



<https://www.facebook.com/AzureUserGroupBulgaria/>

Community



# November

## Getting started with Azure Data Factory



  
**AZURE**  
USER GROUP  
BULGARIA

 **puzl**  
coworking

**29.11.17**  
**19:00**



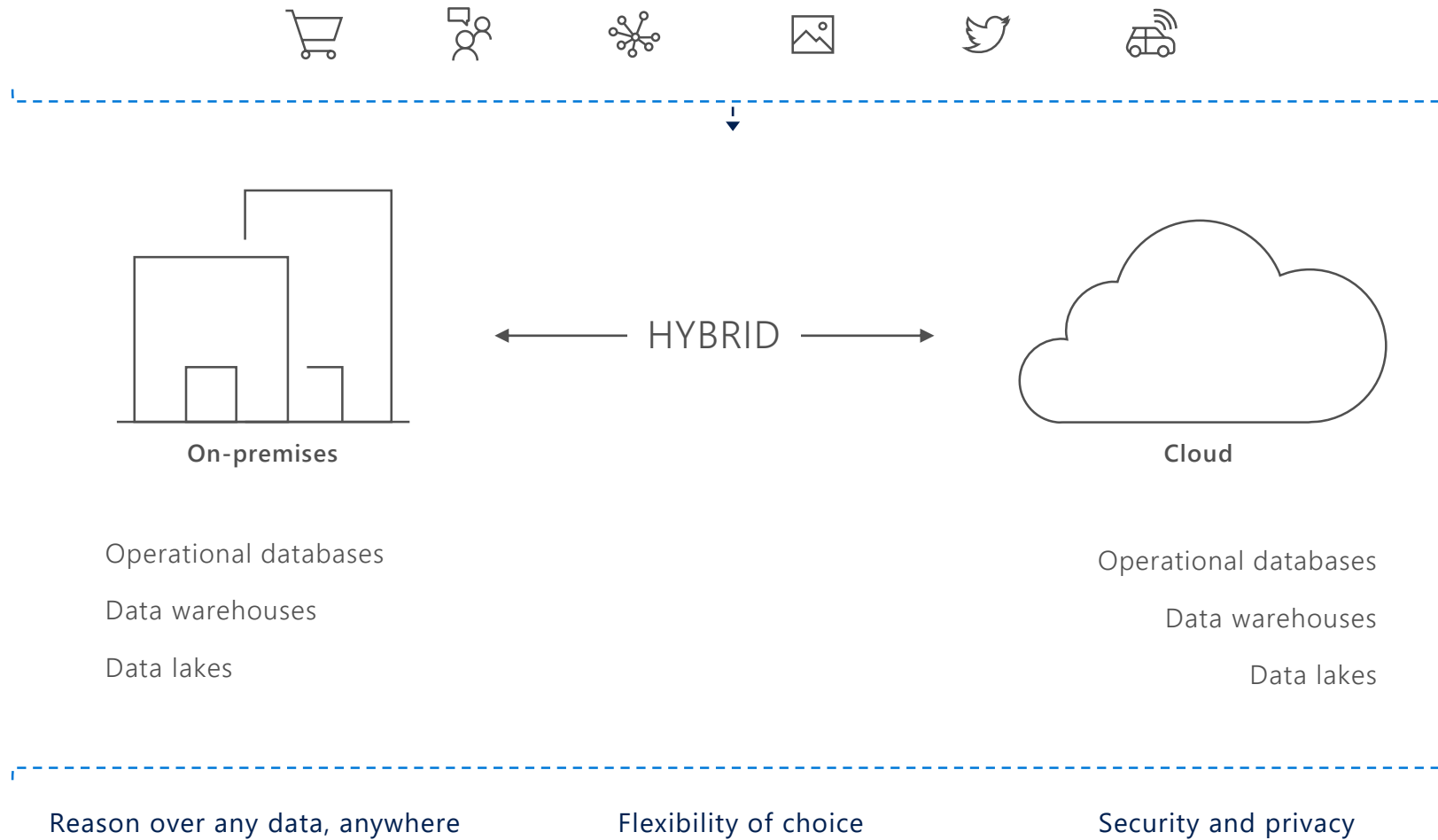
# December



# Azure Data Factory (v2)

Ivan Donev

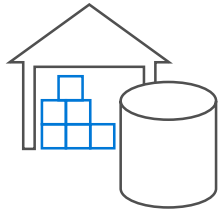
# THE MODERN DATA ESTATE



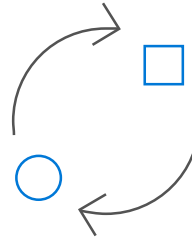
AZURE DATA FACTORY

# A Z U R E   D A T A   F A C T O R Y

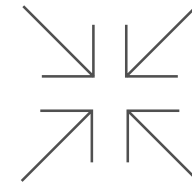
Fully-managed data integration service in the cloud



**Flexible**  
Data integration



**Hybrid**  
Data orchestration

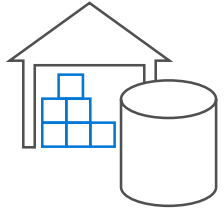


**Data movement**  
As-a-service

---

**Security and compliance**

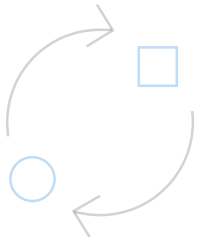




### Flexible

Data integration

Modernize your data warehouse with Azure big data and advanced analytics services such as HDInsight and Data lake Analytics



### Hybrid

Data orchestration

Build custom data-driven SaaS applications unique to your customer data using your language of choice



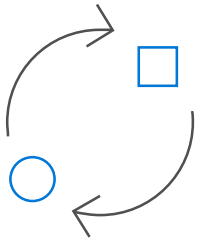
### Data movement

As-a-service

Bring together all your sources of data to understand your customers and drive impactful business decisions



**Flexible**  
Data integration



**Hybrid**  
Data orchestration



**Data movement**  
As-a-service

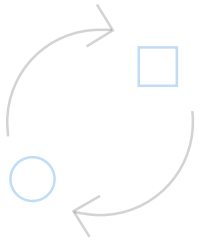
Orchestrate your data pipeline wherever your data lives – in cloud or in self-hosted environment

Meet your security and compliance needs while taking advantage of truly hybrid integration capabilities

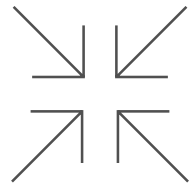
Execute your SQL Server Integration Services (SSIS) packages in the cloud



