

Complex Azure Orchestration

Data Factory in Production



Paul Andrew | Principal Consultant & Solution Architect



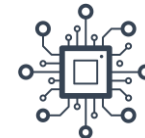
altius



@MrPaulAndrew



In/MrPaulAndrew



MrPaulAndrew.com

DATA:Scotland



Microsoft



Arnold Clark



SentryOne®





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

Complex Azure Orchestration
Data Factory in Production



<https://github.com/mrpaulandrew>

CommunityEvents

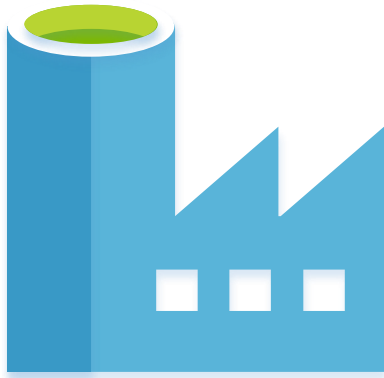
Demo code, content and slides from various community events.

● C++

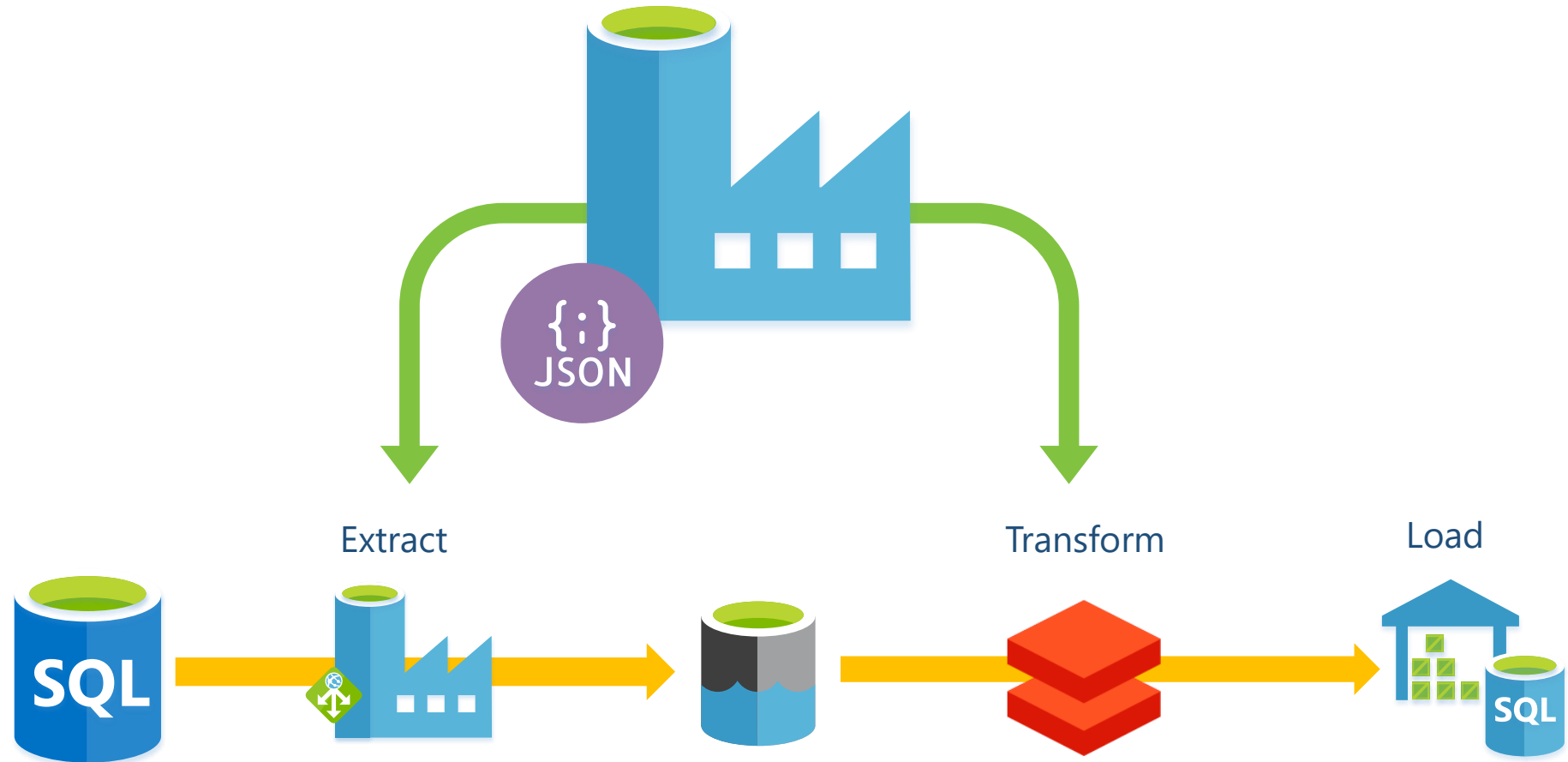
[{Event/Location}-{Month}-{Year}](#)

Azure Data Factory

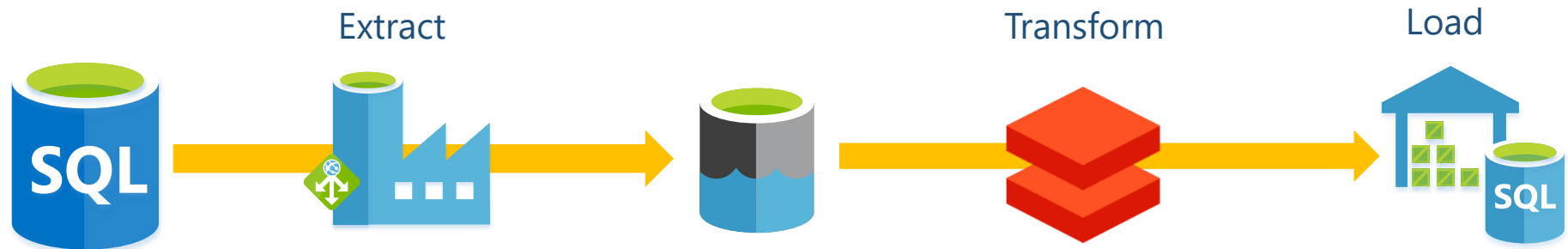
A Quick Overview



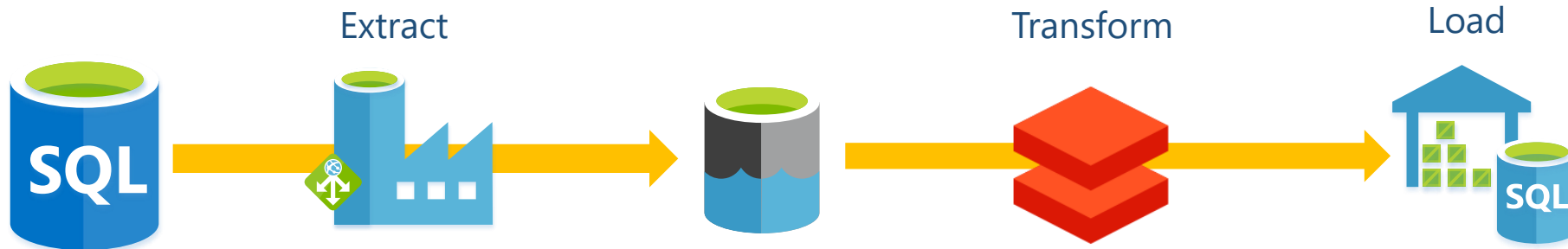
What is Azure Data Factory?



What is Azure Data Factory?



