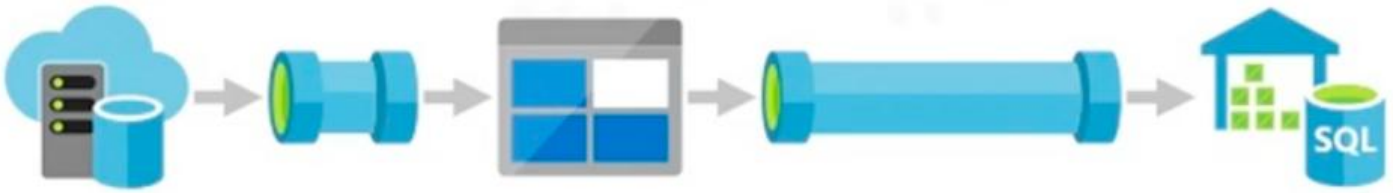
What can you do in Azure Data Factory?



Copy Data

Transform Data

Copy Data

More than 80 connectors to different services are available

Transform Data

Using newly added Data Flow, now Data Factory is complete cloud based ETL tool.

PowerPoint Slide Show - [DataFactory] - PowerPoint

Azure Data Factory

Definition:

Azure Data Factory (ADF) is a hybrid data integration service that enables you to quickly and efficiently create automated data pipelines – without having to write any code!

Slide 3 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint

Azure Data Factory

- Hybrid Data Integration Service
- Simplifies ETL at scale
- Enables modern data integration
- Drag and drop interface
- Over 80 connectors available
- Move, transform and save data
- Managed Service
- Create Data Driver workflows
- Orchestrate and automate data movement
- Transform and store data
- Operationalize the process
- ETL or ELT scenarious

Slide 4 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint

Data Factory on Azure Ecosystem

01 Migration?

Data Factory excels in periodic data loads and transformation instead.

02 Streaming?

ADF can orchestrate, but there are other dedicated services for streaming

03 Transformations?

Data flows for simple ones, but you can use Databricks or HDInsight for more complex transforms

PowerPoint Slide Show - [DataFactory] - PowerPoint

SSIS vs Data Factory

SSIS	Data Factory
More code-free transformations	Much higher scalability
On Premises connectors (e.g excel)	Cloud and SaaS Connectors
	Event based Triggers
	Can use SSIS Packages

Data Factory considerations

Two versions	Build options	Highly integrated	No data storage	Security standards
ADF V2 is the current and improved version	PowerShell, .Net, Python, REST, ARM	DevOps, Key Vault, Monitor, Automation	Need to persist data by the end.	HTTP/TLS whenever possible

Deploying Clusters — Dask docu | Dask: Status | python 3.x - ImportError: cannot | Introduction to Azure Data Facto

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

Bookmarks | Razorpay Dashboard | BitPaper - Features | Home | TL-WR740N | Privacy error | (55) DSC Error in G... | Manage excellatee... | GitHub - Pierian-Da... | Your Account | Learning on Simpli... | Reading list

for business intelligence (BI) applications to consume. Ultimately, through Azure Data Factory, raw data can be organized into meaningful data stores and data lakes for better business decisions.

How does it work?

Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers.

This visual guide provides a high-level overview of the Data Factory architecture:

Deploying Clusters — Dask docu | Dask: Status | python 3.x - ImportError: cannot | Introduction to Azure Data Factor

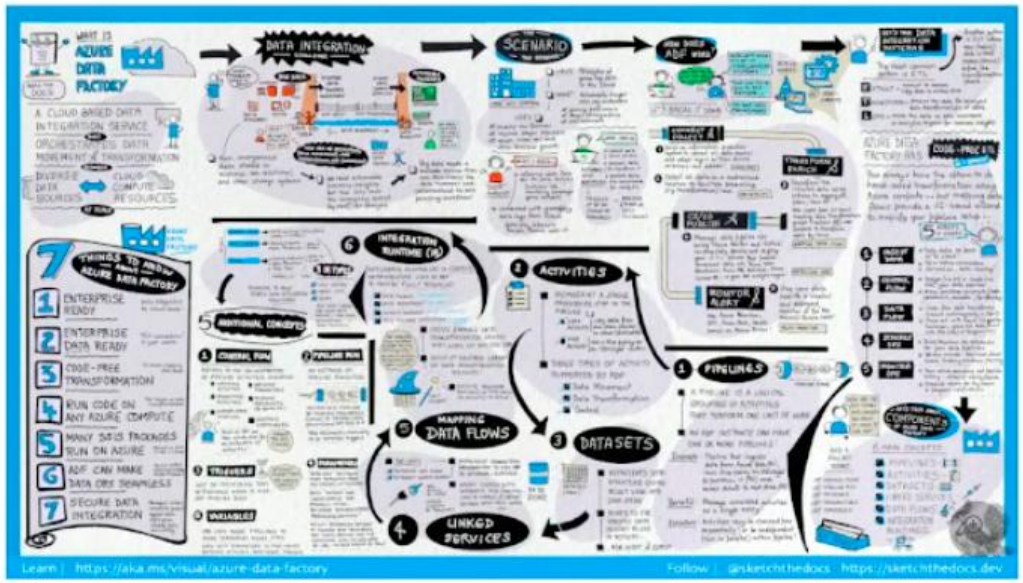
https://docs.microsoft.com/en-us/azure/data-factory/introduction

Bookmarks | Razorpay Dashboard | BitPaper - Features | Home | TL-WR740N | Privacy error | (55) DSC Error in G... | Manage excellatee... | GitHub - Pierian-Da... | Your Account | Learning on Simpli... | Reading list

How does it work?

Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers.

This visual guide provides a high-level overview of the Data Factory architecture:



To see more detail, click the preceding image to zoom in, or browse to the high resolution image.

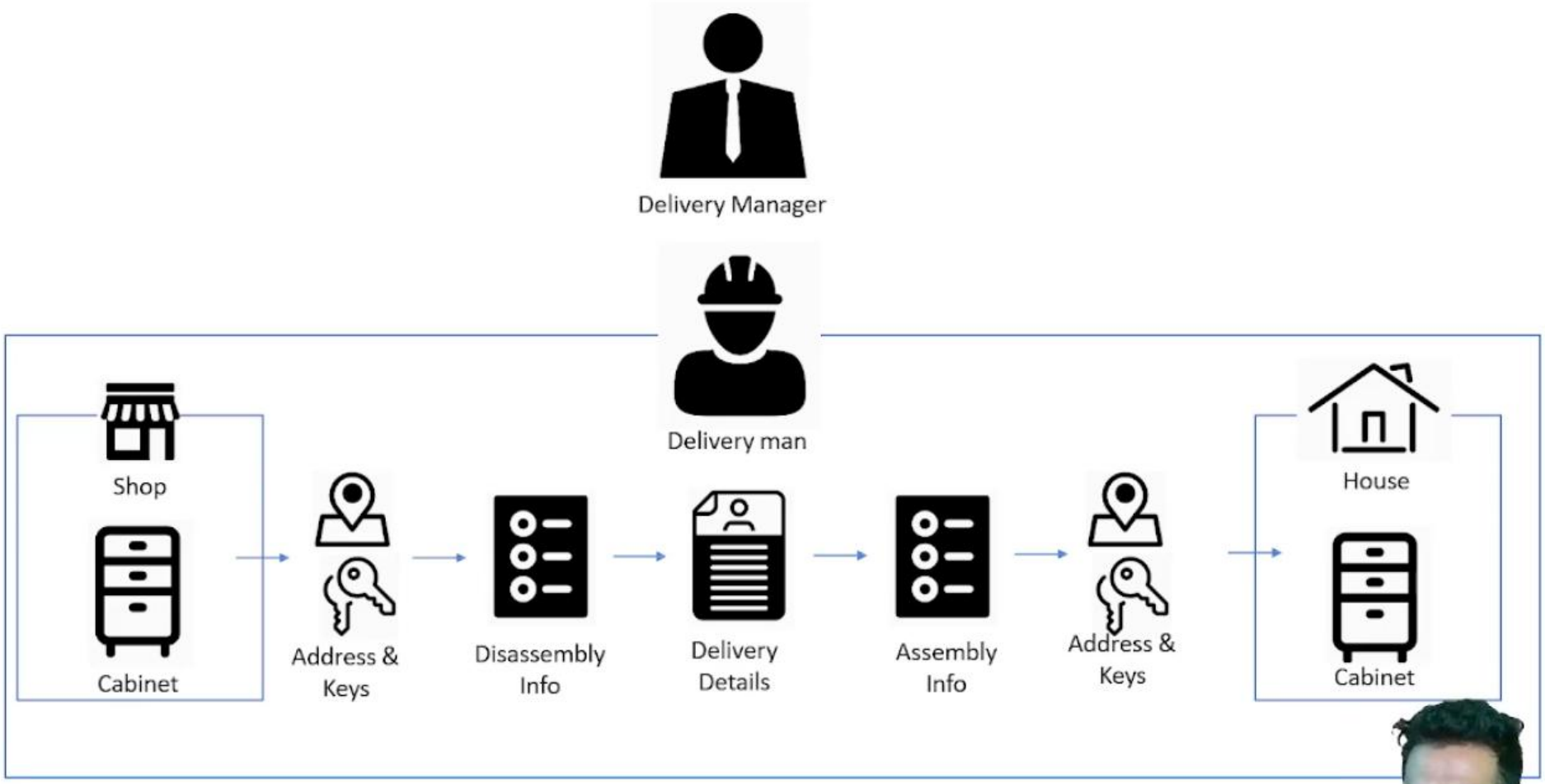
Connect and collect

PowerPoint Slide Show - [DataFactory] - PowerPoint

Azure Data Factory Components

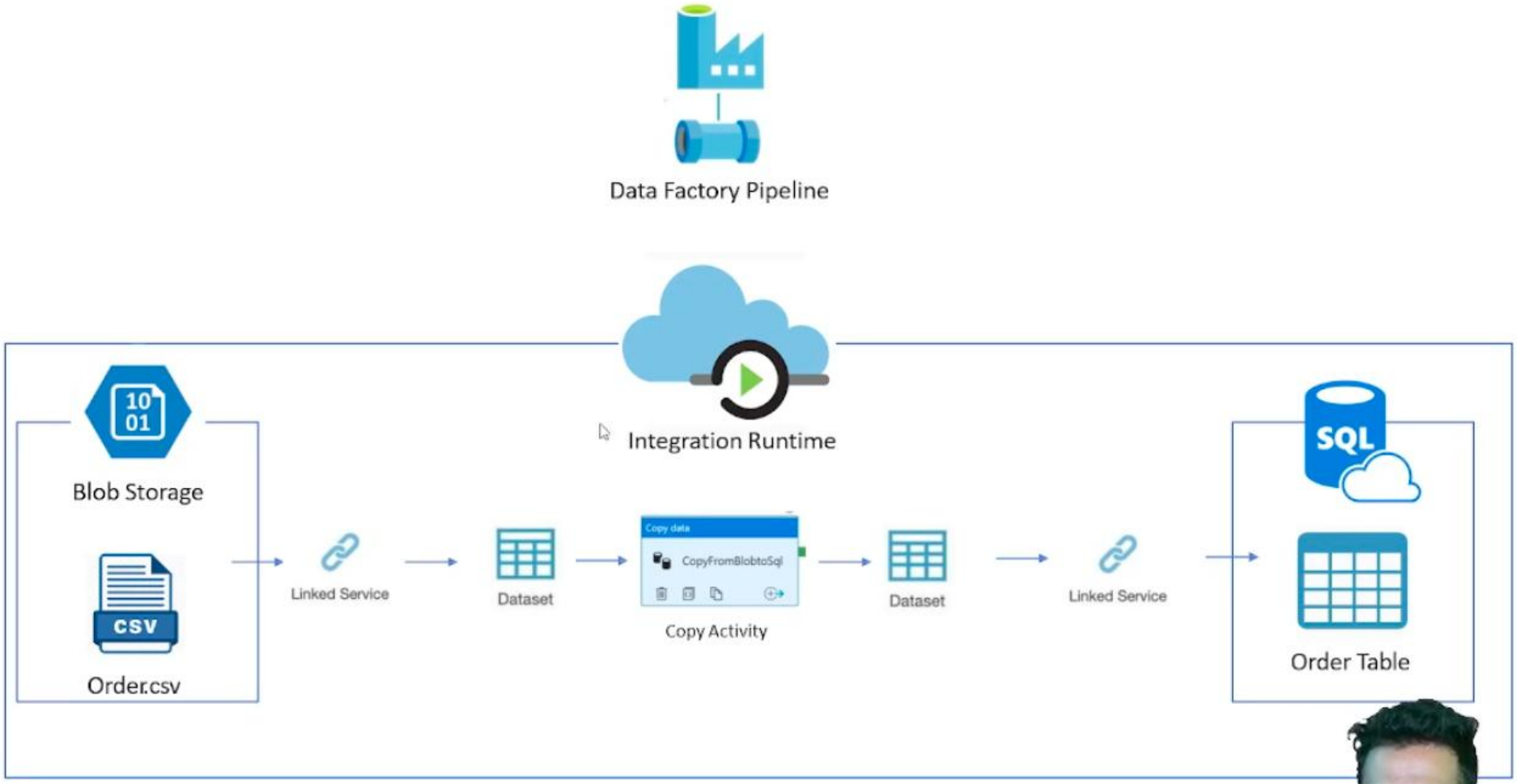
Slide 8 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint



Slide 9 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint



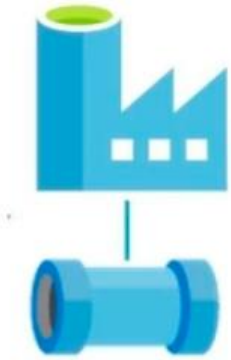
Slide 10 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint

Data Factory vs SSIS

Azure Data Factory	SSIS
Pipeline	Package
Linked Service	Connection manager
Source	Source
Sink	Destination
Activity	Control flow task
Data Flow	Data flow

PowerPoint Slide Show - [DataFactory] - PowerPoint



Data Factory
Pipeline

- Data Factories can contain one ore more pipelines
- Logical group of Activities
- Manage Activities as a set
- One Pipeline can have one or more activities

Slide 12 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint

Azure Data Factory Activities

- Represents a processing step in the pipelines
- Actions to perform on data
  - Ingest data
  - Transform data
  - Store data
- Can be linked
  - Execute sequentially or
  - Run in parallel

