

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***

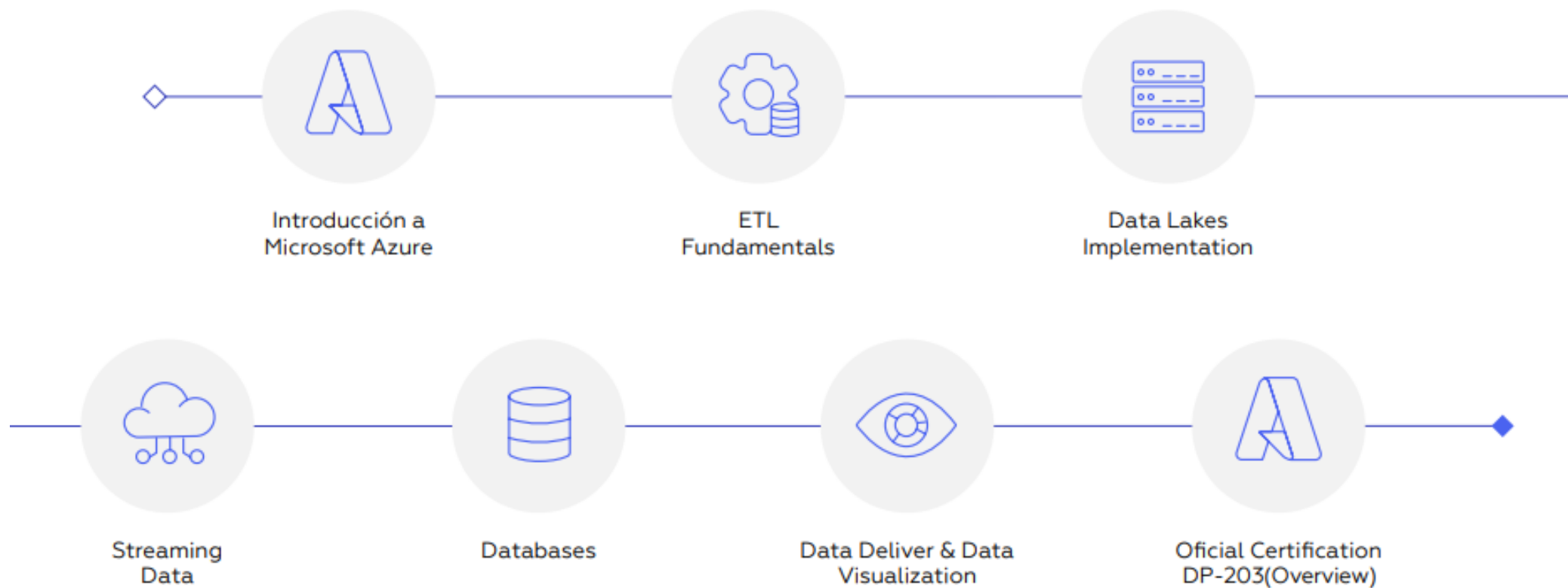
*08:50 PM – 09:00 PM **Pausa Activa***

*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

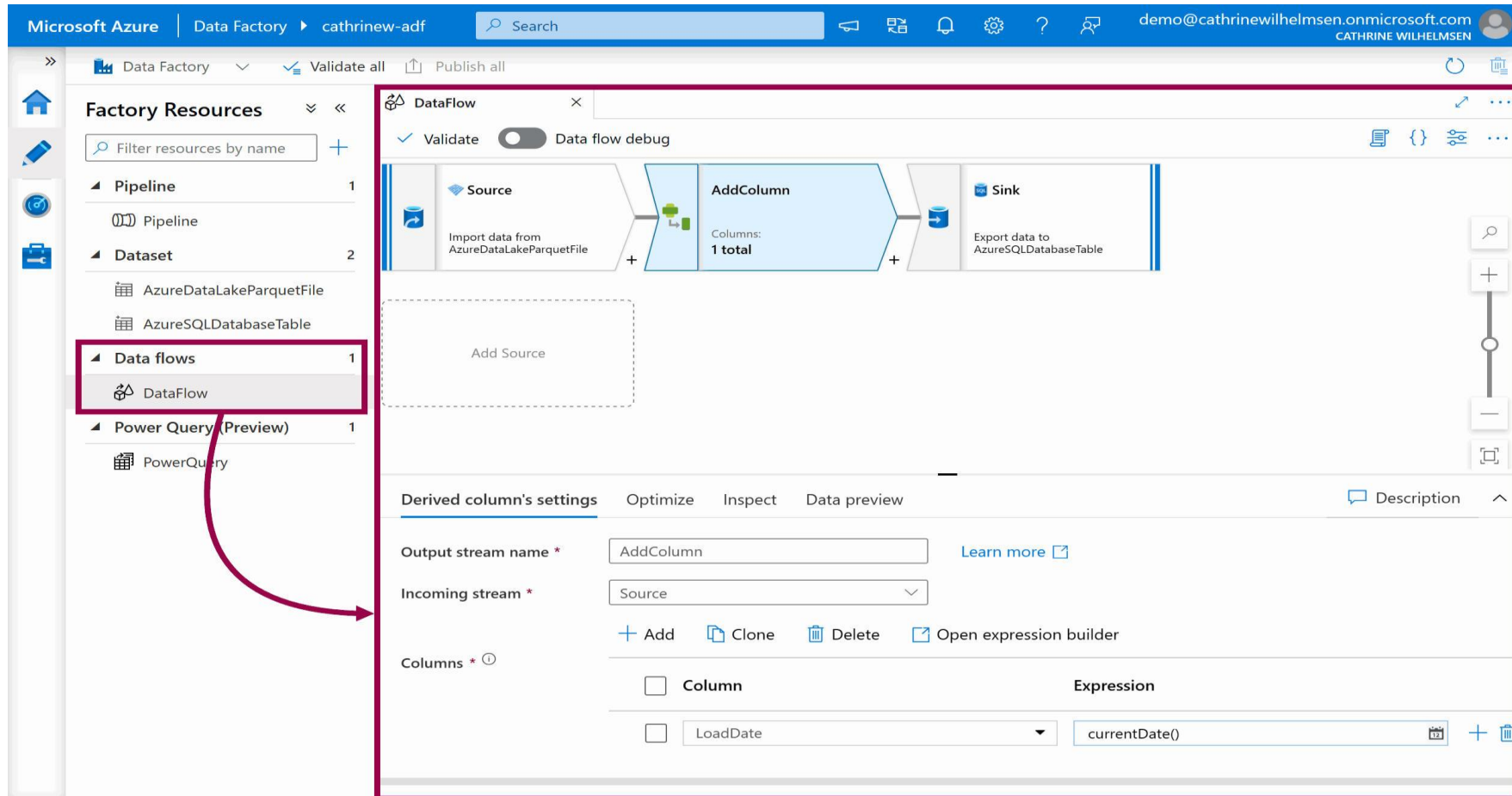
Dataflow ETL en
Batch

02

Laboratorio 03:
Implementación de
un ETL Básico con
