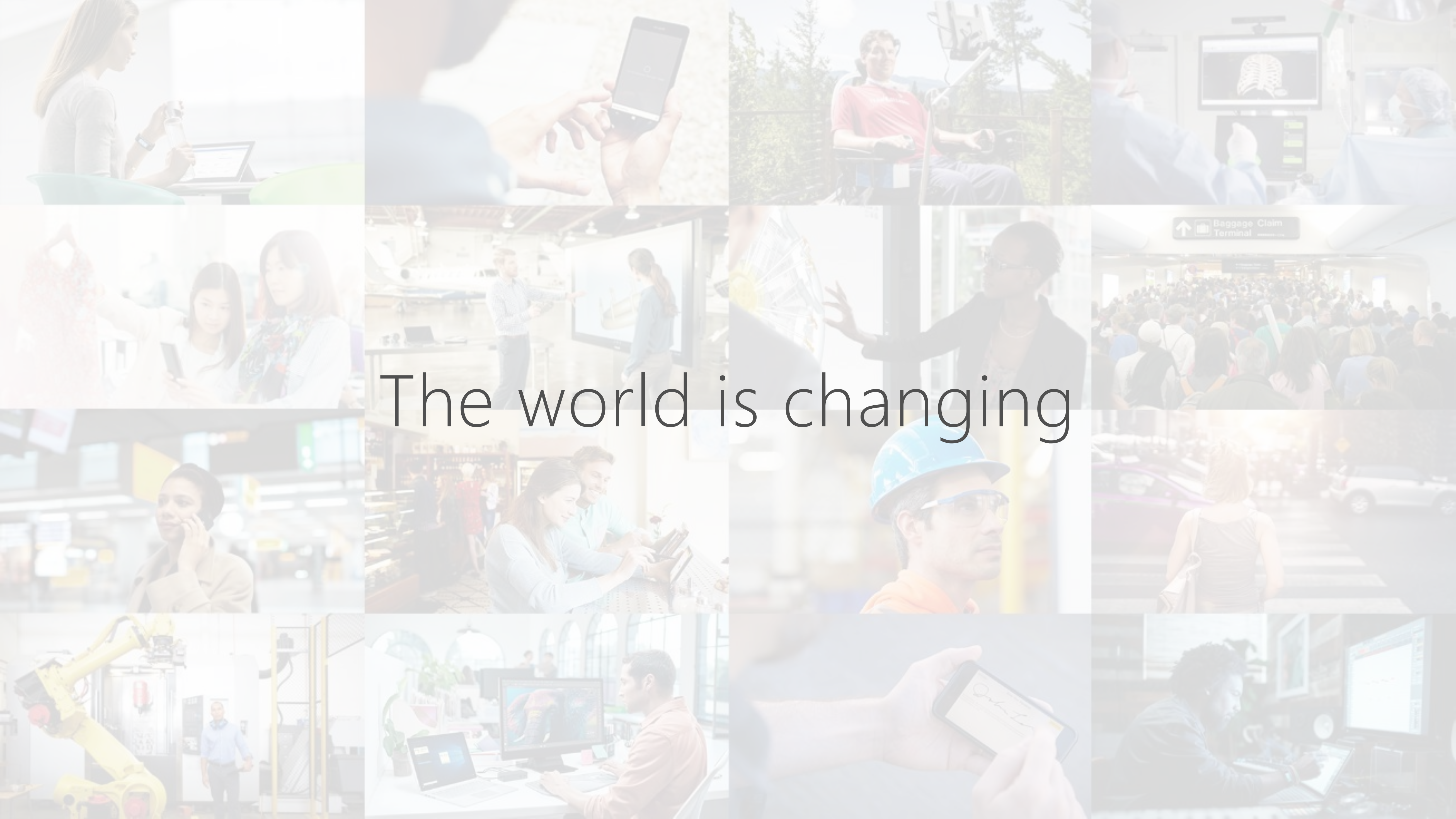
Microsoft

Azure Data Factory

Hybrid data integration, at global scale

Nakul Joshi
njoshi@microsoft.com
Cloud Solution Architect

The world is changing



Organizations that harness data, cloud, and AI outperform

Data will grow to **44 ZB in 2020**

Today, 80% of organizations adopt cloud-first strategies

AI investment increased by **300% in 2017**

There are barriers to getting value from data

⚠ Data silos

⚠ Incongruent data types

⚠ Complexity of solutions

⚠ Multi cloud environment

⚠ Rising costs

Derive real value from your data

Data silos	Incongruent data types	Performance constraints	Complexity of solutions	Rising costs
✓	✓	✓	✓	✓
One hub for all data	Support for diverse types of data	Unlimited data scale	Familiar tools and ecosystem	Lower TCO

On-premises, hybrid, Azure

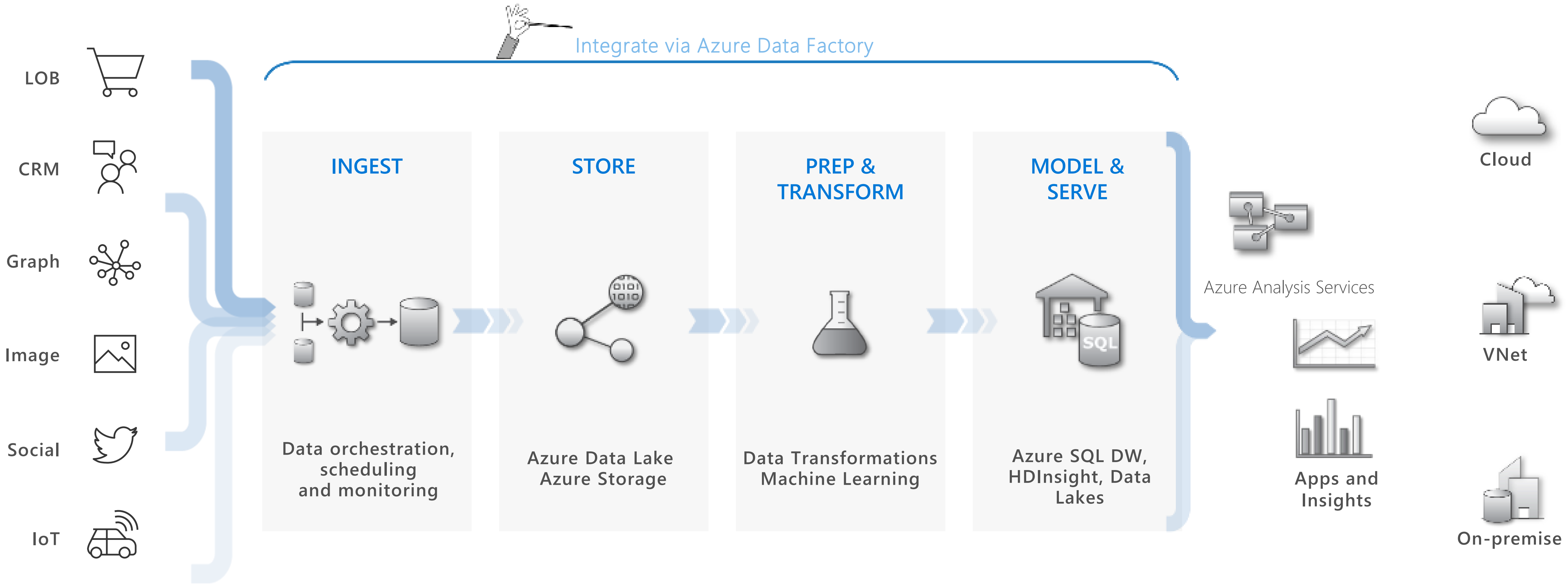
Organizations that harness data, cloud, and AI outperform

Nearly double operating margin

$100M in additional operating income

AZURE DATA FACTORY

Modernize your enterprise data warehouse at scale



Integrate via Azure Data Factory

LOB

CRM

Graph

Image

Social

IoT

INGEST

Data orchestration, scheduling and monitoring

STORE

Azure Data Lake
Azure Storage

PREP & TRANSFORM

Data Transformations
Machine Learning

MODEL & SERVE

Azure SQL DW, HDInsight, Data Lakes

Azure Analysis Services

Apps and Insights

Cloud

VNet

On-premise

AZURE DATA FACTORY

A fully-managed data integration service in the cloud

PRODUCTIVE

- ✓ Drag & Drop UI
- ✓ Codeless Data Movement

HYBRID

- ✓ Orchestrate where your data lives
- ✓ Lift SSIS packages to Azure

SCALABLE

- ✓ Serverless scalability with no infrastructure to manage

TRUSTED

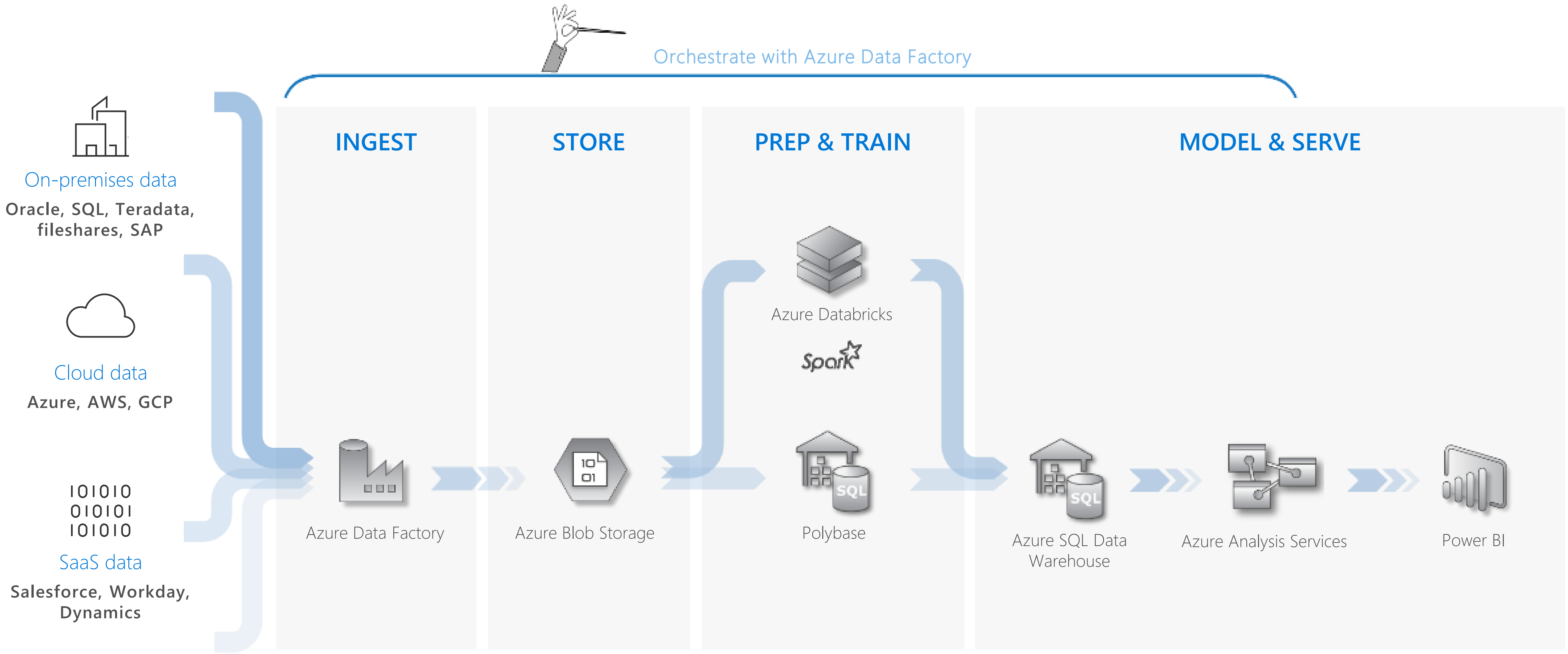
- ✓ Certified compliant Data Movement

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
- Lift and Shift SSIS to the Cloud
  - Execute SSIS packages in ADF Integration Runtime

