

# ESPECIALIZACIÓN

# **Ingeniería de datos con Azure**

Curso: Microsoft Azure & ETL Fundamentals

Docente: Richard Tadeo Zenteno

## REGLAS



Se requiere **puntualidad** para un mejor desarrollo del curso.



Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.



Las preguntas se realizarán **a través del chat** y en caso de que lo requieran **podrán activar el micrófono**.



Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.



**Identificarse** en la sala Zoom con el primer nombre y primer apellido.

## ITINERARIO

*07:00 PM – 07:30 PM      **Soporte técnico DMC***

*07:30 PM – 08:50 PM      **Agenda***

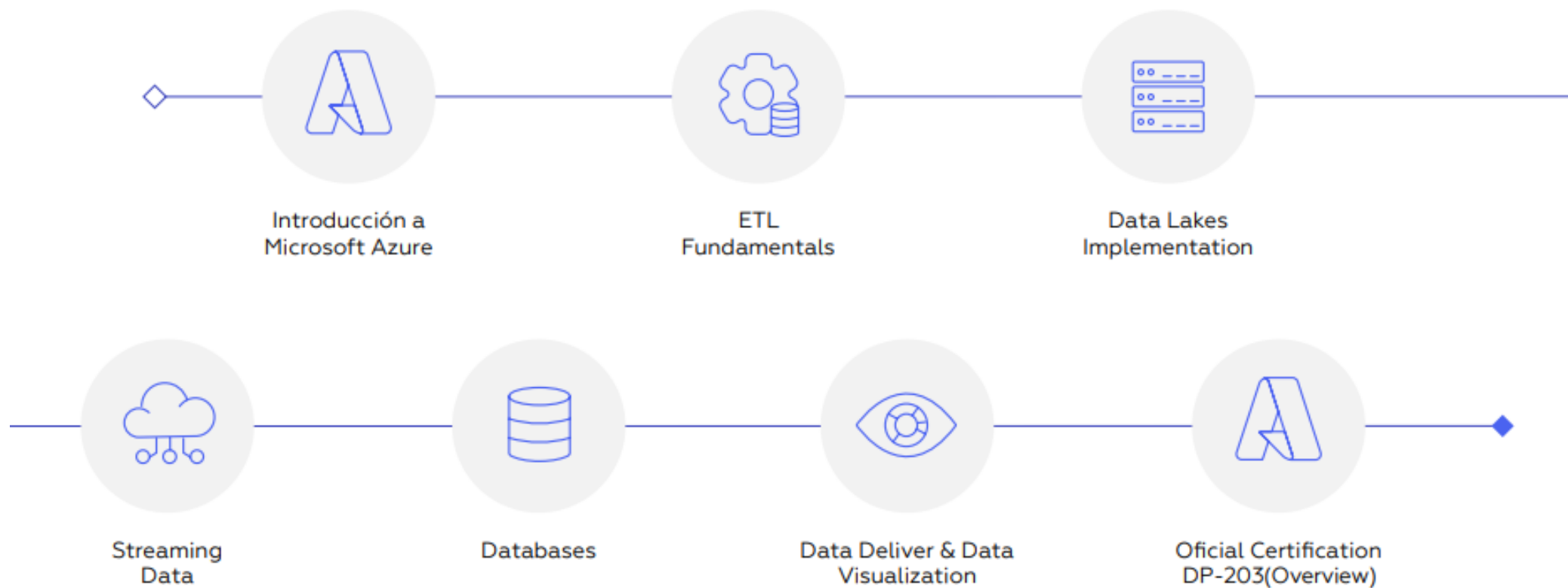
*08:50 PM – 09:00 PM      **Pausa Activa***

*09:00 PM – 10:30 PM      **Agenda***

*Horario de Atención Área Académica y Soporte*

*Lunes a Viernes 09:00 am a 10:30 pm / Sábado 09:00 am a 02:00pm*

# MALLA CURRICULAR



**CERIFICACIÓN FINAL**  
por **Aprobación** de la Especialización en **Ingeniería de Datos con Microsoft Azure** (48 horas académicas)

# CONTENIDO



## Introducción a Microsoft Azure

- Introducción a Cloud Computing. Proveedores de servicios Cloud, On-Premise vs. On-Cloud, principales servicios, descripción de los modelos de costos.
- Identity and Access Management (IAM). Overview de los roles principales, ejemplos de gestión de permisos.



## ETL Fundamentals

- Introducción a las soluciones ETL. Definición, descripción de sus etapas.
- Introducción a los servicios Azure Data Factory y Data Flow. Características generales, casos de uso.
- Taller: Implementación de un ETL Básico con Azure.



## Data Lakes Implementation

- Introducción a Data Lakes. Definición, arquitectura, capas (Raw, Stage, Analytics).
- Introducción a los servicios Azure Blob Storage y Storage Account.
- Taller: Implementación de un Datalake en Azure.

# CONTENIDO



## Streaming Data

- Introducción a procesamiento de datos Batch y Streaming. Diferencias Near-Real-Time y Real-Time.
- Introducción a IoT. Definición, uso de sensores, aplicaciones.
- Revisión de servicios: Azure EventHubs y IoT Hub. Características generales, ejemplos de implementación y uso.
- Taller: Manejo de Streaming al Data.



## Databases

- Introducción a las bases de datos Relacionales y No-Relacionales. Definición, características, casos de uso.
- Azure SQL Database for MariaDB. Descripción y características generales.
- Azure SQL Database for PostgreSQL. Descripción y características generales.
- Azure SQL Database for CosmosDB. Descripción y características generales.
- Taller: Diseño de una base de datos relacional y técnicas para poblarla.



## Data Deliver & Data Visualization

- Azure Synapse Analytics. Propósito del servicio, características generales.
- Fabric. Propósito del servicio, características generales.
- Taller: Conexión de Power BI a servicios de datos de Azure.

## AGENDA

**01**

Introducción a las  
soluciones ETL

**02**

Descripción de  
etapas

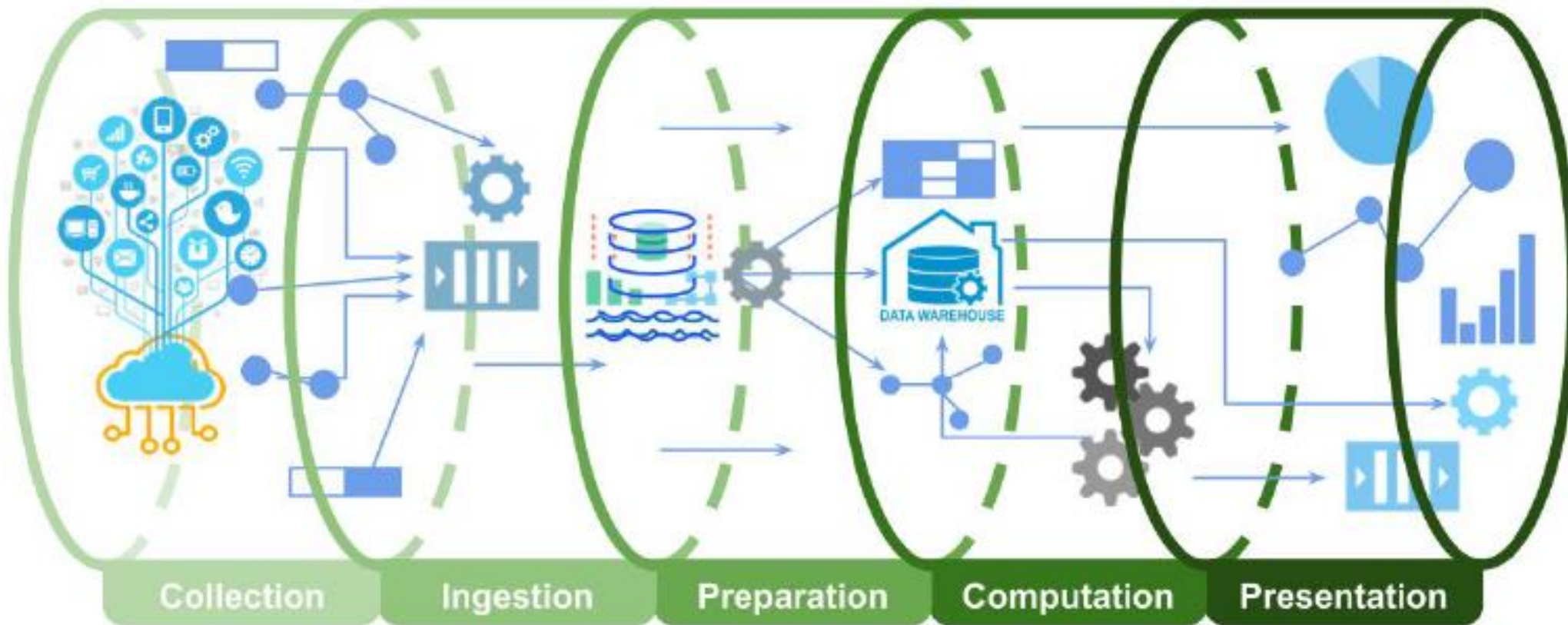
**03**

Introducción a los  
servicios Azure  
Data Factory y Data  
Flow

**04**

Laboratorio 03:  
Ingesta con  
Datafactory

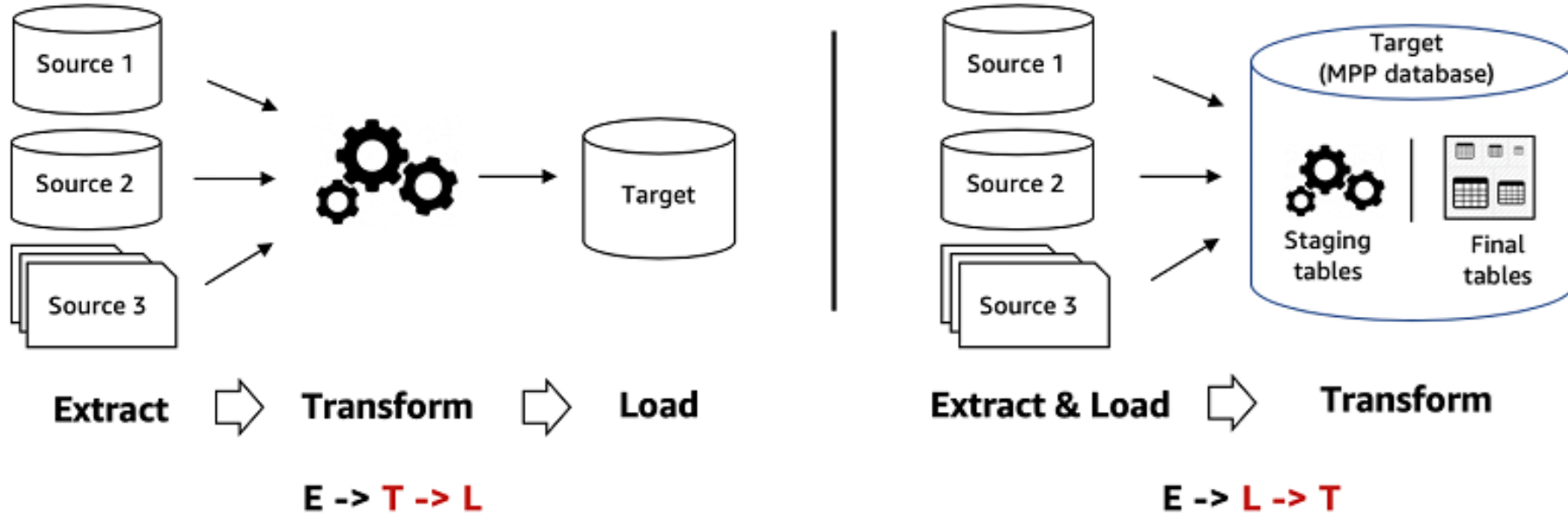
# Data Pipelines



*“Procesos (tuberías) para disponibilizar los datos → información → conocimiento → ventaja competitiva.”*



