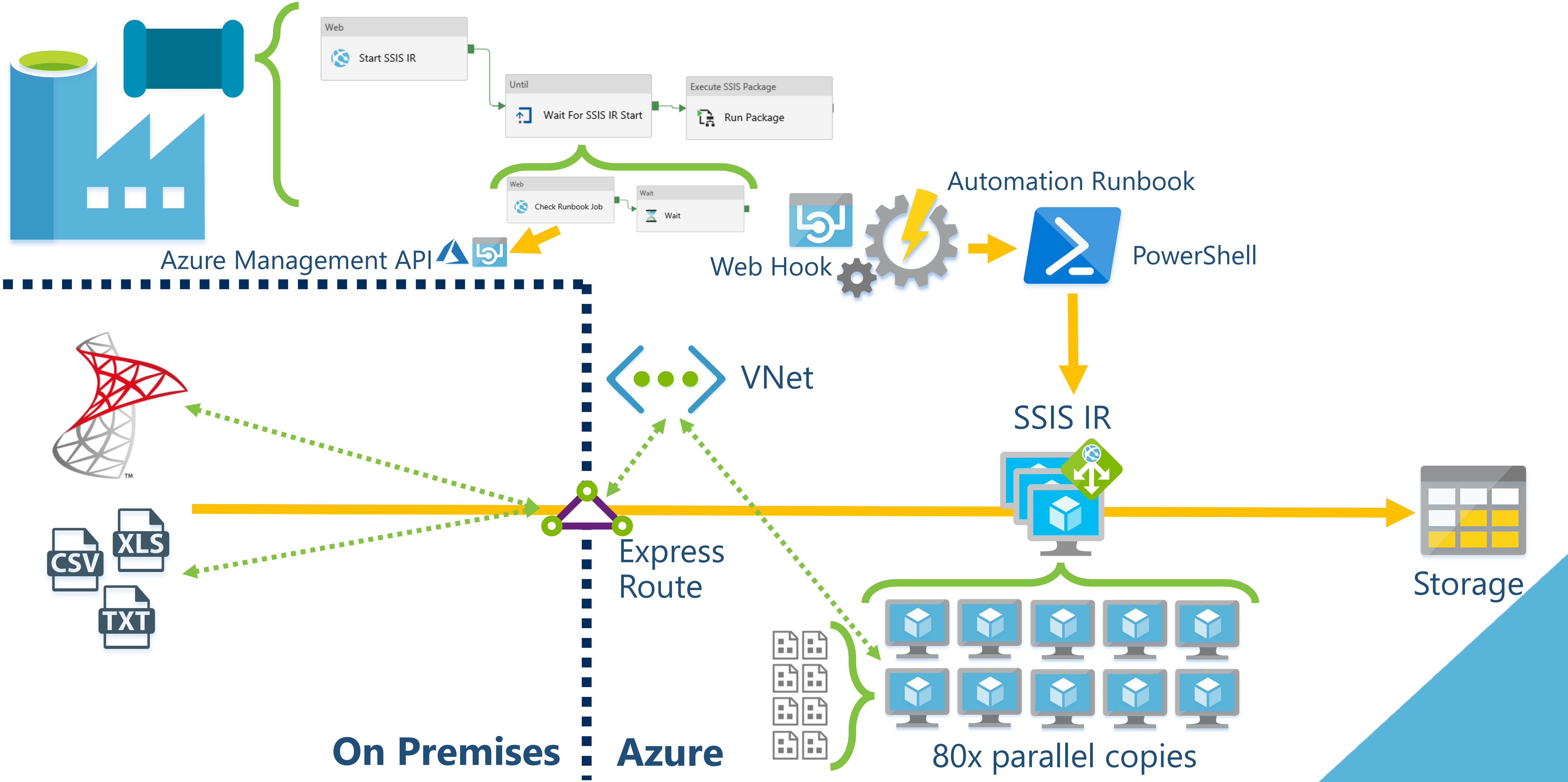
On Premises

Azure

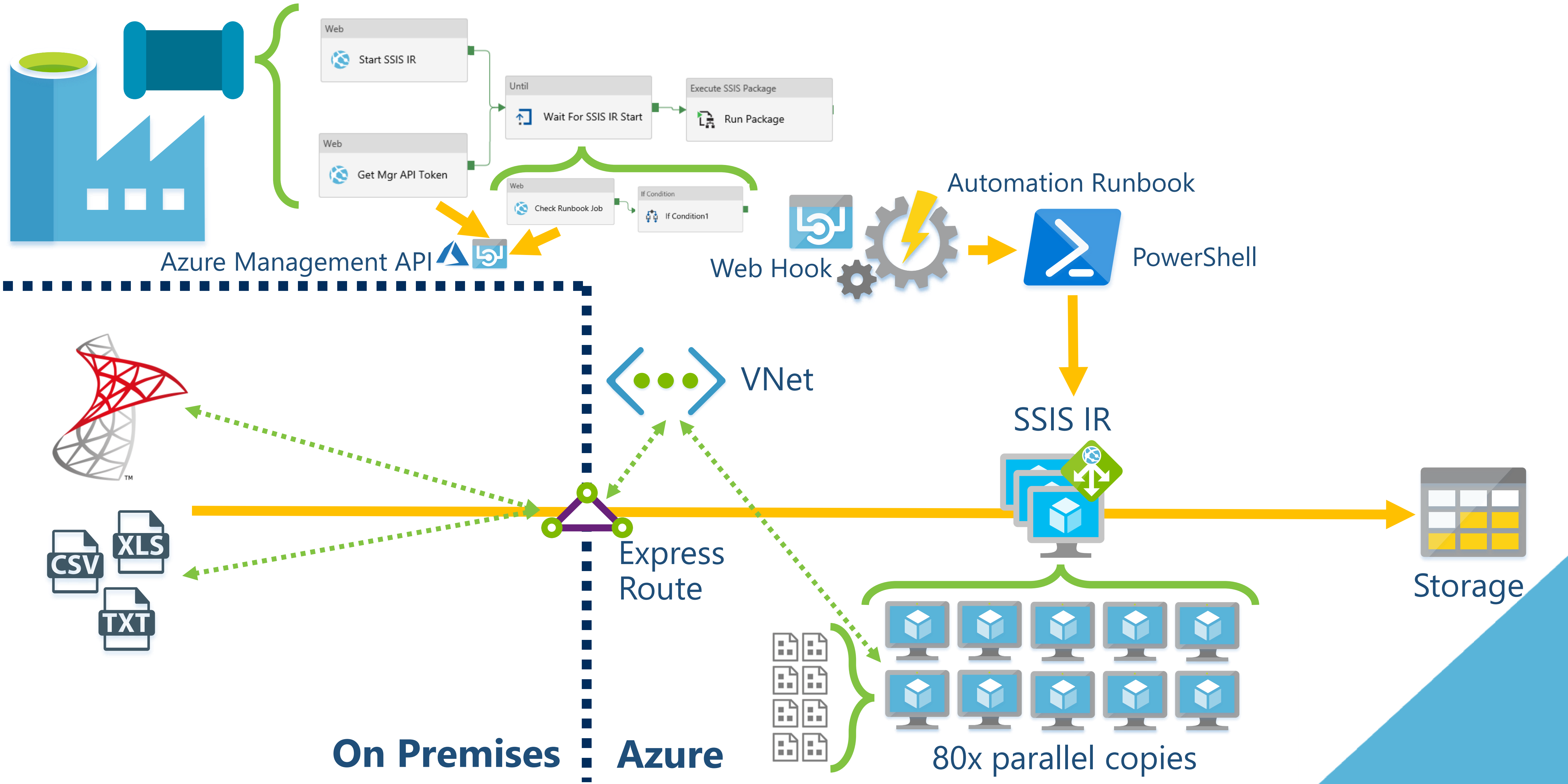
The SSIS IR Start/Stop



The SSIS IR Start/Stop



The SSIS IR Start/Stop



Web

Start SSIS IR

Web

Get Mgr API Token

Until

Wait For SSIS IR Start

Execute SSIS Package

Run Package

Web

Check Runbook Job

If Condition

If Condition1

Azure Management API

Web Hook

Automation Runbook

PowerShell

VNet

SSIS IR

CSV

XLS

TXT

Express Route

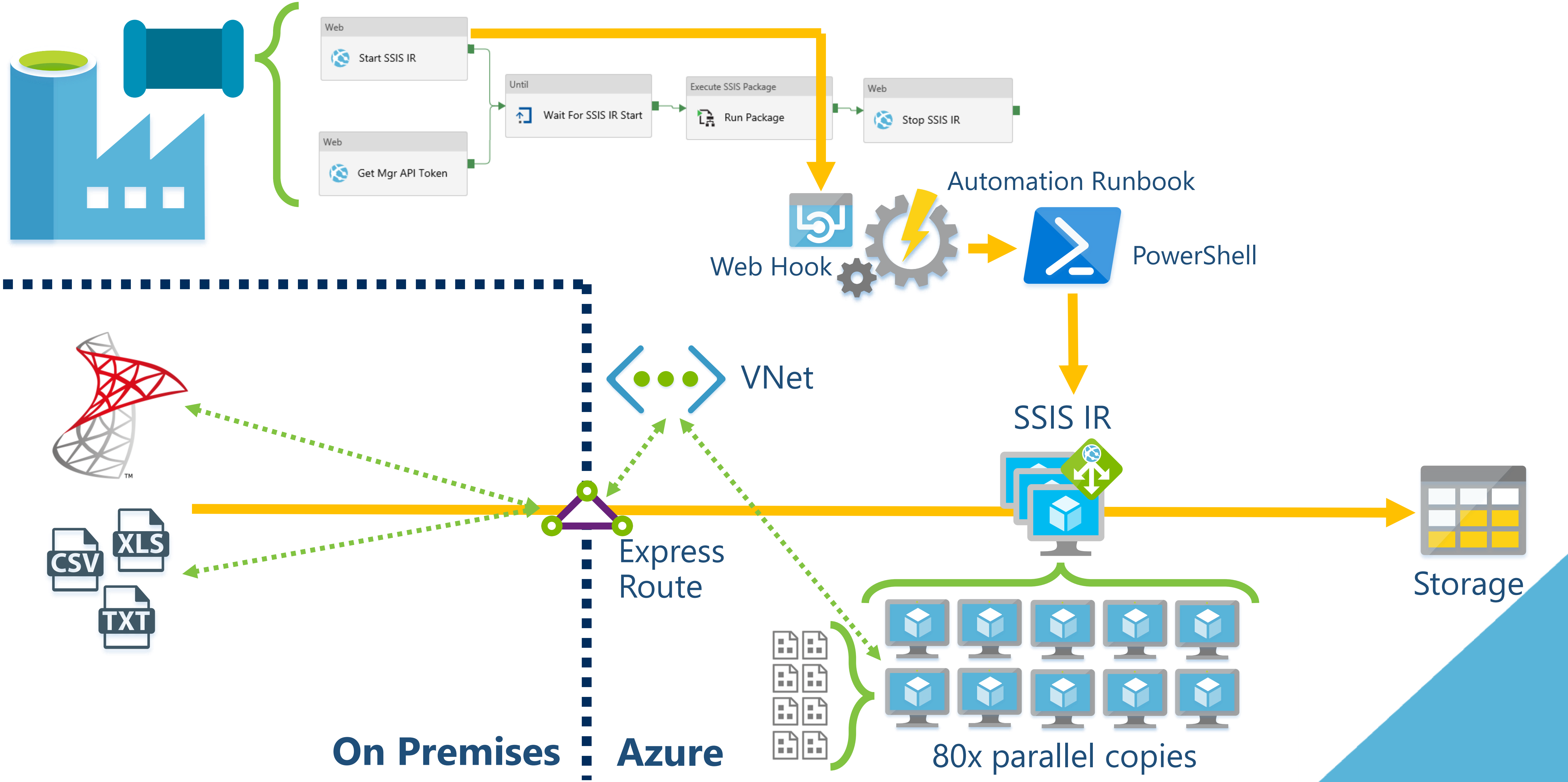
Storage

80x parallel copies

On Premises

Azure

The SSIS IR Start/Stop



Web: Start SSIS IR

Web: Get Mgr API Token

Until: Wait For SSIS IR Start

Execute SSIS Package: Run Package