AZURE DATA FACTORY

Modernize your enterprise data warehouse at scale



Orchestrate with Azure Data Factory

On-premises data

Oracle, SQL, Teradata, fileshares, SAP

Cloud data

Azure, AWS, GCP

SaaS data

Salesforce, Workday, Dynamics

INGEST

Azure Data Factory

STORE

Azure Blob Storage

PREP & TRAIN

Azure Databricks

Spark

Polybase

MODEL & SERVE

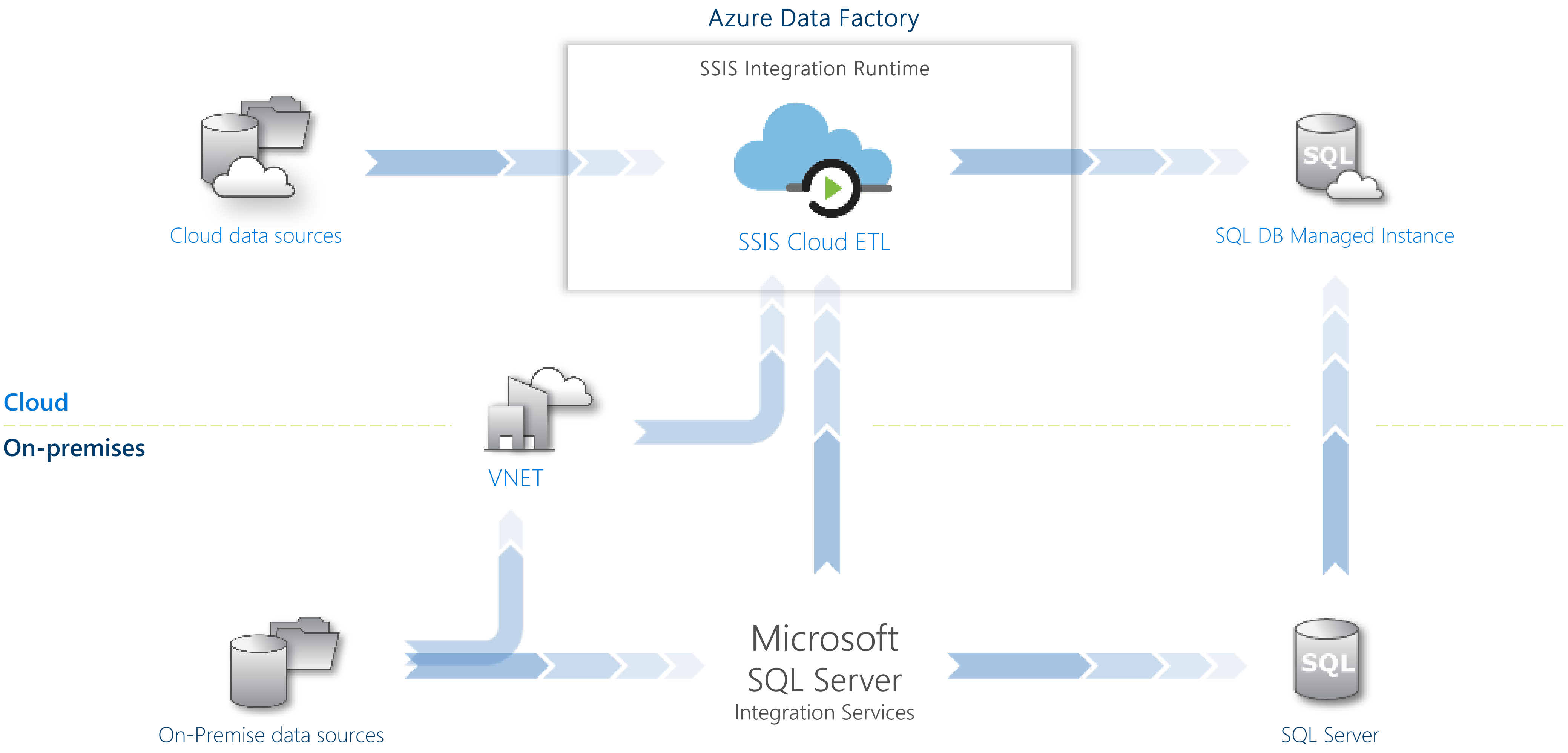
Azure SQL Data Warehouse

Azure Analysis Services

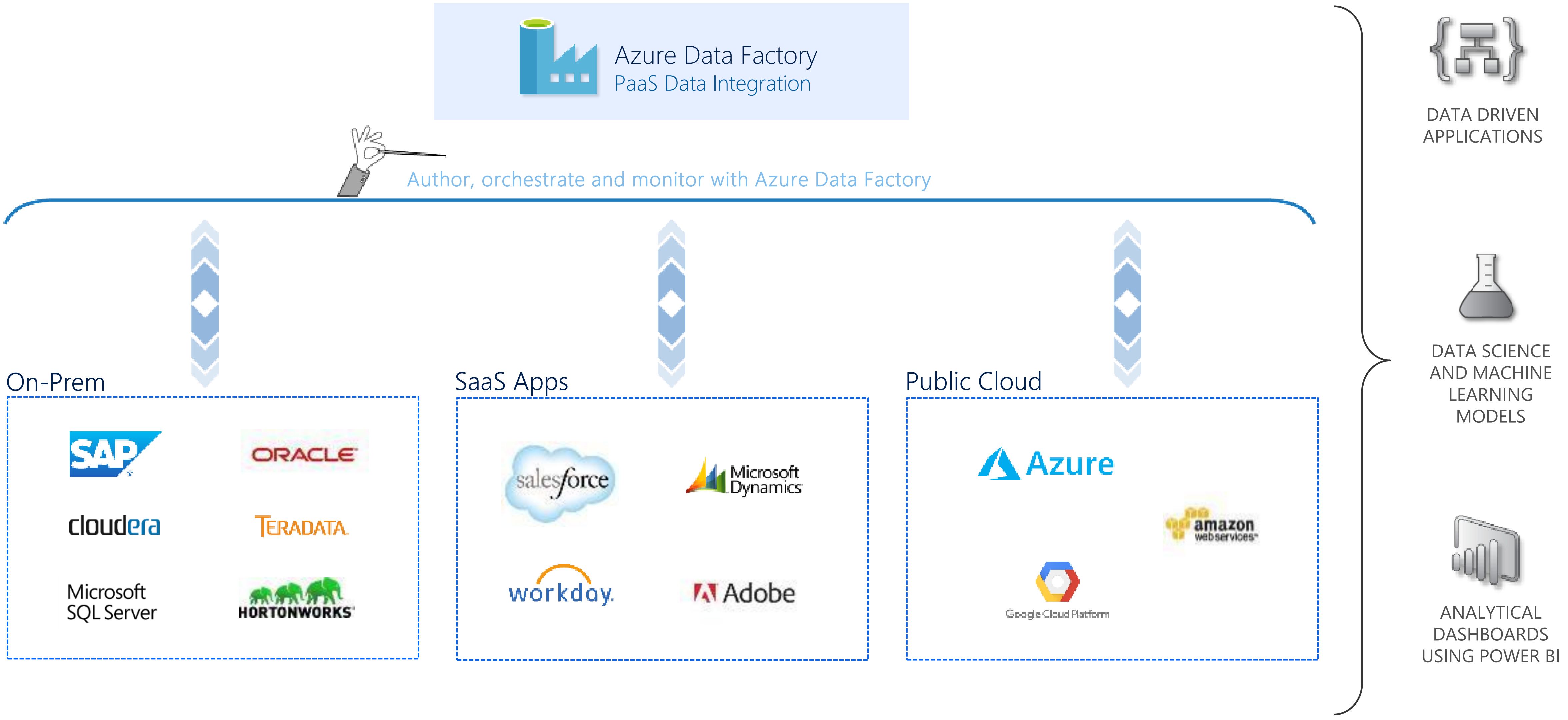
Power BI

Microsoft Azure also supports other **Big Data** services like **Azure HDInsight**, **Azure SQL Database** and **Azure Data Lake** to allow customers to tailor the above architecture to meet their unique needs.

Lift your SQL Server Integration Services (SSIS) packages to Azure



Hybrid and Multi-Cloud Data Integration



Azure Data Factory
PaaS Data Integration

Author, orchestrate and monitor with Azure Data Factory

On-Prem

SaaS Apps

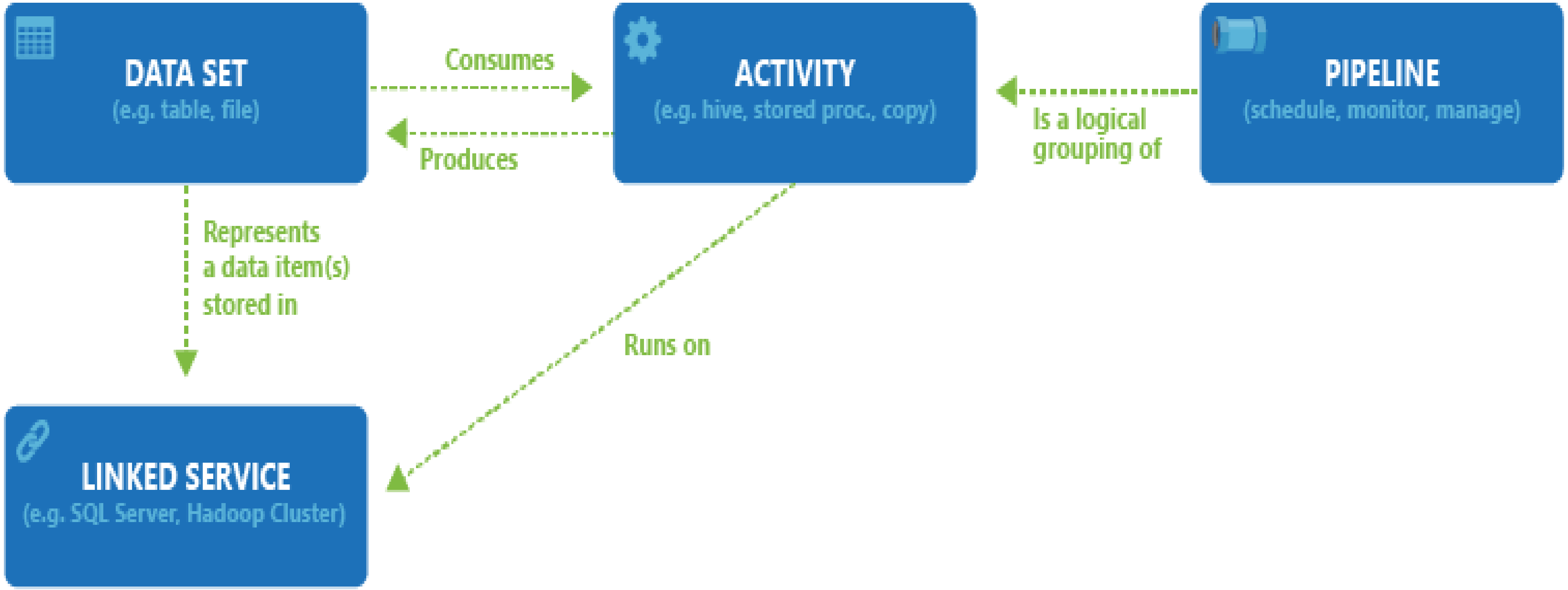
Public Cloud

DATA DRIVEN APPLICATIONS

DATA SCIENCE AND MACHINE LEARNING MODELS

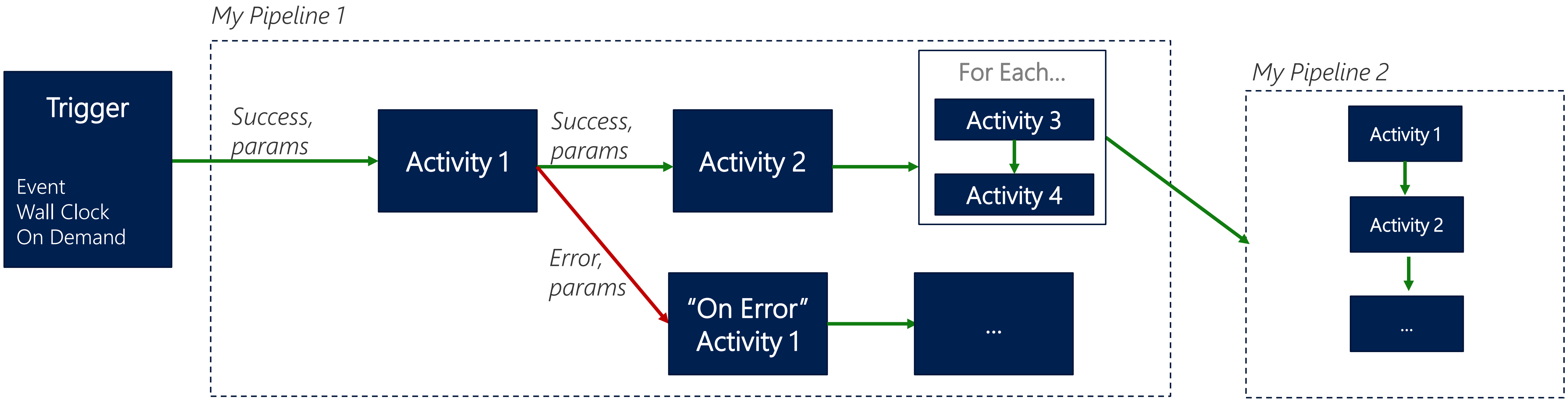
ANALYTICAL DASHBOARDS USING POWER BI

Concepts – Pipelines/Activities/Dataset/Linked Service


DATA SET
(e.g. table, file)
Consumes
ACTIVITY
(e.g. hive, stored proc., copy)
Produces
Is a logical grouping of
PIPELINE
(schedule, monitor, manage)
Represents a data item(s) stored in
Runs on
LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Control Flow Introduced in Azure Data Factory

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



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	Set-AzureRmDataFactoryV2IntegrationRuntime -Name $integrationRuntimeName -Type SelfHosted

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	name": "LookupWaterMarkActivity",
		"type": "Lookup",
		"typeProperties": {
		"source": {
		"type": "SqlSource",
		"sqlReaderQuery": "select * from watermarktable"
		}

On-Demand Spark	Support for on-demand HDI Spark clusters, similar to on-demand Hadoop activities in V1	"type": "HDInsightOnDemand", "typeProperties": { "clusterSize": 2, "clusterType": "spark", "timeToLive": "00:15:00",
SSIS Runtime	Lift & shift, deploy, manage, monitor SSIS packages in the cloud with SSIS Azure IR Service in Azure Data Factory	Start-AzureRmDataFactoryV2IntegrationRuntime -DataFactoryName $DataFactoryName -Name
Code-free UI	Build end-to-end data pipeline solutions for ADF without writing code or JSON	

Access all your data

- 70+ connectors & growing
- Azure IR available in 20 regions
- Hybrid connectivity using self-hosted IR: on-prem & VNet

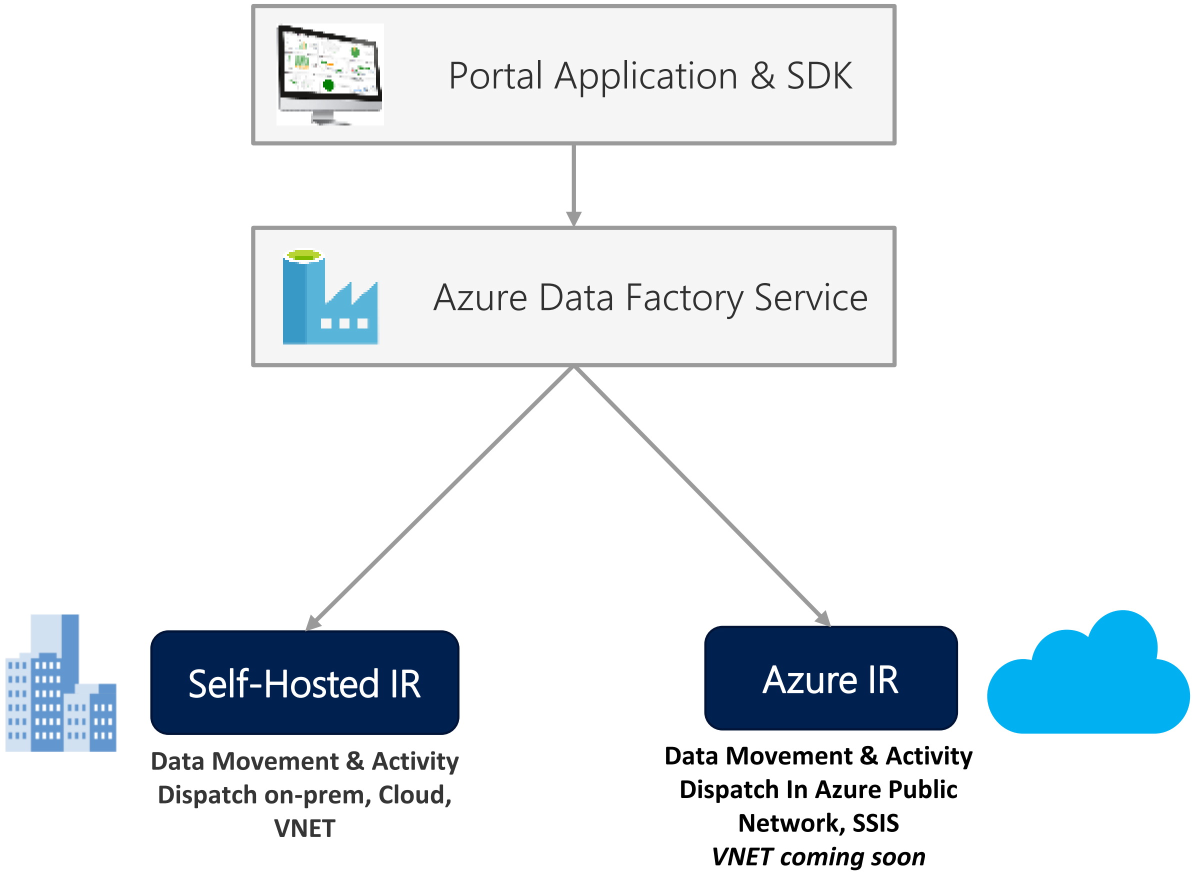
Azure	Database		File Storage	NoSQL	Services and Apps		Generic
Azure Blob Storage	Amazon Redshift	SQL Server	Amazon S3	Couchbase	Dynamics 365	Salesforce	HTTP
Azure Data Lake Store	Oracle	MySQL	File System	Cassandra	Dynamics CRM	Salesforce Service Cloud	OData
Azure SQL DB	Netezza	PostgreSQL	FTP	MongoDB	SAP C4C	ServiceNow	ODBC
Azure SQL DW	SAP BW	SAP HANA	SFTP		Oracle CRM	Hubspot	
Azure Cosmos DB	Google BigQuery	Informix	HDFS		Oracle Service Cloud	Marketo	
Azure DB for MySQL	Sybase	DB2			SAP ECC	Oracle Responsys	
Azure DB for PostgreSQL	Greenplum	MariaDB			Zendesk	Oracle Eloqua	
Azure Search	Microsoft Access	Drill			Zoho CRM	Salesforce ExactTarget	
Azure Table Storage	Hive	Phoenix			Amazon Marketplace	Atlassian Jira	
Azure File Storage	Hbase	Presto			Megento	Concur	
	Impala	Spark			PayPal	QuickBooks Online	
	Vertica				Shopify	Xero	
					GE Historian	Square	

