Complex Azure Orchestration

Data Factory in Production

Paul Andrew | Principal Consultant & Solution Architect

Microsoft Most Valuable Professional

altius

@MrPaulAndrew In/MrPaulAndrew MrPaulAndrew.com

https://github.com/mrpaulandrew

CommunityEvents

Demo code, content and slides from various community events.

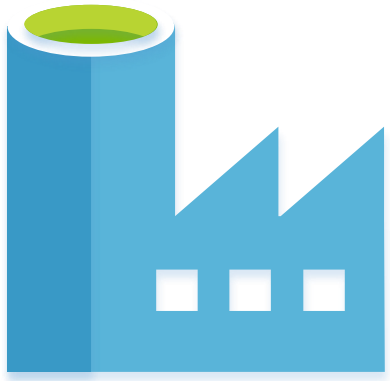
C++

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

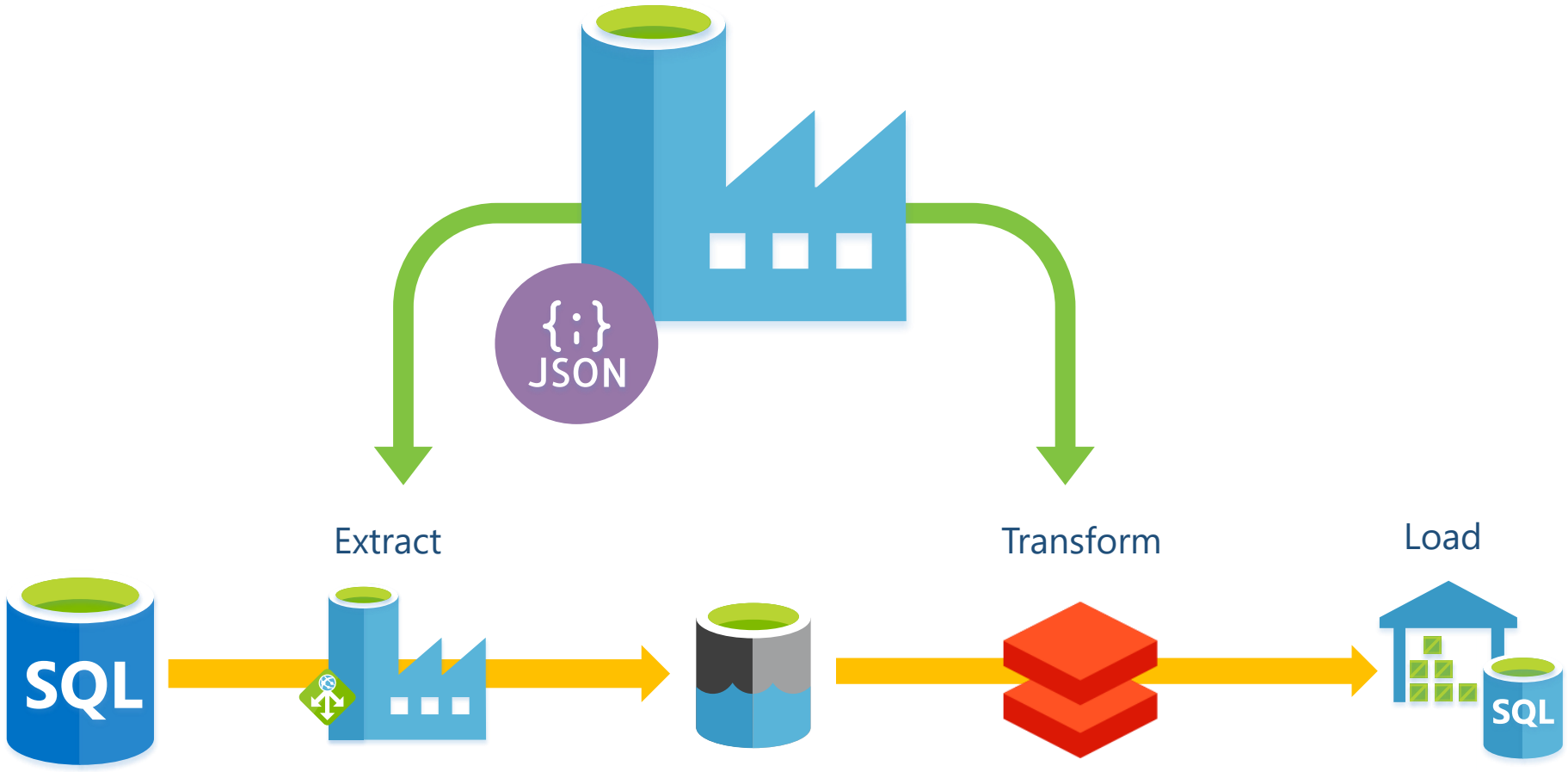
Session Agenda (Short Stories)

- 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

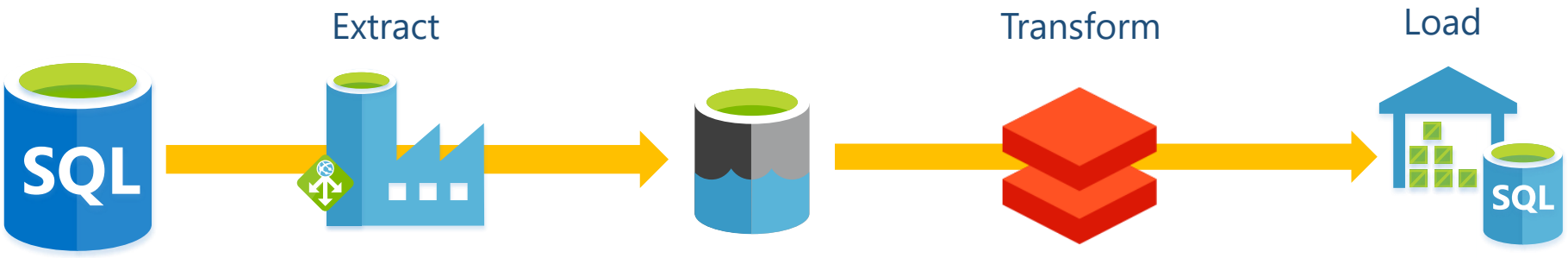
Azure Data Factory
A Quick Overview



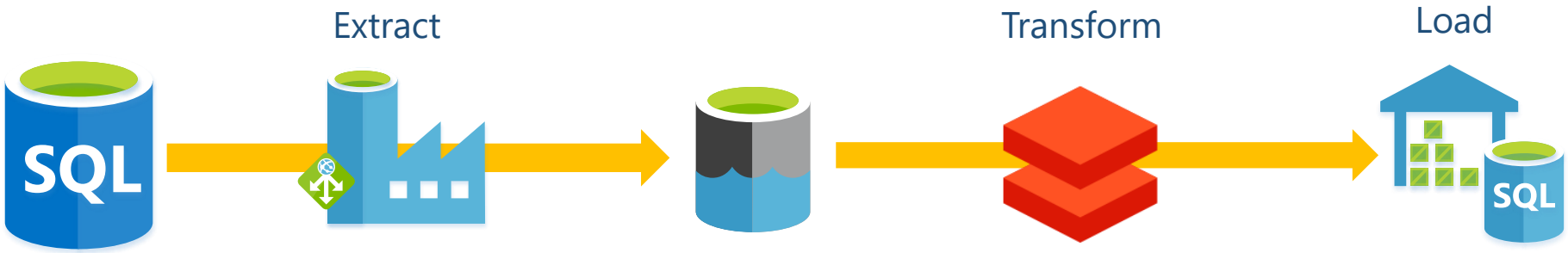
What is Azure Data Factory?



What is Azure Data Factory?



Data Factory Components

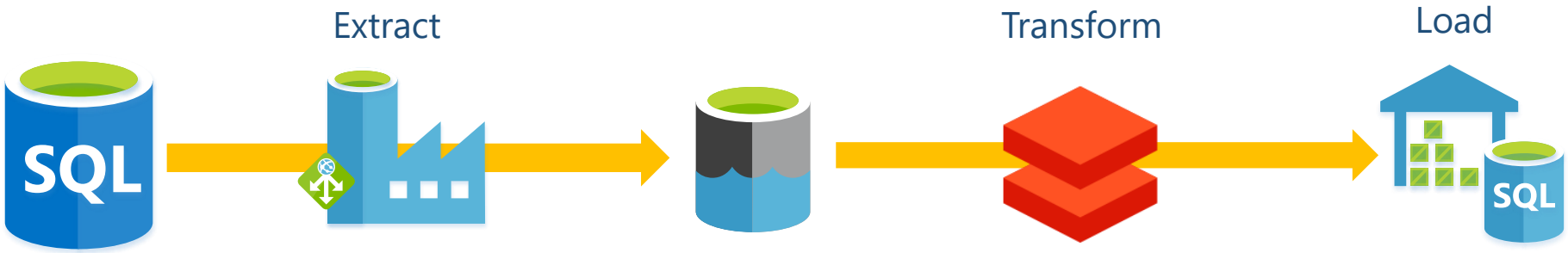


1. **Linked Services** ✓
2. **Data Sets** ✓
3. **Activities** ✓
4. **Pipelines** ✓
5. **Triggers** ✗

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

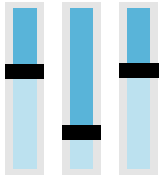


1. **Linked Services** ✓
2. **Data Sets** ✓
3. **Activities** ✓
4. **Pipelines** ✓
5. **Triggers** ✗

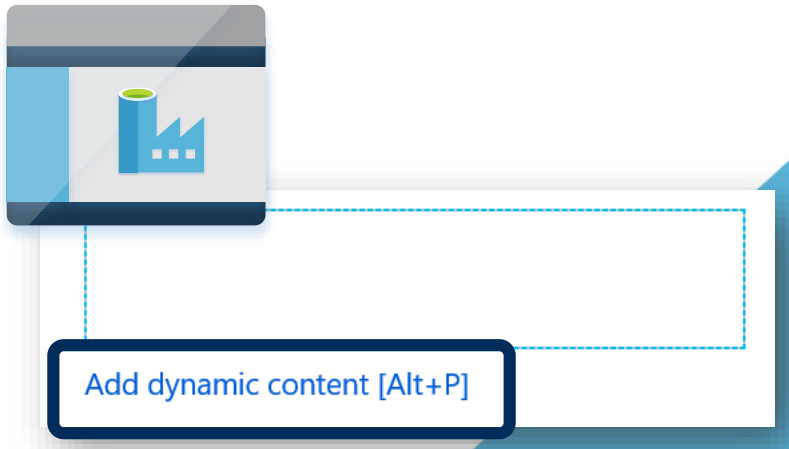
JSON

Expression Builder

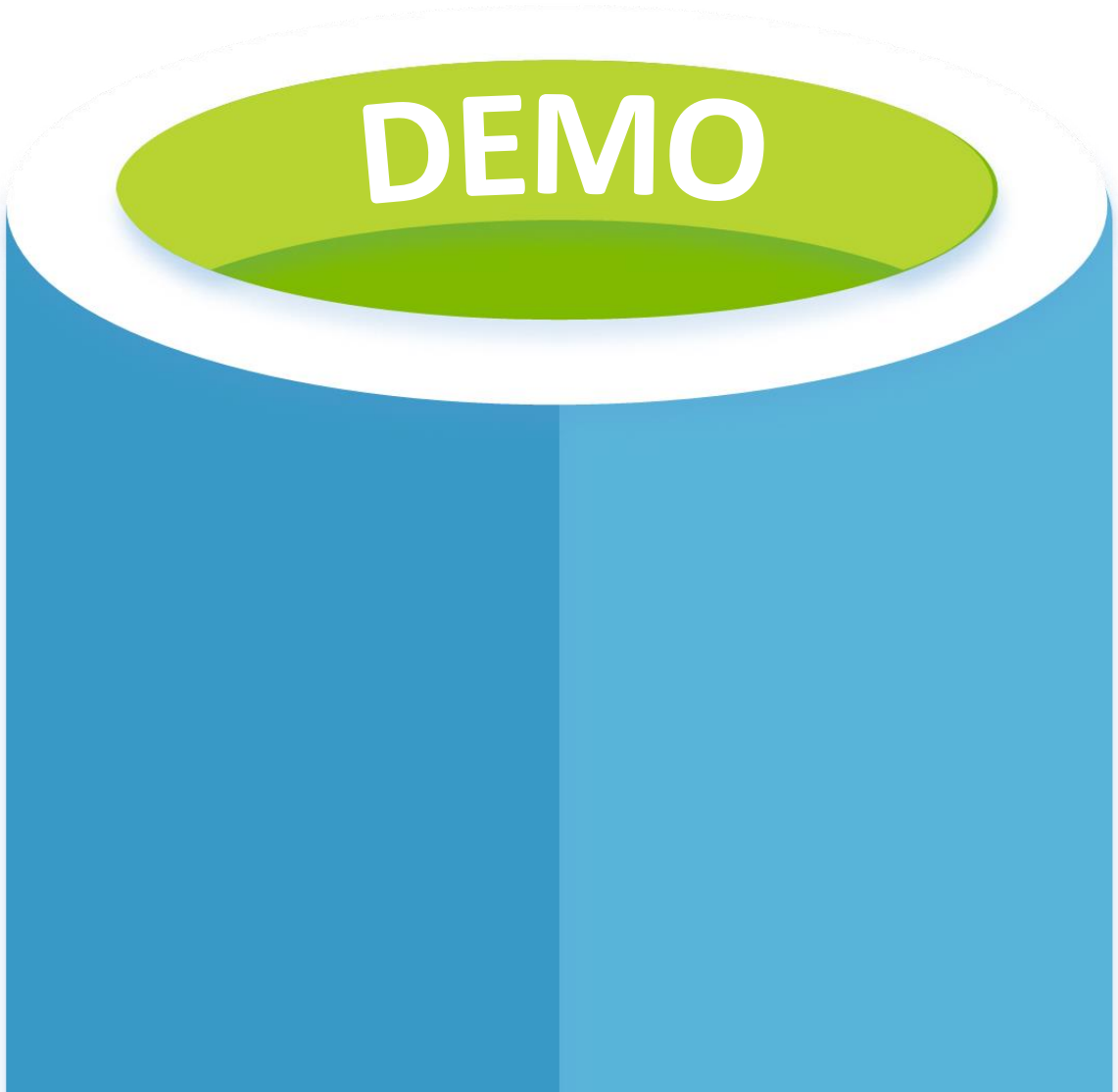
@{.....} ← Parameters
System Variables



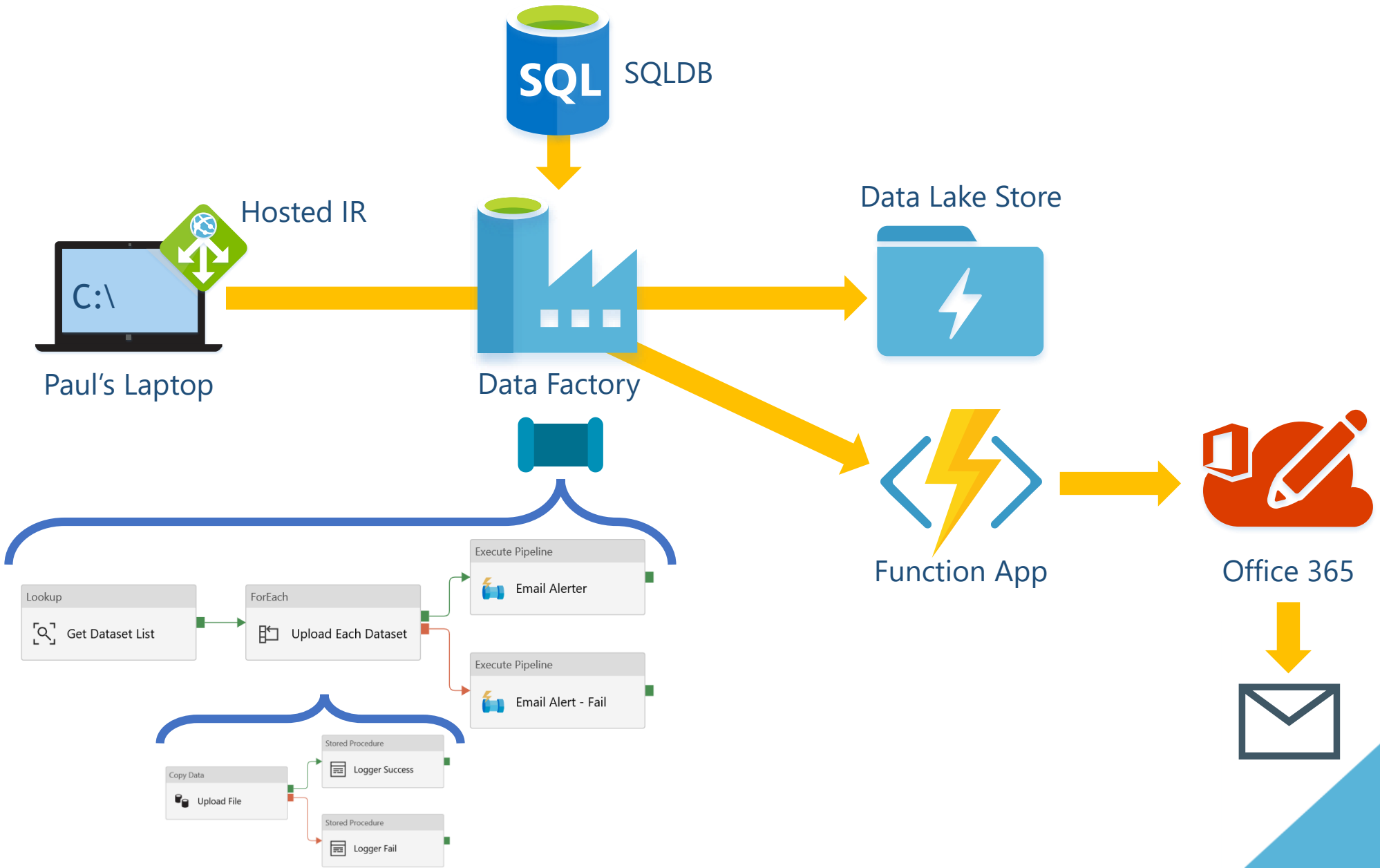
- Collection
- Conversation
- Date
- Logical
- Math
- String