Slide 13 of 36

Activity types

01 Data movement activities

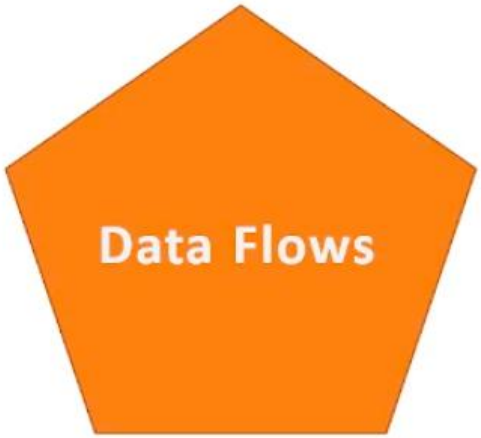
Copy data amongst data stores located on-premises and in the cloud
Data stores – Blob storage, Cosmos DB, Amazon Redshift, Google BigQuery Hive, Maria DB…etc.

02 Data transformation activities

Transform and enrich data
e.g. Hive, Pig, MapReduce, Spark or Databricks

03 Control activities

Control pipeline flow
e.g. ForEach, Web



- Data Flow is a new feature of Azure Data Factory (ADF) that allows you to develop graphical data transformation logic that can be executed as activities within ADF pipelines.
- Two types:
  - Mapping
  - Wrangling

PowerPoint Slide Show - [DataFactory] - PowerPoint

Integration Runtimes

- **Data Integration Capabilities**
  - **Data Flow**
  - **Data Movement**
    - Format conversion, column mapping, serialization/ deserialization etc.
    - Provides the native compute to move data between cloud data stores in a secure, reliable, and high-performance manner.
  - **Activity dispatch** (e.g. Databricks Notebook, HDInsight Hive, pig, spark activity, SP, ADL Analytics U-SQL activity)
  - **SSIS Package execution**

PowerPoint Slide Show - [DataFactory] - PowerPoint

Dataset

- Simply point or reference the data
- Reference data used in an Activity
  - Files
  - Folders
  - Documents
  - Tables

Slide 16 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint

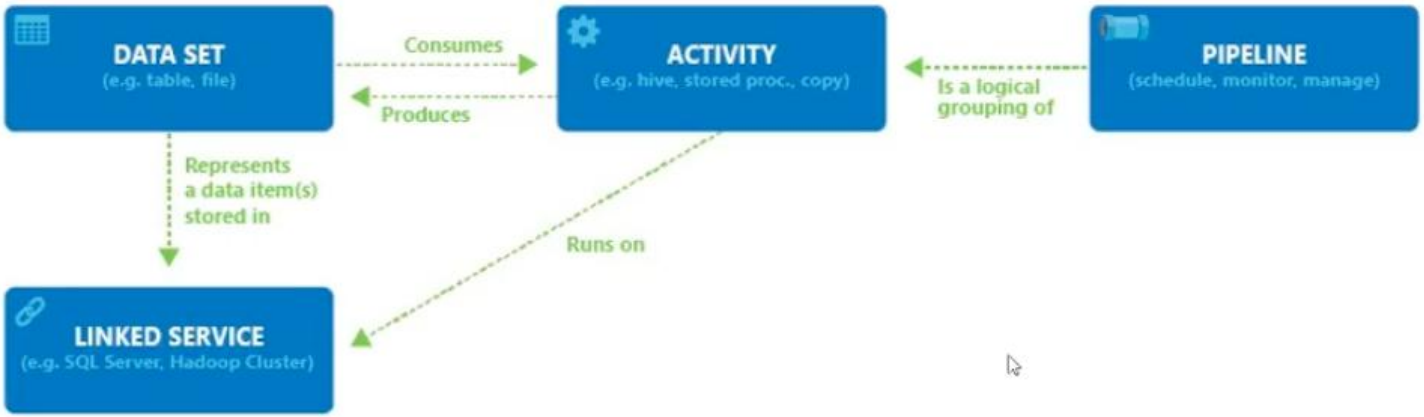
Linked service

- Similar to connection string
- Represent the connection information to connect to external resources
  - Datastores like Azure SQL Server
  - Compute resource e.g. Spark Cluster

Slide 17 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint

ADF Components



PowerPoint Slide Show - [DataFactory] - PowerPoint

Integration Runtimes

- Provides fully managed, serverless compute infrastructure
  - You don't have to worry about infrastructure provision, software installation, patching, or capacity scaling.
  - Pay only for duration of actual use
- Bridges between the activity and linked service
  - Activity defines the action
  - Linked service define the location

PowerPoint Slide Show - [DataFactory] - PowerPoint

Integration Runtimes

Specify the infrastructure to run activities

Azure Integration Runtime

Work on public networks

Responsible for data flows, data movements, and activity dispatches

Self-hosted Integration Runtime

Work on public and private networks

Provide data movement and activity dispatch capabilities

Need to install on on-premises machine or a virtual machine inside private network