Dataflow

Data Flows

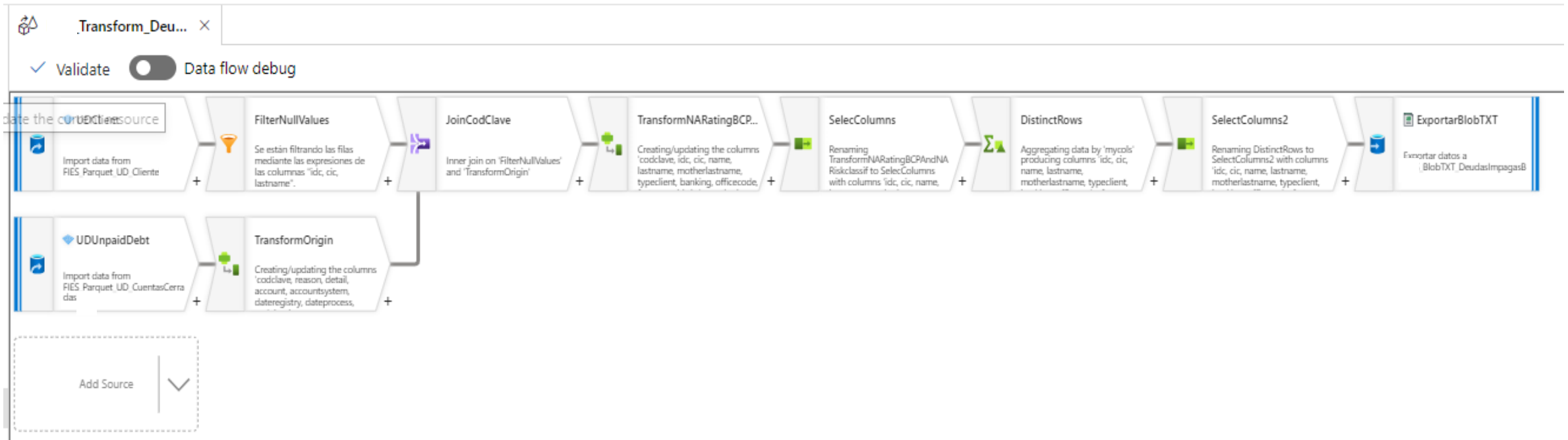
Los flujos de datos son un tipo especial de actividad para crear transformaciones de datos. Puede transformar los datos en varios pasos mediante un editor visual, sin tener que escribir ningún otro código que no sean expresiones de datos.