* Supported file formats: CSV, AVRO, ORC, Parquet, JSON

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
- Built-in Source Control Support
- On-Demand Spark support
- Transform data in Azure Databricks
- Flexible pipeline scheduling with wall-clock, tumbling windows 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
- Graphical UI pipeline builder for a code-free experience

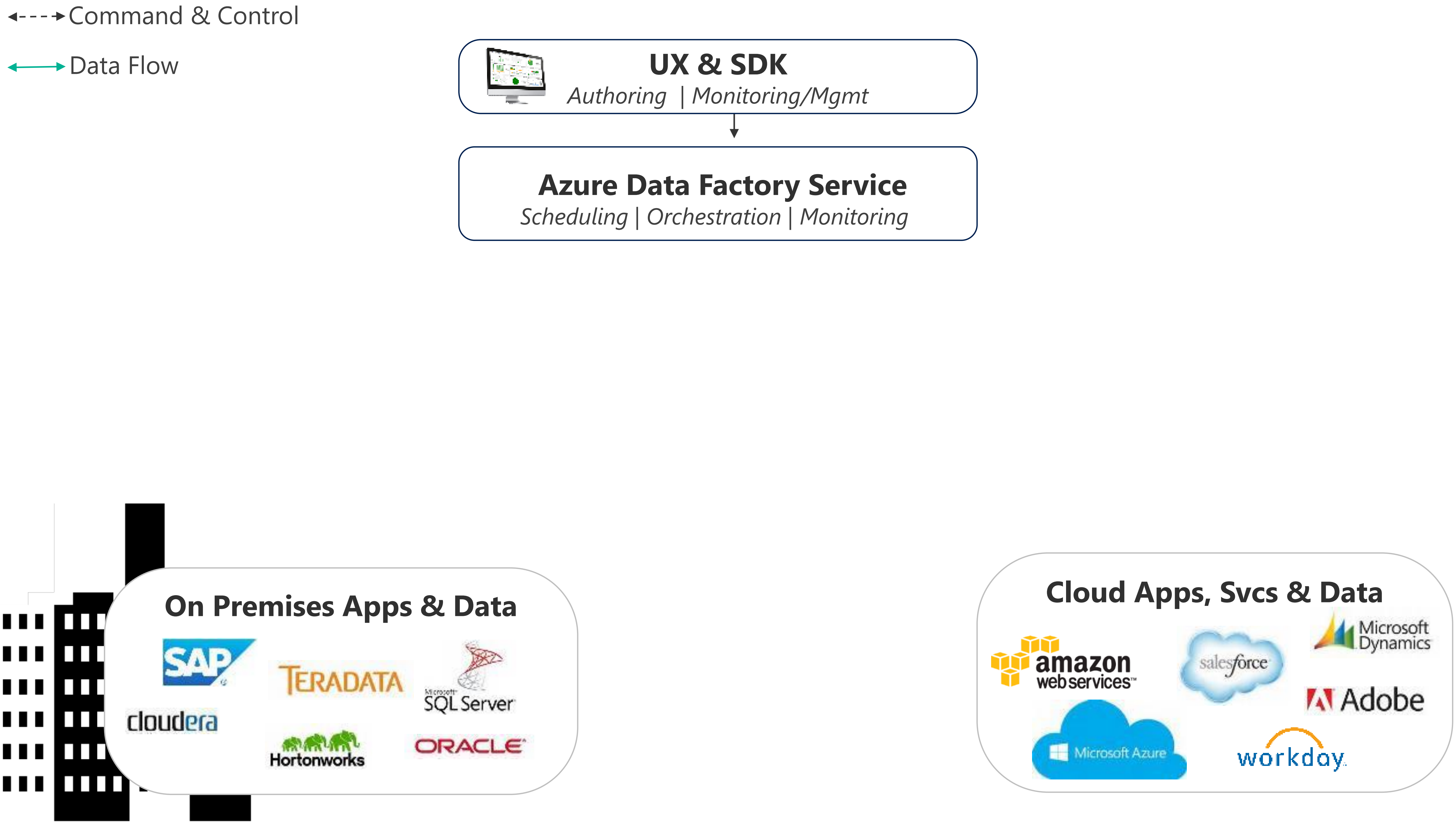
ADF Integration Runtime (IR)



