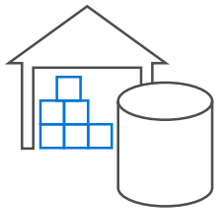
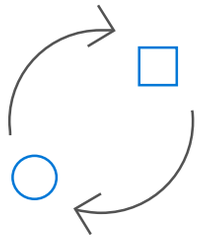
AZURE DATA FACTORY

AZURE DATA FACTORY

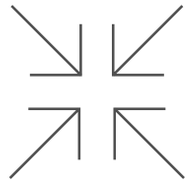
Fully-managed data integration service in the cloud



Flexible
Data integration

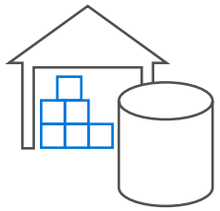


Hybrid
Data orchestration

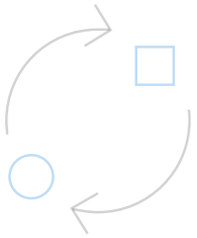


Data movement
As-a-service

Security and compliance



Flexible
Data integration



Hybrid
Data orchestration

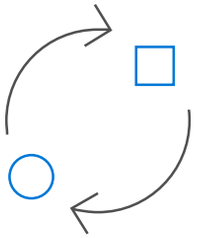
Data movement
As-a-service

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

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

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

Flexible
Data integration



Hybrid
Data orchestration

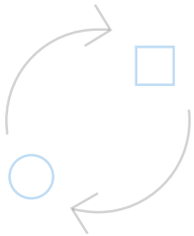
Data movement
As-a-service

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

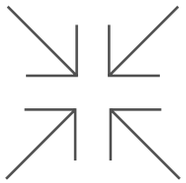
Meet your security and compliance needs while taking advantage of truly hybrid integration capabilities

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

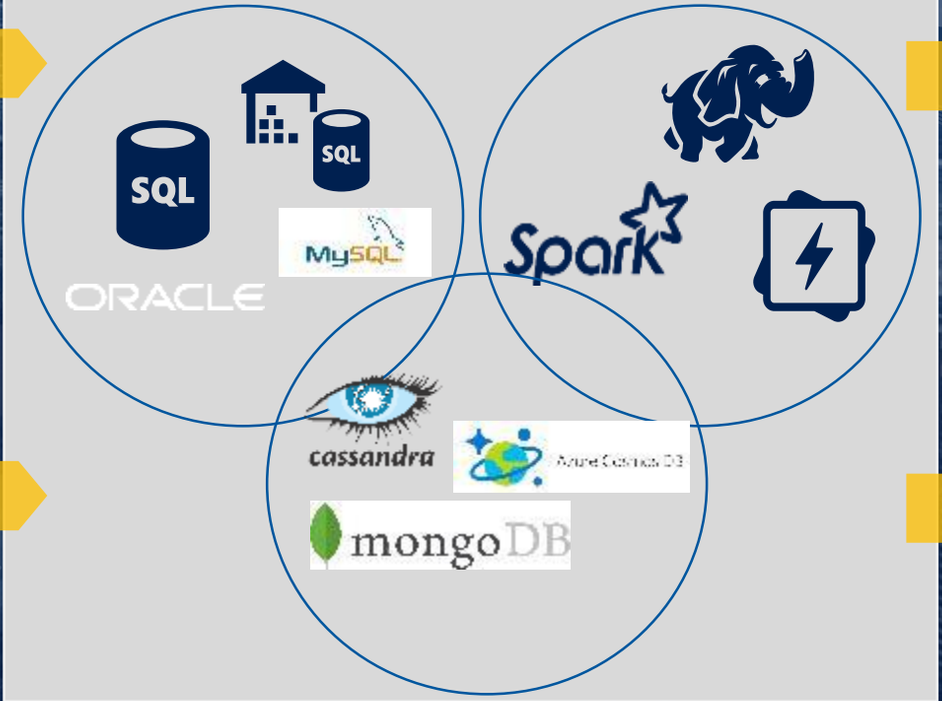
Relational data

OLTP | ERP | CRM | LOB

Non-relational data

Web | Media | Social media | Devices

Data Processing & Movement



Any BI tool

Dashboards | Reporting
Mobile BI | Cubes

Advanced Analytics

Machine Learning
Stream analytics Cognitive | AI

Any language

.NET | Java | R | Python
Ruby | PHP | Scala

CLOUD

V-NET

ON-PREMISE

AZURE DATA FACTORY ORCHESTRATES DATA PIPELINE ACTIVITY WORKFLOW & SCHEDULING

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

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.	{ "name":"MyForEachActivityName", "type":"ForEach", "typeProperties":{ "isSequential":"true", "items":"@pipeline().parameters.mySinkDatasetFolderPathCollection", "activities":[{ "name":"MyCopyActivity", "type":"Copy", "typeProperties": …
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 https://management.azure.com/subscriptions/<subId>/resourceGroups/<resourceGroupName>/providers/Microsoft.DataFactory/factories/<dataFactoryName>/pipelines/<pipelineName>/createRun?api-version=2017-03-01-preview
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	