SSIS Integration Runtime

Supports SSIS package execution

Works on public and private networks

Integration Runtimes

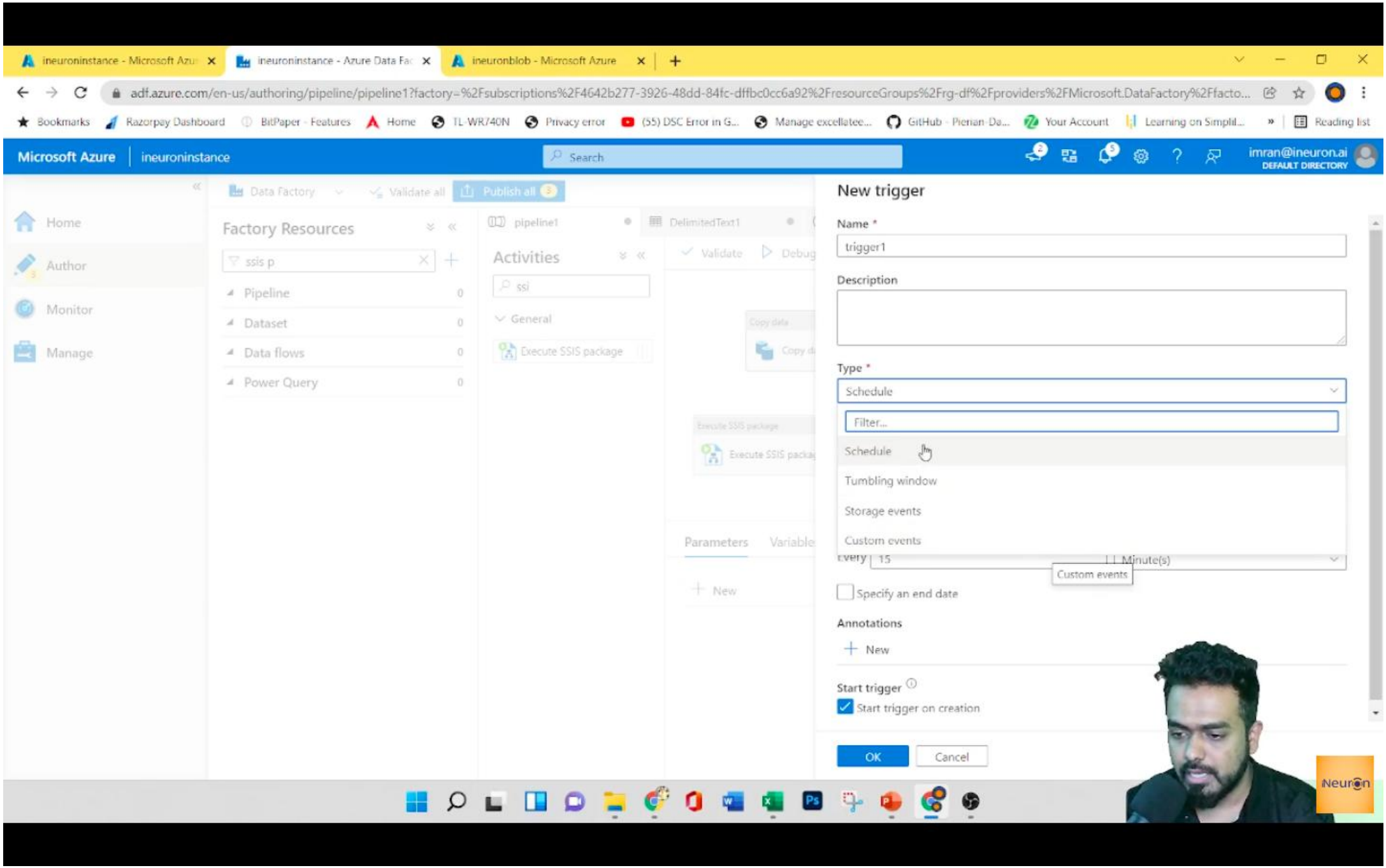
IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

PowerPoint Slide Show - [DataFactory] - PowerPoint

Integration Runtimes

- Default IR – AutoResolveIntegrationRuntime
- Create Azure IR
  - When you want to explicitly define the location of IR
  - Virtually group the activities executions on different IR for management purpose