**Flexible**  
Data integration



**Hybrid**  
Data orchestration



**Data movement**  
As-a-service

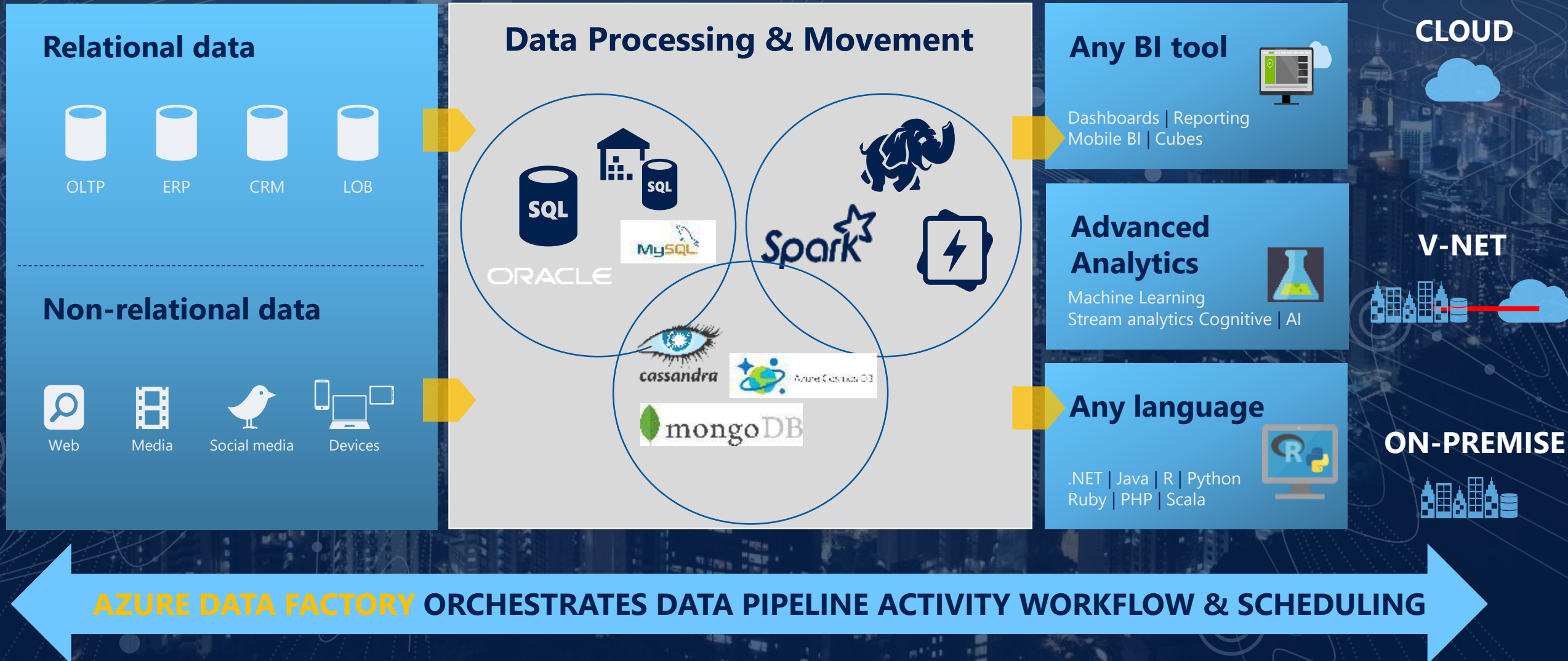
Accelerate integration with managed data movement as-a-service

Improve your TCO with 30+ natively supported connectors across 18 global points of presence

Elastic data movement at scale

Serverless data movement with no infrastructure to manage

# HYBRID DATA INTEGRATION AT SCALE



# ADF: Cloud-First Data Integration Objectives

- Consume hybrid disparate data
  - On-prem + Cloud
  - Grow ADF ecosystem of structured, un-structured, semi-structured data connectors
- Calculate and format data for analytics
  - Transform, aggregate, join, normalize
  - Separate data flow (transformation) from control flow (orchestration)
- Address large-scale Big Data requirements
  - Scale-up or Scale-out data movement and transformation
  - Support multiple processing engines
- Operationalize
  - Support flexible scheduling and triggering mechanism for broad range of use cases
  - Manage & monitor multiple pipelines (via Azure Monitor & OMS)
  - Support secure VNET environments
- Enable SSIS package execution
  - Execute SSIS packages in ADF Integration Runtime

# ADF: Cloud-First Data Integration Scenarios

## Lift and Shift to the Cloud

- Migrate on-prem DW to Azure
- Lift and shift existing on-prem SSIS packages to cloud
- No changes needed to migrate SSIS packages to Cloud service

## DW Modernization

- Modernizing DW arch to reduce cost & scale to needs of big data (volume, variety, etc)
- Flexible wall-clock and triggered event scheduling
- Incremental Data Load

## Build Data-Driven, Intelligent SaaS Application

- C#, Python, PowerShell, ARM support

## Big Data Analytics

- Customer profiling, Product recommendations, Sentiment Analysis, Churn Analysis, Customized offers, customer usage tracking, customized marketing
- On-demand Spark cluster support

## Load your Data Lake

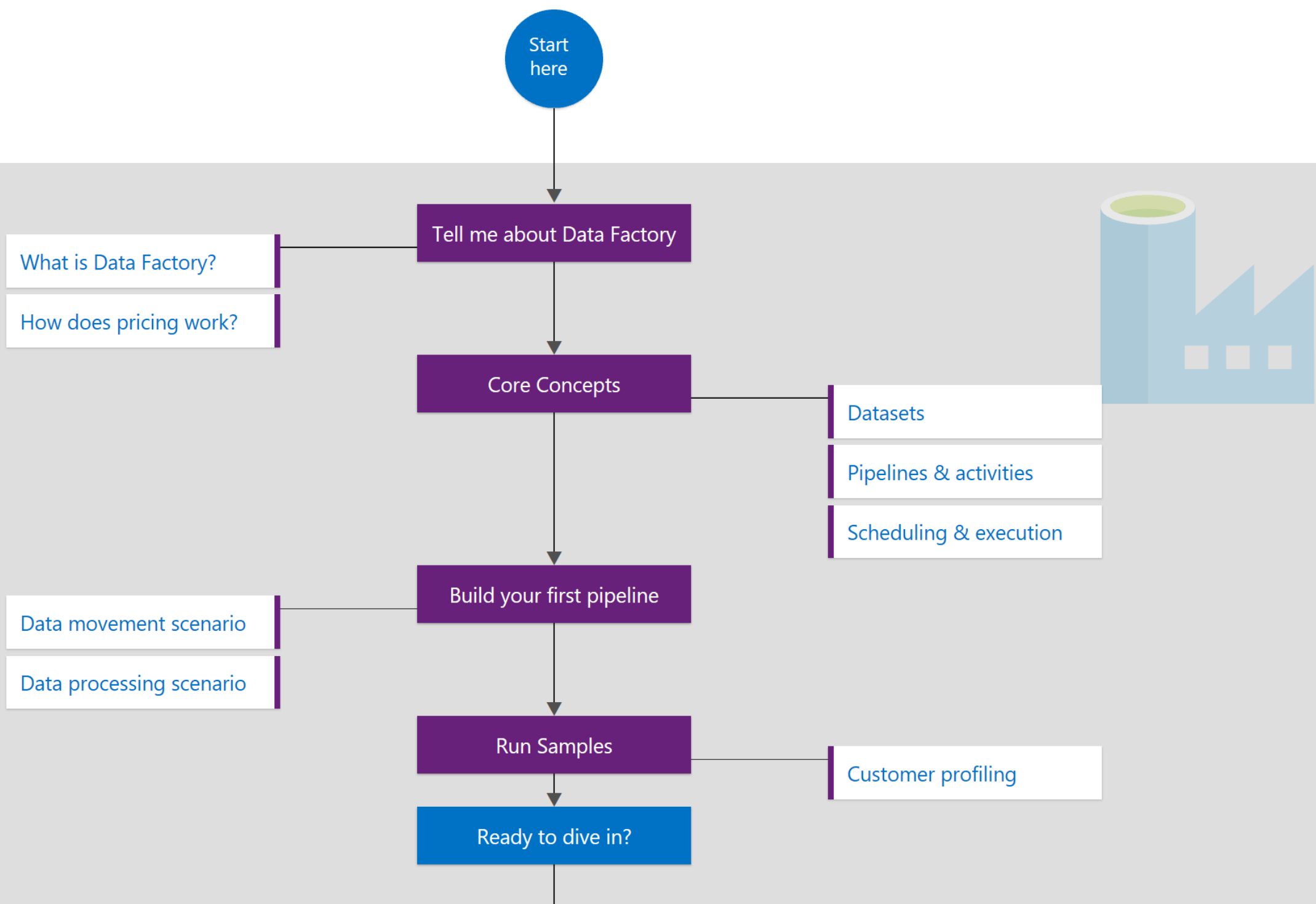
- Separate control-flow to orchestrate complex patterns with branching, looping, conditional processing