# Tipos de Data Pipelines (ETL y ELT)



**Batch** pipelines - runs in set intervals

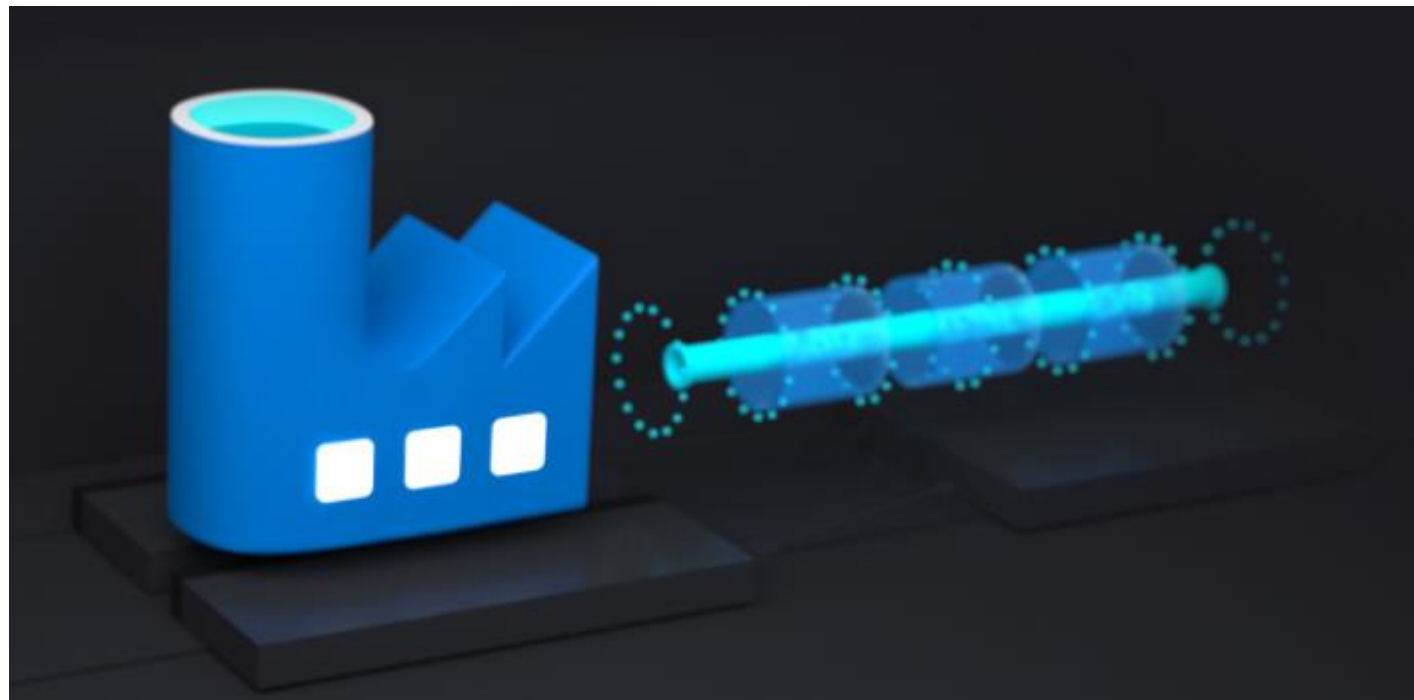
## Running



**Real time** - streaming pipelines

# ¿Qué es Azure Datafactory?

Es un servicio de integración de datos serverless totalmente administrada para la ingesta, preparación y transformación de todos los datos a gran escala.



# ETL sin código como servicio

## INGERIR



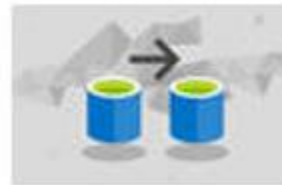
- Datos de copia híbrida locales y de varias nubes
- Más de 90 conectores nativos
- Sin servidor y escalado automático
- Uso de asistente para trabajos de copia rápidos

## FLUJO DE CONTROL



- Diseño de canalizaciones de datos sin código
- Generación de canalizaciones mediante SDK
- Utilización de construcciones de flujo de trabajo: bucles, ramas, ejecución condicional, variables, parámetros...

## FLUJO DE DATOS



- Transformaciones de datos sin código que se ejecutan en Spark
- Escalabilidad horizontal con Azure Integration Runtime
- Generación de flujos de datos mediante SDK
- Diseñadores para ingenieros de datos y análisis de datos

## PROGRAMACIÓN



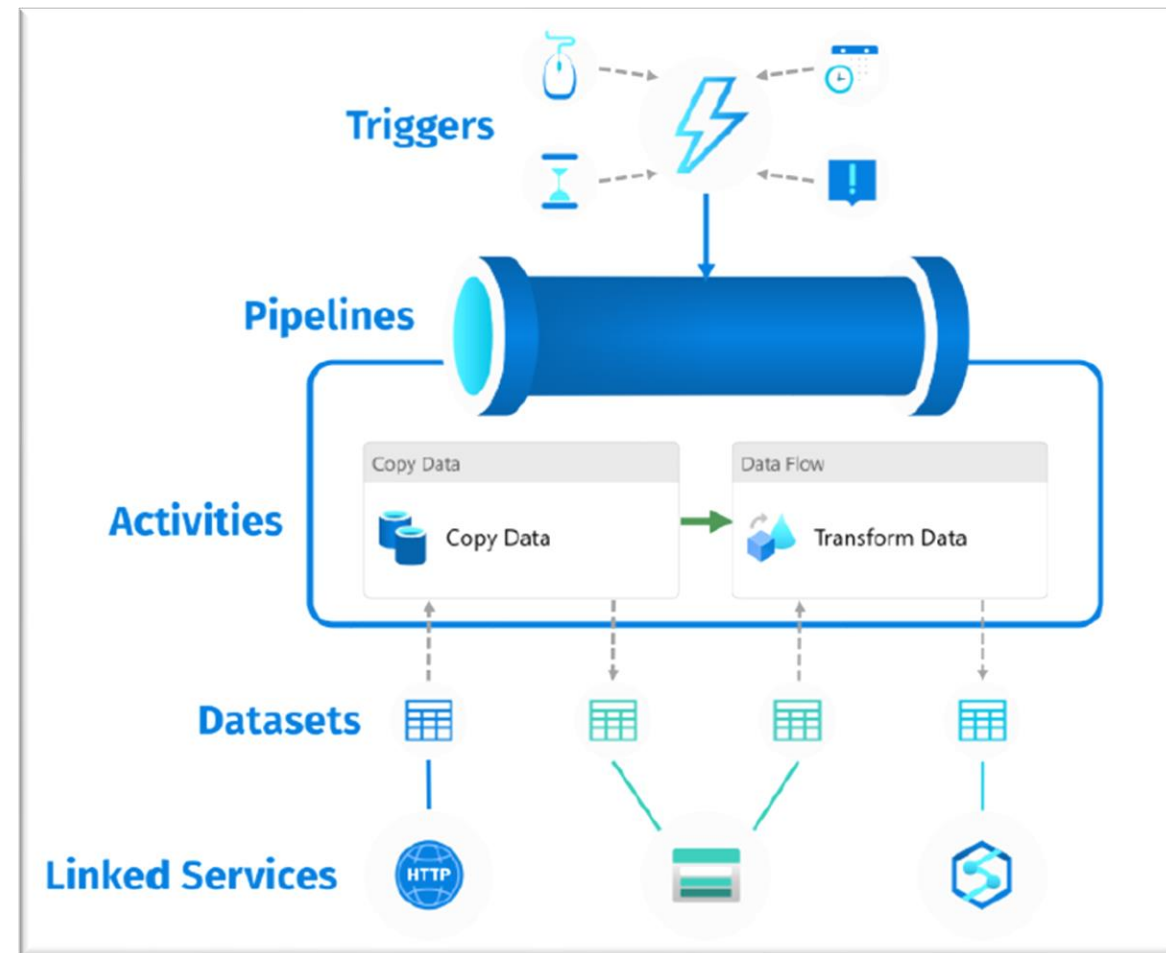
- Creación y mantenimiento de programaciones operativas para las canalizaciones de datos
- Reloj de pared, basado en eventos, ventana de saltos de tamaño constante, encadenadas

## MONITOR



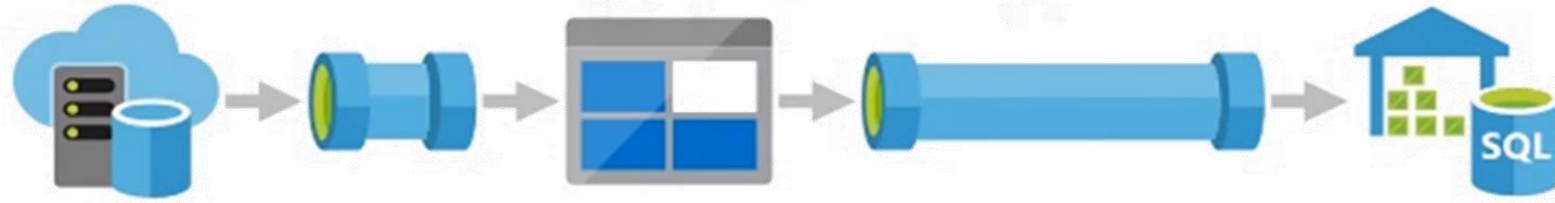
- Visualización de ejecuciones activas e historial de canalizaciones
- Actividad detallada y ejecuciones de flujos de datos
- Establecimiento de alertas y notificaciones

