



# Azure Data Factory v2

PASS Regionalgruppe Hamburg

# Agenda

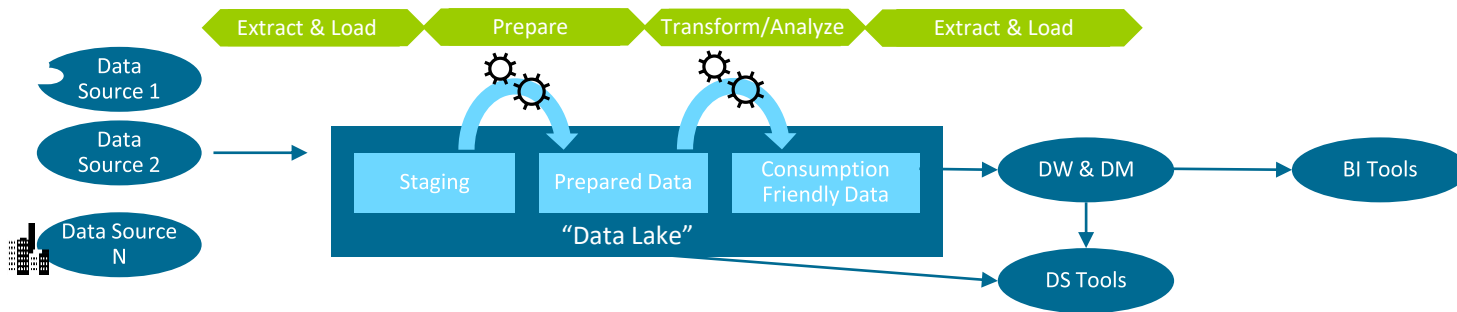


1. Target Scenarios
2. Current State
3. Intro Data Factory v2
4. Triggers
5. Control Flow
6. Integration Runtime
7. SSIS in ADFv2
8. Pricing
9. Roadmap & Q+A