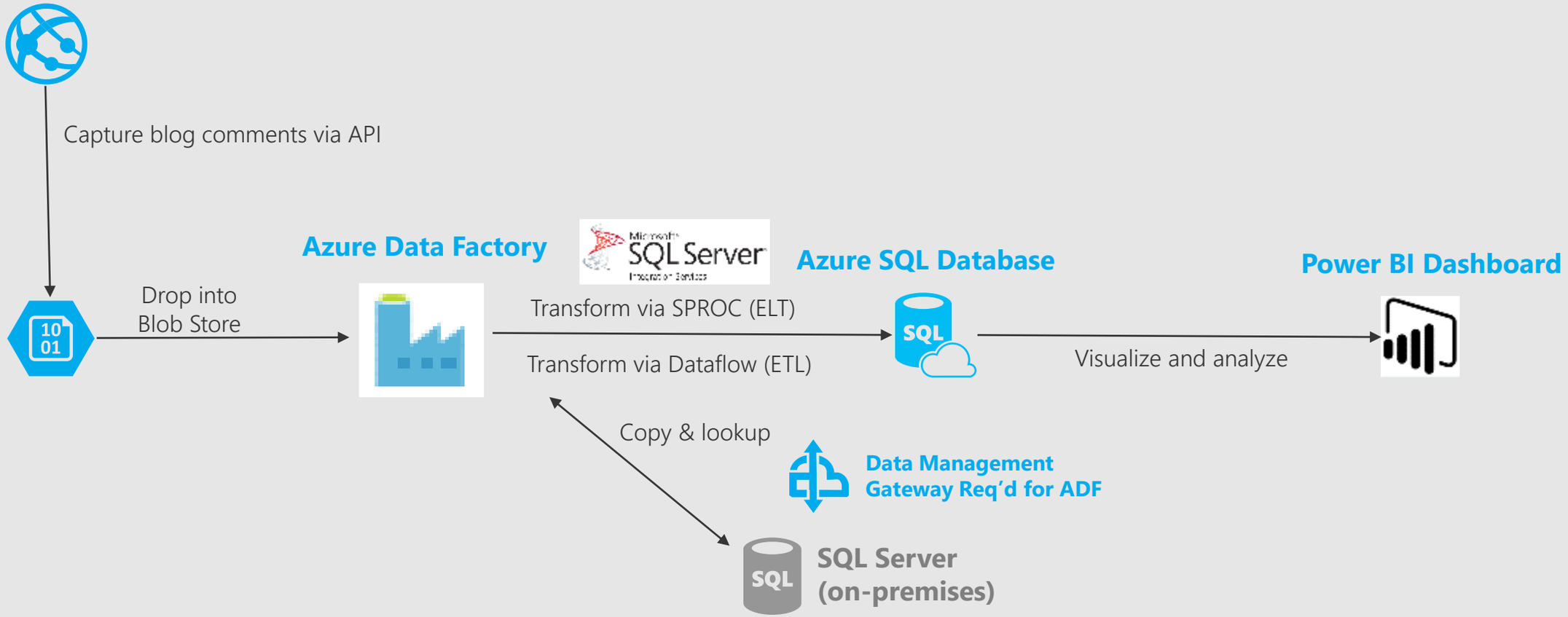
Scheduling	Flexible Scheduling
	Wall-clock scheduling
	Event-based triggers
		"type": "ScheduleTrigger",
		"typeProperties": {
		"recurrence": {
		"frequency": <<Minute, Hour, Day, Week, Year>>,
		"interval": <<int>>, // optional, how often to fire (default to 1)
		"startTime": <<datetime>>,
		"endTime": <<datetime>>,
		"timeZone": <<default UTC>>
		"schedule": { // optional (advanced scheduling specifics)
		"hours": [<<0-24>>],
		"weekDays": ": [<<Monday-Sunday>>],
		"minutes": [<<0-60>>],
		"monthDays": [<<1-31>>],
		"monthlyOccurences": [
		{
		"day": <<Monday-Sunday>>,
		"occurrence": <<1-5>>

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	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. A **Dataset** 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 A **Linked Service** 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.	Accessing parameters of other activities Using expressions @parameters("{name of parameter}") @activity("{Name of Activity}").output.RowsCopied
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	