Web: Stop SSIS IR

Web Hook

Automation Runbook

PowerShell

VNet

SSIS IR

Express Route

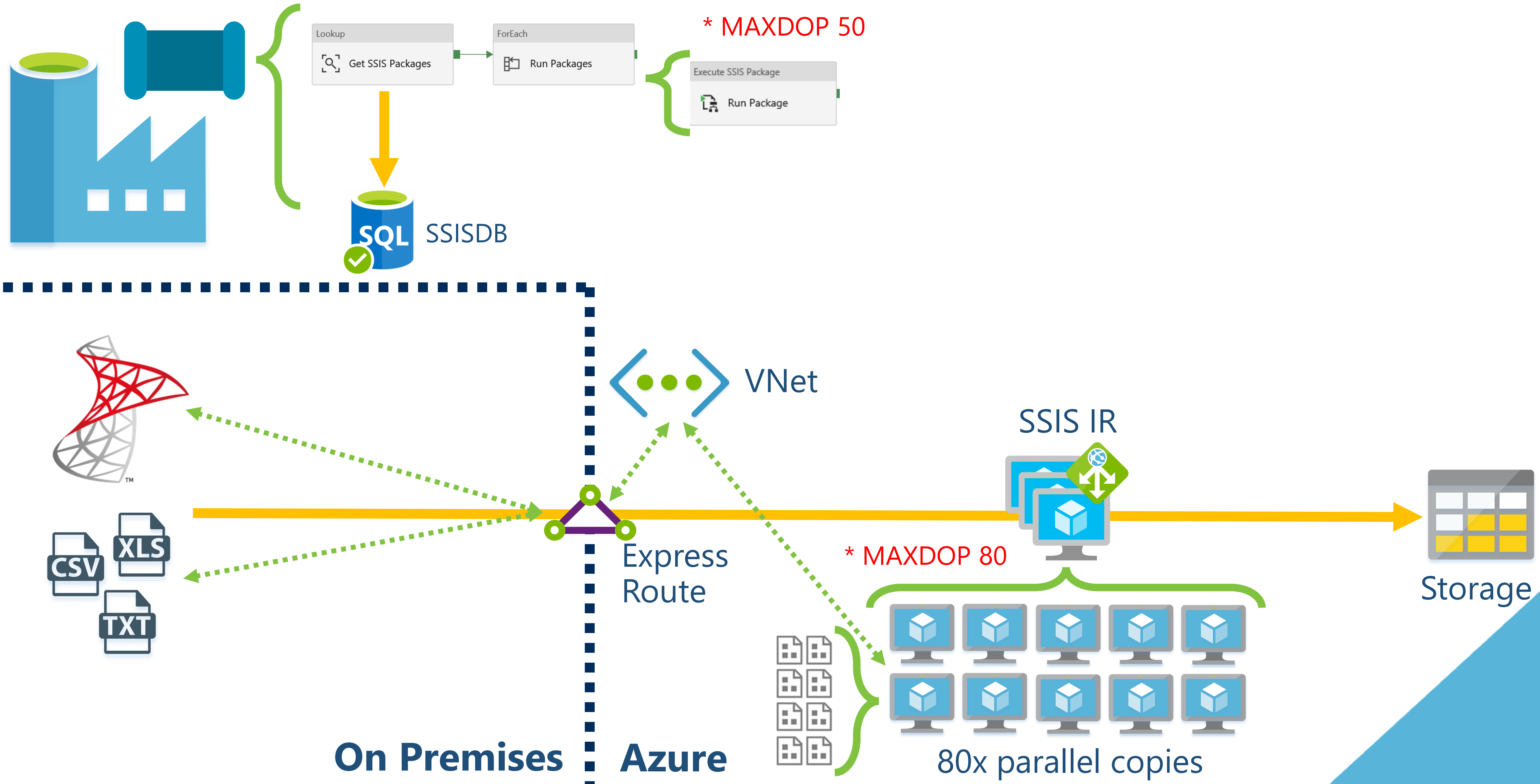
Storage

80x parallel copies

On Premises

Azure

The SSIS IR Parallelism



Lookup: Get SSIS Packages

ForEach: Run Packages

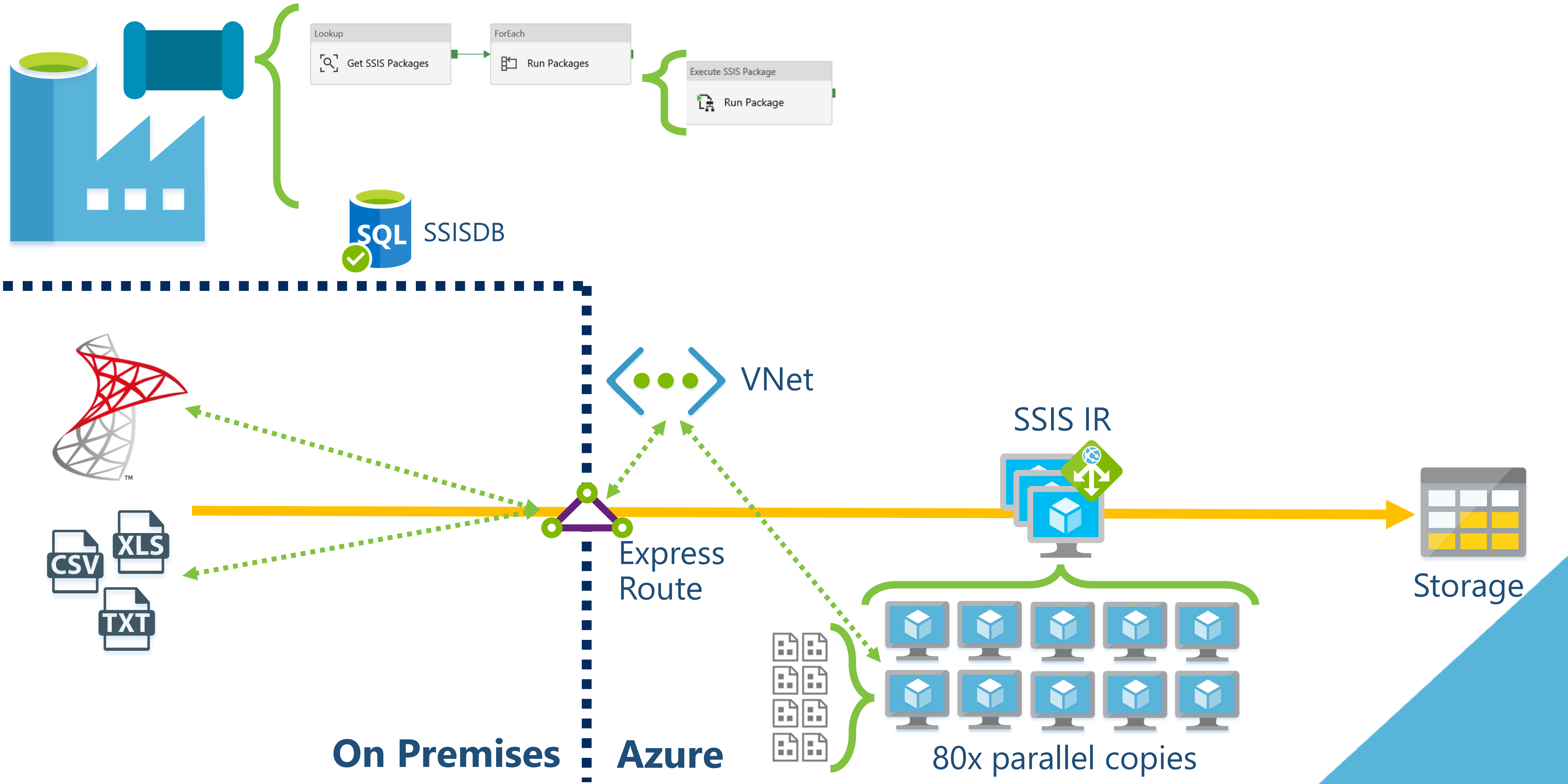
Execute SSIS Package: Run Package

* MAXDOP 50

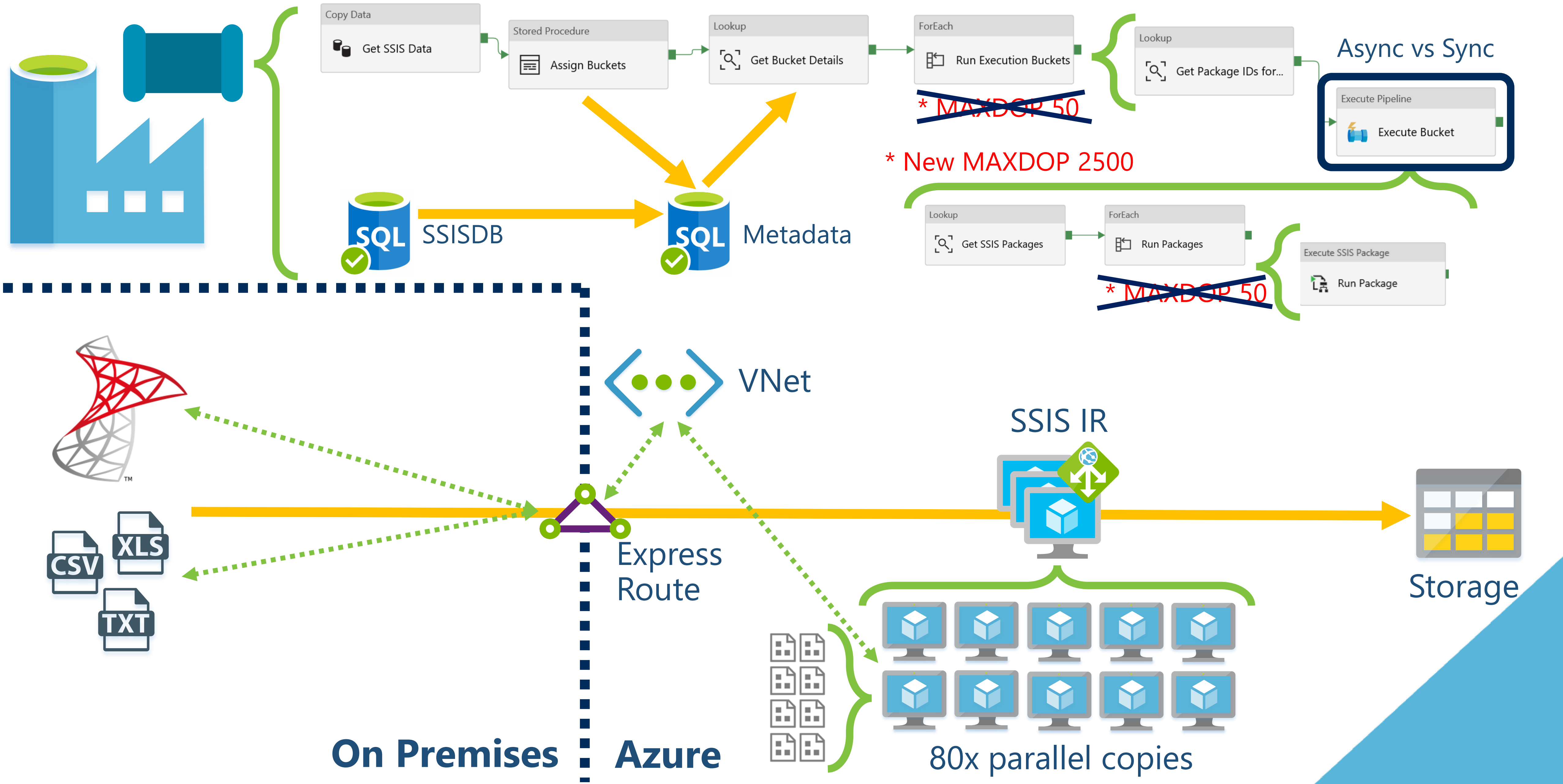
* MAXDOP 80

80x parallel copies

The SSIS IR Parallelism



The SSIS IR Parallelism



Copy Data: Get SSIS Data

Stored Procedure: Assign Buckets

Lookup: Get Bucket Details

ForEach: Run Execution Buckets

* ~~MAXDOP 50~~

Lookup: Get Package IDs for...