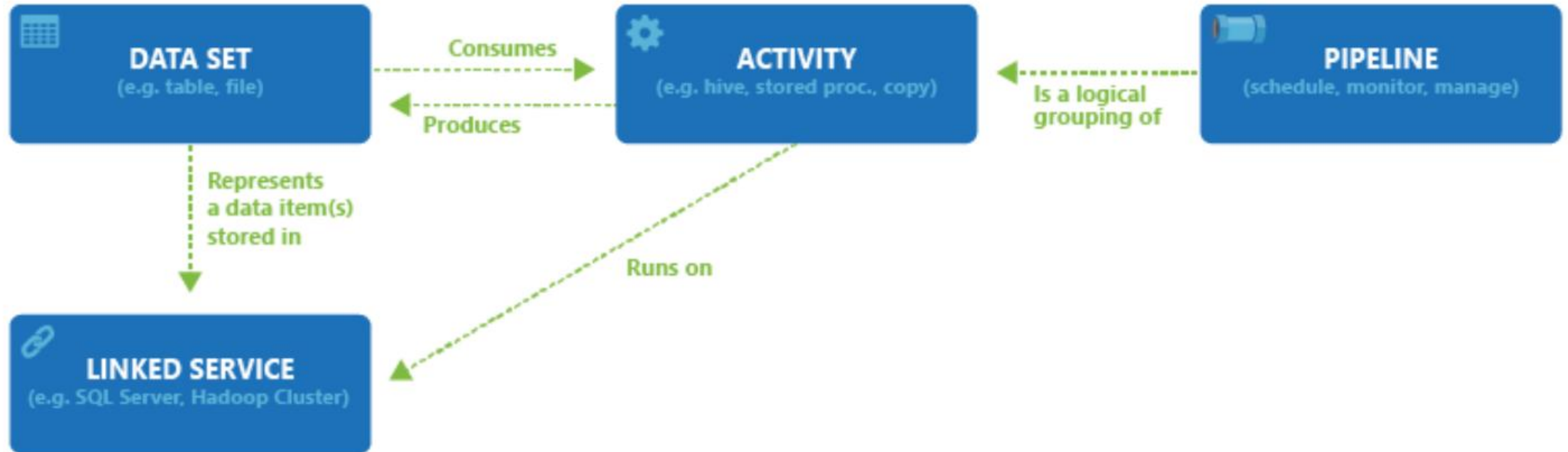
# ADF V2 Improvements

- Integration Runtimes (IR) replace DMG, provide data movement and activity dispatch on-prem or in the cloud
- Supports resources within virtual networks
- Integration Runtime includes SSIS option to lift & shift SSIS packages to the Cloud
- Separation of “control flow” & “data flow” capabilities for more flexible pipeline management
  - Looping, conditionals, dependencies, parameters
- Python SDK
- On-Demand Spark support
- Flexible pipeline scheduling with wall-clock and triggered executions
- Expanded use cases: From primarily time window-oriented pipelines, to trigger-based on-demand for more flexible ETL and data integration orchestrations
- (Coming in GA) UX pipeline and data transformation builder code-free experience

# Basic terminology



# Datasets and linked services in Azure Data Factory



# Pipeline execution

Each run has unique ID

- 10:00, 11:00 and 11:30 are three different runs
- Manually or triggered
  - Trigger: wall-clock;
  - Currently no support for event-based triggers

# Manual execution

## POST

<https://management.azure.com/subscriptions/mySubId/resourceGroups/myResourceGroup/providers/Microsoft.DataFactory/factories/myDataFactory/pipelines/copyPipeline/createRun?api-version=2017-03-01-preview>



# Manual execution

## PowerShell

```
Invoke-AzureRmDataFactoryV2Pipeline -DataFactory $df -  
PipelineName "Adfv2QuickStartPipeline" -ParameterFile  
.\PipelineParameters.json
```

# Manual execution

.NET

```
client.Pipelines.CreateRunWithHttpMessagesAsync  
    (resourceGroup, dataFactoryName, pipelineName,  
parameters)
```

The background is a high-angle, aerial photograph of a dense urban skyline, likely Hong Kong, with numerous skyscrapers and buildings. The image is overlaid with a semi-transparent dark blue filter. White, thin-lined geometric patterns, including concentric circles, dots, and wavy lines, are scattered across the image, particularly around the text and in the corners, giving it a technical or digital feel.

Let's create our first  
ADF job

# New ADF V2 Concepts

Concept	Description	Sample
Control Flow	Orchestration of pipeline activities that includes chaining activities in a sequence, branching, conditional branching based on an expression, parameters that can be defined at the pipeline level and arguments passed while invoking the pipeline on demand or from a trigger. Also includes custom state passing and looping containers, I.e. For-each, Do-Until iterators.	<pre>{ "name": "MyForEachActivityName",   "type": "ForEach",   "typeProperties": { "isSequential": "true",     "items": "@pipeline().parameters.mySinkDatasetFolderPathCollection",     "activities": [       {         "name": "MyCopyActivity", "type": "Copy", "typeProperties": ...</pre>
Runs	A Run is an instance of the pipeline execution. Pipeline Runs are typically instantiated by passing the arguments to the parameters defined in the Pipelines. The arguments can be passed manually or properties created by the Triggers.	POST <a href="https://management.azure.com/subscriptions/&lt;subId&gt;/resourceGroups/&lt;resourceGroupName&gt;/providers/Microsoft.DataFactory/factories/&lt;dataFactoryName&gt;/pipelines/&lt;pipelineName&gt;/createRun?api-version=2017-03-01-preview">https://management.azure.com/subscriptions/&lt;subId&gt;/resourceGroups/&lt;resourceGroupName&gt;/providers/Microsoft.DataFactory/factories/&lt;dataFactoryName&gt;/pipelines/&lt;pipelineName&gt;/createRun?api-version=2017-03-01-preview</a>
Activity Logs	Every activity execution in a pipeline generates activity start and activity end logs event	
Integration Runtime	Replaces DMG as a way to move & process data in Azure PaaS Services, self-hosted or on prem or IaaS Works with VNETs Enables SSIS package execution	
Scheduling	Flexible Scheduling Wall-clock scheduling Event-based triggers	<pre>"type": "ScheduleTrigger", "typeProperties": {   "recurrence": {     "frequency": "&lt;&lt;Minute, Hour, Day, Week, Year&gt;&gt;",     "interval": "&lt;&lt;int&gt;&gt;", // optional, how often to fire (default to 1)     "startTime": "&lt;&lt;datetime&gt;&gt;",     "endTime": "&lt;&lt;datetime&gt;&gt;",     "timeZone": "&lt;&lt;default UTC&gt;&gt;"   },   "schedule": { // optional (advanced scheduling specifics)     "hours": [&lt;&lt;0-24&gt;&gt;],     "weekDays": "": [&lt;&lt;Monday-Sunday&gt;&gt;],     "minutes": [&lt;&lt;0-60&gt;&gt;],     "monthDays": [&lt;&lt;1-31&gt;&gt;],     "monthlyOccurrences": [       {         "day": "&lt;&lt;Monday-Sunday&gt;&gt;",         "occurrence": "&lt;&lt;1-5&gt;&gt;"</pre>