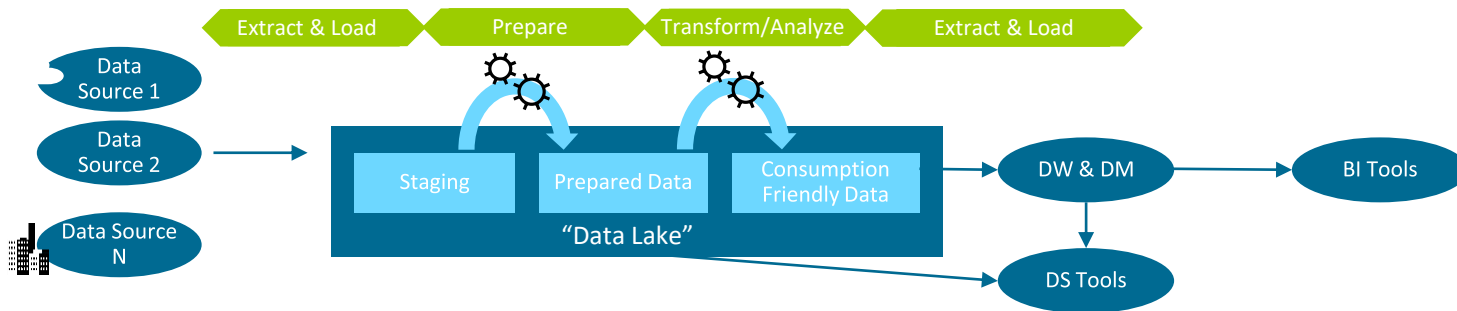
# Target Scenarios

for Data Integration in Azure

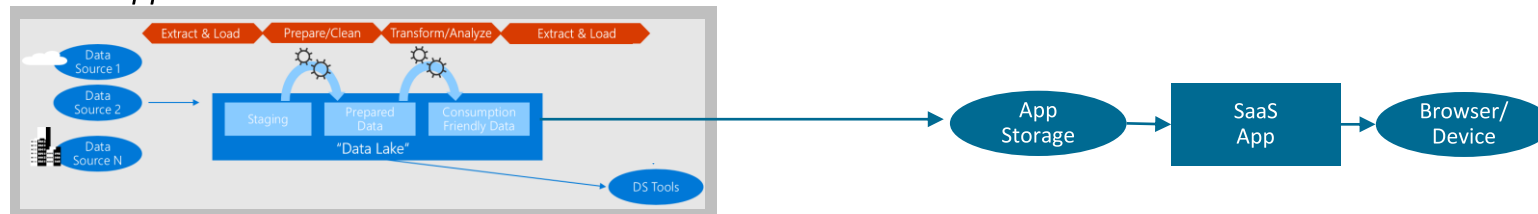
## Modern Data Warehouse



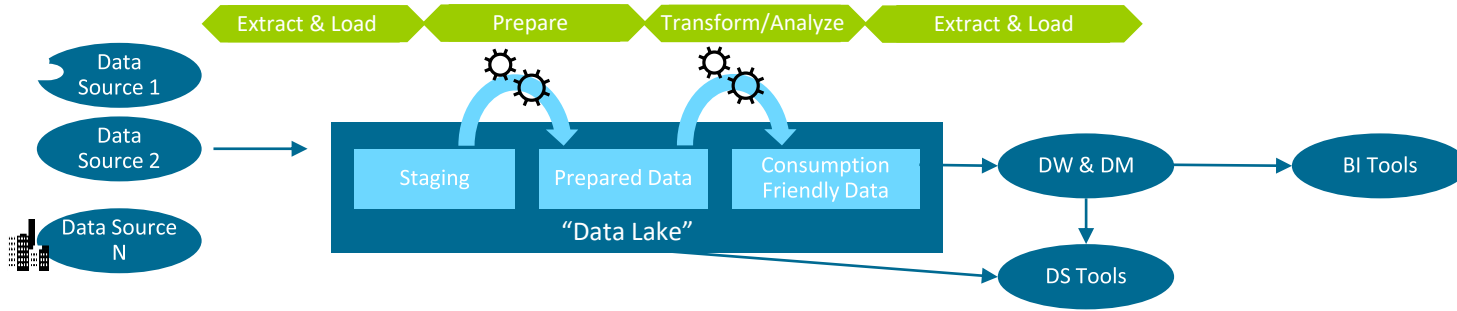
## Modern Data Warehouse



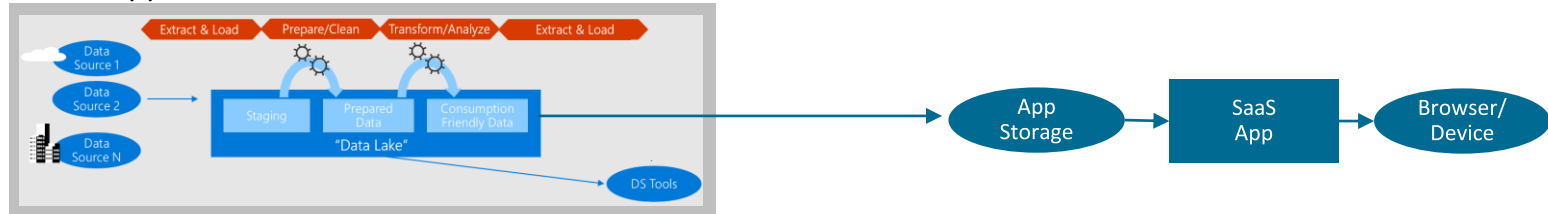
## Data-driven SaaS Application



## Modern Data Warehouse



## Data-driven SaaS Application



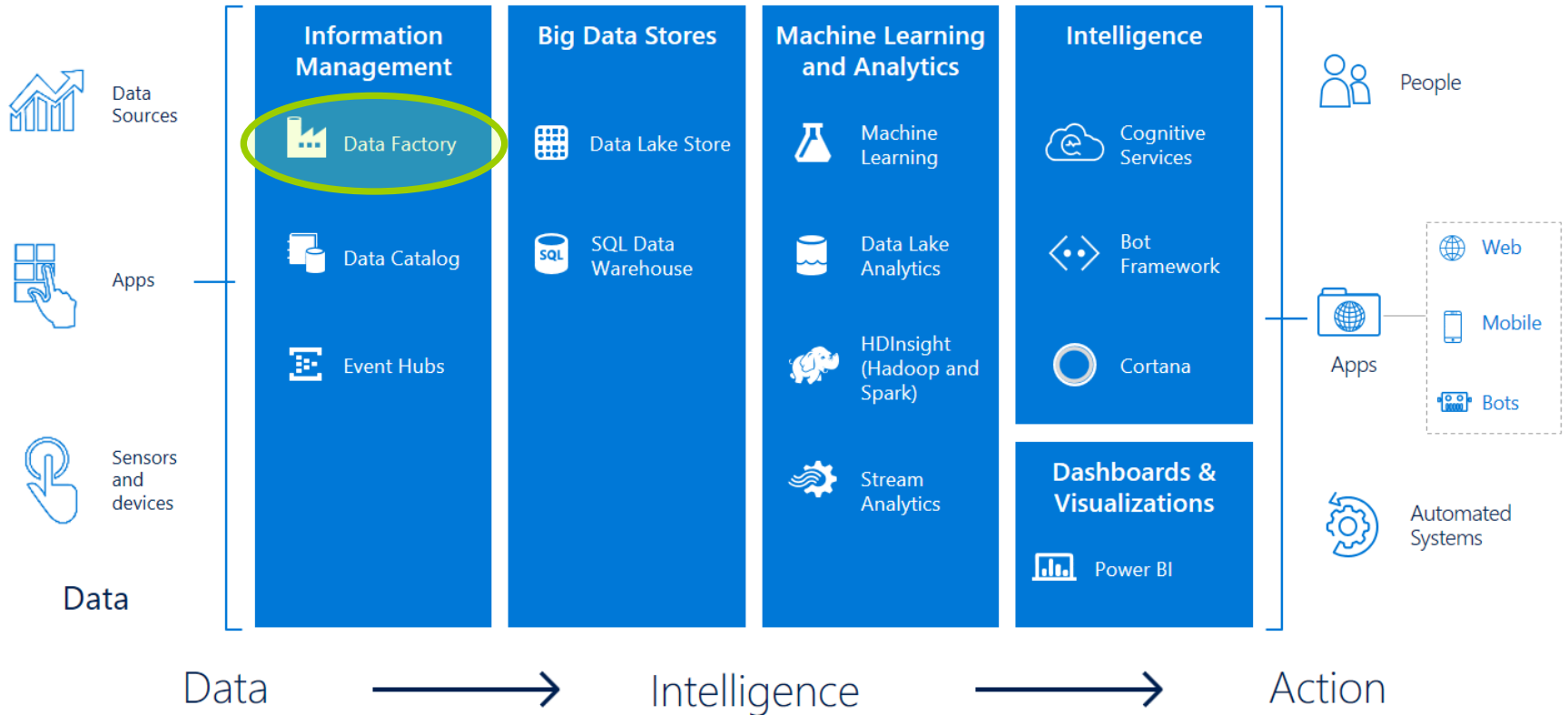
## Lift my existing SSIS packages to the cloud



# Current State

Data Management on the  
Microsoft Data Platform

# Part of Cortana Analytics Suite





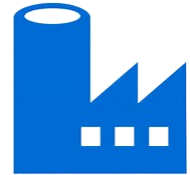
## Azure Data Factory (ADF)

Provides orchestration, data movement and monitoring services

Orchestration model: **time series processing**

**Hybrid Data movement as a Service** w/ many connectors

Programmatic authoring, visual monitoring (.NET, Powershell)



## SSIS: server software for ETL

Focused on ETL to/from SQL Server

Scale up data transformation engine

Visual authoring of control and data flow

Rich ecosystem (BIML, Task Libraries, etc)



- I need to create **on-demand** and/or **event-triggered** pipelines
- I need to create a **delta-processing** pipeline
- I need rich **orchestration constructs** to model my unique requirements (e.g. facilitate efficient restatements)
- I need to **reliably** work with **all my data across cloud, on prem, SaaS apps, etc**
  - **Data movement** at scale
  - Data pipelines spanning services, servers, cloud/onprem, etc
- ISVs: Need full **programmatic access**, using the **languages** and **runtimes** we are comfortable with
- My data integration team are **not developers**, I need **visual tools**
- **Productivity** of ETL is key, it is still 70%+ of overall solution time
- Scale to **1000's** pipelines
- How do I leverage my **existing SSIS investments** in the cloud?

# Data Factory v1 in Azure Portal

**Actions**

- Author and deploy
- Copy data (PREVIEW)
- Monitor & Manage
- Sample pipelines
- Diagram
- Metrics and operations

**Contents**

Datasets	Pipelines	Linked services
2	2	4
With errors: 0	Scheduled: 1 One time: 1	Data Stores: 4 Integration runtime: 0

[See more](#)

**Copy Data (hvwdatafactory)**

- Properties
- Source
- Destination
- Settings
- Summary
- Deployment

### Properties

Enter name and description for the copy data task and specify task.

Task name (required):

Task description:

Task cadence or Task schedule:

- Run once now
- Run regularly on schedule

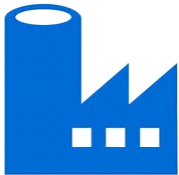


# Intro Data Factory v2

In Preview since Sep 2017

## Data Factory

*Managed Data Integration Service*



## New Pipeline Model

Rich pipeline orchestration

Triggers – on-demand, schedule, event

## Data Movement as a Service

Cloud, Hybrid

30 connectors provided

## SSIS Package Execution

In a managed cloud environment

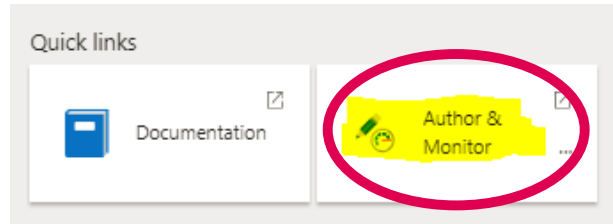
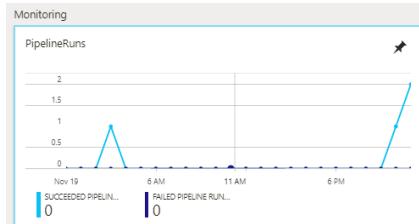
Use familiar tools, SSMS & SSDT

## Author & Monitor

Programmability (Python, .NET, Powershell, etc)

Visual Tools (coming soon)

# Data Factory v2 in Azure Portal



inovexdatafactory | Monitor Pipeline Runs ▾

Refresh

Last 24 Hours 11/19/2017 12:01 AM - 11/20/2017 12:01 AM ▾

Time Zone (UTC+01:00) Berlin ▾

All Succeeded In Progress Failed

Pipeline Name ▾	Actions	Run Start ▾	Duration	Triggered By	Status	Parameters
DemoCfForEachPipeline		11/19/2017, 11:21:33 PM	00:00:43	Manual	✔ Succeeded	@
DemoCfForEachPipeline		11/19/2017, 11:18:27 PM	00:00:43	Manual	✔ Succeeded	@
DemoCfForEachPipeline		11/19/2017, 10:22:11 PM	00:00:46	Manual	✔ Succeeded	@
DemoIfConPipeline		11/19/2017, 2:59:59 AM	00:00:24	ScheduleTrigger	✔ Succeeded	@

Since 15th  
Jan 2018!

