Using Azure Data Factory for ETL Operations

Emilio Melo

PLURALSIGHT AUTHOR

www.cloudadvantage.tech

Module Overview

Runtimes, Linked Services, Datasets and Triggers

Demo: Building the Pipeline

Data Flow transformations

Integration Runtimes

Data Flow execution

Data Movement execution

Dispatch of Activities

SSIS Package execution

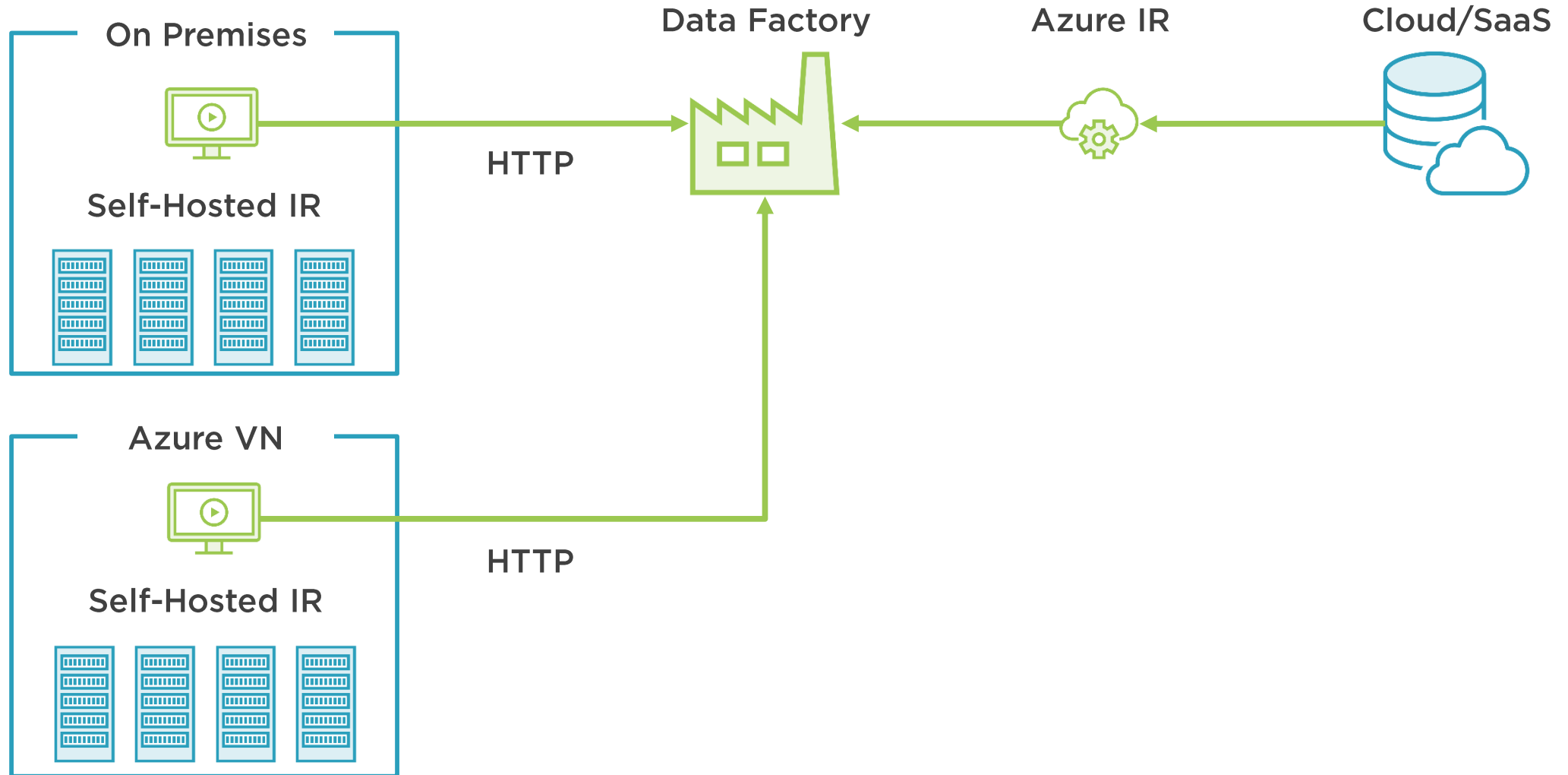# Azure Integration Runtime

Data movement between public endpoints

Only runtime for Data Flows

Serverless, consumption-based (DIUs)