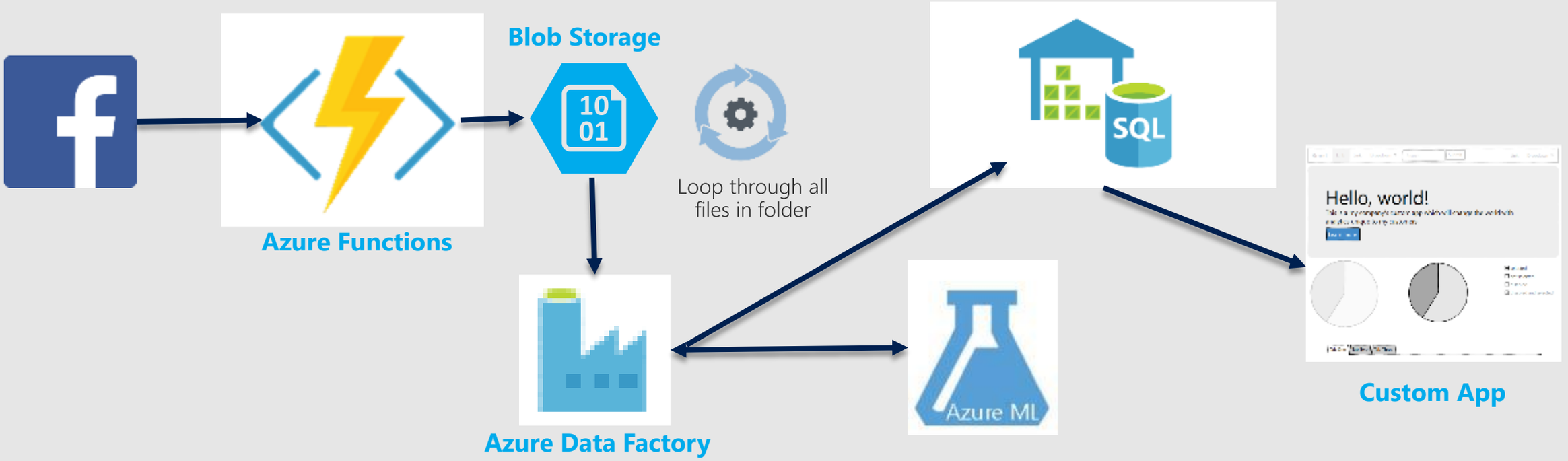
Patterns & Scenarios for ADF V2

Hybrid Data Integration Pattern 1:
Analyze blog comments

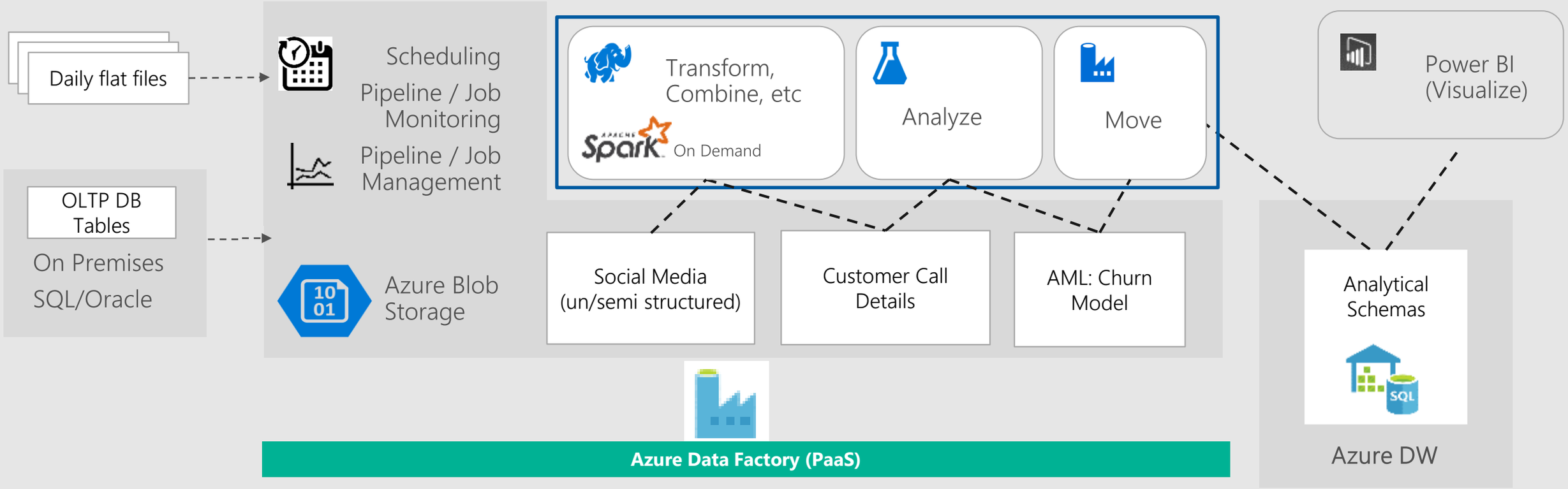


Hybrid Data Integration Pattern 2:

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

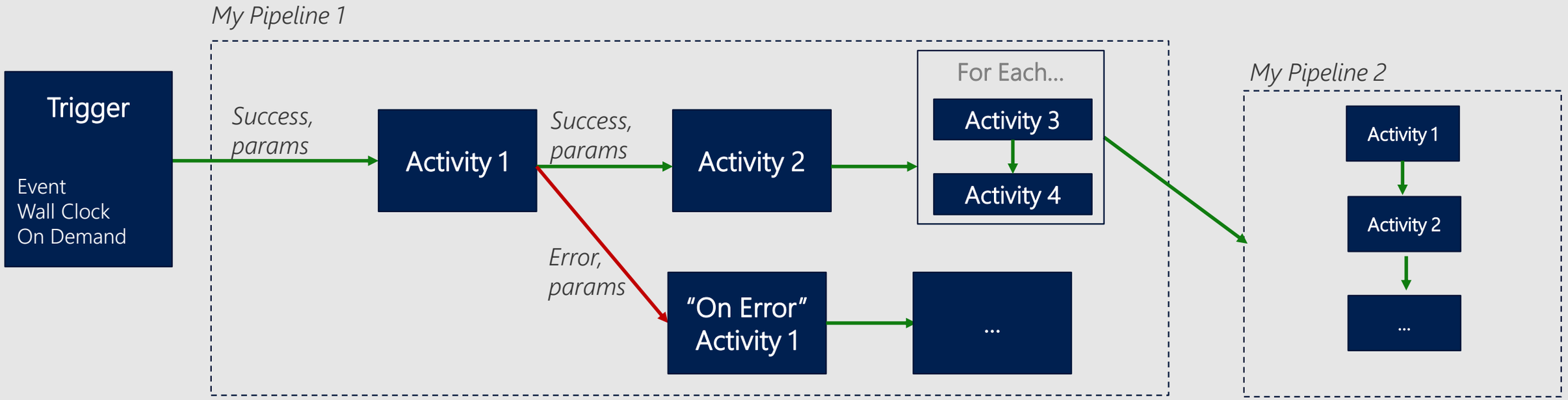


Hybrid Data Integration Pattern 3:
Modern Data Warehouse



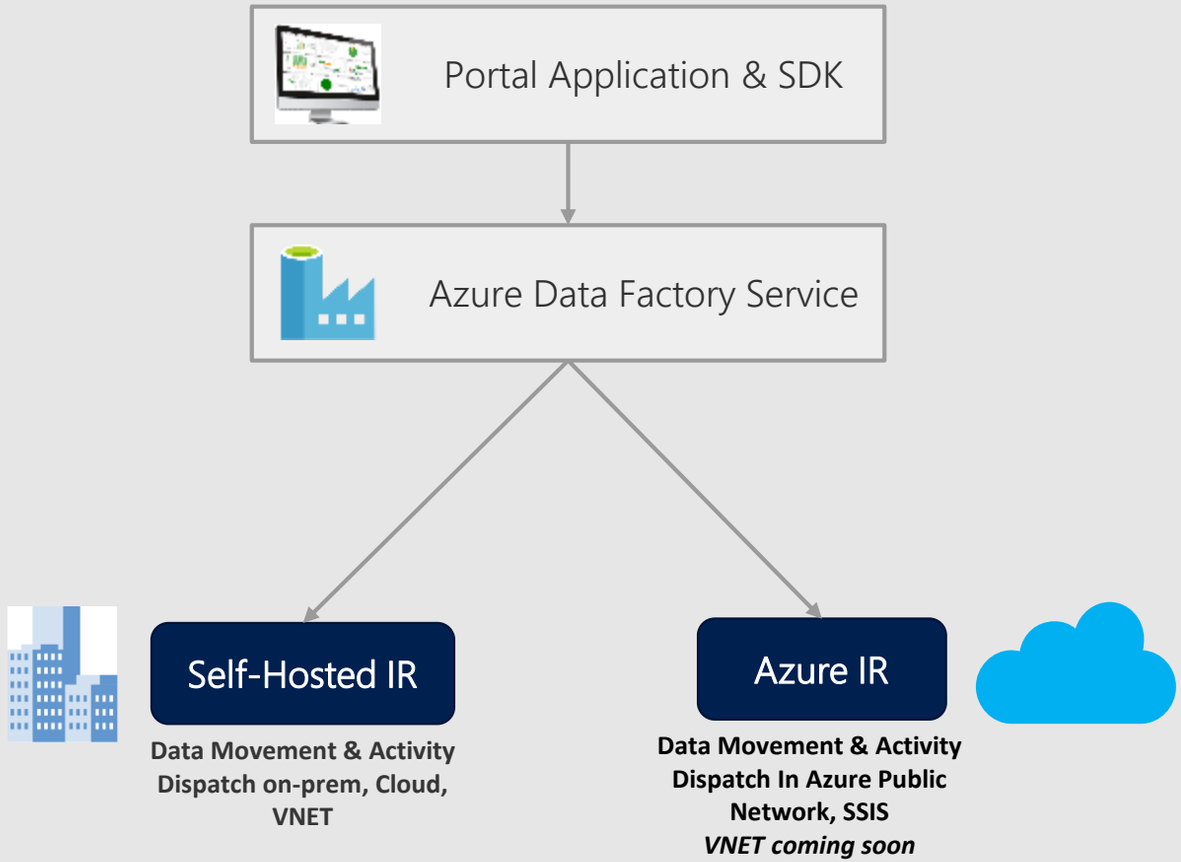
ADFv2: Control Flow

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)