SUPPORT + TROUBLESHOOTING

- Diagnose and solve problems >
- Activity log >
- New support request >

GENERAL

- Properties >

GETTING STARTED

- Quick start >

RESOURCE MANAGEMENT

- Tags >
- Locks >
- Users >



Azure Data Factory

# Let's get started



Create pipeline



Copy Data



Configure SSIS Integration Runtime



Configure Git Repository

## Overview



Overview Video



Introduction to Data Factory



Lift & shift SSIS packages



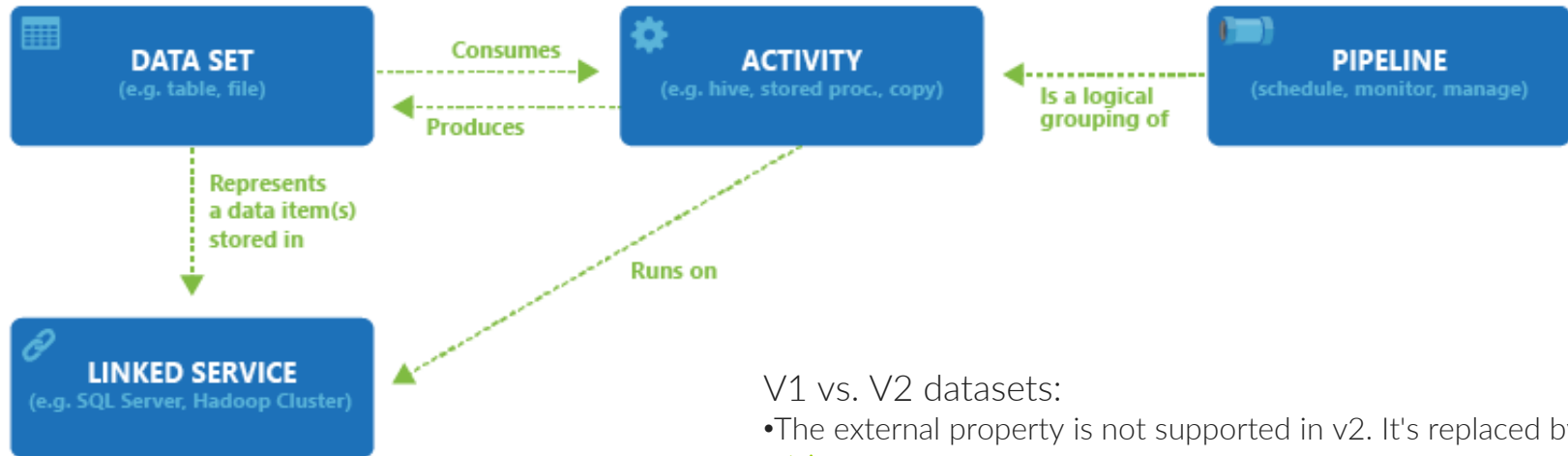
# Demo ADFv2

## Azure Portal – look and feel



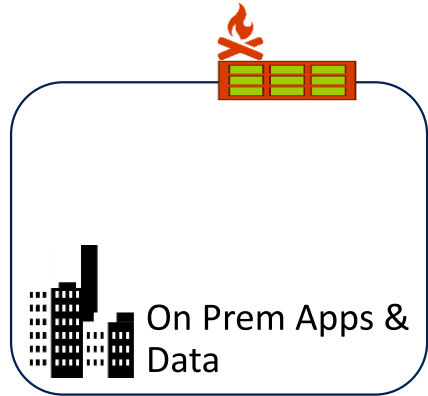
# Data Factory Essentials

## Artefacts in Data Factory



V1 vs. V2 datasets:

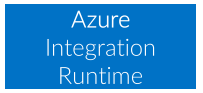
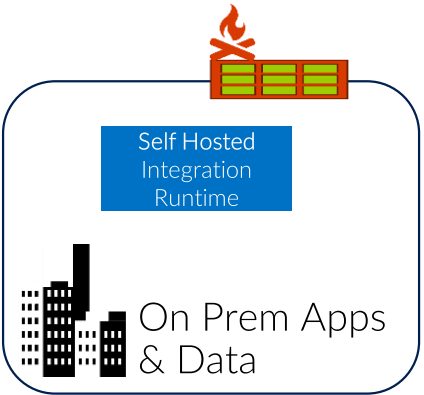
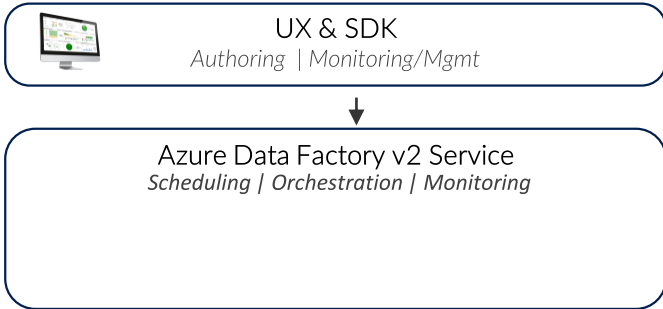
- The external property is not supported in v2. It's replaced by a [trigger](#).
- The policy and availability properties are not supported in V2. The start time for a pipeline depends on [triggers](#).
- Scoped datasets (datasets defined in a pipeline) are not supported in V2.



Cloud Svcs, Apps &  
Data

←--→ Command and Control

↔ Data



Cloud Svcs, Apps & Data

# Data Factory

A data integration account  
Location of orchestration, service metadata

# Integration Runtime (IR)

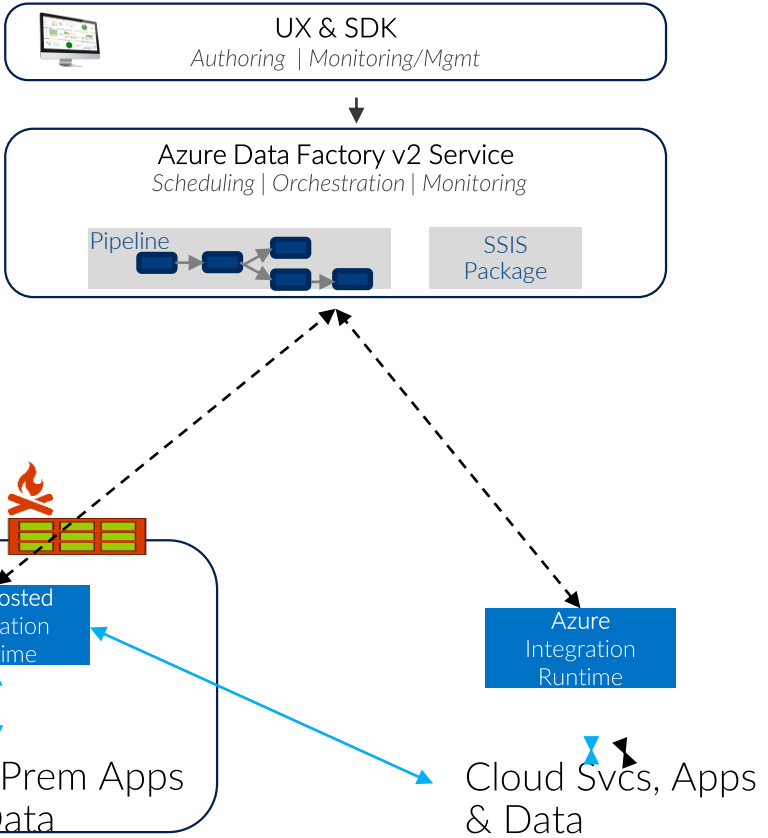
ADF's execution engine

Three core capabilities:

- data movement
- pipeline activity execution
- SSIS package execution

←--→ Command and Control

↔ Data



## Data Factory

A data integration account.  
Location of orchestration, service metadata

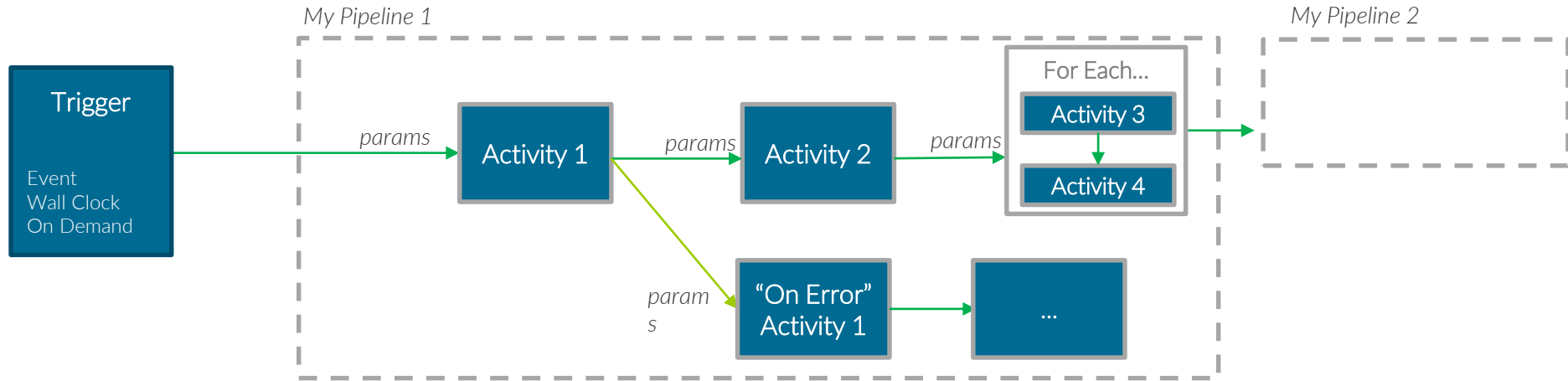
## Integration Runtime (IR)