The screenshot displays the Microsoft Azure Data Factory DataFlow editor. The left sidebar shows the 'Factory Resources' tree with 'Data flows' selected. The main canvas shows a DataFlow pipeline with a Source, a transformation named 'AddColumn', and a Sink. The 'AddColumn' transformation is selected, and its settings are shown in the bottom pane. The settings include 'Output stream name' (AddColumn), 'Incoming stream' (Source), and 'Columns' (LoadDate). The 'Columns' section shows a table with 'Column' and 'Expression' headers, with 'LoadDate' mapped to 'currentDate()'. A red box highlights the 'Data flows' section in the left sidebar, and a red arrow points from it to the 'AddColumn' transformation.

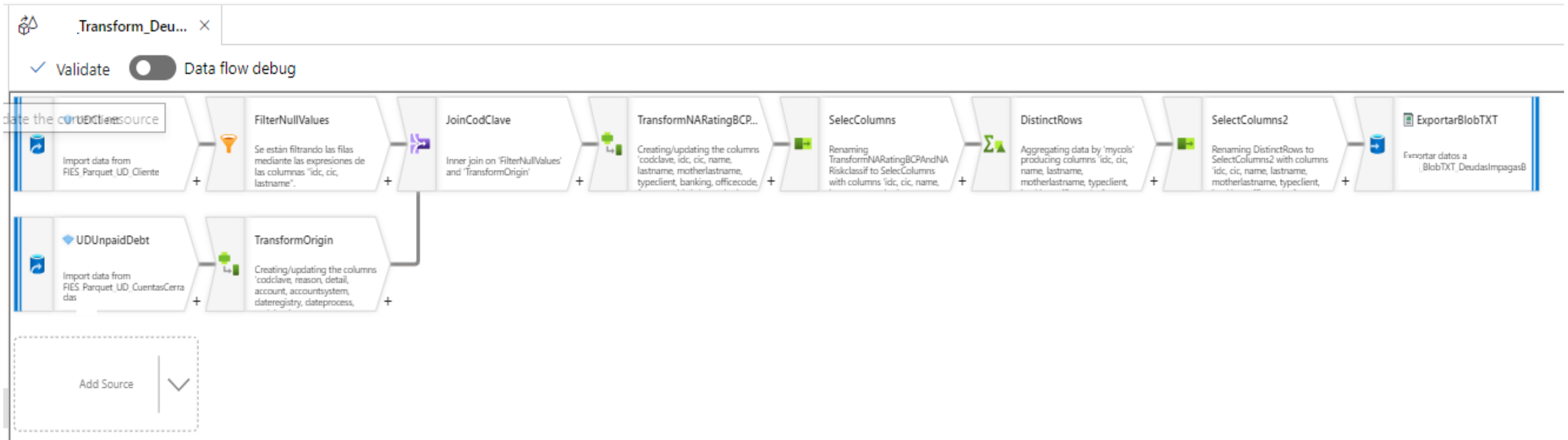
Data Flows

- **Propósito:** Estas son actividades dentro de los flujos de datos mapeados que transforman los datos a medida que se mueven a través del conducto.
- **Transformaciones comunes:** Incluye actividades como uniones, agregaciones, filtros y búsquedas dentro de un flujo de datos.