# New ADF V2 Concepts

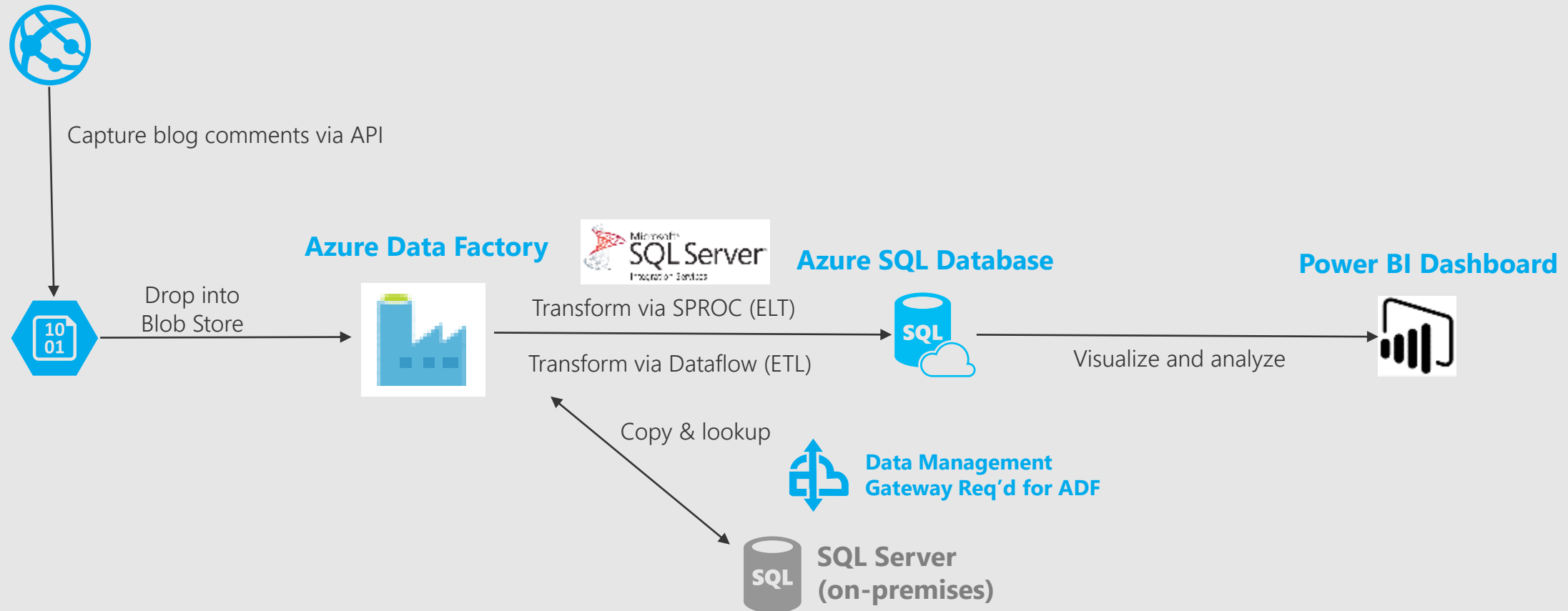
Concept	Description	Sample
On-Demand Execution	Instantiate a pipeline by passing arguments as parameters defined in a pipeline and execute from script / REST / API.	Invoke-AzureRmDataFactoryV2PipelineRun -DataFactory \$df -PipelineName "Adfv2QuickStartPipeline" -ParameterFile .\PipelineParameters.json
Parameters	<p>Name-value pairs defined in the pipeline. Arguments for the defined parameters are passed during execution from the run context created by a Trigger or pipeline executed manually. Activities within the pipeline consume the parameter values.</p> <p>A <b>Dataset</b> is a strongly typed parameter and a reusable/referenceable entity. An activity can reference datasets and can consume the properties defined in the Dataset definition</p> <p>A <b>Linked Service</b> is also a strongly typed parameter containing the connection information to either a data store or a compute environment. It is also a reusable/referenceable entity.</p>	<p>Accessing parameters of other activities Using expressions</p> <p>@parameters("{name of parameter}")</p> <p>@activity("{Name of Activity}").output.RowsCopied</p>
Incremental Data Loading	Leverage parameters and define your high-water mark for delta copy while moving dimension or reference tables from a relational store either on premises or in the cloud to load the data into the lake	

# Patterns & Scenarios for ADF V2



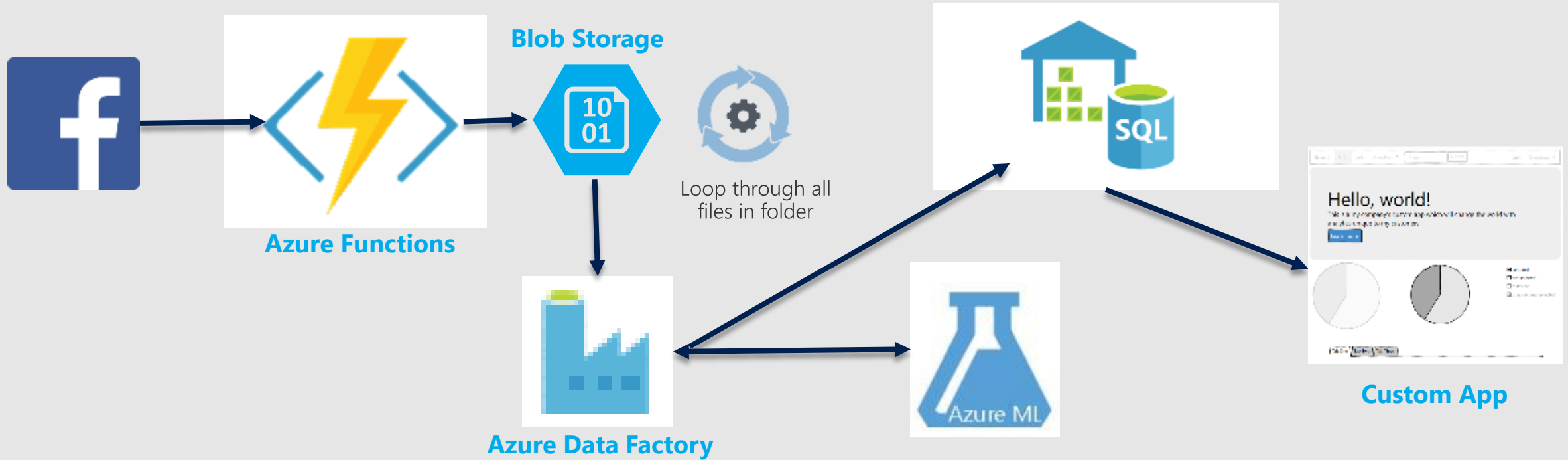
# Hybrid Data Integration Pattern 1:

## Analyze blog comments



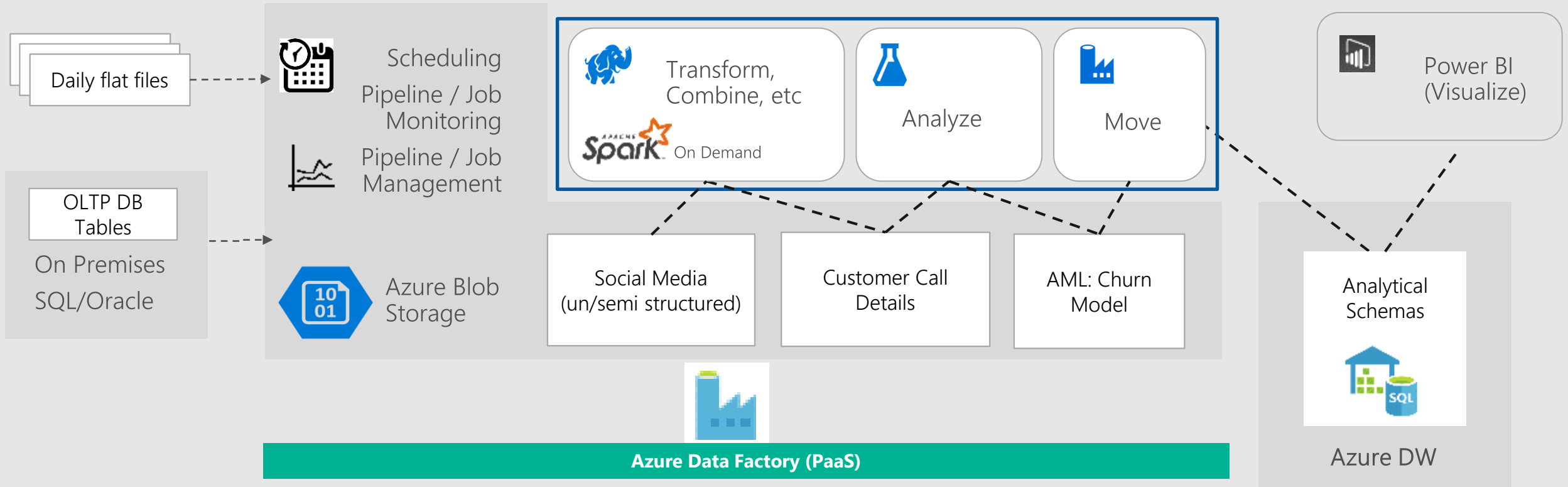
# Hybrid Data Integration Pattern 2:

Data-Driven SaaS App: Sentiment Analysis w/Machine Learning



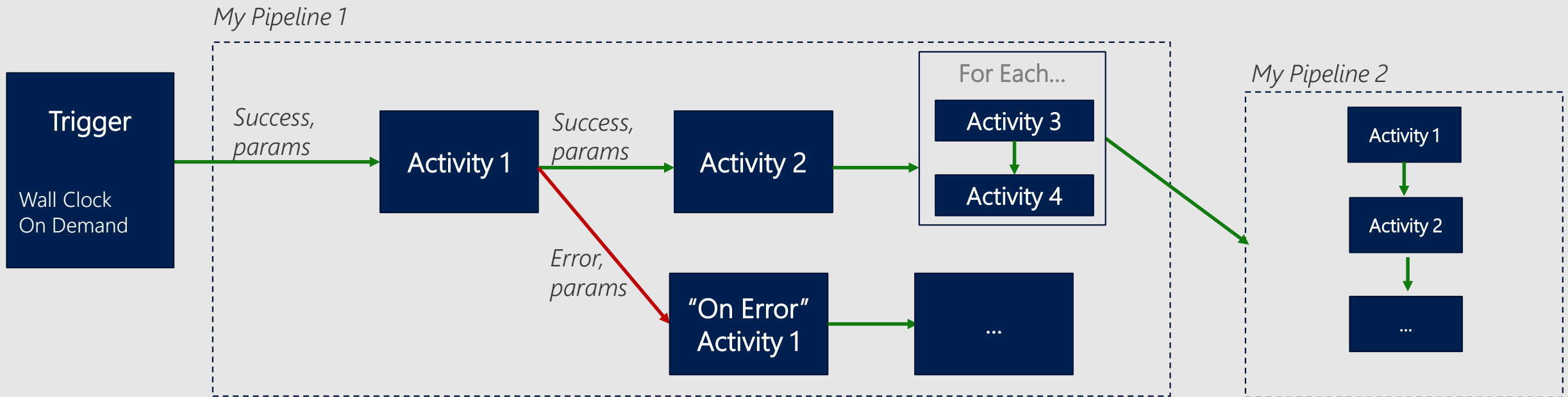
# Hybrid Data Integration Pattern 3:

## Modern Data Warehouse



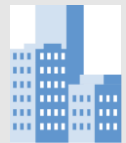
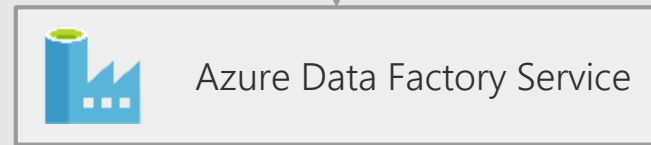
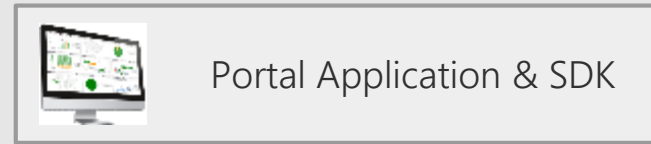
# ADFv2: Control Flow

Coordinate pipeline activities into finite execution steps to enable looping, conditionals and chaining while separating data transformations into individual data flows



# Integration Runtime

# ADF Integration Runtime (IR)



## Self-Hosted IR

Data Movement & Activity  
Dispatch on-prem, Cloud,  
VNET

## Azure IR

Data Movement & Activity  
Dispatch In Azure Public  
Network, SSIS  
*VNET coming soon*



- ADF compute environment with multiple capabilities:
  - Activity dispatch & monitoring
  - Data movement
  - SSIS package execution
- To integrate data flow and control flow across the enterprises' hybrid cloud, customer can instantiate multiple IR instances for different network environments:
  - On premises (similar to DMG in ADF V1)
  - In public cloud
  - Inside VNet
- Bring a consistent provision and monitoring experience across the network environments



←----→ Command & Control

↔ Data Flow



## UX & SDK

*Authoring | Monitoring/Mgmt*



## Azure Data Factory Service

*Scheduling | Orchestration | Monitoring*

### On Premises Apps & Data



TERADATA



cloudera



ORACLE

### Cloud Apps, Svcs & Data



Adobe

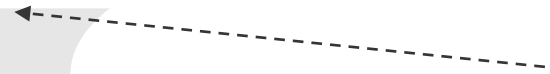
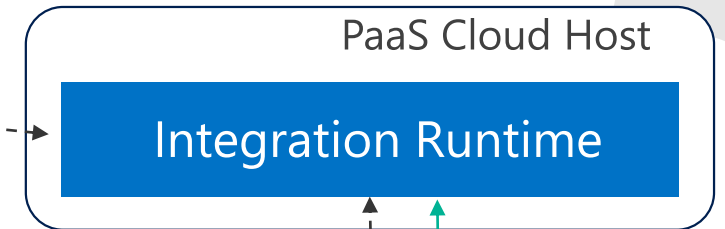
workday

←----→ Command & Control

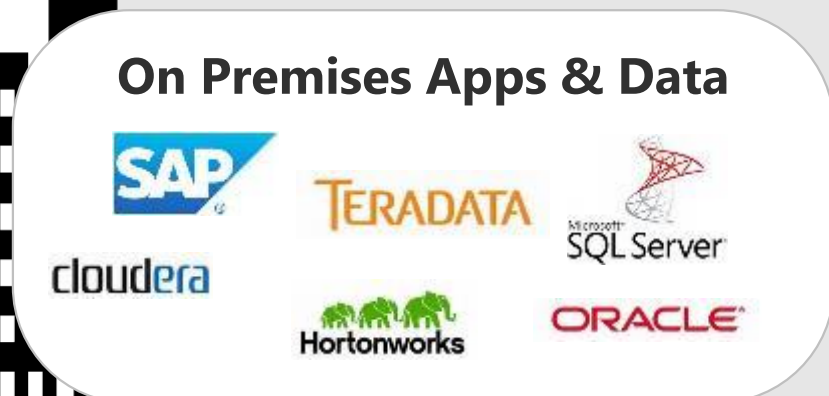
↔ Data Flow



## Azure Cloud



## On Premises Apps & Data

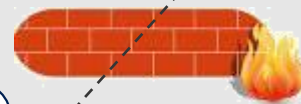
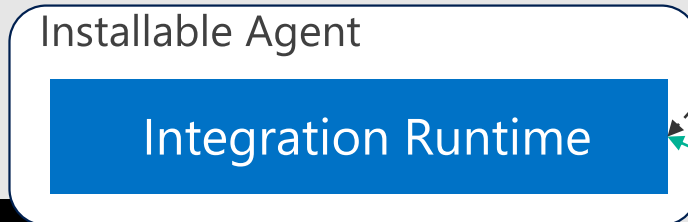
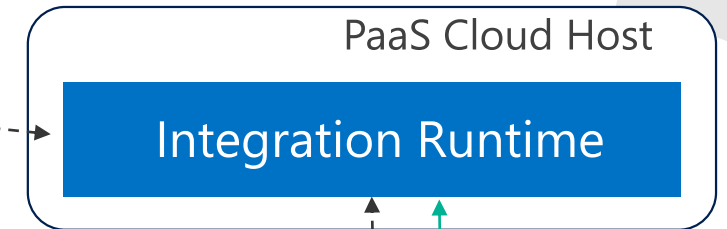