ADF's execution engine

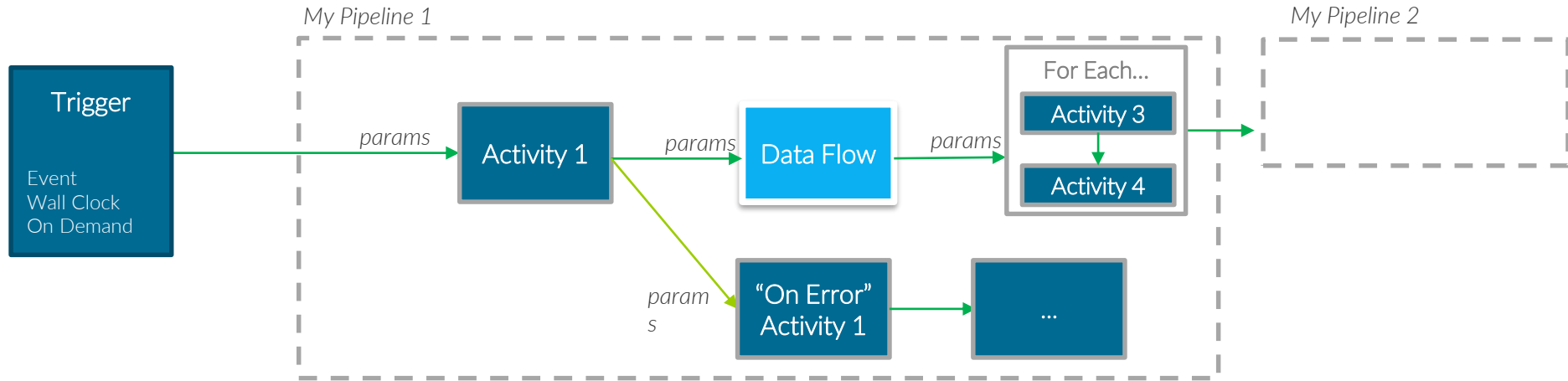
Three core capabilities:

- data movement
- pipeline activity execution
- SSIS package execution

# ADFv2 Pipelines



# ADFv2 Pipelines



# Expressions & Parameters

## Getting dynamic using inline expressions

1. rich new set of custom inner syntax in JSON
2. parameters as first class citizens in the service to support expressions factory wide
3. @ symbol starts expressions: e.g.

`"name": "@parameters('password')"`

4. Different types of functions available:

String (substring..), Collection (union..), Logic (less than),  
Conversation (array..), math (add..), Date (addminutes..)

e.g.: `replace('the old string', 'old', 'new')`

# System variables

1. Variables could be used in expressions, 2 scopes:

- Pipeline Scoped:

e.g. `@pipeline().RunId` shows the ID of the specific pipeline run

- Trigger Scoped:

e.g. `trigger().startTime` shows the time when the trigger actually fired to invoke the pipeline run





# Data Movement aka “Copy Activity”

## Scalable

per job elasticity (cloud data movement units)

Up to 1 GB/s

Set parallelism of threads used in source and target

## Simple

Visually author or via code (Python, .Net, etc)

Serverless, no infrastructure to manage

Staged copy (compress/decompress in hybrid scenarios, SQL DW load using polybase, bypass firewall restrictions)

## Access all your data

30+ connectors provided and growing (cloud, on premises, SaaS)

Data Movement as a Service: 17 points of presence world wide

Self-hostable Integration Runtime for hybrid movement

# Development

How to develop in ADFv2?

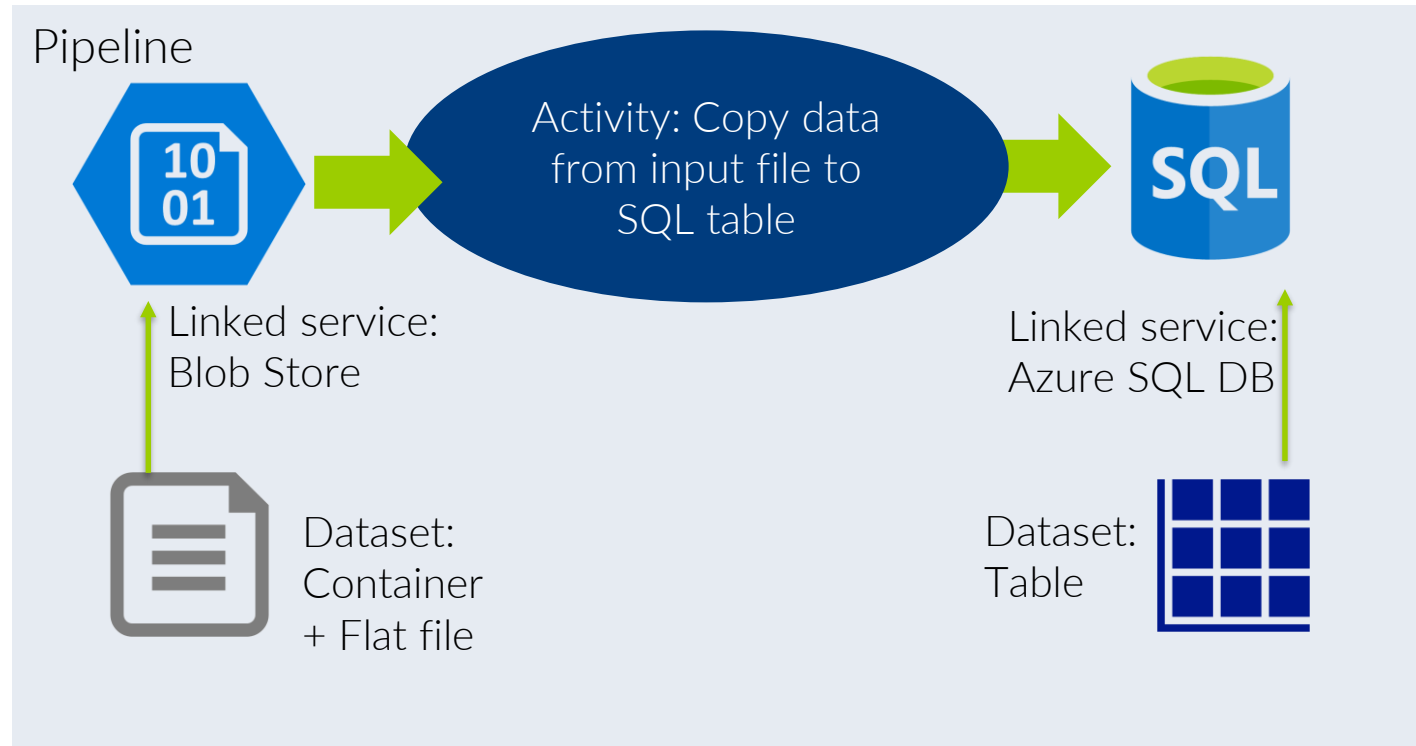
1. .net: Create ADF Objects and Deploy to ADFv2 .net using .net
2. PowerShell: Create ADF Objects and Deploy to ADFv2
3. Edit & PowerShell: Create ADF Objects per copy and paste and Deploy json artefacts using Powershell
4. **Visual Authoring in Azure Portal (since 15th Jan 2018!)**
5. Visual Studio Project (unkown)
6. BIML (private preview)

