DP-203 Microsoft Azure Data Engineer

Day 6 - Azure Data Factory

30th July 2021

Vishwamitra Kala

databoq®

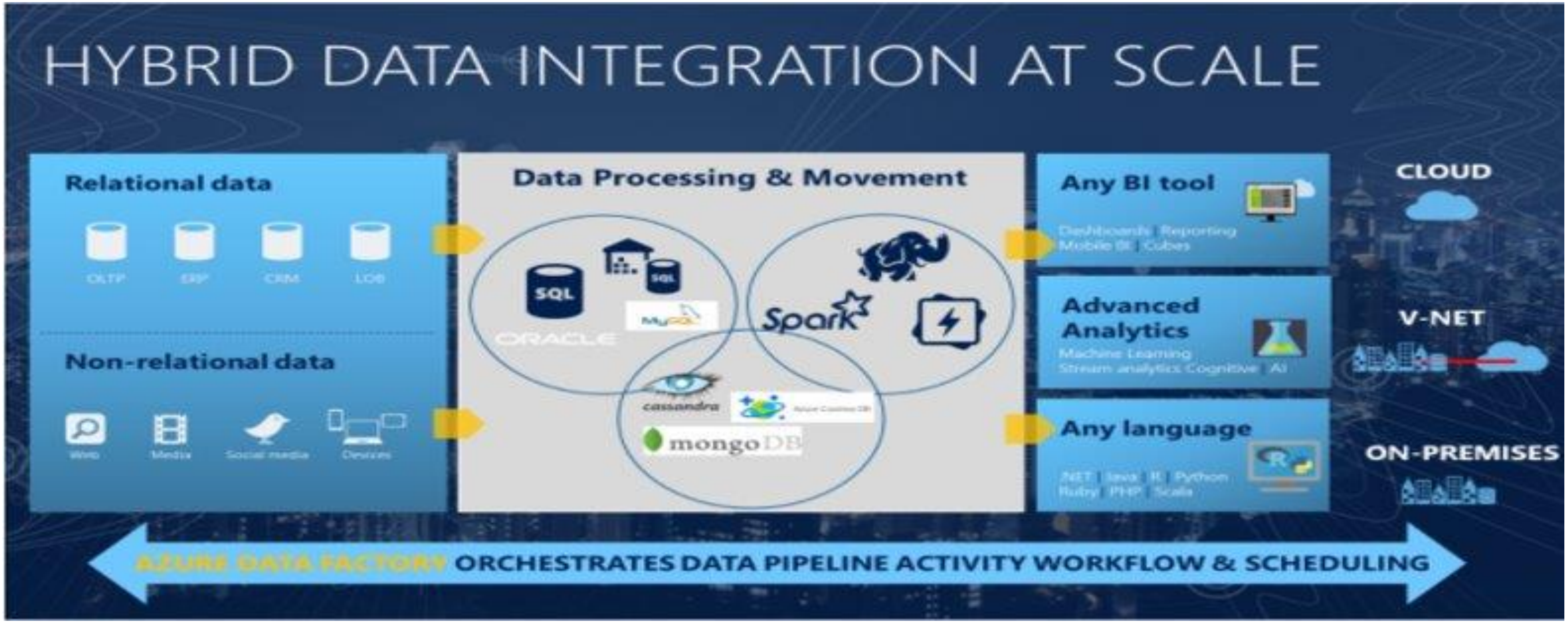
Understanding ADF

- Azure Data Factory is the cloud-based ETL and data integration service that allows you to create data-driven workflows for orchestrating data movement and transforming data at scale.

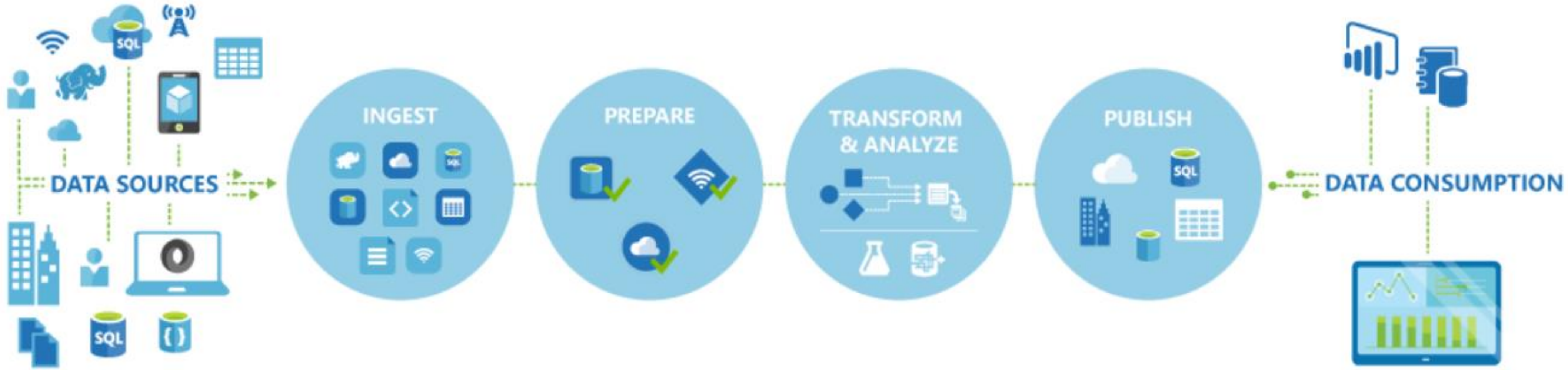
- Create and schedule data-driven workflows (called pipelines) that can ingest data from disparate data stores.