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

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

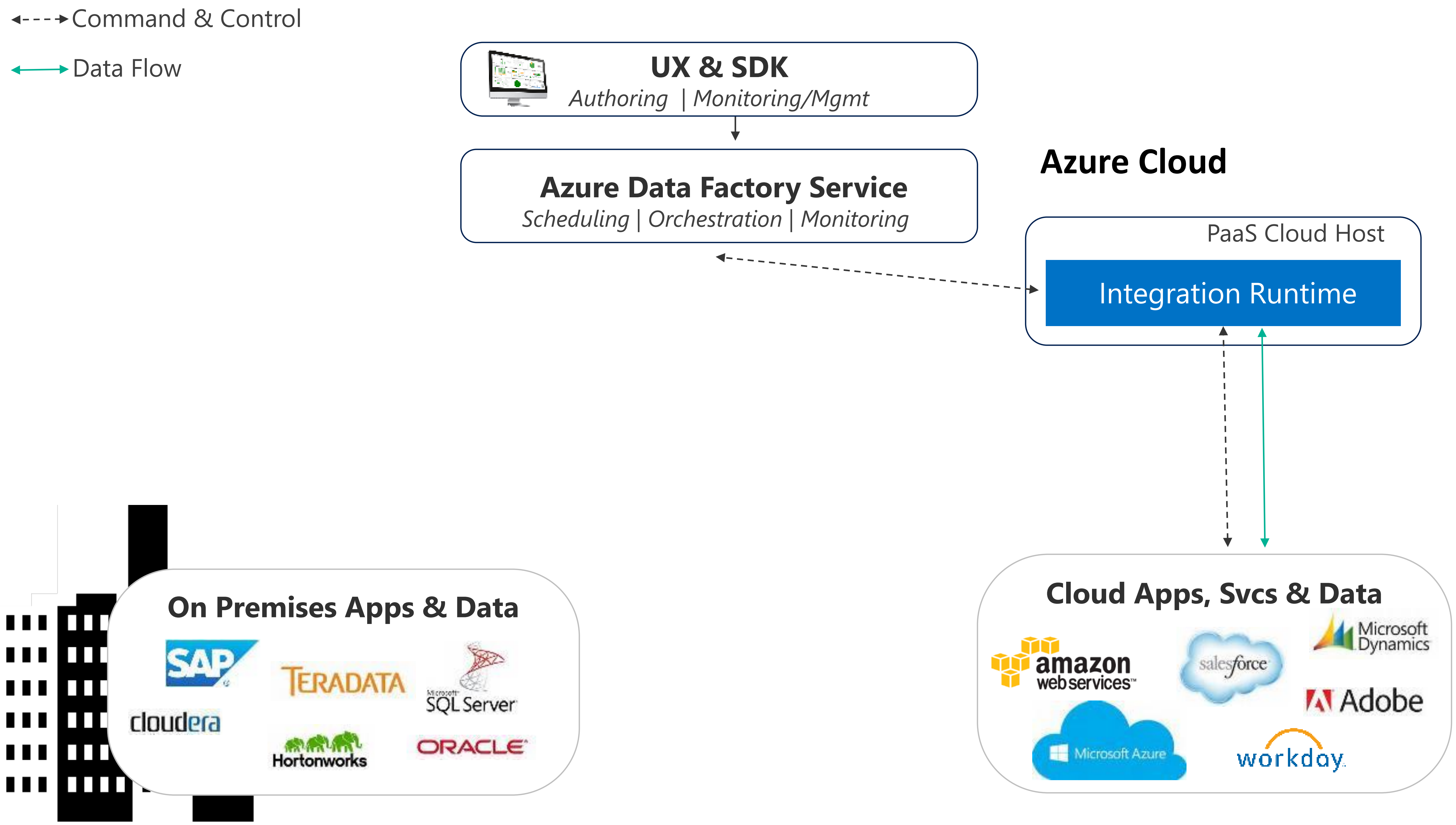
Authoring | Monitoring/Mgmt

Azure Data Factory Service

Scheduling | Orchestration | Monitoring

On Premises Apps & Data

Cloud Apps, Svcs & Data



Command & Control

Data Flow

UX & SDK

Authoring | Monitoring/Mgmt

Azure Data Factory Service

Scheduling | Orchestration | Monitoring

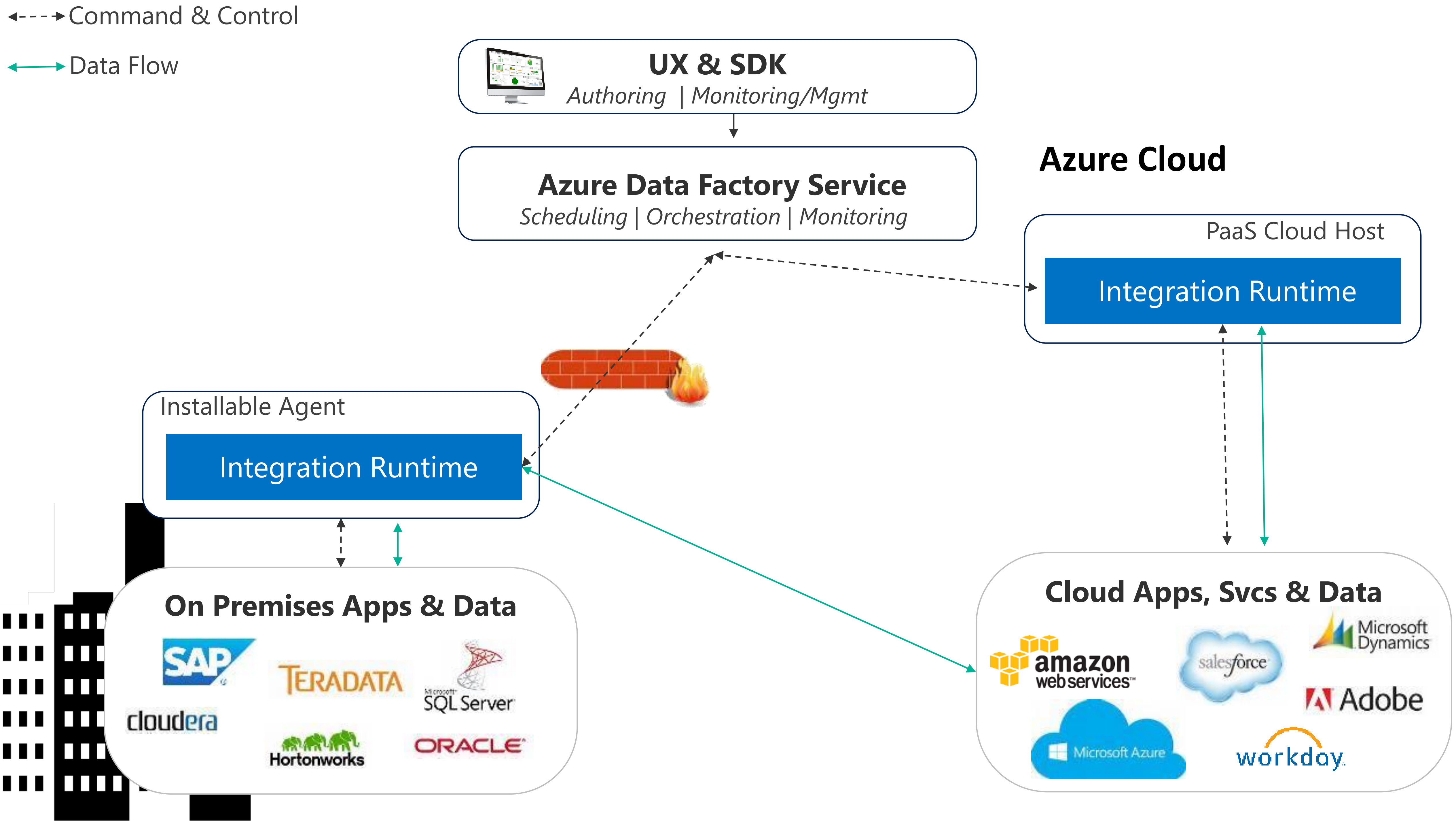
Azure Cloud

PaaS Cloud Host

Integration Runtime

On Premises Apps & Data

Cloud Apps, Svcs & Data



Command & Control

Data Flow

UX & SDK

Authoring | Monitoring/Mgmt

Azure Data Factory Service

Scheduling | Orchestration | Monitoring

Azure Cloud

PaaS Cloud Host

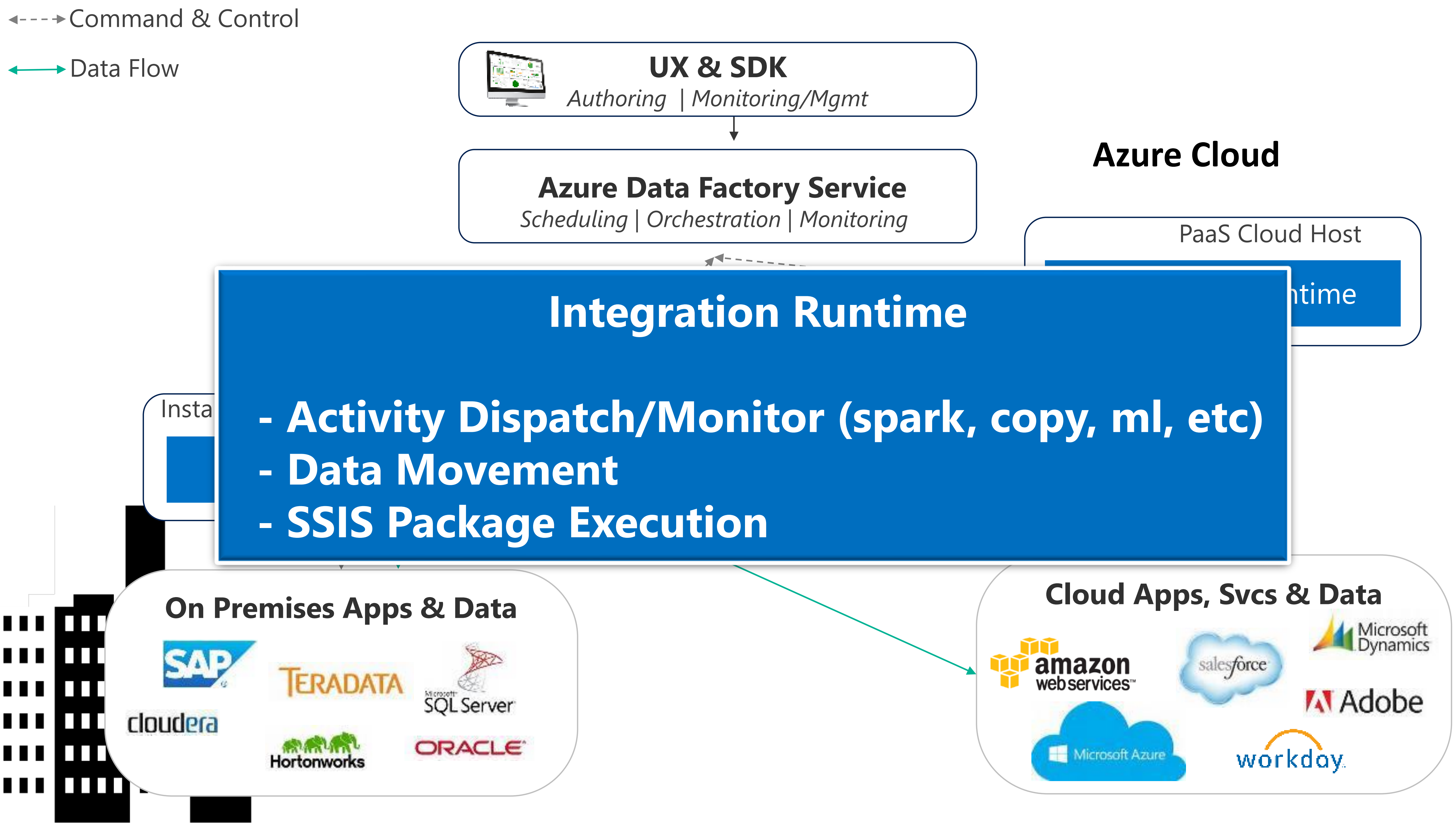
Integration Runtime

Installable Agent

Integration Runtime

On Premises Apps & Data

Cloud Apps, Svcs & Data



Command & Control

Data Flow

UX & SDK

Authoring | Monitoring/Mgmt

Azure Data Factory Service

Scheduling | Orchestration | Monitoring

Azure Cloud

PaaS Cloud Host

ntime

Insta

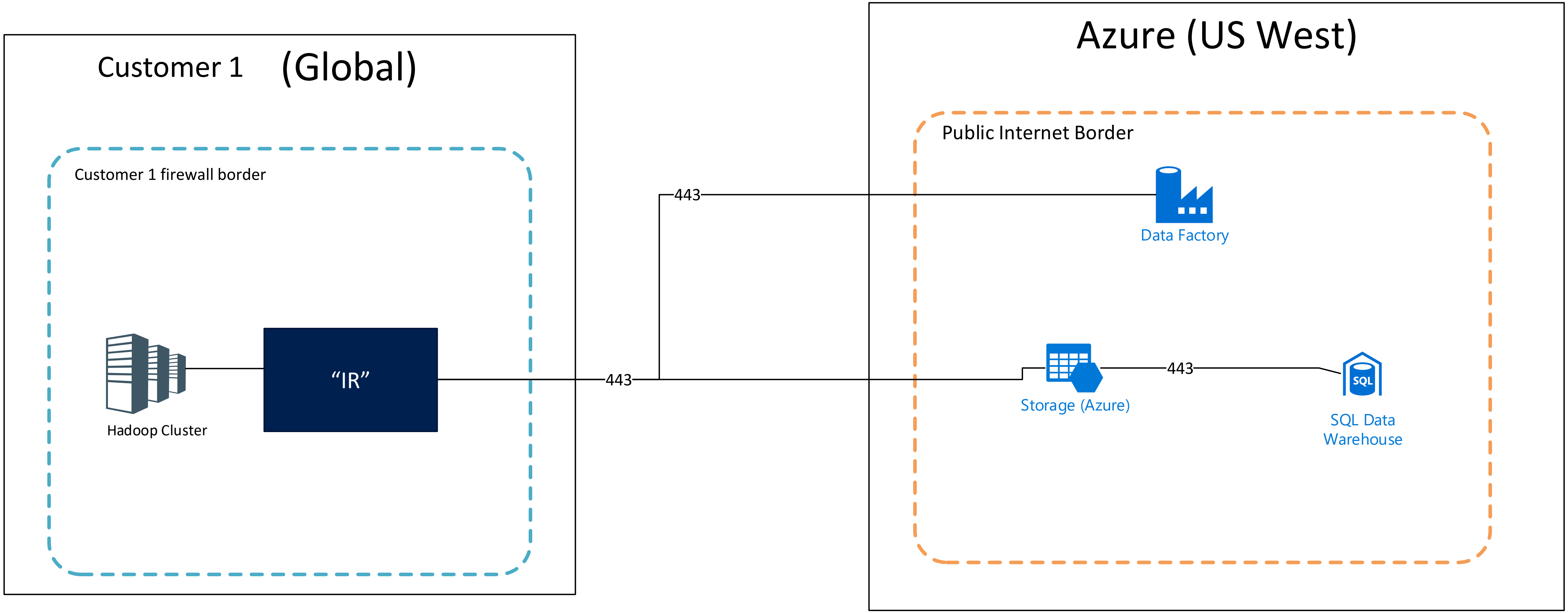
Integration Runtime

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

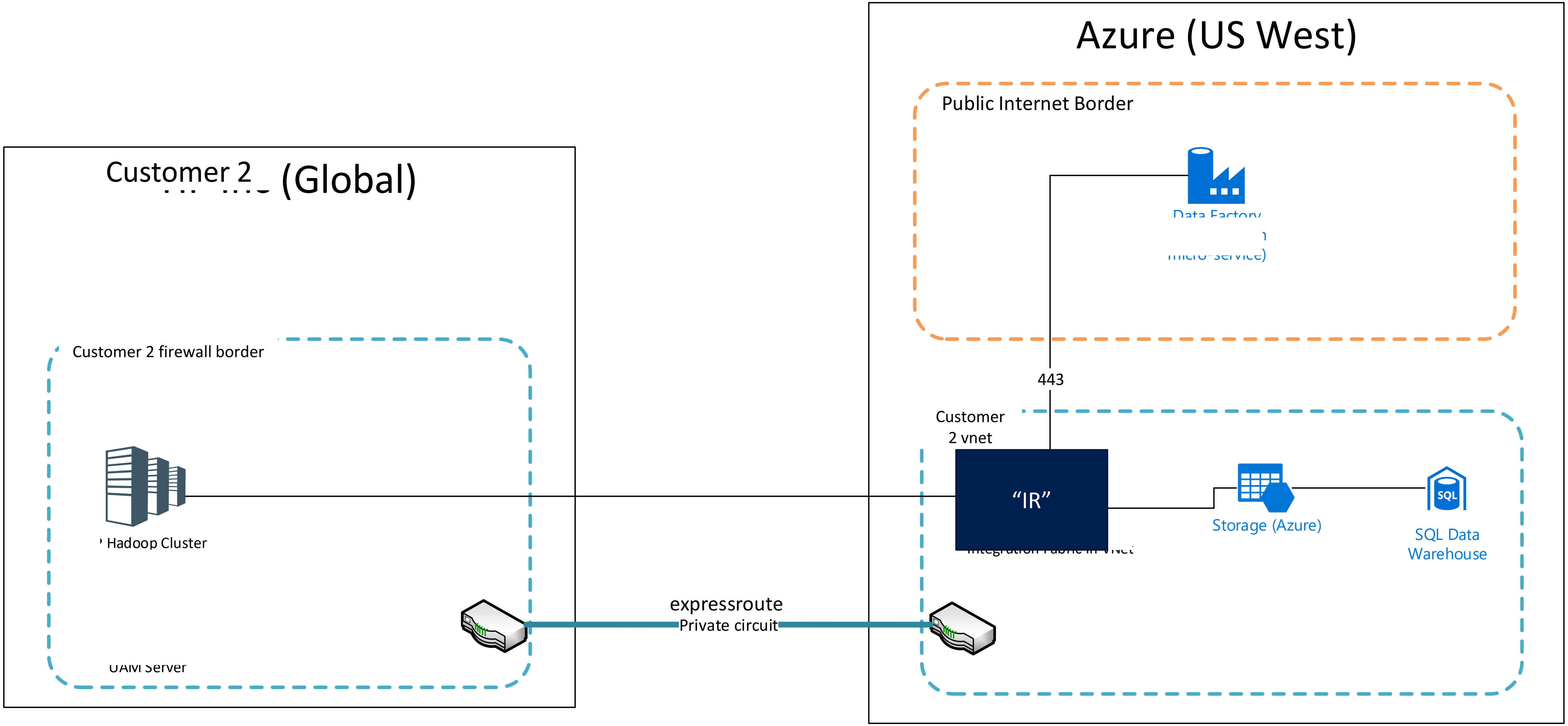
On Premises Apps & Data

Cloud Apps, Svcs & Data

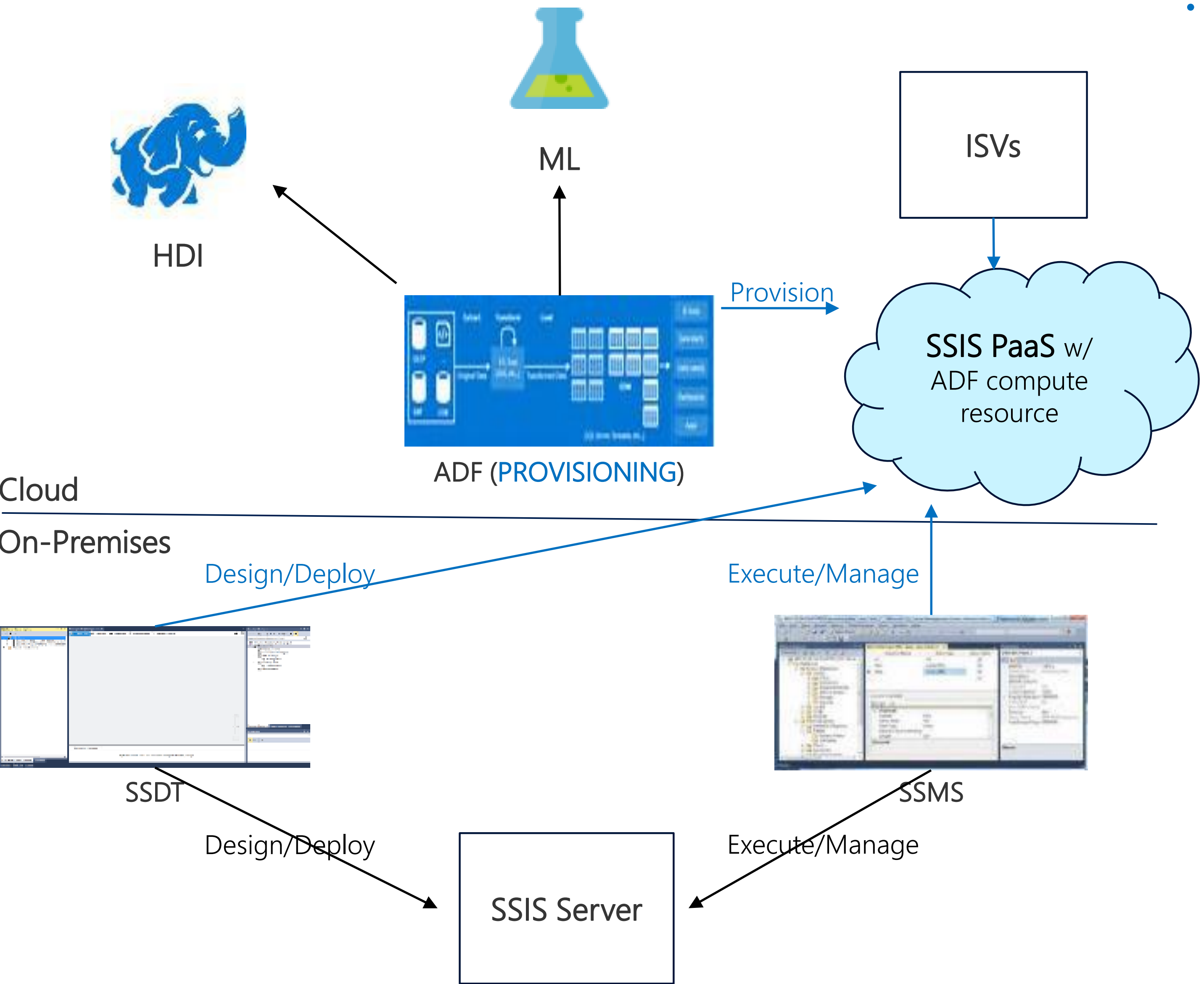
Azure Data Factory “Integration Runtime” deployed on premises for transformation and then moved to cloud