# Componentes de Azure Datafactory



# ¿Qué podemos hacer con Azure Data Factory?

- Batch
- Real-Time
- Ingesta
- Data Transfer



- Batch
- Real-Time

## Copy Data

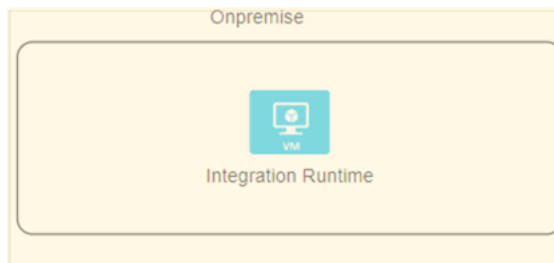
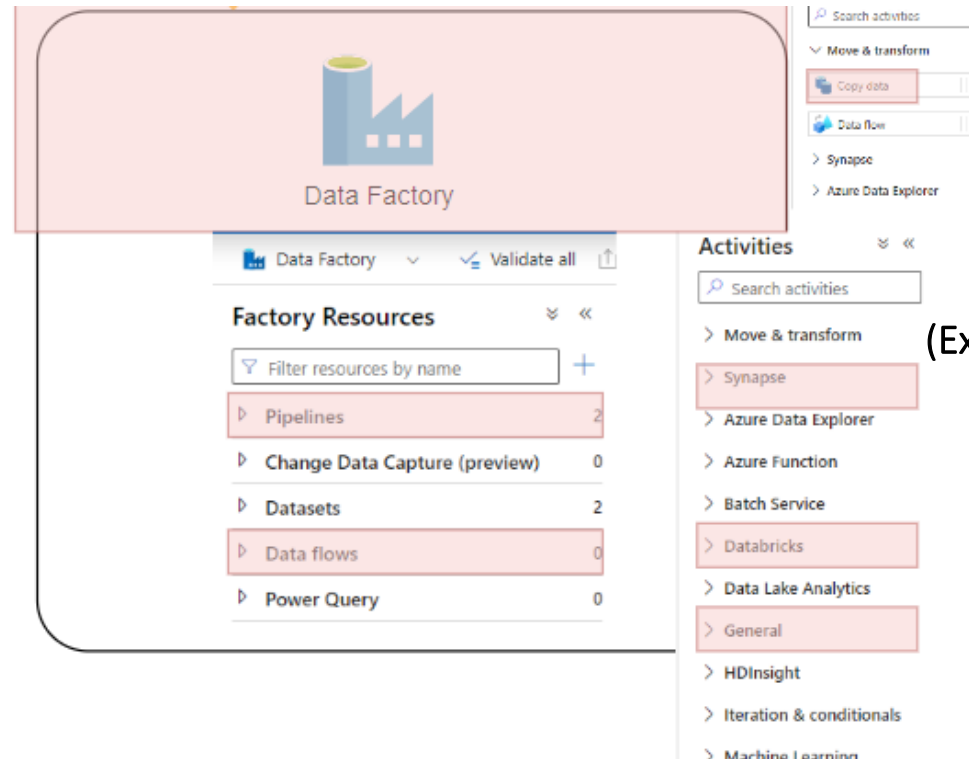
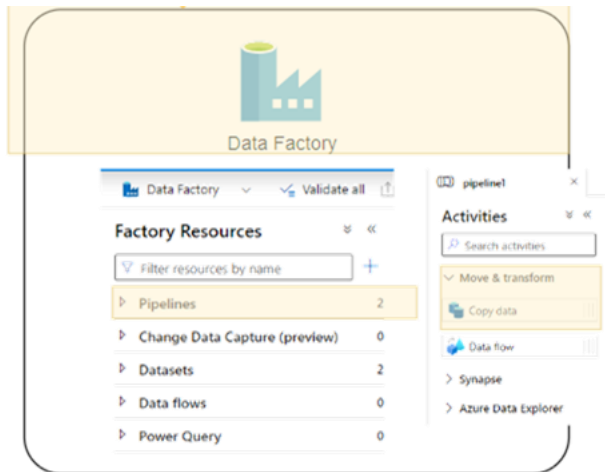
## Transform Data

ELT

ETL

(Extract - Load - Transform)

(Extract - Transform - Load)