- Build complex ETL processes that transform data visually with data flows or by using compute services such as Azure HDInsight Hadoop, Azure Databricks, and Azure Synapse Analytics.

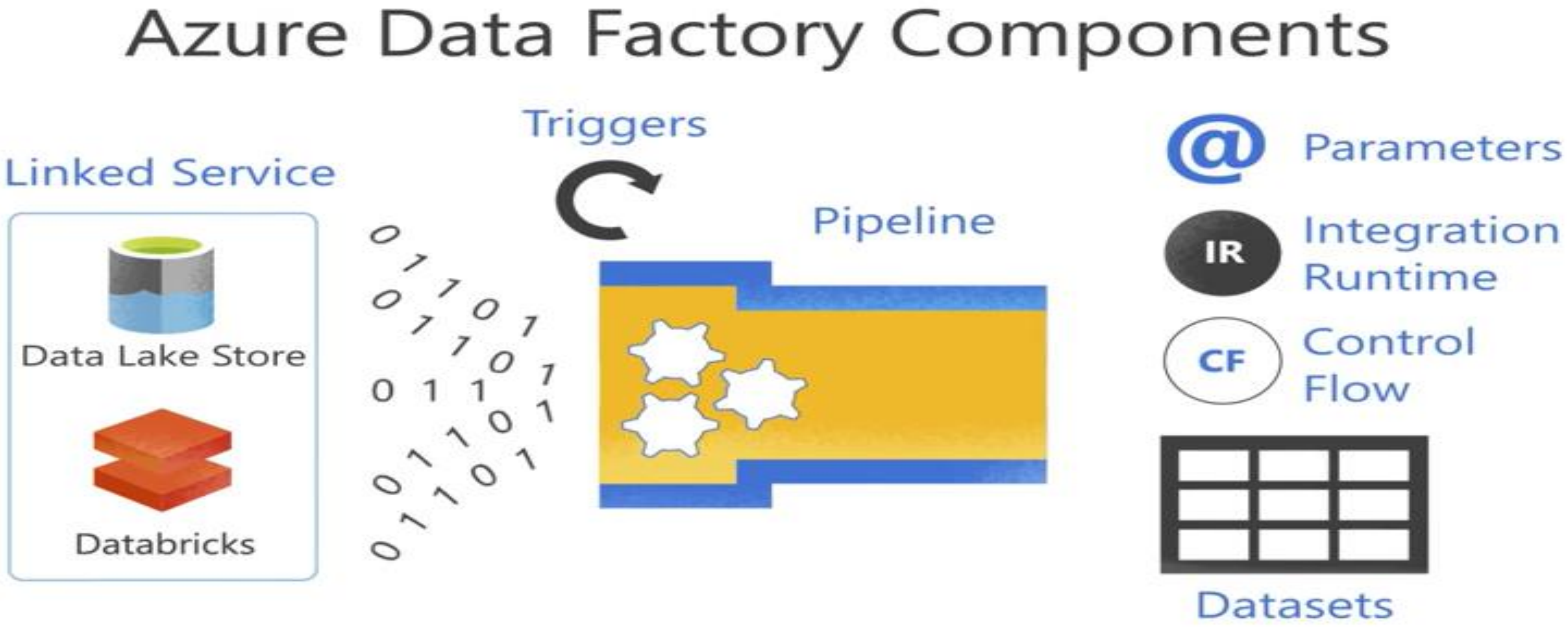
ADF as an Orchestrator



Data Factory Process



Data Factory Components



Set-up Azure Data Factory - Demo

- Name: The name of the Azure Data Factory instance
- Subscription: The subscription in which the ADF instance is created
- Resource group: The resource group where the ADF instance will reside
- Version: select V2 for the latest features
- Location: The datacenter location in which the instance is stored

Azure Data Factory security

- To create Data Factory instances, the user account *must be a member of the contributor or owner role, or an administrator of the Azure subscription.*
- To create and manage Data Factory objects including datasets, linked services, pipelines, triggers, integration runtimes and manage resources with PowerShell or the SDK you *must belong to the Data Factory Contributor role at the resource group level or above.*
- Data Factory Contributor role gives following permissions :
  - Create, edit, and delete data factories and child resources including datasets, linked services, pipelines, triggers, and integration runtimes.
  - Deploy Resource Manager templates. Resource Manager deployment is the deployment method used by Data Factory in the Azure portal.
  - Manage App Insights alerts for a data factory.
  - At the resource group level or above, lets users deploy Resource Manager template.
  - Create support tickets.

Create linked services - Demo

- Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.

- There are over 100 connectors that can be used to define a linked service.