Async vs Sync

Execute Pipeline: Execute Bucket

* New MAXDOP 2500

Lookup: Get SSIS Packages

ForEach: Run Packages

* ~~MAXDOP 50~~

Execute SSIS Package: Run Package

SSISDB

Metadata

VNet

SSIS IR

Express Route

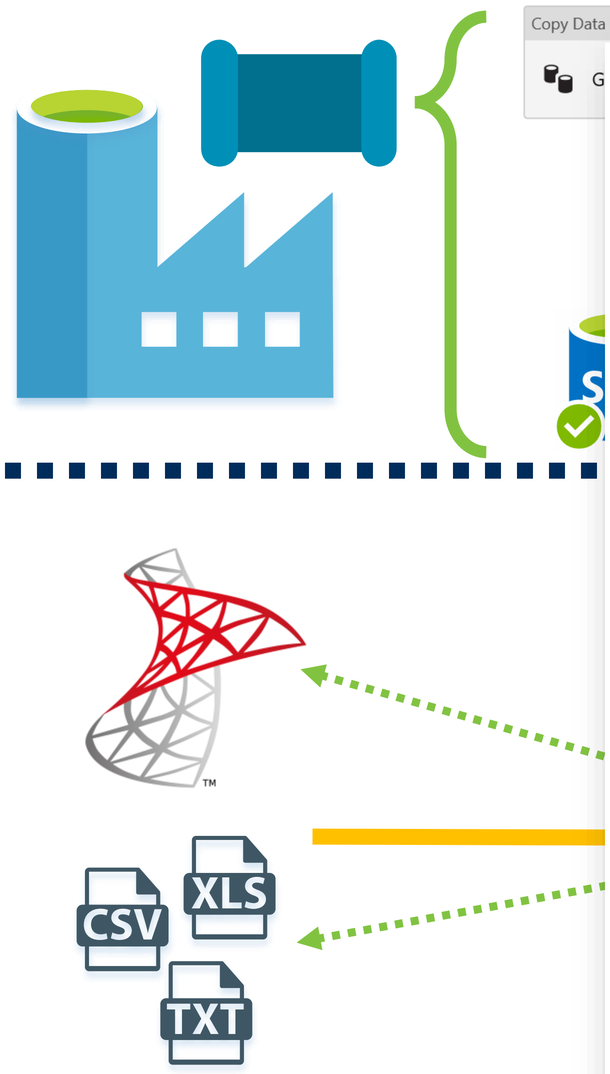
Storage

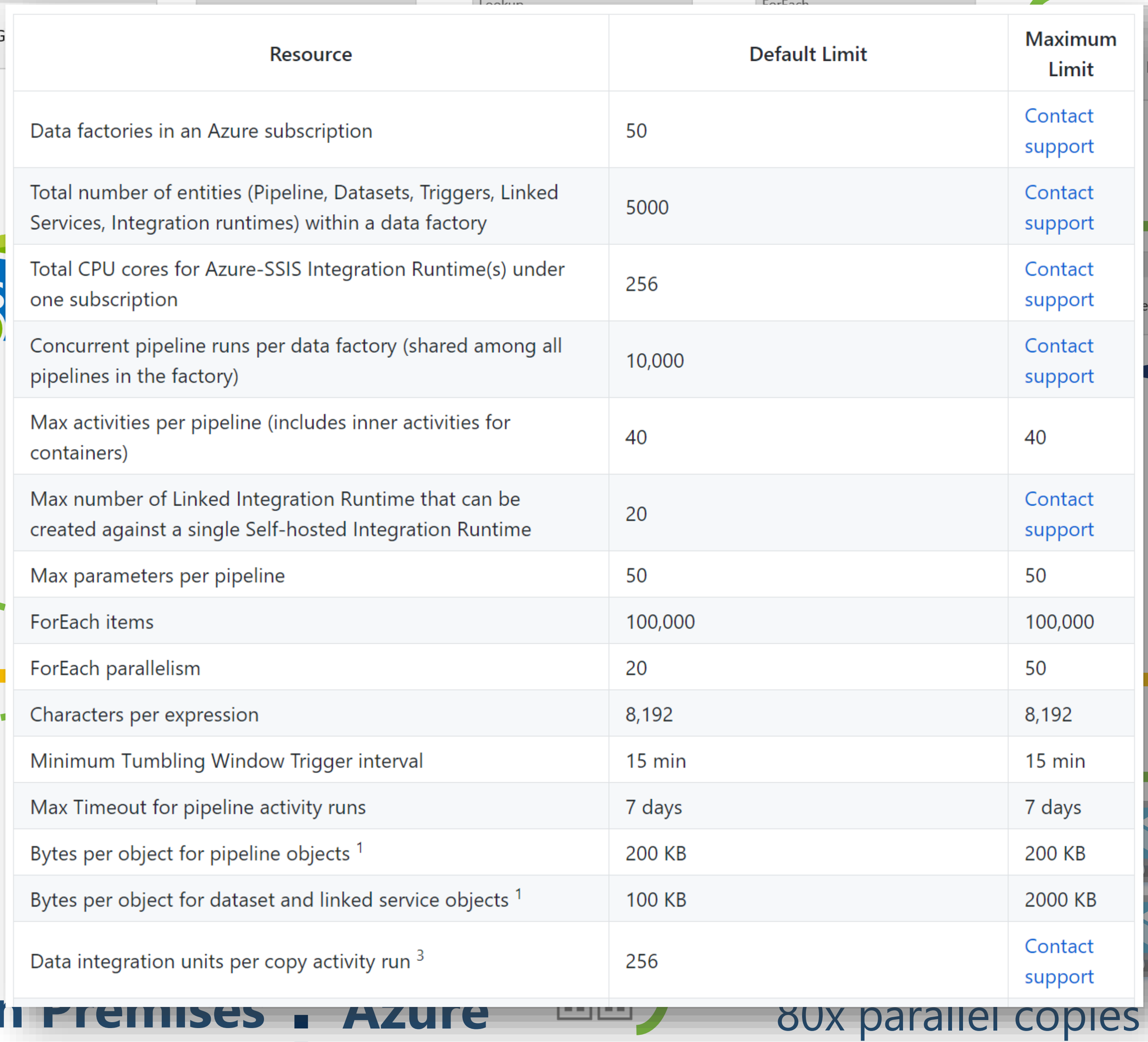
80x parallel copies

On Premises

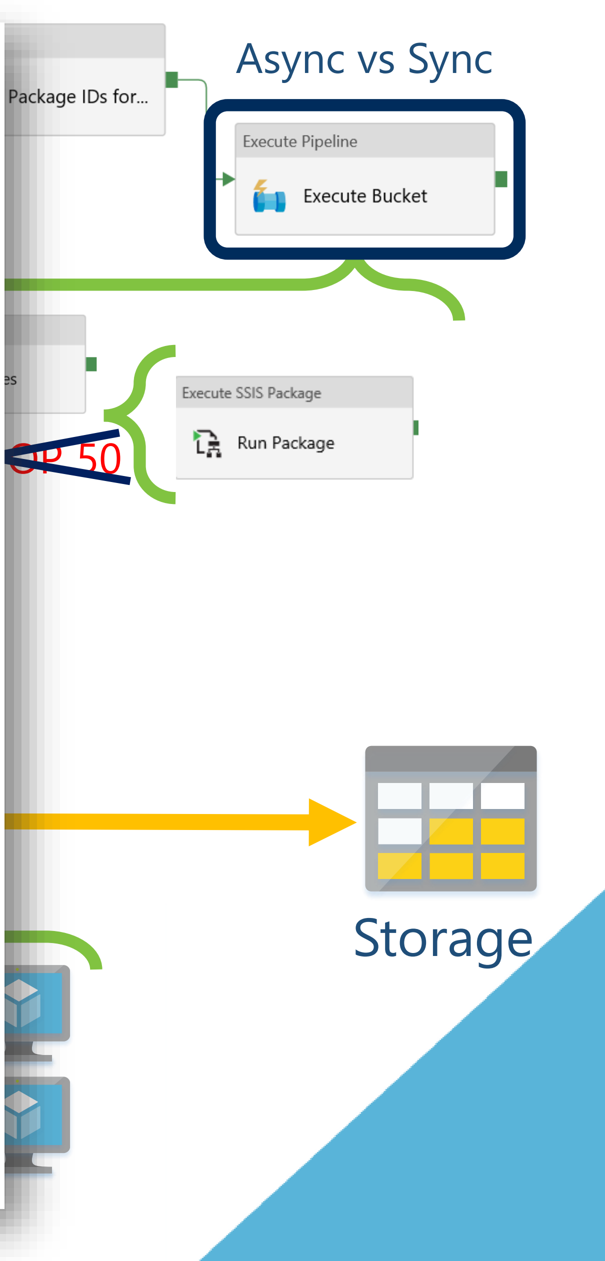
Azure

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

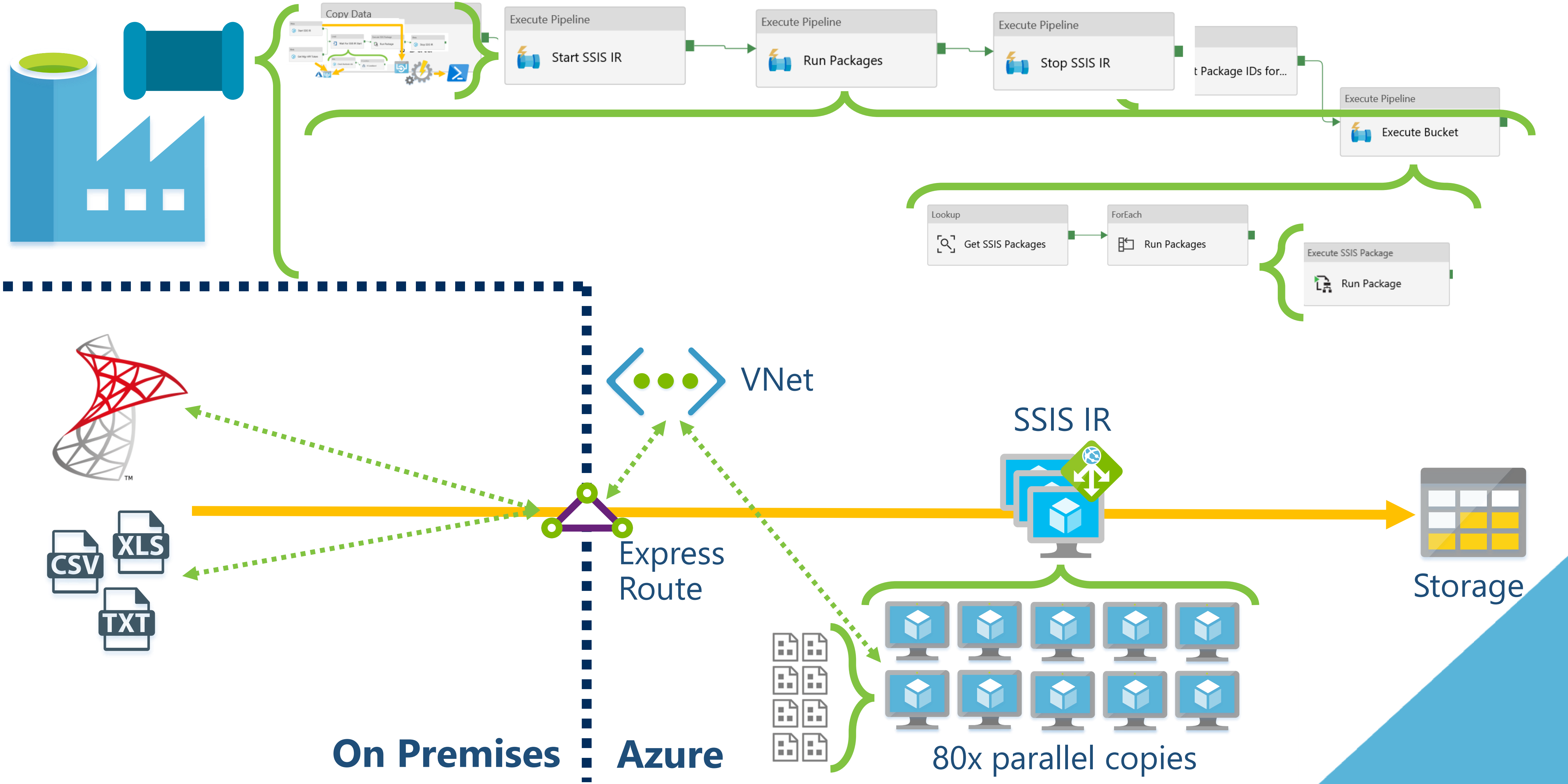


On Premises : Azure

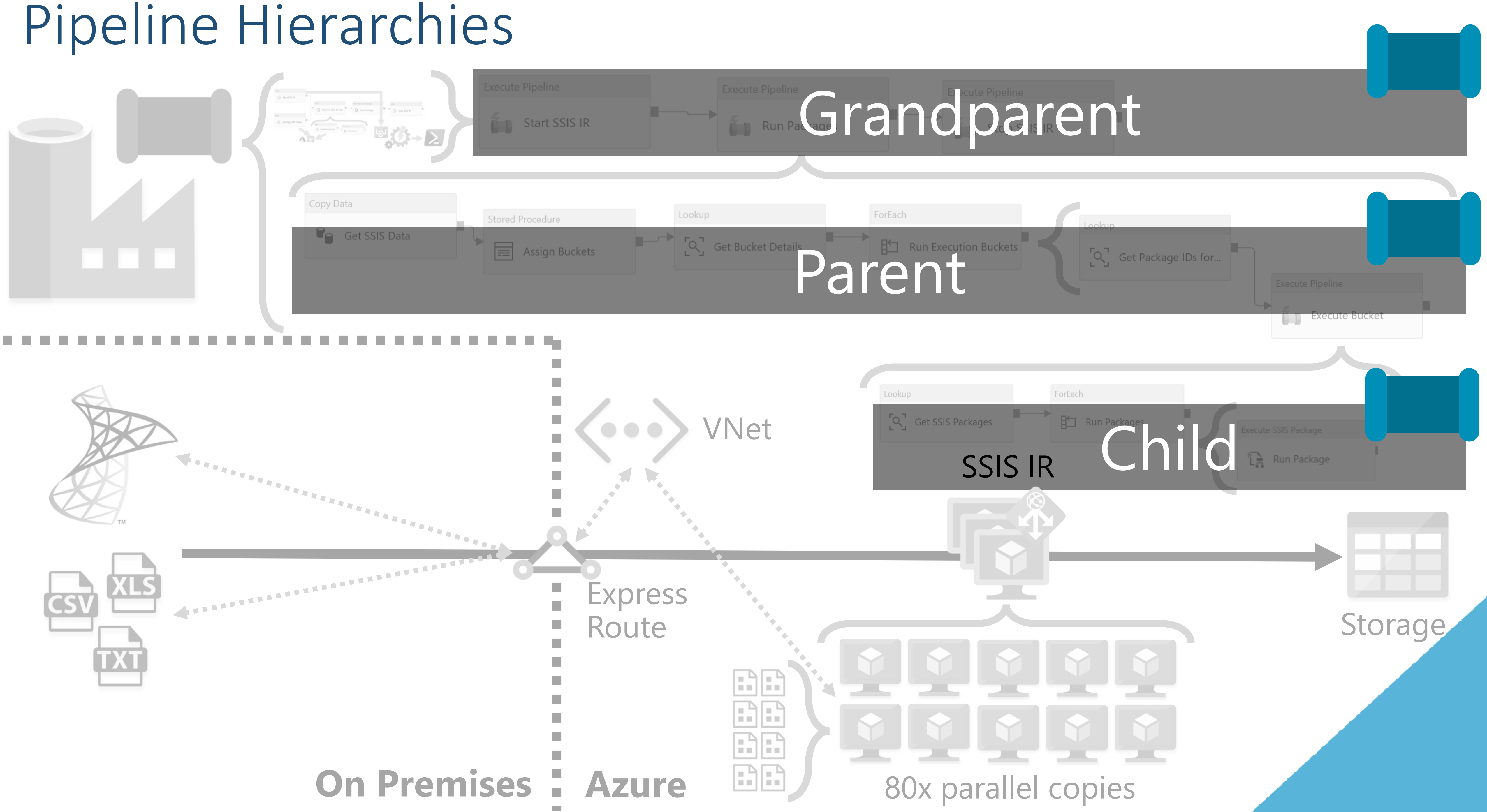