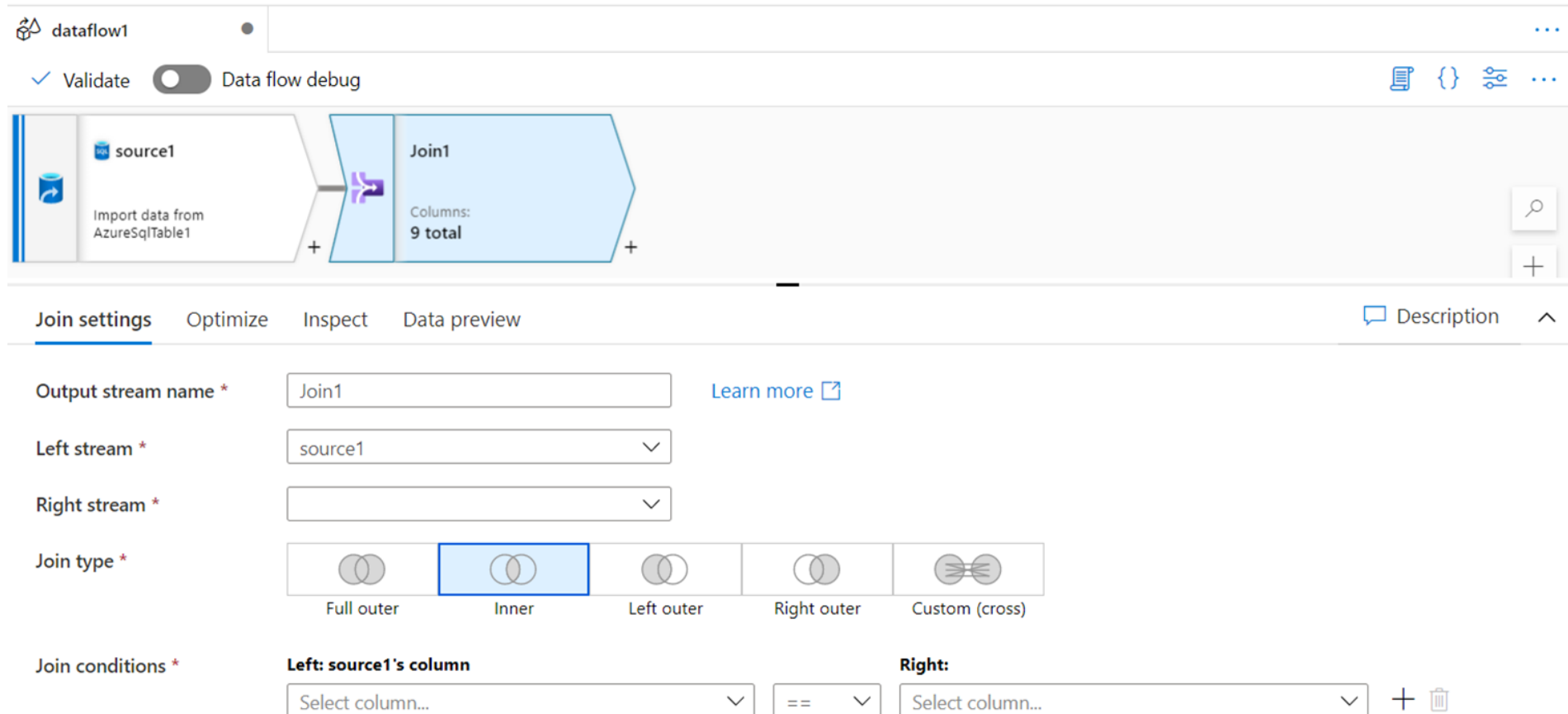
Data Flows

- **Caso de uso:** Cuando necesitas limpiar, agregar o remodelar datos antes de cargarlos en el destino final.
- Realizar operaciones de configuración complejas utilizando la interfaz de usuario del flujo de datos para configurar las transformaciones de datos.



Data Flows - Join

- **Descripción:** Combina datos de múltiples fuentes en una sola salida.
- **Uso:** Se utiliza para integrar datos de diferentes orígenes en un solo conjunto de datos.



The screenshot shows the Azure Data Studio interface for a Data Flow task named 'Join1'. The task is configured to join data from 'source1' (Import data from AzureSqlTable1) with another stream. The output stream is named 'Join1' and has 9 total columns. The 'Join settings' tab is active, showing the following configuration:

- Output stream name ***: Join1
- Left stream ***: source1
- Right stream ***: (Empty)
- Join type ***: Inner (Selected)
- Join conditions ***:
 - Left: source1's column
 - Right: (Empty)

The 'Join type' section shows five options: Full outer, Inner (selected), Left outer, Right outer, and Custom (cross). The 'Join conditions' section shows a table with columns for 'Left: source1's column' and 'Right: (Empty)', with a comparison operator '==' selected.

Data Flows - Split

- **Descripción:** Divide un conjunto de datos en múltiples flujos basados en condiciones específicas.
- **Uso:** Útil para dirigir diferentes subconjuntos de datos a diferentes destinos o para aplicar diferentes transformaciones.

✓ Validate ☐ Data flow debug

stream01
Columns: 9 total

stream02
Add cases to distribute the data in multiple groups

Conditional split settings | Optimize | Inspect | Data preview

Output stream name * ConditionalSplit1 [Learn more](#)

Incoming stream * source1

Split on ☒ First matching condition ☐ All matching conditions

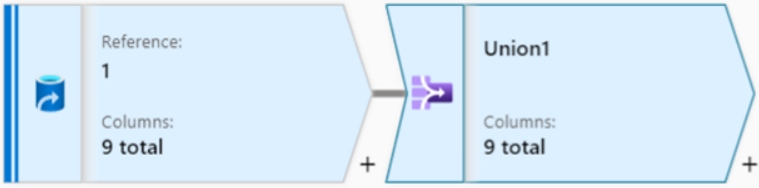
Split condition

Stream names	Condition
stream01	ANY +

Data Flows - Union

- **Descripción:** Combina datos de múltiples flujos en un solo flujo.
- **Uso:** Útil para consolidar datos de diferentes orígenes en un solo conjunto de datos.

✓ Validate
⏻ Data flow debug
📄 {} ⚙️ ...



Union settings
Optimize
Inspect
Data preview
Description ^

Output stream name * [Learn more](#)

Incoming stream *

Union by * ⓘ ☒ Name ☐ Position

Union with **Streams**

Data Flows - Lookup

- **Descripción:** Permite buscar y recuperar datos de otra tabla o fuente de datos.
- **Uso:** Se utiliza para enriquecer los datos con información adicional de otras fuentes.