# Demo ADFv2

## Data Movement: Copy Activity

# Demo Data Movement

## Copy Activity



# Triggers

How to start a pipeline

# Triggers

How do pipelines get started

1. on-demand
2. Wall-clock Schedule
3. Tumbling Window (aka time-slices in v1)
4. *Event (not yet available)*

# Run pipeline on-demand

## 1. Power Shell:

Invoke-AzureRmDataFactoryV2Pipeline + Parameters

## 2. Rest API:

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

## 3. .NET:

client.Pipelines.CreateRunWithHttpMessagesAsync(+ parameters)

## 4. Azure Portal

Click

# Run pipeline by schedule

```
{
  "properties": {
    "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": "UTC"
      }
      "schedule": {                  // optional (advanced scheduling specifics)
        "hours": [<<0-24>>],
        "weekDays": "": [<<Monday-Sunday>>],
        "minutes": [<<0-60>>],
        "monthDays": [<<1-31>>],
        "monthlyOccurrences": [
          {
            "day": <<Monday-Sunday>>,
            "occurrence": <<1-5>>
          }
        ]
      }
    }
  }
}
```

Structure of Scheduler Trigger

```
{
  "properties": {
    "name": "MyTrigger",
    "type": "ScheduleTrigger",
    "typeProperties": {
      "recurrence": {
        "frequency": "Hour",
        "interval": 1,
        "startTime": "2017-11-01T09:00:00-08:00",
        "endTime": "2017-11-02T22:00:00-08:00"
      }
    },
    "pipelines": [{
      "pipelineReference": {
        "type": "PipelineReference",
        "referenceName": "SQLServerToBlobPipeline"
      },
      "parameters": {}
    },
    {
      "pipelineReference": {
        "type": "PipelineReference",
        "referenceName": "SQLServerToAzureSQLPipeline"
      },
      "parameters": {}
    }
  ]
}
```

Sample Scheduler Trigger



# Gain insights from ADLA pipeline & recurring jobs

## New

- Superset of original jobs view
- Adds grouping of jobs by pipelines & recurrences
- Jobs and consumption trends per pipeline
- Quickly identify pipelines and jobs to troubleshoot
- Quickly compare failed jobs with “last known good” instance
- Manage pipeline cost, improve efficiency and predict future cost

## How to use

- Create ADF v2 pipelines containing ADLA U-SQL activities
- Pipelines and Recurrences automatically appear in ADLA portal
- Submit and monitor pipeline/recurring jobs using Azure PowerShell, ADLA SDK and REST APIs

## New Pipeline Jobs View



All Jobs				Pipeline Jobs		Recurring Jobs	
PIPELINE JOBS	JOB	AU HOURS SUCC...	AU HOURS FAIL...				
Customer Categorization	161	10.357	3.634				
Customer Usage Trend Ana...	12	0.095	0.635				
Service Reliability Report	3	0.057	0				
Service Availability Report	3	0.069	0				
Billing Pipeline	3	0.057	0				
Service Demand Prediction	6	1.292	0.125				
ADFADLAE2Test2	3	0.106	0				
ADFADLAE2Test	5	0.091	0				

# Demo ADFv2 Scheduler Trigger

# Control Flow

# Activities

## Known from v1 - Data Transformation Activities

Data transformation activity	Compute environment
Hive	HDInsight [Hadoop]
Pig	HDInsight [Hadoop]
MapReduce	HDInsight [Hadoop]
Hadoop Streaming	HDInsight [Hadoop]
Spark	HDInsight [Hadoop]
Machine Learning activities: Batch Execution and Update Resource	Azure VM
Stored Procedure	Azure SQL, Azure SQL Data Warehouse, or SQL Server
U-SQL	Azure Data Lake Analytics

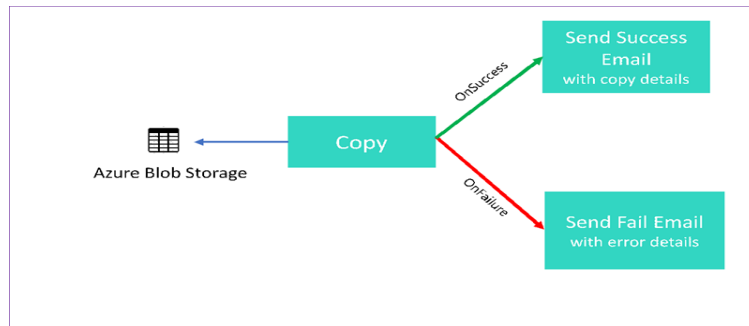
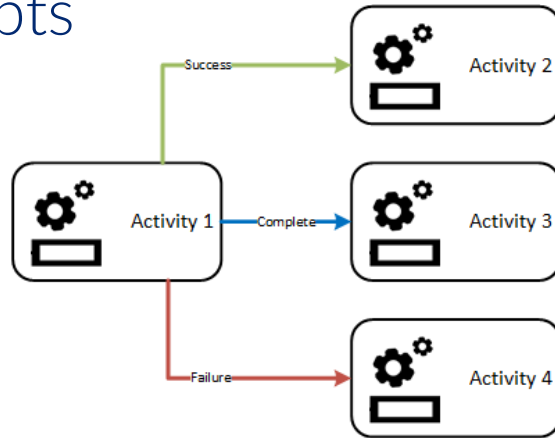
# Activities

## New! Control Flow Activities

Control activity	Description
Execute Pipeline Activity	allows a Data Factory pipeline to invoke another pipeline.
ForEachActivity	used to iterate over a collection and executes specified activities in a loop.
WebActivity	call a custom REST endpoint and pass datasets and linked services
Lookup Activity	look up a record/ table name/ value from any external source to be referenced by succeeding activities. Could be used for incremental loads!
Get Metadata Activity	retrieve metadata of any data in Azure Data Factory e.g. did another pipeline finish
Do Until Activity	similar to Do-Until looping structure in programming languages.
If Condition Activity	do something based on condition that evaluates to true or false.

# Activities

## Concepts



## Branching

Dependencies of activities in a pipeline

Possible constraints:

- On success
- On failure
- On completion
- On skip

Also custom 'if' conditions will be available for branching based expressions

# Visual authoring experience

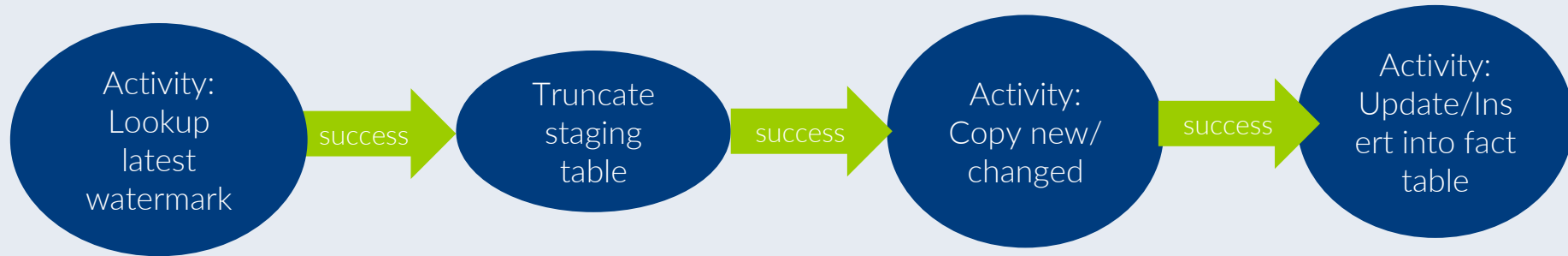
(only private preview)