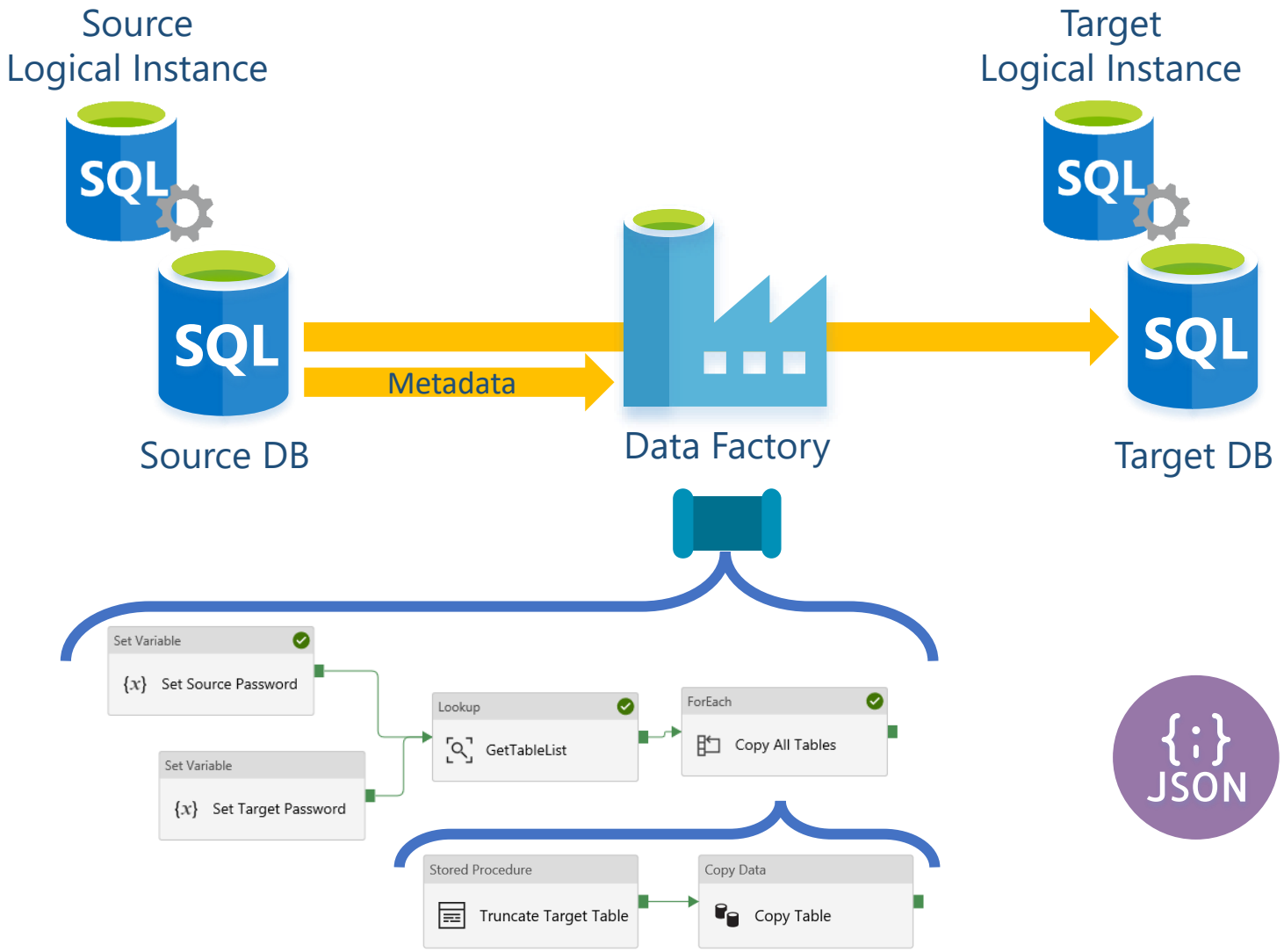
Dynamic Data Factory Pipelines



Demo Architecture 1



Demo Architecture 2



1x Linked Service

1x Dataset

https://mrpaulandrew.com/2018/11/15/using-data-factory-parameterised-linked-services/

Extending Data Factory with 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

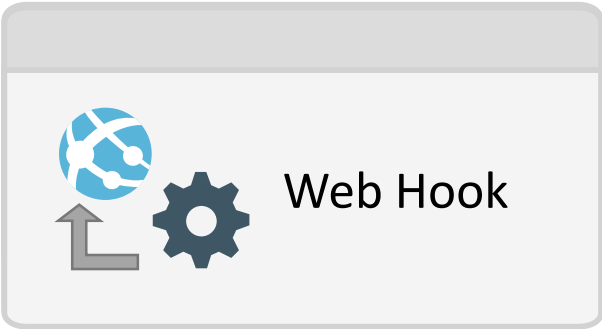
https://mrpaulandrew.com/2019/06/18/azure-data-factory-web-hook-vs-web-activity/

Web Hook vs Web Activity

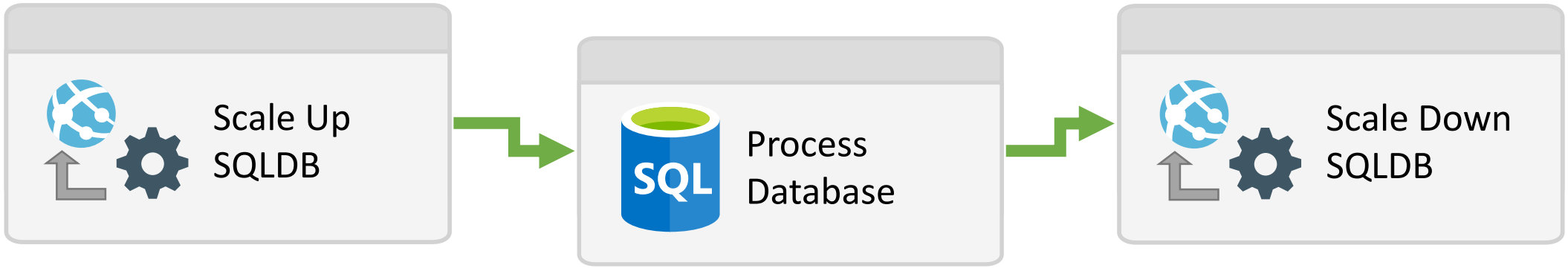
Asynchronous	Synchronous
Web	Web Hook
PUT POST GET DELETE	POST
1 Minute Timeout	Configurable Timeout
Retry Capabilities	No Retry
Linked Services Datasets	No Artifact Support
One Way Call	Call Back URL

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

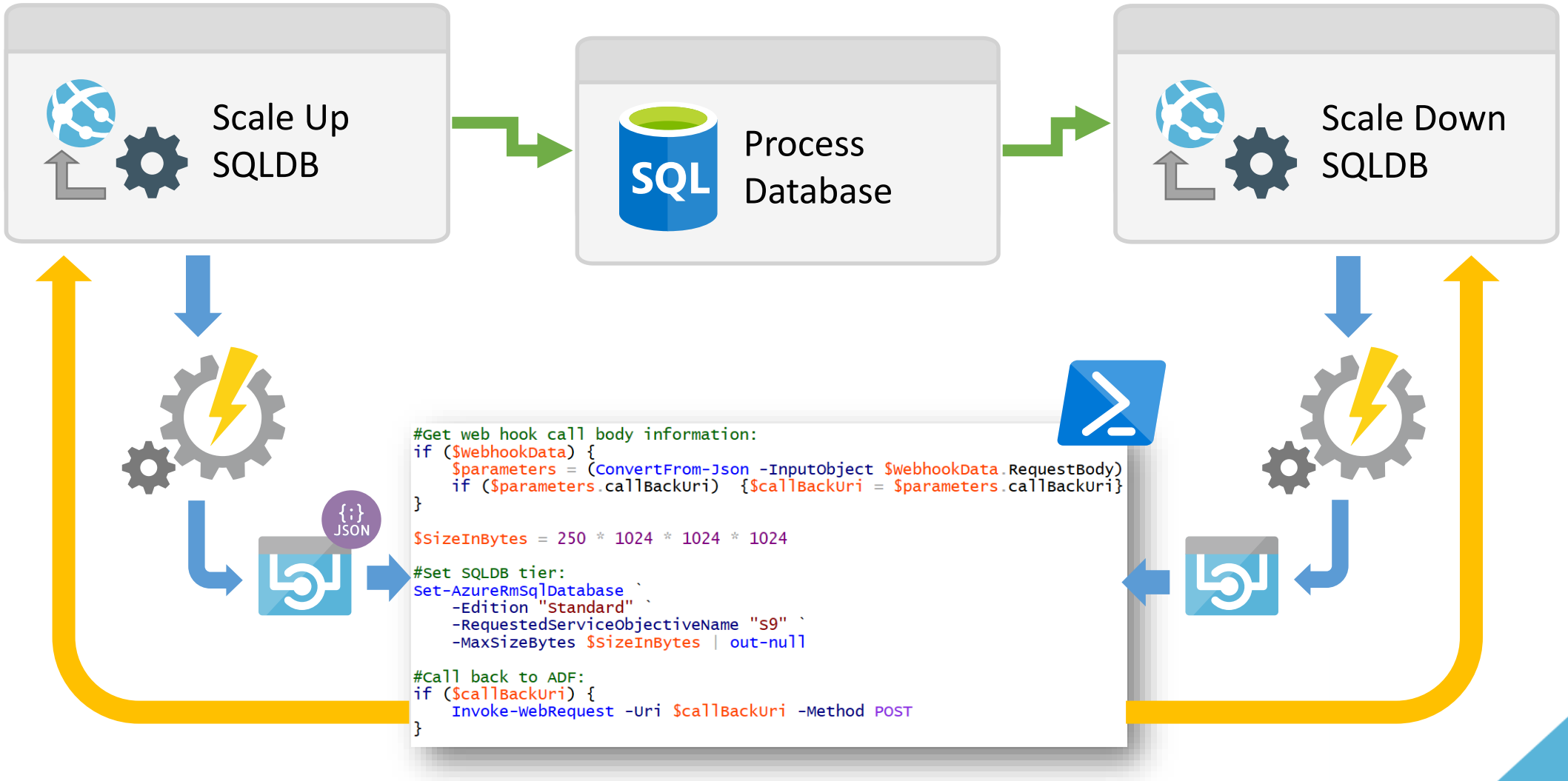
Web Hook vs Web Activity



Web Hook vs Web Activity

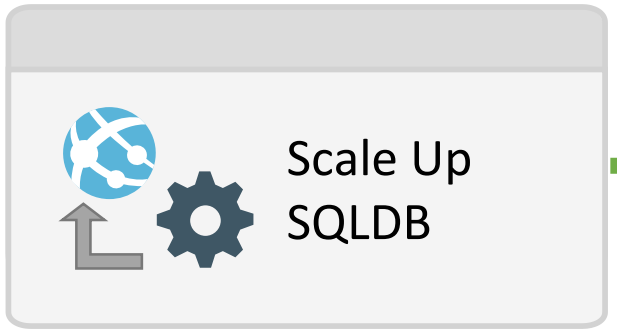


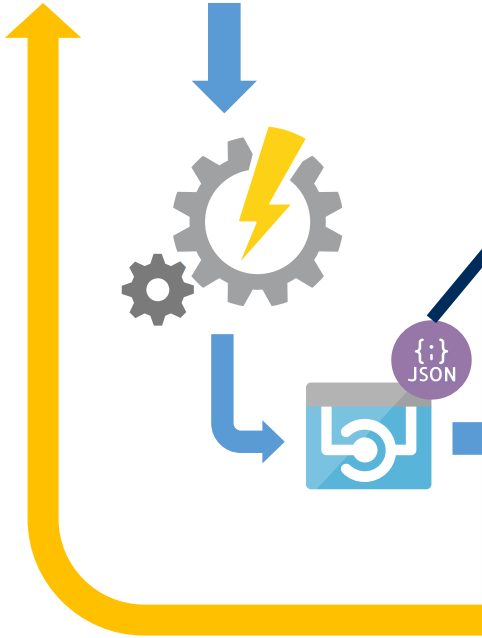
Web Hook vs Web Activity




```
#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






```
#Get web hook call
if ($webhookData) {
    $parameters = $webhookData.RequestBody | ConvertFrom-Jsonn
    if ($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
}
```

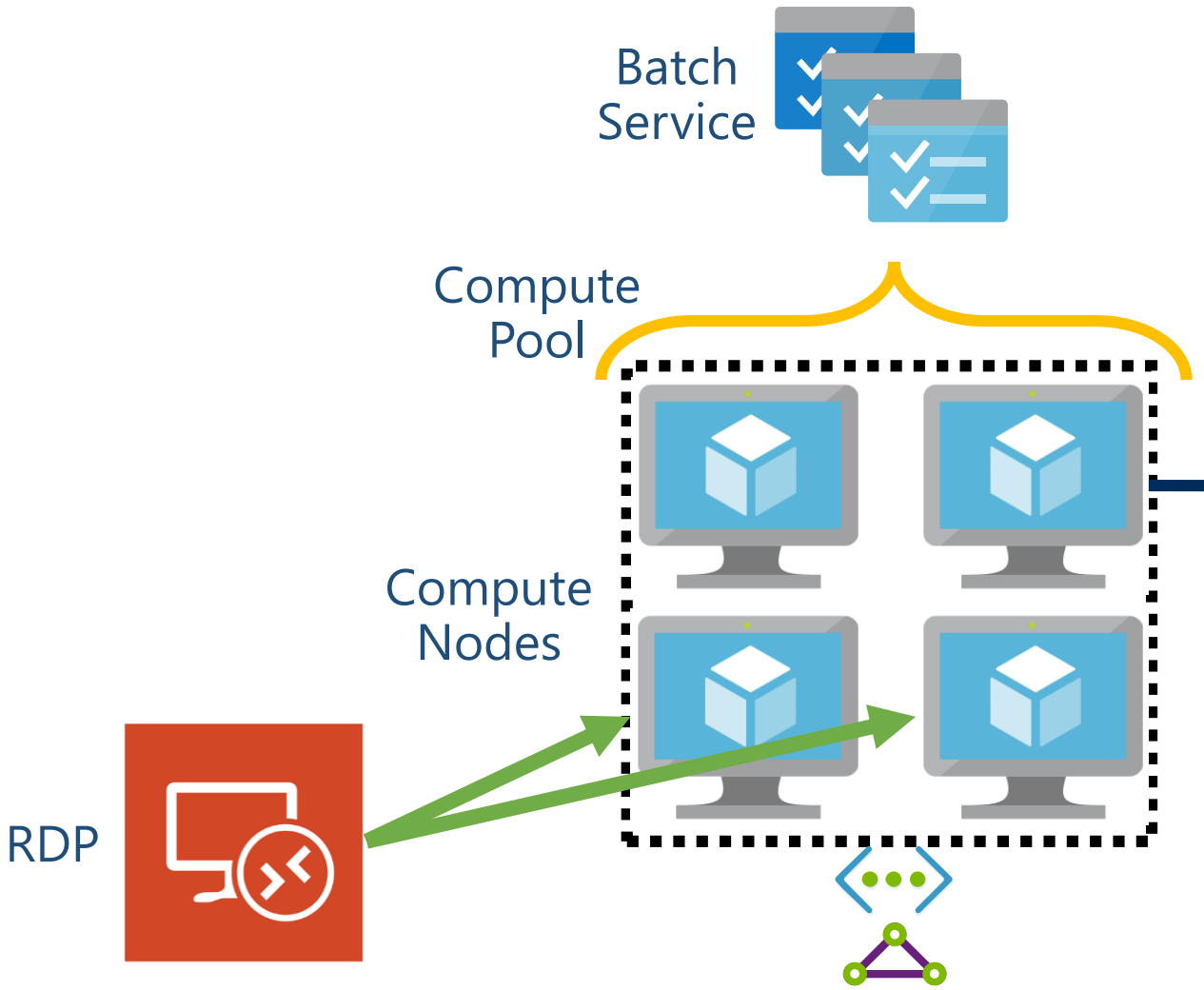
```
{
    "WebhookName":"SQLDBScaler",
    "RequestBody":{
        "count": 1,
        "value": [
            {
                "EnvironmentName": "AzureCloud",
                "ResourceGroupName": "ResourceGroup01",
                "ServerName": "SQLInstance01",
                "DatabaseName": "Database01",
                "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

Extending Data Factory
with Custom Activities

Azure Batch Service

Scale out compute delivered using PaaS technology with IaaS underneath.