✓ Validate ☐ Data flow debug

Lookup1

Reference: 1
Columns: 9 total

Lookup1
Columns: 9 total

Lookup settings Optimize Inspect Data preview

Output stream name * Lookup1 [Learn more](#)

Primary stream * source1

Lookup stream *

Match multiple rows ☐ ⓘ

Match on * Any row

Lookup conditions *

Left: source1's column

Select column... == Select column...

Right:

Select column...

Data Flows - Derived Column

- **Descripción:** Crea nuevas columnas basadas en expresiones o cálculos.
- **Uso:** Permite enriquecer los datos con información adicional derivada de las columnas existentes.

✓ Validate
⏻ Data flow debug
📄 {} ⚙️ ...

source1
 Import data from AzureSqlTable1

+

DerivedColumn1
 Columns: 9 total

+

🔍 +

Derived column's settings

Optimize

Inspect

Data preview

🗨 Description ^

Output stream name *

DerivedColumn1

[Learn more](#)

Incoming stream *

source1

+ Add

📄 Clone

🗑 Delete

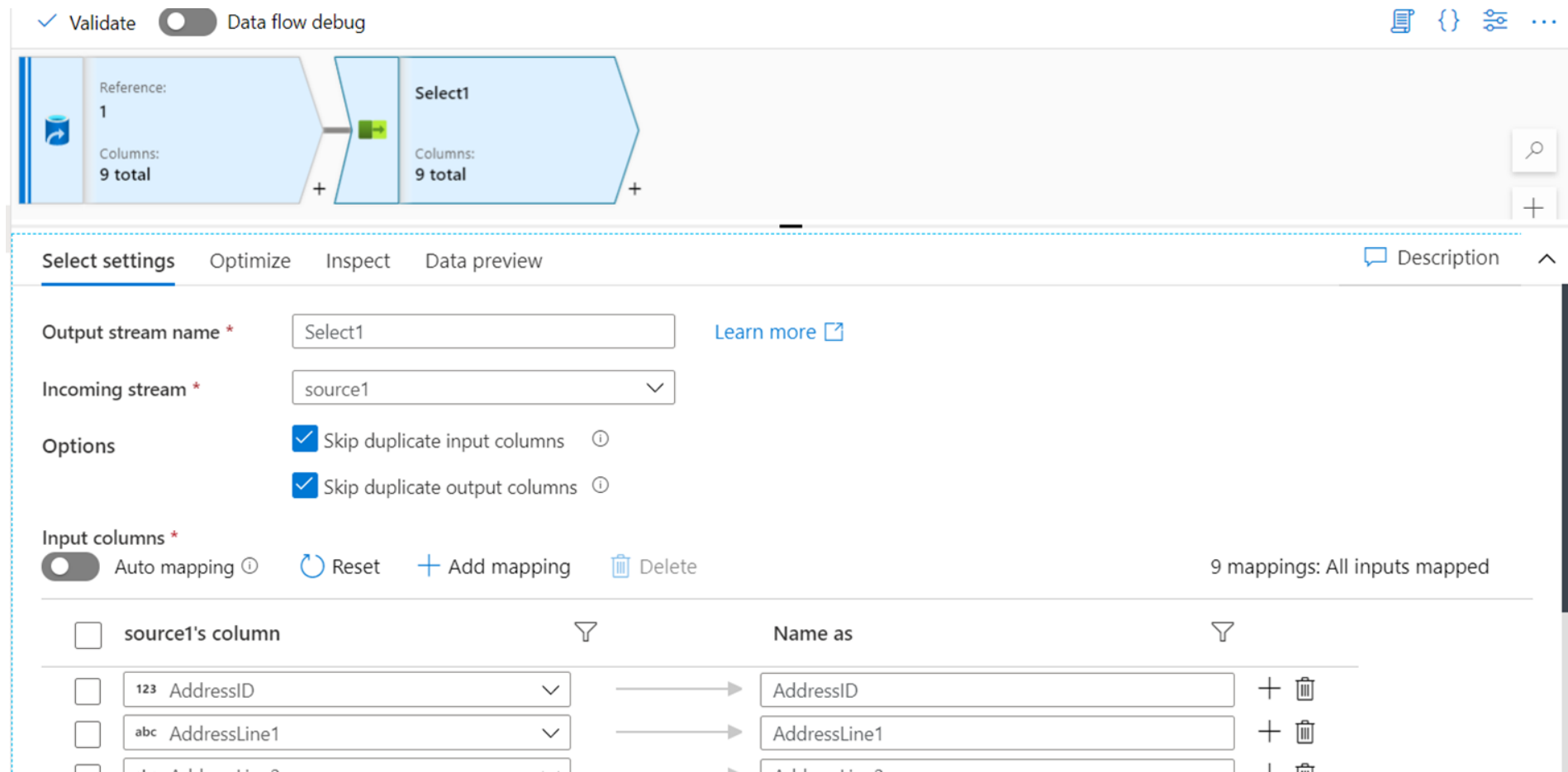
🔗 Open expression builder

Columns * ⓘ

<input type="checkbox"/>	Column	Expression
<input type="checkbox"/>	! Add or select a column...	<div>Enter expression...</div> <div>ANY + 🗑</div>

Data Flows - Select

- **Descripción:** Permite seleccionar columnas específicas de un conjunto de datos.
- **Uso:** Útil para reducir el número de columnas en un conjunto de datos, manteniendo solo las necesarias.