- Two Types:

  - Data Stores - **Example - Azure SQL Database, Azure Blob Storage, etc.**

  - Compute Resources - **Example - Spark Cluster**

Create Datasets - Demo

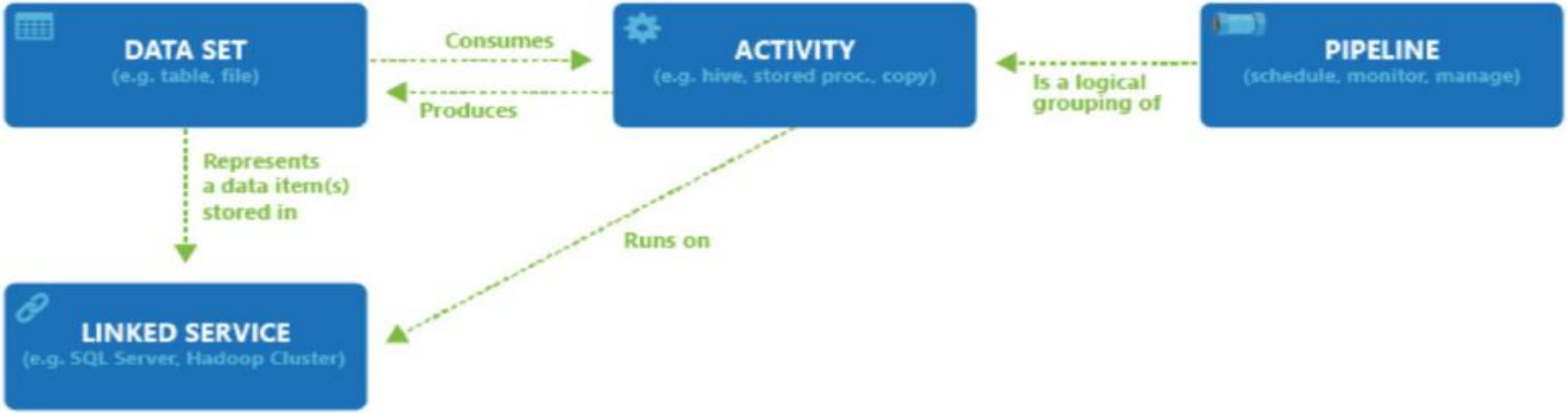
- Its a named view of data that simply points or references the data you want to use in your activities as inputs and outputs.

- Datasets identify data within different data stores, such as tables, files,folders, and documents. For example, an Azure Blob dataset specifies the blob container and folder in Blob storage from which the activity should read the data.

- A dataset in Data Factory can be defined as an object within the Copy Data Activity, as a separate object,

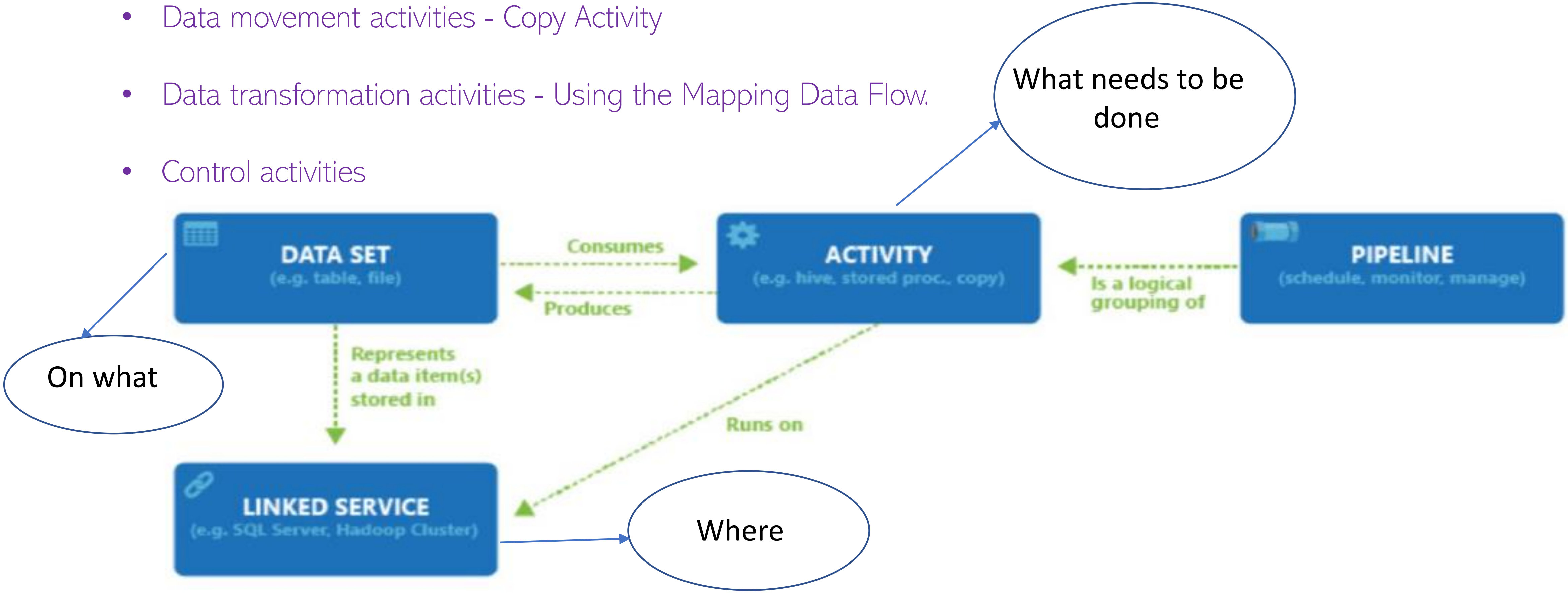
Data factory activities and pipelines – Demo