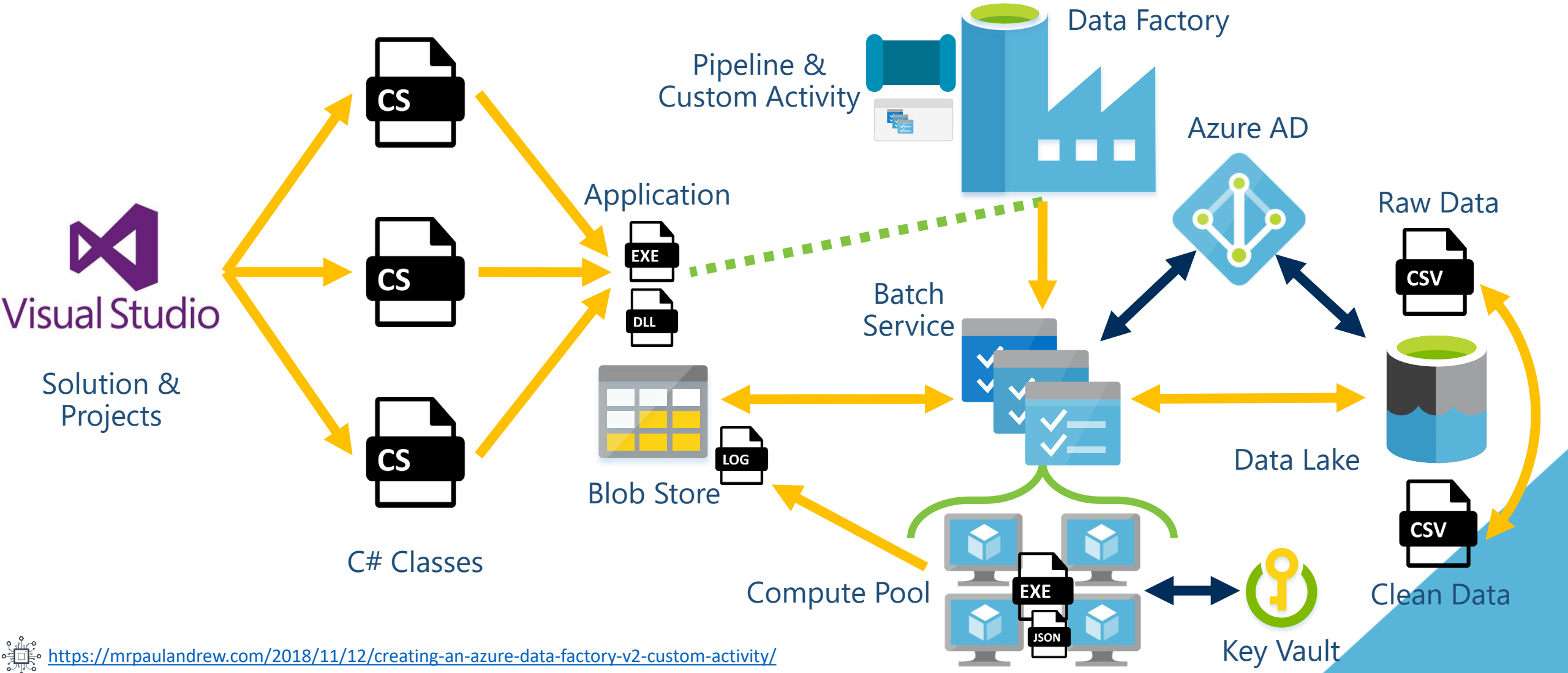
VM node size set per compute pool:

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.



https://mrpaulandrew.com/2018/11/12/creating-an-azure-data-factory-v2-custom-activity/

Extensibility Conclusions

C#

Do Stuff

10 minutes of execution
unless using durable functions

C#

Custom
Activity

Auto scale out compute &
Scale up per compute node

REST

Web

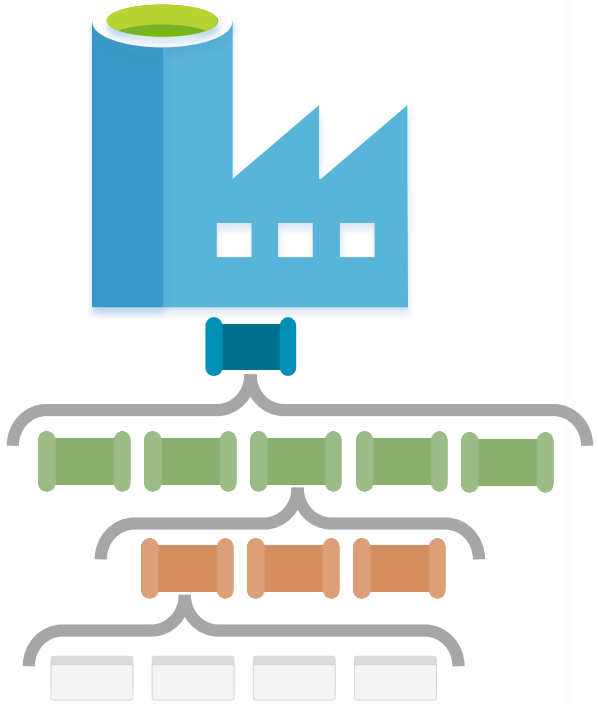
Asynchronous, limited control

REST

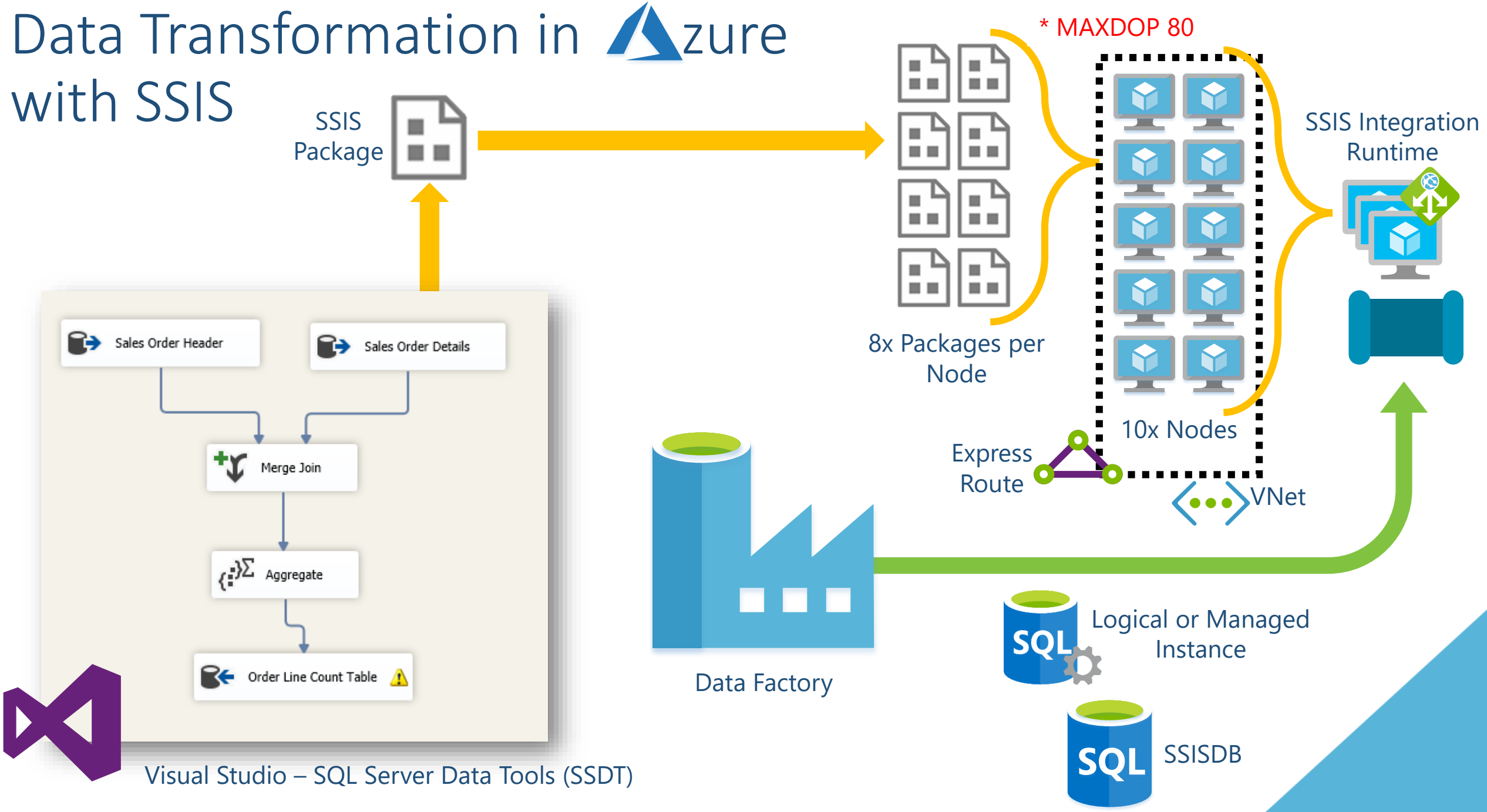
Web Hook

Synchronous, call back control

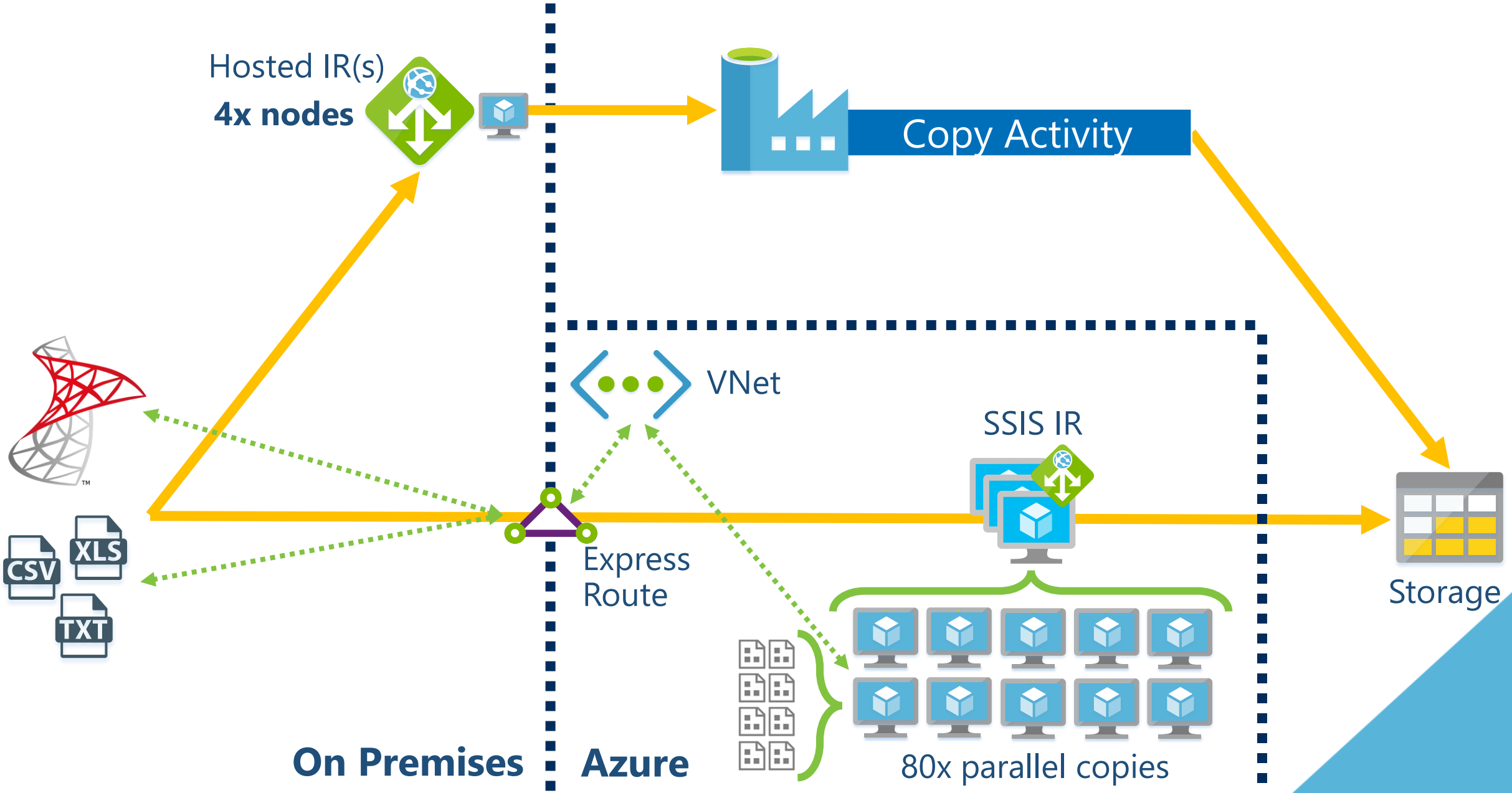
Scale Out ~~Execution~~ Everything!



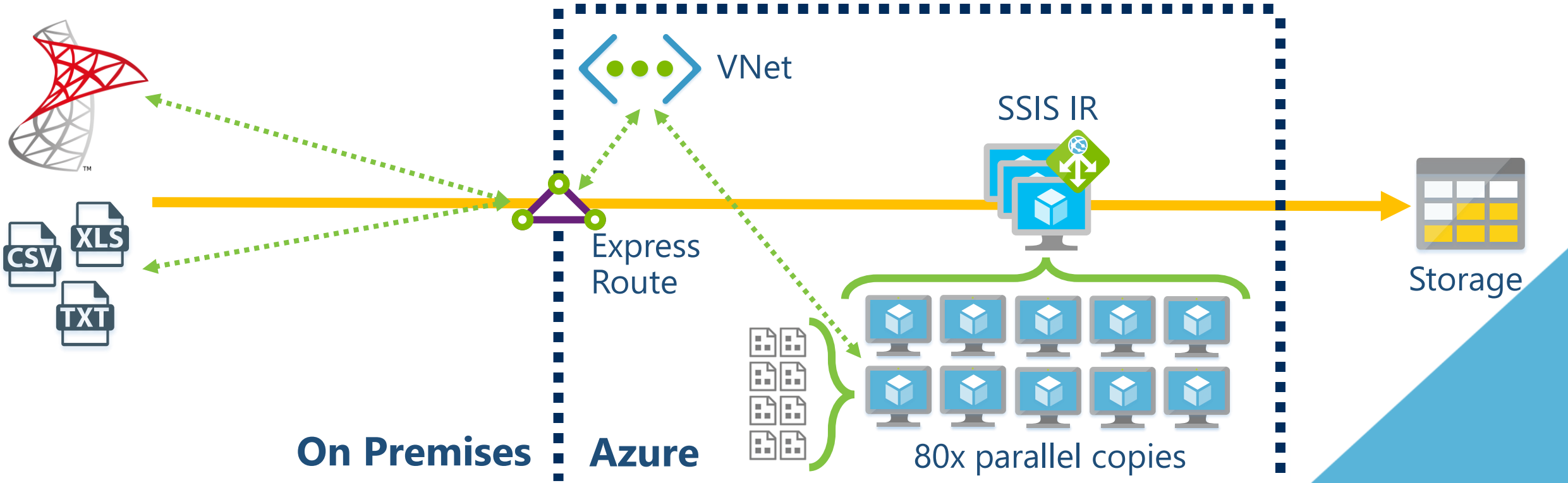
Data Transformation in Azure with SSIS



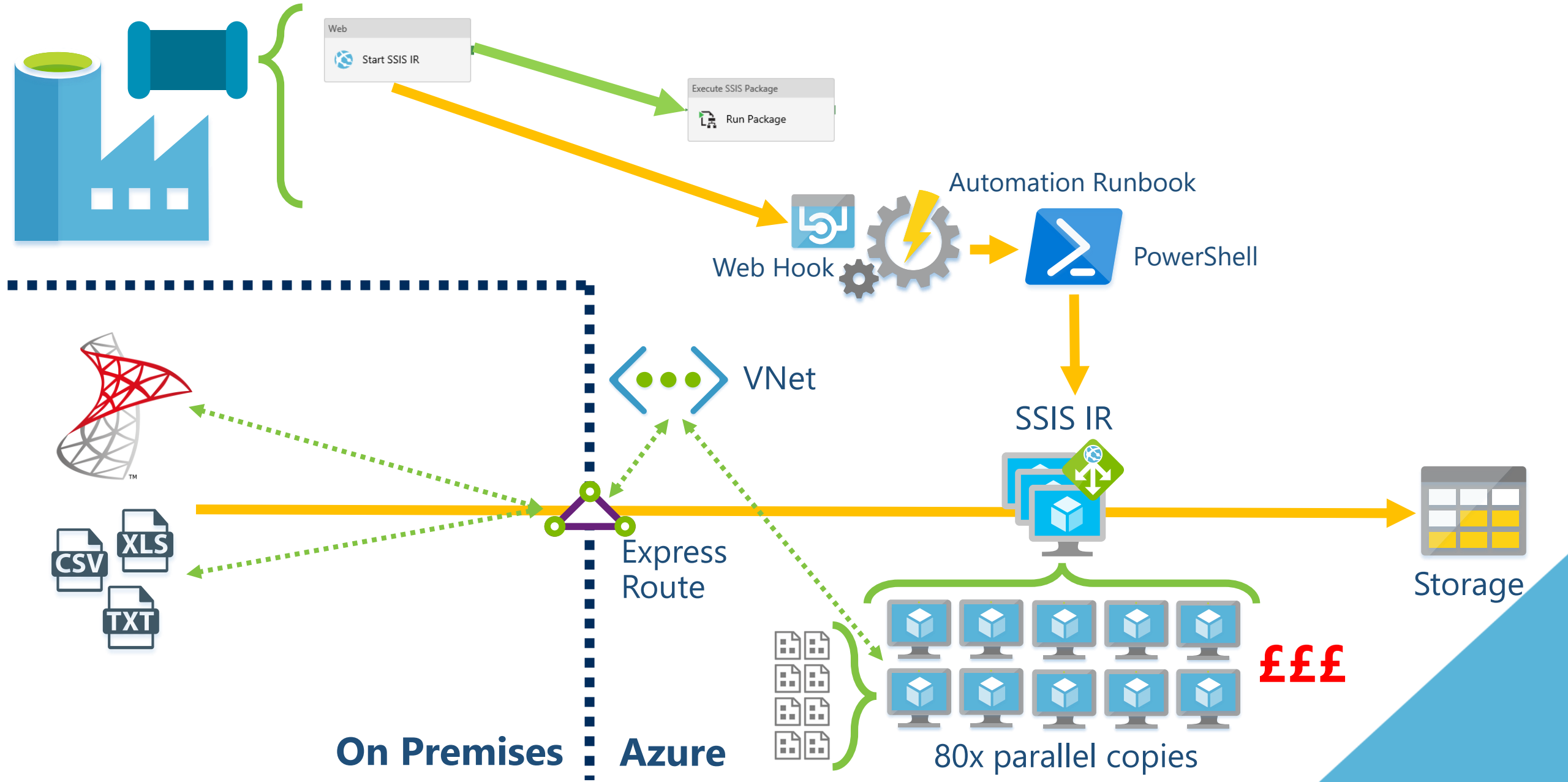
The SSIS IR vs Hosted IR with Express Route