AutoResolveIntegrationRuntime

Self-hosted Integration Runtime

On Premises

Self-Hosted IR

HTTP

Data Factory

Azure IR

Cloud/SaaS

Azure VN

Self-Hosted IR

HTTP

Azure-SSIS

Enables SSIS Packages on ADF

Lift-and-shift scenarios

SSIS Activities only
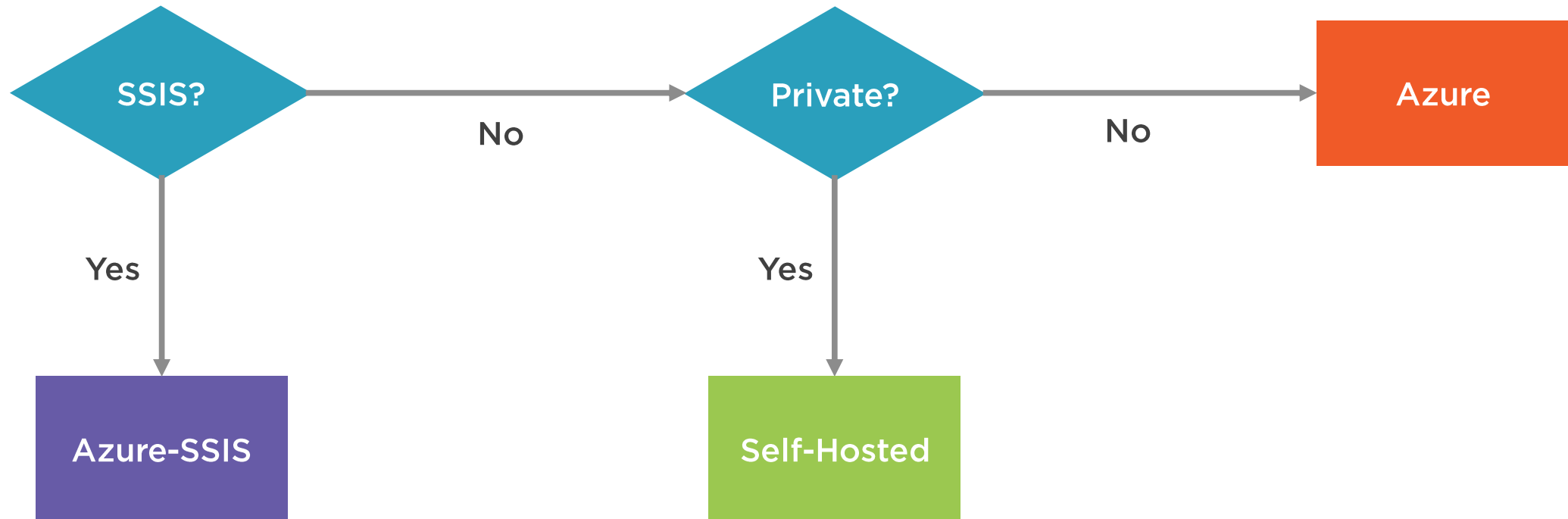
IaaS level (VM)

BYOL possible (AHUB)

Consider stopping IR when idle

SSISDB support on Azure SQL

Main Data Factory Elements

Linked Services and Datasets

Linked Services are like Connection Strings

Two types:

- Data Stores
- External Compute

Datasets are about data structure

Examples:

- SQL Database/Table
- Blob Storage/File

Triggers

Schedule	Tumbling Window	Event-based
Ideal for periodic packages	Ideal for time-sliced data	Fired based on Blob creation or deletion

Tumbling Window vs. Schedule

Tumbling Window	Schedule
Supports backfill scenarios	Forward-looking only
Runs at fixed intervals	Runs on dates configured
Retry, concurrency and system variables	"Fire and forget"
1-1 relationship with pipelines	Many-many relationship with pipelines
Use on time-series data	Use for periodic runs

Copy Activity

More than just copy

83 Connectors and counting

6 different formats

Enterprise Features

Parameters and Variables

Flexibility

Create once, customize as needed

System Variables

Out-of-the-box options for use

User variables

Create and operate your own

Data Flows

Transform options

SSIS, Data Flows and external compute

Treated as Activity

Just drag and drop to authoring canvas

Two Types

Mapping and Wrangling Data Flows

Only Azure IR

Data must be in Azure

Highly Scalable

Runs Spark Clusters on background

Module Summary

Creating and configuring:

- Integration Runtimes
- Triggers
- Activities
- Data Flows