- Data movement activities - Copy Activity
- Data transformation activities - Using the Mapping Data Flow.
- Control activities


DATA SET
(e.g. table, file)
Consumes
Produces
ACTIVITY
(e.g. hive, stored proc., copy)
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)

Data factory activities and pipelines – Demo

- Data movement activities - Copy Activity
- Data transformation activities - Using the Mapping Data Flow.
- Control activities



What needs to be done

DATA SET
(e.g. table, file)

Consumes

Produces

ACTIVITY
(e.g. hive, stored proc., copy)

Is a logical grouping of

PIPELINE
(schedule, monitor, manage)

On what

Represents a data item(s) stored in

Runs on

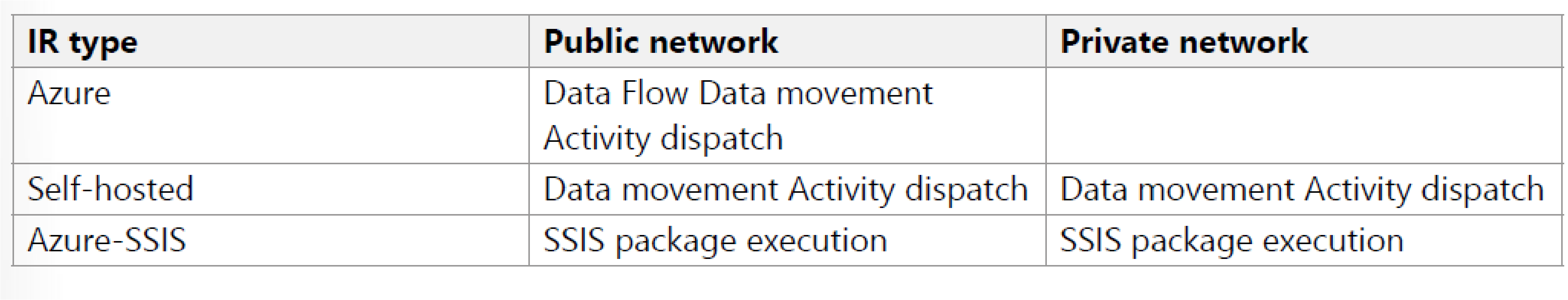
LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Where

Integration runtime types

- Three type:
  - Azure
  - Self-hosted
  - Azure-SSIS

- We can explicitly define the Integration Runtime setting in the *connectVia* property, if this is not defined, then the default Integration Runtime is used with the property set to *Auto-Resolve*.



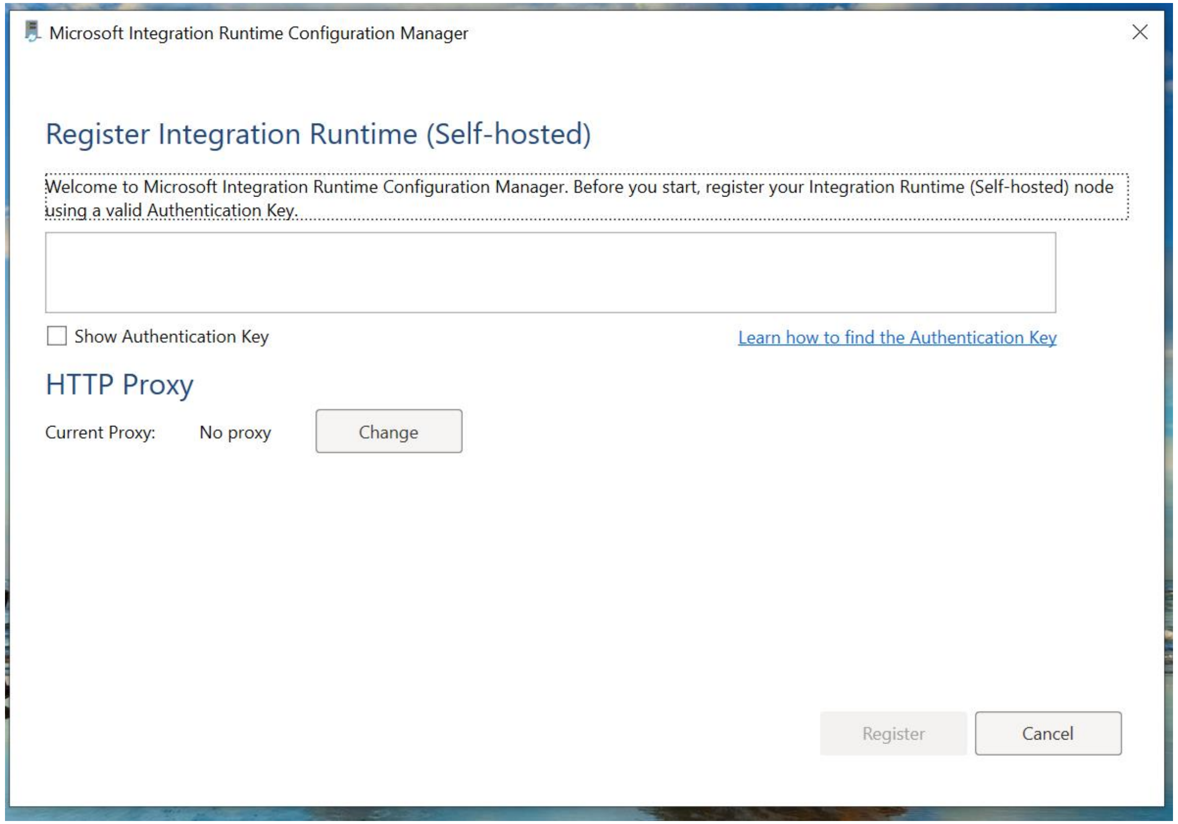
IR type	**Public network**	**Private network**
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Data Integration Capabilities