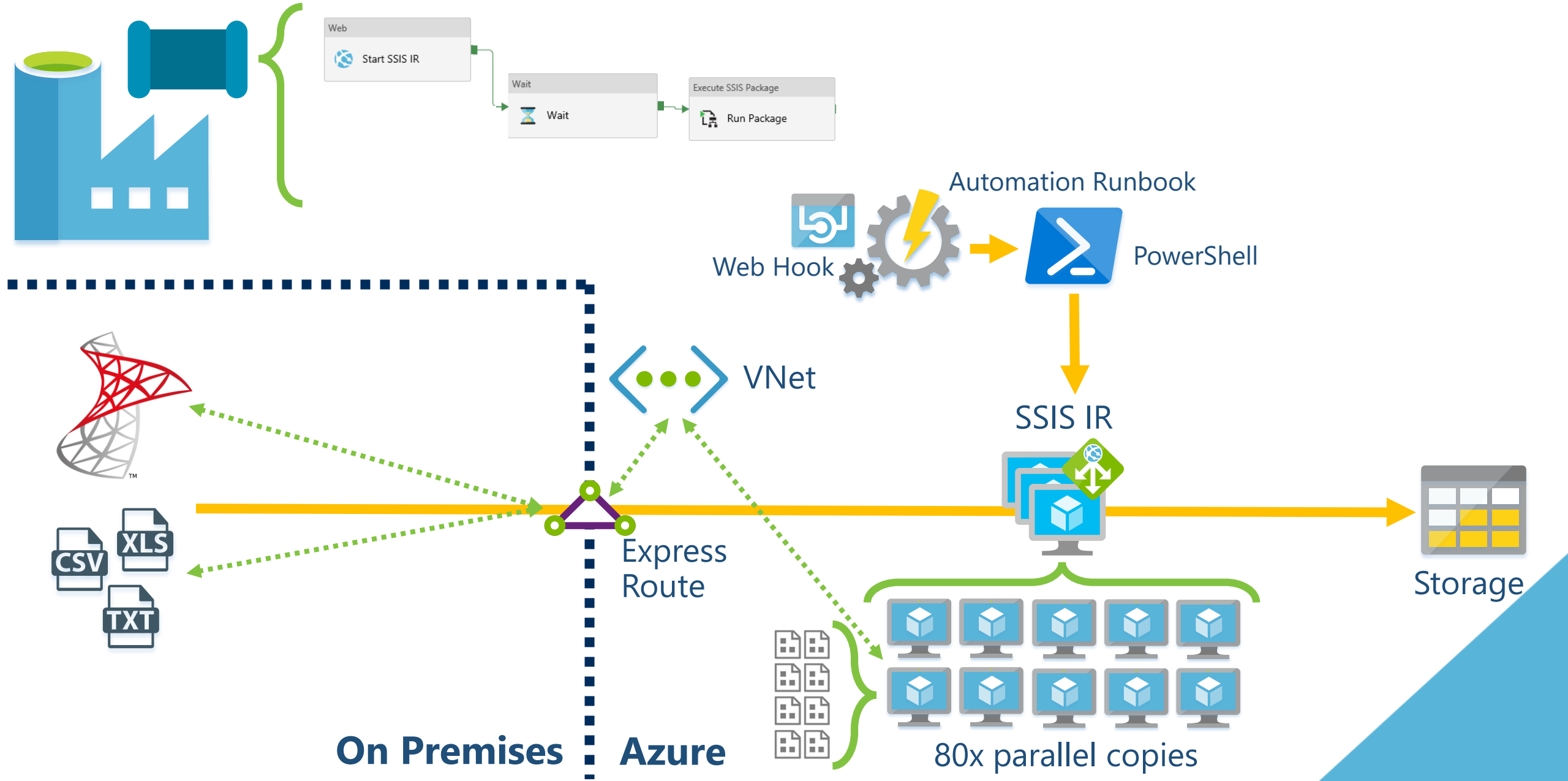
The SSIS IR Start/Stop



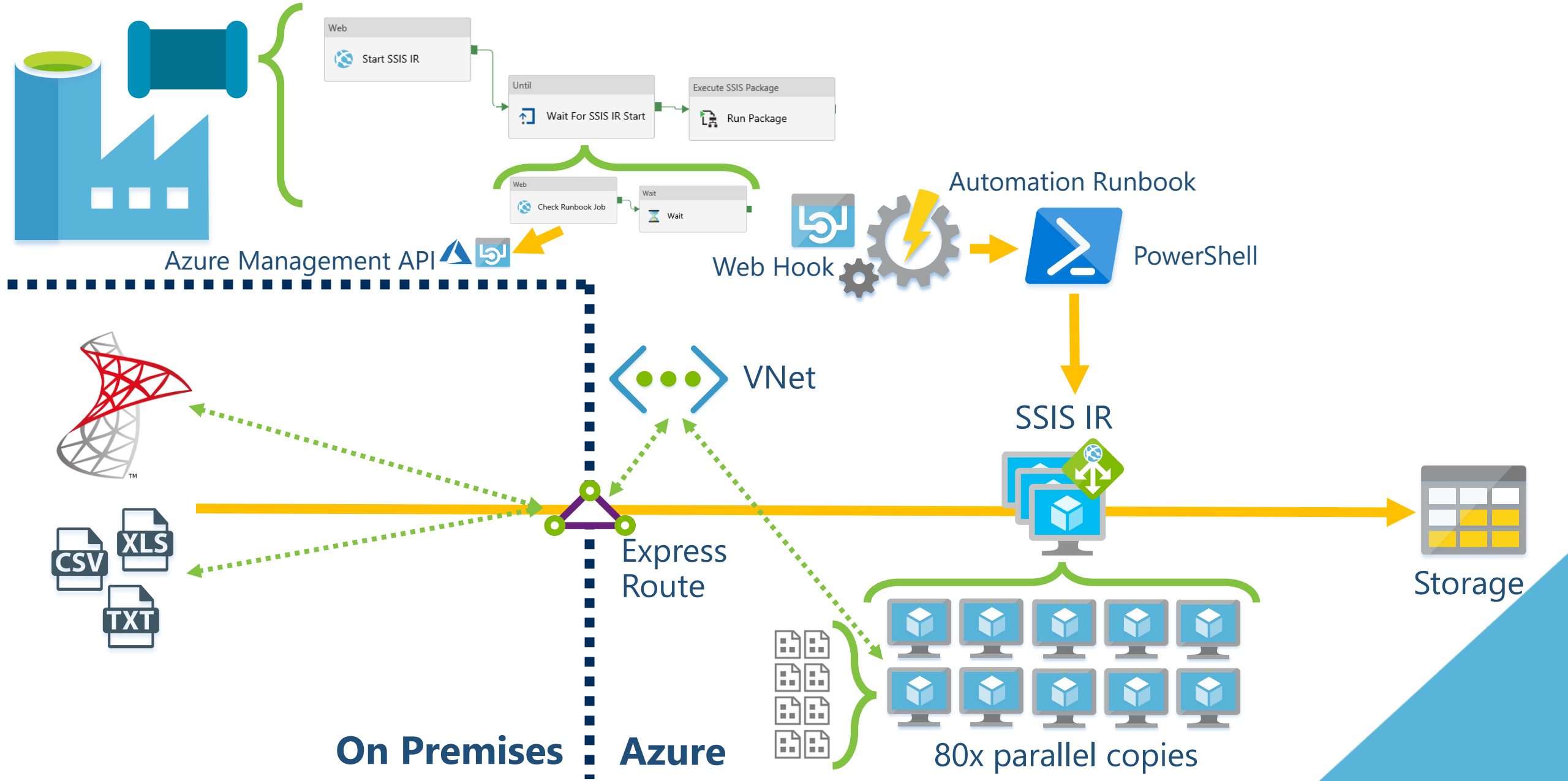
The SSIS IR Start/Stop



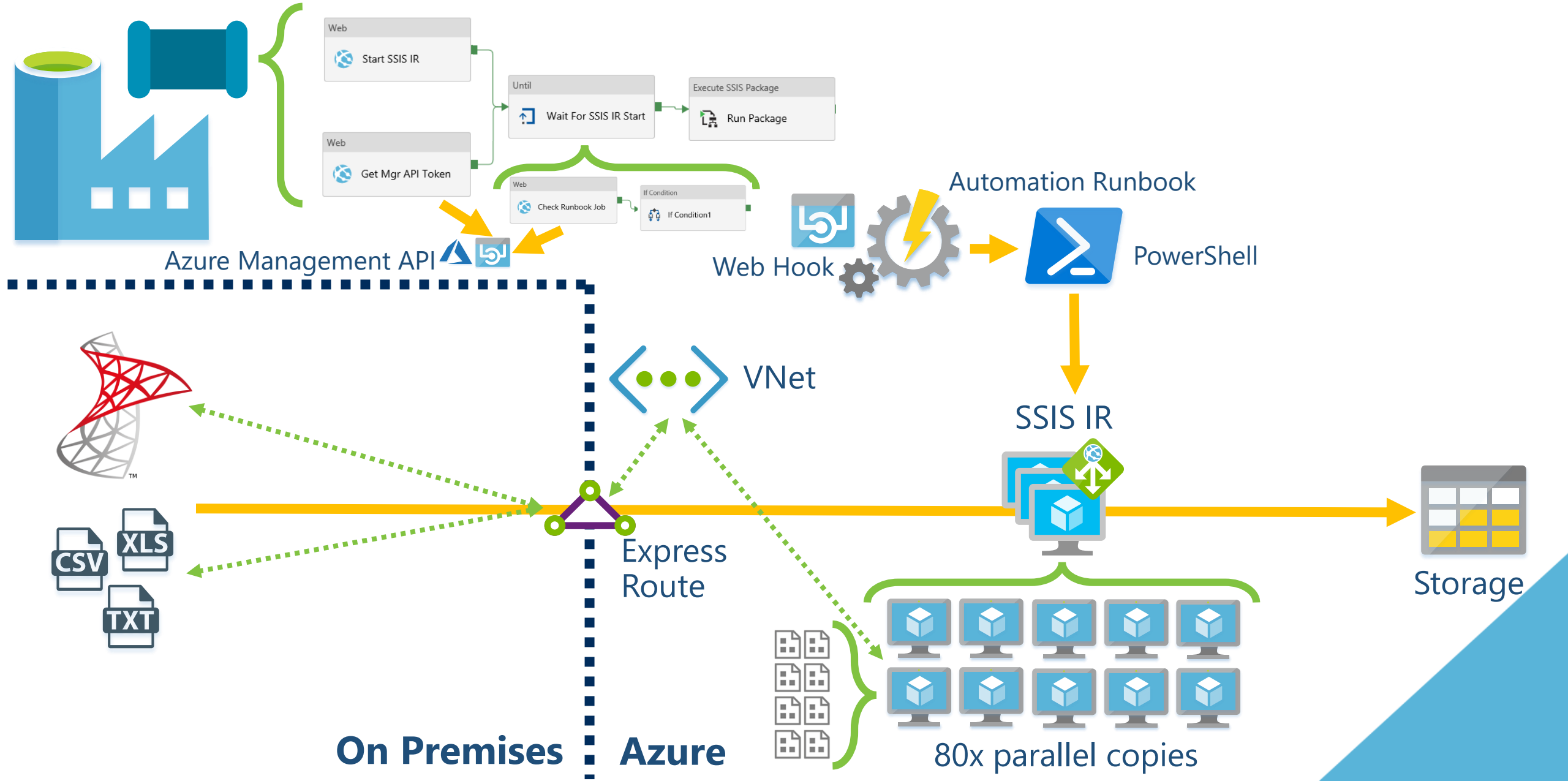
The SSIS IR Start/Stop



The SSIS IR Start/Stop

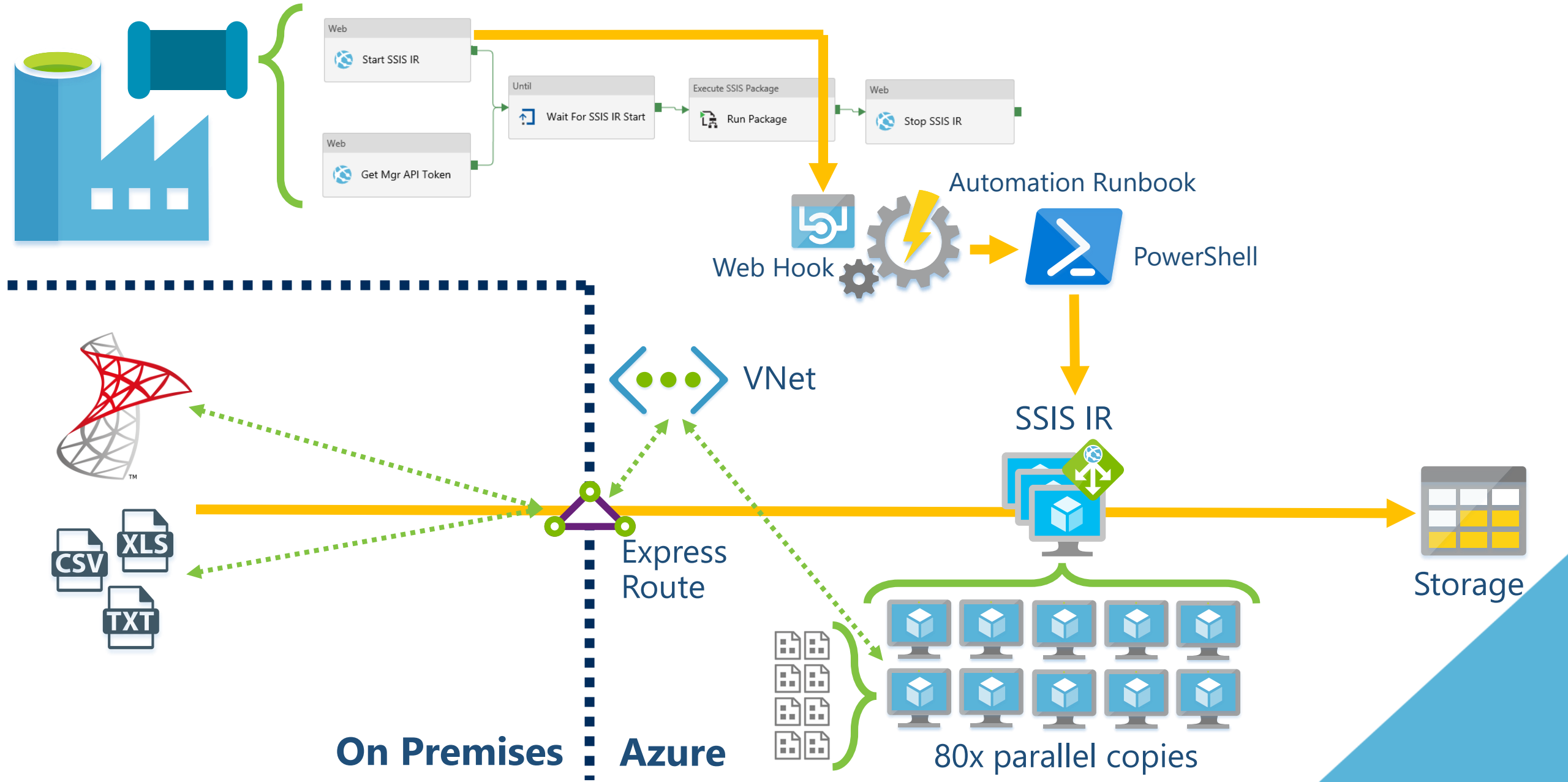


The SSIS IR Start/Stop

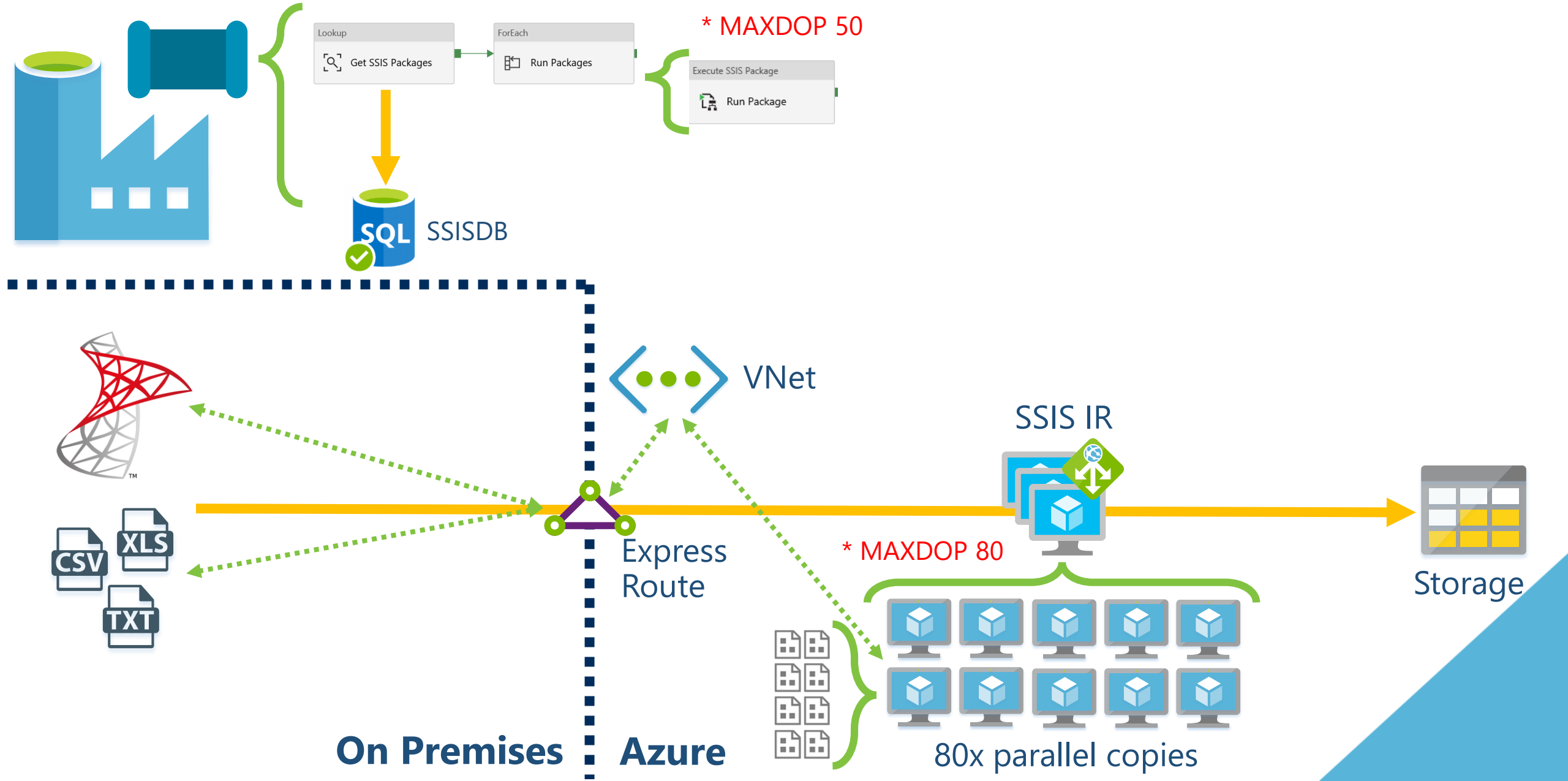


The SSIS IR Start/Stop