80x parallel copies

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

SSIS IR & Package Complete Orchestration Solution



Pipeline Hierarchies



Grandparent

Parent

Child

SSIS IR

VNet

Express Route

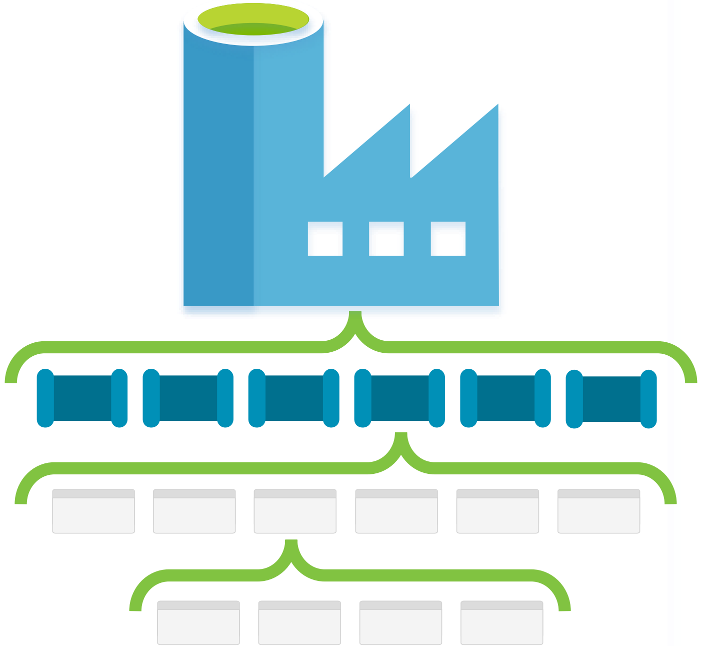
Storage

80x parallel copies

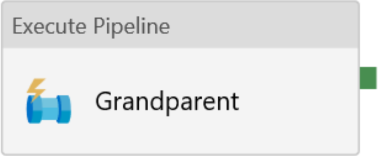
On Premises

Azure

Scale Out ~~Execution~~ Everything!



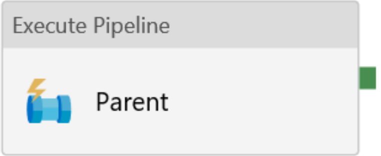
Pattern Summary



Execute Pipeline

Grandparent

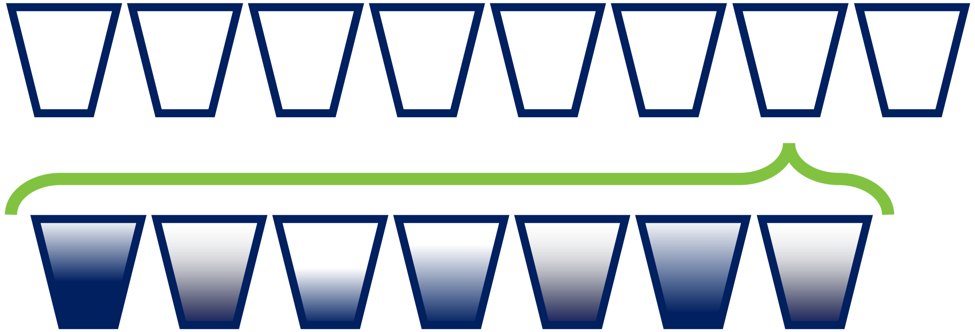
High Level Control Flow and Pipeline Triggers