←----→ Command & Control

↔ Data Flow



**Azure Cloud**

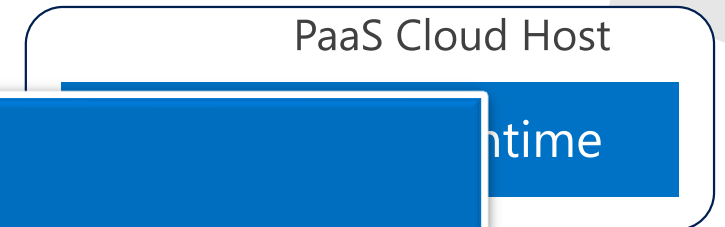


←-----> Command & Control

↔ Data Flow




**Azure Cloud**



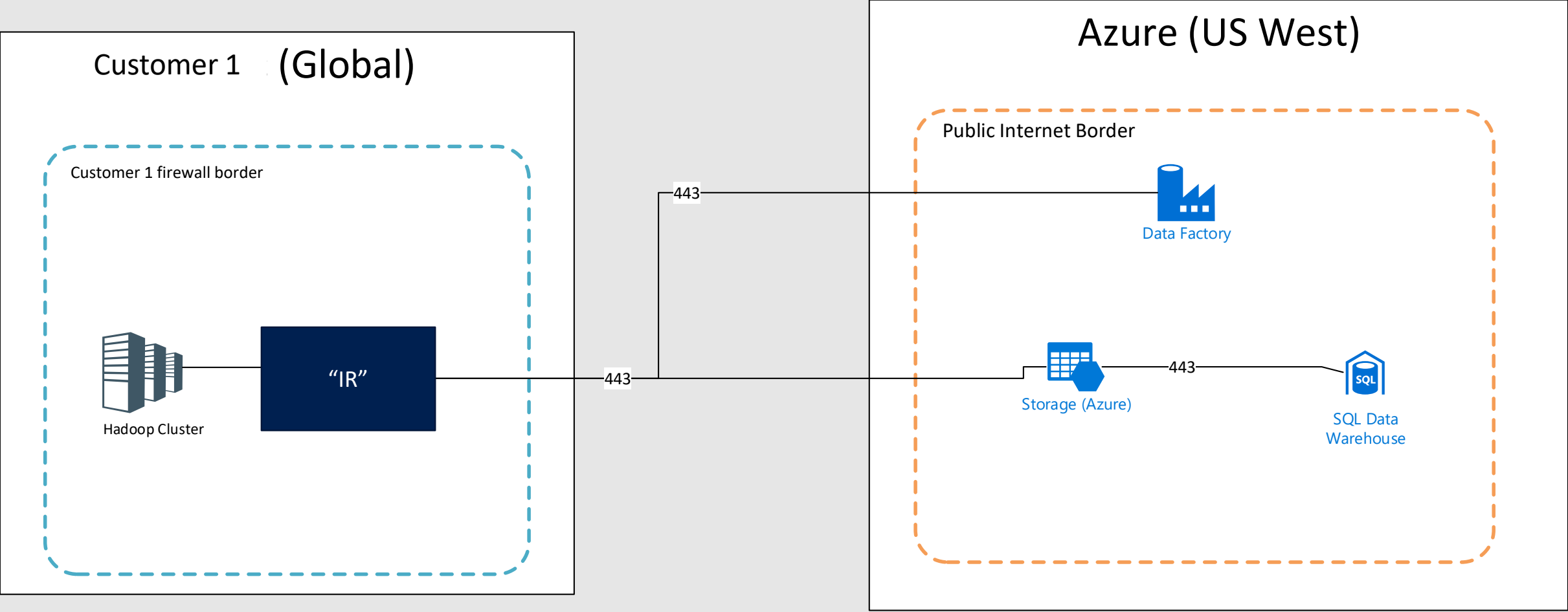
**Integration Runtime**

- Activity Dispatch/Monitor (spark, copy, ml, etc)
- Data Movement
- SSIS Package Execution