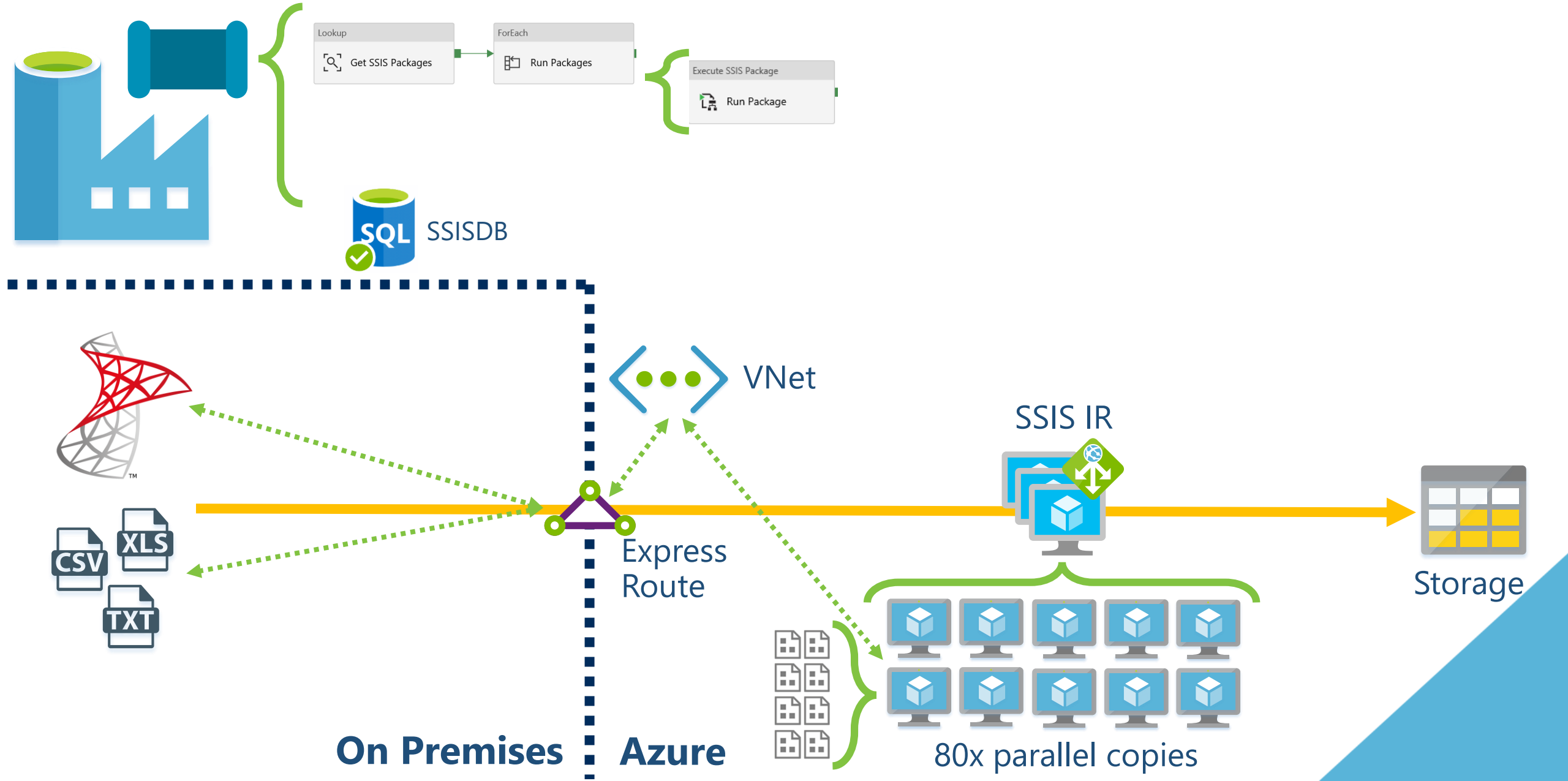
https://mrpaulandrew.com/2018/06/19/how-to-control-the-ssis-ir-from-your-adfv2-pipelines/



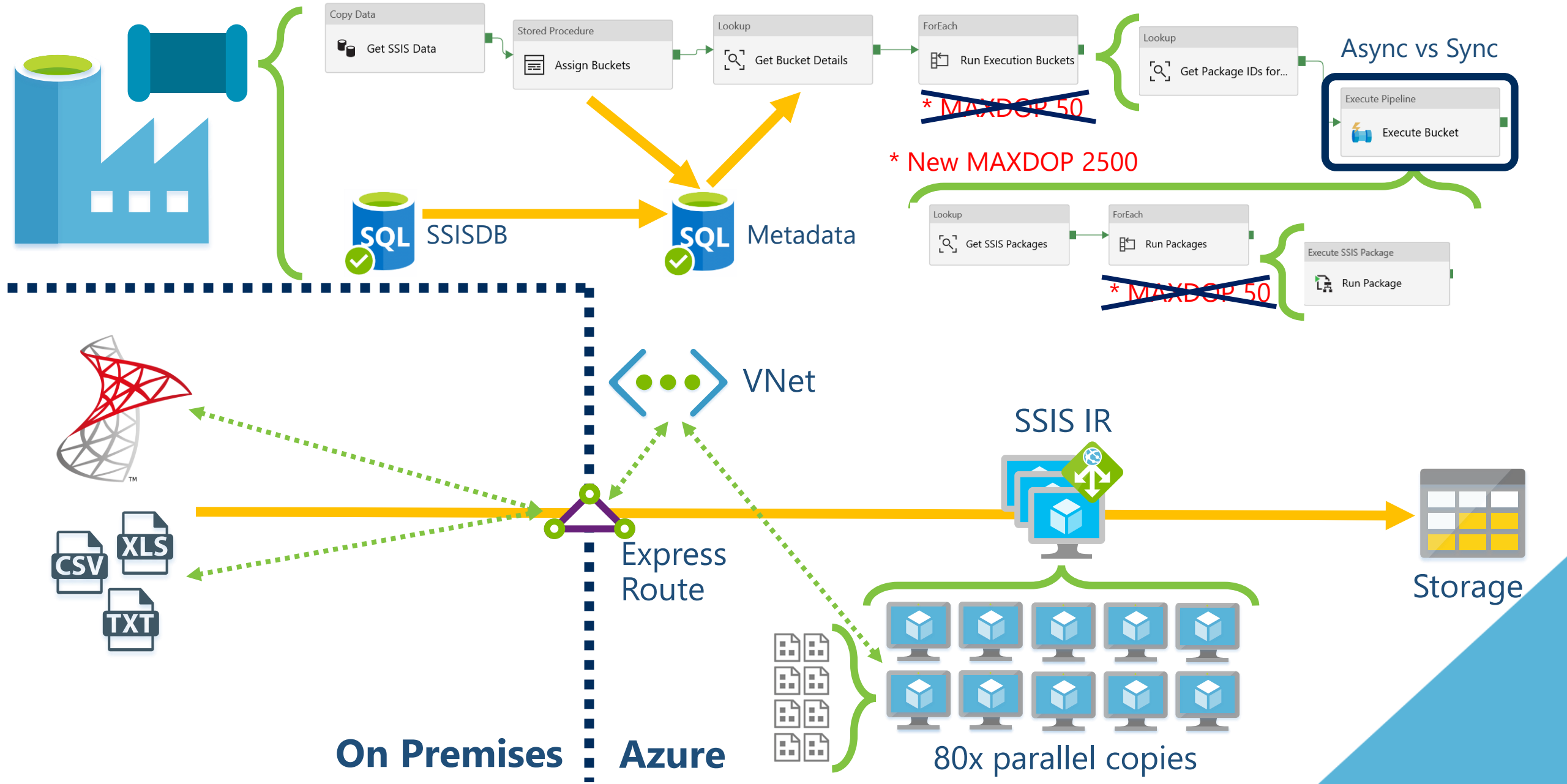
The SSIS IR Parallelism



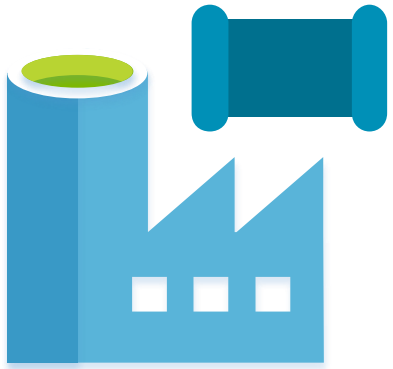
The SSIS IR Parallelism



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

Copy Data

Package IDs for...