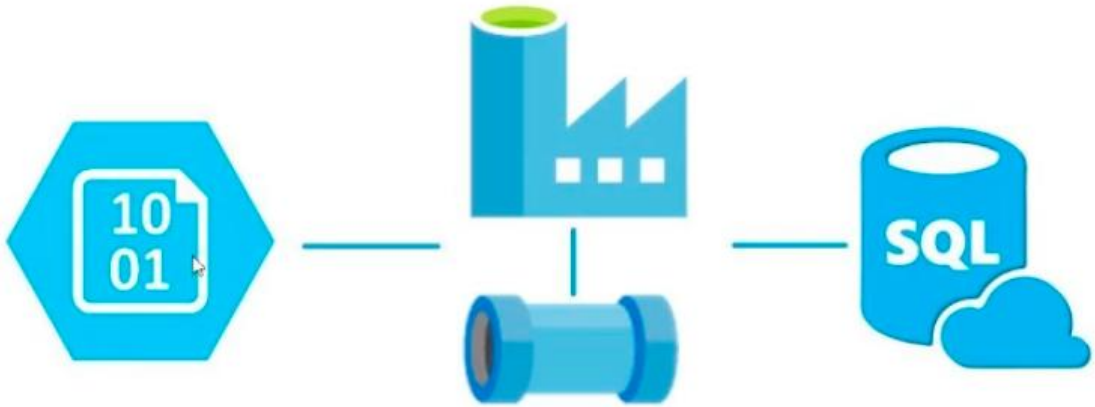
Slide 25 of 36


ineuroninstance - Microsoft Azure
ineuroninstance - Azure Data Fac
ineuronblob - Microsoft Azure
adf.azure.com/en-us/authoring/pipeline/pipeline1?factory=%2Fsubscriptions%2F4642b277-3926-48dd-84fc-dffbc0cc6a92%2FresourceGroups%2Frg-df%2Fproviders%2FMicrosoft.DataFactory%2Ffacto...
Bookmarks
Razorpay Dashboard
BitPaper - Features
Home
TL-WR740N
Privacy error
(55) DSC Error in G...
Manage excellatee...
GitHub - Pierian-Da...
Your Account
Learning on Simpli...
Reading list
Microsoft Azure
ineuroninstance
Search
imran@ineuron.ai
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Home
Author
Monitor
Manage
Factory Resources
ssis p
Pipeline
Dataset
Data flows
Power Query
pipeline1
DelimitedText1
Activities
ssi
General
Execute SSIS package
Validate
Debug
Parameters
Variable
New
New trigger
Name *
trigger1
Description
Type *
Schedule
Filter...
Schedule
Tumbling window
Storage events
Custom events
Every
15
Minute(s)
Custom events
Specify an end date
Annotations
New
Start trigger
Start trigger on creation
OK
Cancel
Neuron

- **Execute pipeline**
- **Many to many relationship b/w pipeline and trigger**
- **Three types of Trigger**
  - **Schedule Trigger** – Invoke pipeline on a wall-clock schedule
  - **Tumbling Window Trigger –** Operates on a periodic interval, also retain state
  - one-to-one relationship
  - Advance configuration options - Dependencies, delay, retry, concurrency
  - Properties - trigger().outputs.WindowStartTime/WindowEndTime
  - **Event-based Trigger –** trigger pipeline in response to an event
    - e.g. Arrival/deletion of file in Blob storage
    - Event trigger with Azure Event Grid Service
    - Properties – triggerBody().folderPath/fileName

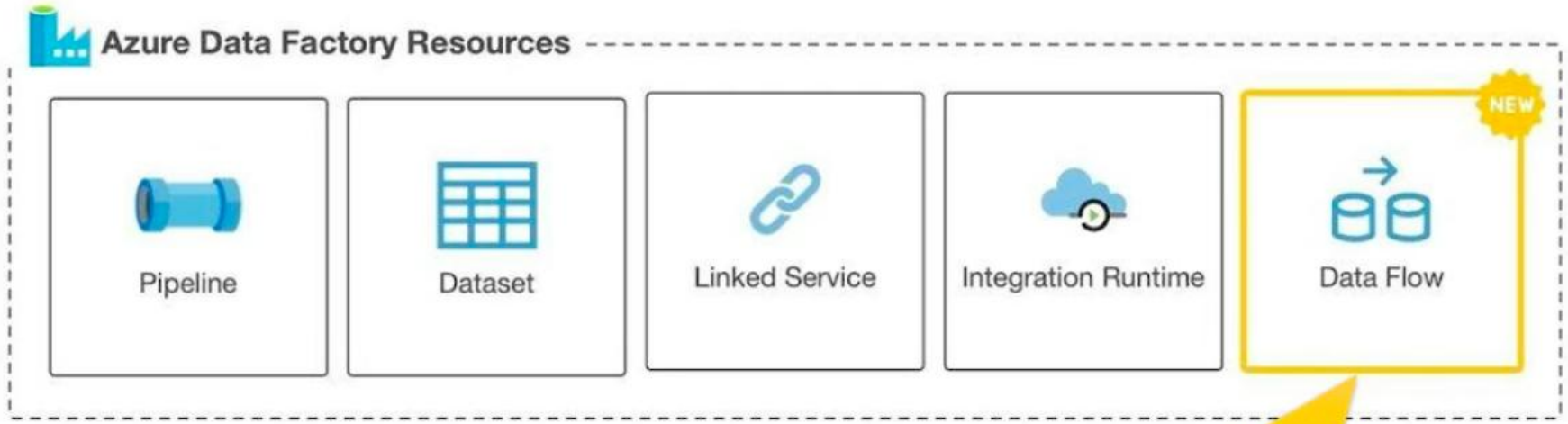
PowerPoint Slide Show - [DataFactory] - PowerPoint