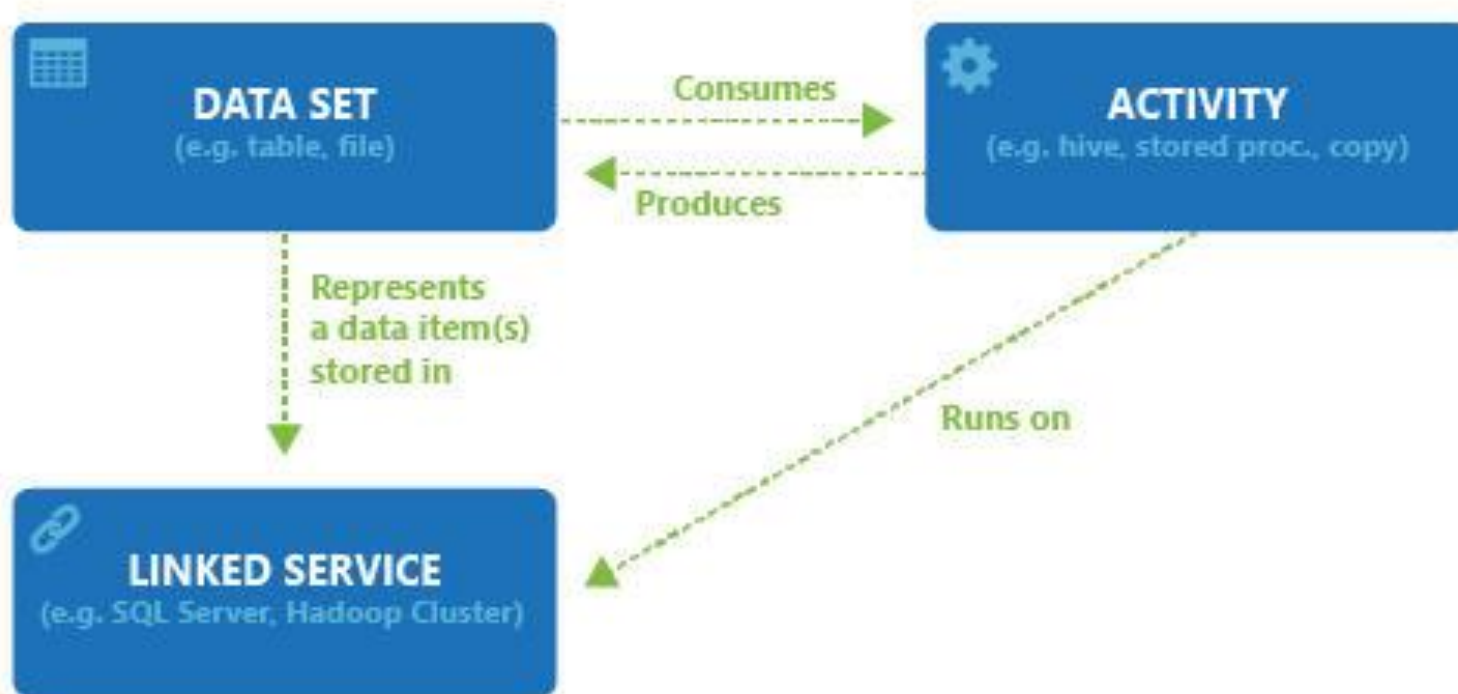


# Data Sources, Linked Services & Datasets

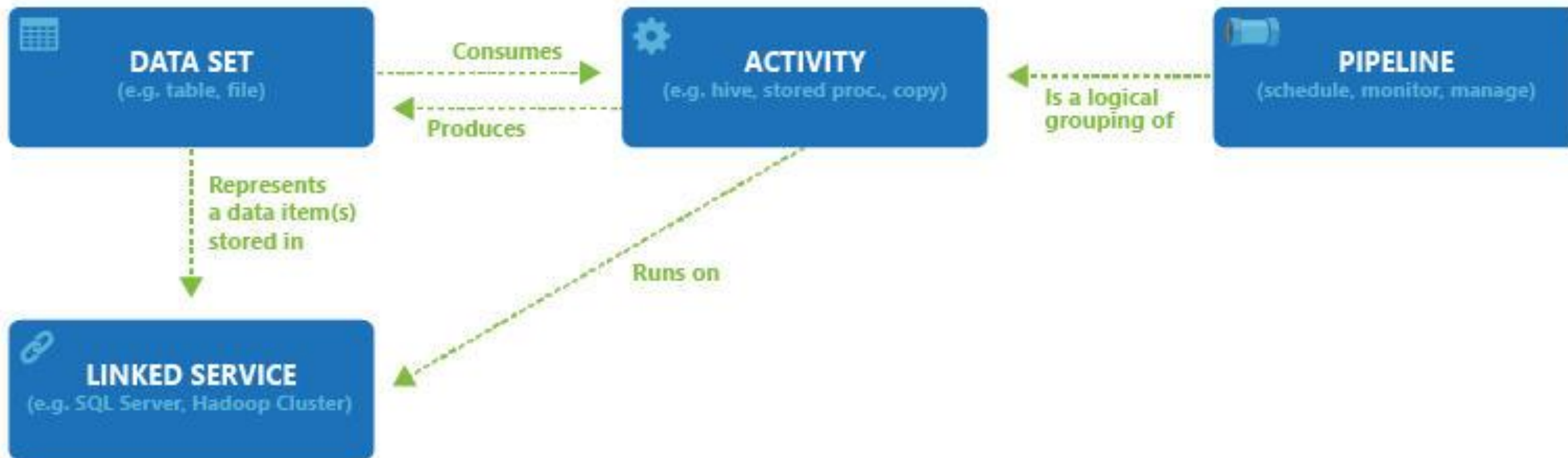




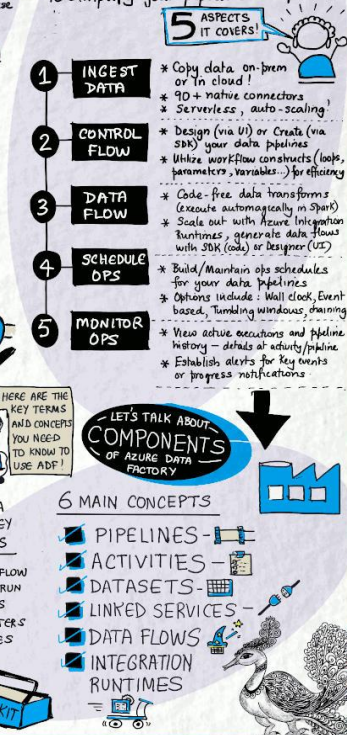
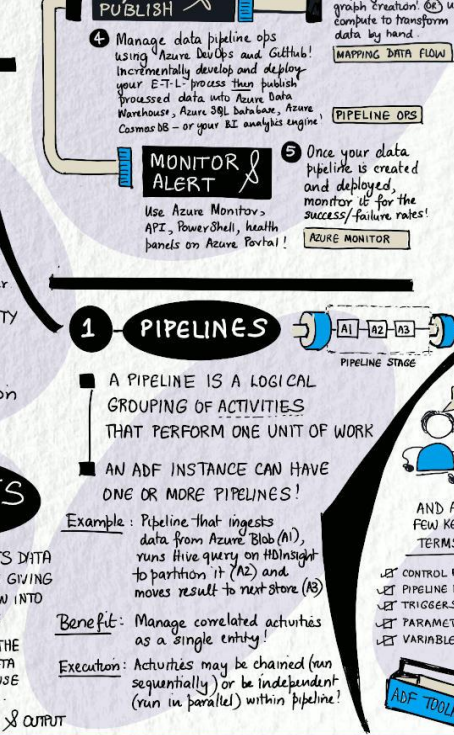
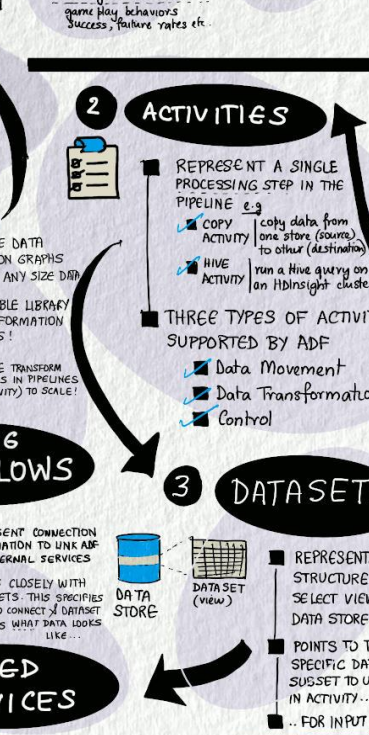
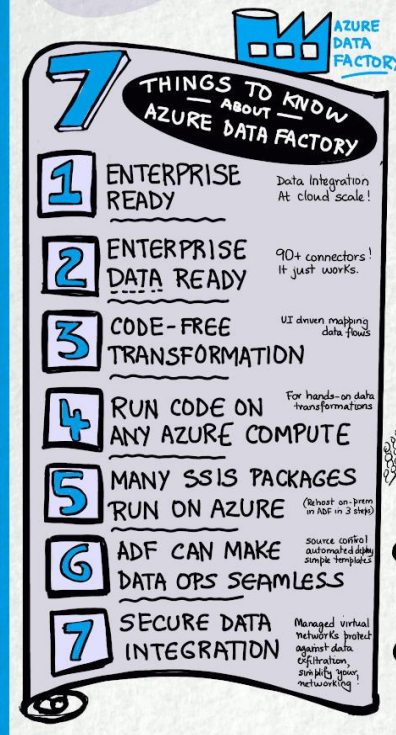
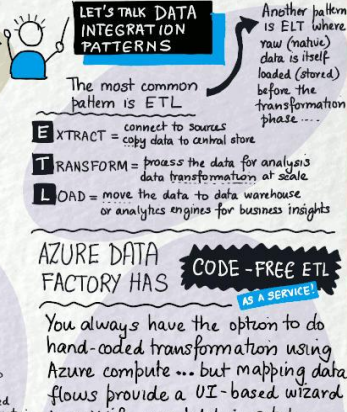
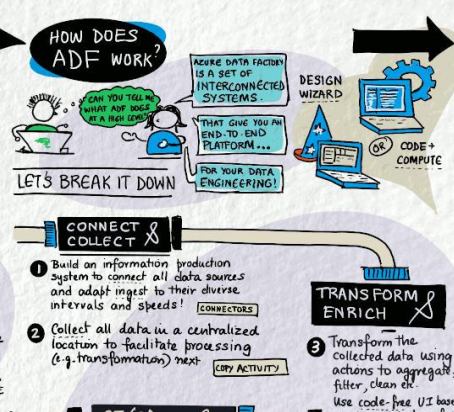
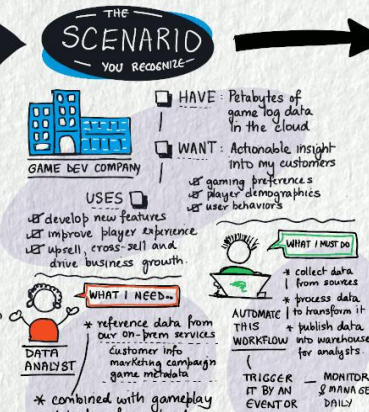
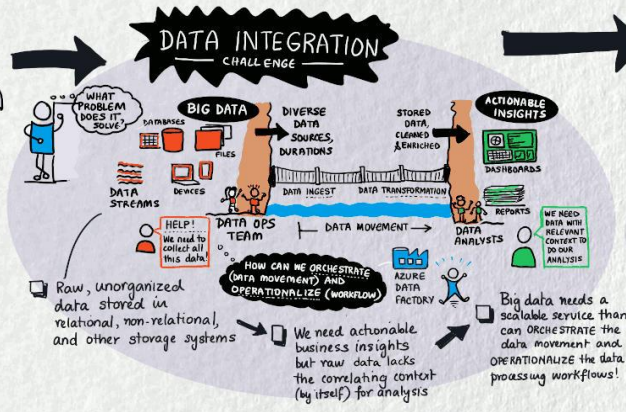
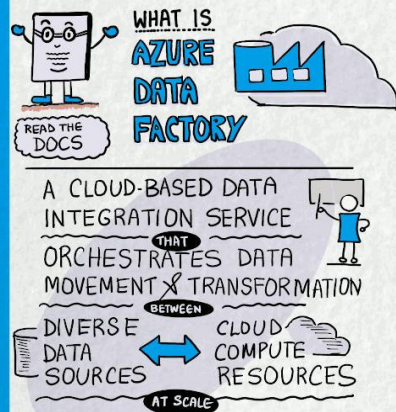
# Activities



# Pipelines



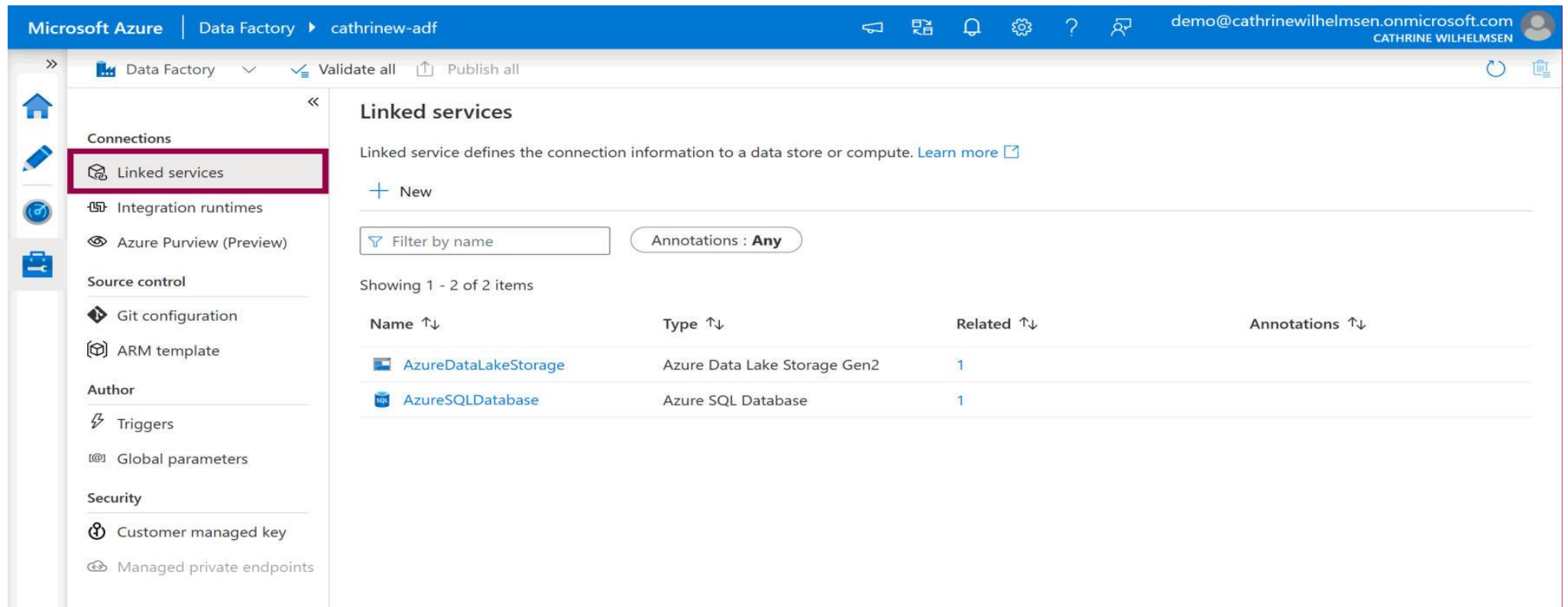






# Linked Services

- Los Linked Services son las conexiones que se realizan hacia un origen de datos, es una especie de Connection Manager como en SSIS.
- Para aplicar seguridad sobre estos, podrías utilizar los Key Vaults.



Microsoft Azure | Data Factory | cathrinew-adf

demo@cathrinewilhelmsen.onmicrosoft.com CATHRINE WILHELMSEN

Data Factory | Validate all | Publish all

**Connections**

- Linked services
- Integration runtimes
- Azure Purview (Preview)

**Source control**

- Git configuration
- ARM template

**Author**

- Triggers
- Global parameters

**Security**

- Customer managed key
- Managed private endpoints

## Linked services

Linked service defines the connection information to a data store or compute. [Learn more](#)

+ New

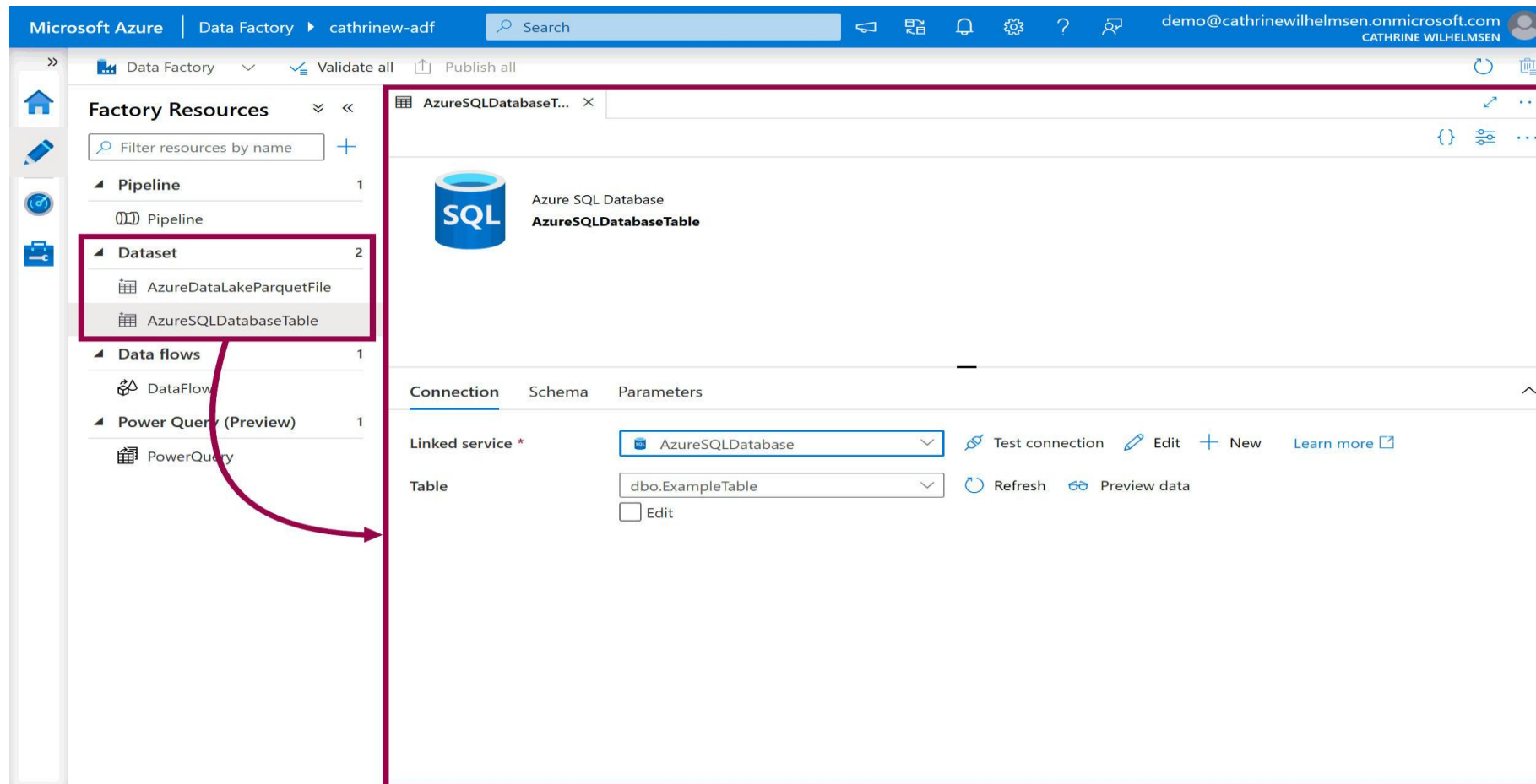
Filter by name Annotations : Any

Showing 1 - 2 of 2 items

Name ↑↓	Type ↑↓	Related ↑↓	Annotations ↑↓
<a href="#">AzureDataLakeStorage</a>	Azure Data Lake Storage Gen2	1	
<a href="#">AzureSQLDatabase</a>	Azure SQL Database	1	