Execute Pipeline

Execute Bucket

Execute SSIS Package

Run Package

50

Async vs Sync

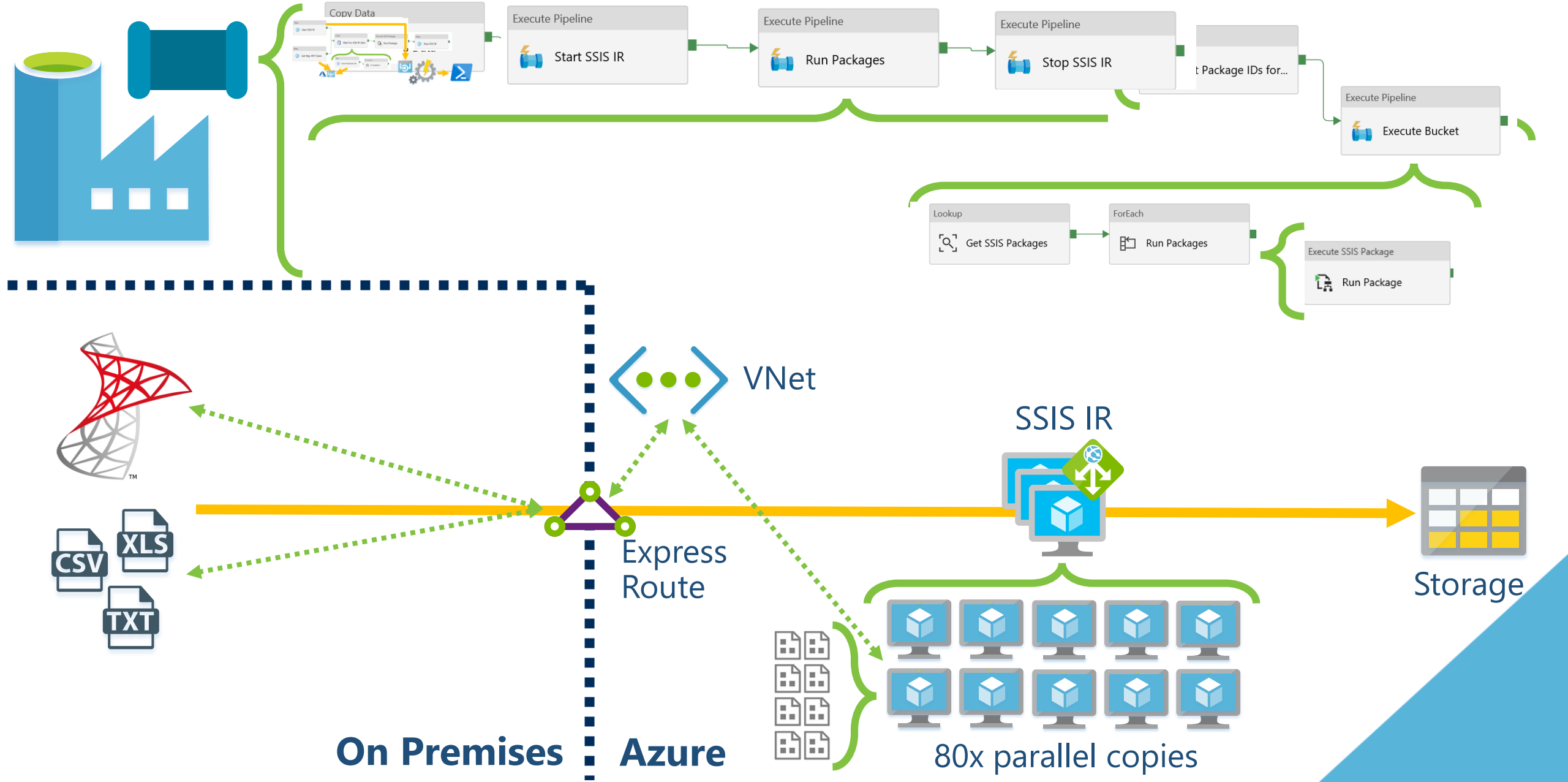
Storage

On Premises : Azure

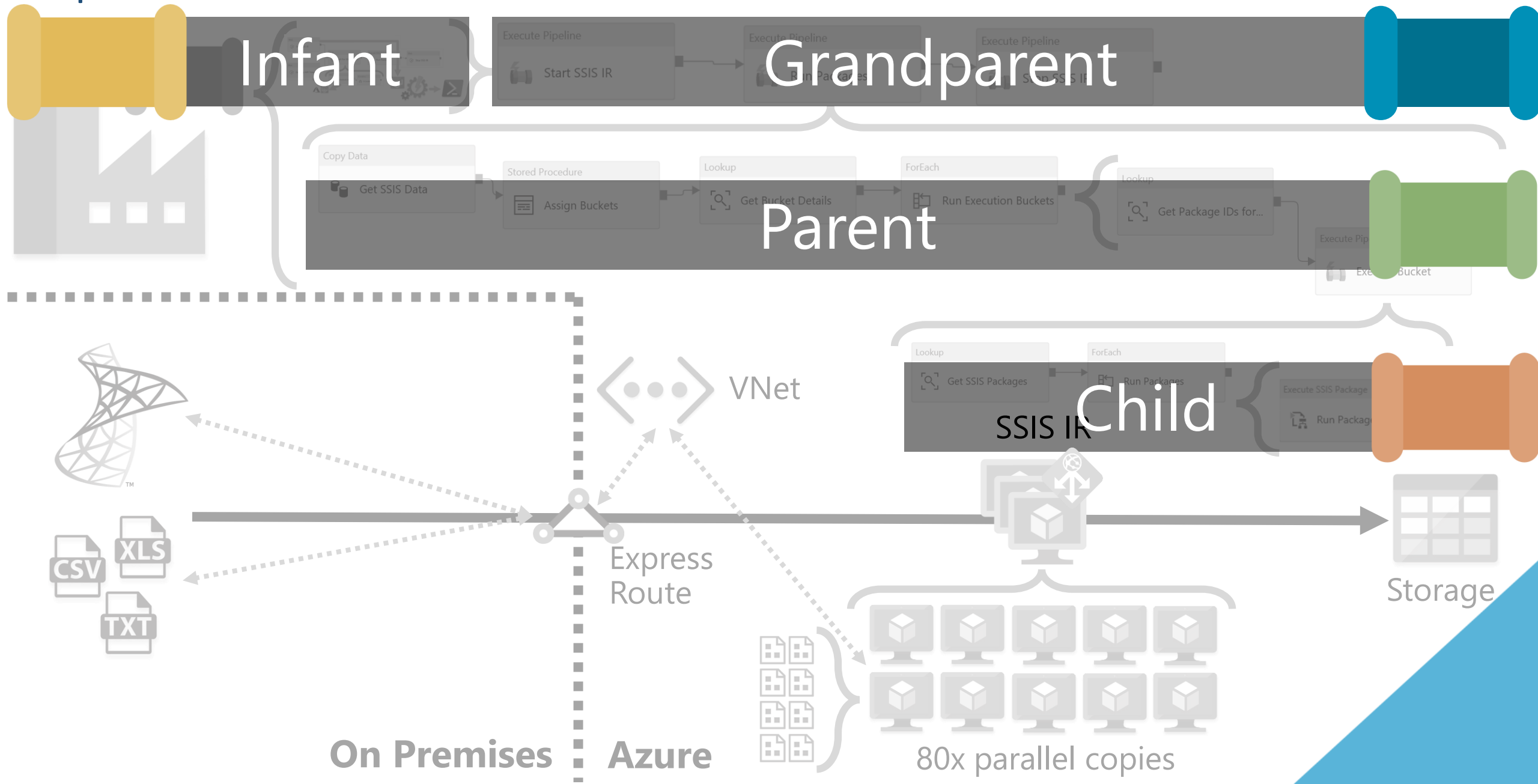
80x parallel copies

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

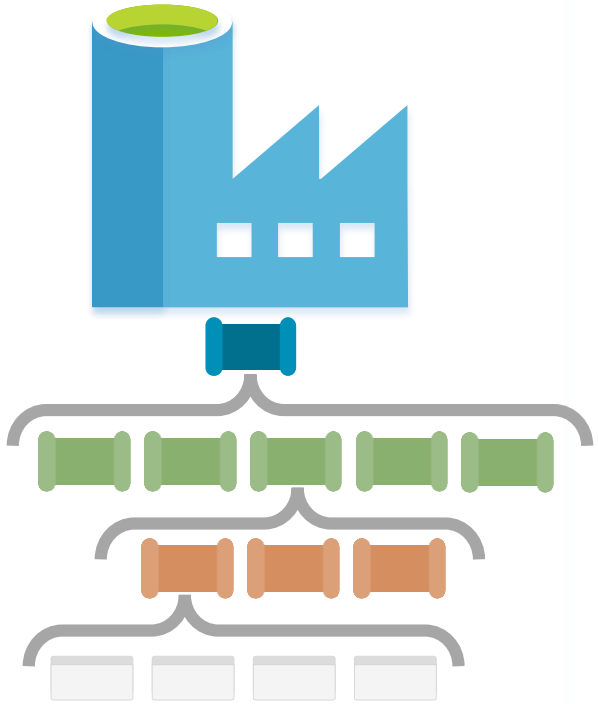
SSIS IR & Package Complete Orchestration Solution



Pipeline Hierarchies



Scale Out ~~Execution~~ Everything!



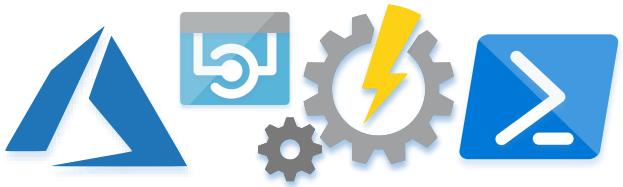
Pipeline Hierarchies – Design Pattern

Grand-parent

- **Attached triggers**, top level bootstrap.
- Group processes and **control dependencies**.

Parent

- **Control resources**, scaling and state.
- Manage **parallelism, stage 1**.



Child

- Call execution **activities**.
- Manage **parallelism, stage 2**.

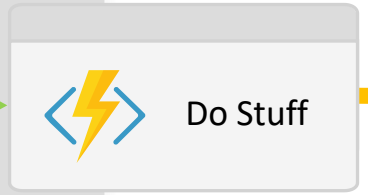
Infant

- Utilities, **boiler plate** operations.
- Error handler.

https://mrpaulandrew.com/2019/09/25/azure-data-factory-pipeline-hierarchies-generation-control/

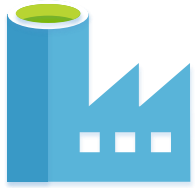
Solution Bootstrapping

Bootstrapping



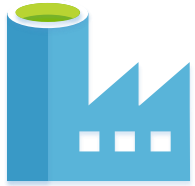
Functions

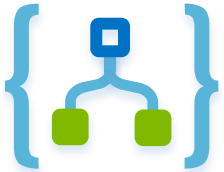

```
Console.WriteLine("Creating pipeline run...");
Dictionary<string, object> parameters = new Dictionary<string,
object> {
                    {"inputPath", inputBlobPath }, {
"outputPath", outputBlobPath }
                    };
CreateRunResponse runResponse =
client.Pipelines.CreateRunWithHttpMessagesAsync(resourceGroup,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



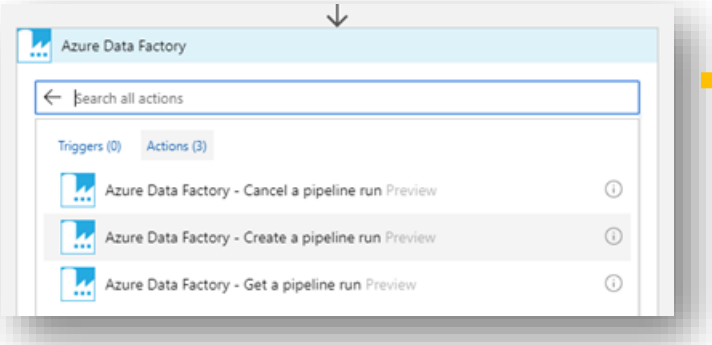
Automation Runbooks

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

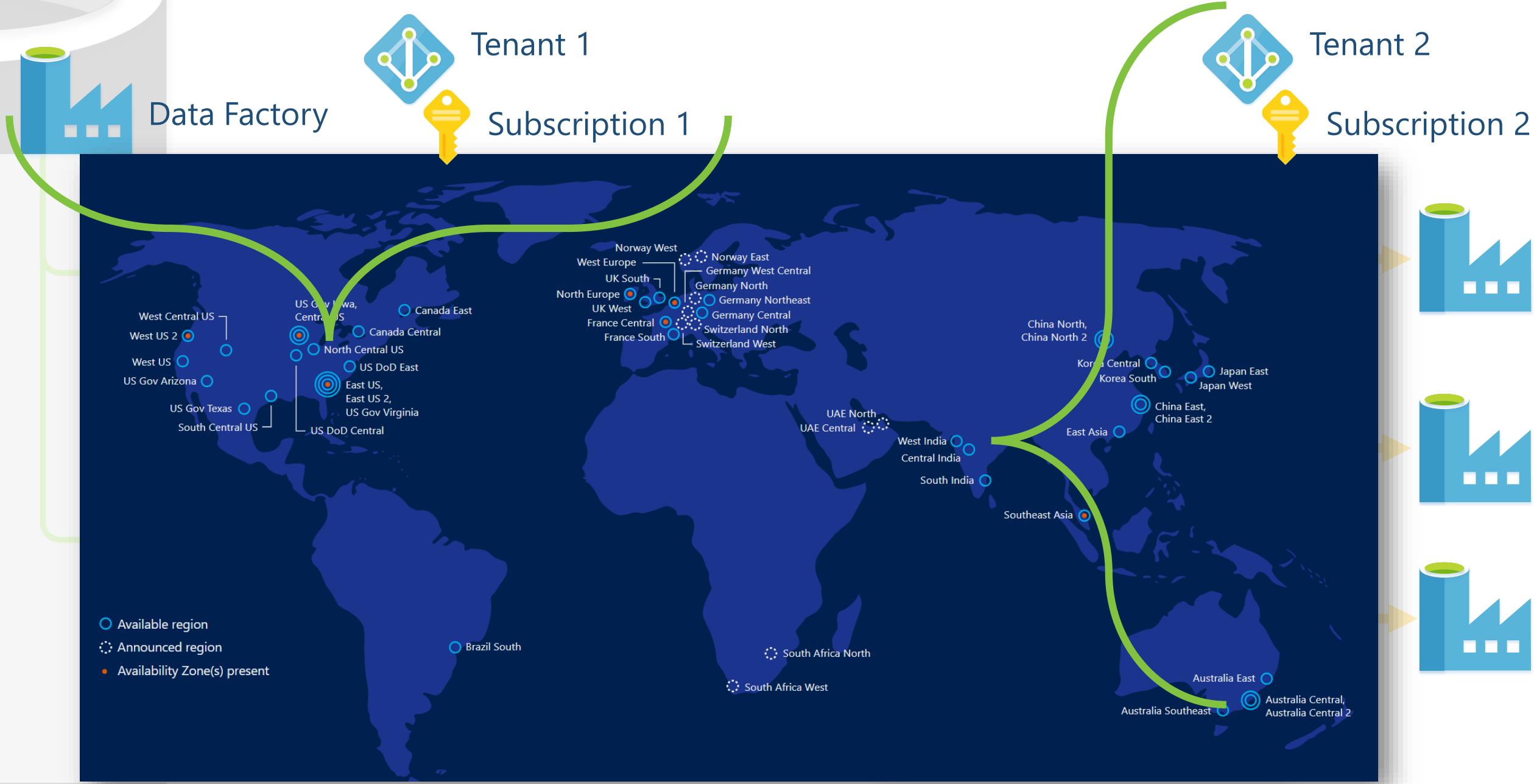




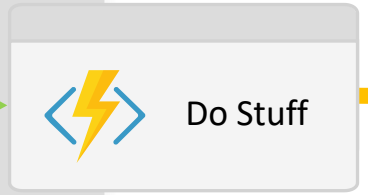
Logic Apps



Bootstrapping

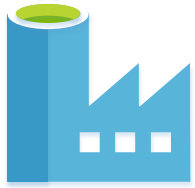


Bootstrapping



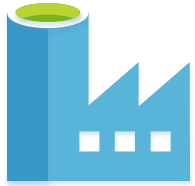
Functions

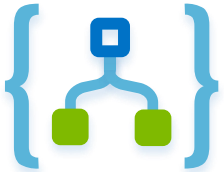

```
Console.WriteLine("Creating pipeline run...");
Dictionary<string, object> parameters = new Dictionary<string,
object> {
                    {"inputPath", inputBlobPath }, {
"outputPath", outputBlobPath }
                    };
CreateRunResponse runResponse =
client.Pipelines.CreateRunWithHttpMessagesAsync(resourceGroup,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



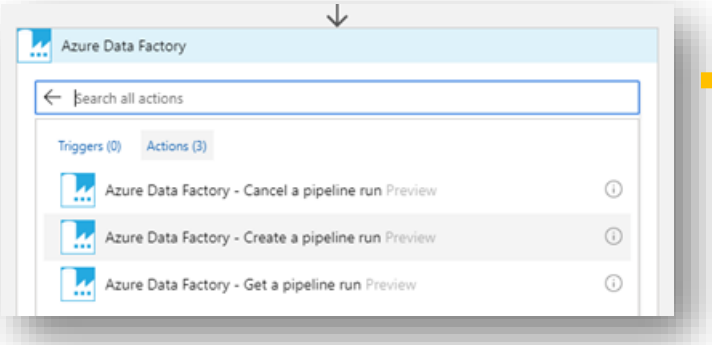
Automation Runbooks

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

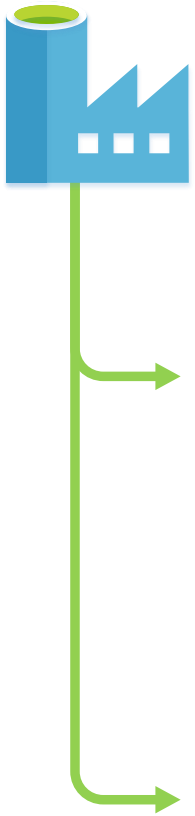




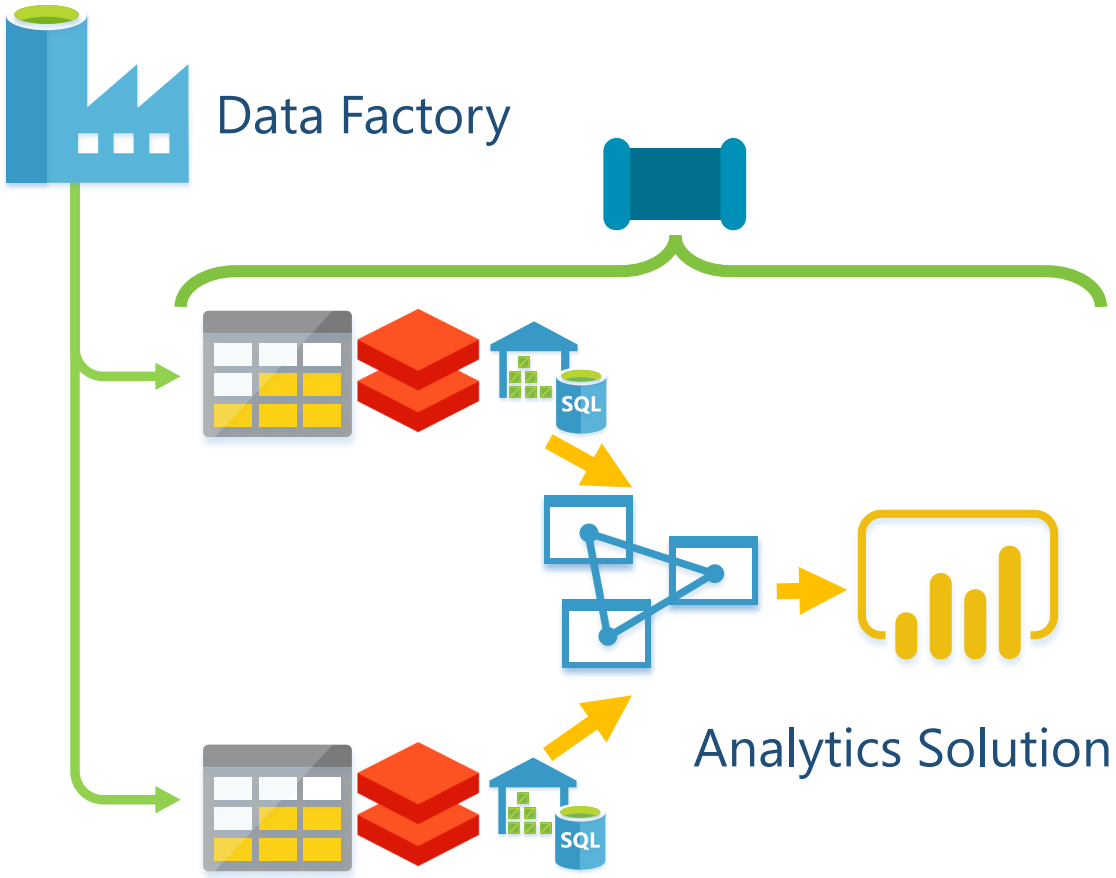
Logic Apps



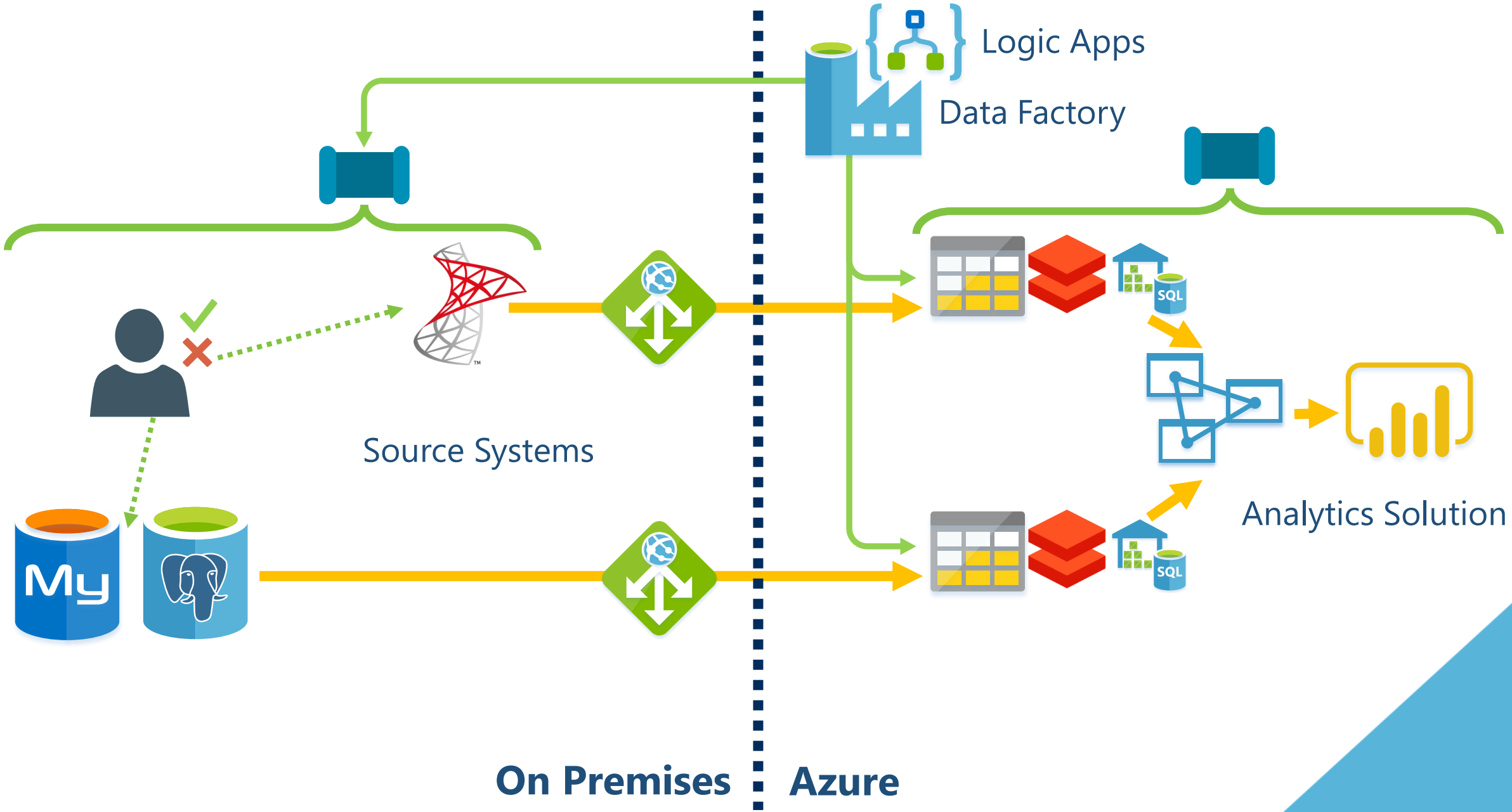
Bootstrapping – Wider Analytics Solution