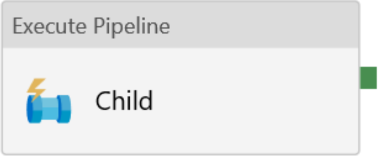
Execute Pipeline

Parent

Platform Component Control



Manage Parallel Activities



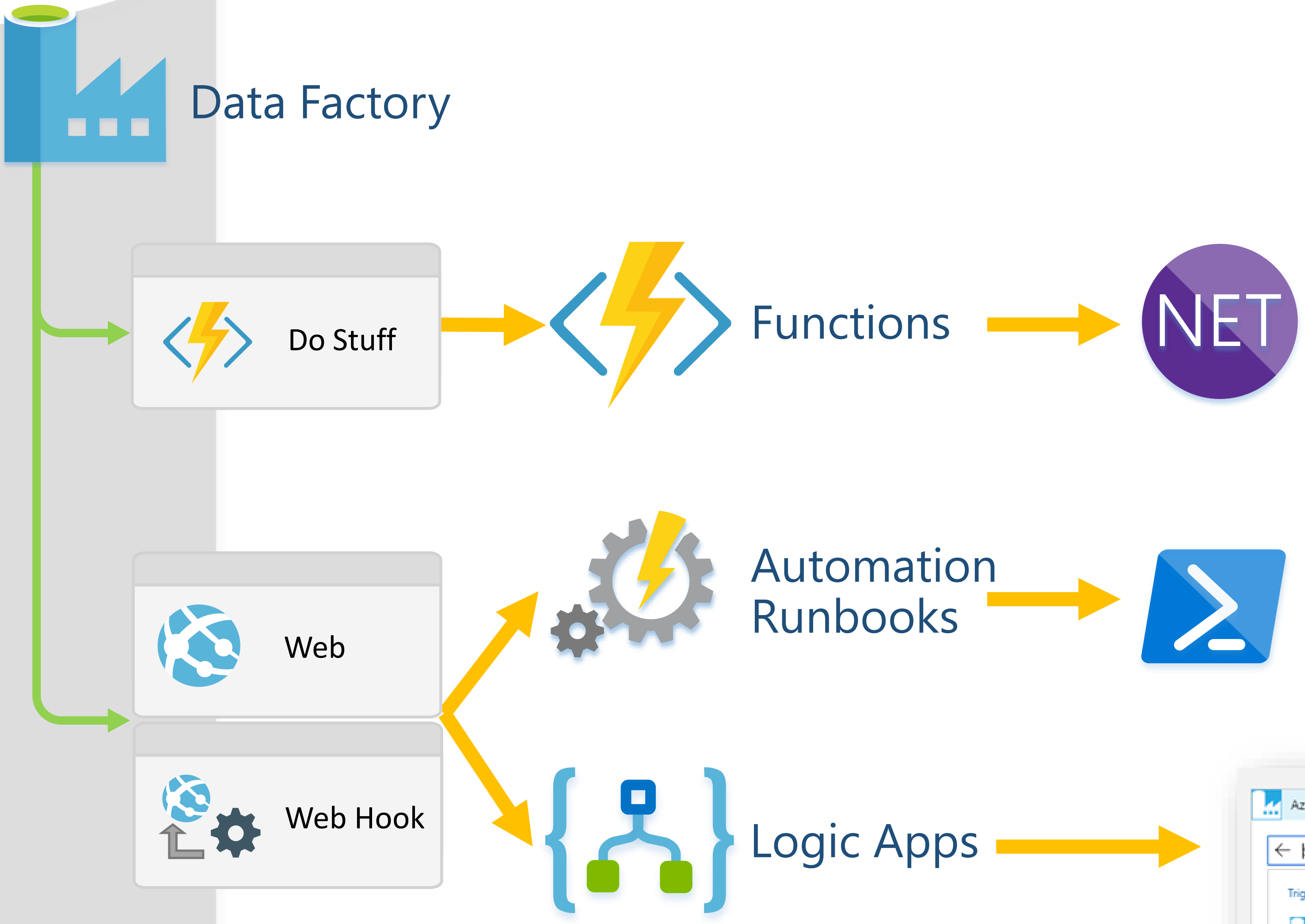
Execute Pipeline

Child

Service Level Executions

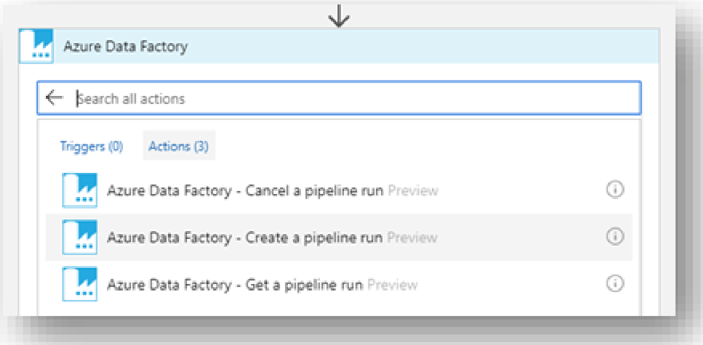
Solution Bootstrapping

Bootstrapping



```
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);
```

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


Azure Data Factory

Search all actions

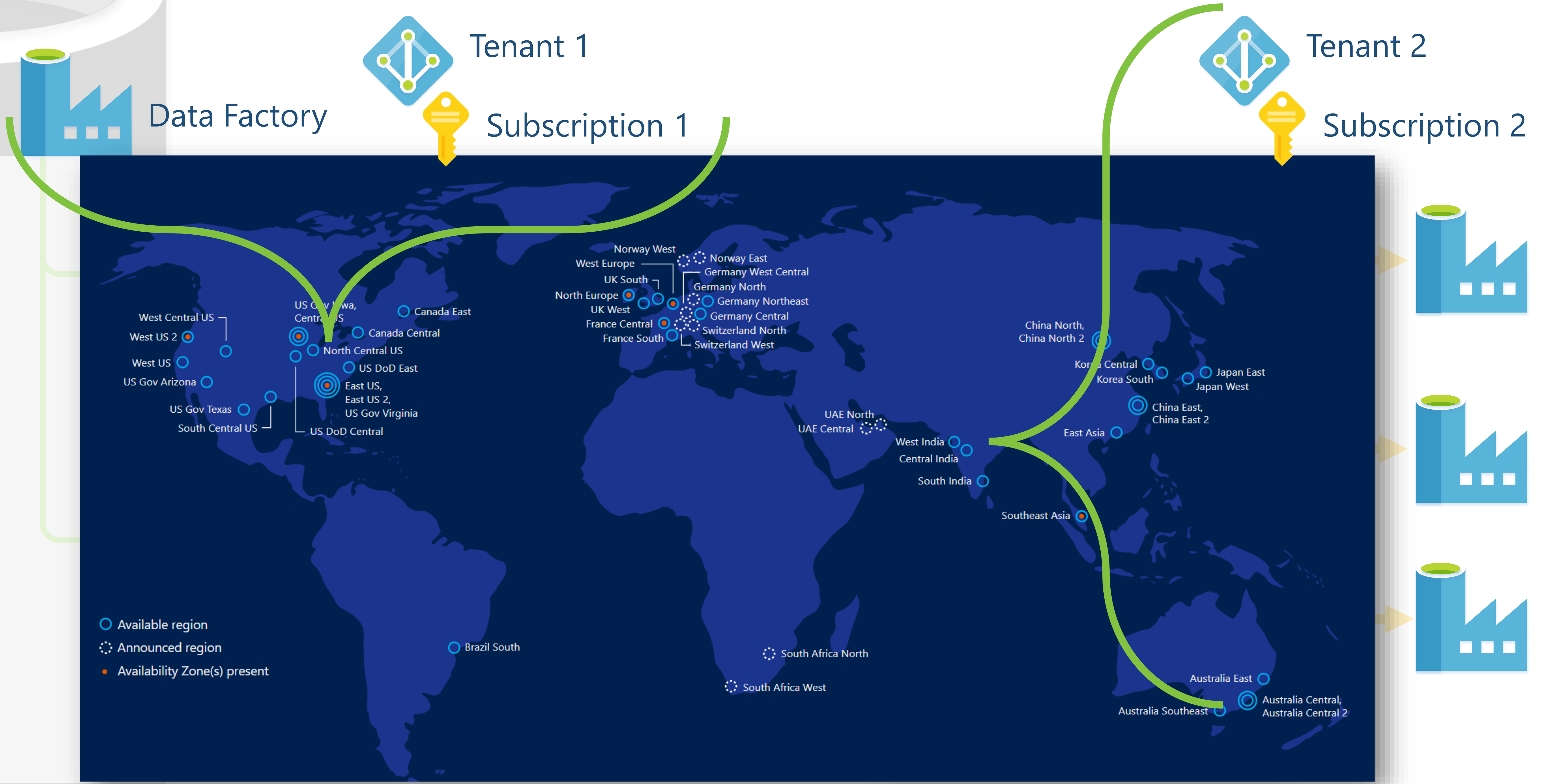
Triggers (0) Actions (3)