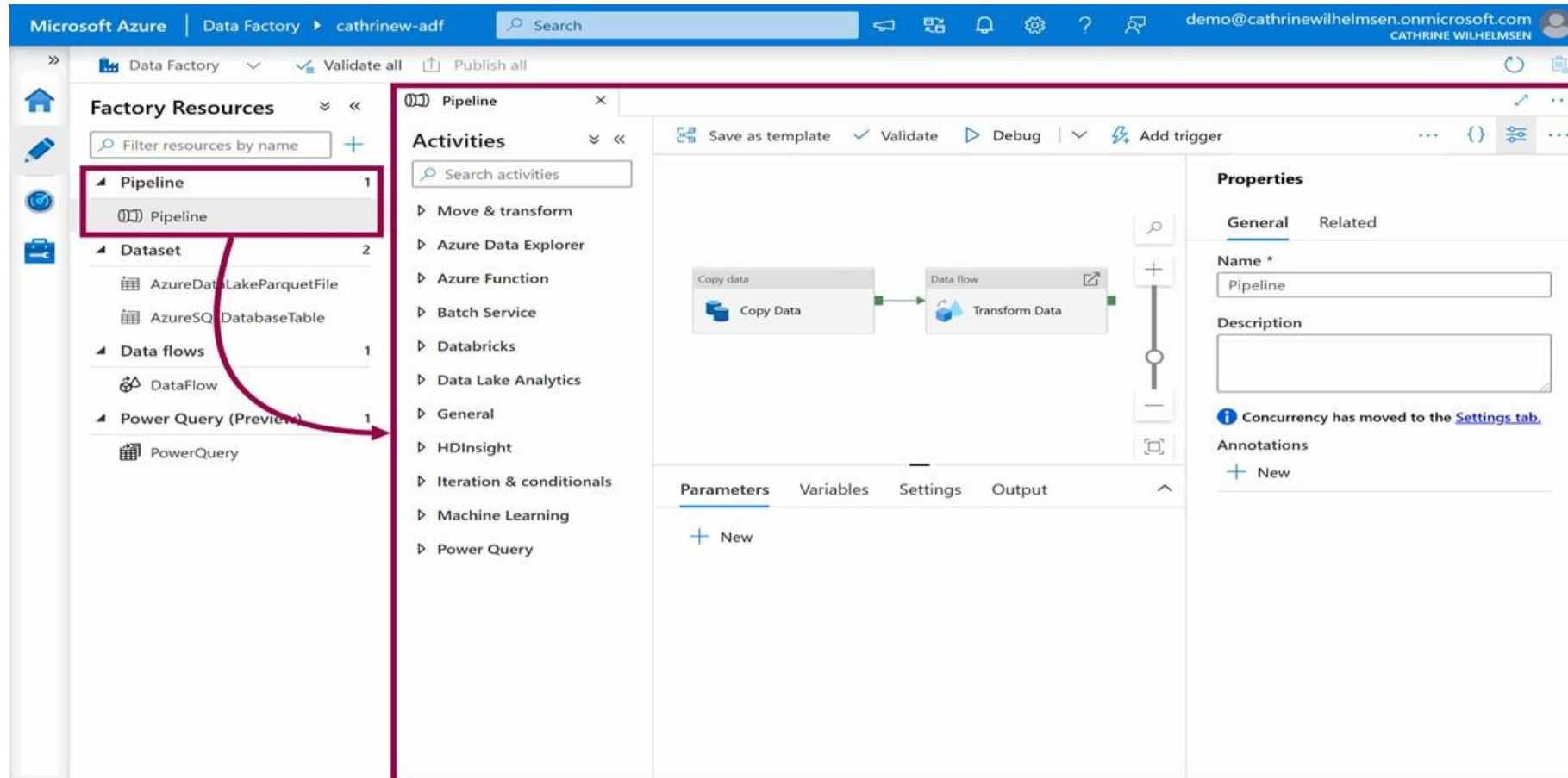
# Datasets

- Son los componentes que guardan la estructura de origen o de destino de un determinado conjunto de datos. Estos conjuntos de datos pueden ser Tables o Querys.



# Pipeline

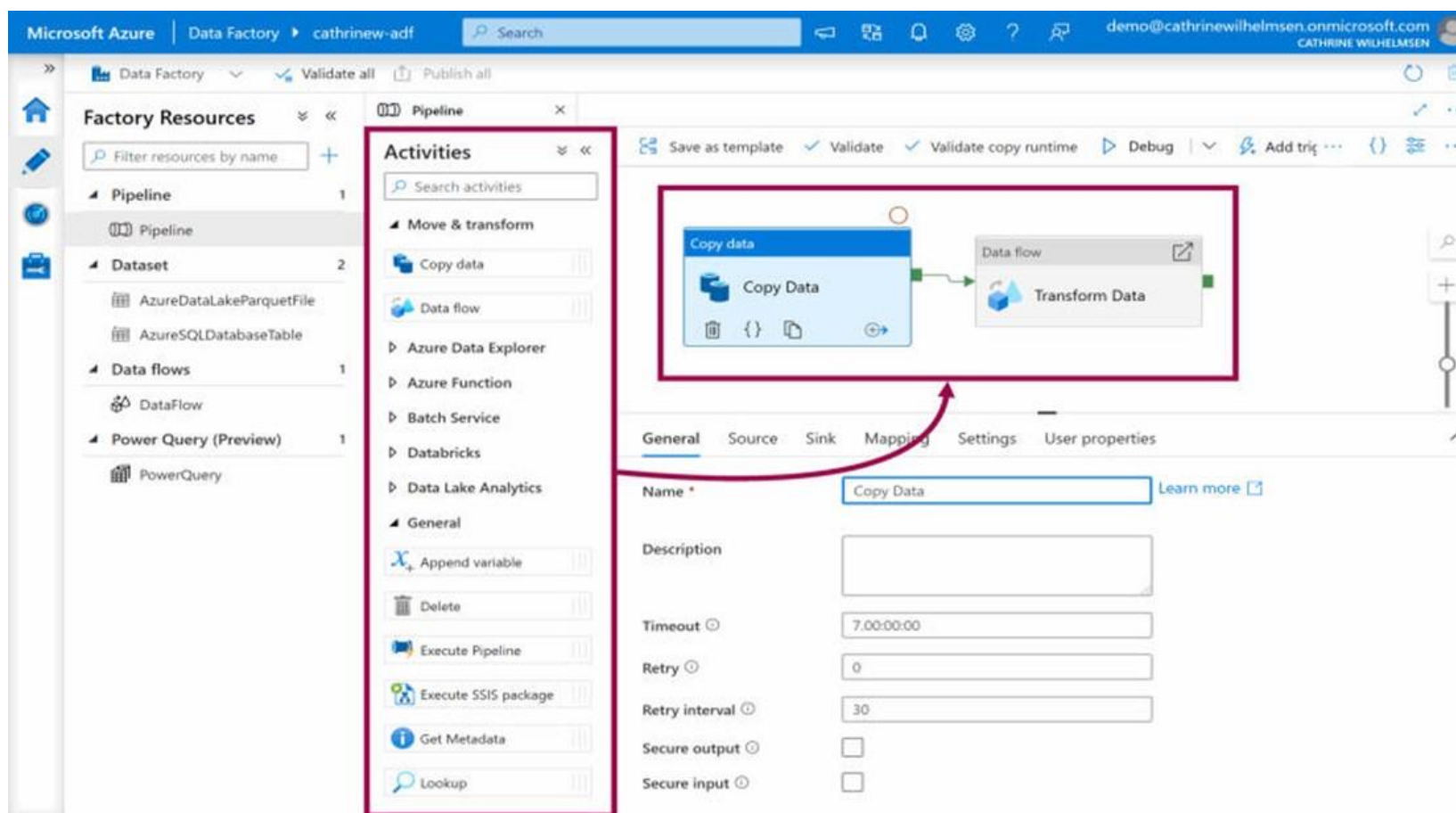
- Un pipeline es aquel componente que me permite agrupar varios actividades y poder invocarlos , invocando tan solo al pipeline



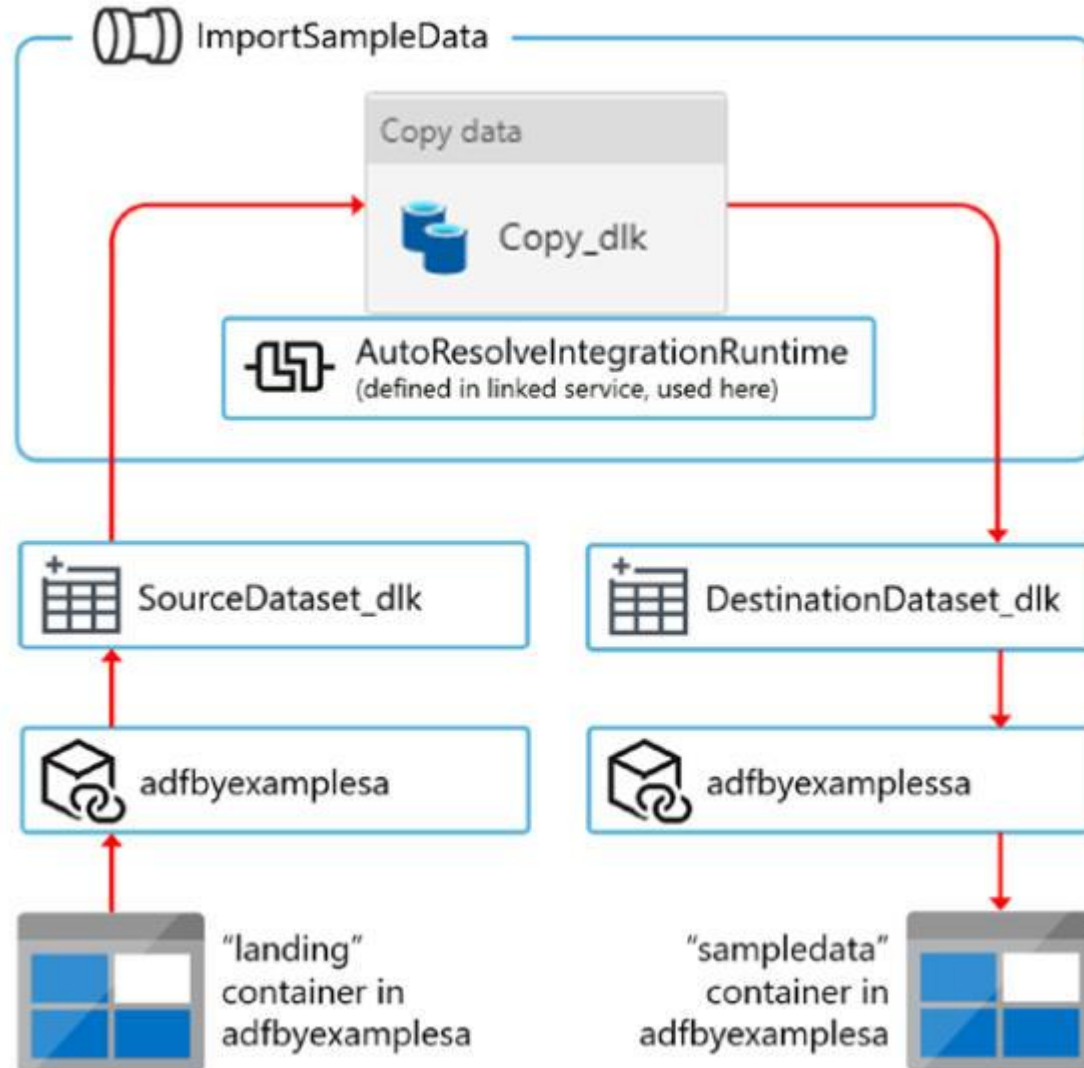
The screenshot displays the Microsoft Azure Data Factory interface. On the left, the 'Factory Resources' pane shows a list of resources: Pipeline (1), Dataset (2), Data flows (1), and Power Query (Preview) (1). The 'Pipeline' resource is highlighted with a red box. A red arrow points from this box to the 'Pipeline' activity in the 'Activities' pane. The 'Activities' pane lists various activities: Move & transform, Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. The main canvas shows a pipeline diagram with a 'Copy data' activity connected to a 'Data flow' activity. The 'Properties' pane on the right shows the 'General' tab with fields for Name and Description. The 'Name' field is labeled 'Pipeline'.