It provides the following data integration capabilities across different network environments, including:

- Data Flow: Execute a Data Flow in managed Azure compute environment.
- Data movement: Copy data across data stores in public network and data stores in private network (on-premises or virtual private network). It provides support for built-in connectors, format conversion, column mapping, and performant and scalable data transfer.
- Activity dispatch: Dispatch and monitor transformation activities running on a variety of compute services such as Azure Databricks, Azure HDInsight, Azure Machine Learning, Azure SQL Database,SQL Server, and more.
- SSIS package execution: Natively execute SQL Server Integration Services (SSIS) packages in a managed Azure compute environment.

Self Hosted Integration runtime



Microsoft Integration Runtime Configuration Manager

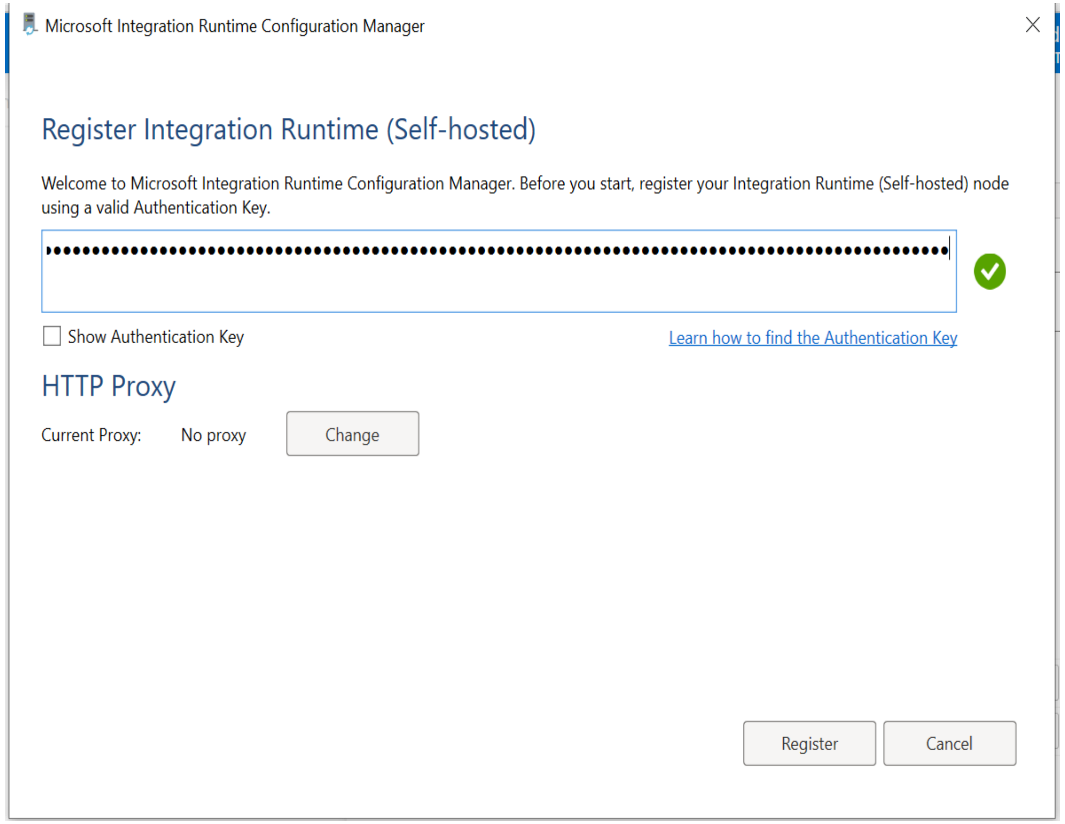
Register Integration Runtime (Self-hosted)

Welcome to Microsoft Integration Runtime Configuration Manager. Before you start, register your Integration Runtime (Self-hosted) node using a valid Authentication Key.

□ Show Authentication Key

HTTP Proxy

Current Proxy: No proxy



Microsoft Integration Runtime Configuration Manager

Register Integration Runtime (Self-hosted)

Welcome to Microsoft Integration Runtime Configuration Manager. Before you start, register your Integration Runtime (Self-hosted) node using a valid Authentication Key.