- 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

SAP

TERADATA

Microsoft SQL Server

cloudera

Hortonworks

ORACLE

Cloud Apps, Svcs & Data

amazon web services

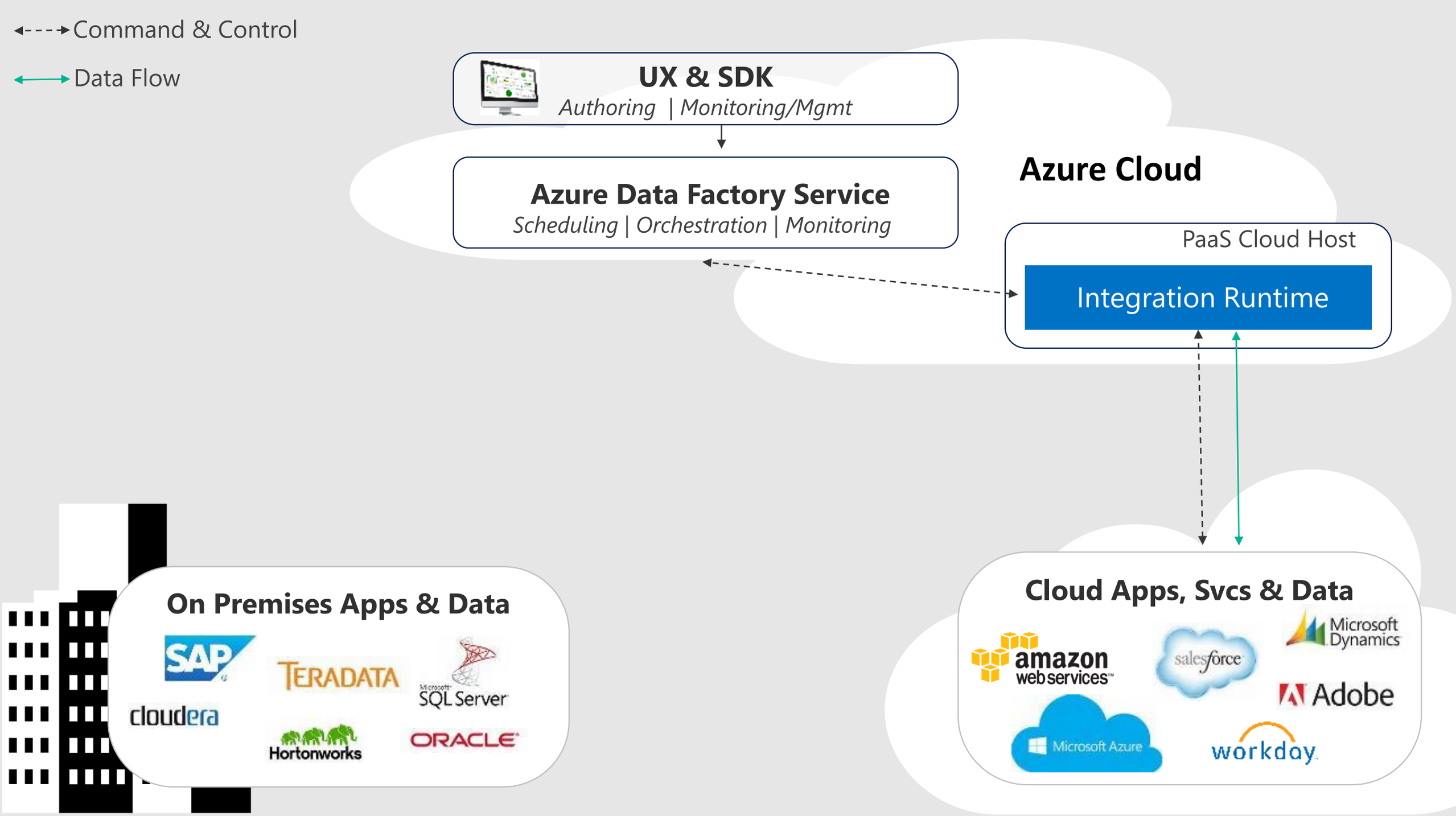
salesforce

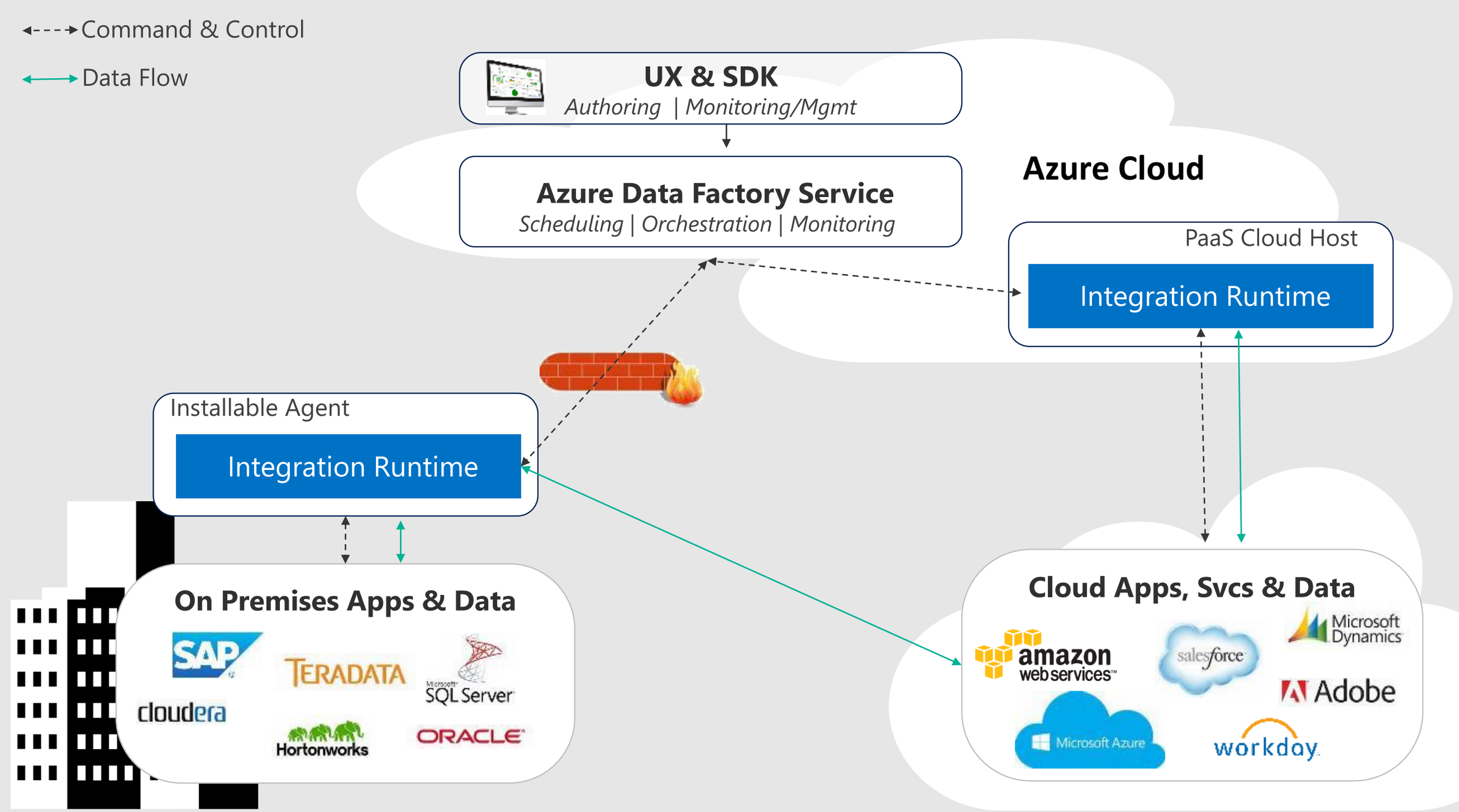
Microsoft Dynamics

Adobe

Microsoft Azure

workday.