Data Factory Components



1 **Linked Services** ✓

2 **Data Sets** ✓

3 **Activities** ✓

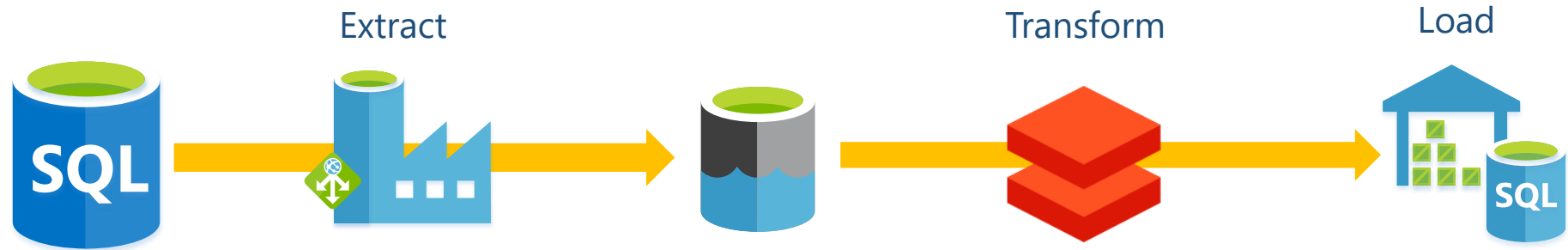
4 **Pipelines** ✓

5 **Triggers** ✗

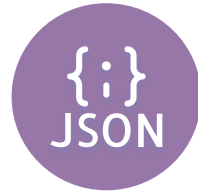
{:}
JSON

```
{
  "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

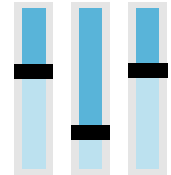


- 1 **Linked Services** ✓
- 2 **Data Sets** ✓
- 3 **Activities** ✓
- 4 **Pipelines** ✓
- 5 **Triggers** ✗



Expression Builder

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



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

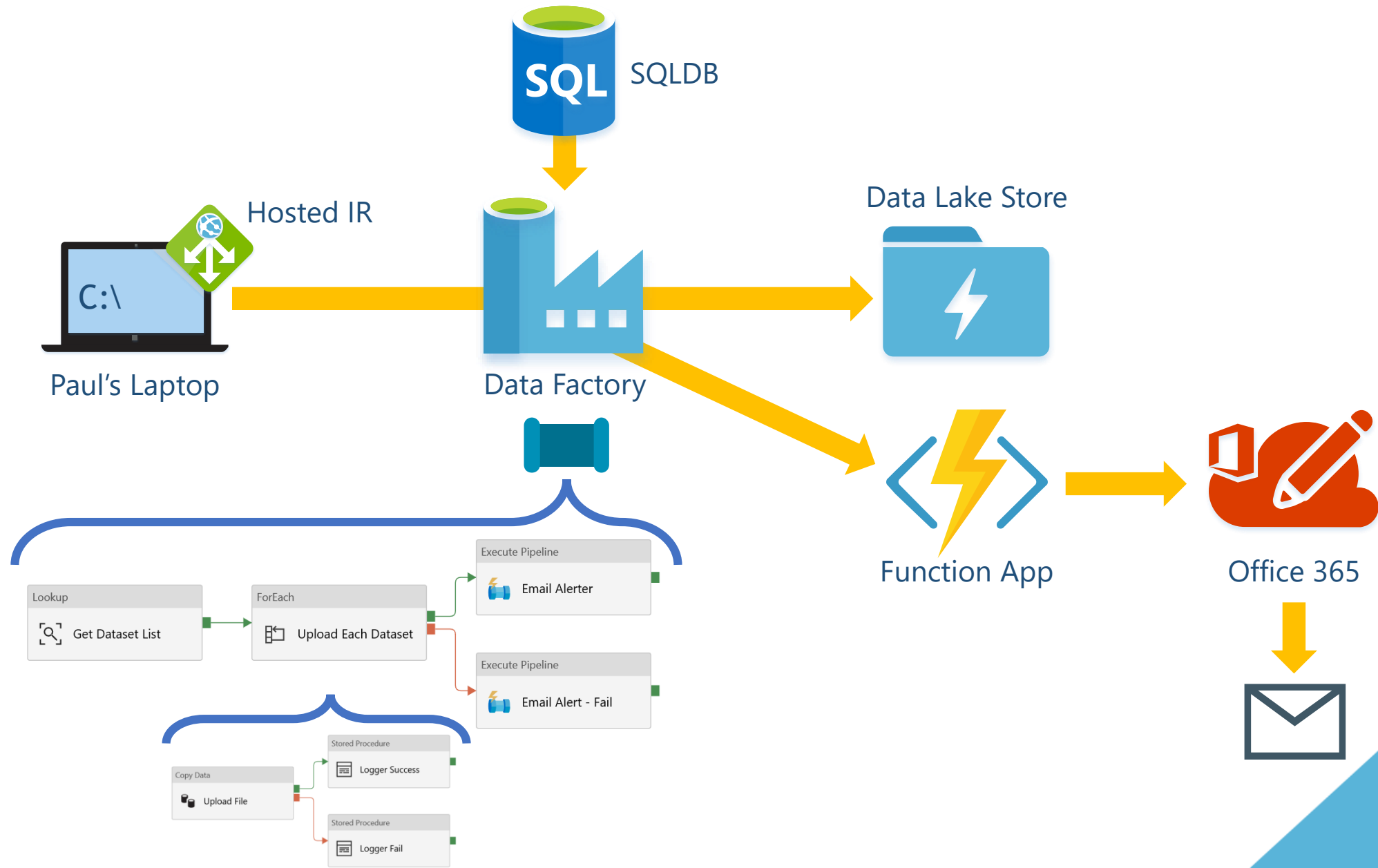


Add dynamic content [Alt+P]

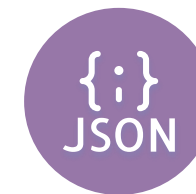
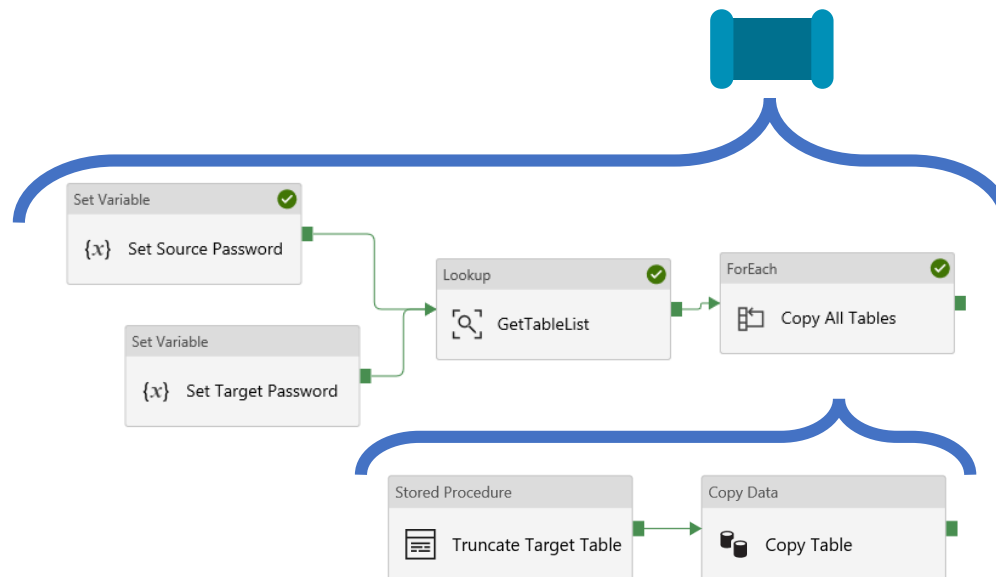
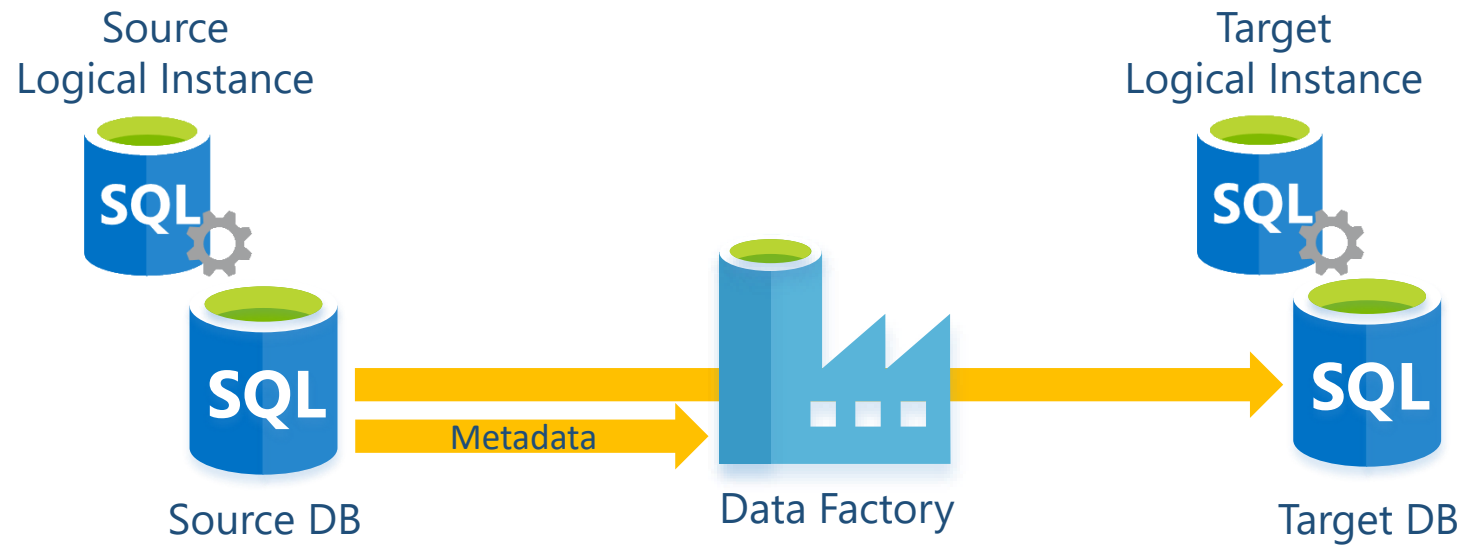
Dynamic Data Factory Pipelines



Demo Architecture 1



Demo Architecture 2



1x Linked Service