Azure Data Factory - Cancel a pipeline run Preview

Azure Data Factory - Create a pipeline run Preview

Azure Data Factory - Get a pipeline run Preview

Bootstrapping



Data Factory

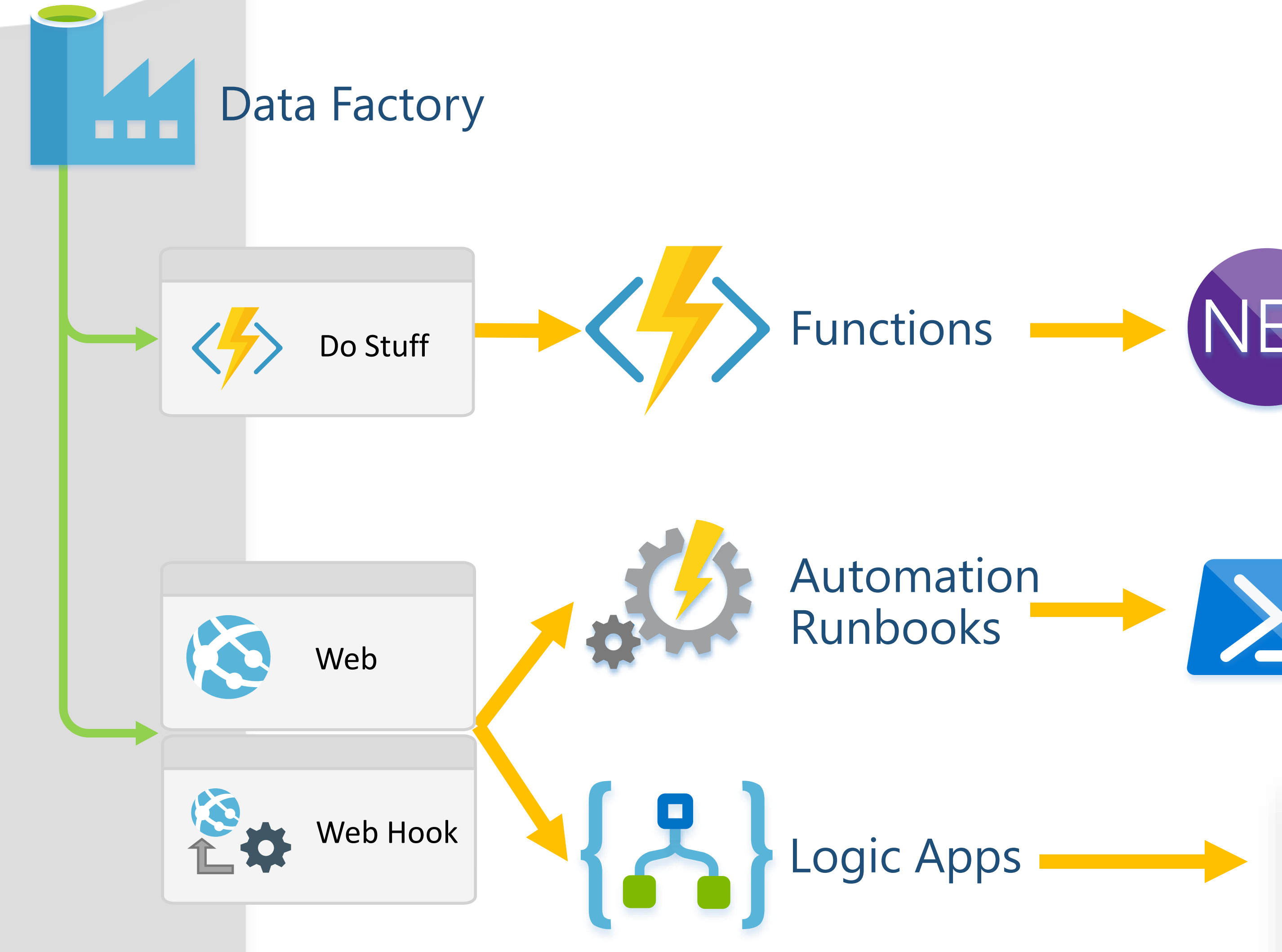
Tenant 1

Subscription 1

Tenant 2

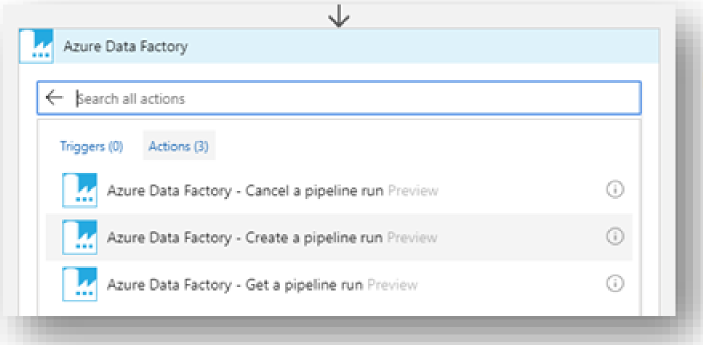
Subscription 2

Bootstrapping



```
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);
```

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



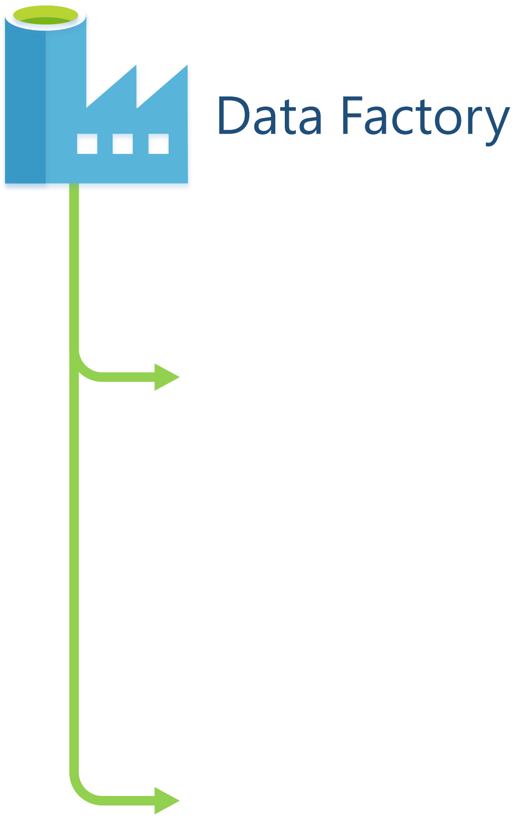
Azure Data Factory

Search all actions

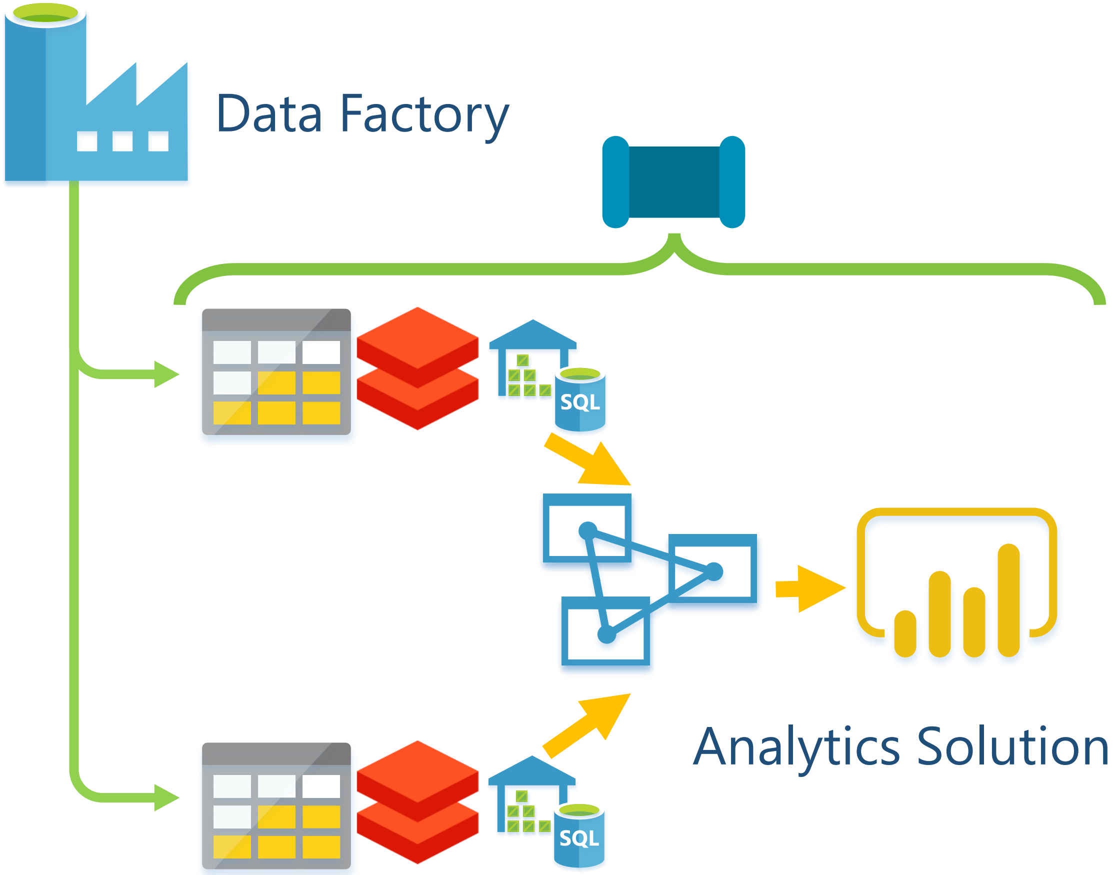
Triggers (0) Actions (3)

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

Bootstrapping – Wider Analytics Solution



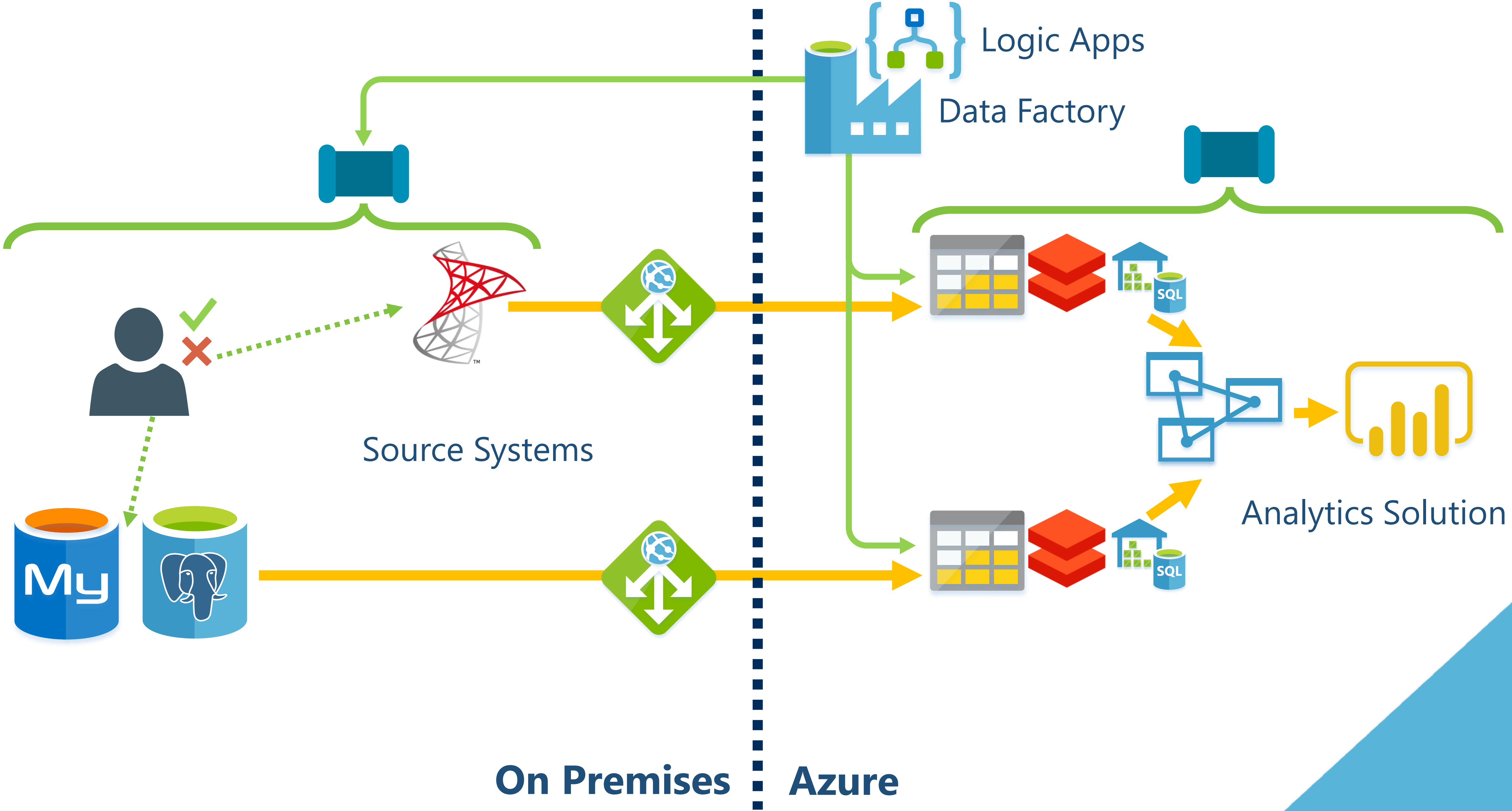
Bootstrapping – Wider Analytics Solution