Bootstrapping – Wider Analytics Solution

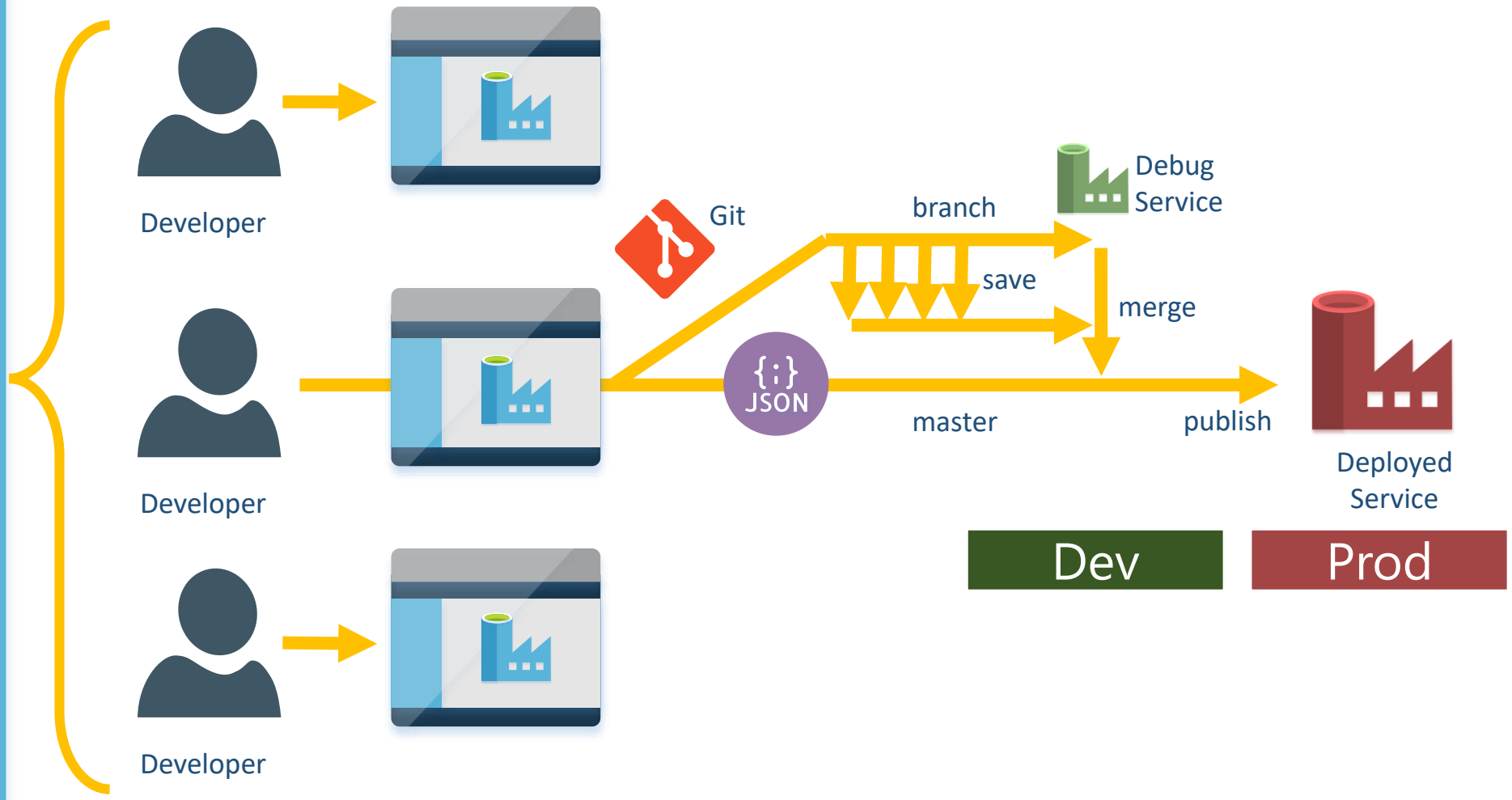


Bootstrapping – Wider Analytics Solution

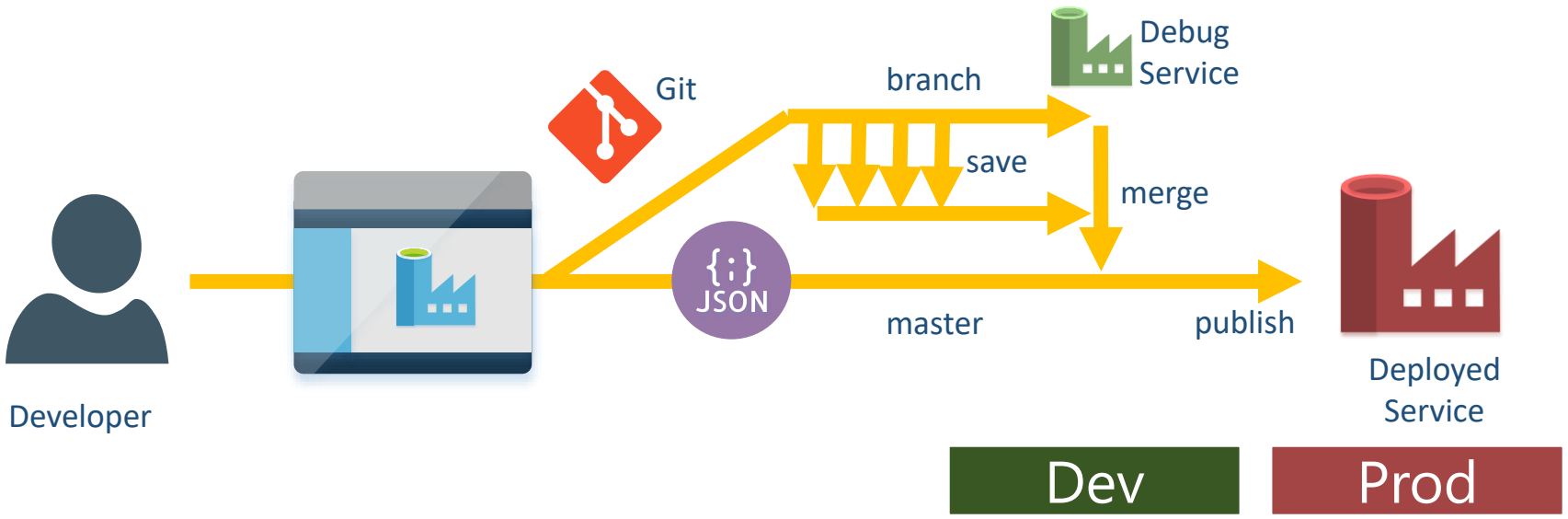


Data Factory DevOps – CI/CD

Data Factory Continuous Integration

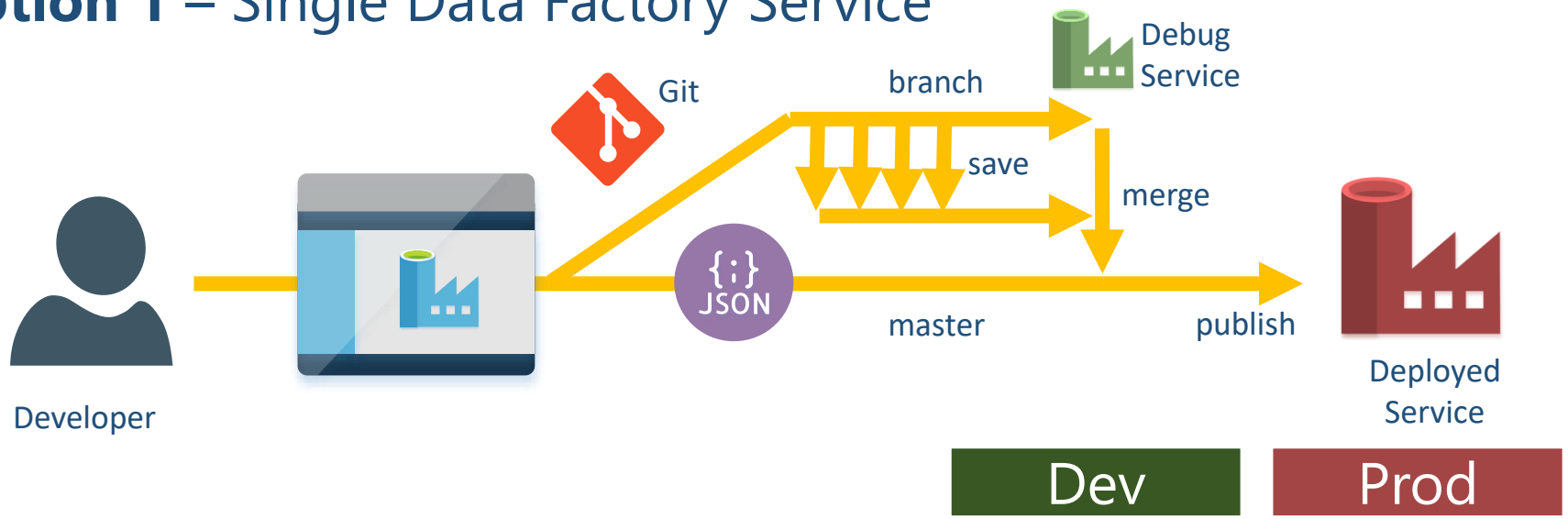


Data Factory Continuous Delivery

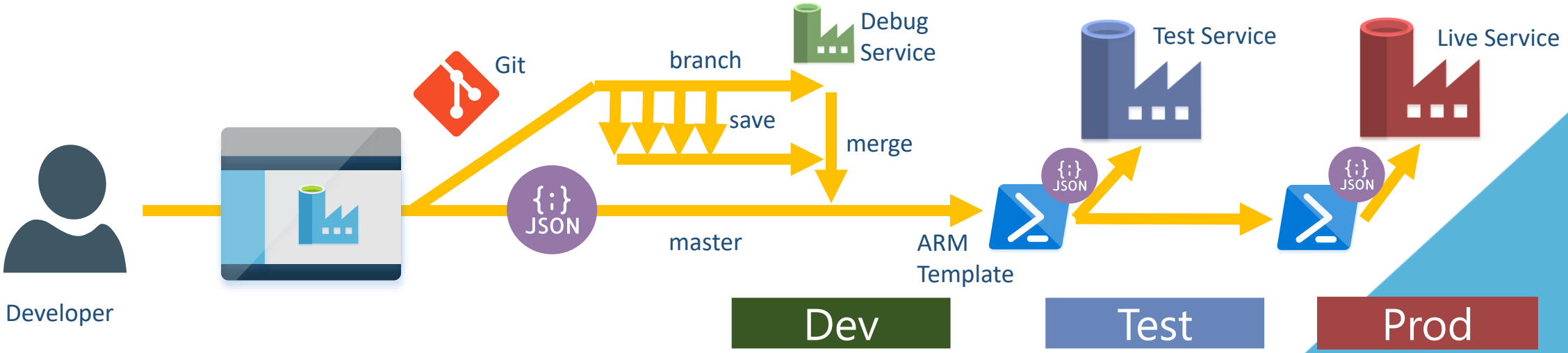


Data Factory Continuous Delivery

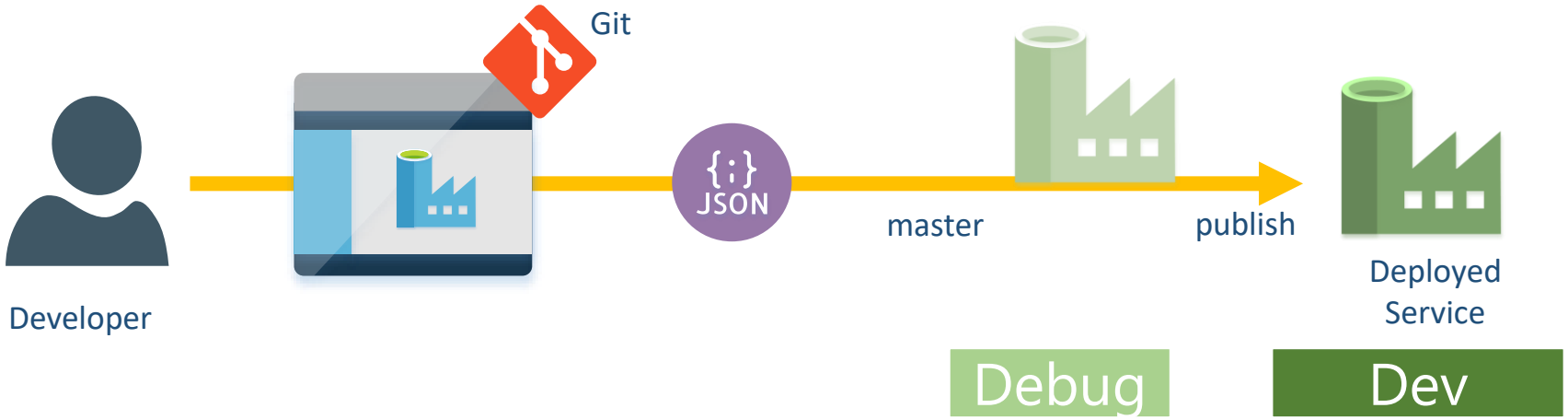
Option 1 – Single Data Factory Service



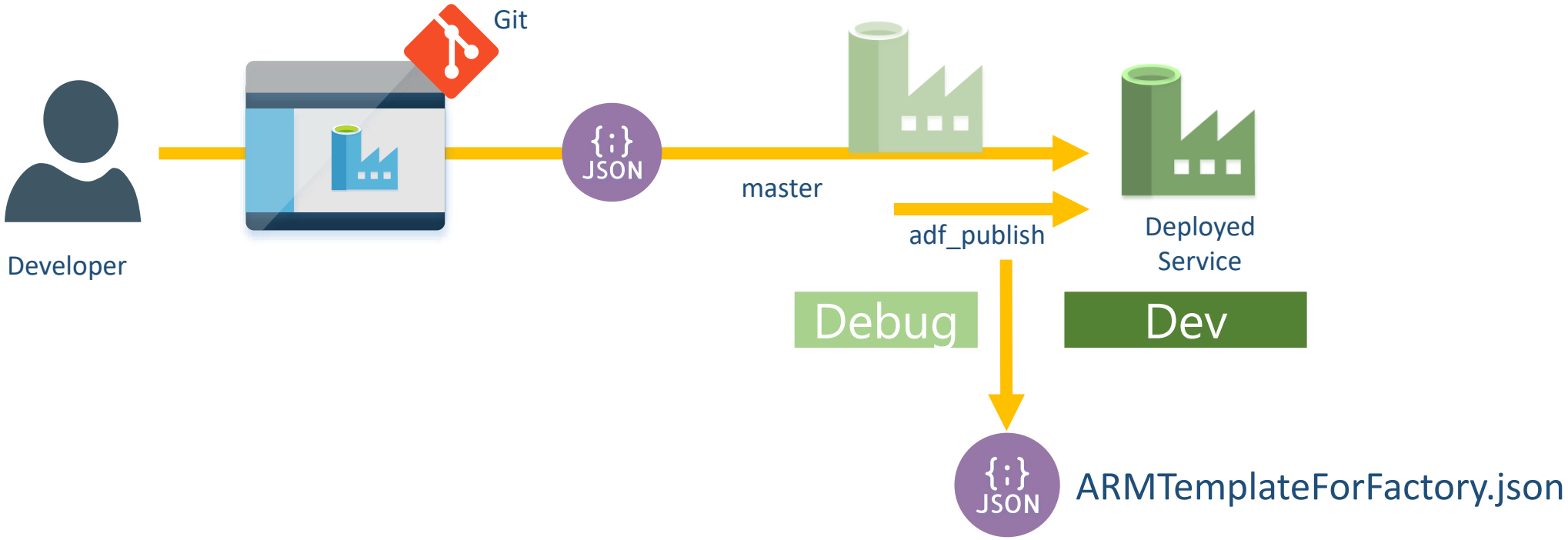
Option 2 – ARM Templates for Multiple Data Factory Services