1. Control Flow
2. Data Flow Column Mapping
3. Manage IR Runtimes (later more)

# Demos Control Flow

## Lookup Activity for delta load

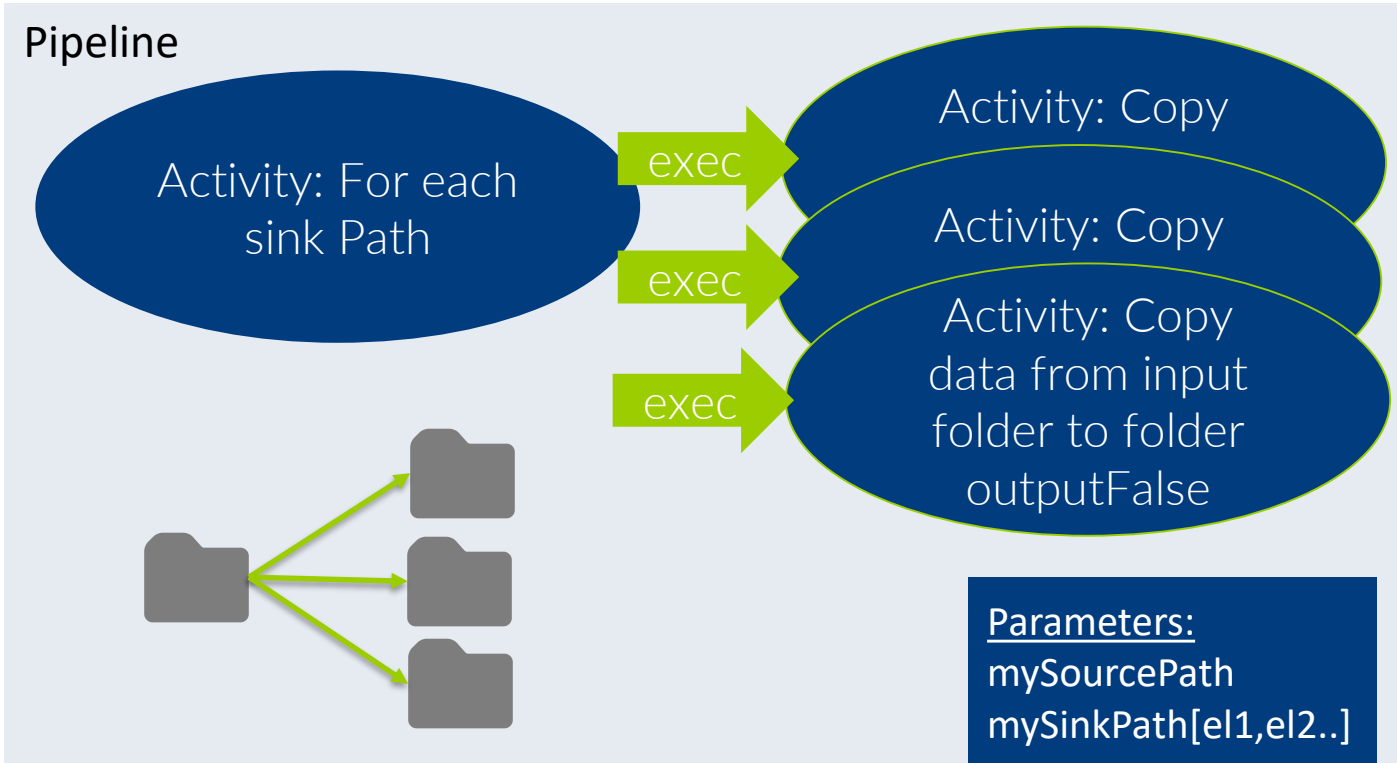
### Pipeline





# Demos Control Flow

## For Each Activity



# Integration Runtimes

3 types explained

# Integration runtime

## Different capabilities

### **1. Data Movement**

Move data between data stores, built-in connectors, format conversion, column mapping, and performant and scalable data transfer

### **2. Activity Dispatch**

Dispatch and monitor transformation activities (e.g. Stored Proc on SQL Server, Hive on HD Insight..)

### **3. SSIS package execution**

Execute SSIS packages

# Combinations of IR types, networks and capabilities

IR type	Public network	Private network
Azure (Default)	Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

# Integration runtimes

## 1. Azure Integration Runtime

- move data between cloud data stores
- fully managed
- serverless compute service (PaaS) on Azure
- cost will occur only for time of duration
- user could define data movement units
- compute size auto scaled for copy jobs

# Integration runtimes

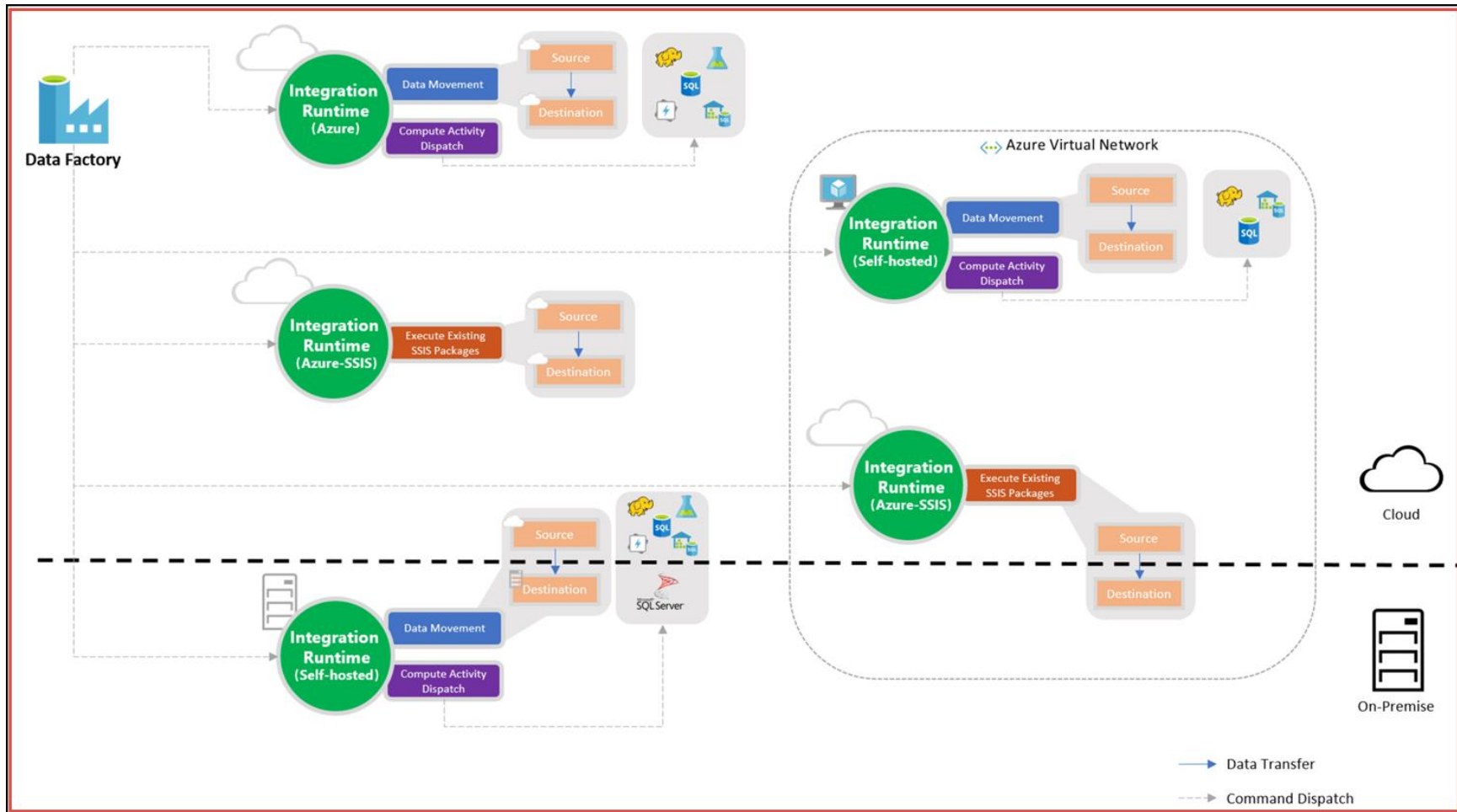
## 2. Self-hosted Integration Runtime

- perform data integration securely in a private network environment w/o direct line-of-sight from the public cloud environment
- Installed on-premises in your environment
- Supports copy activity between a cloud data stores and a data store in private network
- Supports dispatching the transform activities
- Works in your corporate network or virtual private network
- Only Outbound http based connections to open internet
- Scale out supported