Command & Control
Data Flow
UX & SDK
Authoring | Monitoring/Mgmt
Azure Data Factory Service
Scheduling | Orchestration | Monitoring
Azure Cloud
PaaS Cloud Host
Integration Runtime
On Premises Apps & Data
SAP
TERADATA
Microsoft SQL Server
cloudera
Hortonworks
ORACLE
Cloud Apps, Svcs & Data
amazon web services
salesforce
Microsoft Dynamics
Adobe
Microsoft Azure
workday.


Command & Control
Data Flow
UX & SDK
Authoring | Monitoring/Mgmt
Azure Data Factory Service
Scheduling | Orchestration | Monitoring
Azure Cloud
PaaS Cloud Host
Integration Runtime
Installable Agent
Integration Runtime
On Premises Apps & Data
SAP
TERADATA
Microsoft SQL Server
cloudera
Hortonworks
ORACLE
Cloud Apps, Svcs & Data
amazon web services
salesforce
Microsoft Dynamics
Adobe
Microsoft Azure
workday

Command & Control

Data Flow

UX & SDK

Authoring | Monitoring/Mgmt

Azure Data Factory Service

Scheduling | Orchestration | Monitoring

Azure Cloud

PaaS Cloud Host

ntime

Insta

Integration Runtime

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

On Premises Apps & Data

SAP

TERADATA

Microsoft SQL Server

cloudera

Hortonworks

ORACLE

Cloud Apps, Svcs & Data

amazon web services

salesforce

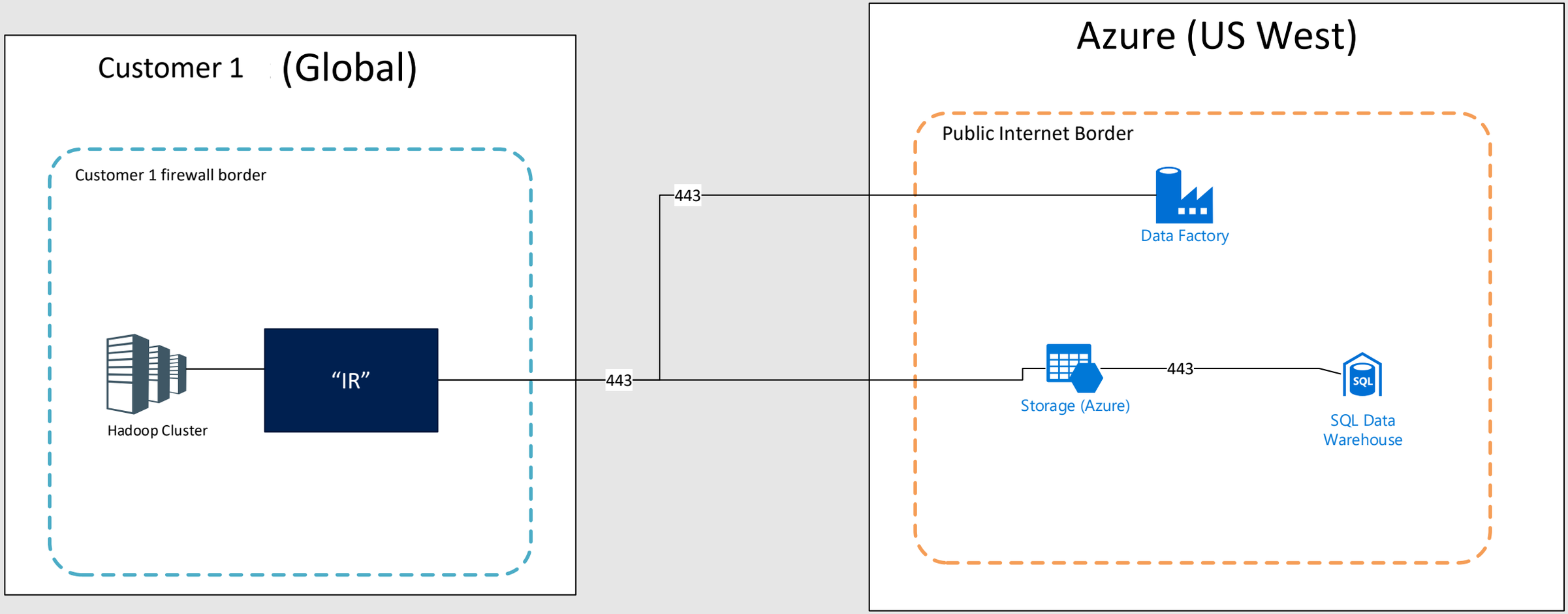
Microsoft Dynamics

Adobe

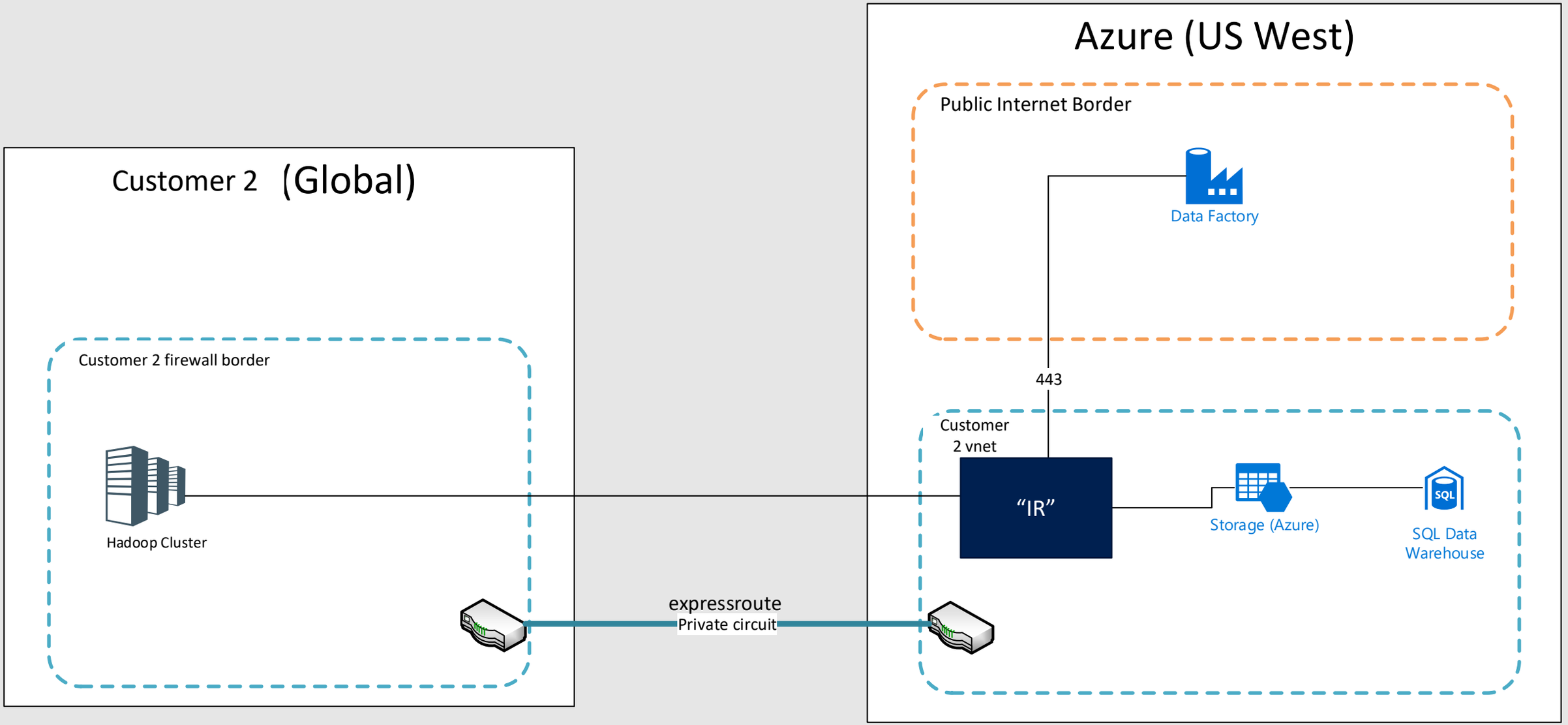
Microsoft Azure

workday

Azure Data Factory “Integration Runtime” deployed on premises for transformation and then moved to cloud