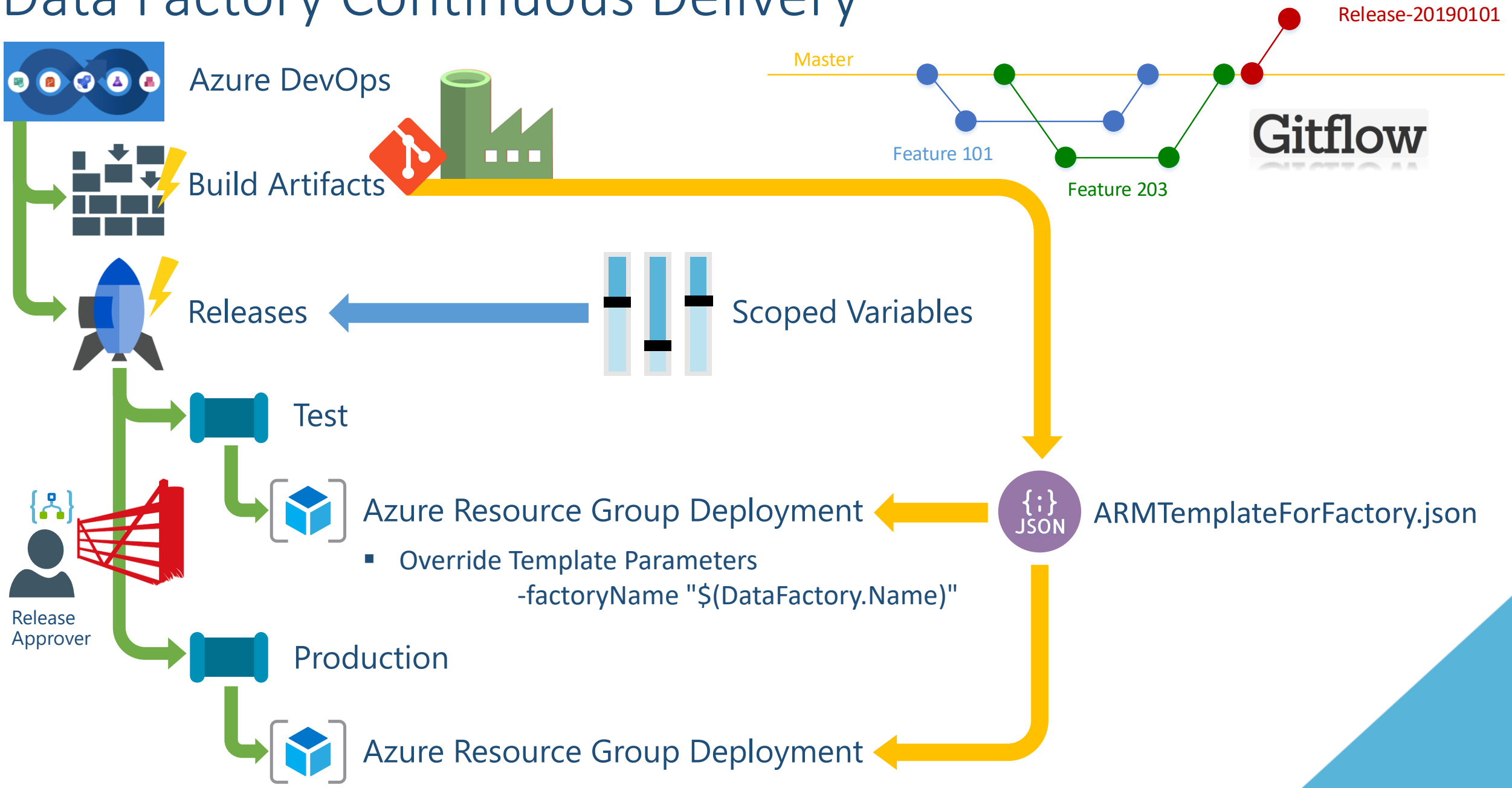
Data Factory Publish



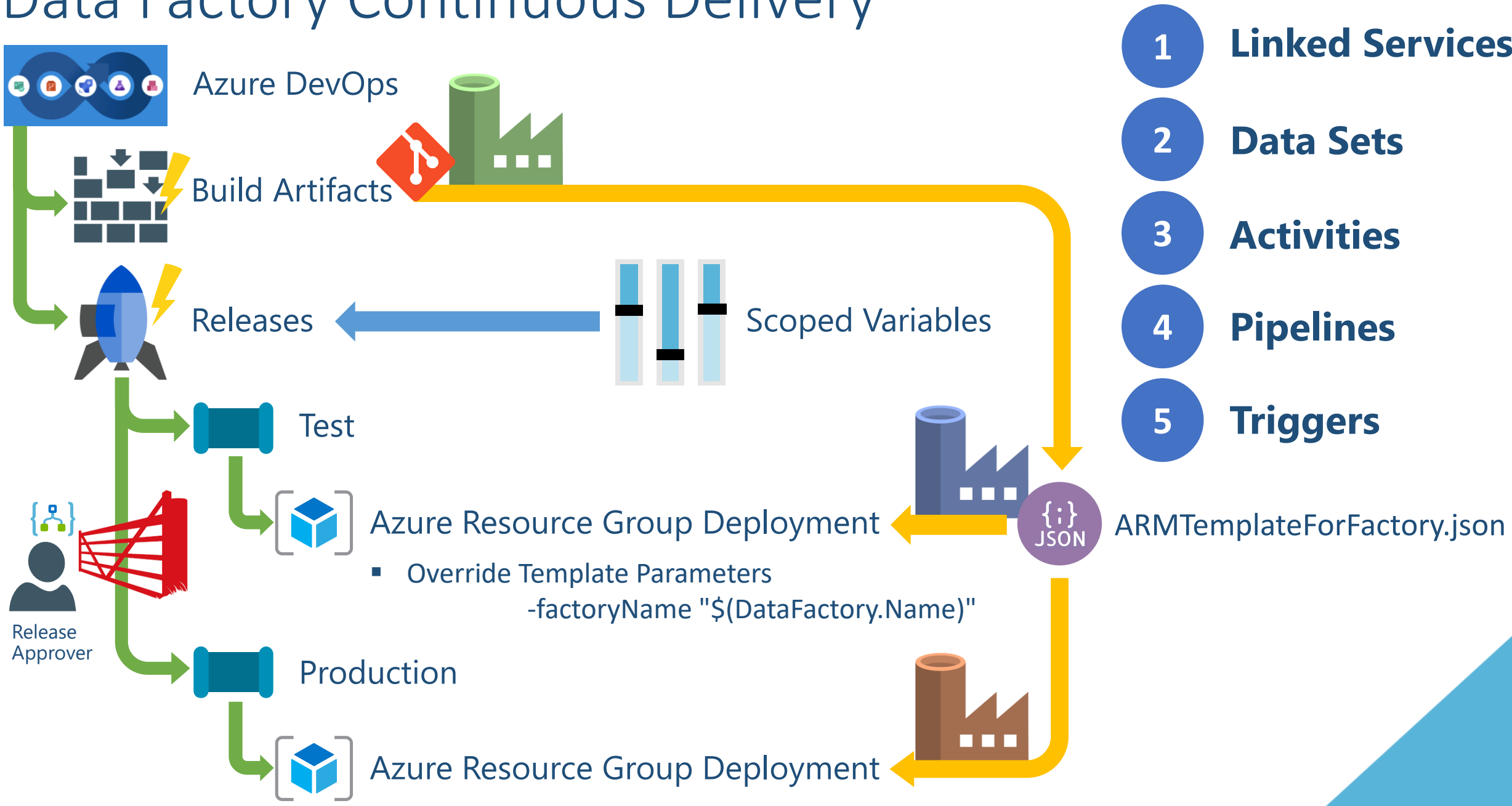
Data Factory Publish



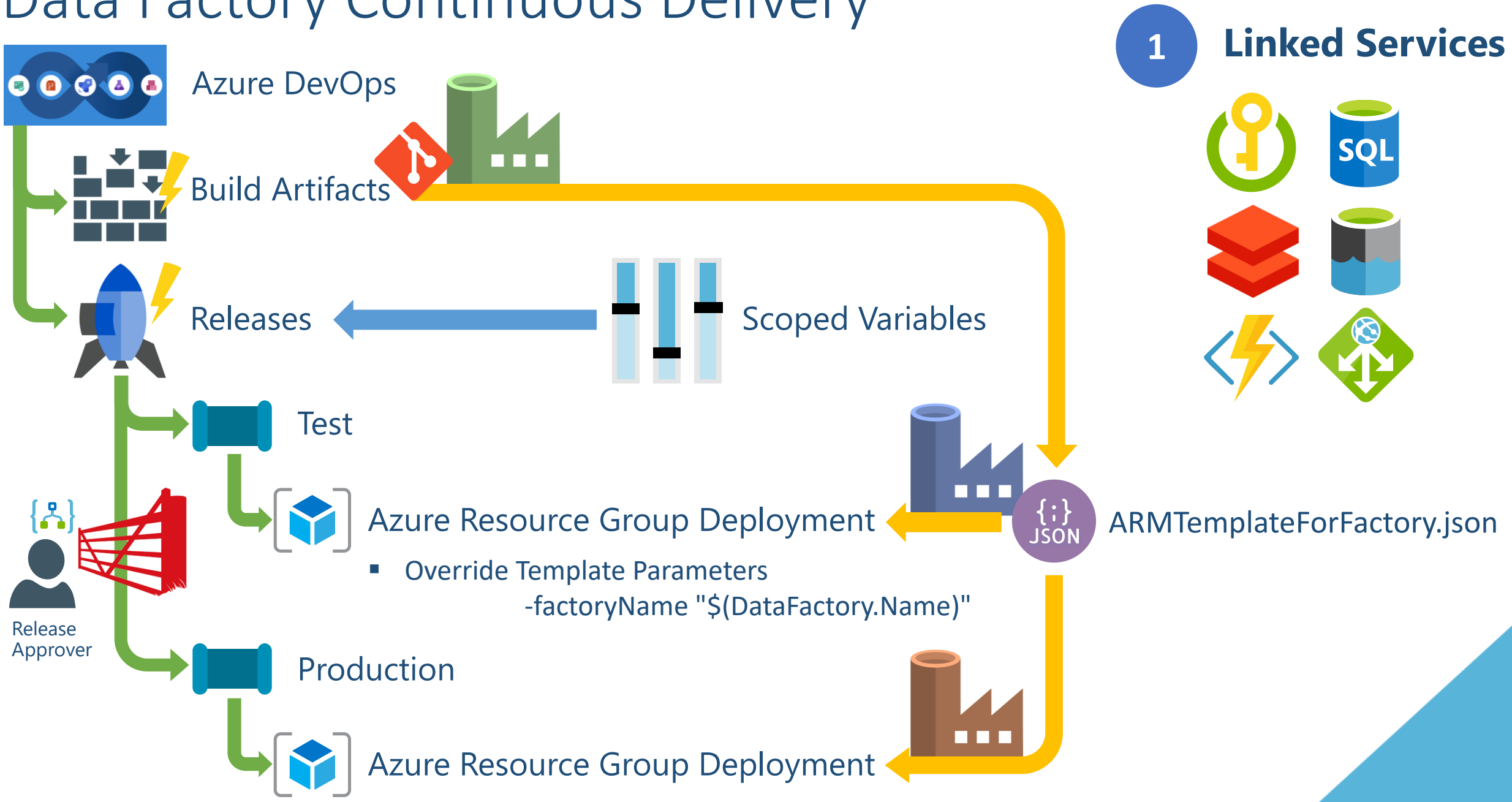
Data Factory Continuous Delivery



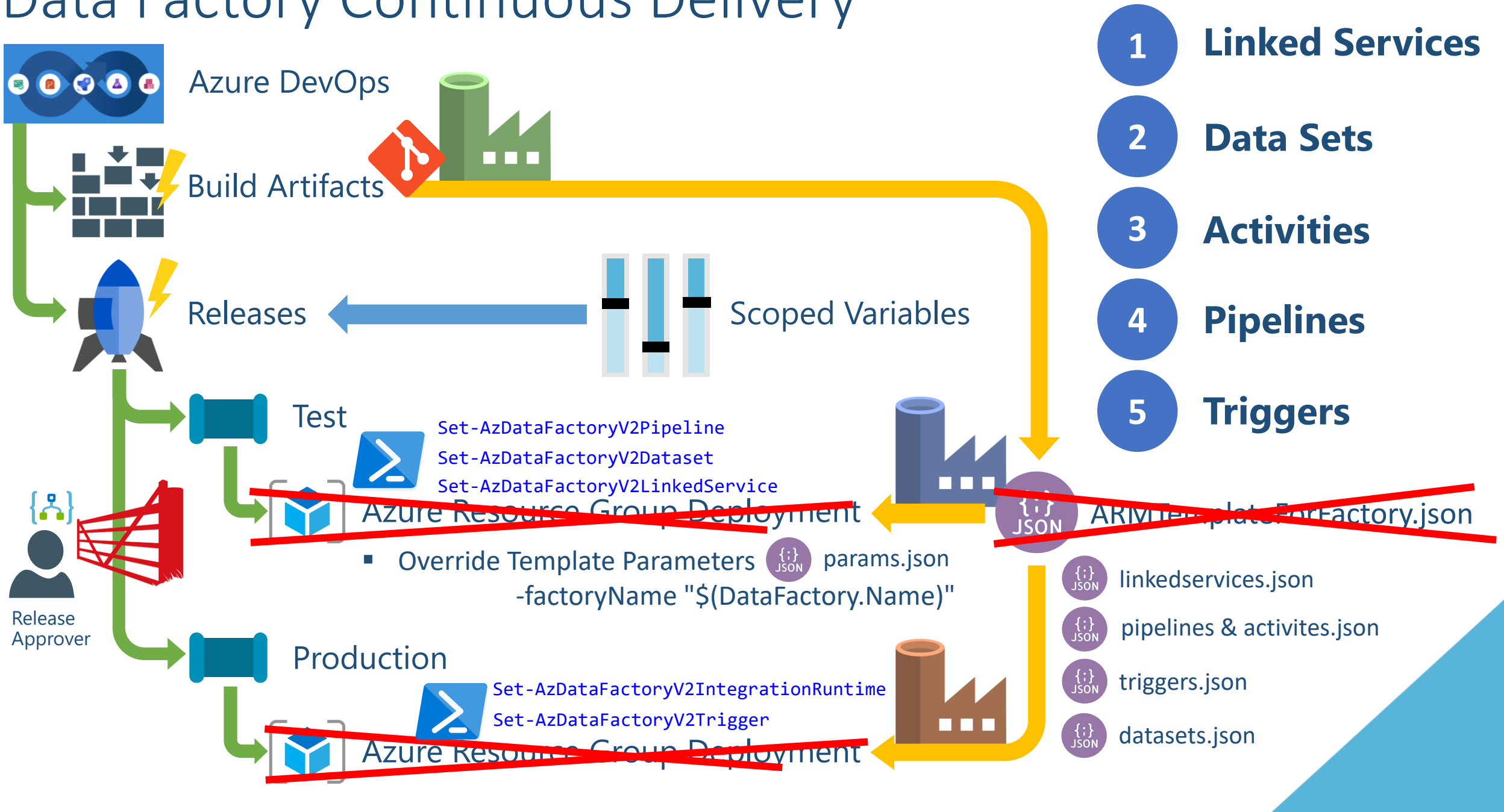
Data Factory Continuous Delivery



Data Factory Continuous Delivery

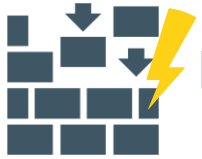


Data Factory Continuous Delivery



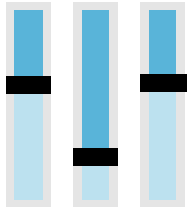
Data Factory Continuous Delivery - Bonus Option 3

Azure DevOps

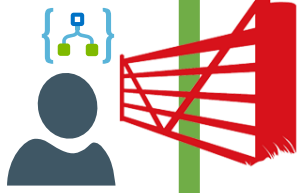


Build Artifacts

Releases



Scoped Variables



Release Approver

Test

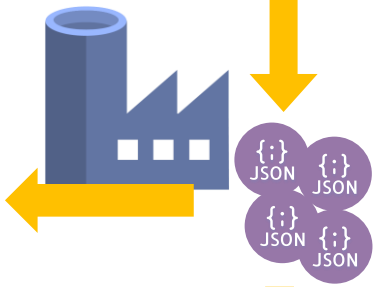
Run PowerShell params.json

```
Set-AzDataFactoryV2Pipeline
Set-AzDataFactoryV2Dataset
Set-AzDataFactoryV2LinkedService
```

Production

Run PowerShell params.json

```
Set-AzDataFactoryV2IntegrationRuntime
Set-AzDataFactoryV2Trigger
```



linkedservices.json
pipelines & activites.json
datasets.json
triggers.json

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

Caveats

1) Handle own dependencies.

2) Handle own removals.

Data Factory DevOps Summary

What is your code branching strategy?

Which source control tool to use?

How many environments do we want?

What deployment method do we want to use?

What built artefacts are we going to use?...

OR

How much control do you want?

ARMTemplate
ForFactory.json

linkedservices.json
pipelines & activites.json
datasets.json
triggers.json

Session Agenda (Short Stories)

- 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

altius

Blog: mrpaulandrew.com
Email: paul@mrpaulandrew.com

Twitter: @mrpaulandrew
LinkedIn: In/mrpaulandrew

GitHub: github.com/mrpaulandrew