# Integration runtimes

## 3. Azure-SSIS Integration Runtime

- fully managed cluster of Azure VMs for native execution of SSIS packages.
- Access to on-premises data access using Vnet (classic in preview)
- SSIS Catalog on Azure SQL DB or SQL Managed Instance
- scale up: set node size
- scale out: number of nodes
- reduce costs by start/stop of service

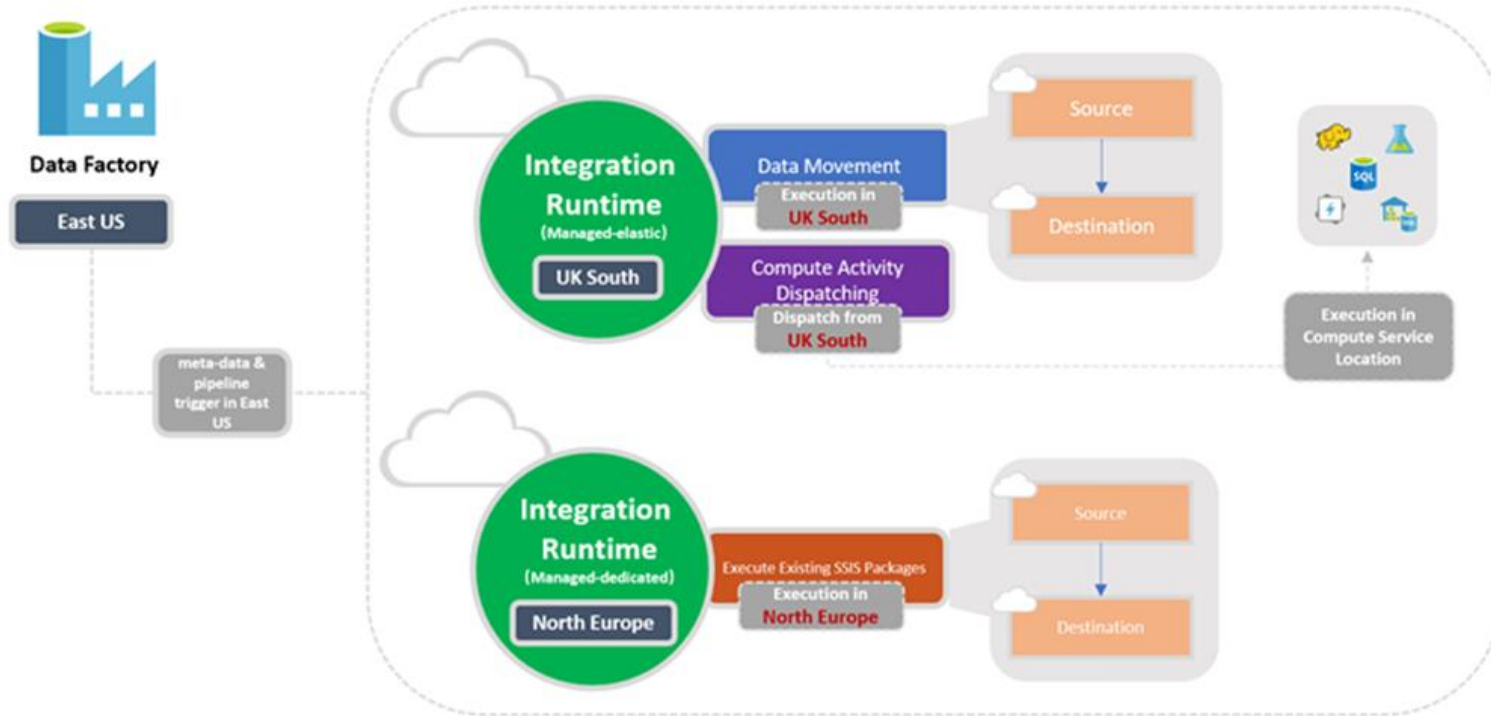




# Determining which IR to use

- IR is referenced over linked service in the data factory
- Transformation Activity: target compute needs linked service
- Copy Activity: source and sink need linked service, the computation IR is determined automatically (see details on msdn)
- Integration runtime locations can differ from its Data Factory location which uses it

# Samples ADF and IR locations



# Demo ADFv2 Integration Runtimes

# SSIS in ADFv2

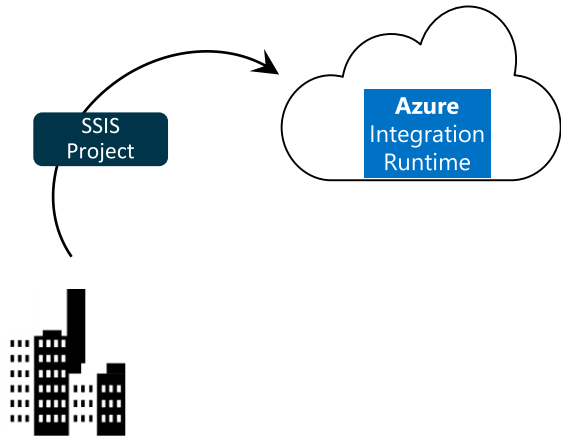
Lift existing SSIS projects to the cloud

# Managed Cloud Environment

Pick # nodes & node size

Resizable

SQL Standard Edition, Enterprise coming soon



## Compatible

Same SSIS runtime across Windows, Linux, Azure Cloud

## SSIS + SQL Server

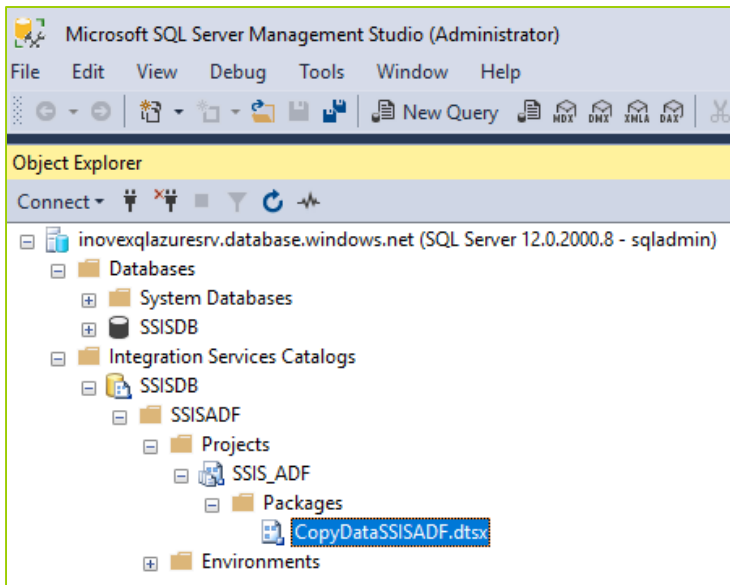
SQL Managed instance + SSIS (in ADFv2)

Access on premises data via VNet

## Get Started

Hourly pricing (no SQL Server license required)

Use existing license (coming soon)



## SSIS DB in SQL Azure

Integration Services Catalog in SQL Server Management Studio on PaaS DB

Execution Overview

Filter: Result: All: (3 more)

Result	Duration (sec)	Package Name	Task Name	Execution Path
Succeeded	17.297	CopyDataSSISADF.dtsx	CopyDataSSISADF	<a href="#">\CopyDataSSISADF</a>
Succeeded	17.234	CopyDataSSISADF.dtsx	Data Flow Task	<a href="#">\CopyDataSSISADF\Data Flow Task</a>

## Statistics as usual

Same Database structure and reports as on-premise SSIS Runtime

# Scaleable Integration Services

How to scale up/out using 3 Settings on Azure SSIS IR

1. Configurable number of nodes on which SSIS is executed

```
$AzureSSISNodeSize = "Standard_A4_v2" # minimum size, others avail.
```

2. Configurable size of nodes

```
$AzureSSISNodeNumber = 2 #between 1 and 10 nodes*
```

3. Configurable maximum parallel executions per node

```
$AzureSSISMaxParallelExecutionsPerNode = 2 # between 1-8*
```

Im Powershell CommandLet: Set-AzureRmDataFactoryV2IntegrationRuntime  
or use the gui in portal