# Activities

- Son los principales componentes de Azure Data Factory. Estos componentes permiten armar toda la malla de procesos por donde va a correr nuestro ETL. Hay diferentes tipos de activities:



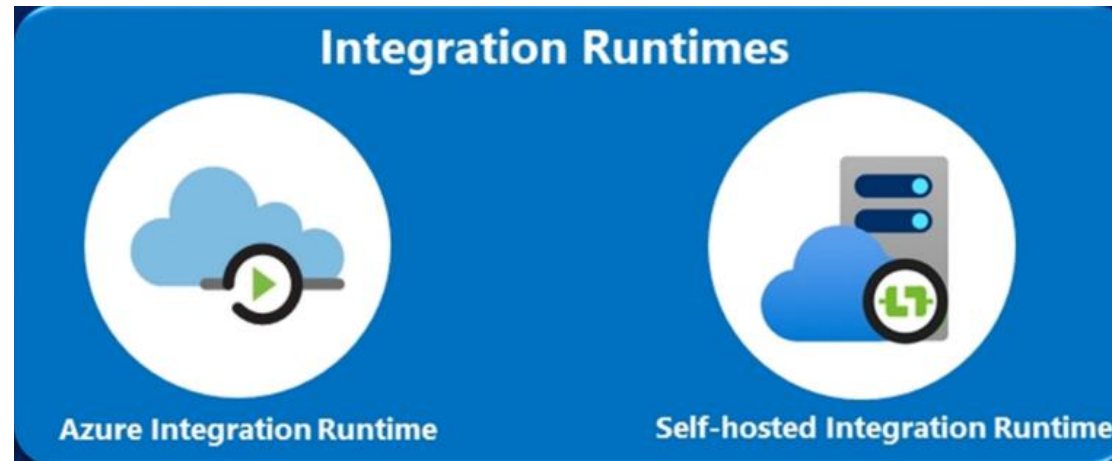
# Integration Runtime



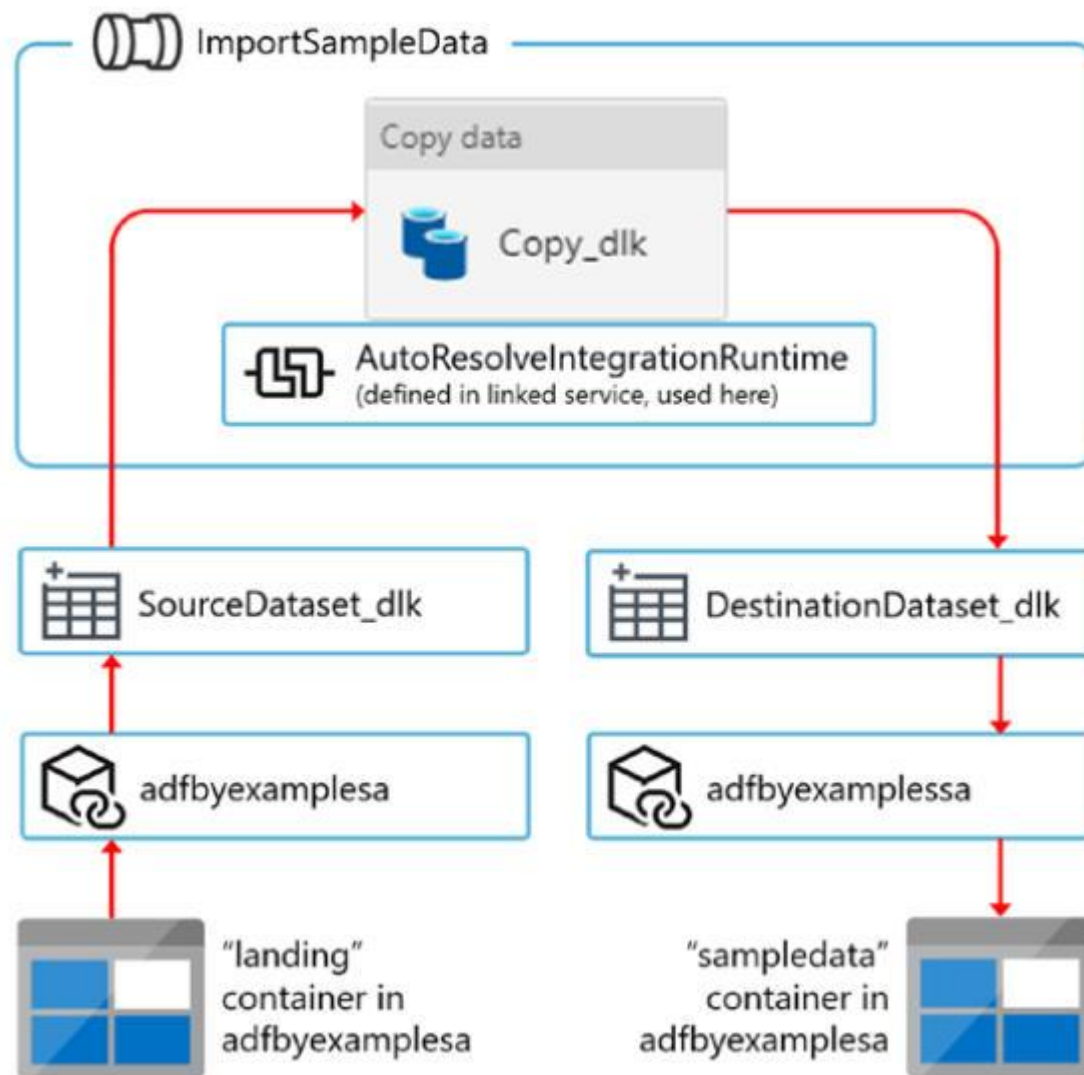
- Una instancia de Integration Runtime proporciona el puente entre la actividad y los servicios vinculados. De esta manera, la actividad puede realizarse en la región más cercana posible al almacén de datos o servicio de proceso de destino de la manera con mayor rendimiento, a la vez que se satisfacen las necesidades de seguridad y cumplimiento.

# Integration Runtime

Una instancia de Integration Runtime proporciona el puente entre la actividad y los servicios vinculados. De esta manera, la actividad puede realizarse en la región más cercana posible al almacén de datos o servicio de proceso de destino de la manera con mayor rendimiento, a la vez que se satisfacen las necesidades de seguridad y cumplimiento.