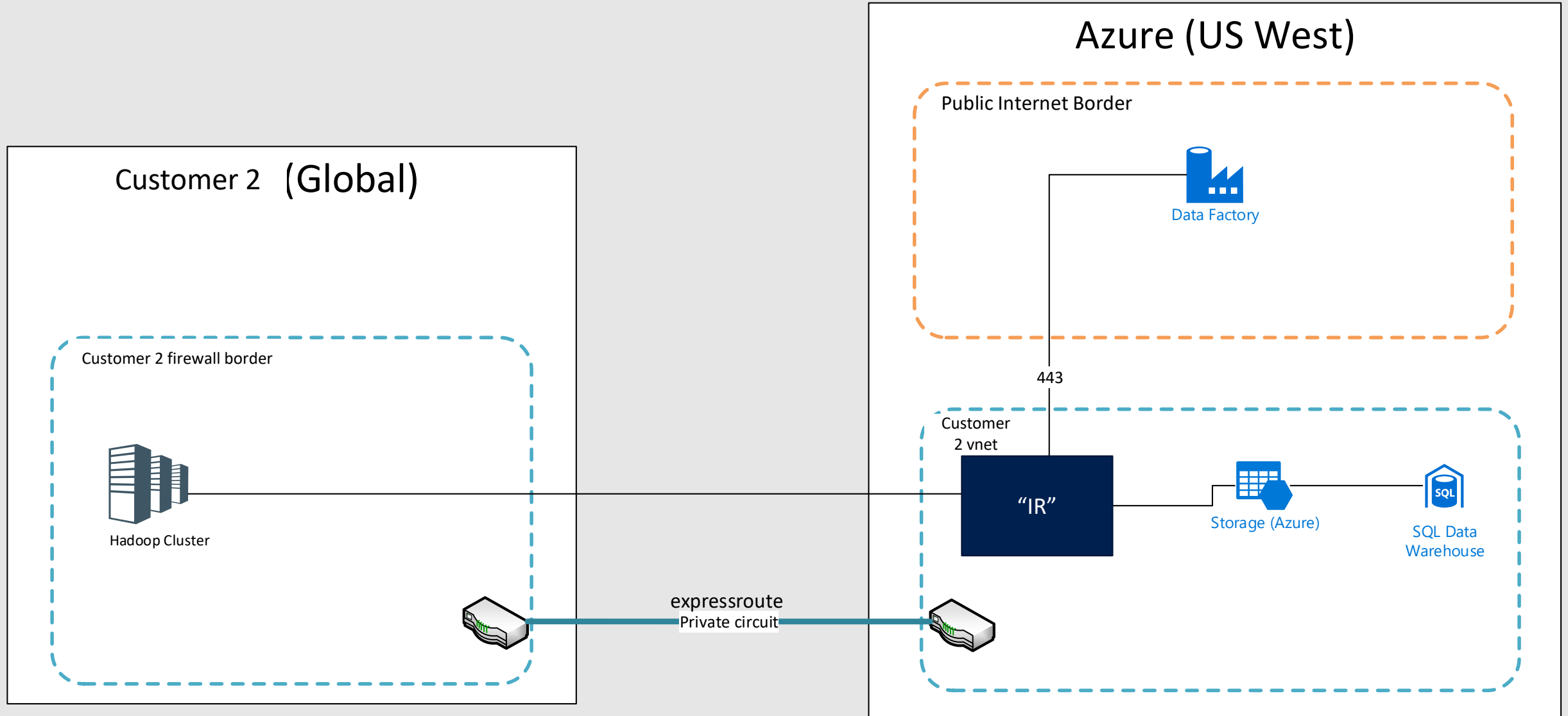
Insta



# Azure Data Factory "Integration Runtime" deployed on premises for transformation and then moved to cloud

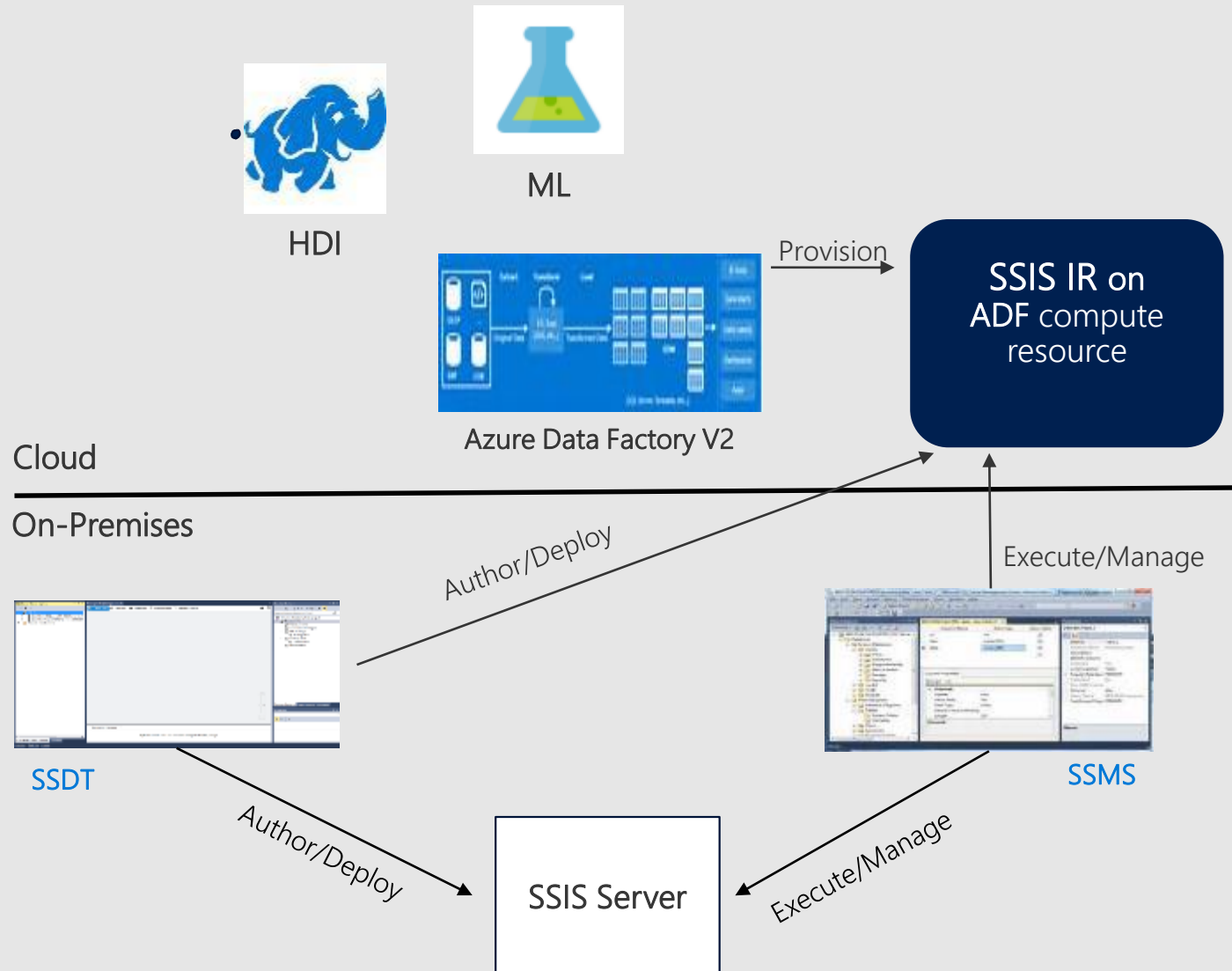


# Azure Data Factory "Integration Runtime" deployed inside VNet



SSIS in ADF

# SSIS as an Integration Runtime



- Provision SSIS IR via Data Factory
- Use SQL Server Data Tools (SSDT) to author/deploy SSIS packages
- Use SQL Server Management Studio (SSMS) to execute/manage SSIS packages
- Target SSIS customers who want to move all/part of their on-premises workloads and just "lift & shift" many existing packages to Azure
- Independent Software Vendors (ISVs) can build extensions/Software as a Service (SaaS) on SSIS Everest