The screenshot shows the 'Select' transform configuration in Google Cloud Data Studio. At the top, there's a 'Validate' button and a 'Data flow debug' toggle. Below this, a visual representation of the data flow shows an input stream 'source1' with 9 columns feeding into a 'Select1' transform, which also has 9 columns. The 'Select settings' tab is active, showing the 'Output stream name' as 'Select1' and the 'Incoming stream' as 'source1'. Under 'Options', both 'Skip duplicate input columns' and 'Skip duplicate output columns' are checked. The 'Input columns' section shows 'Auto mapping' is turned on, and a table lists the mappings from 'source1's column' to the output column names.

source1's column	Name as
123 AddressID	AddressID
abc AddressLine1	AddressLine1


Data Flows - Aggregation

- **Descripción:** Permite realizar operaciones de agregación como sumas, promedios, conteos, etc.
- **Uso:** Ideal para resumir datos y obtener métricas clave.

✓ Validate
⏻ Data flow debug
📄 {} ⚙️ ...


 Reference:
1
Columns:
9 total

+


 Aggregate1
Columns:
0 total

+

🔍 +

Aggregate settings
Optimize
Inspect
Data preview
Description ^

Output stream name *

[Learn more](#)


Incoming stream *

Group by

Aggregates

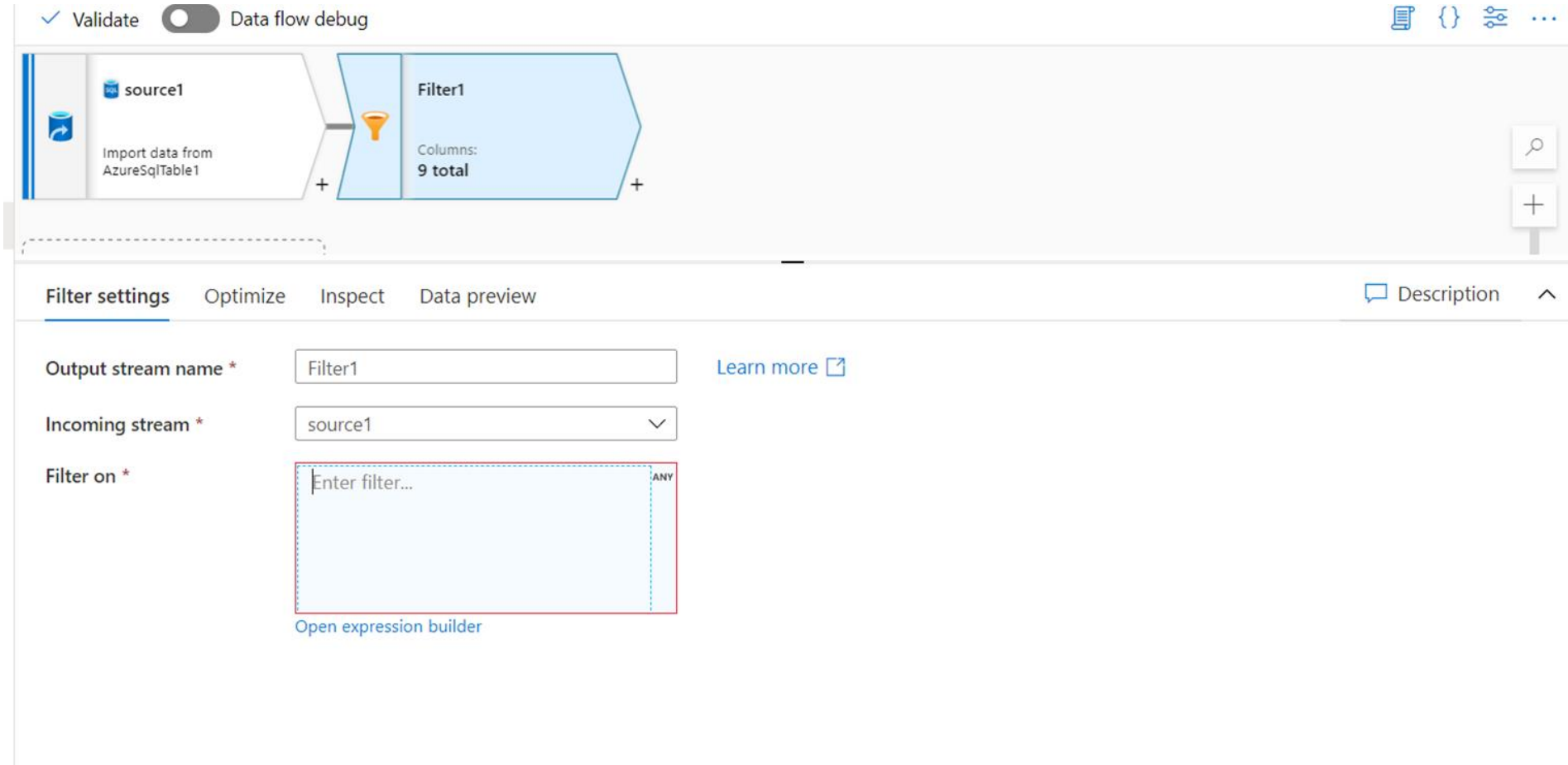
Columns

Name as

+
 

Data Flows - Filter

- **Descripción:** Filtra filas basadas en condiciones específicas.
- **Uso:** Útil para eliminar datos no deseados o irrelevantes antes de realizar otras transformaciones.



The screenshot displays the 'Filter' data flow configuration in Azure Data Studio. At the top, there are tabs for 'Filter settings', 'Optimize', 'Inspect', and 'Data preview', with 'Filter settings' being the active tab. A 'Description' button is located on the right. The main configuration area includes:

- Output stream name ***: A text field containing 'Filter1'.
- Incoming stream ***: A dropdown menu showing 'source1'.
- Filter on ***: A large text area with a placeholder 'Enter filter...'. To its right is a small 'ANY' button.