Azure Data Factory “Integration Runtime” deployed inside VNet

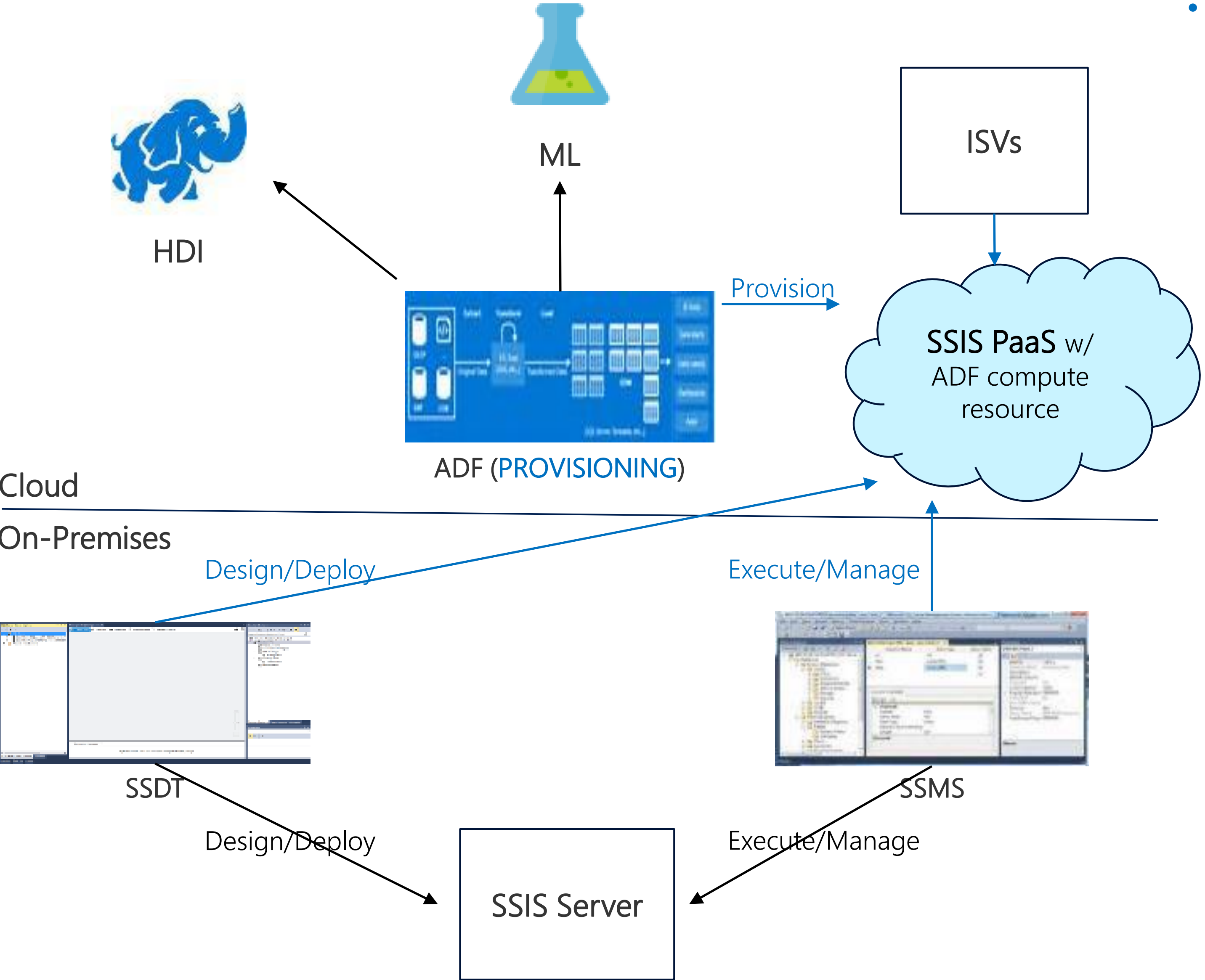


Microsoft ETL/ELT Services in Azure



- Introducing Azure-SSIS IR: Managed cluster of Azure VMs (nodes) dedicated to run your SSIS packages and no other activities
  - You can scale it up/out by specifying the node size /number of nodes in the cluster
  - You can bring your own Azure SQL Database (DB)/Managed Instance (MI) server to host the catalog of SSIS projects/packages (**SSISDB**) that will be attached to it
  - You can join it to a Virtual Network (VNet) that is connected to your on-prem network to enable on-prem data access
  - Once provisioned, you can enter your Azure SQL DB/MI server endpoint on SSDT/SSMS to deploy SSIS projects/packages and configure/execute them just like using SSIS on premises

Microsoft ETL/ELT Services in Azure



- Customer cohorts for Phase 1:

1. "SQL Migrators"

These are SSIS customers who want to retire their on-prem SQL Servers and migrate all apps + data (“complete/full lift & shift”) into Azure SQL MI – For them, SSISDB can be hosted by Azure SQL MI inside VNet

2. "ETL Cost Cutters”

These are SSIS customers who want to lower their operational costs and gain High Availability (HA)/scalability for just their ETL workloads w/o managing their own infra (“partial lift & shift") – For them, SSISDB can be hosted by Azure SQL DB in the public network

Deployment Methods

- SSIS PaaS supports the project deployment model used in SSIS 2012/later versions
  - Projects built in the legacy package deployment model used in SSIS 2008/earlier versions can be converted into this model via SSDT/SSMS using Integration Services Project Conversion Wizard
  - Packages built in SSIS 2008/earlier versions can be upgraded to the latest version supported by SSIS PaaS via SSDT/SSMS using SSIS Package Upgrade Wizard
  - In this model, the whole project needs to be deployed after any package changes – An incremental package deployment feature will be provided in the near future
  - Projects containing environment references/run-time parameters can be saved into project deployment files (.ispac extension)
  - Projects are deployed into SSISDB hosted by Azure SQL DB/MI server, packages are run by creating/starting jobs via SSISDB sprocs that will be executed on Azure-SSIS IR, and execution logs are written back into SSISDB

Deployment Methods

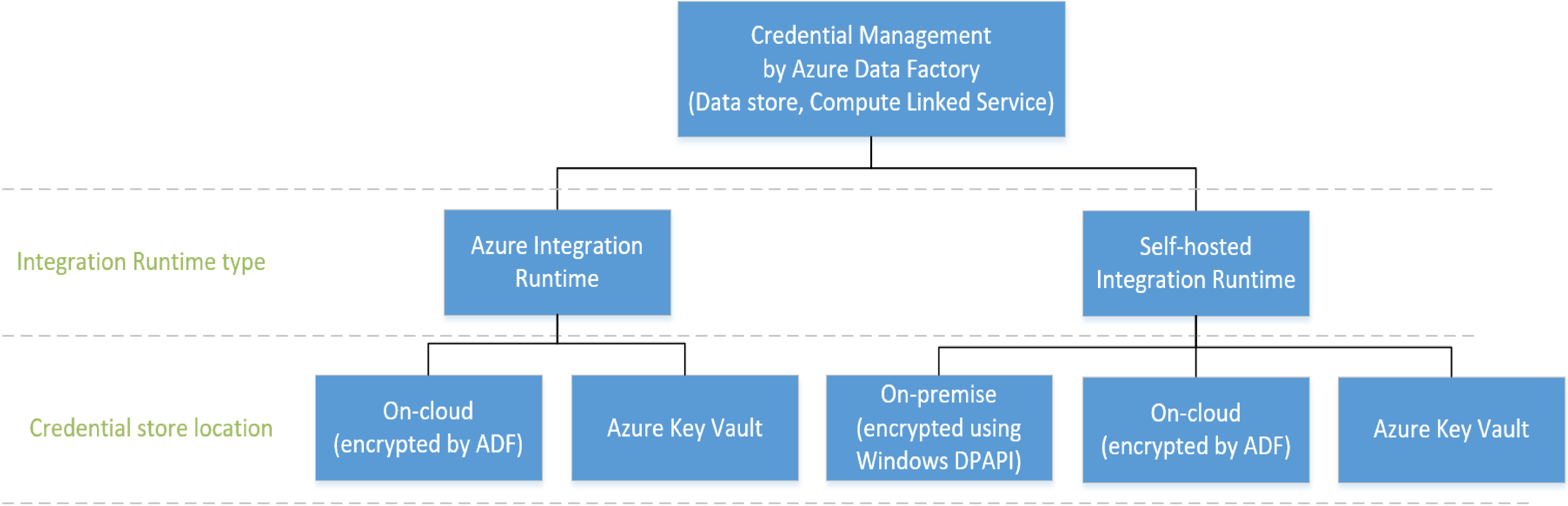
- SSIS projects can be deployed via SSDT/SSMS using Integration Services Deployment Wizard

- SSIS projects can be deployed via Command Line Interface (CLI)
  - Run isdeploymentwizard.exe from the command prompt (TBD)

- SSIS projects can be deployed via custom code/PSH using SSIS Managed Object Model (MOM) .NET SDK/API
  - Microsoft.SqlServer.Management.IntegrationServices.dll is installed in .NET Global Assembly Cache (GAC) with SQL Server/SSMS installation

- SSIS projects can be deployed via T-SQL scripts executing SSISDB sprocs
  - Execute SSISDB sproc [catalog].[deploy_project]

Credential Management (Linked service)



Credential Management by Azure Data Factory (Data store, Compute Linked Service)

Integration Runtime type

Azure Integration Runtime

Self-hosted Integration Runtime

Credential store location

On-cloud (encrypted by ADF)

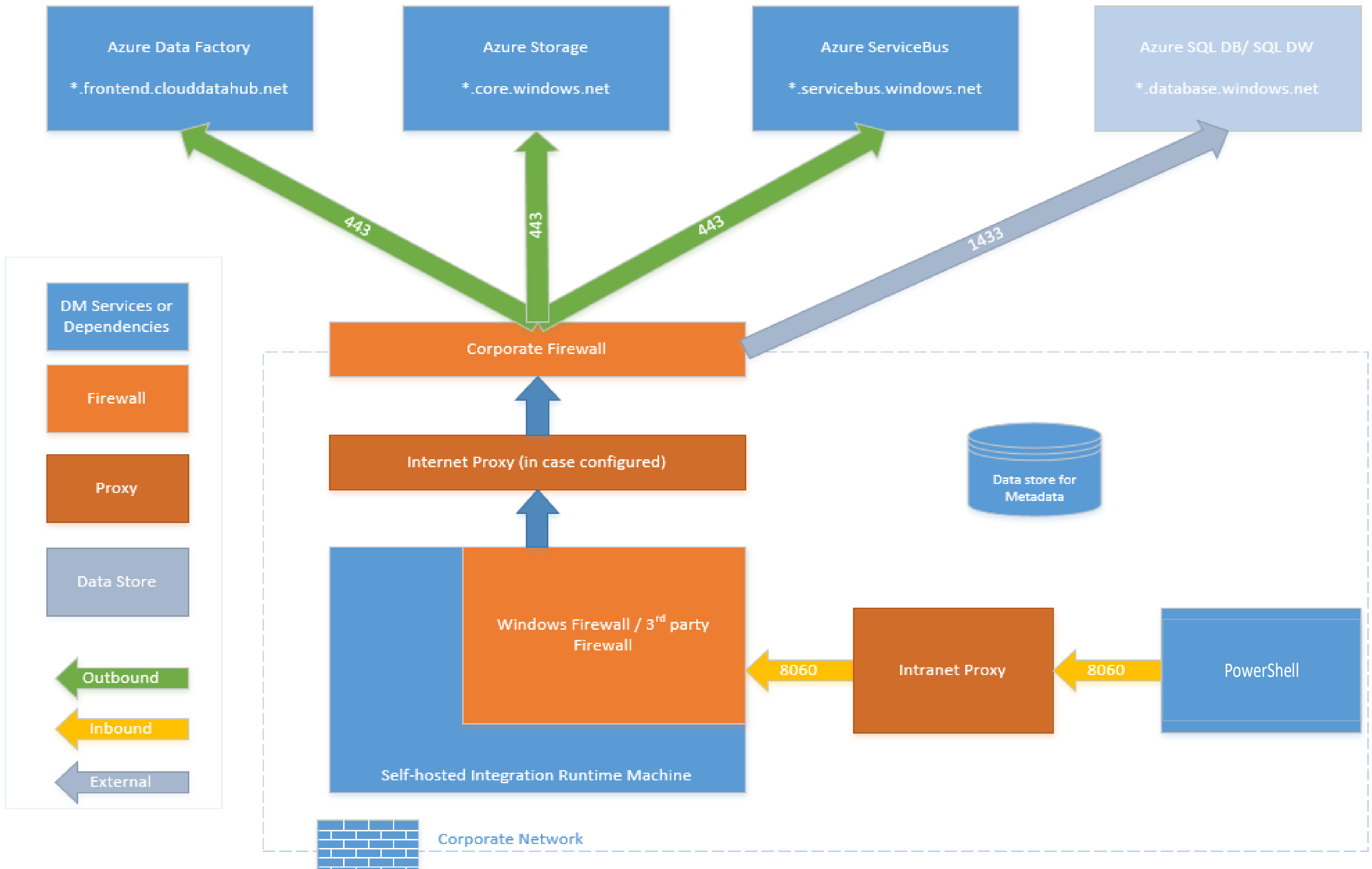
Azure Key Vault

On-premise (encrypted using Windows DPAPI)

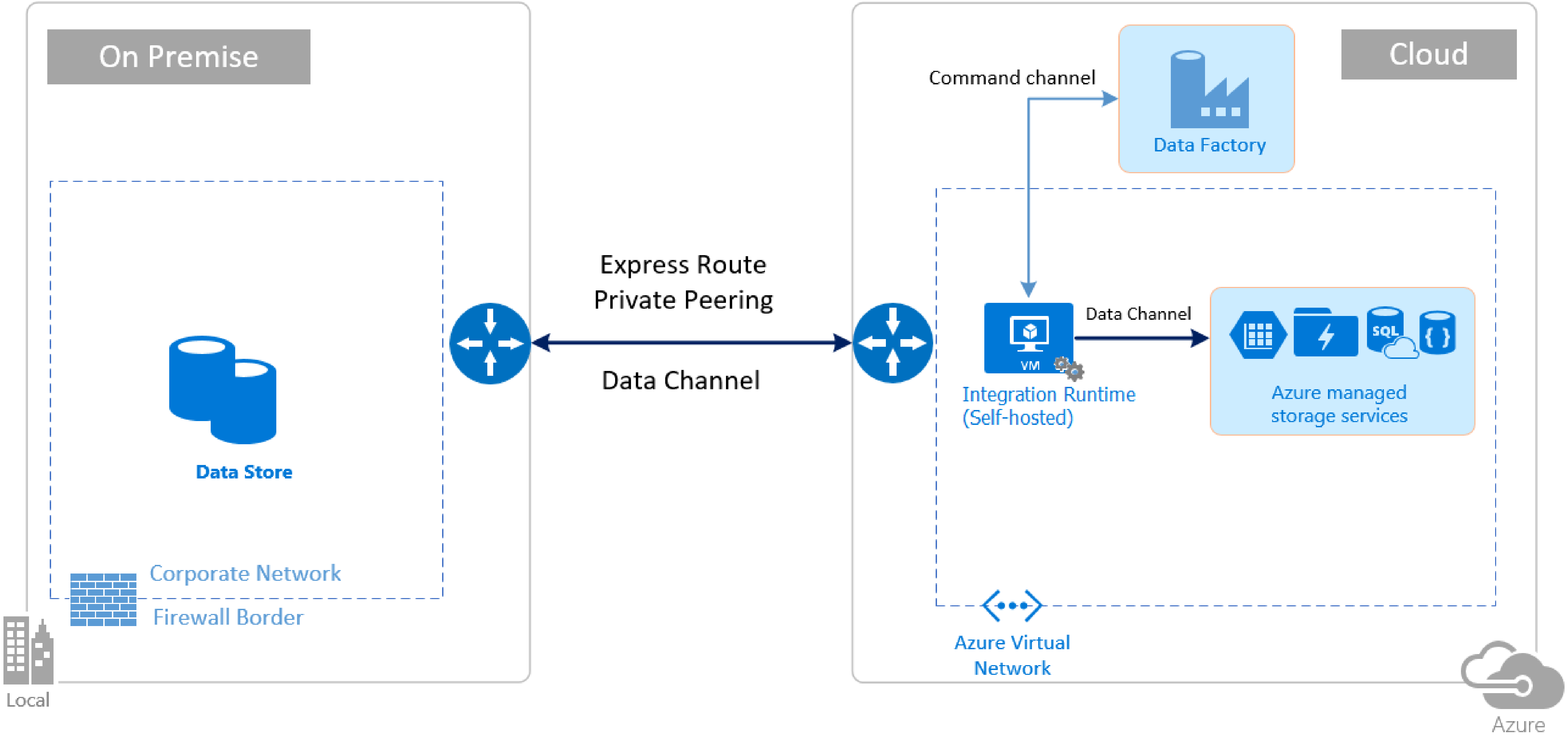
On-cloud (encrypted by ADF)