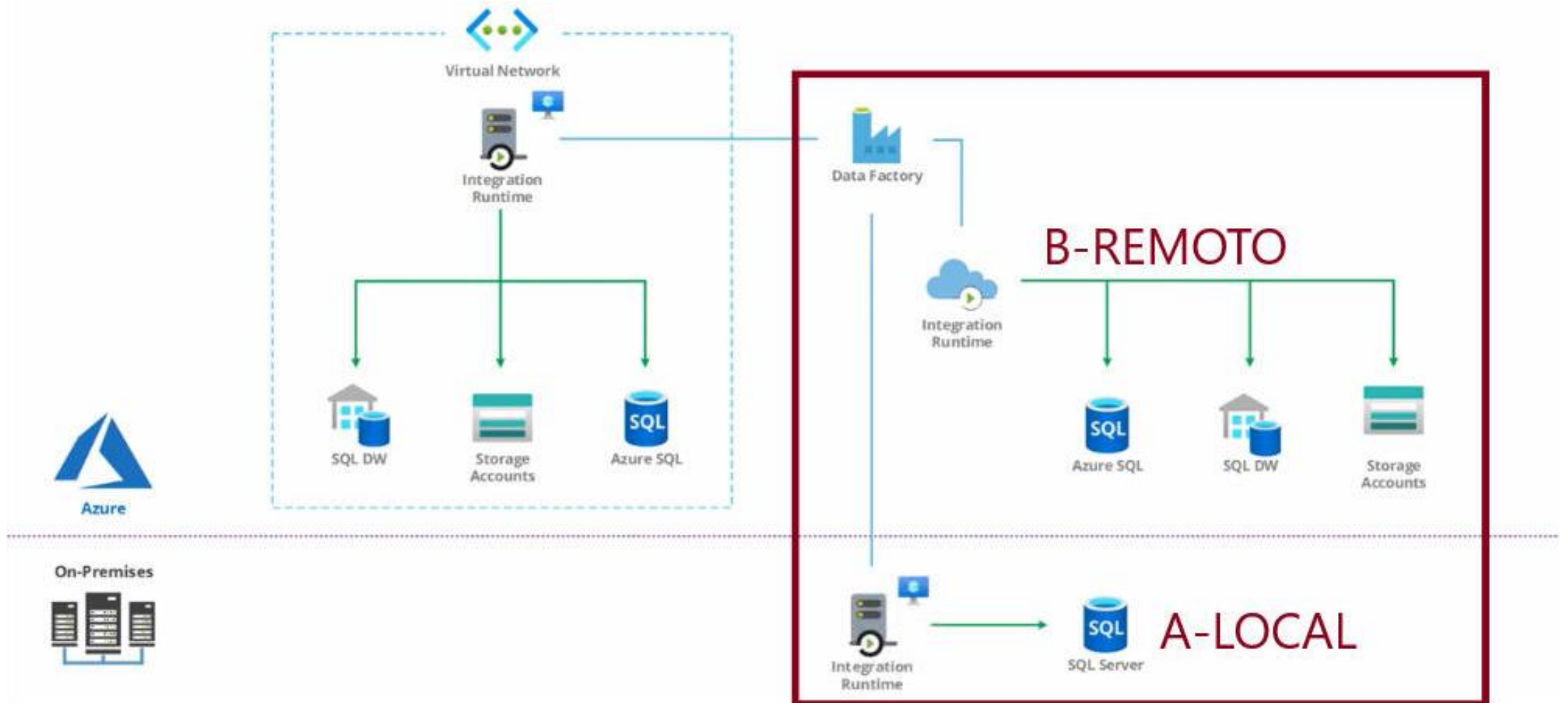


# Integration Runtime - AutoResolveIntegrationRuntime





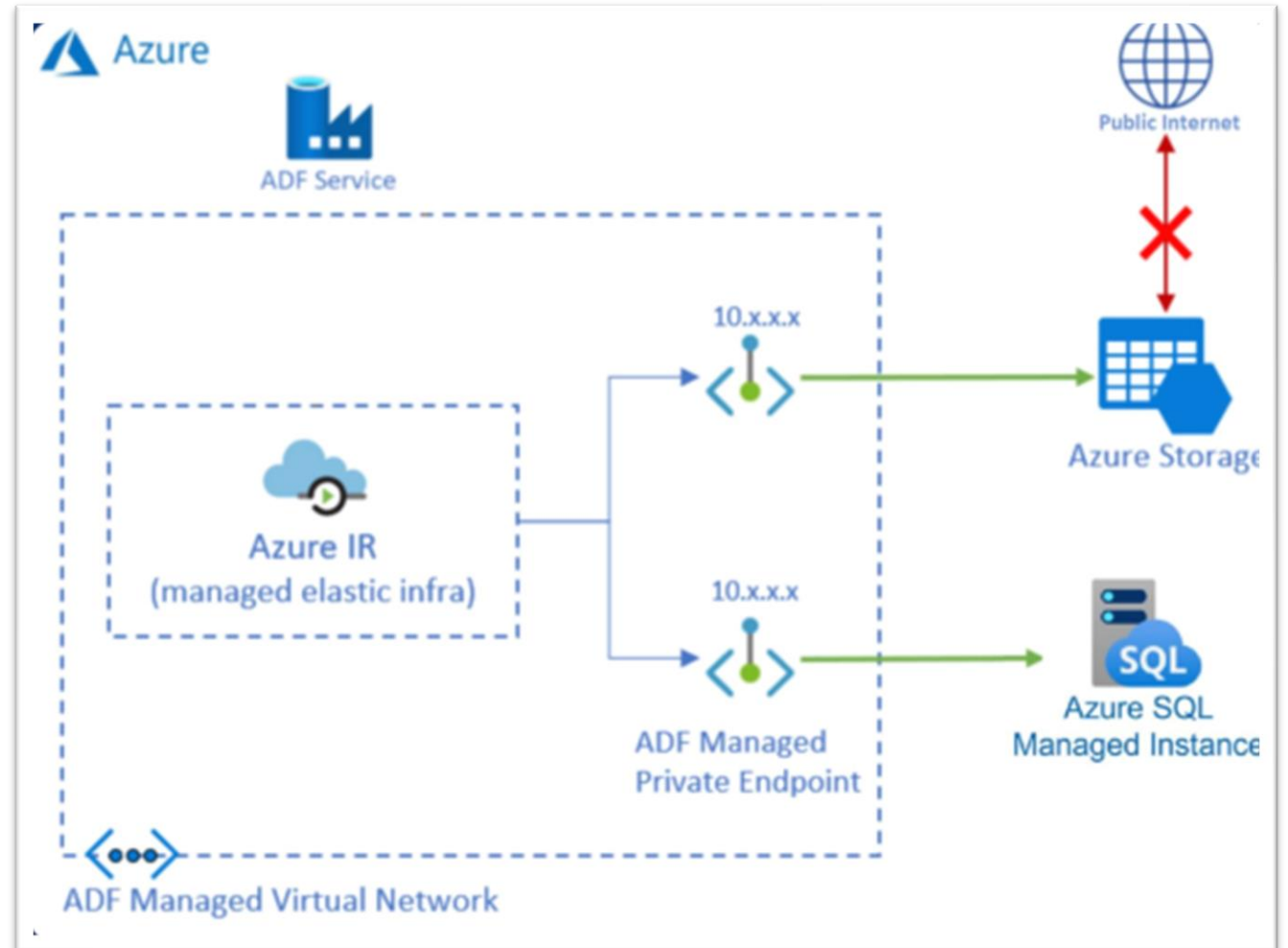
# Integration Runtime – Self-Hosted



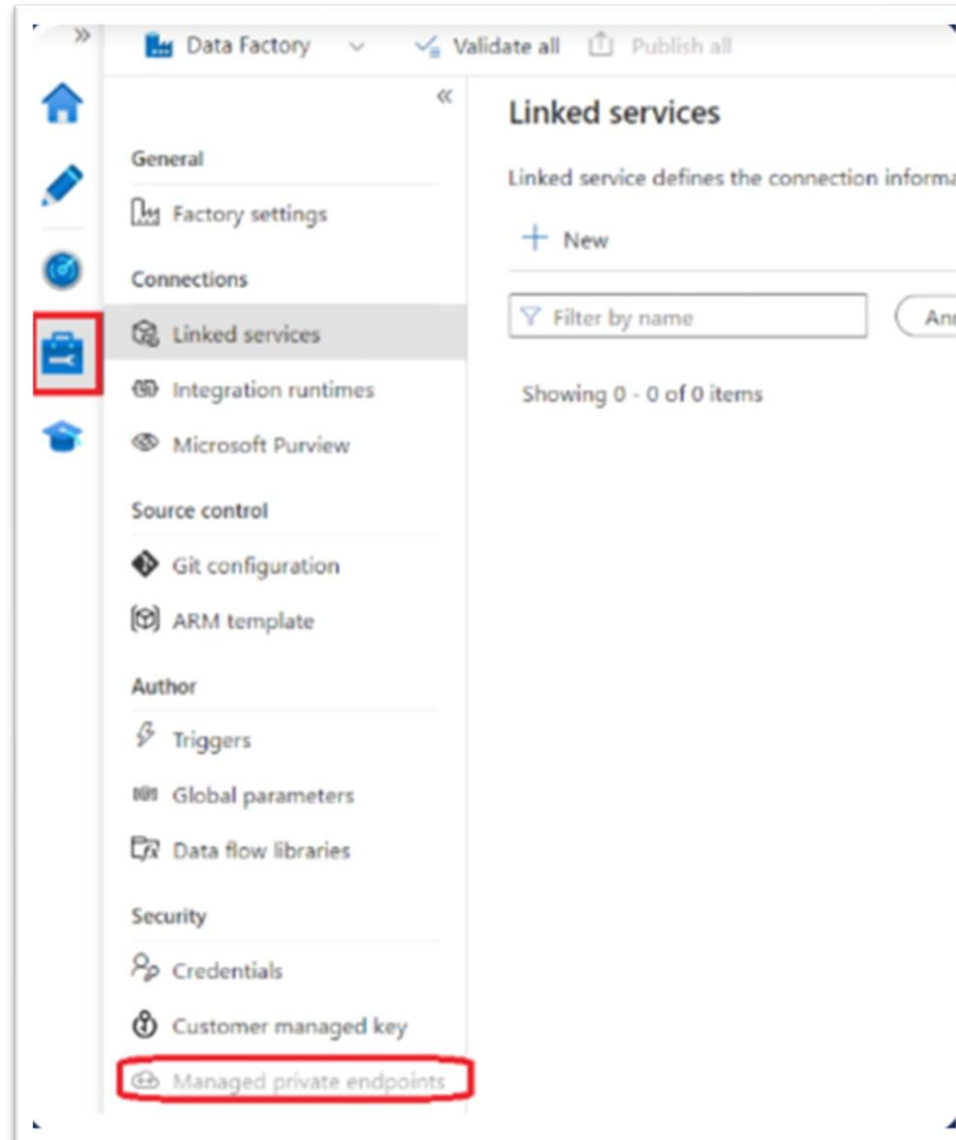
# Integration Runtime – Managed Virtual Network

## Managed Private Endpoint

Los puntos de conexión privados administrados en Azure Data Factory son conexiones privadas a recursos de Azure, gestionadas automáticamente por Data Factory. No necesitas configurar tu propia red virtual o subred, ya que Microsoft se encarga de ello.