Below the text area is a link that says 'Open expression builder'. Above the configuration area, a visual pipeline diagram shows a source named 'source1' (labeled 'Import data from AzureSqlTable1') connected to a filter node labeled 'Filter1' (labeled 'Columns: 9 total'). The interface also features a 'Validate' checkbox (checked) and a 'Data flow debug' toggle (disabled) at the top left, and various utility icons at the top right.

Data Flows - Sort

- **Descripción:** Ordena los datos según criterios definidos, como orden ascendente o descendente.
- **Uso:** Facilita la organización de datos para análisis posteriores.

The screenshot shows the configuration interface for the 'Sort' widget in Google Cloud Data Studio. At the top, there are tabs for 'Sort settings', 'Optimize', 'Inspect', and 'Data preview', with 'Sort settings' being the active tab. Below the tabs, the configuration fields are as follows:

- Output stream name ***: A text input field containing 'Sort1'.
- Incoming stream ***: A dropdown menu showing 'source1'.
- Options ***: Two checkboxes, 'Case insensitive' and 'Sort only within partition', both of which are currently unchecked.
- Sort conditions ***: A section for defining sort criteria. It includes:
 - source1's column**: A dropdown menu with 'Select column...' selected.
 - Order**: A dropdown menu with 'Ascending' selected.
 - Nulls first**: A checkbox that is checked.

At the top of the interface, there is a 'Validate' button (checked) and a 'Data flow debug' toggle (unchecked). On the right side, there are icons for documentation, code, settings, and a search icon. Below the main configuration area, there is a 'Description' tab and a search icon.

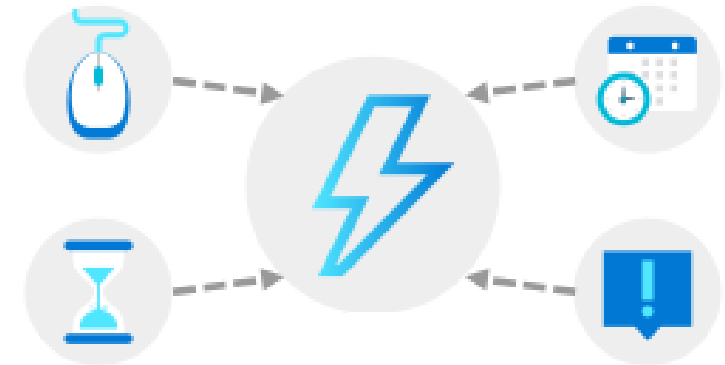
Data Factory - Trigger

Un **Trigger** es una característica que le permite automatizar la ejecución de sus pipelines basándose en condiciones o horarios específicos. Los triggers determinan cuándo debe correr un pipeline, permitiendo un mayor control y eficiencia en los procesos de integración y transformación de datos.

Tipos de Triggers en Azure Data Factory:

- Scheduled Trigger
- Tumbling Window Trigger
- Storage Event Trigger
- Custom Event Trigger

Triggers



Data Factory - Trigger

Add triggers

Choose trigger...