1x Dataset

Web Activities vs Web Hook Activities



Web Hook vs Web Activity



PUT
POST
GET
DELETE

POST

1 Minute Timeout

Configurable Timeout

Retry Capabilities

No Retry




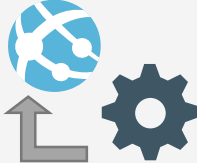

Linked Services
Datasets

No Artifact Support

One Way Call

Call Back URL

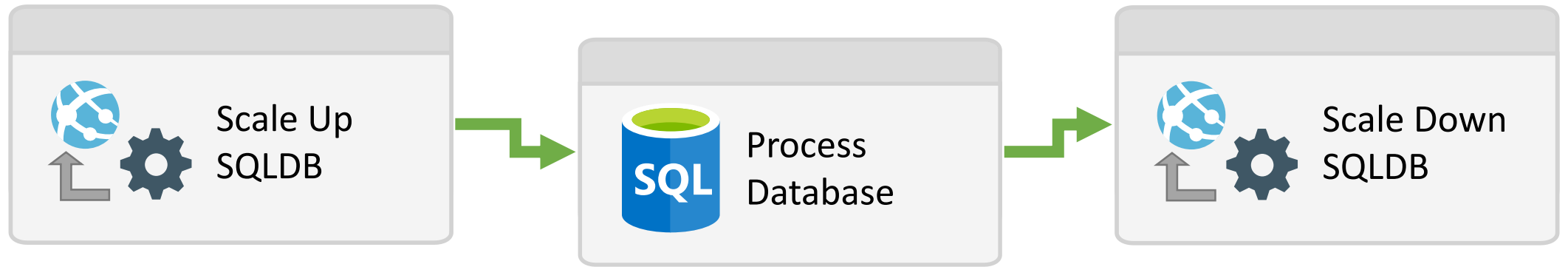
Web Hook vs Web Activity

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

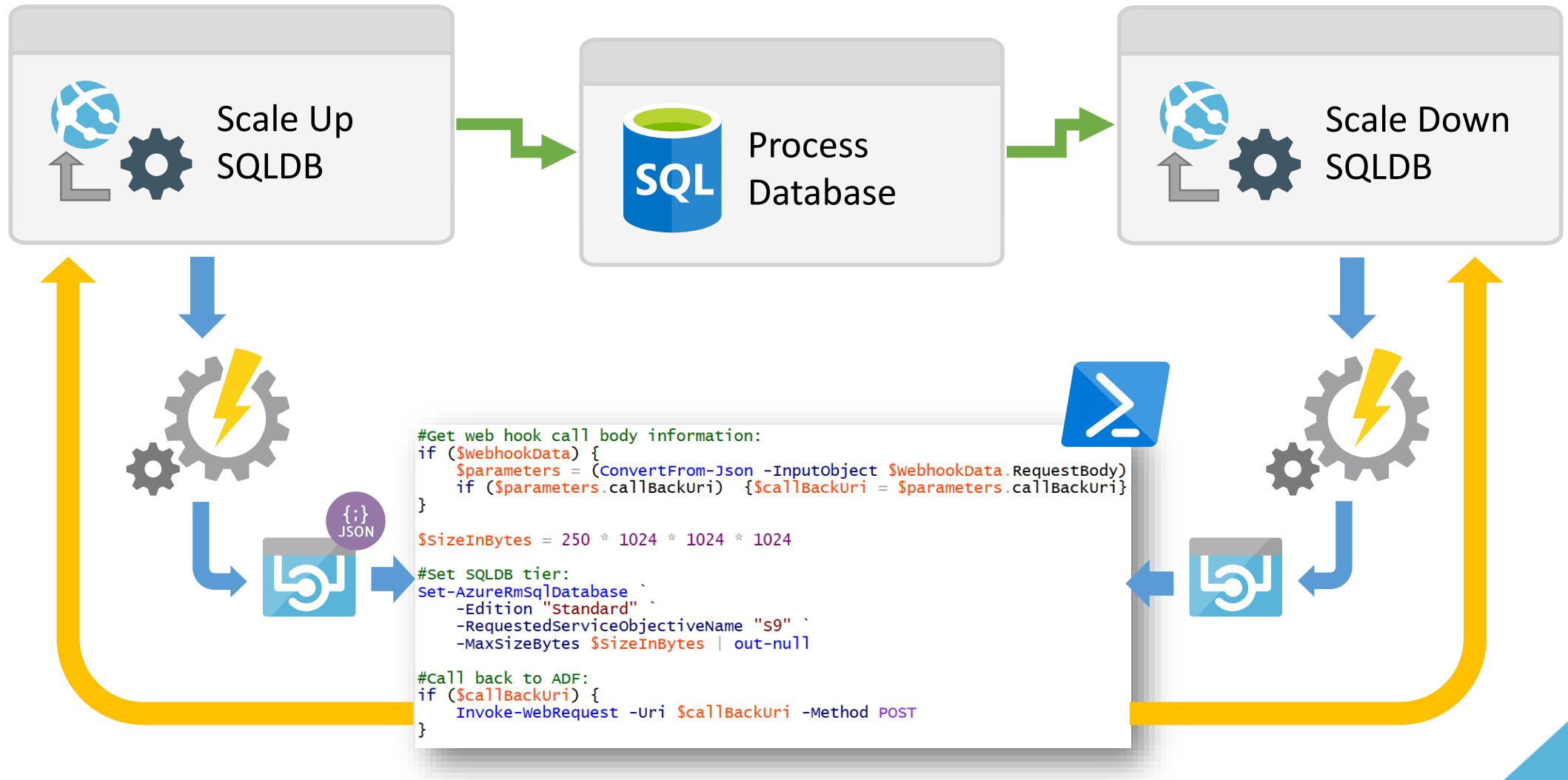
Web Hook vs Web Activity



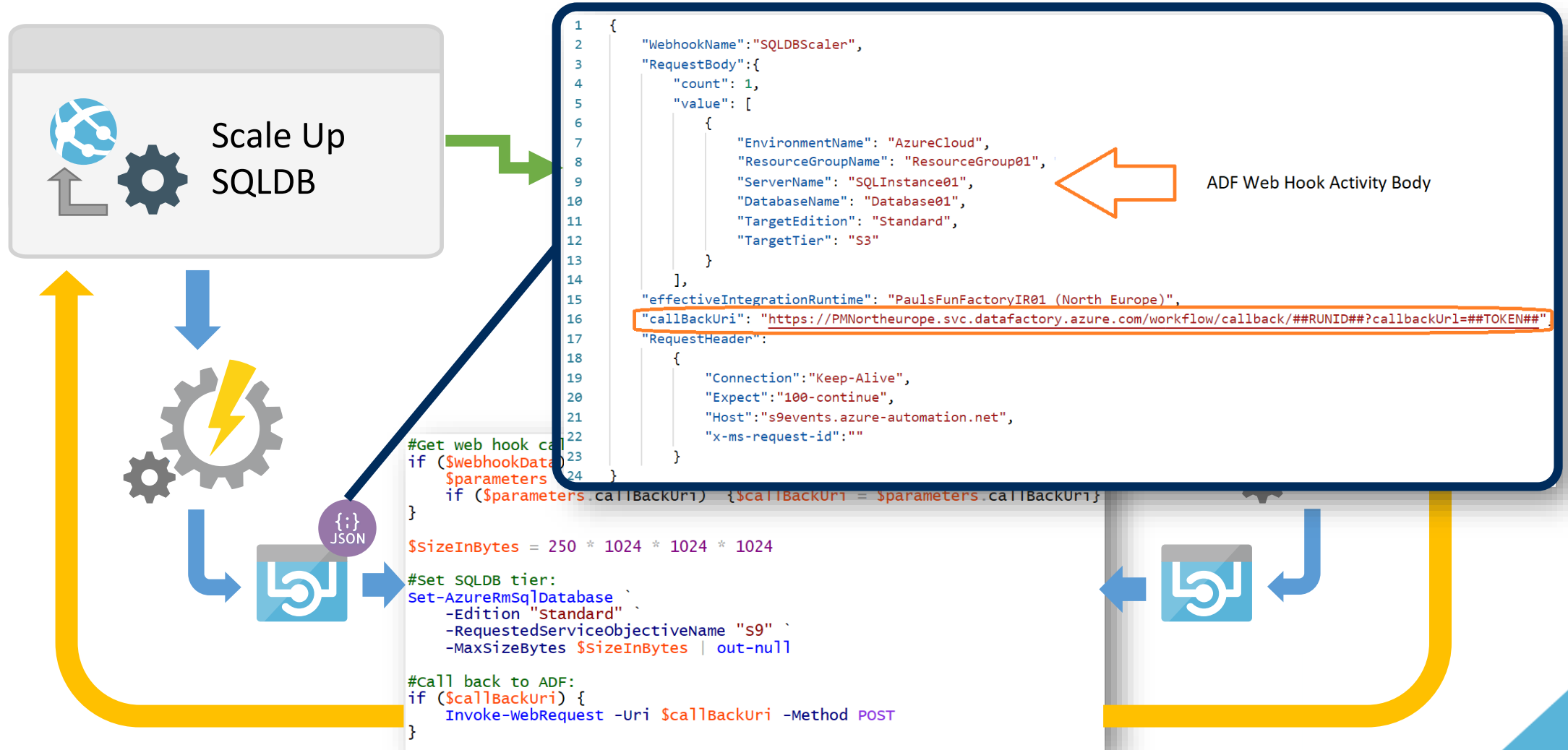
Web Hook vs Web Activity



Web Hook vs Web Activity



Web Hook vs Web Activity

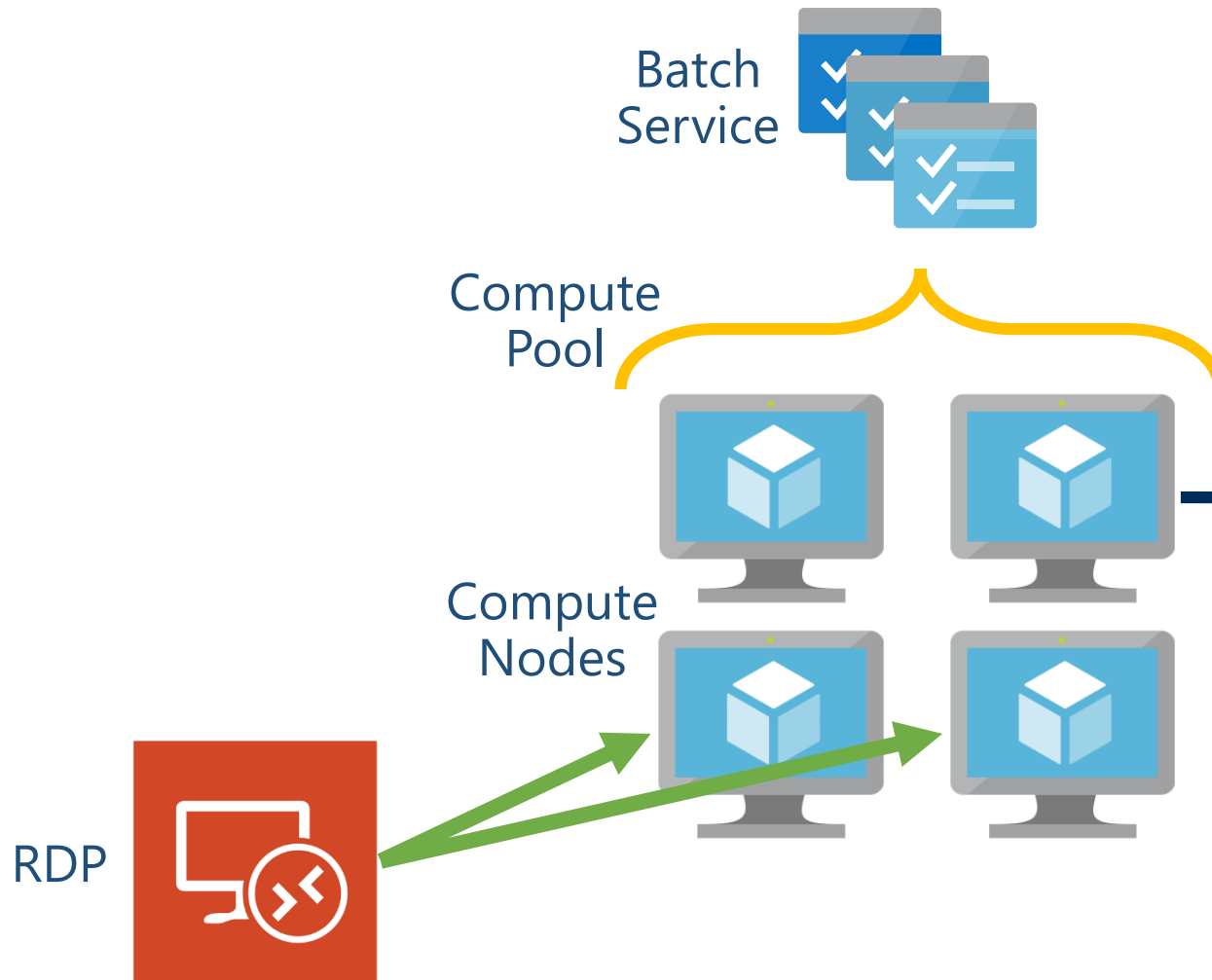


Custom Activities



Azure Batch Service

Scale out compute delivered using PaaS technology with IaaS underneath.



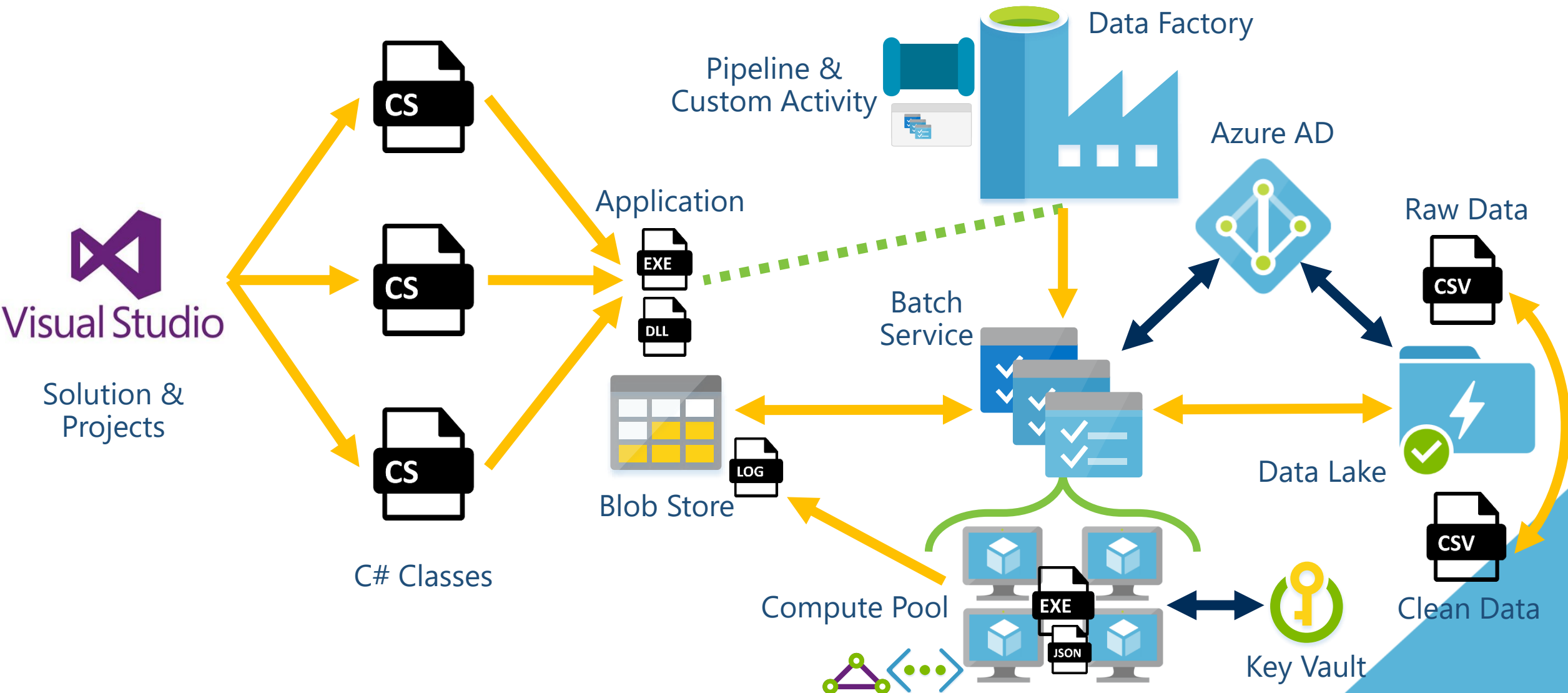
VM node size set per compute pool:

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

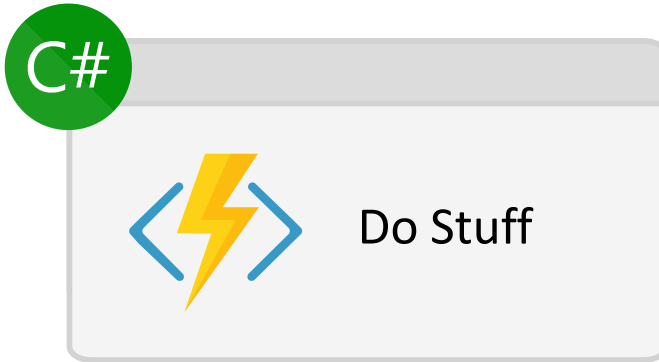
- ▶ 1 compute node = 1 virtual machine.
- ▶ 1 job per compute node.
- ▶ Max of 4 tasks per node.
- ▶ OS on D drive, not C.
- ▶ Special environment variables.

Building a Custom Activity

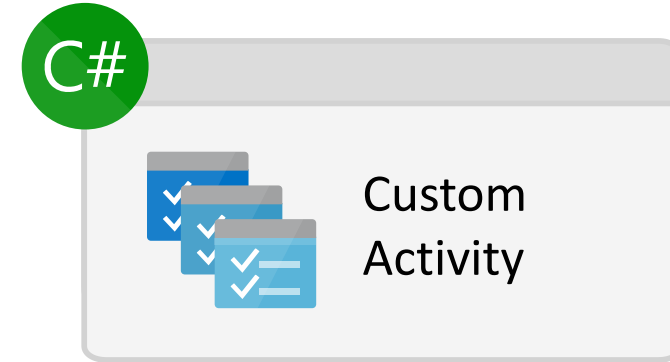
A .Net Console App Executed Using Azure Batch Service.



Extensibility Conclusions



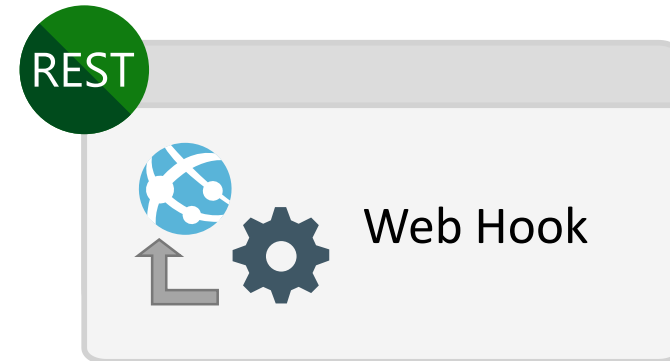
10 minutes execution



Auto scale out compute &
Scale up per compute node

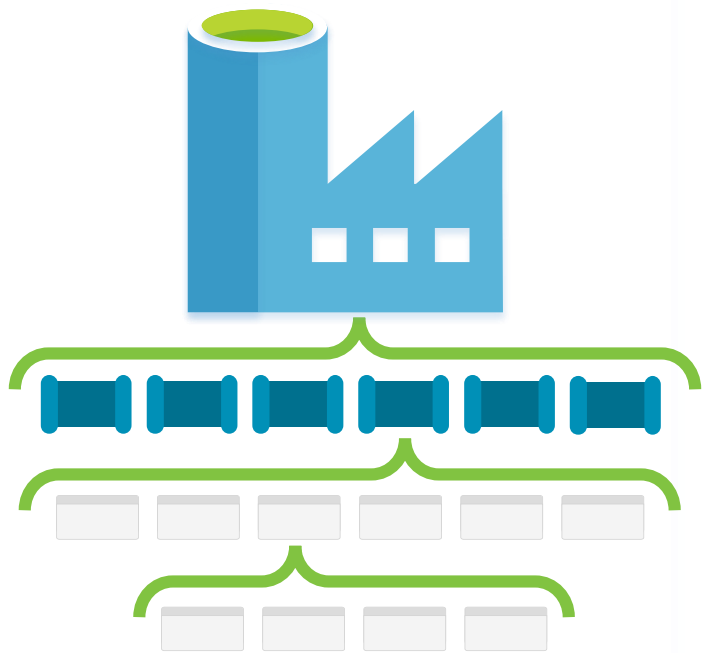


Asynchronous

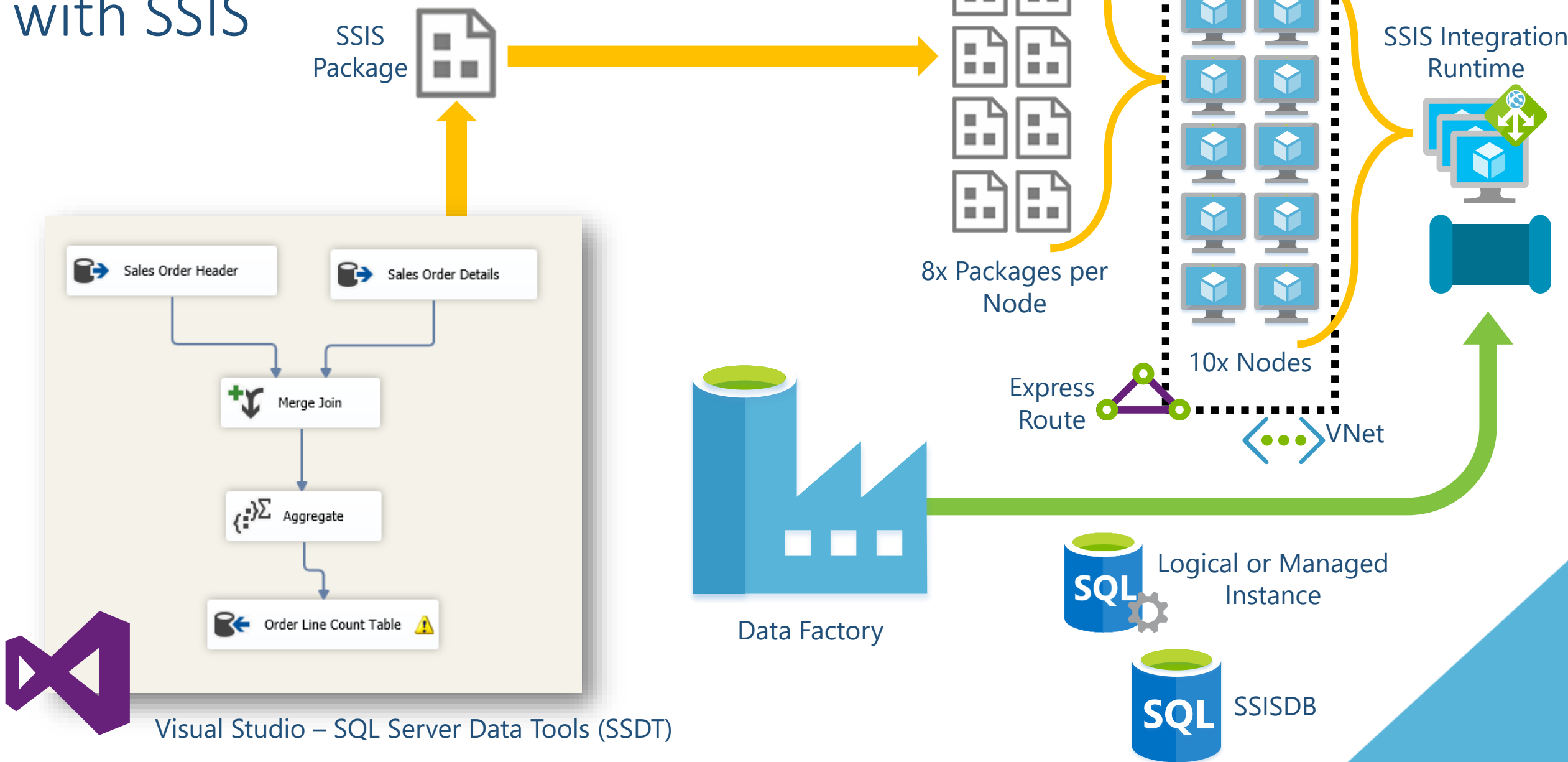


Synchronous
(Call back)

Scale Out ~~Execution~~ Everything!



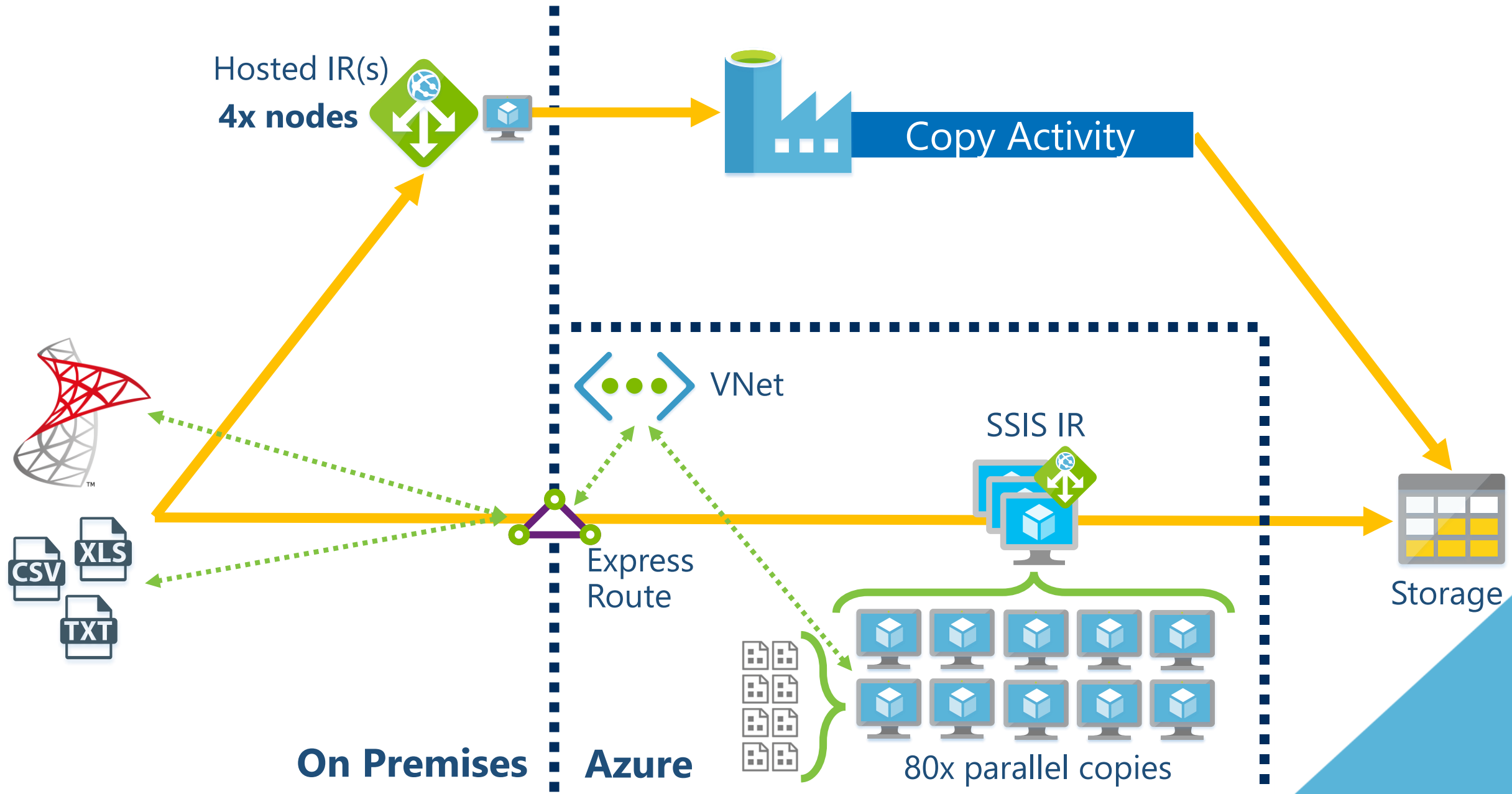
Data Transformation in zure with SSIS



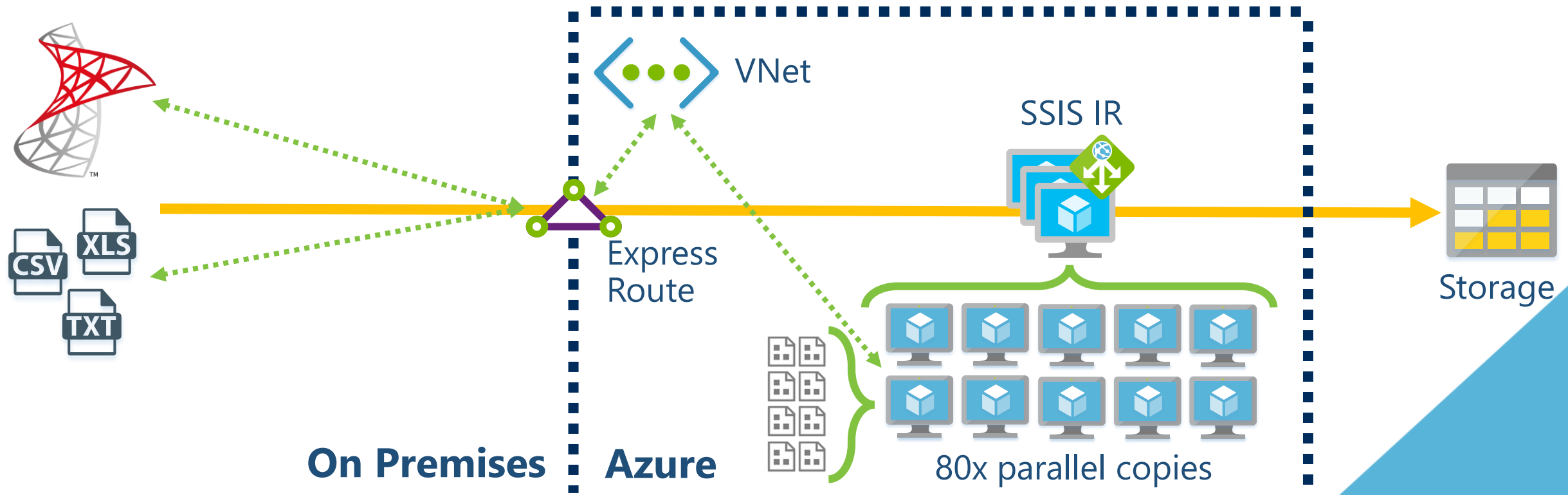
Visual Studio – SQL Server Data Tools (SSDT)