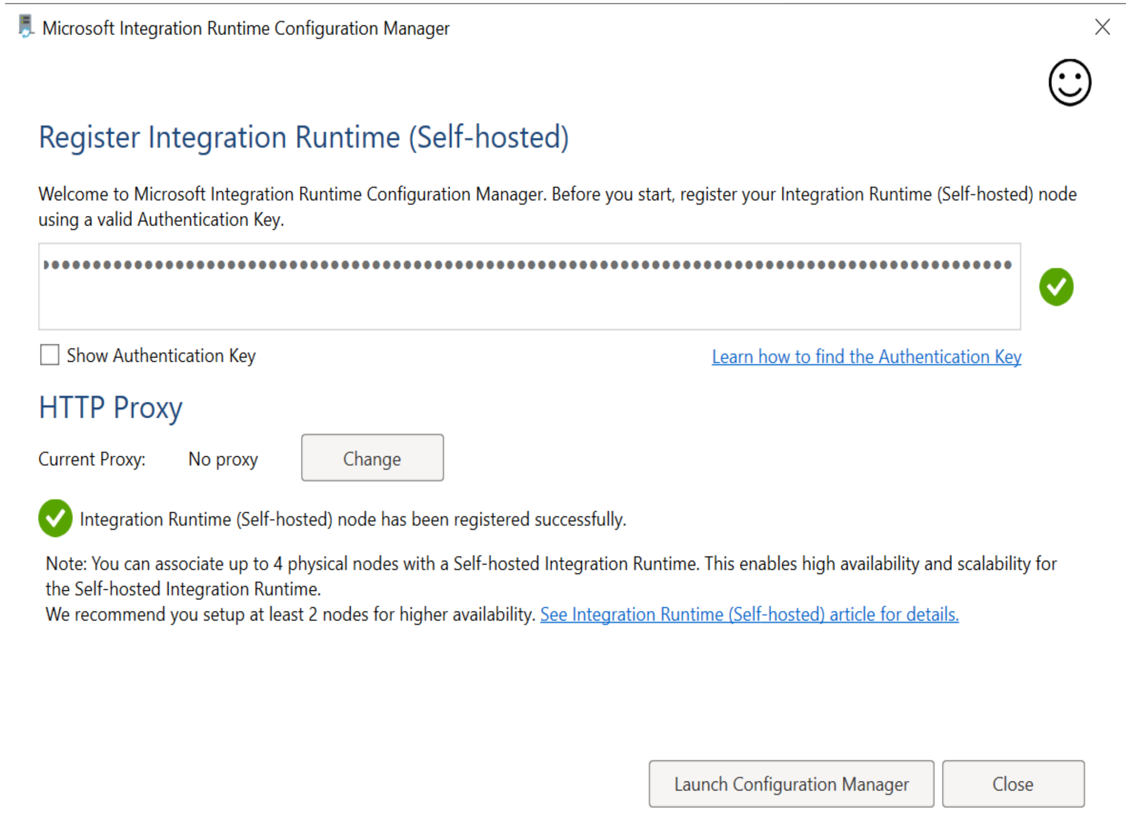
•• ✓

□ Show Authentication Key

HTTP Proxy

Current Proxy: No proxy

Self-Hosted Integration runtime



Microsoft Integration Runtime Configuration Manager

Register Integration Runtime (Self-hosted)

Welcome to Microsoft Integration Runtime Configuration Manager. Before you start, register your Integration Runtime (Self-hosted) node using a valid Authentication Key.

••• ✓

□ Show Authentication Key

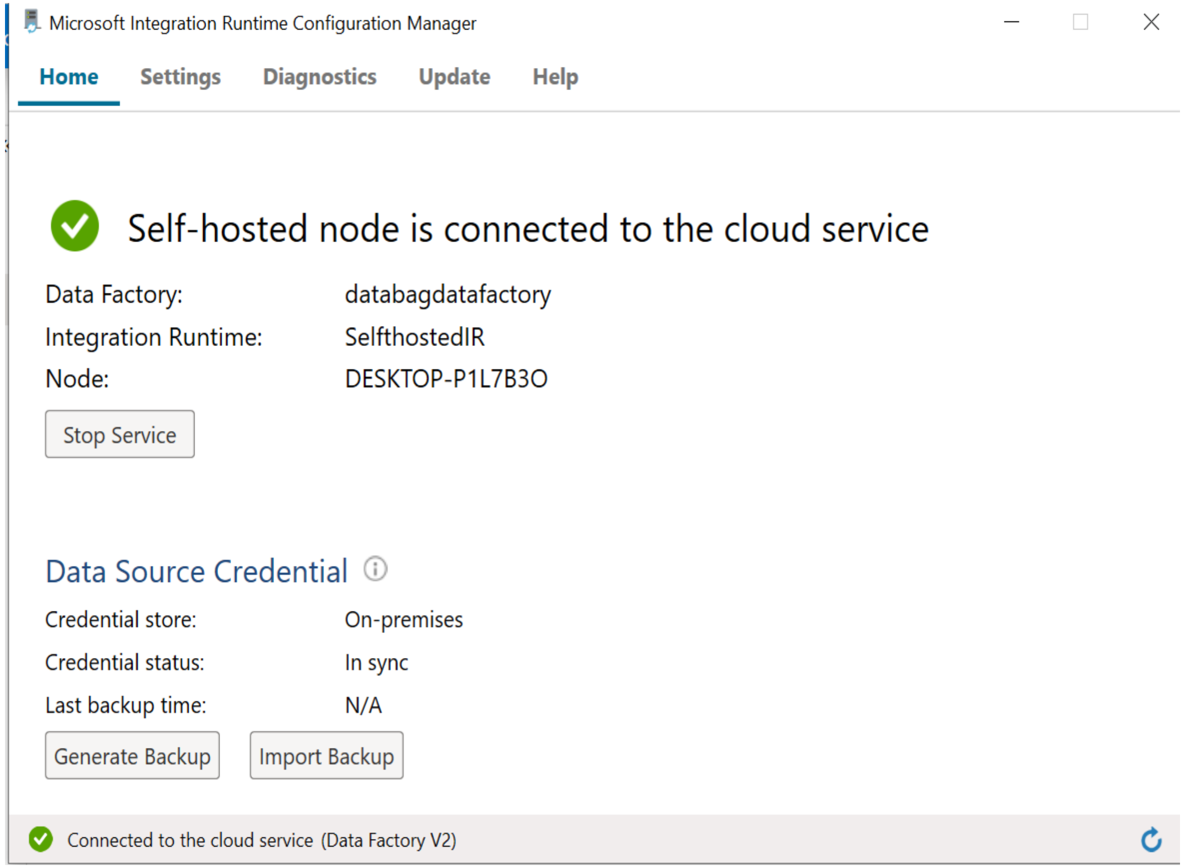
Learn how to find the Authentication Key