Azure Key Vault

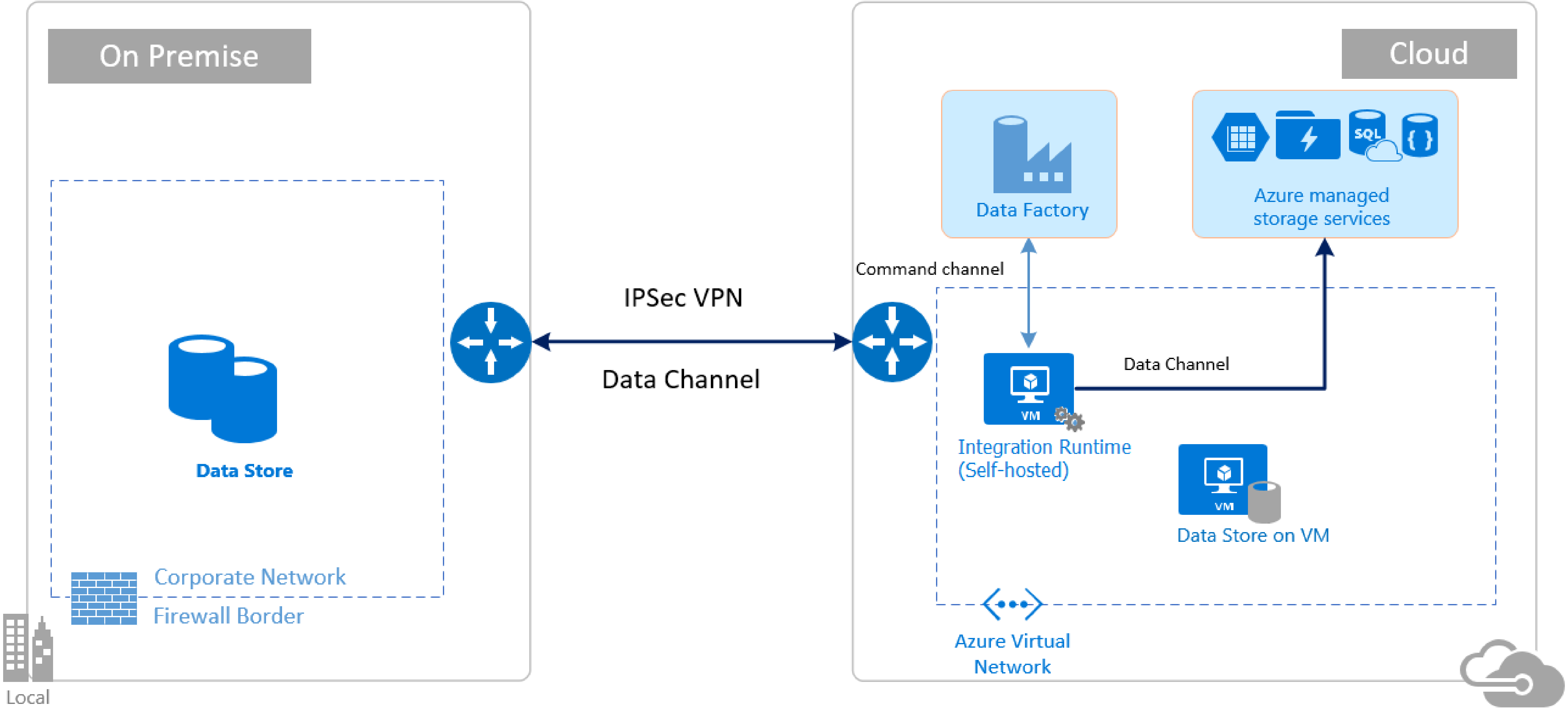
Self-hosted IR – Firewall Requirements



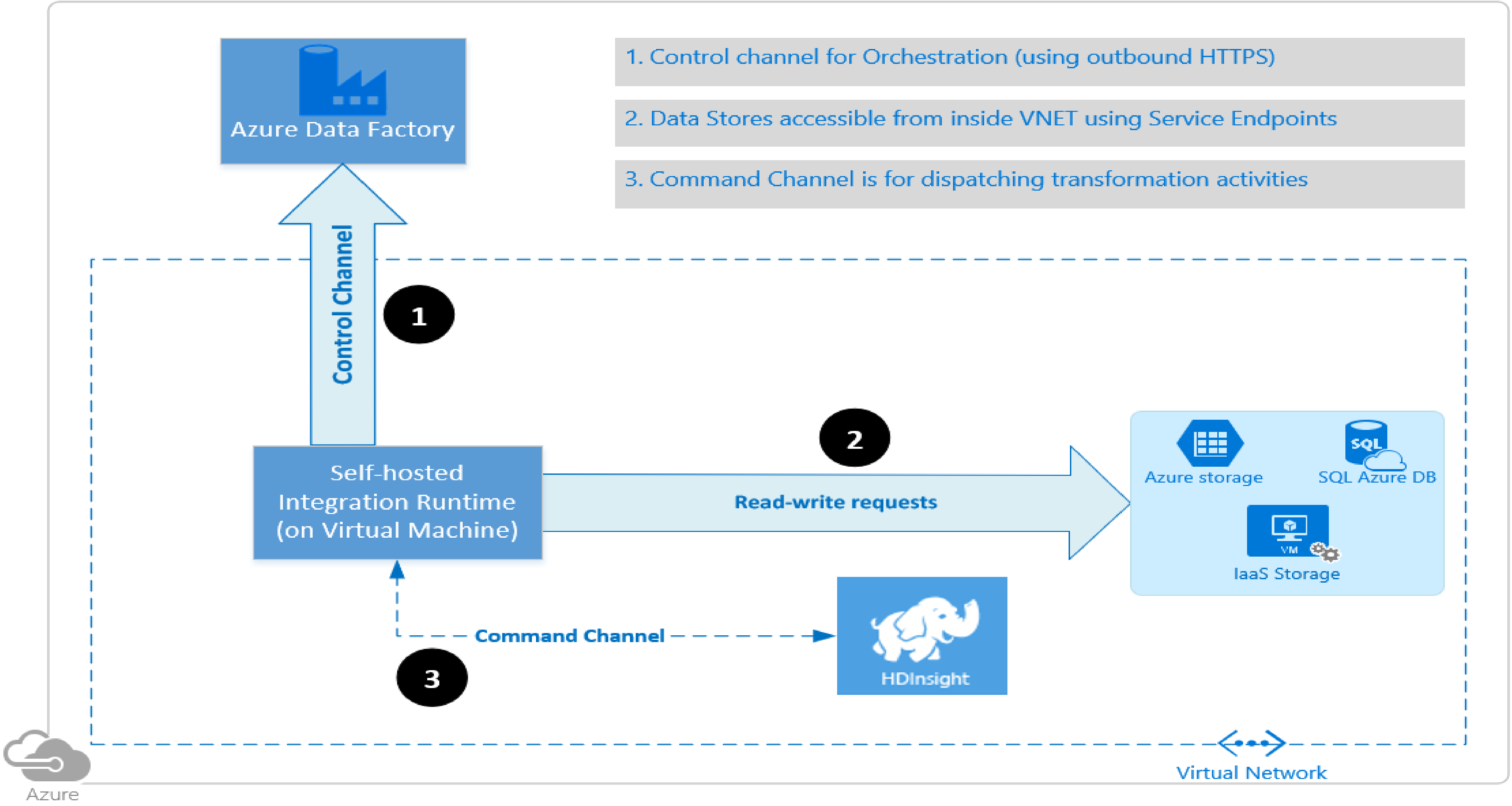
Network Topology (with ExpressRoute)



Network Topology (with VPN)



VNet



1. Control channel for Orchestration (using outbound HTTPS)

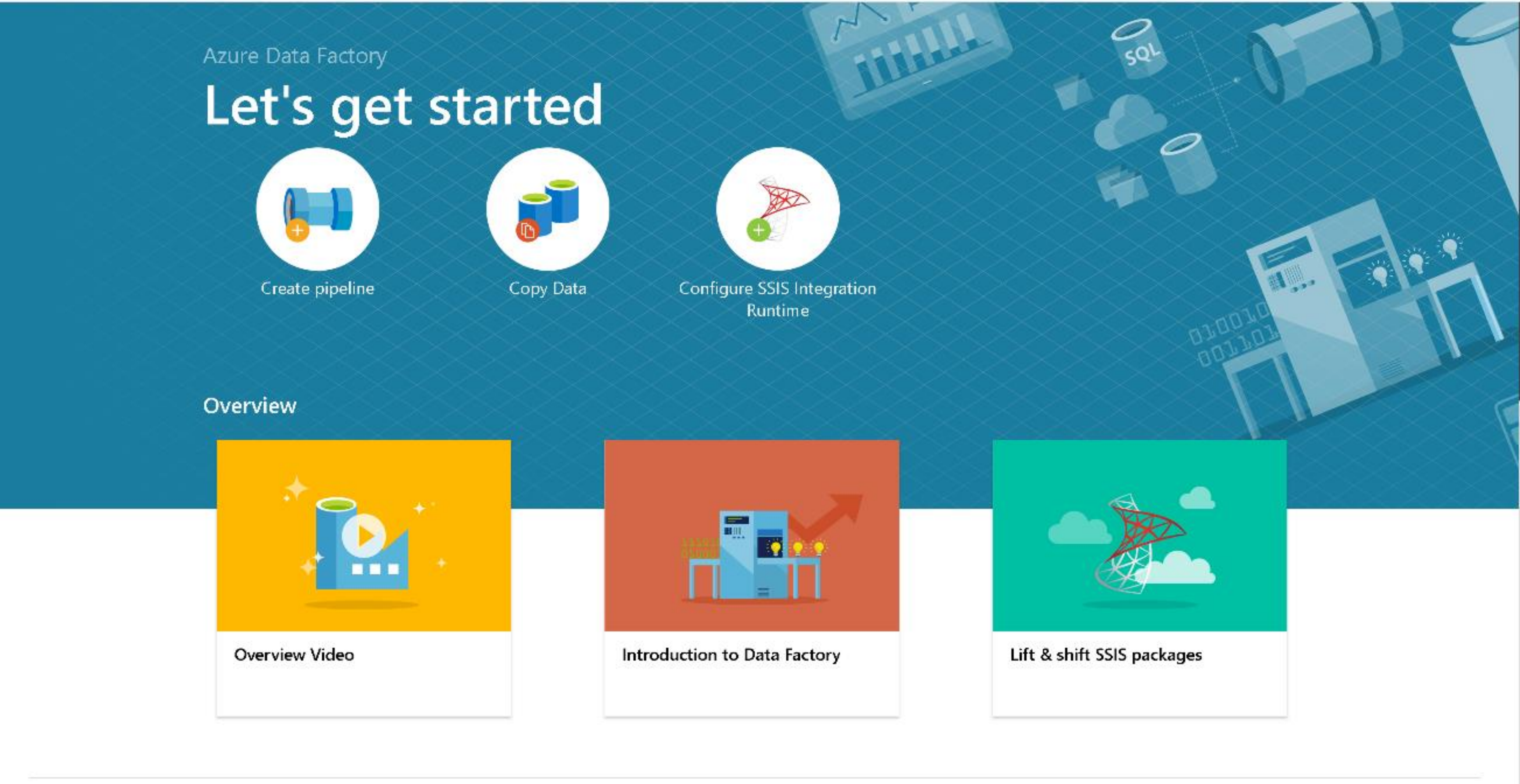
2. Data Stores accessible from inside VNET using Service Endpoints

3. Command Channel is for dispatching transformation activities

Demo

- Walkthrough of the UI with provisioning
- Demo Copy Activity – Blob to Blob
- Provision Self Hosted IR
- Demo Copy Activity – On Prem to Blob with Self Hosted IR
- Demo SSIS Runtime
- Demo ADB Activity with format conversion

Questions?