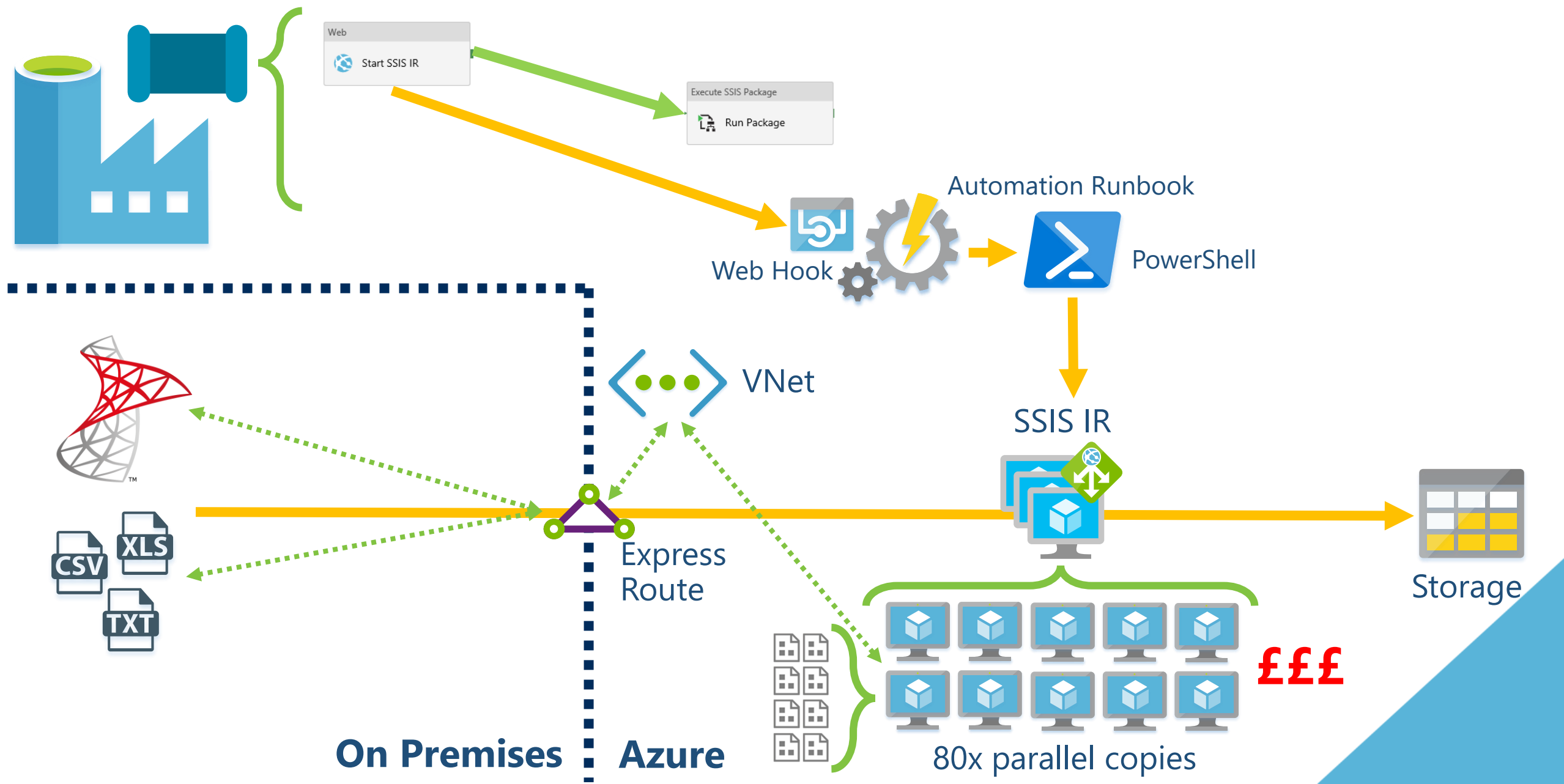
The SSIS IR vs Hosted IR with Express Route



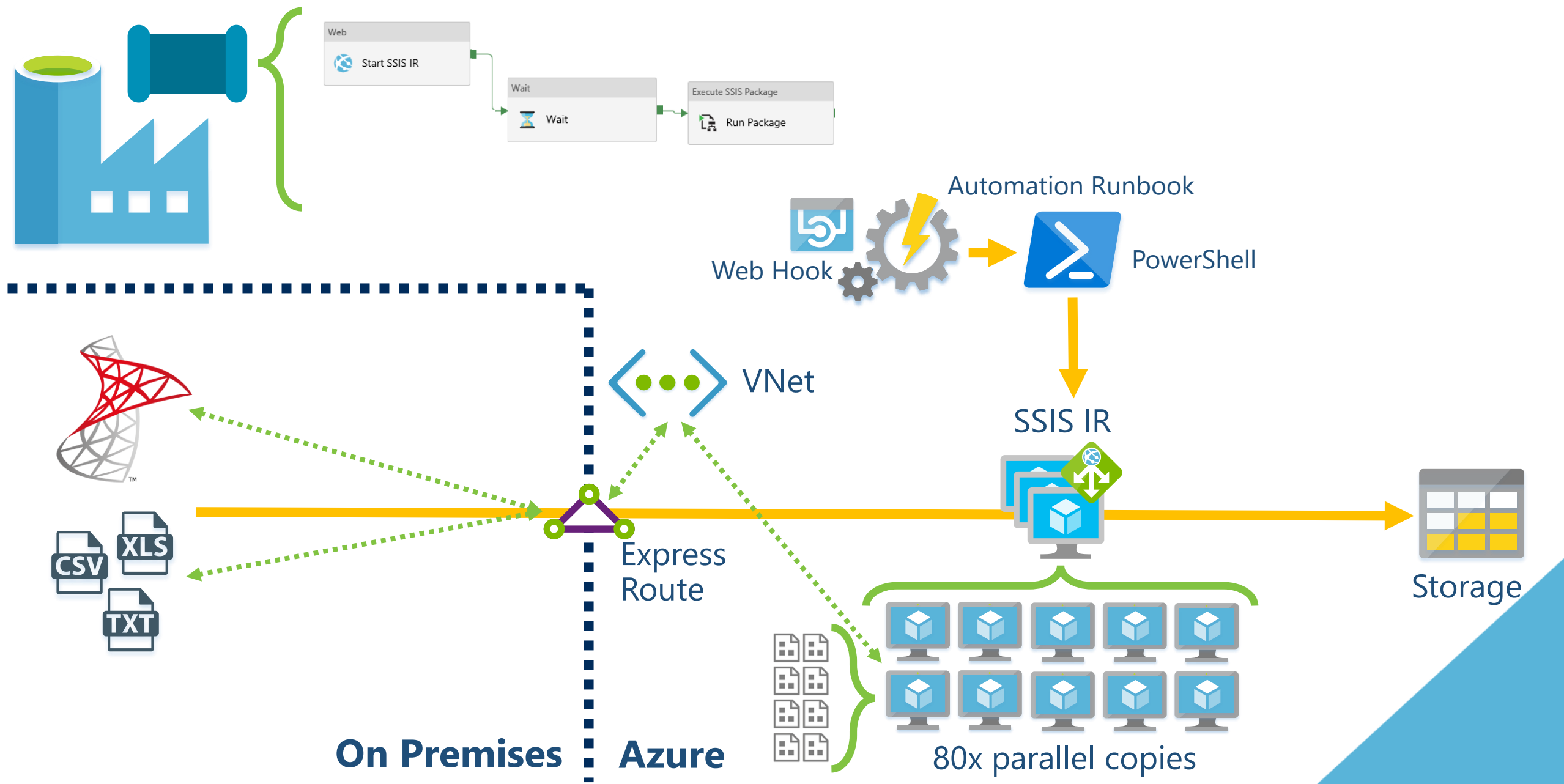
The SSIS IR Start/Stop



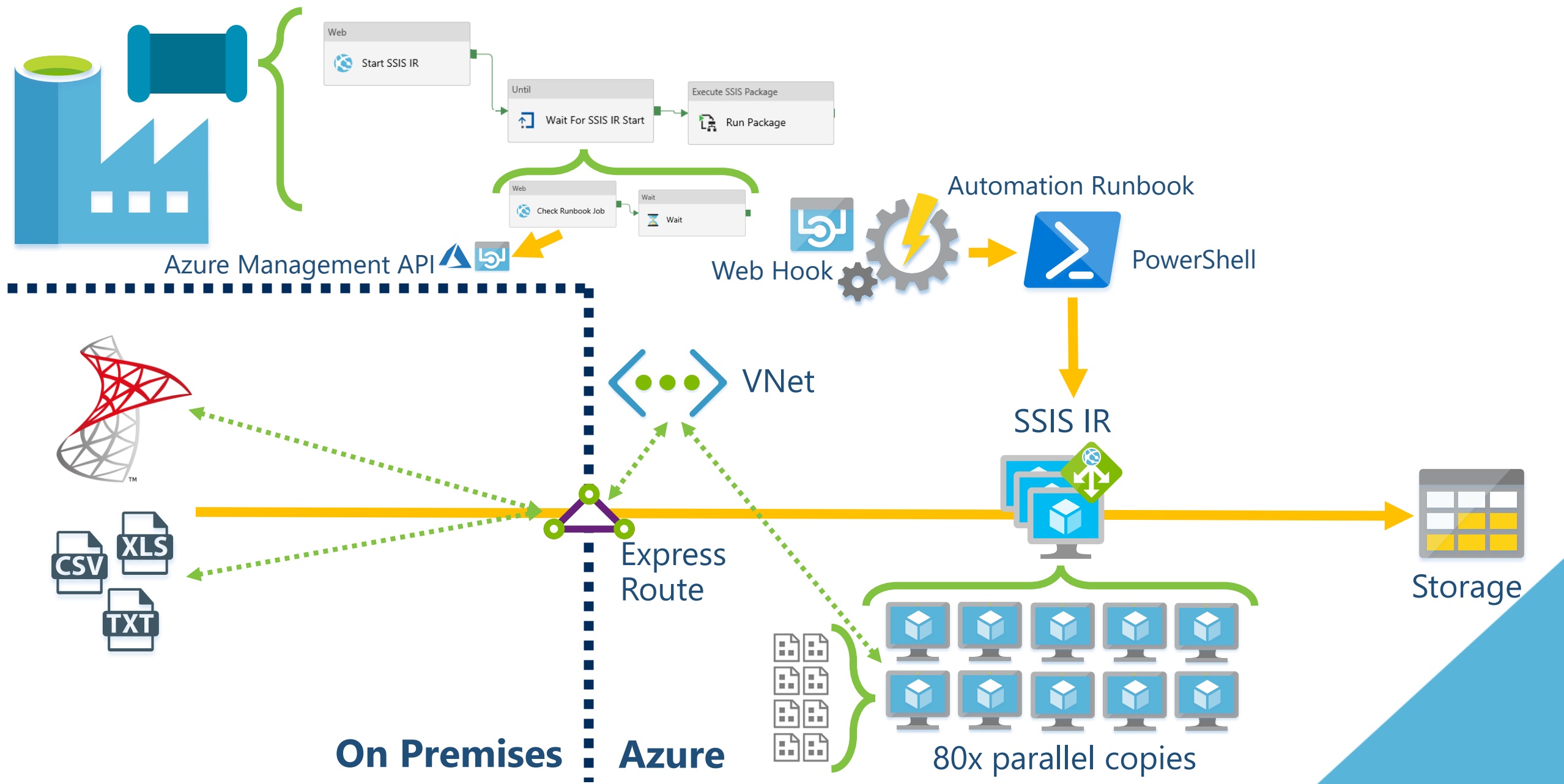
The SSIS IR Start/Stop



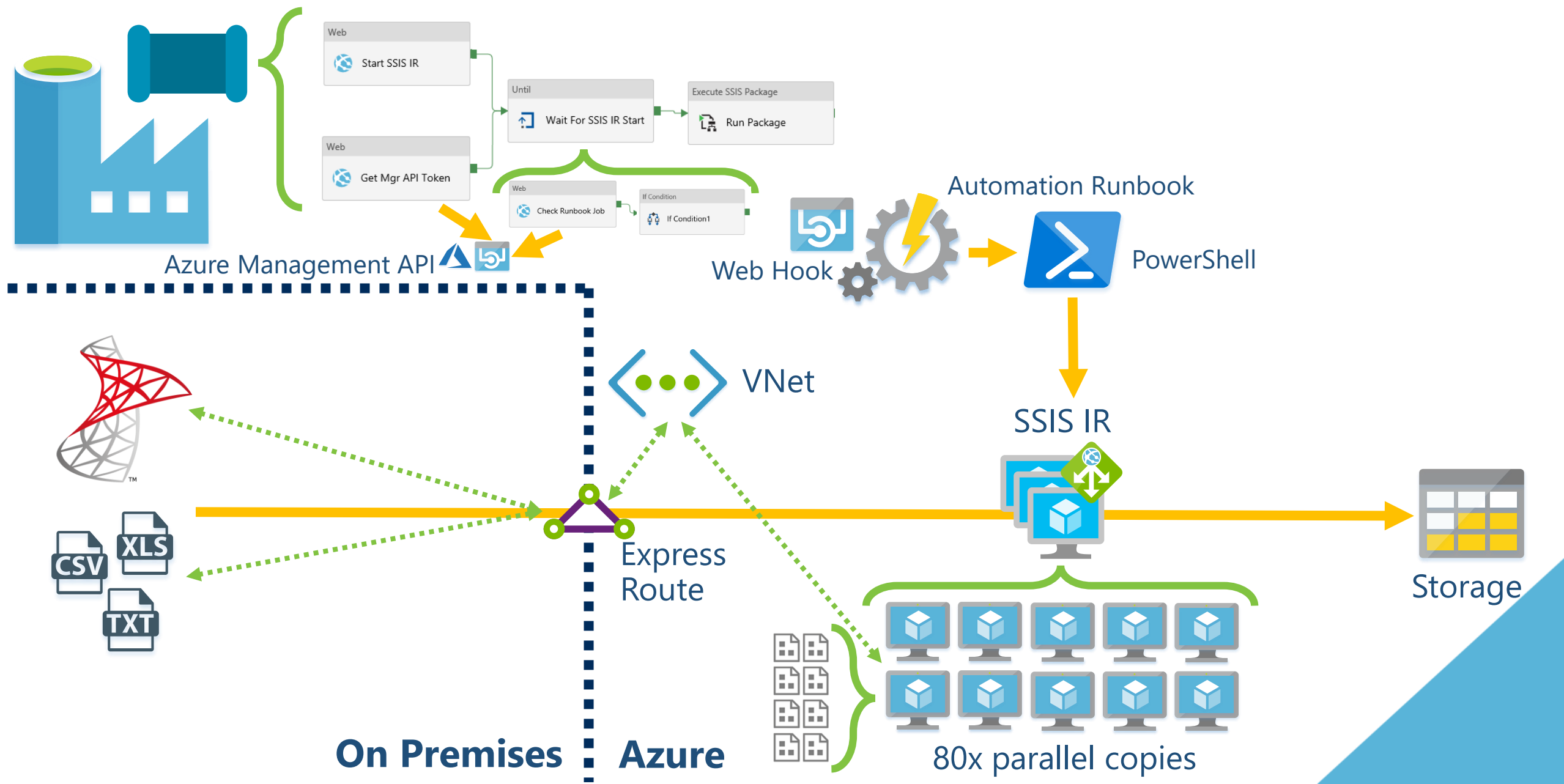
The SSIS IR Start/Stop



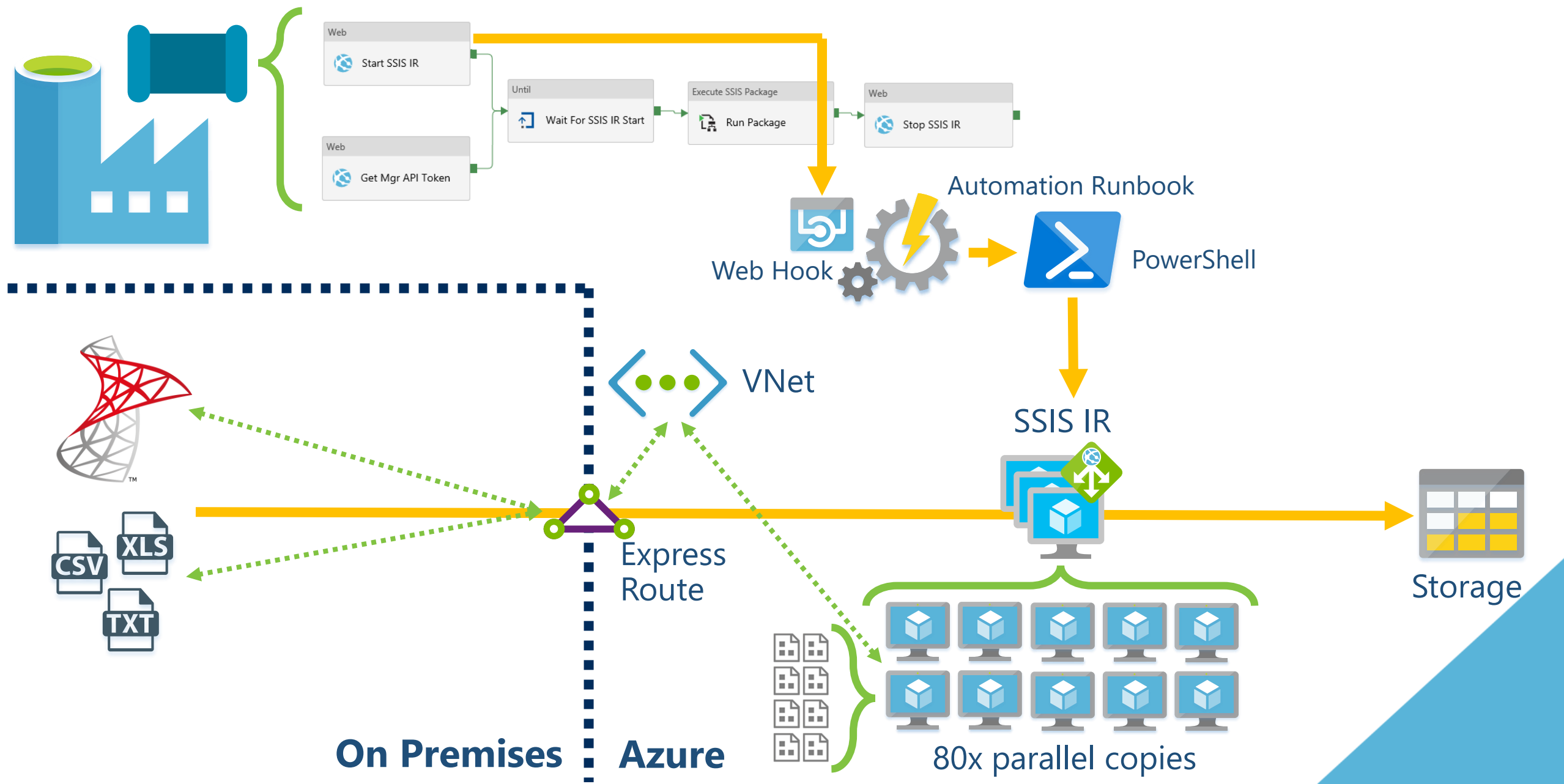
The SSIS IR Start/Stop



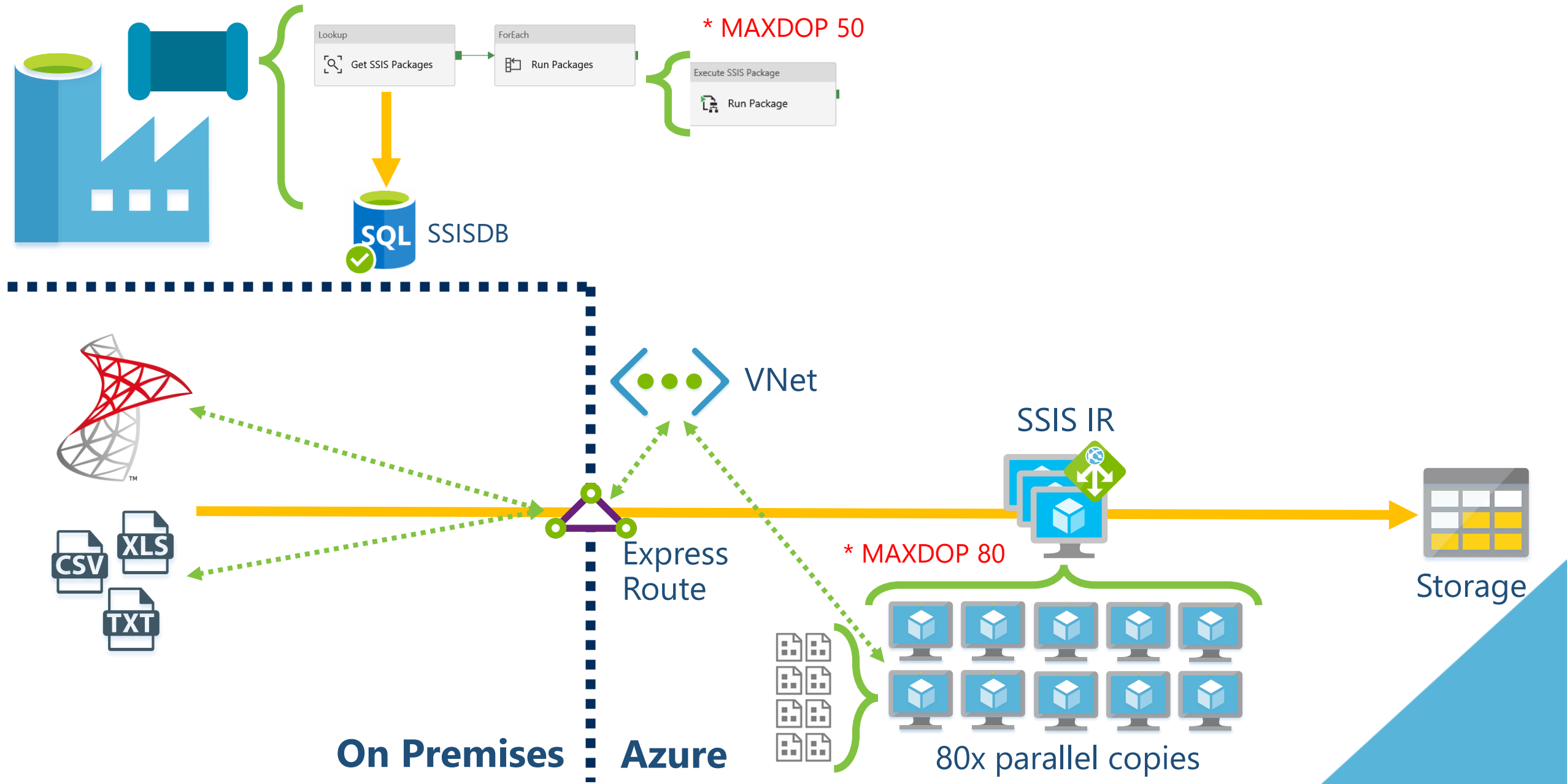
The SSIS IR Start/Stop



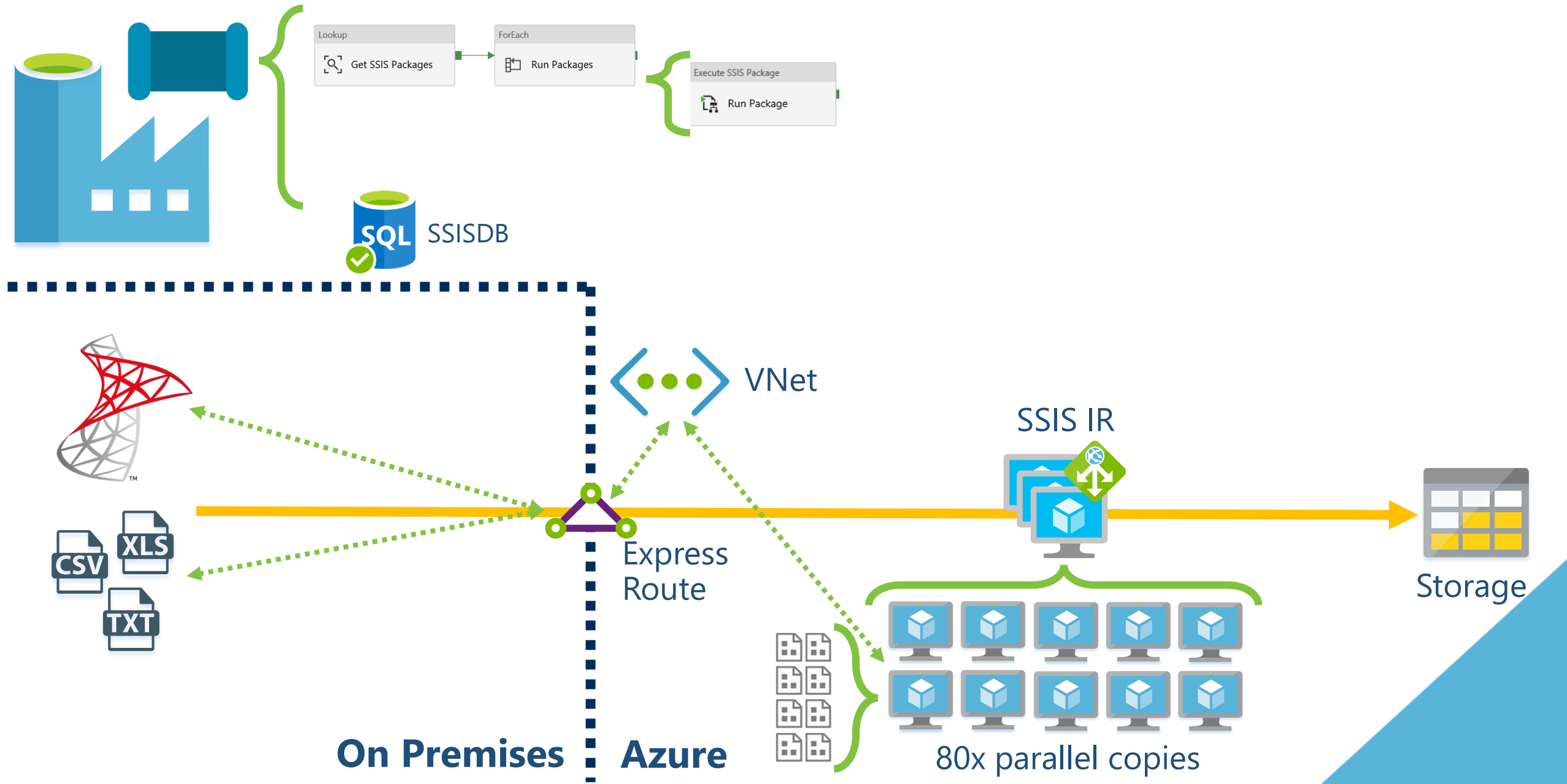
The SSIS IR Start/Stop



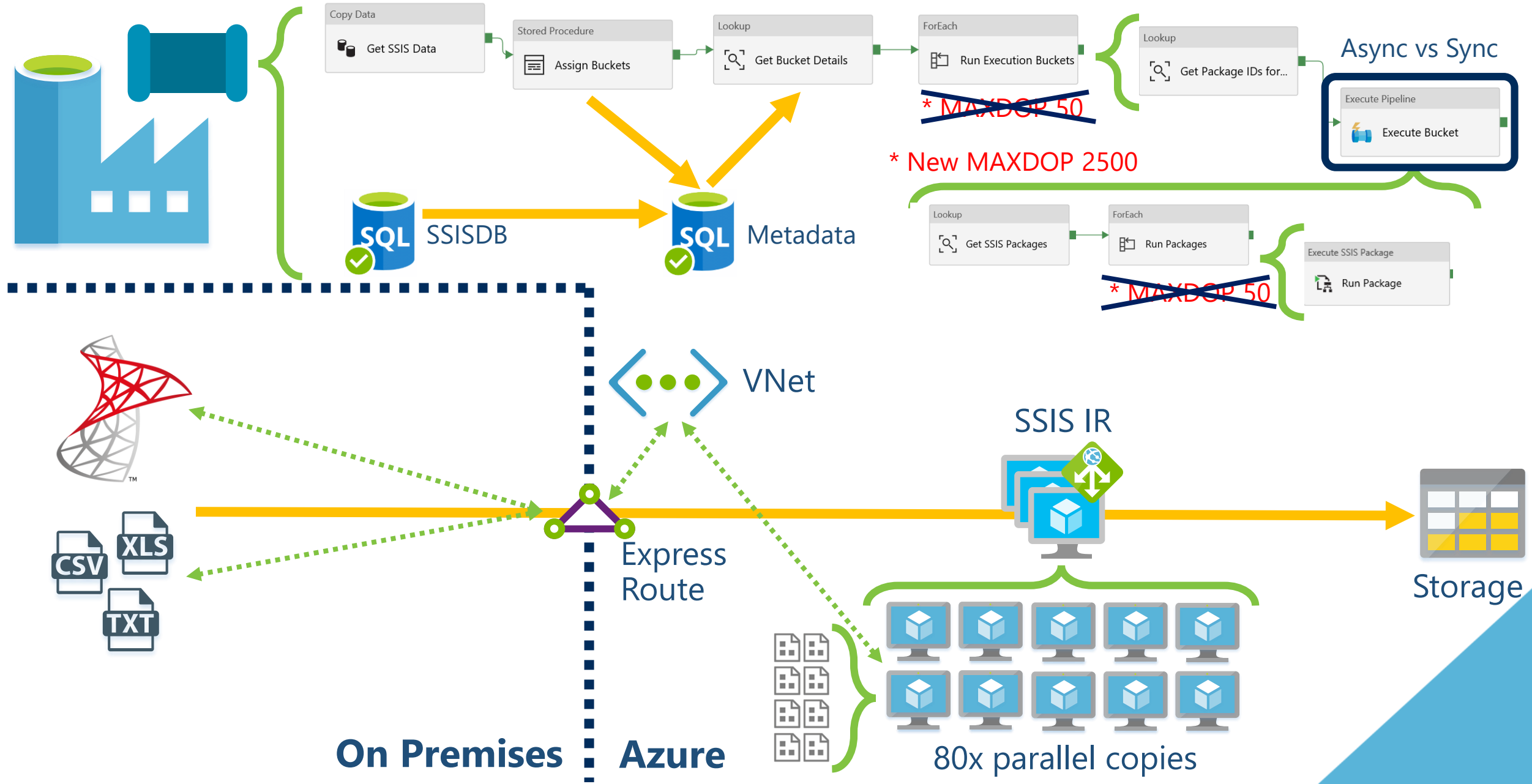
The SSIS IR Parallelism



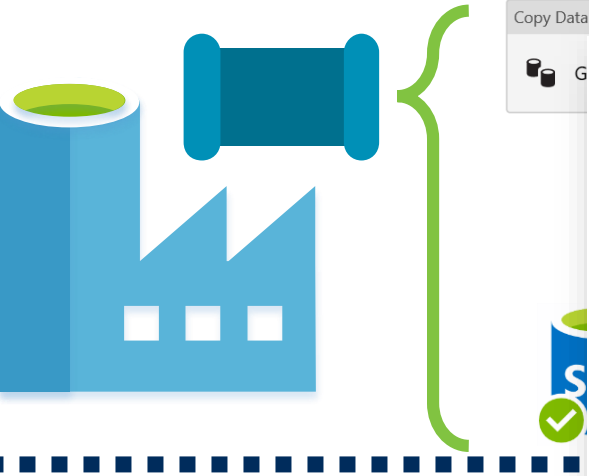
The SSIS IR Parallelism



The SSIS IR Parallelism

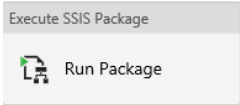
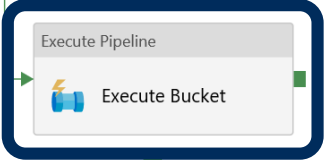


The SSIS IR Parallelism



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

Async vs Sync



OP 50

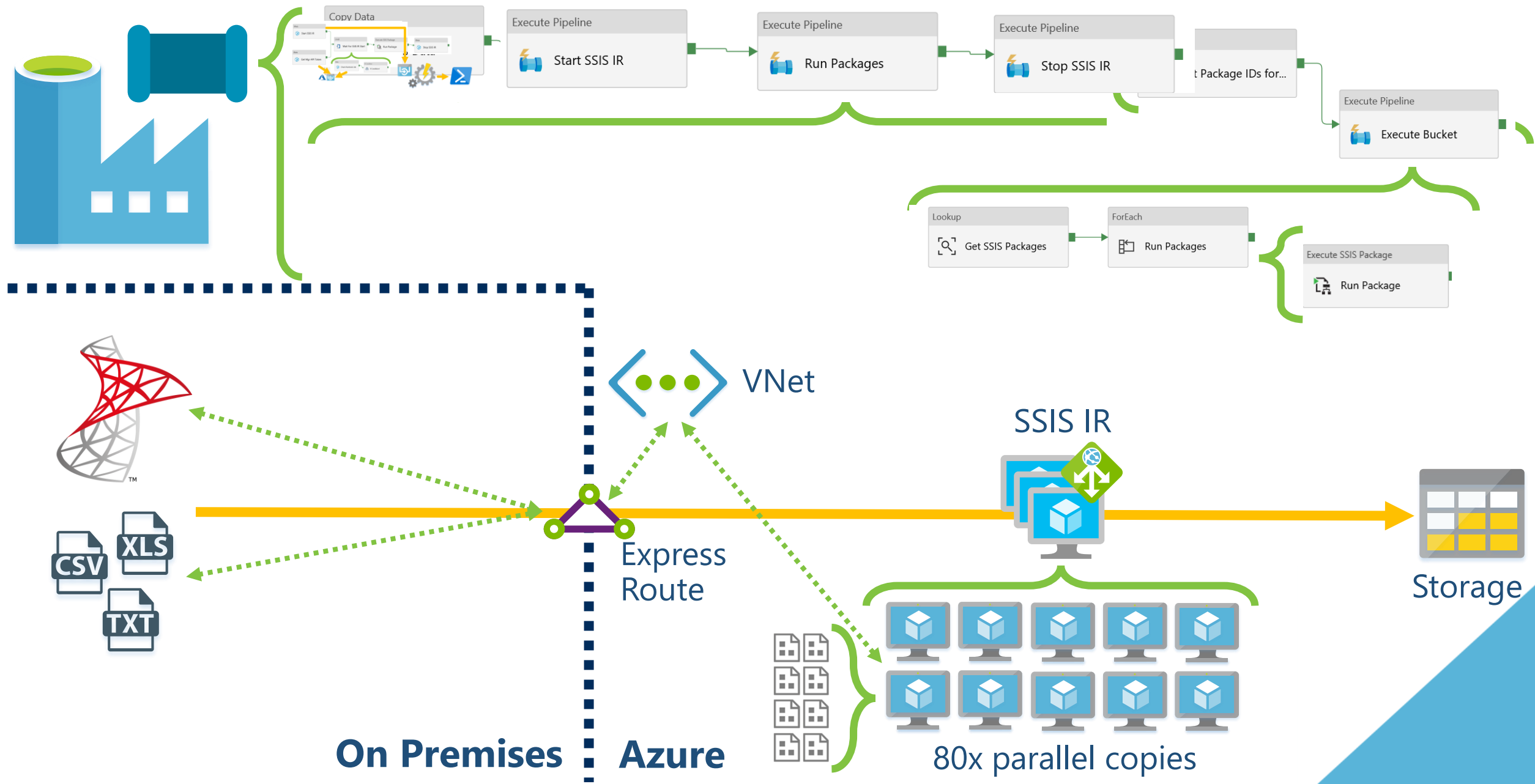