HTTP Proxy

Current Proxy: No proxy [Change]

✓ Integration Runtime (Self-hosted) node has been registered successfully.

Note: You can associate up to 4 physical nodes with a Self-hosted Integration Runtime. This enables high availability and scalability for the Self-hosted Integration Runtime.
We recommend you setup at least 2 nodes for higher availability. See Integration Runtime (Self-hosted) article for details.



Microsoft Integration Runtime Configuration Manager

✓ Self-hosted node is connected to the cloud service

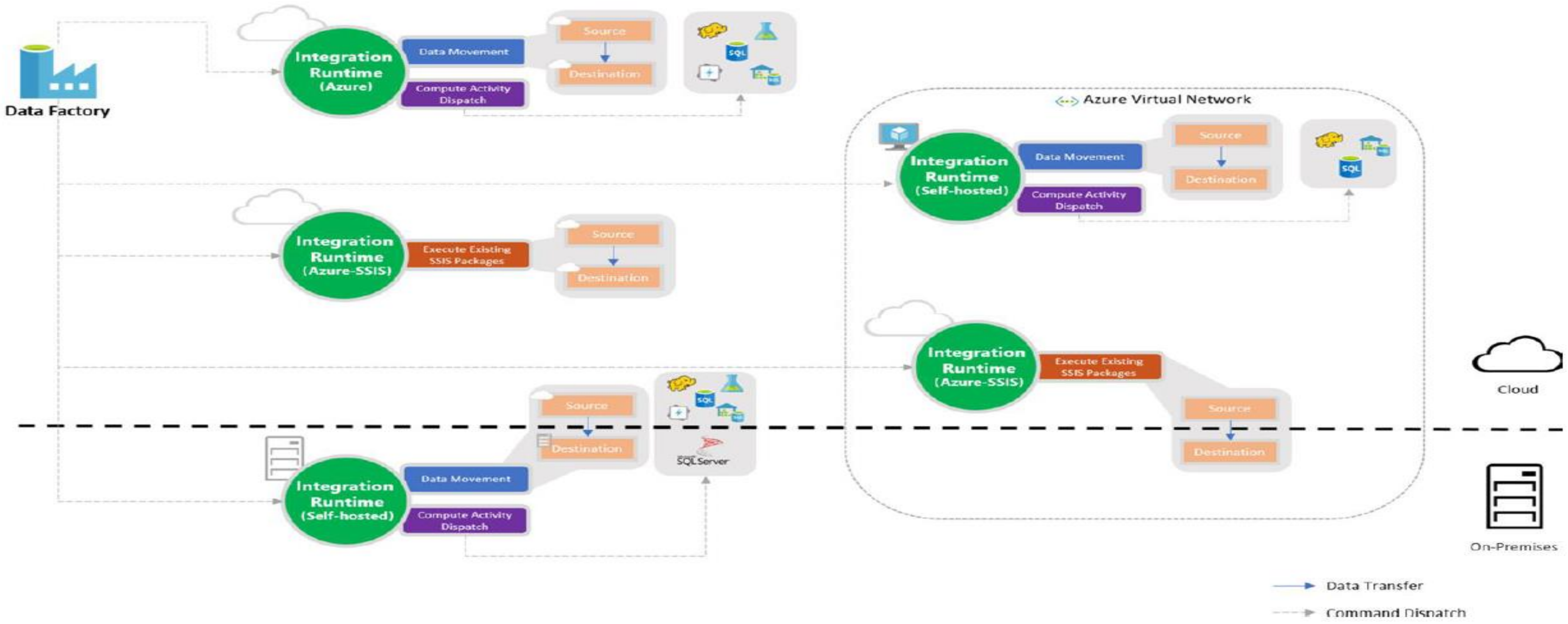
Data Factory:	databagdatafactory
Integration Runtime:	SelfhostedIR
Node:	DESKTOP-P1L7B3O

Data Source Credential

Credential store:	On-premises
Credential status:	In sync
Last backup time:	N/A

✓ Connected to the cloud service (Data Factory V2)

Which Integration runtime to use?