Data Factory

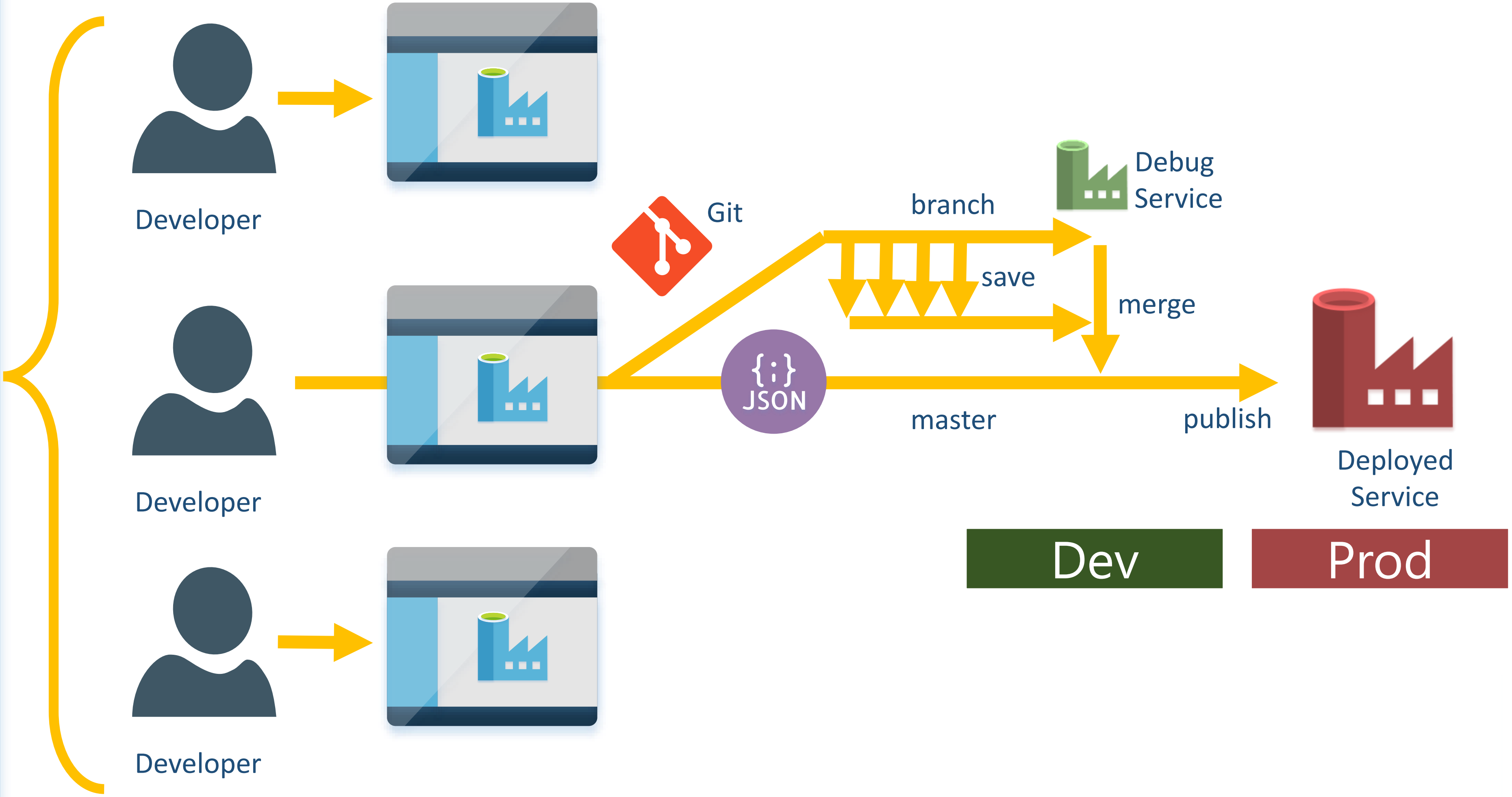
Analytics Solution

Bootstrapping – Wider Analytics Solution

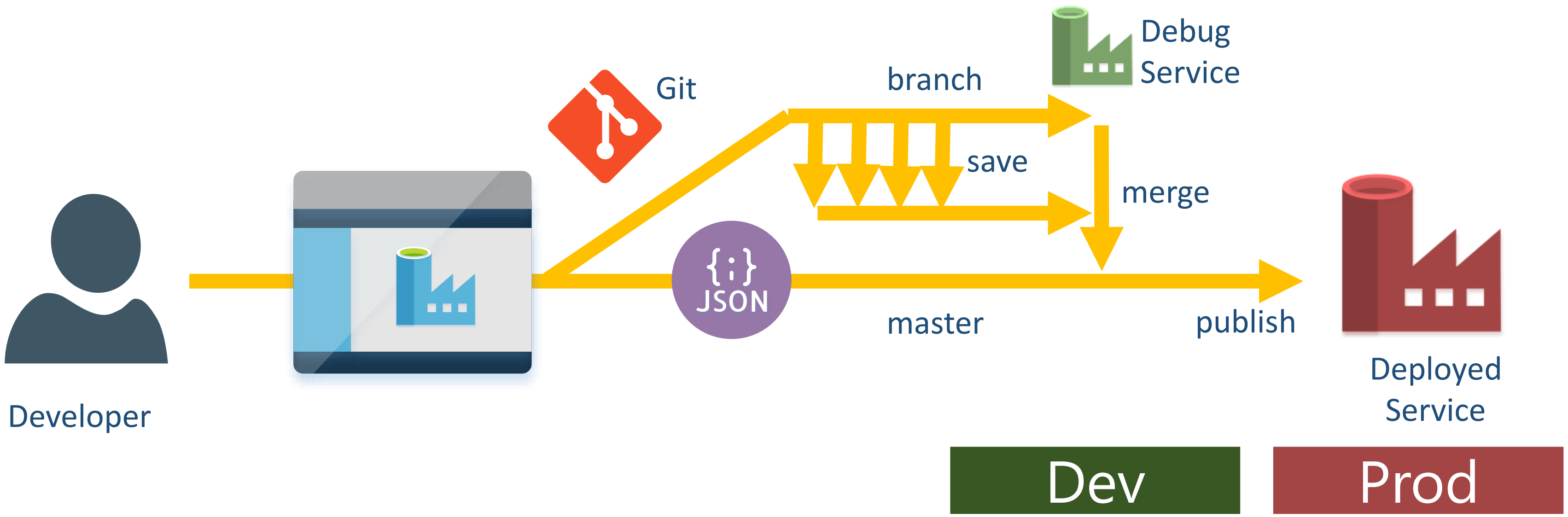


Data Factory DevOps – CI/CD

Data Factory Continuous Integration

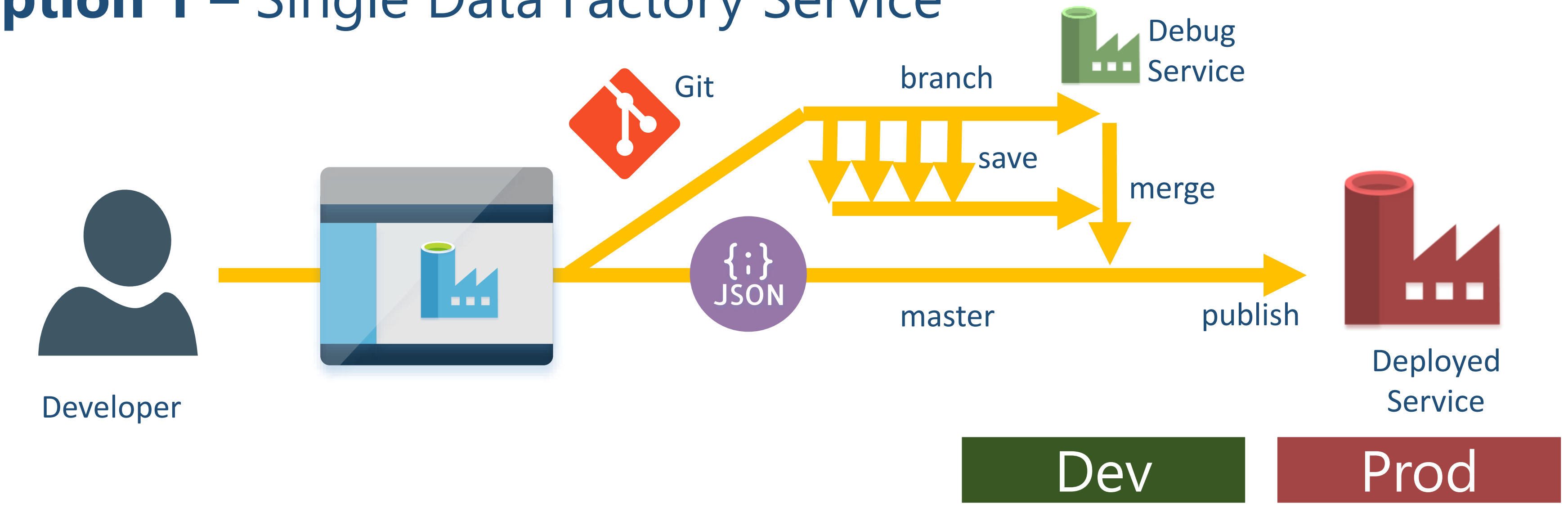


Data Factory Continuous Delivery

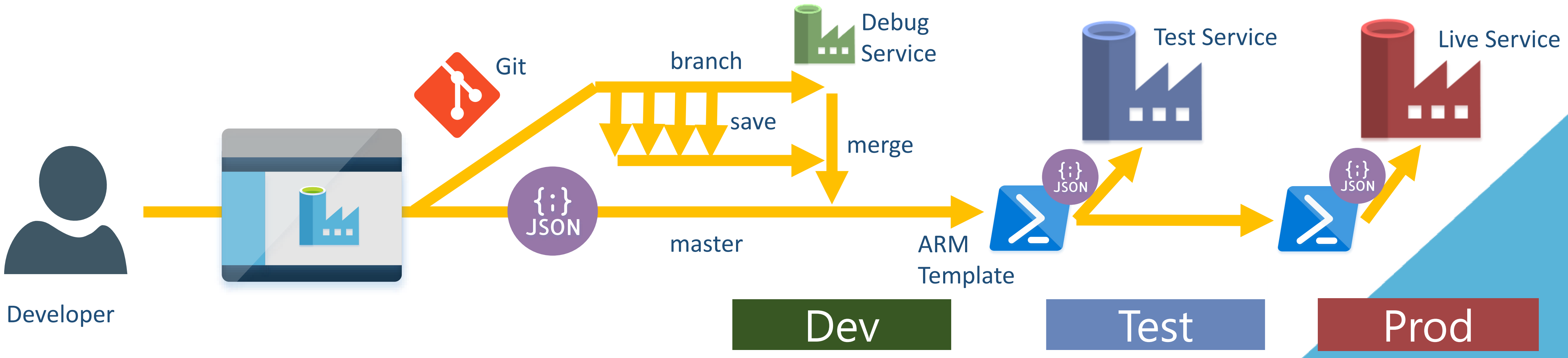


Data Factory Continuous Delivery

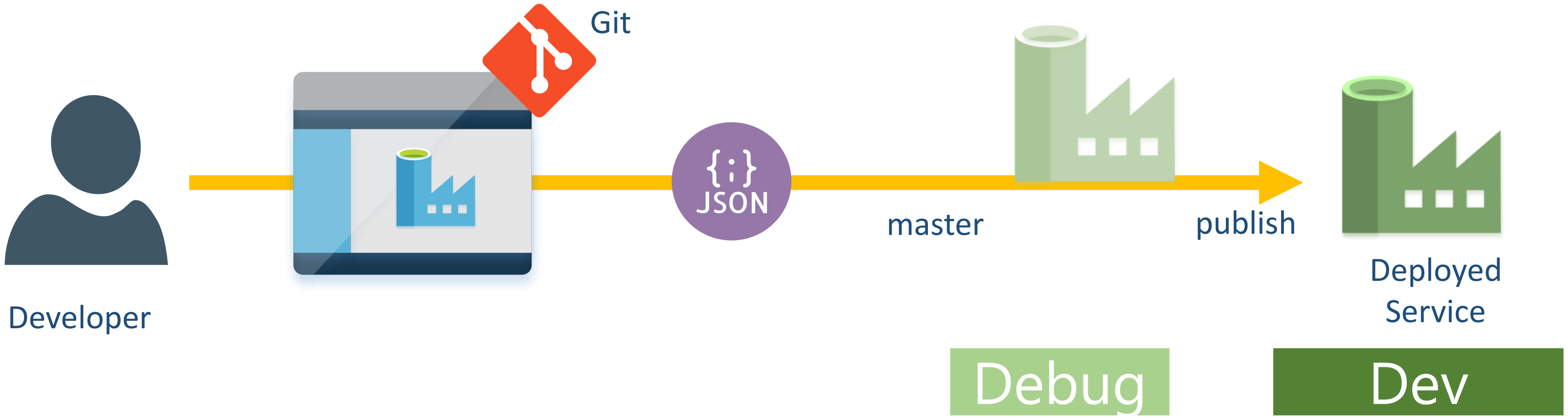
Option 1 – Single Data Factory Service



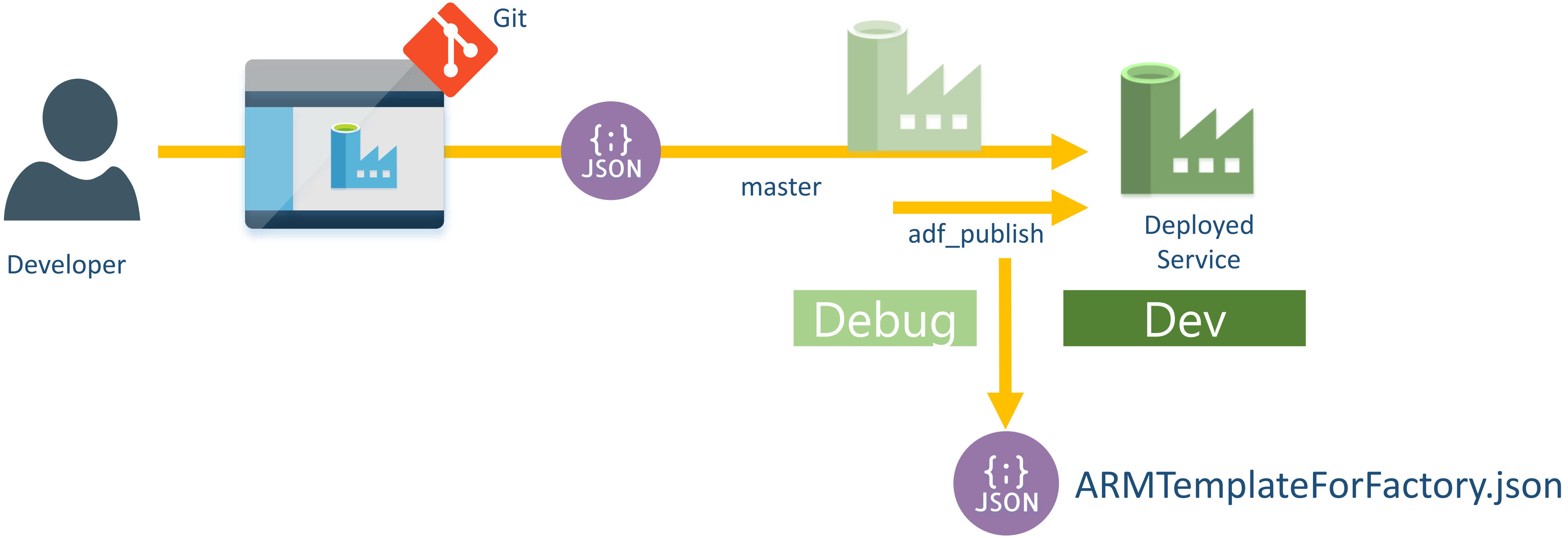
Option 2 – ARM Templates for Multiple Data Factory Services