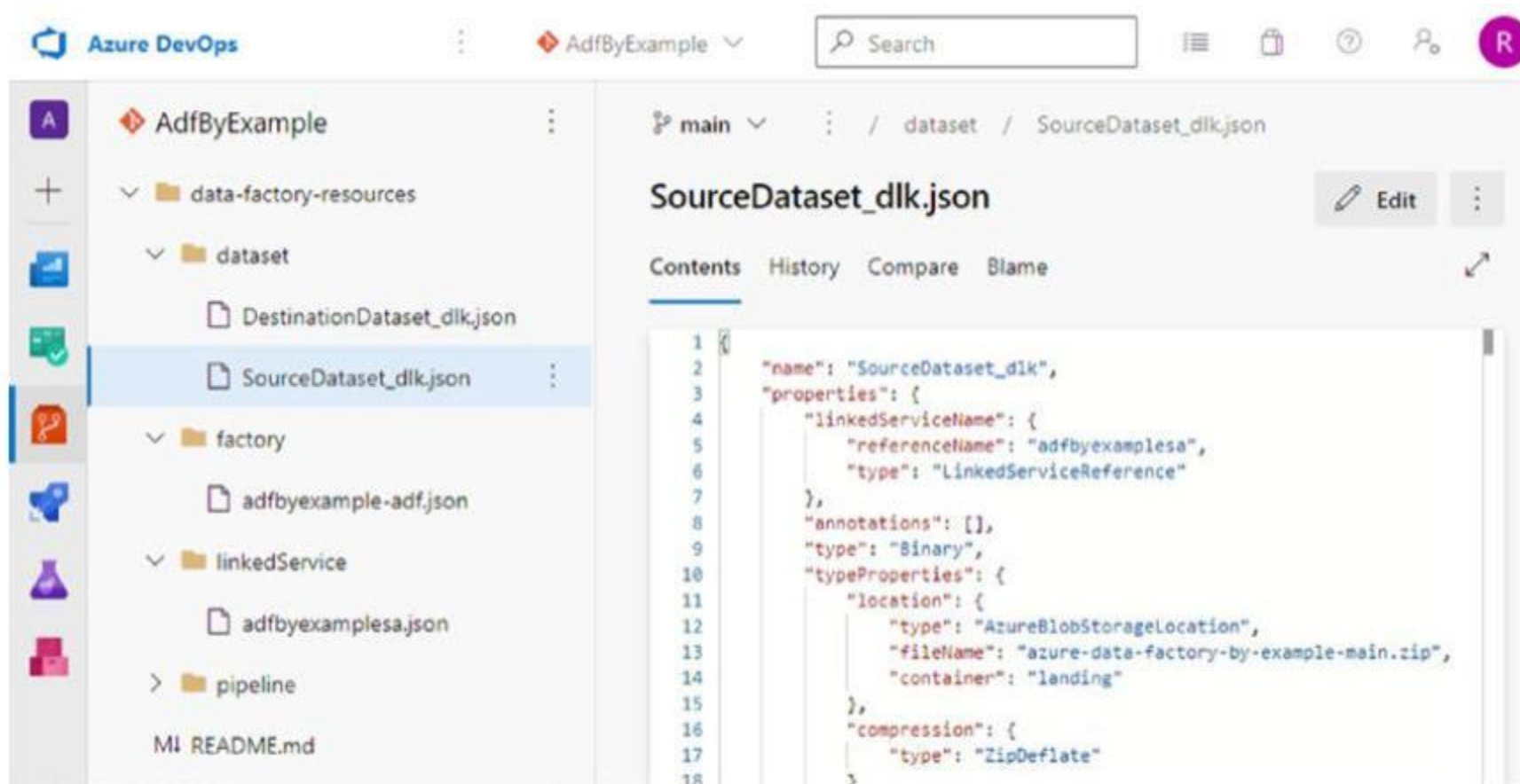


# Integration Runtime – Managed Virtual Network



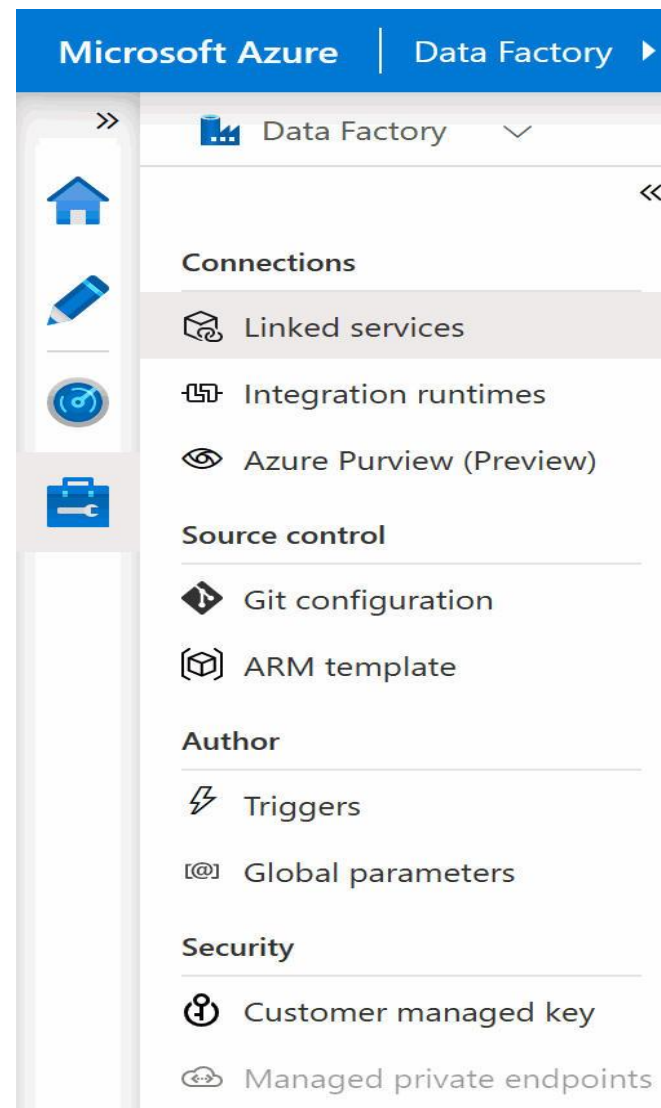
# CI/CD y Publicación

Data Factory ofrece compatibilidad total con CI/CD de sus canalizaciones de datos mediante Azure DevOps y GitHub. Esto le permite desarrollar y distribuir incrementalmente los procesos ETL antes de publicar el producto terminado.



# Componentes de Azure Data Factory en la página de administración

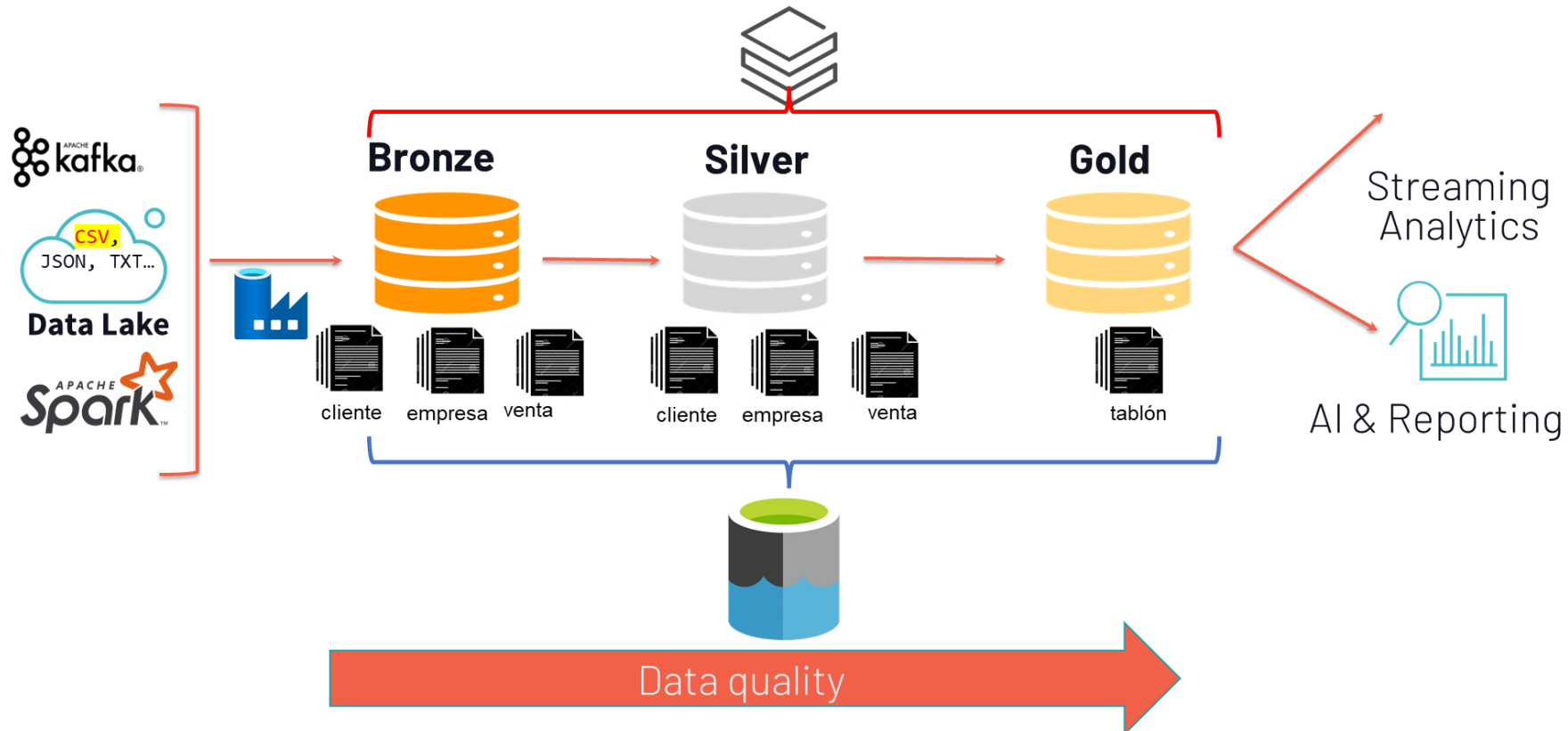
En el lado izquierdo de la página de Administración , verá los componentes y servicios que puede crear y configurar. Nos centraremos en dos de los componentes principales de esta publicación: los servicios vinculados y los disparadores.



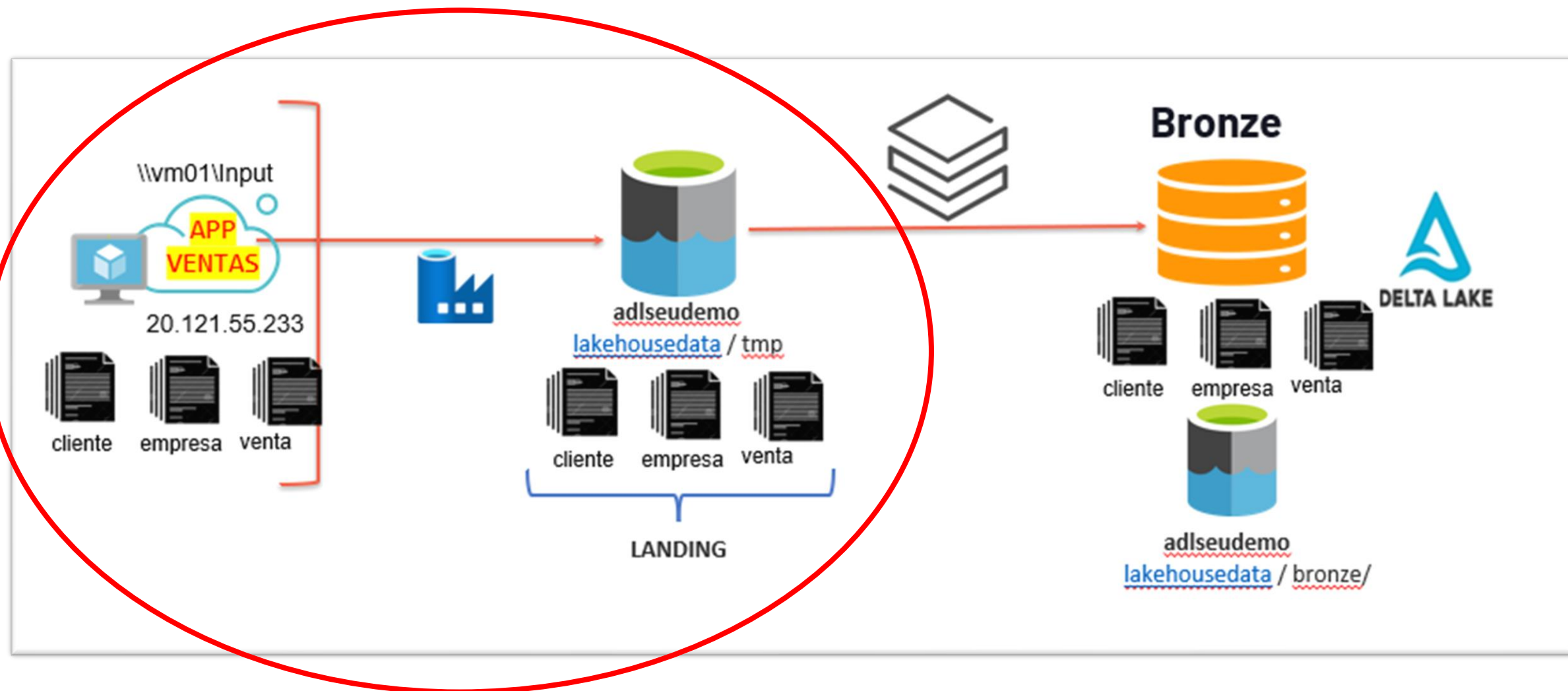
## LABORATORIO: Ingesta con Datafactory (Arquitectura Delta)

### Caso de Uso - Análisis descriptivo:

Una analista de negocio de una empresa retail desea analizar los datasets de Cliente, Empresa, Transacción de su sistema de ventas y generar un reporte automático donde se tenga las Ventas por rango de edades de los clientes y la empresa a las que pertenecen los clientes.



## LABORATORIO: Ingesta con Datafactory



# RONDAS DE PREGUNTAS





**¡GRACIAS!**