Storage

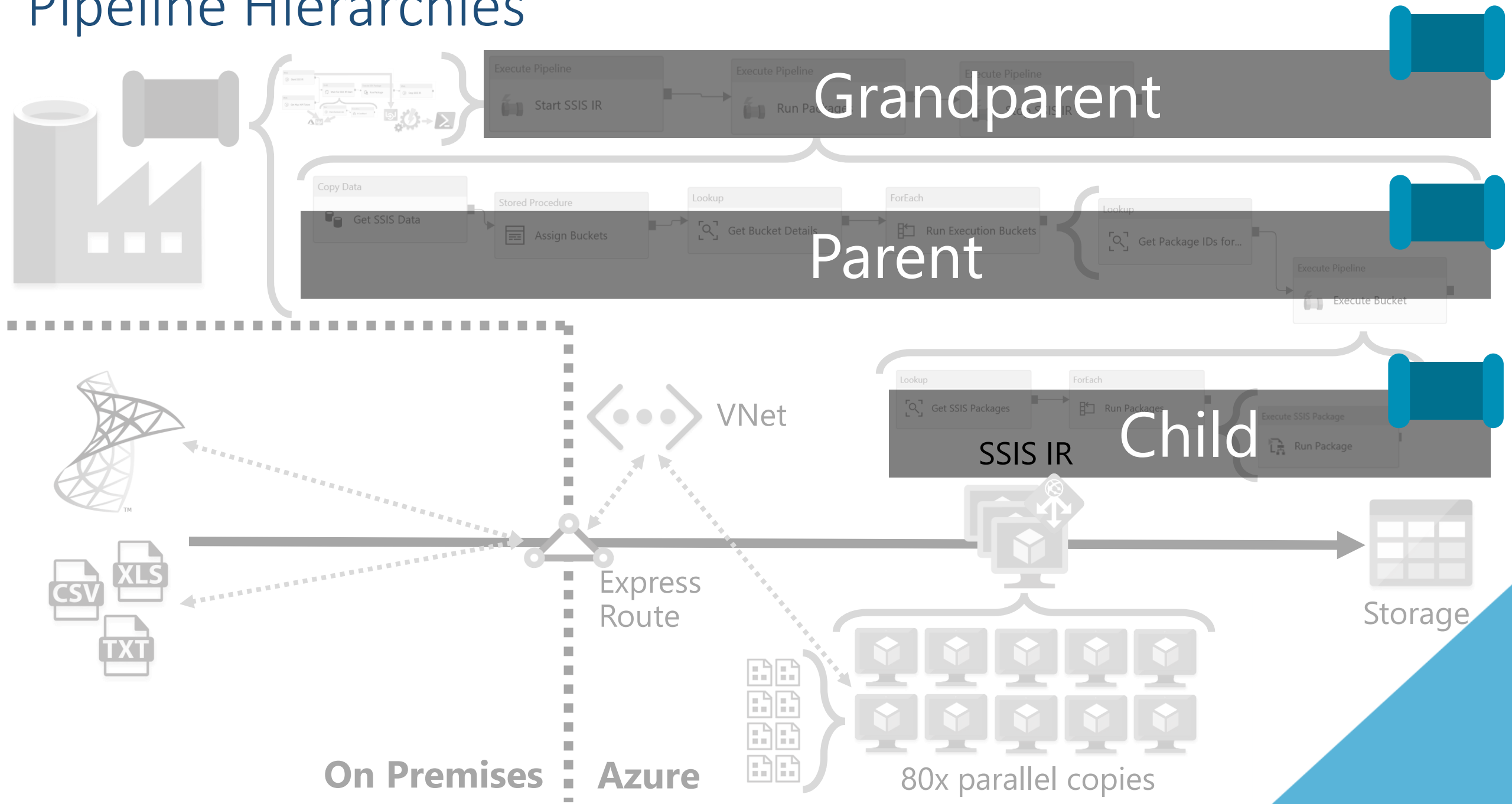
<https://github.com/MicrosoftDocs/azure-docs/blob/master/includes/azure-data-factory-limits.md>

On Premises : Azure 80x parallel copies

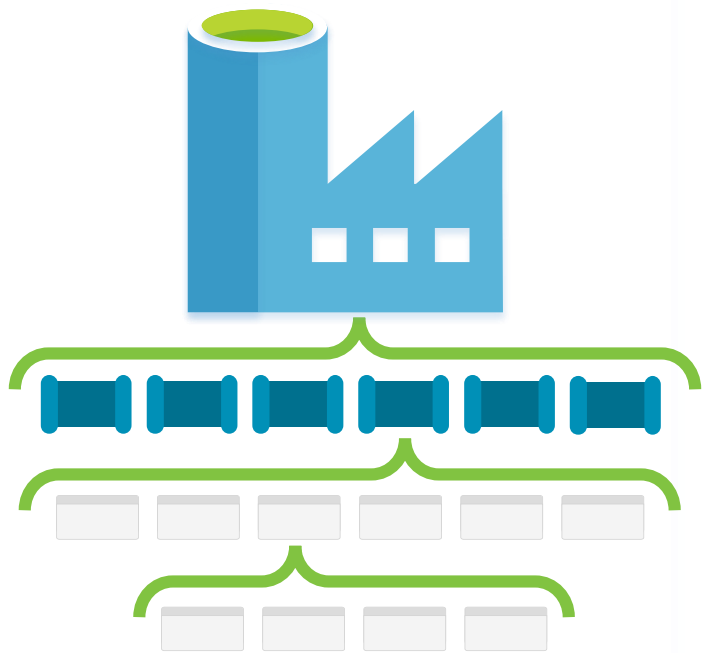
SSIS IR & Package Complete Orchestration Solution



Pipeline Hierarchies




Scale Out ~~Execution~~ Everything!



Pattern Summary


Execute Pipeline

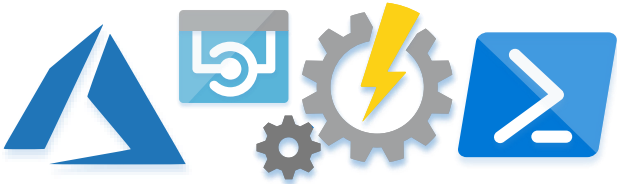
 Grandparent



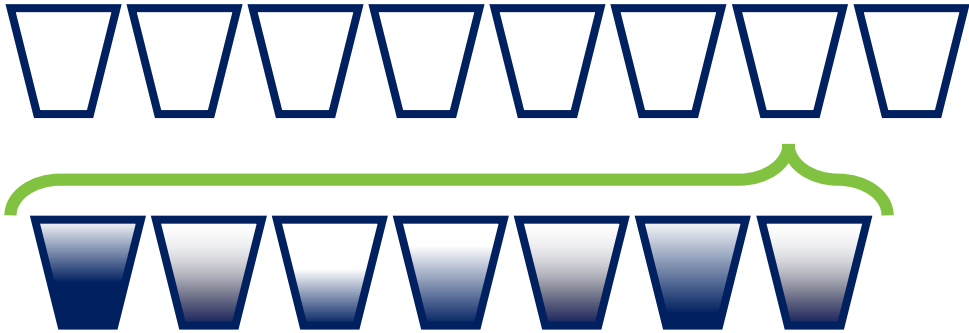
Control Flow and Pipeline Triggers

Execute Pipeline

 Parent




Platform Component Control



Manage Parallel Activities

Execute Pipeline

 Child

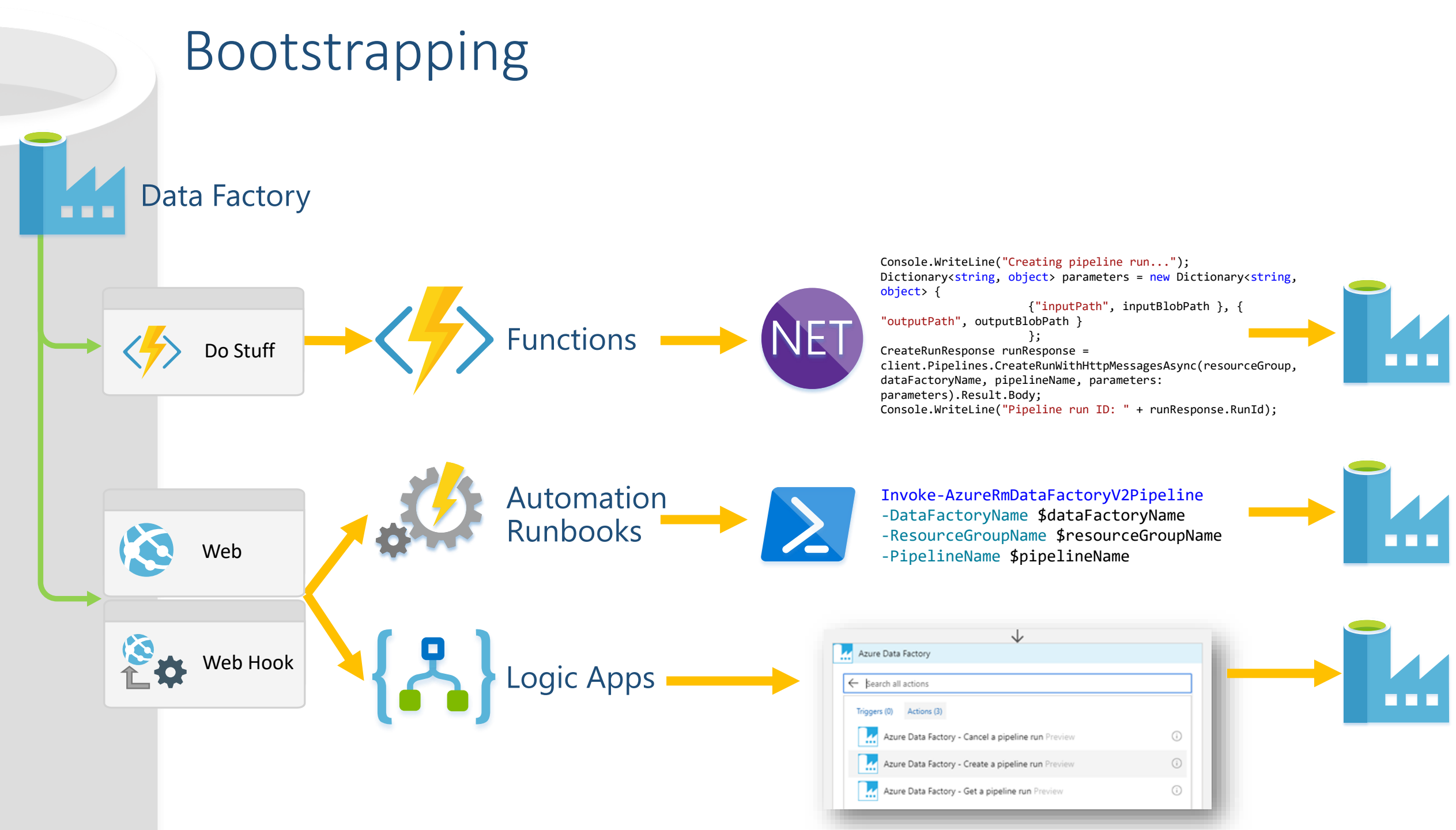


Service Level Executions

Solution Bootstrapping



Bootstrapping



Bootstrapping



Data Factory



Tenant 1



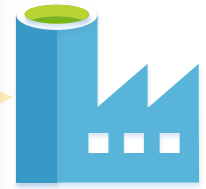
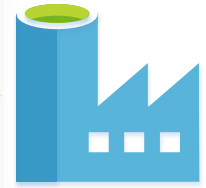