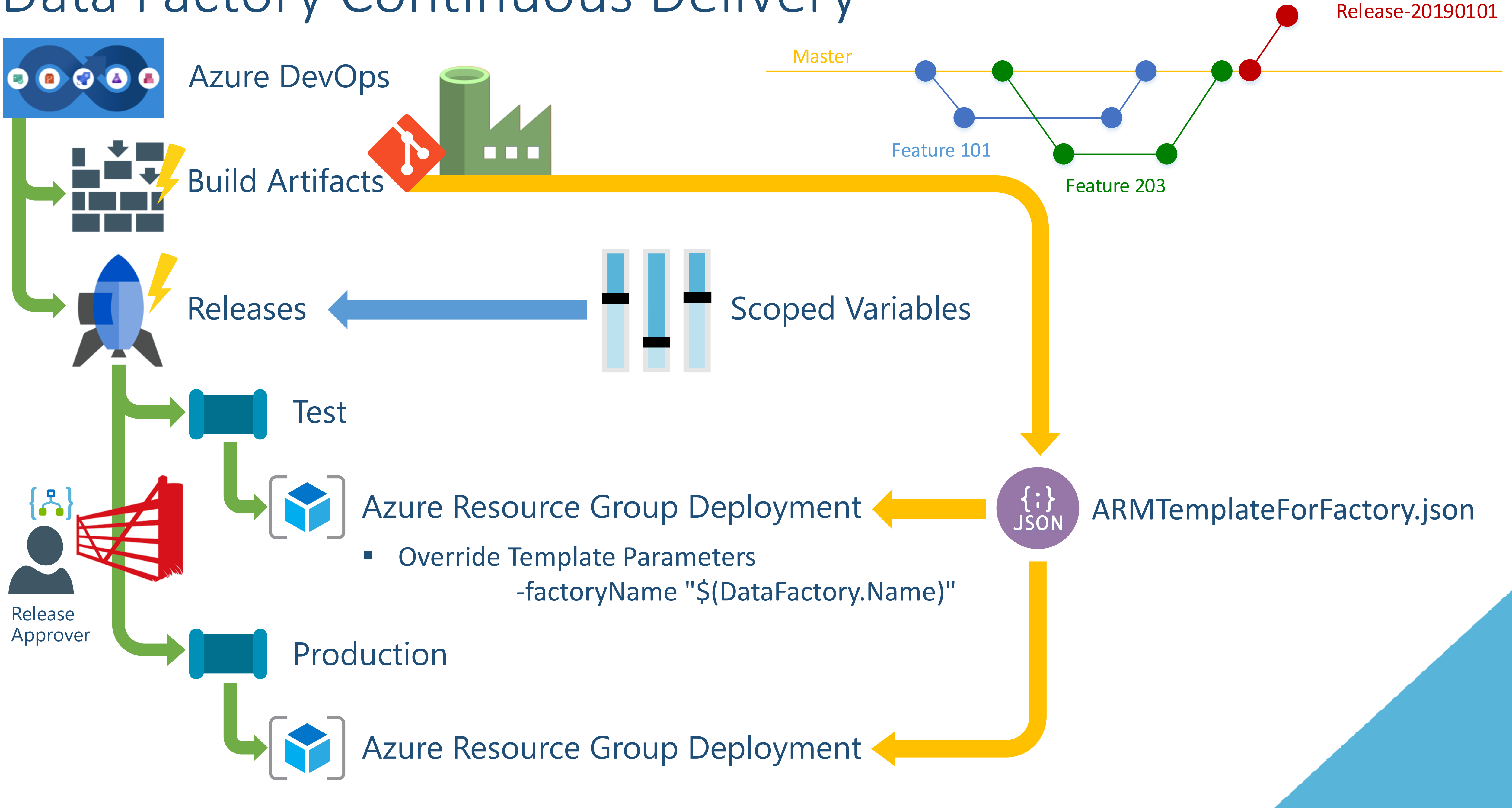
Data Factory Publish



Data Factory Publish



Data Factory Continuous Delivery

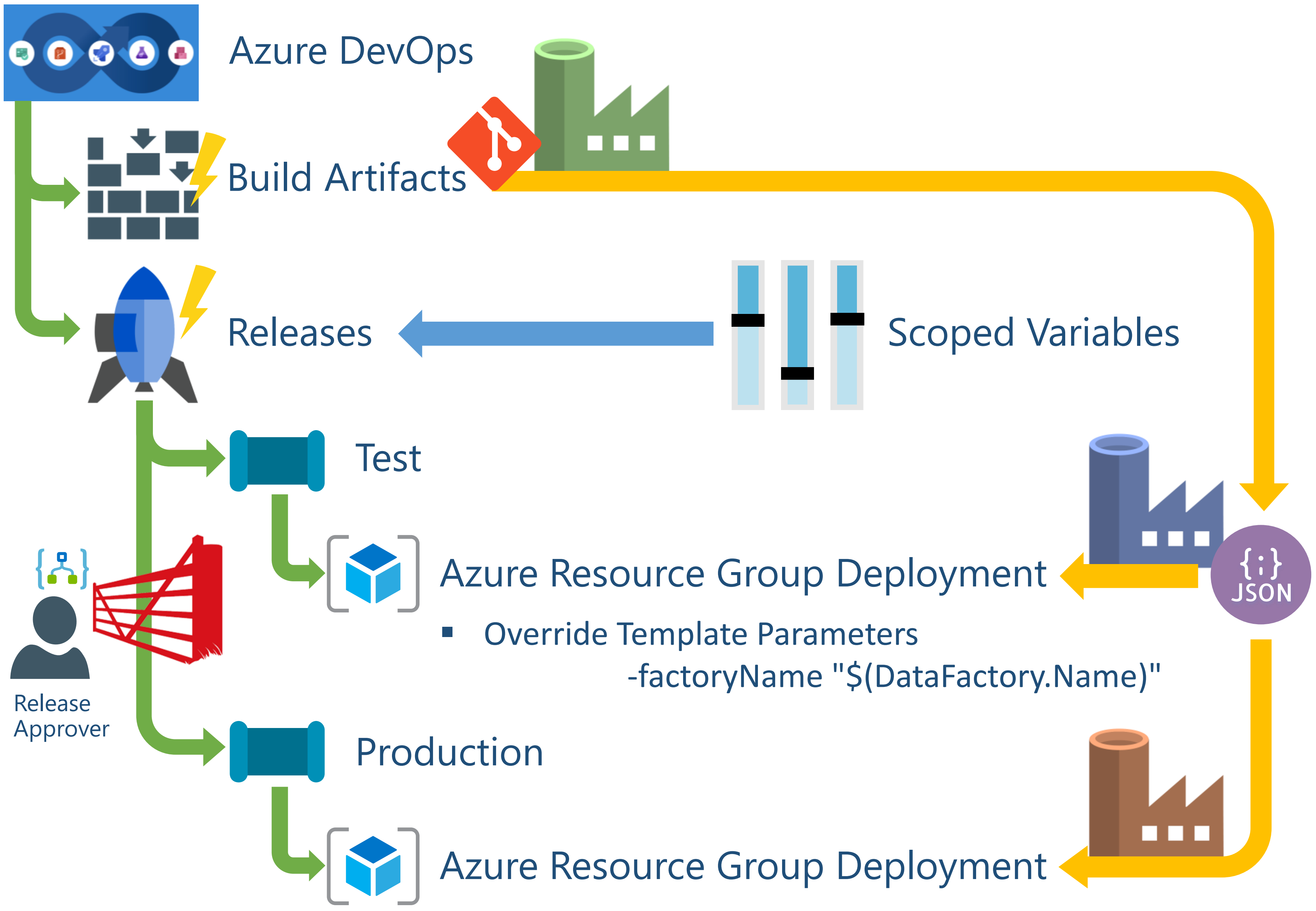


- Override Template Parameters

  -factoryName "$(DataFactory.Name)"

Data Factory Continuous Delivery

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



Azure DevOps

Build Artifacts

Releases

Scoped Variables

Test

Azure Resource Group Deployment

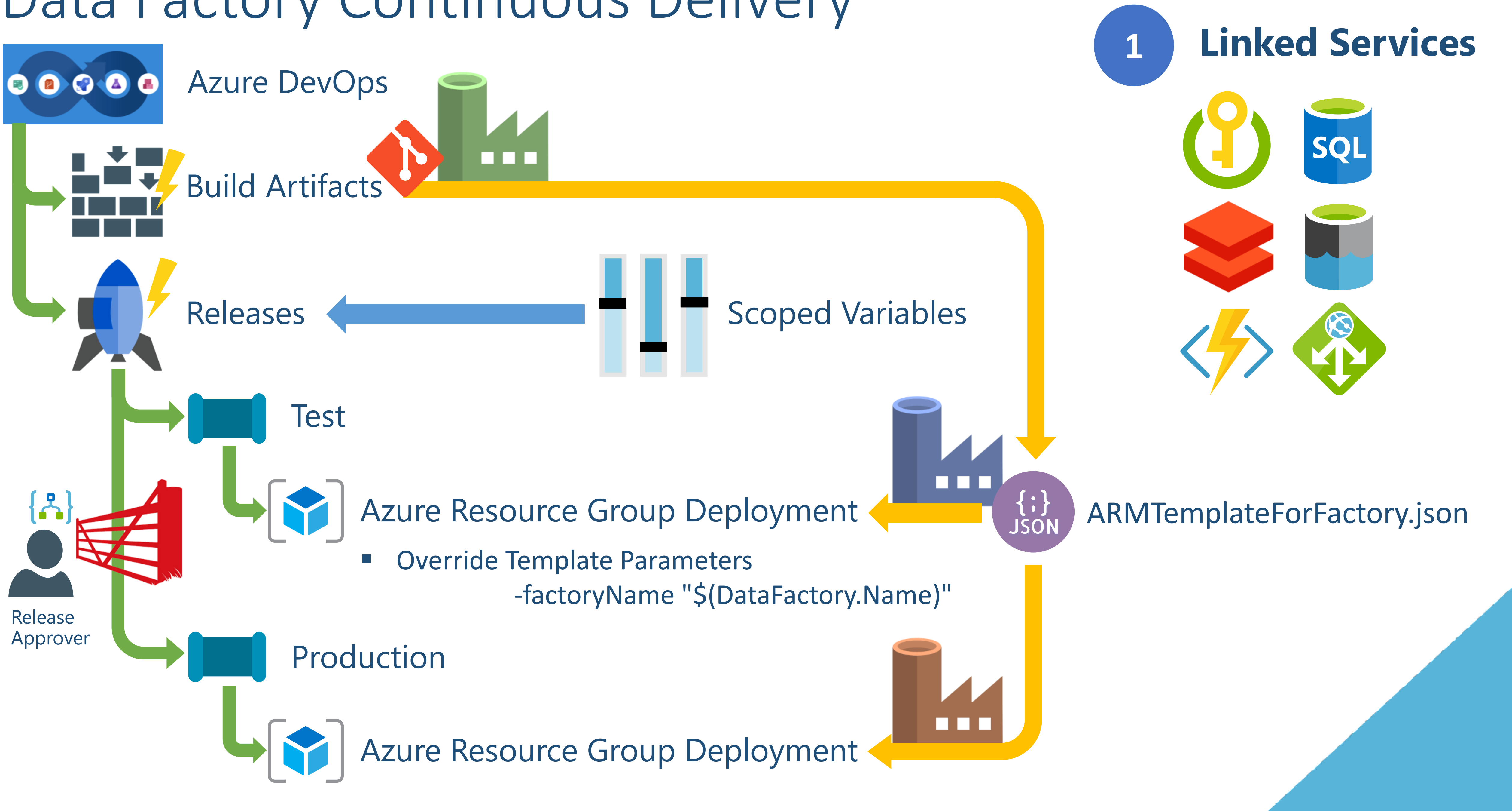
- Override Template Parameters
  -factoryName "$(DataFactory.Name)"

Production

Azure Resource Group Deployment

ARMTemplateForFactory.json

Data Factory Continuous Delivery



1 **Linked Services**

Azure DevOps

Build Artifacts

Releases

Scoped Variables

Test

Azure Resource Group Deployment

- Override Template Parameters

  -factoryName "$(DataFactory.Name)"

ARMTemplateForFactory.json

Release Approver

Production

Azure Resource Group Deployment

Session Agenda

- Data Factory – A Quick Overview ✓
- Dynamic Pipelines ✓
- Extending Data Factory ✓
  - Web Activities
  - Custom Activities
- True Scale Out Execution ✓
  - SSIS Integration Runtime
- Data Factory – In Production ✓
  - Bootstrapping
  - DevOps

Thank you for listening…

Paul Andrew

MVP Microsoft® Most Valuable Professional

Blog:	mrpaulandrew.com
Email:	paul@mrpaulandrew.com
Twitter:	@mrpaulandrew
LinkedIn:	In/mrpaulandrew
GitHub:	github.com/mrpaulandrew