Search

+ New

New trigger

Name *
trigger1

Description

Type *
Schedule

Filter...

Schedule

Tumbling window

Storage events

Custom events

every 15 Minute(s)

☐ Specify an end date

Annotations

OK Cancel

Edit trigger

Name *
trigger1

Description

Type *
ScheduleTrigger

Start date * ⓘ
9/29/2024, 8:22:00 PM

Time zone * ⓘ
Bogota, Lima, Quito (UTC-5)

Recurrence * ⓘ
Every 5 Minute(s)

☒ Specify an end date

End On * ⓘ
9/29/2024, 8:30:00 PM

OK Cancel

¡Sigue aprendiendo!

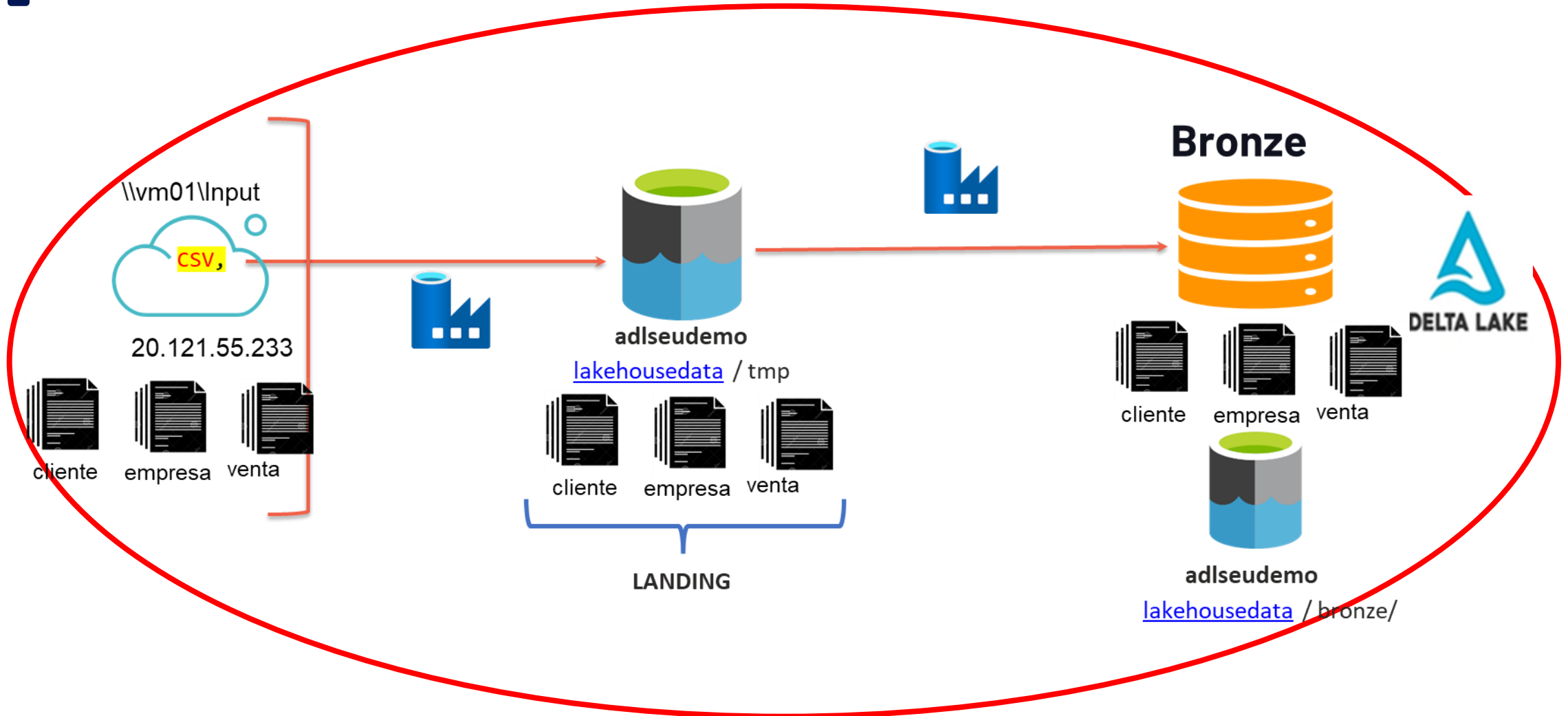
[1. Introducción a las transformaciones en el flujo de datos de asignación - Azure Data Factory & Azure Synapse | Microsoft Learn](#)

[2. Data Flow Transformations in Azure Data Factory \(sqlshack.com\)](#)

[3. Transform data using a mapping data flow - Azure Data Factory | Microsoft Learn](#)

[4. Select transformation in mapping data flow - Azure Data Factory & Azure Synapse | Microsoft Learn](#)

LABORATORIO: ETL con Data Factory



RONDAS DE PREGUNTAS



¡GRACIAS!