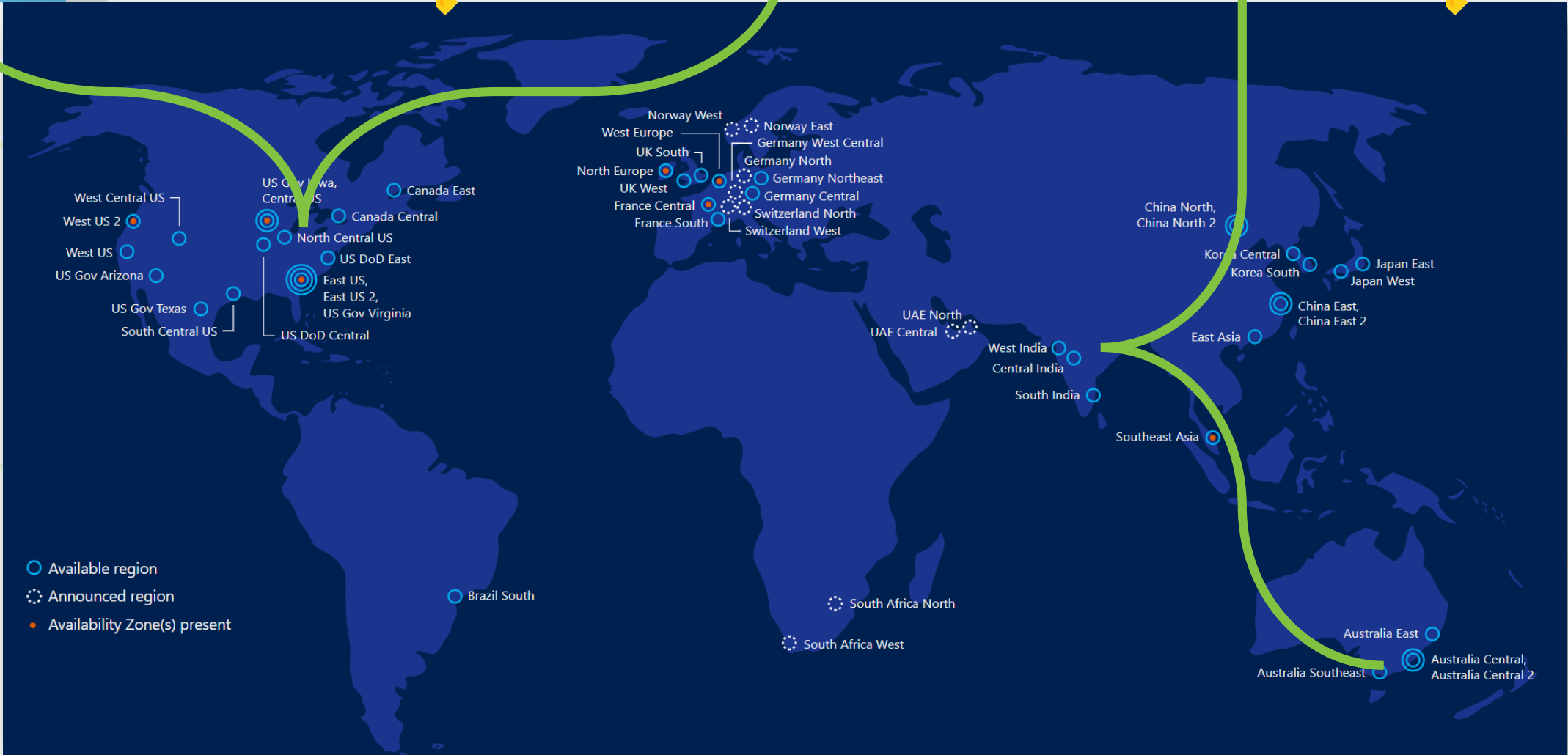
Subscription 1



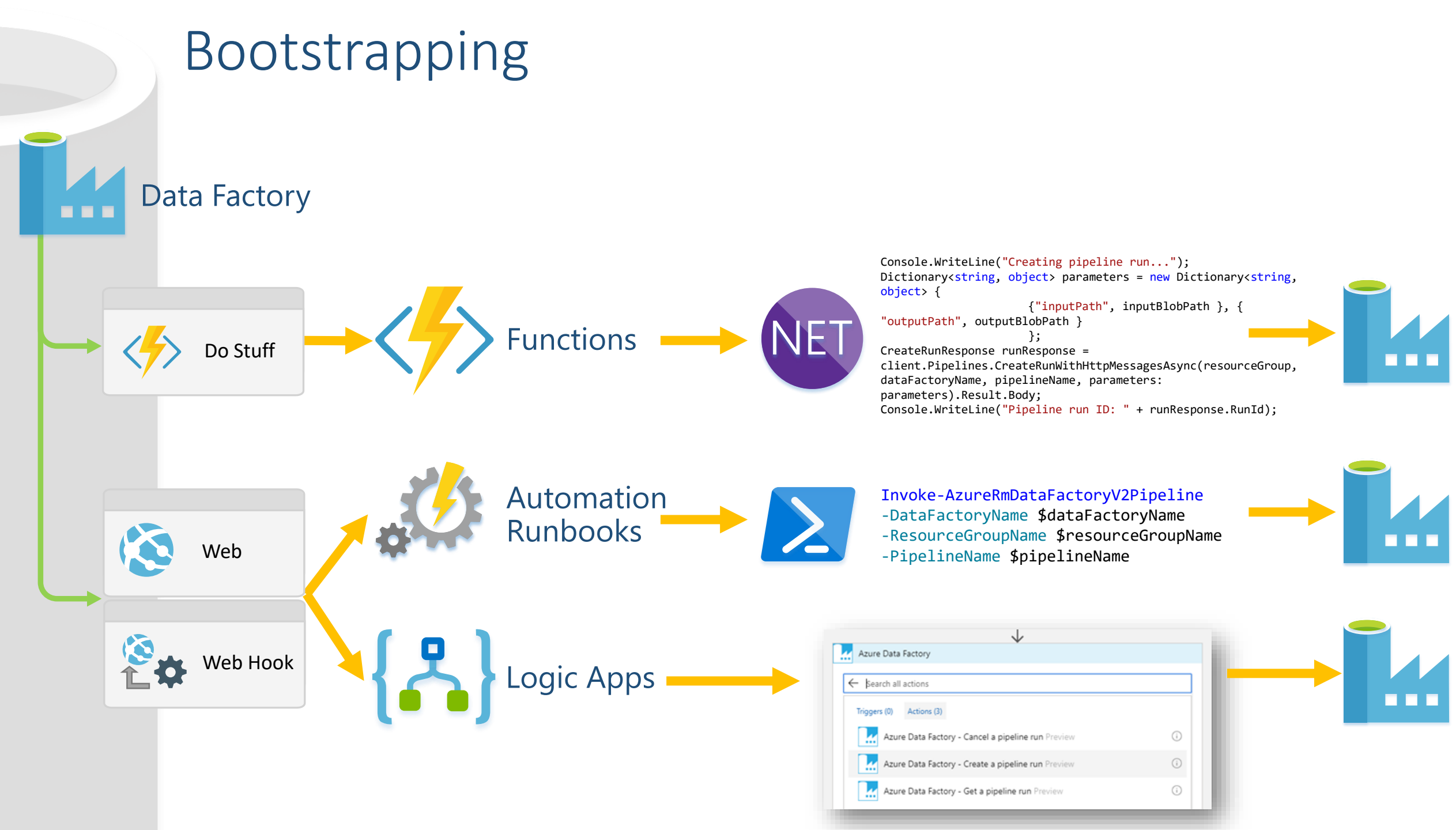
Tenant 2



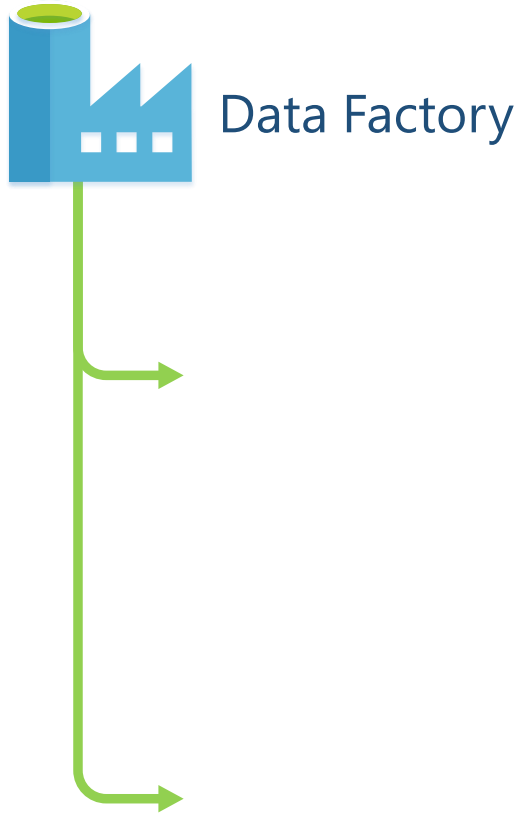
Subscription 2



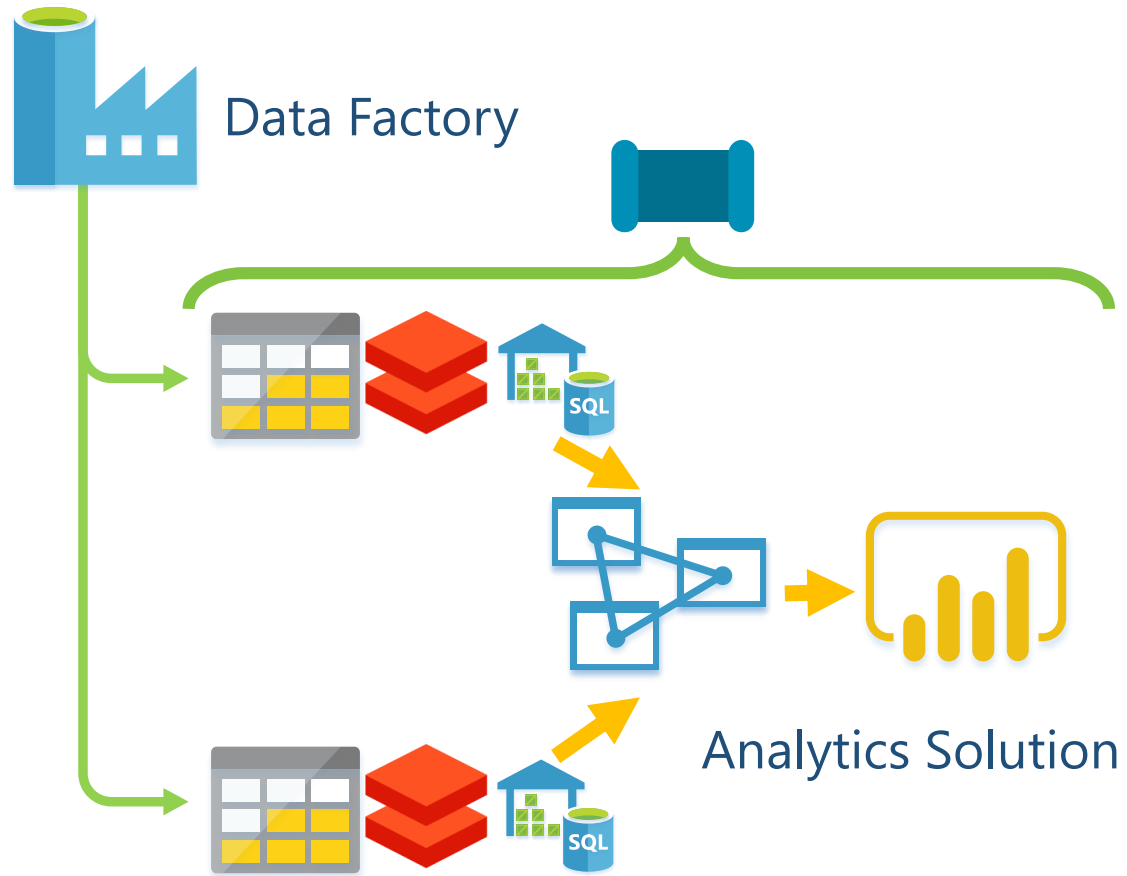
Bootstrapping



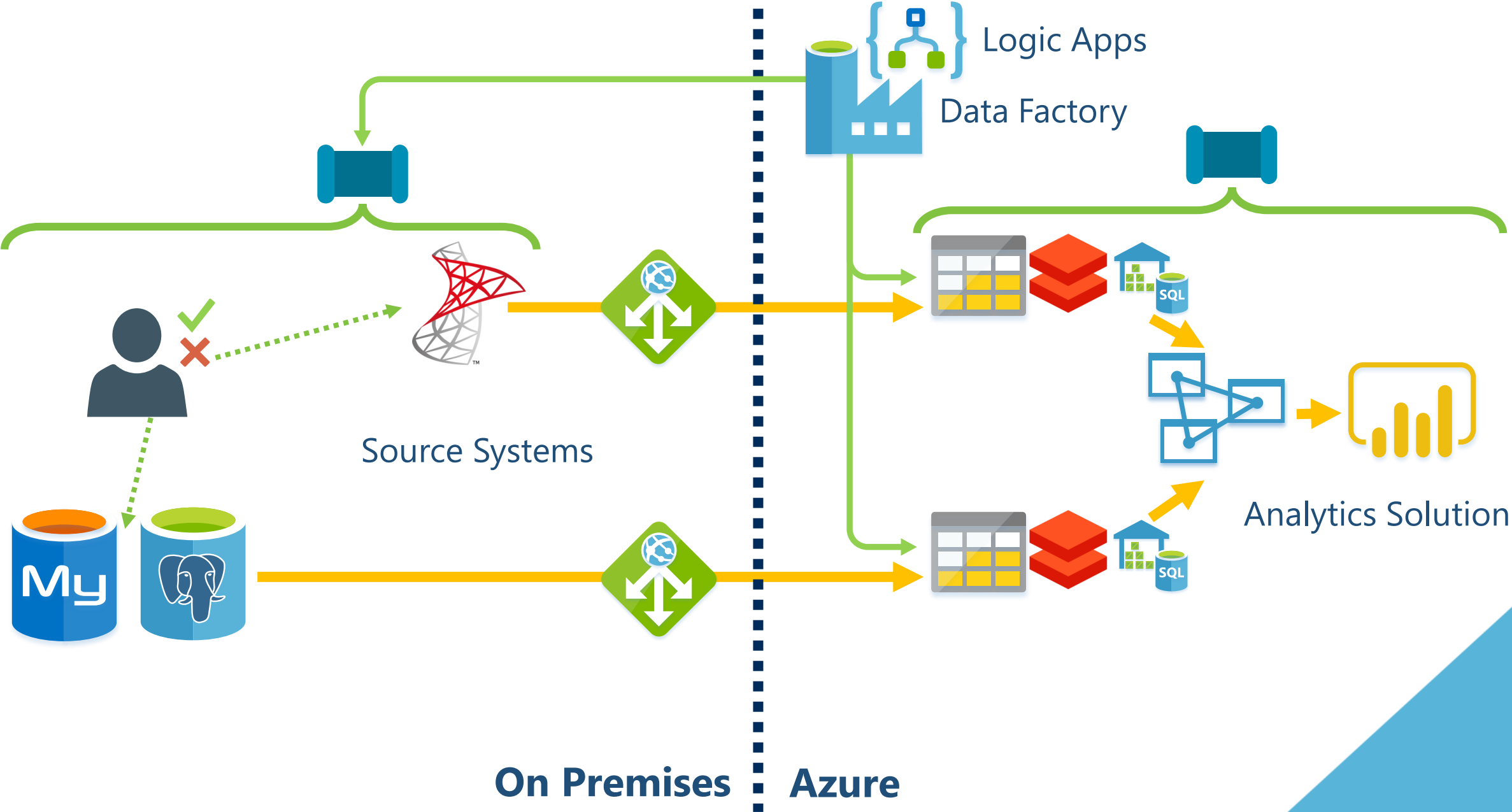
Bootstrapping – Wider Analytics Solution



Bootstrapping – Wider Analytics Solution



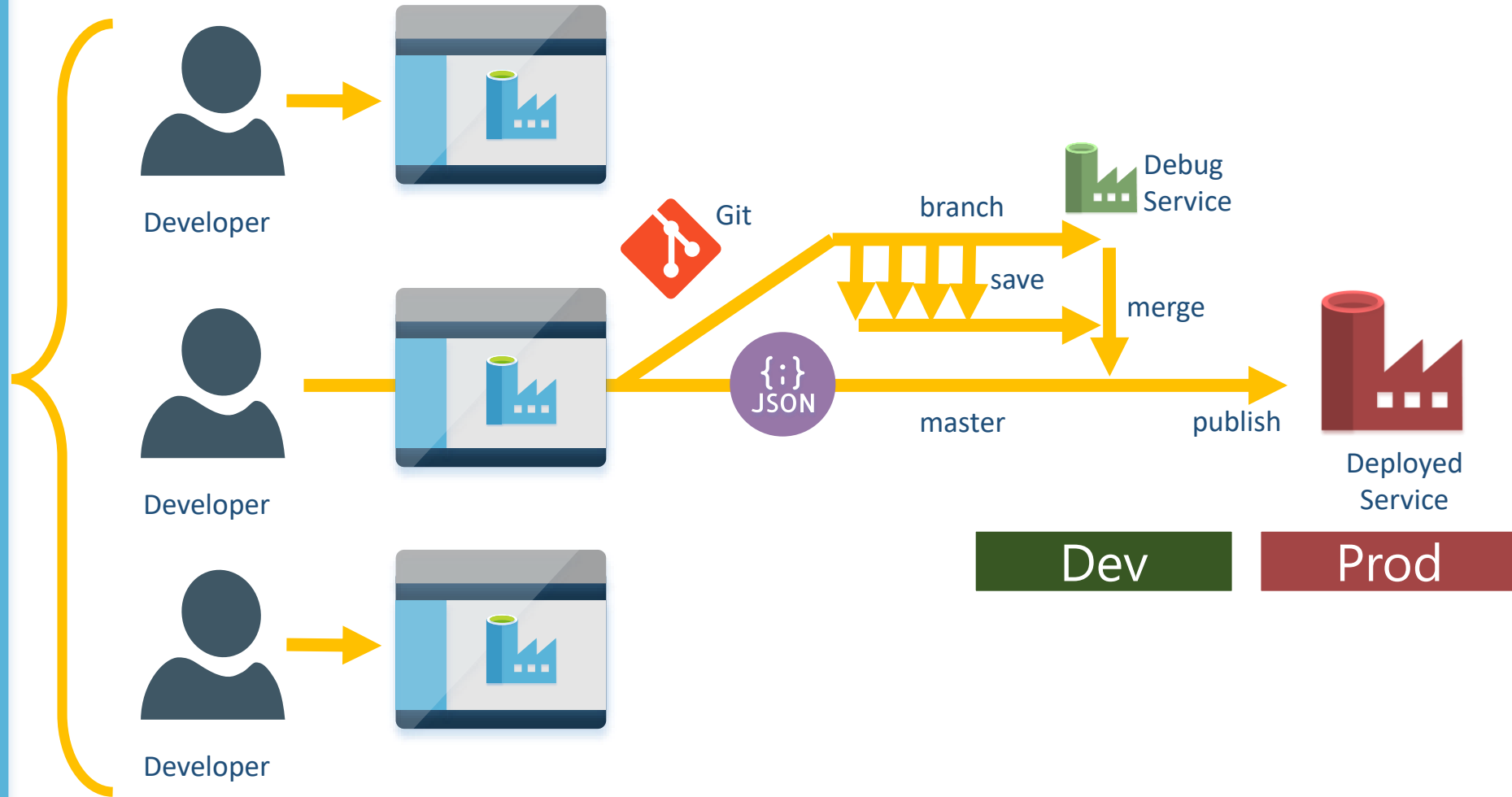
Bootstrapping – Wider Analytics Solution



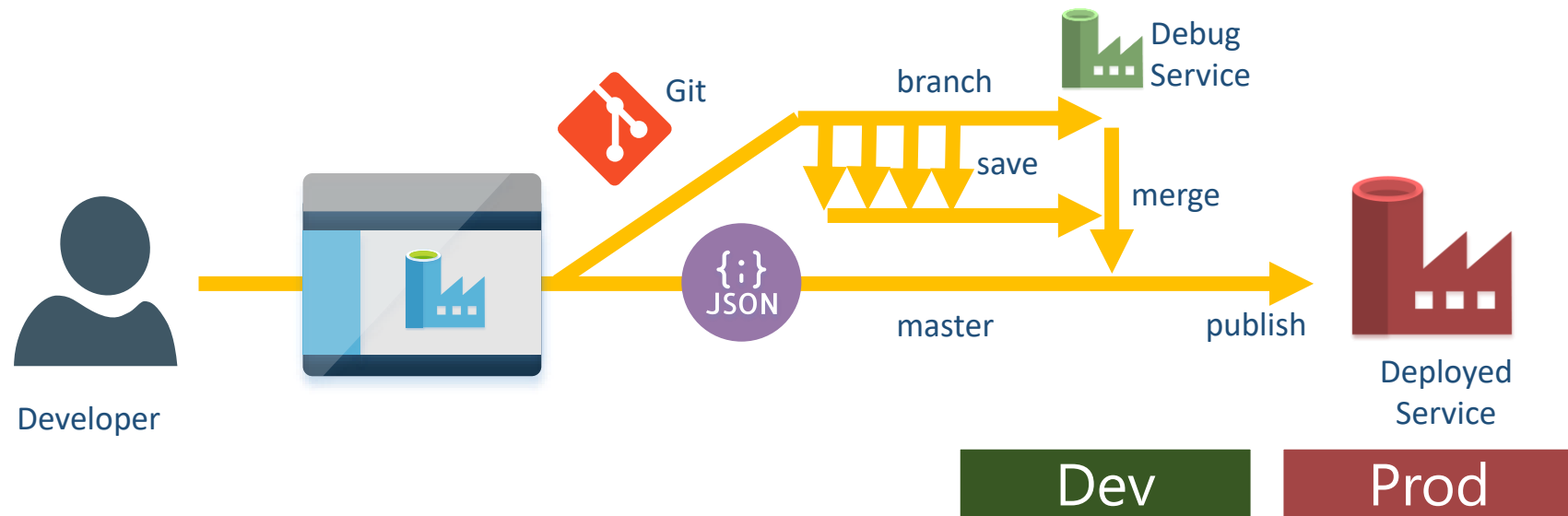
Data Factory DevOps – CI/CD



Data Factory Continuous Integration

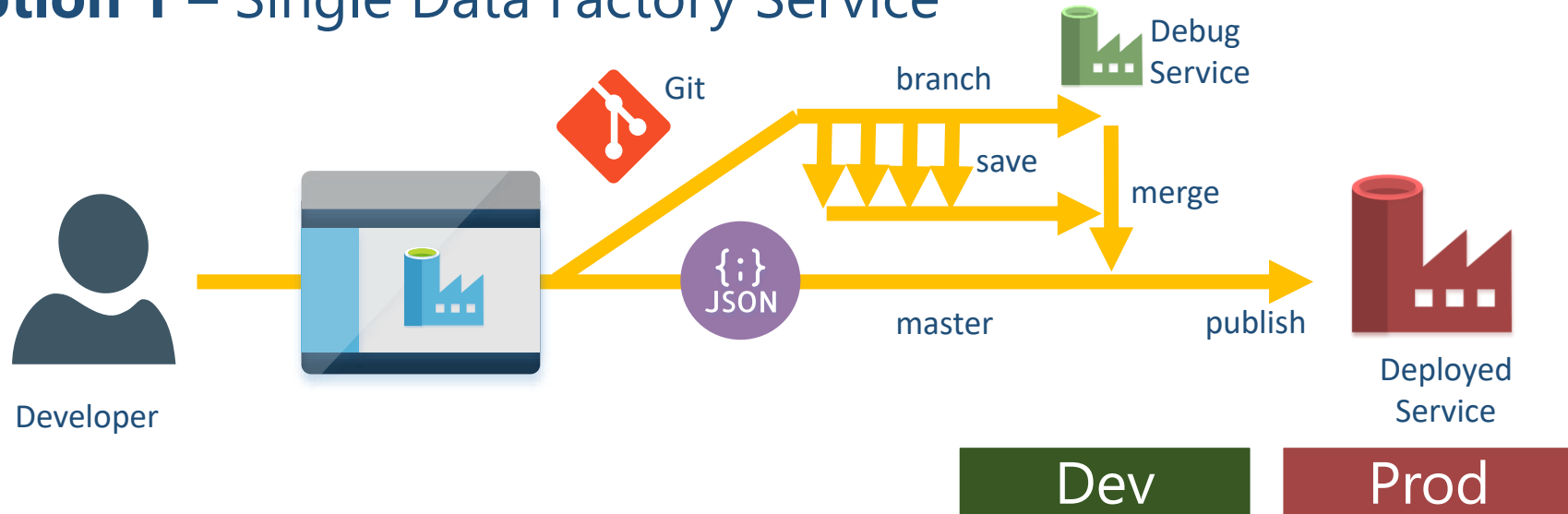


Data Factory Continuous Delivery

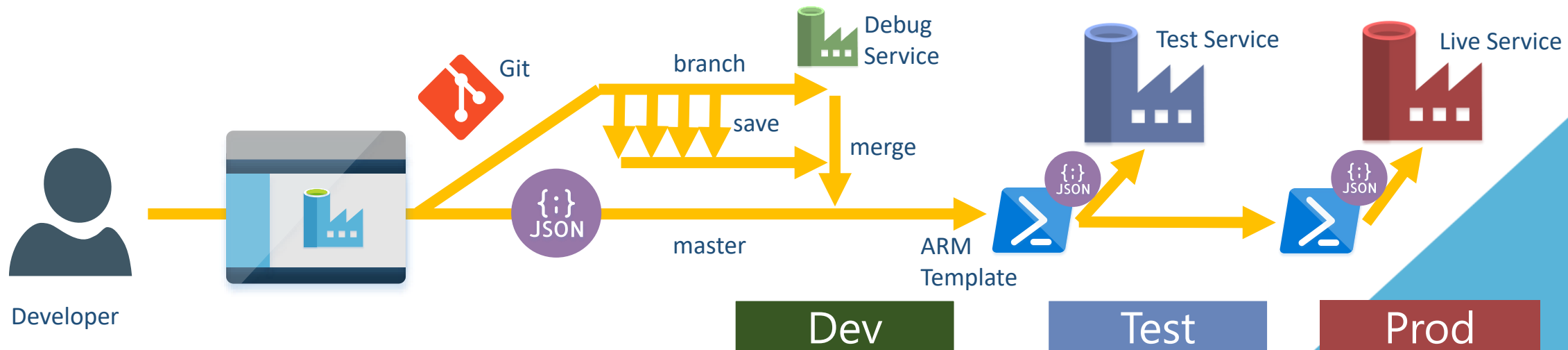


Data Factory Continuous Delivery

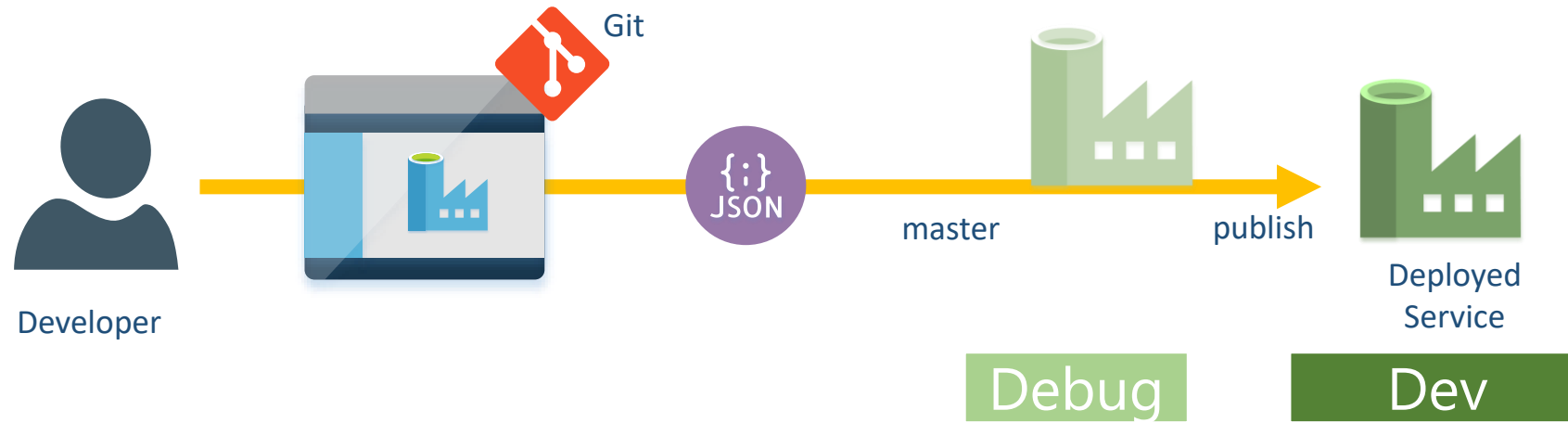
Option 1 – Single Data Factory Service



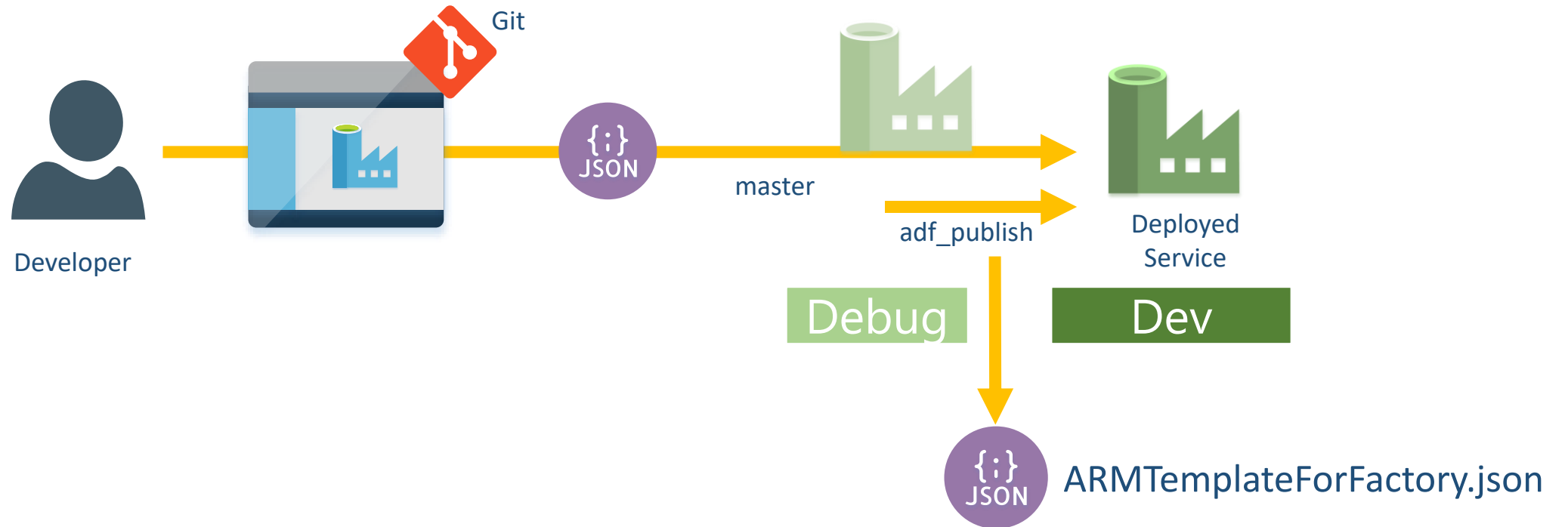