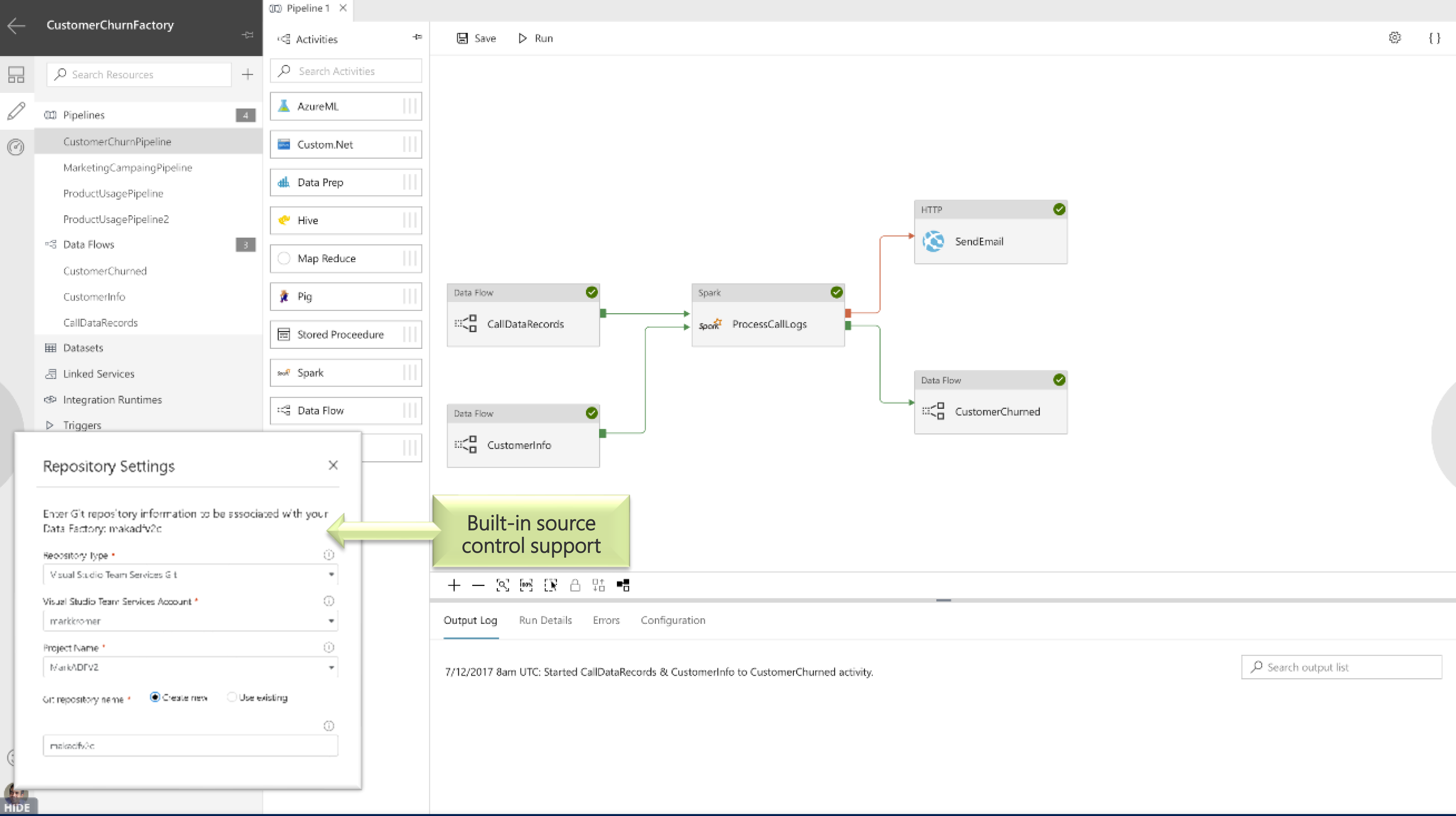
Azure Data Factory

Let's get started

Create pipeline
Copy Data
Configure SSIS Integration Runtime

Overview

Overview Video
Introduction to Data Factory
Lift & shift SSIS packages



Repository Settings

Enter Git repository information to be associated with your Data Factory: makadfv2c

Repository type *: Visual Studio Team Services Git

Visual Studio Team Services Account *: markkromer

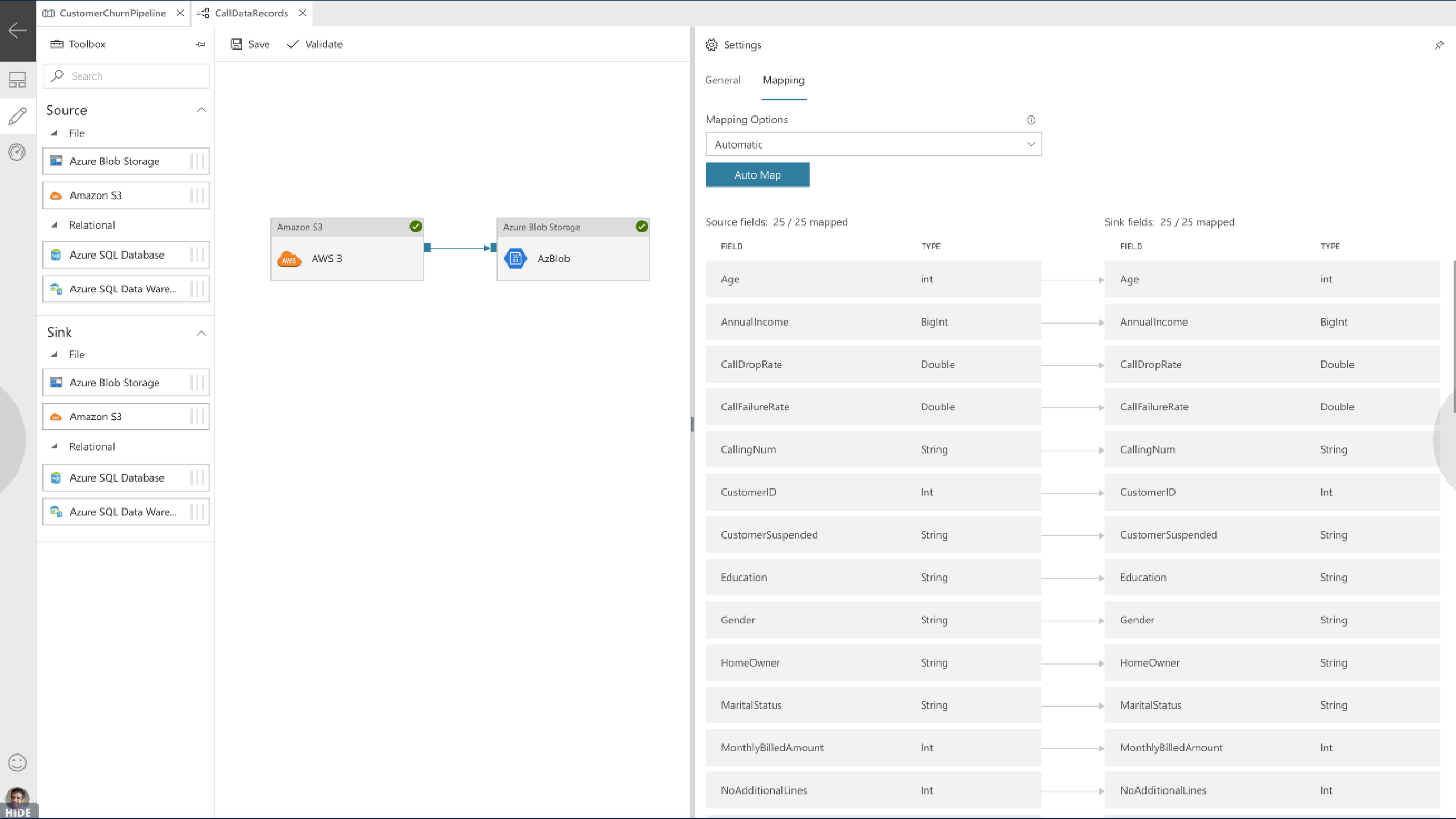
Project Name *: MarkADFV2

Git repository name *: ● Create new ○ Use existing

makadfv2c

7/12/2017 8am UTC: Started CallDataRecords & CustomerInfo to CustomerChurned activity.

Built-in source control support

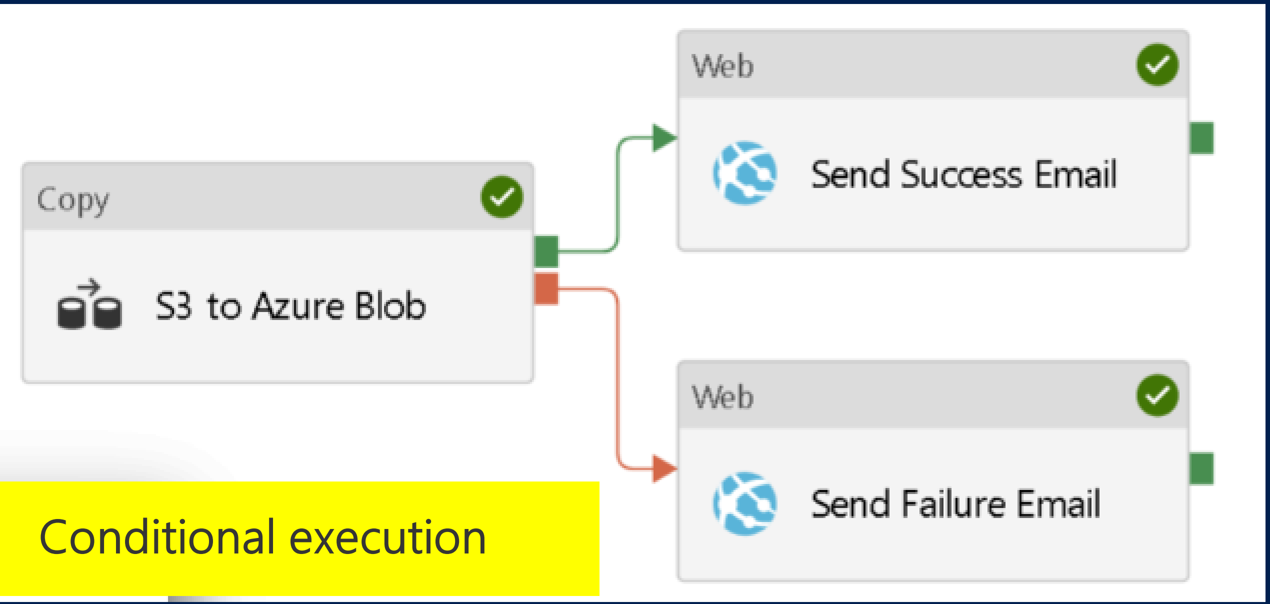


Mapping Options: Automatic

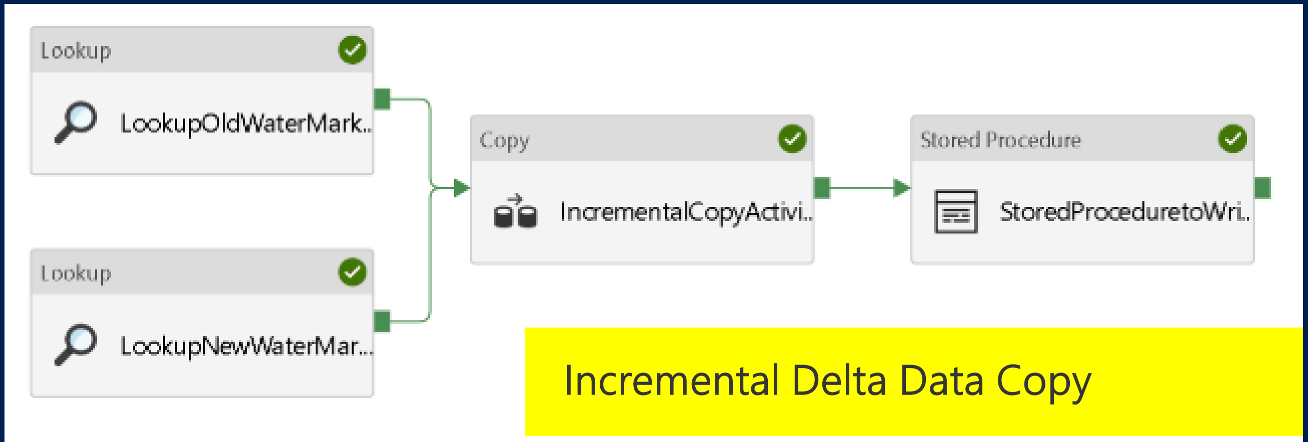
Source fields: 25 / 25 mapped

Sink fields: 25 / 25 mapped

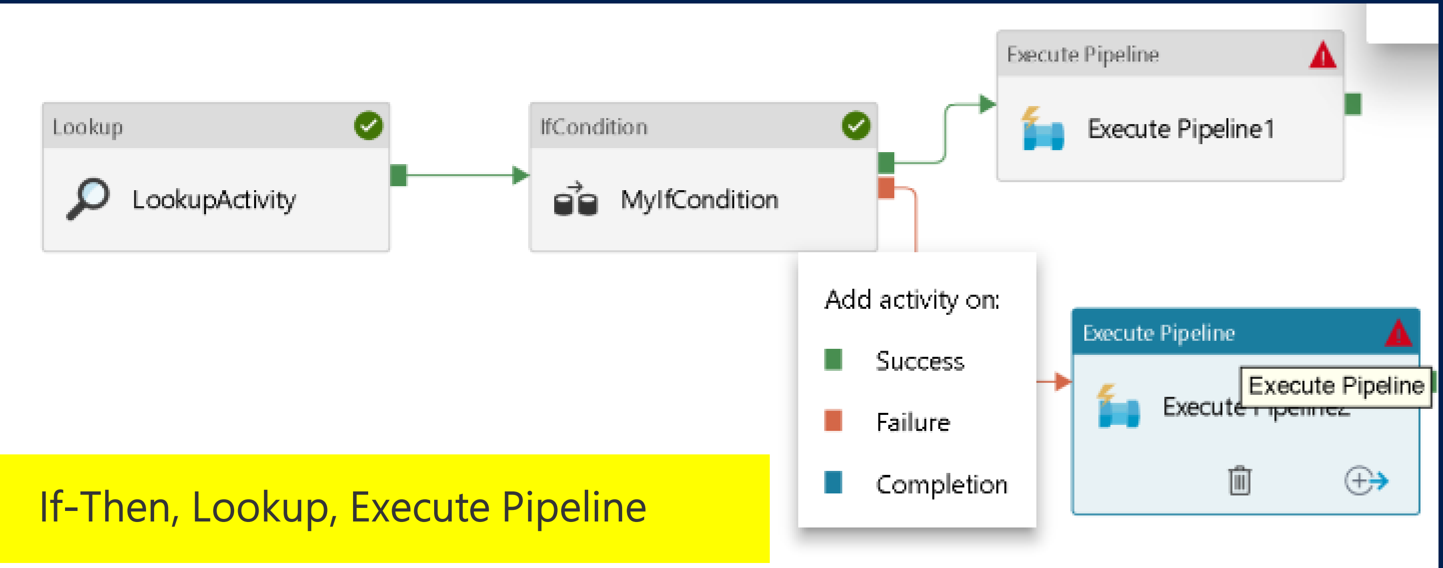
FIELD	TYPE	FIELD	TYPE
Age	int	Age	int
AnnualIncome	BigInt	AnnualIncome	BigInt
CallDropRate	Double	CallDropRate	Double
CallFailureRate	Double	CallFailureRate	Double
CallingNum	String	CallingNum	String
CustomerID	Int	CustomerID	Int
CustomerSuspended	String	CustomerSuspended	String
Education	String	Education	String
Gender	String	Gender	String
HomeOwner	String	HomeOwner	String
MaritalStatus	String	MaritalStatus	String
MonthlyBilledAmount	Int	MonthlyBilledAmount	Int
NoAdditionalLines	Int	NoAdditionalLines	Int