Demo: Copy Activity



PowerPoint Slide Show - [DataFactory] - PowerPoint

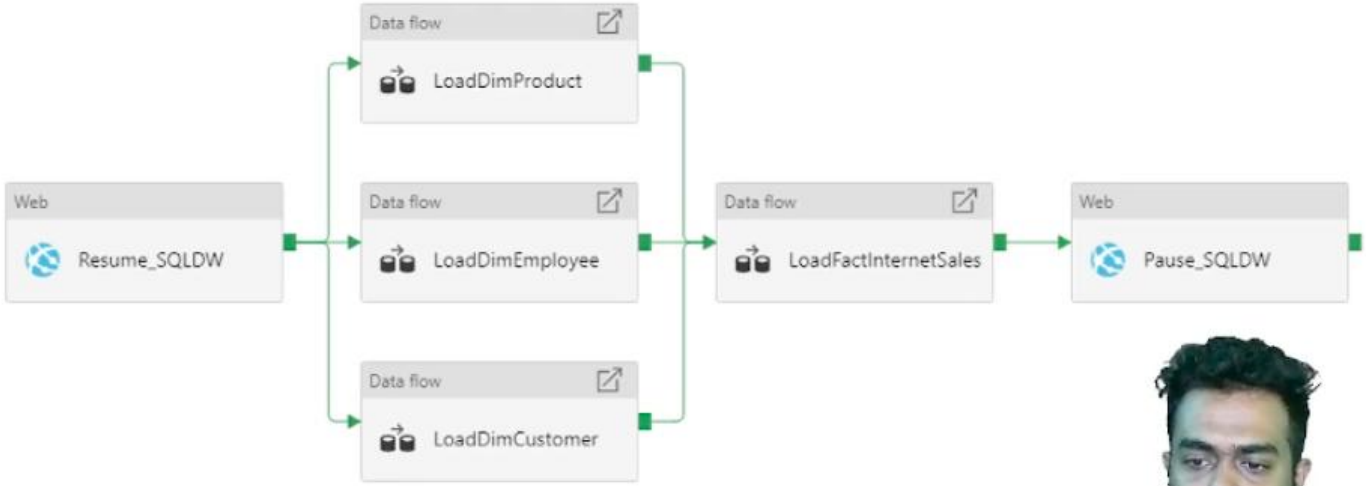
Data Flows



PowerPoint Slide Show - [DataFactory] - PowerPoint

Example of the SSIS Control Flow tab for loading our data mart tables:

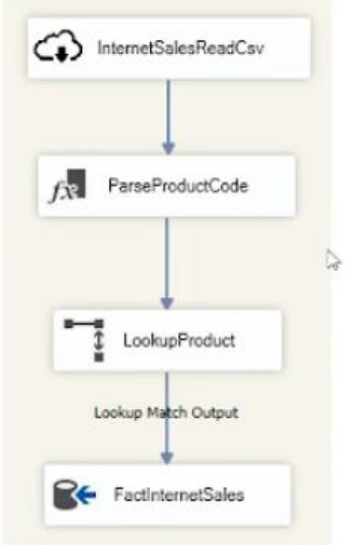
Example of the ADF Pipeline for loading our data mart tables:



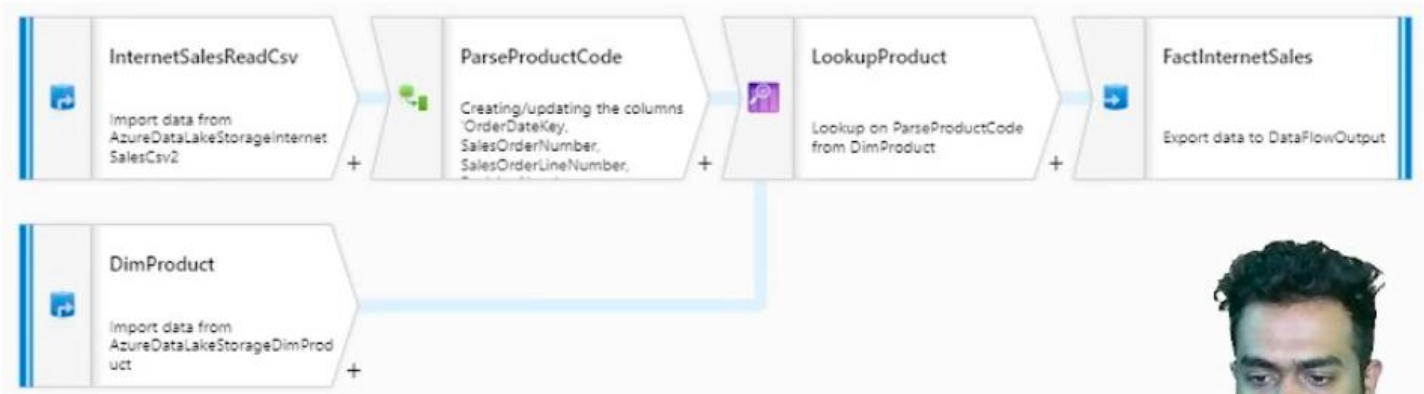
Slide 30 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint

Example of SSIS Data Flow tab for loading the FactInternetSales table:

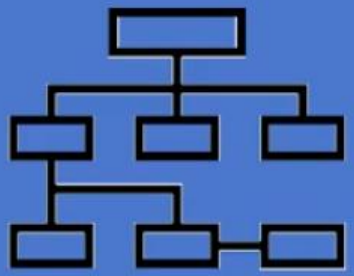


Example of ADF Mapping Data Flows for loading the FactInternetSales table:



Slide 31 of 36

PowerPoint Slide Show - [DataFactory] - PowerPoint



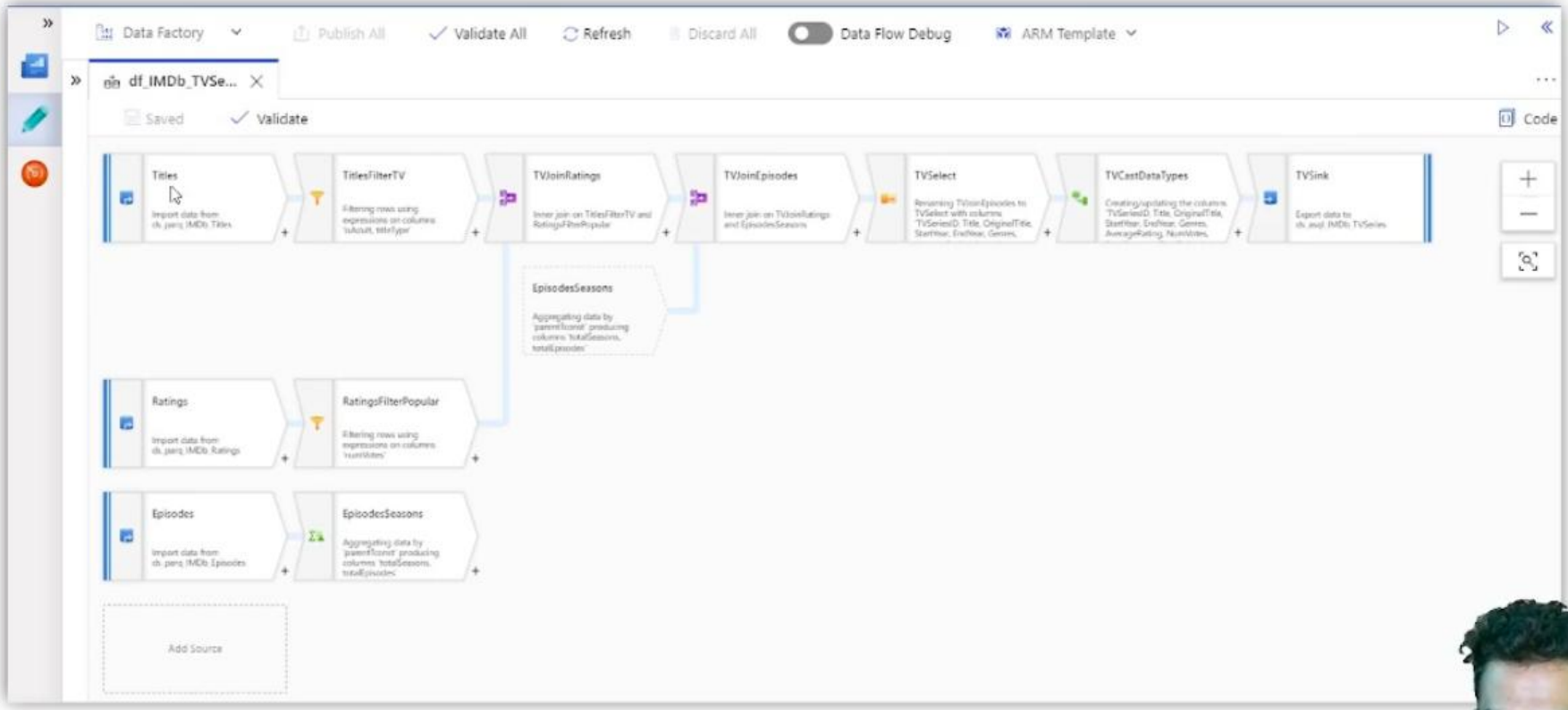
Data Flow

Mapping Data flow – Transform Data (Known data and schema)

Wrangling Data flow – Prepare and explore data using power query (known or unknown datasets)

PowerPoint Slide Show - [DataFactory] - PowerPoint

Mapping Data Flows



PowerPoint Slide Show - [DataFactory] - PowerPoint

Data flows behind the scene

- Behind the scene Data flow will execute on Azure Databricks using Spark
- ADF internally handles all the code translation, spark optimization and execution of transformation