Option 2 – ARM Templates for Multiple Data Factory Services



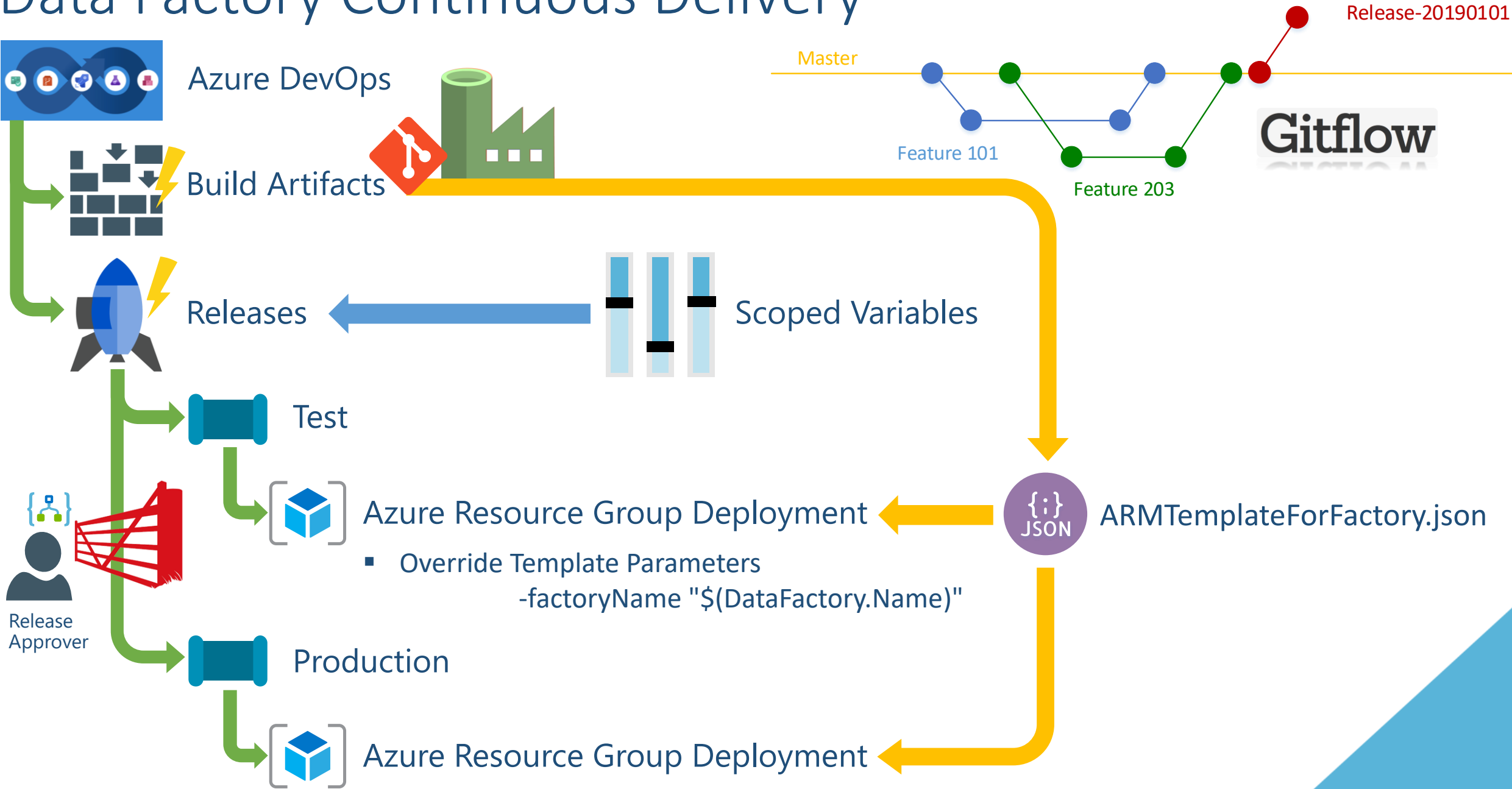
Data Factory Publish



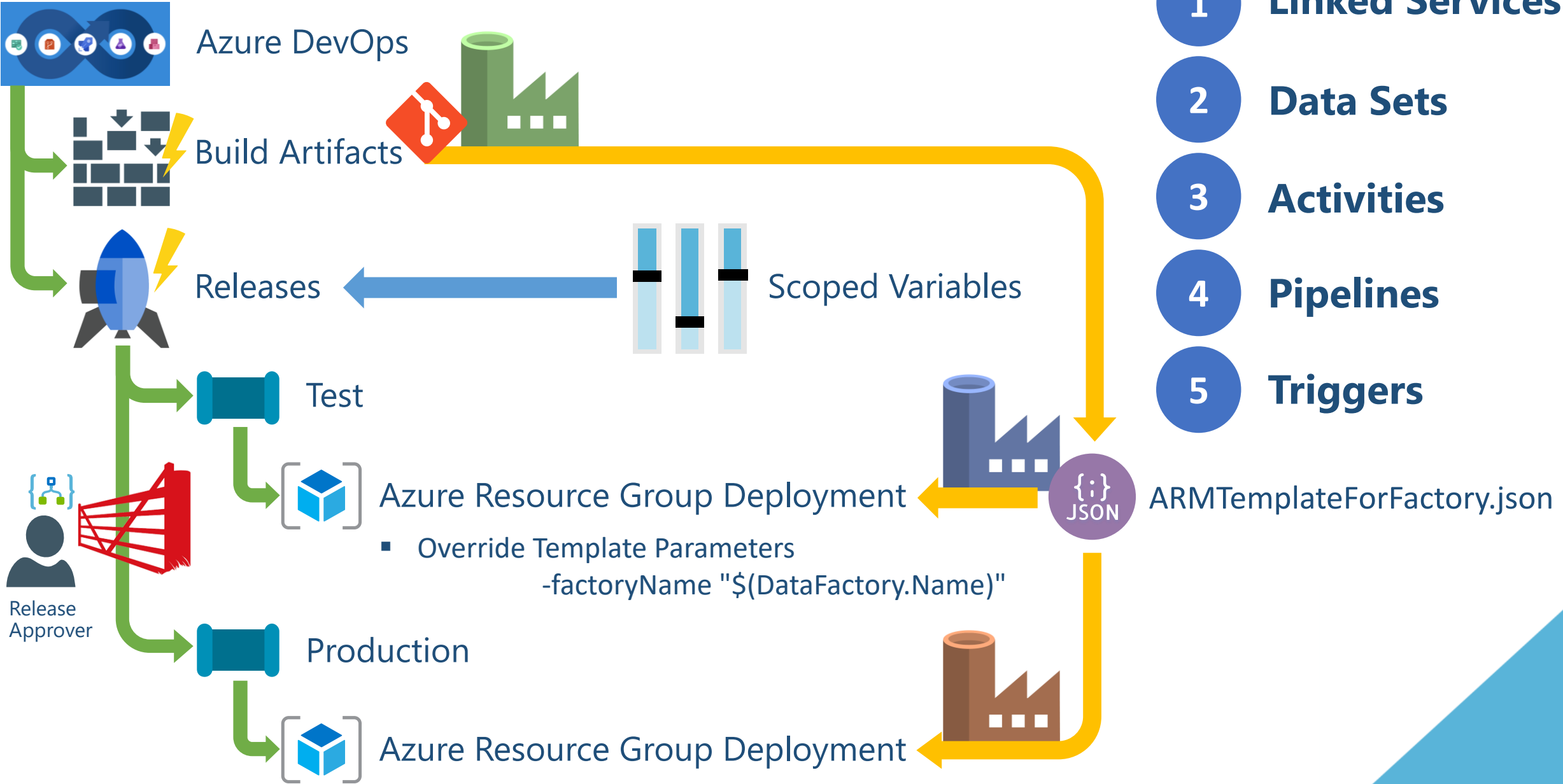
Data Factory Publish



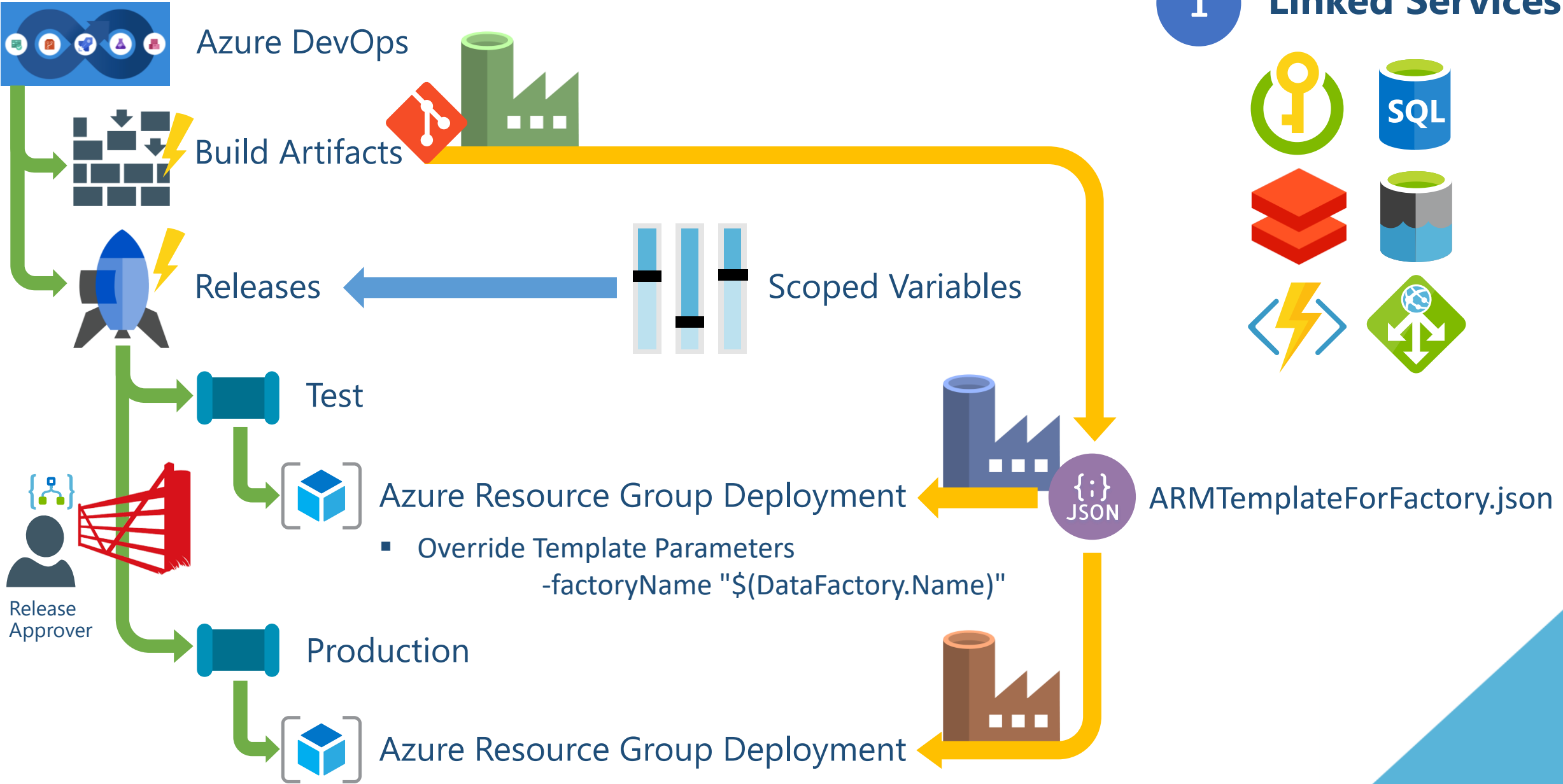
Data Factory Continuous Delivery



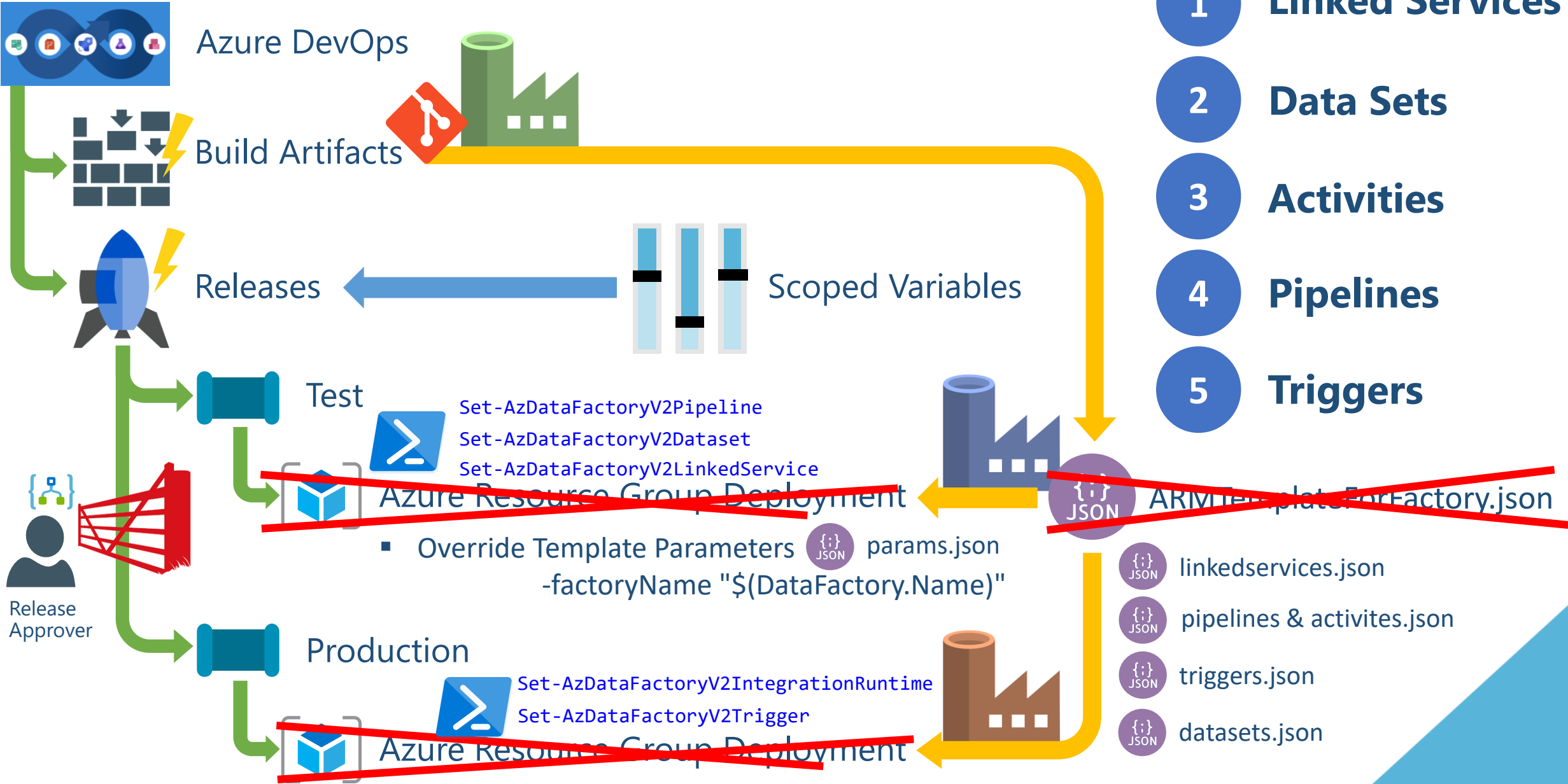
Data Factory Continuous Delivery



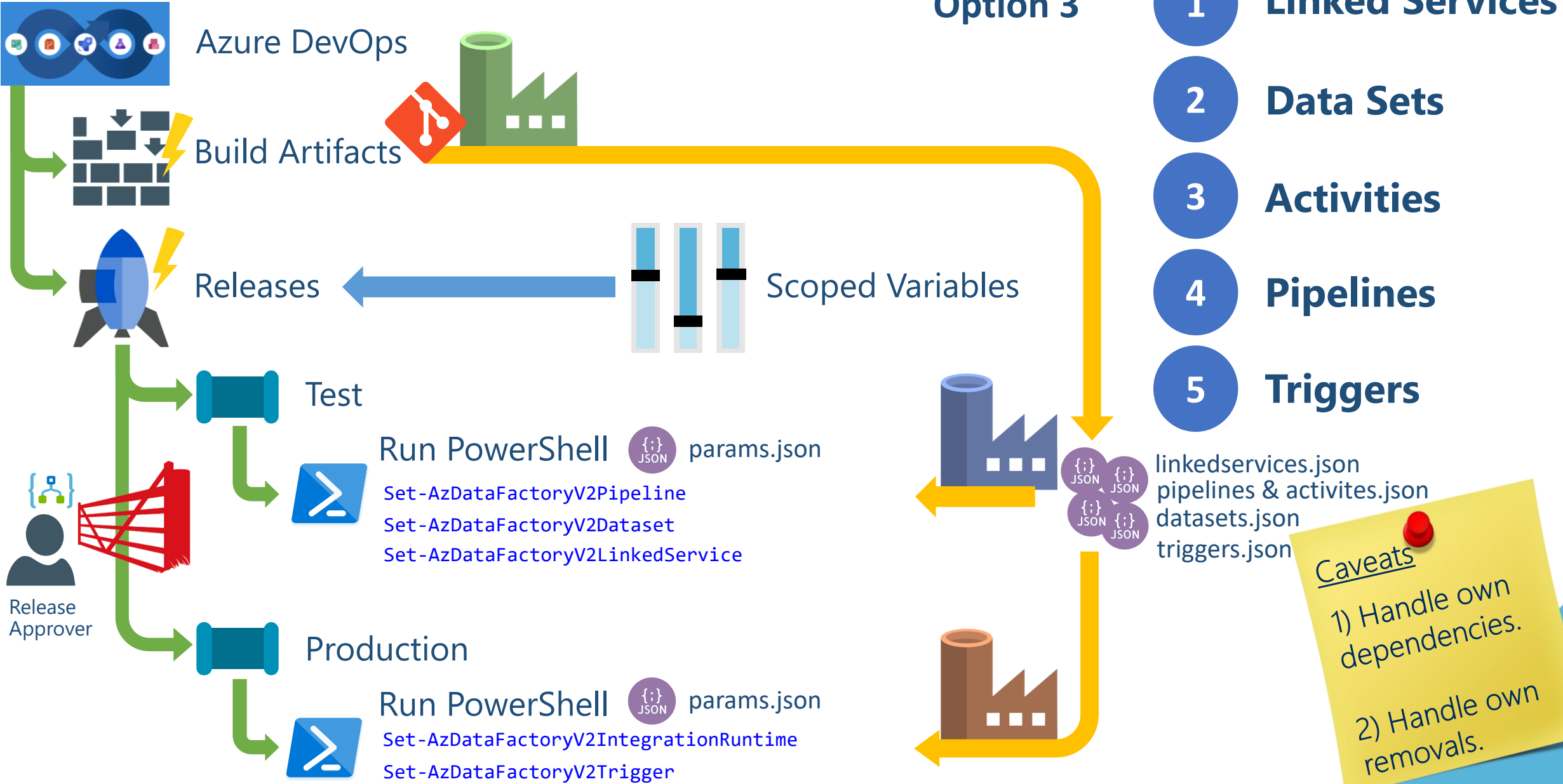
Data Factory Continuous Delivery



Data Factory Continuous Delivery



Data Factory Continuous Delivery - Bonus Option 3



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

Complex Azure Orchestration
Data Factory in Production

Thank you for listening...

Paul Andrew



altius

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

Twitter: [@mrpaulandrew](https://twitter.com/mrpaulandrew)
LinkedIn: [In/mrpaulandrew](https://in.linkedin.com/in/mrpaulandrew)

GitHub: github.com/mrpaulandrew