# Notes from the field

1. Connect in SSMS directly to the DB SSISDB to see SSIS Catalog
2. Deploy from Visual Studio only in Project Deployment Mode, workaround SSMS Import for single packages
3. In Preview no SSIS 3rd Party components supported (e.g. Theobald for SAP, cozyroc..)
4. Use same location for Azure-SSIS IR and the used (SQL Azure) DB for SSIS Catalog



# Execution Methods

1. SSIS packages can be executed via SSMS
2. SSIS packages can be executed via CLI
  - › Run dtexec.exe from the command prompt (TBD)
3. SSIS packages can be executed via custom code/PSH using SSIS MOM .NET SDK/API
  - › Microsoft.SqlServer.Management.IntegrationServices.dll is installed in .NET GAC with SQL Server/SSMS installation
4. SSIS packages can be executed via T-SQL scripts executing SSISDB sprocs
  - › Execute SSISDB sprocs [catalog].[create\_execution] + [catalog].[set\_execution\_parameter\_value] + [catalog].[start\_execution]

# Scheduling Methods

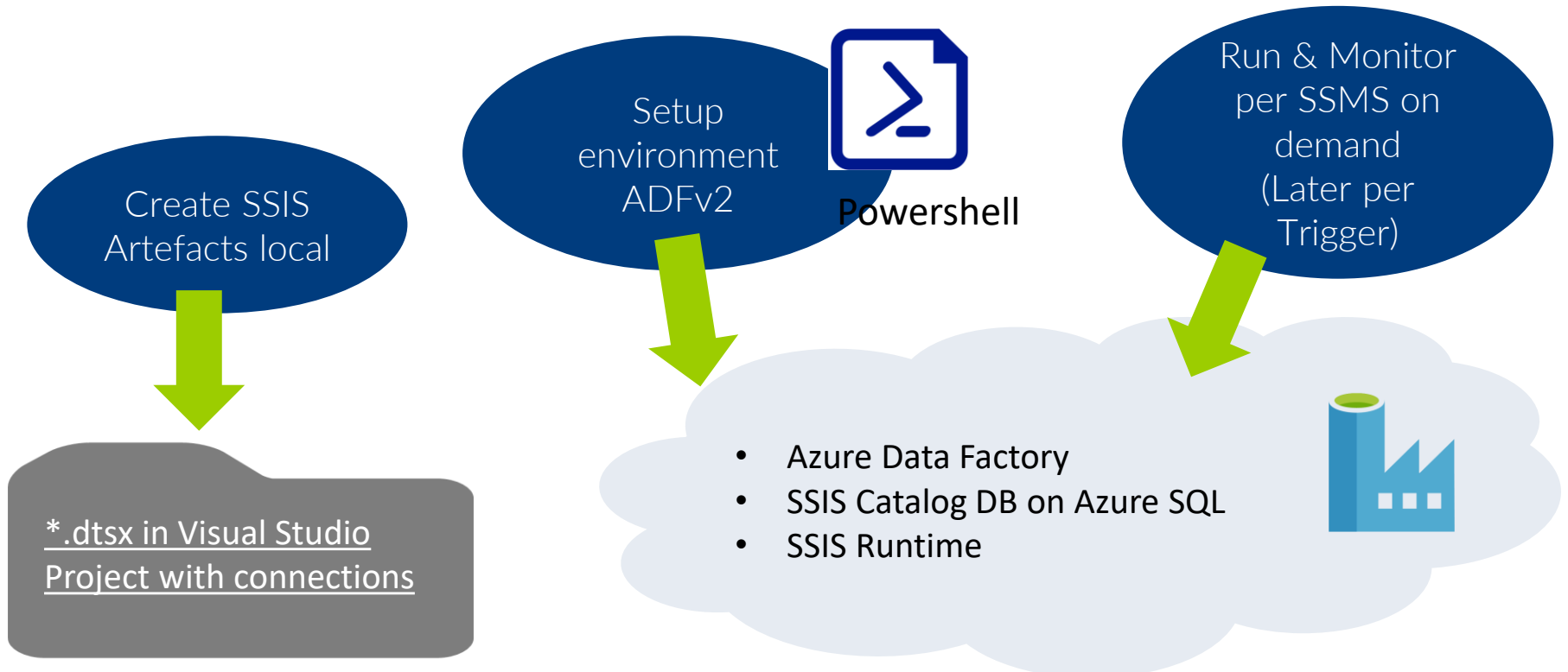
1. SSIS package executions can be directly/explicitly scheduled via ADFv2 App (Work in Progress)
  - › For now, SSIS package executions can be indirectly/implicitly scheduled via ADFv1/v2 Sproc Activity
2. If you use Azure SQL MI server to host SSISDB
  - › SSIS package executions can also be scheduled via Azure SQL MI Agent (Extended Private Preview)
3. If you use Azure SQL DB server to host SSISDB
  - › SSIS package executions can also be scheduled via Elastic Jobs (Private Preview)
4. If you keep on-prem SQL Server
  - › SSIS package executions can also be scheduled via on-prem SQL Server Agent

# Demo

## Integration Services in ADFv2

# SSIS Demo

Setup and manage environment, deploy packages



# Pricing

How much is the fish?

# Pricing

Different factors for billing

1. Number of activities run
2. Volume of data moved
3. SQL Server Integration Services (SSIS) compute hours
4. Whether a pipeline is active or not

# Pricing

## 1. Orchestration:

- › Activity runs in Azure IR :
- › 0,464 € per 1.000 runs / 0,422 € post 50.000 runs
- › Activity runs in Self-Hosted IR: 0,633 € per 1,000 runs

## 2. Volume in Data Movement

- › Azure IR: 0,106 € per hour
- › Self-hosted IR: 0,043 € per hour
- › + Outbound data transfer charges

# Pricing

## 3. Azure-SSIS Integration Runtime:

- › SSIS usage is charged by the hour
- › Depends on VM size
- › Min: "SSIS D1 v2" 182 €/ month (1 core, 3,5 GB RAM, 50 GB temp storage)
- › Max: "SSIS D4 v2" 737 €/month (8 cores, 28 GB RAM, 400 GB temp storage)
- › + SQL Azure DB costs for SSIS catalogue (e.g. Basic 4,2 €/month..)

## 4. Inactive pipelines

- › Pipelines not triggered and zero runs for a week
- › 0,338 € per month



# Roadmap & Q+A

# Roadmap

2018

- Visual experiences (data transform)
- Data movement (connectivity, scale, ... )
- Further SSIS integration
- Data Discovery

# Links and further informatoin

1. Microsoft documentation:

<https://docs.microsoft.com/en-us/azure/data-factory/>

2. Powershell cmdlets for ADF and ADFv2 for v5.0.0

[https://docs.microsoft.com/en-us/powershell/module/azurermdatadf/?view=azurermps-5.0.0&viewFallbackFrom=azurermps-5.0.0#data\\_factories](https://docs.microsoft.com/en-us/powershell/module/azurermdatadf/?view=azurermps-5.0.0&viewFallbackFrom=azurermps-5.0.0#data_factories)

3. Very good blog article about Azure Data Factory V2:

<https://www.purplefrogssystems.com/paul/2017/09/whats-new-in-azure-data-factory-version-2-adfv2/>

4. MS Ignite Sessions:

<https://myignite.microsoft.com/videos/55421>

<https://myignite.microsoft.com/sessions/55271?source=sessions>

inovex



inovex ist ein IT-Projekthaus  
mit dem Schwerpunkt „Digitale Transformation“:

Product Ownership · Datenprodukte  
Web · Apps · Smart Devices · BI  
Big Data · Data Science · Search  
Replatforming · Cloud · DevOps  
Data Center Automation & Hosting  
Trainings · Coachings

inovex gibt es in Karlsruhe · Pforzheim ·  
Stuttgart · München · Köln · Hamburg

Und natürlich unter [www.inovex.de](http://www.inovex.de)

Wir nutzen Technologien,  
um unsere Kunden glücklich zu machen.  
*Und uns selbst.*