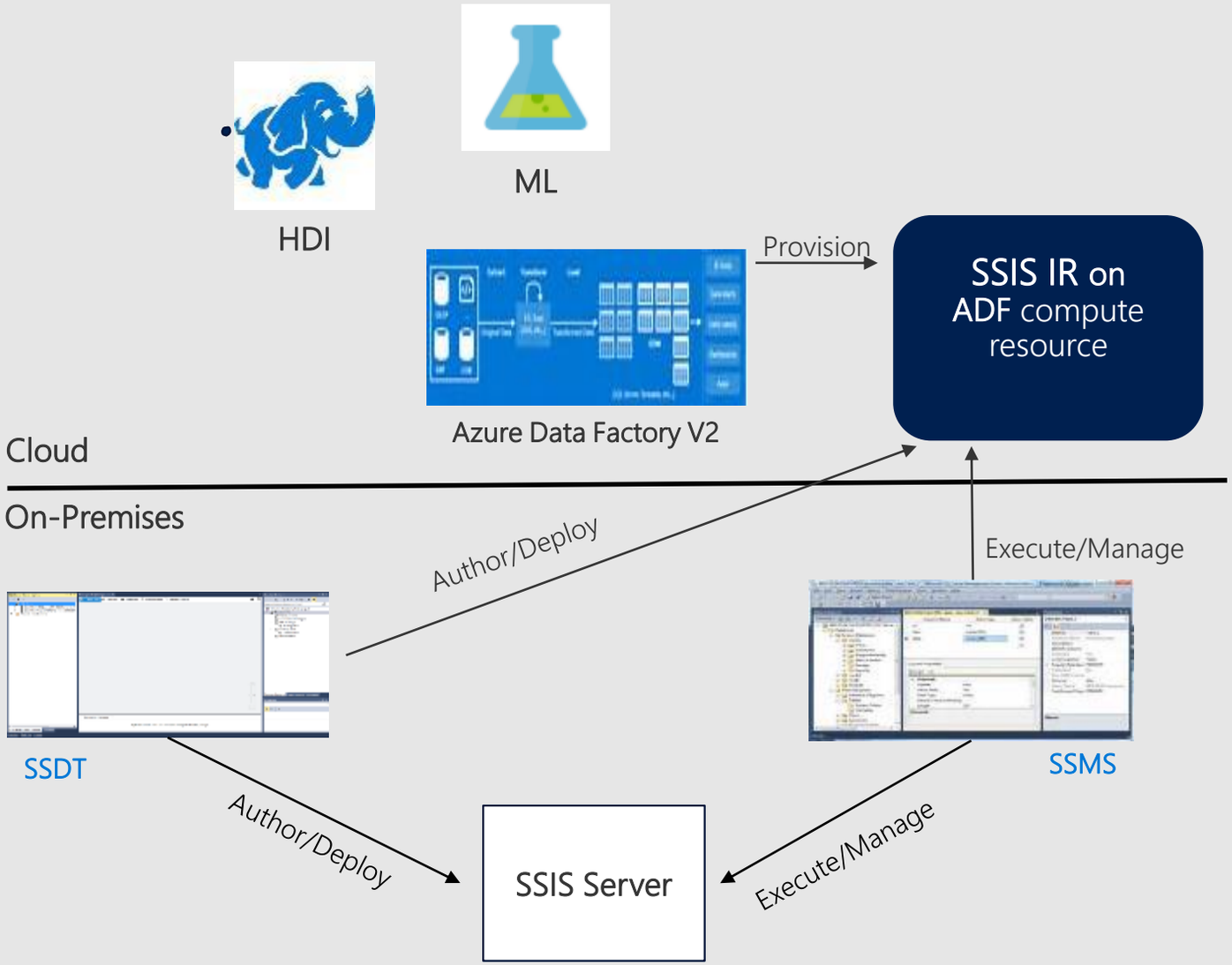


Azure Data Factory “Integration Runtime” deployed inside VNet



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

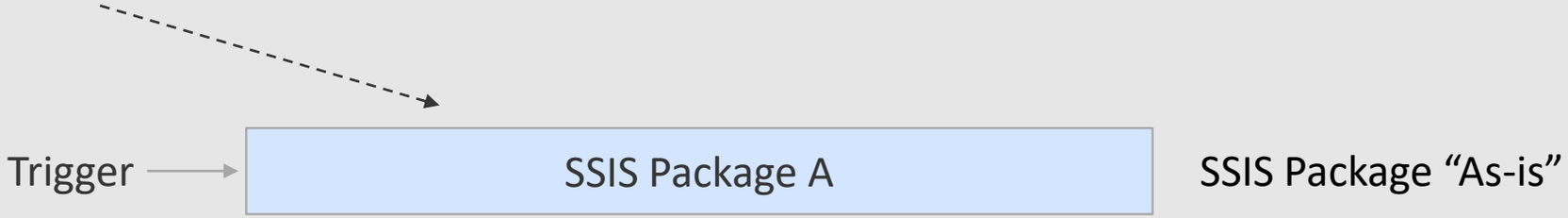
1. SSIS on Prem (to SQL Svr)

- *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

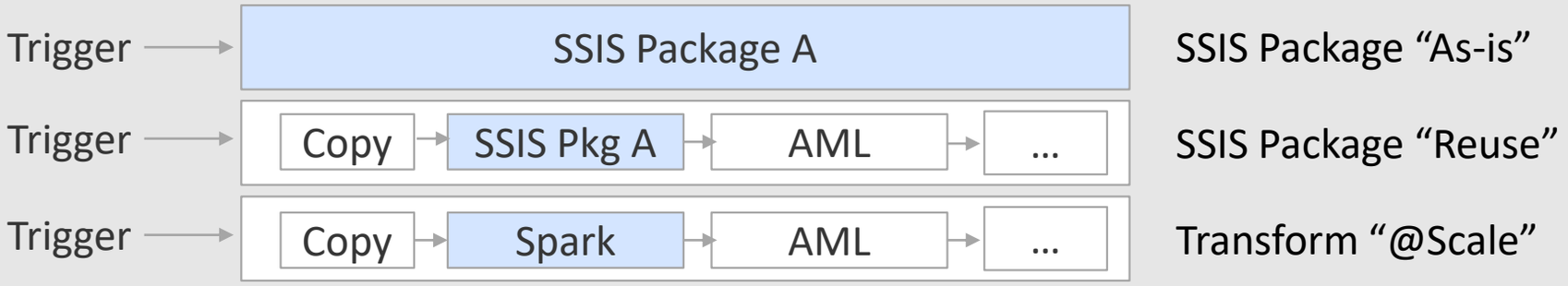
1. SSIS on Prem (to SQL Svr)

- *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

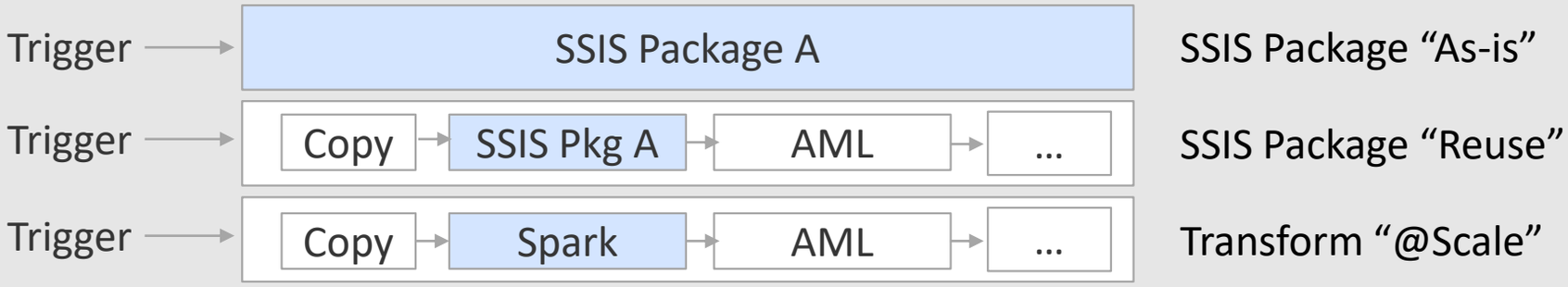


Traditional ETL

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**

Modern DW & Data Driven SaaS

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)**