## Traditional ETL



1. SSIS on Prem (to SQL Svr)



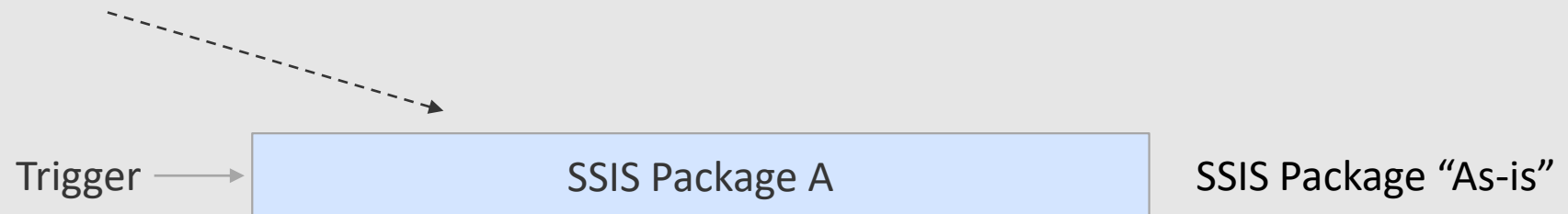
- *Lift & Shift w/ compatibility*

2. SSIS on IaaS (to SQL on IaaS | Az DB)



- *Want PaaS benefits (scale, no VM mgmt, etc)*
- *Mix reuse & modernization*

3. SSIS Runtime in ADF





## Traditional ETL



1. SSIS on Prem (to SQL Svr)



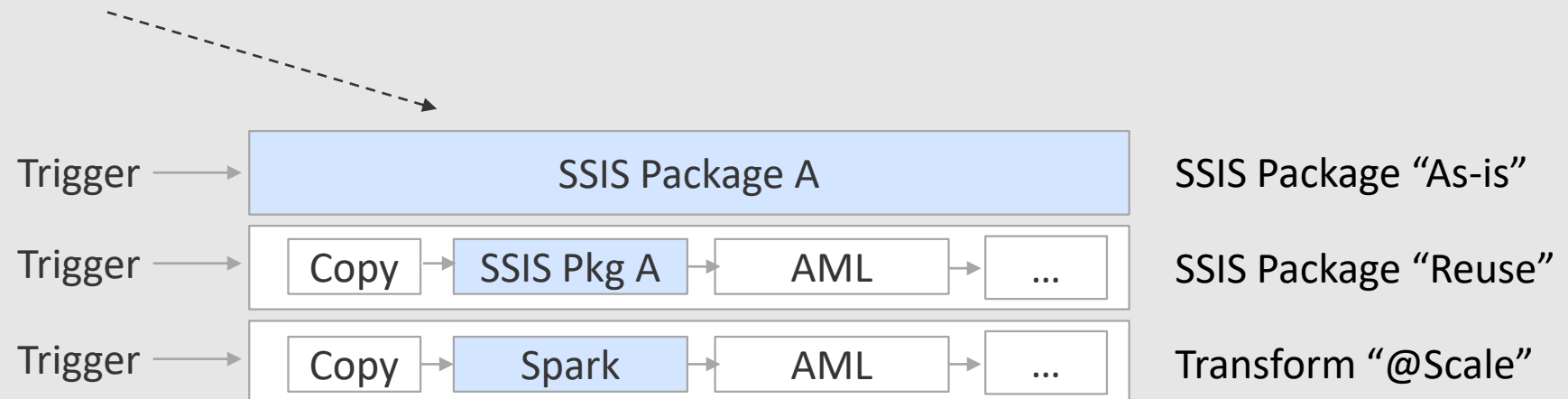
- *Lift & Shift w/ compatibility*

2. SSIS on IaaS (to SQL on IaaS | Az DB)



- *Want PaaS benefits (scale, no VM mgmt, etc)*
- *Mix reuse & modernization*

3. SSIS Runtime in ADF





### Traditional ETL



### Modern DW & Data Driven SaaS

1. SSIS on Prem (to SQL Svr)

- *Lift & Shift w/ compatibility*

2. SSIS on IaaS (to SQL on IaaS | Az DB)

- *Want PaaS benefits (scale, no VM mgmt, etc)*
- *Mix reuse & modernization*

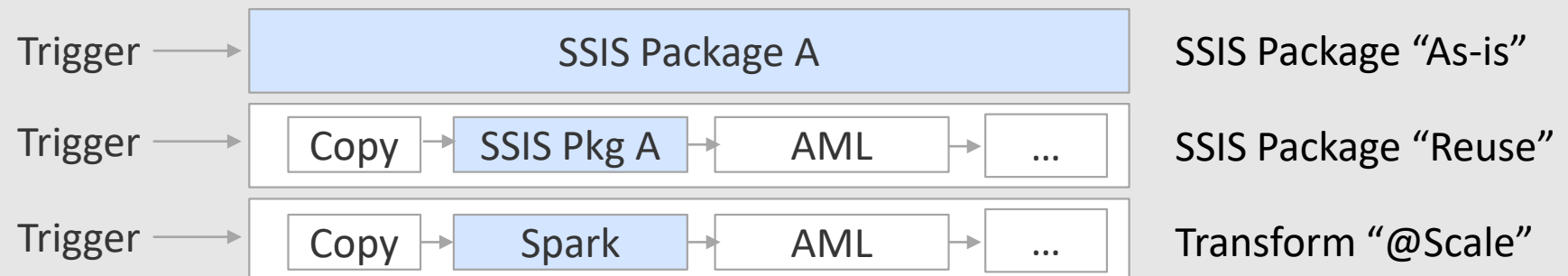
3. SSIS Runtime in ADF

1. ADF V1 (Time-series, Tumbling Window)  
(scenarios: log processing, etc.)

*Migrate v1 to v2:*

- *Much more flexible model*
- *Similar or cheaper in many scenarios*

2. ADF v2 (PaaS: Control Flow, Data Flow)



# ADF V2 Pricing (Preview Prices)

# Orchestration

Units: Activity Runs

Activity Run: Single run or re-run of an activity or trigger

Where the activity is orchestrated	Price
Azure Integration Runtime (Public Network environment)	0 – 50K runs: \$0.55/1k runs 50K+ runs: \$0.50/1k runs
Self-Hosted Integration Runtime (Private Network environment)	\$0.75/1K runs

# Data Movement

Units: DMU

DMU: Unit of measure used to copy data from source to sink

	Price
Azure Integration Runtime (Public Network environment)	\$0.125/hr
Self-Hosted Integration Runtime (Private Network environment)	\$0.05/hr

Inactive Pipelines (Pipelines not associated with a trigger with zero runs for a week) : \$0.20 / week

# Azure Integration Runtime for SSIS

VM	Price for Standard Edition \$/hr	Price for Enterprise Edition \$/hr
A4 v2 (4core)	\$0.420	\$0.956
A8 v2 (8core)	\$0.862	\$1.935
D1 v2 (1core)	\$0.296	\$0.832
D2 v2 (2core)	\$0.397	\$0.933
D3 v2 (4core)	\$0.599	\$1.136
D4 v2 (8core)	\$1.199	\$2.271

The ADF Azure-SSIS IR is a fully-managed service to execute your SSIS packages. You can specify the VM type and the number of nodes for your dedicated IR environment. Note that you have to pay for a SQL Azure DB instance separately in order to host the SSIS catalog in the cloud. If you are using the Azure Data Factory to move data outside Azure network, you are required to pay for the amount of the data egress.

Note that these VM prices are discounted rates from the list prices of SQL Server VMs.

Enterprise SKUs for SSIS will not be available during public preview.

# Scenario 1: Data movement across different data sources

Suppose you have a pipeline that has:

1. Copy Activity 1 moves data from On-Premises SQL Server to Azure Blob to be consumed later by a customized application
2. Copy Activity 2 moves data from Azure Blob to Azure SQL Database

Assume that the copy from On-Premises SQL Server to Azure Blob takes 2 hours. And the copy from Azure Blob to Azure SQL Database takes 1 hour with 2 DMUs. The pipeline is executed 30 times in a month.

*Copy from On-Premises SQL Server to Azure Blob:*

Data Movement: 30 activity runs X 2 hours duration for every run X \$0.05 = \$3

Activity Runs on Self-Hosted Integration Runtime (30 runs) + \$.75

Subtotal = \$3.75

*Copy from Azure Blob to Azure SQL Database:*

30 activity runs X 1 hour duration for every run X 2 DMUs x \$.125 = \$7.50

Activity Runs on Azure Integration Runtime (30 runs) + \$0.55

Subtotal = \$8.05

Total Price (per month) \$3.75 + \$8.05 = \$11.80

# Scenario 2: Orchestrating data transformation activities on compute services

Suppose you have a pipeline that has:

- 1. Spark Activity to run Spark application on an Azure HDInsight cluster in Azure Virtual Network
- 2. Azure Data Lake Analytics U-SQL Activity to execute U-SQL on Azure Data Lake Analytics

Assume the pipeline was executed for 30 times in one month.

Spark Activity Data Movement    \$0  
Activity Runs on Self-Hosted Integration Runtime (30 runs) +    \$.75

Subtotal = \$.75

Azure Data Lake Analytics U-SQL Activity Data Movement    \$0  
Activity Runs on Azure Integration Runtime (30 runs) + \$.55

Subtotal = \$.55

Total Price (per month)            \$1.30

# Scenario 3: Lift & Shift SSIS Packages

Suppose you have set up a SQL Server Standard and SSIS on a local 4-cores physical machine to run your daily ETL workload. Your ETL workload includes 100 SSIS packages and each of these packages takes about 5 mins to run. These packages are executed once daily and it requires about 8 hours each day to complete all the packages.

In order to lift and shift your SSIS package execution to Azure Data Factory, you need to provision the SSIS dedicated Integration Runtime via the Azure Data Factory portal

During the provision, you need to select the VM Type that match closely to the compute power you use on-premises. In this example, we will choose the A4 v2 type and 1 node to match the compute power on-premises. You are also required to create a SQL Azure Database instance for hosting the SSIS catalog on Azure.

We choose the S0 Standard for this example. SSIS Integration Runtime is a dedicated pool so the compute resource is dedicated only for you (not shared with other user). You can choose to keep the dedicated integration runtime pool 24/7 available or you can choose to spin it up and down

Integration Runtime Dedicated for 24/7	Cost
SSIS Integration Runtime – A4 v2	\$0.42 / hr
X 1 node	
X 24 hours / day (dedicated pool)	
X 31 days / month	
SSIS Integration Runtime 24/7 cost =	\$312.48 / month
SQL Azure for SSIS Catalog – S0	\$0.02 / hr
X 24 hours / day	
X 31 days / month	
SQL Azure for SSIS Catalog Cost =	\$15.03 / month
Total monthly cost to run SSIS ETL workload	\$327.51 / month

Integration Runtime Dedicated for the execution time only	Cost
SSIS Integration Runtime – A4 v2	\$0.42 / hr
X 1 node	
X 8 hours / day (dedicated pool)	
X 31 days / month	
SSIS Integration Runtime 24/7 cost =	\$104.16 / month
SQL Azure for SSIS Catalog – S0	\$0.0202 / hr
X 24 hours / day	
X 31 days / month	
SQL Azure for SSIS Catalog Cost =	\$15.03 / month
ADF to orchestrate the start / stop SSIS Integration runtime	
2 custom activities – start / stop IR	2 activity runs / day
X 31 days / month	
Total ADF Activity Cost	\$.55 / month for 1 – 1000 runs
Total monthly cost to run SSIS ETL workload	\$119.74 / month