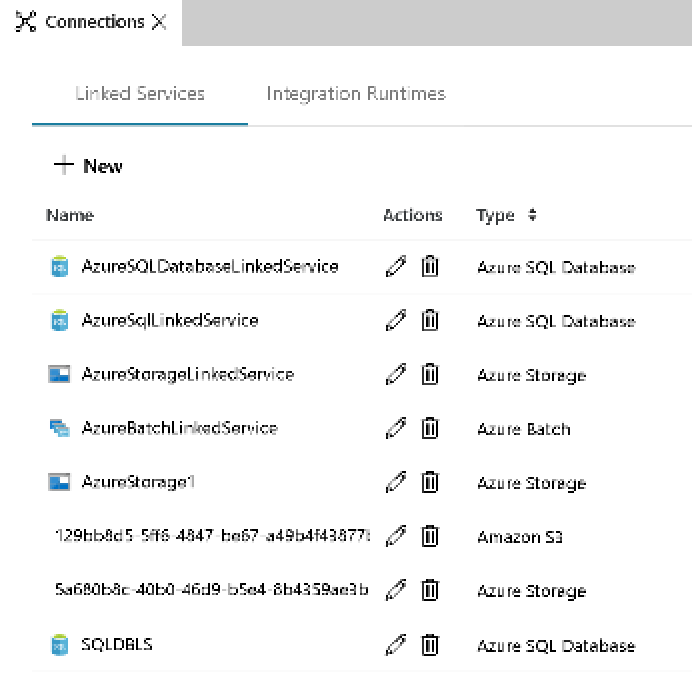
Conditional execution



Incremental Delta Data Copy



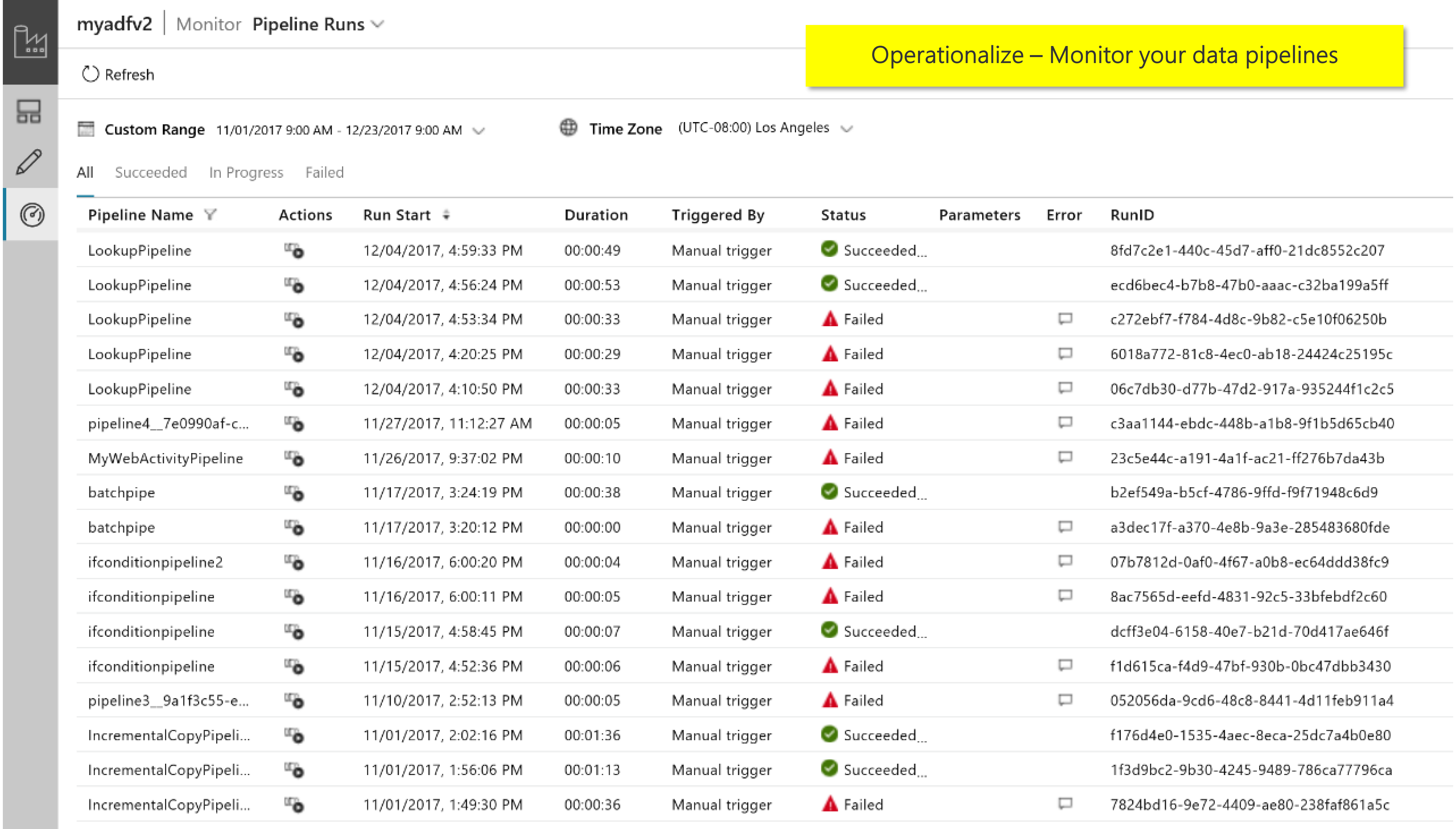
If-Then, Lookup, Execute Pipeline



Name	Actions	Type
AzureSQLDatabaseLinkedService		Azure SQL Database
AzureSqlLinkedService		Azure SQL Database
AzureStorageLinkedService		Azure Storage
AzureBatchLinkedService		Azure Batch
AzureStorage1		Azure Storage
129bb8d5-5ff6-4847-be67-a49b4f43877		Amazon S3
5a680b8c-40b0-46d9-b5e4-8b4359ae3b		Azure Storage
SQLDBLS		Azure SQL Database

Connection Managers

Operationalize – Monitor your data pipelines

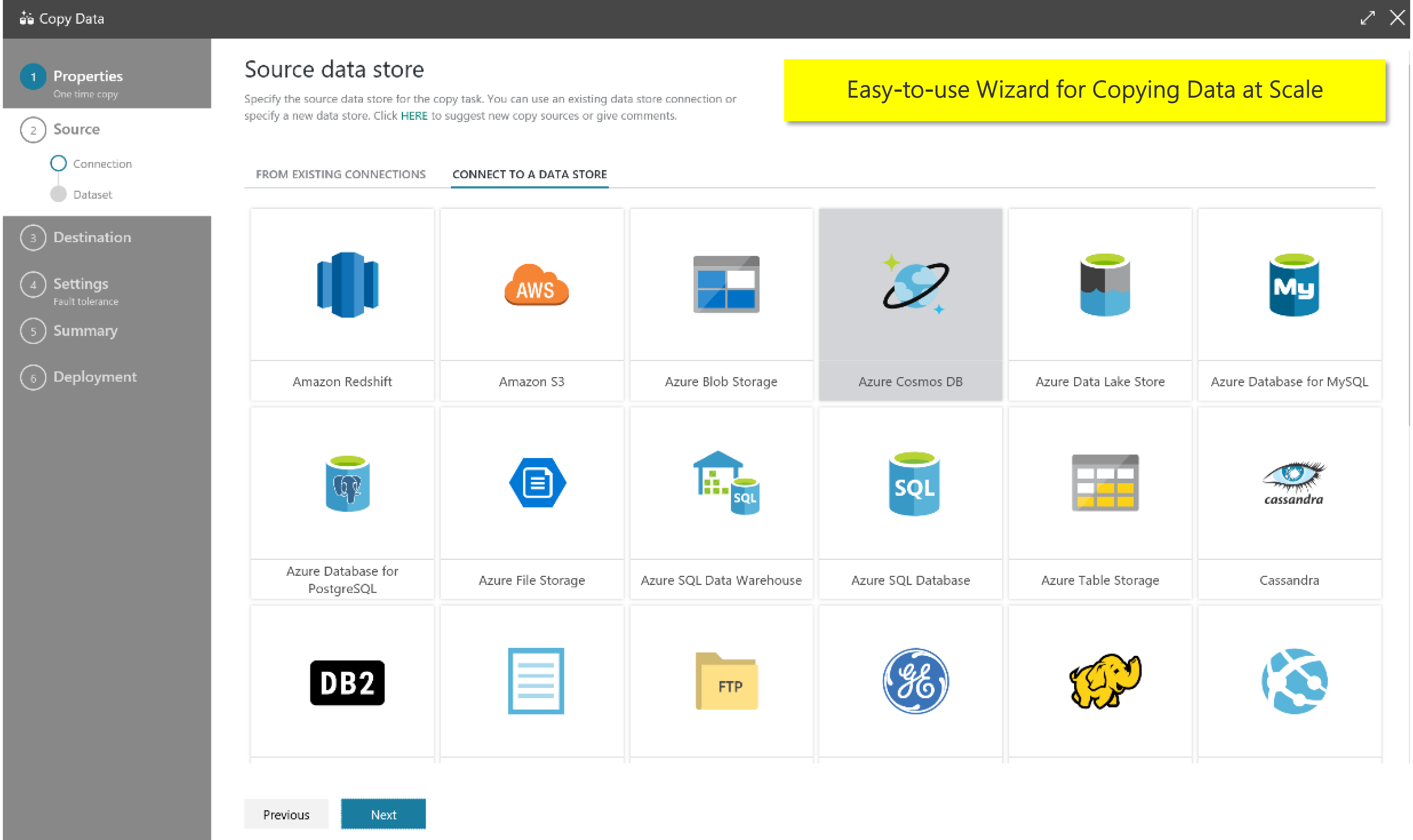


Custom Range 11/01/2017 9:00 AM - 12/23/2017 9:00 AM

Time Zone (UTC-08:00) Los Angeles

Pipeline Name	Actions	Run Start	Duration	Triggered By	Status	Parameters	Error	RunID
LookupPipeline	⚙	12/04/2017, 4:59:33 PM	00:00:49	Manual trigger	✓ Succeeded...			8fd7c2e1-440c-45d7-aff0-21dc8552c207
LookupPipeline	⚙	12/04/2017, 4:56:24 PM	00:00:53	Manual trigger	✓ Succeeded...			ecd6bec4-b7b8-47b0-aaac-c32ba199a5ff
LookupPipeline	⚙	12/04/2017, 4:53:34 PM	00:00:33	Manual trigger	⚠ Failed		🗨	c272ebf7-f784-4d8c-9b82-c5e10f06250b
LookupPipeline	⚙	12/04/2017, 4:20:25 PM	00:00:29	Manual trigger	⚠ Failed		🗨	6018a772-81c8-4ec0-ab18-24424c25195c
LookupPipeline	⚙	12/04/2017, 4:10:50 PM	00:00:33	Manual trigger	⚠ Failed		🗨	06c7db30-d77b-47d2-917a-935244f1c2c5
pipeline4__7e0990af-c...	⚙	11/27/2017, 11:12:27 AM	00:00:05	Manual trigger	⚠ Failed		🗨	c3aa1144-ebdc-448b-a1b8-9f1b5d65cb40
MyWebActivityPipeline	⚙	11/26/2017, 9:37:02 PM	00:00:10	Manual trigger	⚠ Failed		🗨	23c5e44c-a191-4a1f-ac21-ff276b7da43b
batchpipe	⚙	11/17/2017, 3:24:19 PM	00:00:38	Manual trigger	✓ Succeeded...			b2ef549a-b5cf-4786-9ffd-f9f71948c6d9
batchpipe	⚙	11/17/2017, 3:20:12 PM	00:00:00	Manual trigger	⚠ Failed		🗨	a3dec17f-a370-4e8b-9a3e-285483680fde
ifconditionpipeline2	⚙	11/16/2017, 6:00:20 PM	00:00:04	Manual trigger	⚠ Failed		🗨	07b7812d-0af0-4f67-a0b8-ec64ddd38fc9
ifconditionpipeline	⚙	11/16/2017, 6:00:11 PM	00:00:05	Manual trigger	⚠ Failed		🗨	8ac7565d-eefd-4831-92c5-33bfebdf2c60
ifconditionpipeline	⚙	11/15/2017, 4:58:45 PM	00:00:07	Manual trigger	✓ Succeeded...			dcff3e04-6158-40e7-b21d-70d417ae646f
ifconditionpipeline	⚙	11/15/2017, 4:52:36 PM	00:00:06	Manual trigger	⚠ Failed		🗨	f1d615ca-f4d9-47bf-930b-0bc47dbb3430
pipeline3__9a1f3c55-e...	⚙	11/10/2017, 2:52:13 PM	00:00:05	Manual trigger	⚠ Failed		🗨	052056da-9cd6-48c8-8441-4d11feb911a4
IncrementalCopyPipeli...	⚙	11/01/2017, 2:02:16 PM	00:01:36	Manual trigger	✓ Succeeded...			f176d4e0-1535-4aec-8eca-25dc7a4b0e80
IncrementalCopyPipeli...	⚙	11/01/2017, 1:56:06 PM	00:01:13	Manual trigger	✓ Succeeded...			1f3d9bc2-9b30-4245-9489-786ca77796ca
IncrementalCopyPipeli...	⚙	11/01/2017, 1:49:30 PM	00:00:36	Manual trigger	⚠ Failed		🗨	7824bd16-9e72-4409-ae80-238faf861a5c

Easy-to-use Wizard for Copying Data at Scale



Source data store

Specify the source data store for the copy task. You can use an existing data store connection or specify a new data store. Click HERE to suggest new copy sources or give comments.

Customer Insights

- SSIS is a traditional ETL tool that comes bundled with SQL Server on premises
  - Has been around for more than 10 years
  - Some customers have started to lift & shift their ETL workloads to the cloud to reduce their on-prem infra, but found managing Infrastructure as a Service (**IaaS**)/VMs challenging

Customer Insights

- SSIS is a traditional ETL tool that comes bundled with SQL Server on premises
  - Has been around for more than 10 years
  - Some customers have started to lift & shift their ETL workloads to the cloud to reduce their on-prem infra, but found managing Infrastructure as a Service (**IaaS**)/VMs challenging

- Azure Data Factory (ADF) is a modern ELT tool that moves/copies data and dispatches transformations for Big Data Analytics in the cloud
  - Some gaps in ELT workflows can be filled w/ code-free authoring of transformations/built-in tasks from SSIS
  - Some customers have started to combine ADF with SSIS on IaaS/VMs, but found managing IaaS/VMs challenging

Customer Insights

- SSIS is a traditional ETL tool that comes bundled with SQL Server on premises
  - Has been around for more than 10 years
  - Some customers have started to lift & shift their ETL workloads to the cloud to reduce their on-prem infra, but found managing Infrastructure as a Service (**IaaS**)/VMs challenging
- Azure Data Factory (ADF) is a modern ELT tool that moves/copies data and dispatches transformations for Big Data Analytics in the cloud
  - Some gaps in ELT workflows can be filled w/ code-free authoring of transformations/built-in tasks from SSIS
  - Some customers have started to combine ADF with SSIS on IaaS/VMs, but found managing IaaS/VMs challenging
- Evolution of a cloud-first product: SSIS on premises -> IaaS -> PaaS
  - The stage is set for SSIS **PaaS**…