Data Factory
Integration Runtime (Azure)
Data Movement
Compute Activity Dispatch
Source
Destination
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
Source
Destination
SQL Server
Azure Virtual Network
Integration Runtime (Self-hosted)
Data Movement
Compute Activity Dispatch
Source
Destination
Integration Runtime (Azure-SSIS)
Execute Existing SSIS Packages
Source
Destination
Cloud
On-Premises
Data Transfer
Command Dispatch

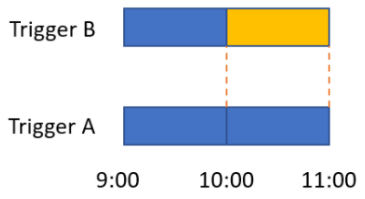
Triggers

- It is used to execute Pipeline

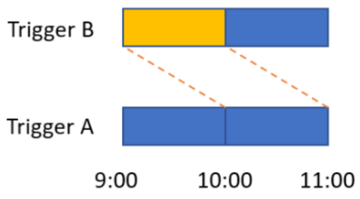
- 3 Different types of Triggers:
  - Schedule – Invoke pipeline based on minutes, hours, days, months
  - Tumbling Window – Here we can define dependencies
  - Event – Trigger pipeline in response to a event

Triggers

Dependency offset

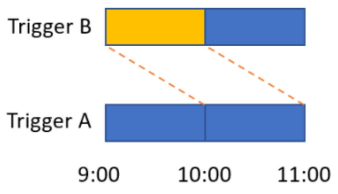


Dependency: A -> B
Offset: 0

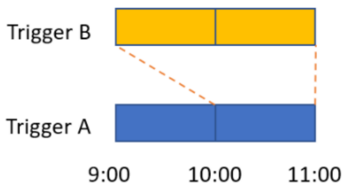


Dependency: A -> B
Offset: -1 hour

Dependency size

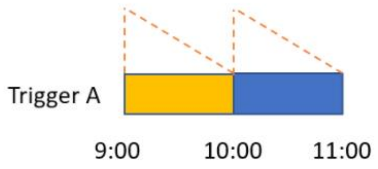


Dependency: A -> B
Offset: -1 hour
Size: unspecified

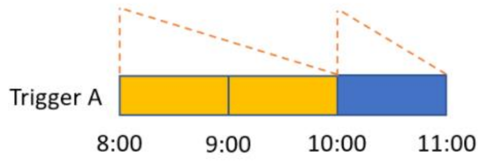


Dependency: A -> B
Offset: -1 hour
Size: 2 hours

Self-dependency

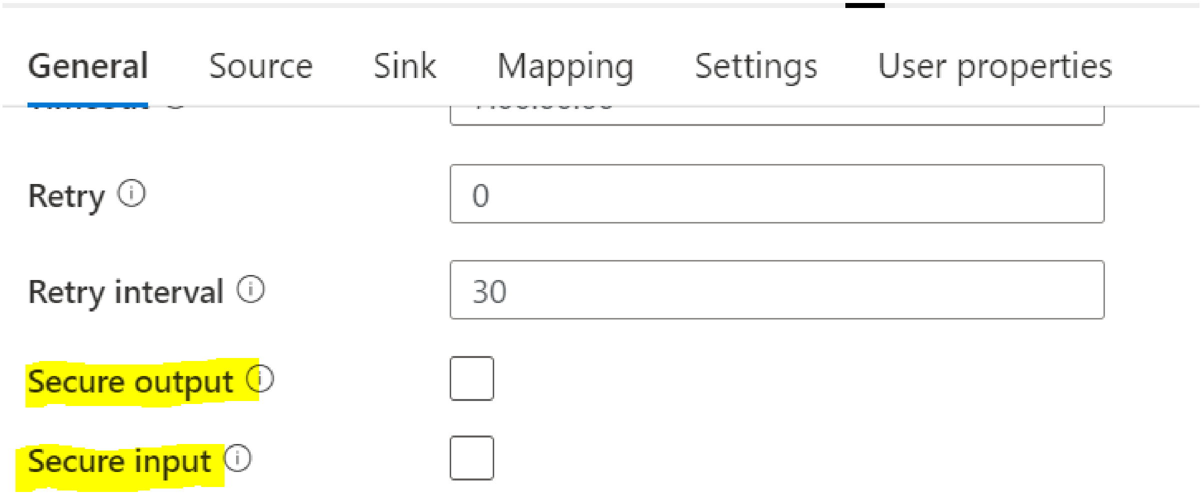


Dependency: A -> A
Offset: -1 hour
Size: unspecified



Dependency: A -> A
Offset: -2 hour
Size: 2 hours

Secure Input and Output,User Properties & Parameters -Demo



Retry ⓘ 0

Retry interval ⓘ 30

Secure output ⓘ □

Secure input ⓘ □

Perform code-free transformation at scale with Azure Data Factory

Learning Objective:

- Use Mapping Data Flow
- Debug Mapping Data Flow
- Use Wrangling Data

Mapping Data Flow- Demo

- Using Mapping Data Flow
- Debug Mapping Data Flow
- Behind the scene Data Flow executes on Azure Databricks using Spark
- ADF internally handles all the code translations, spark optimizations and execution of transformation.

Wrangling Data Flow- Demo

- Using Power Query for Wrangling Data

Monitoring ADF

- Alerts
- Metrics
- Diagnostic Settings

Thank you!