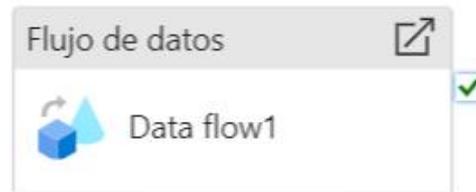




Ejemplo de ETL paso a paso con la herramienta Azure Data Factory



IMPORTANTE

© Copyright **Lituus Spa**

Este documento es propiedad de Lituus Spa y su contenido es confidencial. Este no puede ser reproducido, en su totalidad o parcialmente, ni mostrado a terceros, ni utilizado para otros propósitos que los que han originado su entrega, sin el previo permiso escrito de **Lituus Spa**.

Archivo a procesar

Financial Sample – Archivo Excel

1	Segment	Country	Product	Discount Band	Units Sold	Manufacturi	Sale Price	Gross Sales	Discounts	Sales	COGS	Profit	Dat
2	Government	Canada	Carretera	None	1618,5	\$ 3,00	\$ 20,00	\$ 32.370,00	\$ -	\$ 32.370,00	\$ 16.185,00	\$ 16.185,00	01
3	Government	Germany	Carretera	None	1321	\$ 3,00	\$ 20,00	\$ 26.420,00	\$ -	\$ 26.420,00	\$ 13.210,00	\$ 13.210,00	01
4	Midmarket	France	Carretera	None	2178	\$ 3,00	\$ 15,00	\$ 32.670,00	\$ -	\$ 32.670,00	\$ 21.780,00	\$ 10.890,00	01
5	Midmarket	Germany	Carretera	None	888	\$ 3,00	\$ 15,00	\$ 13.320,00	\$ -	\$ 13.320,00	\$ 8.880,00	\$ 4.440,00	01
6	Midmarket	Mexico	Carretera	None	2470	\$ 3,00	\$ 15,00	\$ 37.050,00	\$ -	\$ 37.050,00	\$ 24.700,00	\$ 12.350,00	01
7	Government	Germany	Carretera	None	1513	\$ 3,00	\$ 350,00	\$ 529.550,00	\$ -	\$ 529.550,00	\$ 393.380,00	\$ 136.170,00	01
8	Midmarket	Germany	Montana	None	921	\$ 5,00	\$ 15,00	\$ 13.815,00	\$ -	\$ 13.815,00	\$ 9.210,00	\$ 4.605,00	01
9	Channel Partners	Canada	Montana	None	2518	\$ 5,00	\$ 12,00	\$ 30.216,00	\$ -	\$ 30.216,00	\$ 7.554,00	\$ 22.662,00	01
10	Government	France	Montana	None	1899	\$ 5,00	\$ 20,00	\$ 37.980,00	\$ -	\$ 37.980,00	\$ 18.990,00	\$ 18.990,00	01
11	Channel Partners	Germany	Montana	None	1545	\$ 5,00	\$ 12,00	\$ 18.540,00	\$ -	\$ 18.540,00	\$ 4.635,00	\$ 13.905,00	01
12	Midmarket	Mexico	Montana	None	2470	\$ 5,00	\$ 15,00	\$ 37.050,00	\$ -	\$ 37.050,00	\$ 24.700,00	\$ 12.350,00	01
13	Enterprise	Canada	Montana	None	2665,5	\$ 5,00	\$ 125,00	\$ 333.187,50	\$ -	\$ 333.187,50	\$ 319.860,00	\$ 13.327,50	01
14	Small Business	Mexico	Montana	None	958	\$ 5,00	\$ 300,00	\$ 287.400,00	\$ -	\$ 287.400,00	\$ 239.500,00	\$ 47.900,00	01
15	Government	Germany	Montana	None	2146	\$ 5,00	\$ 7,00	\$ 15.022,00	\$ -	\$ 15.022,00	\$ 10.730,00	\$ 4.292,00	01
16	Enterprise	Canada	Montana	None	345	\$ 5,00	\$ 125,00	\$ 43.125,00	\$ -	\$ 43.125,00	\$ 41.400,00	\$ 1.725,00	01
17	Midmarket	United States of America	Montana	None	615	\$ 5,00	\$ 15,00	\$ 9.225,00	\$ -	\$ 9.225,00	\$ 6.150,00	\$ 3.075,00	01
18	Government	Canada	Paseo	None	292	\$ 10,00	\$ 20,00	\$ 5.840,00	\$ -	\$ 5.840,00	\$ 2.920,00	\$ 2.920,00	01
19	Midmarket	Mexico	Paseo	None	974	\$ 10,00	\$ 15,00	\$ 14.610,00	\$ -	\$ 14.610,00	\$ 9.740,00	\$ 4.870,00	01
20	Channel Partners	Canada	Paseo	None	2518	\$ 10,00	\$ 12,00	\$ 30.216,00	\$ -	\$ 30.216,00	\$ 7.554,00	\$ 22.662,00	01
21	Government	Germany	Paseo	None	1006	\$ 10,00	\$ 350,00	\$ 352.100,00	\$ -	\$ 352.100,00	\$ 261.560,00	\$ 90.540,00	01

Creación data factory

- Una vez dentro del portal <https://portal.azure.com/> el primer paso es crear una Factoría de Datos (Data Factory), se debe seleccionar ese servicio ingresando “Factoría de Datos” en el buscador.

The screenshot shows the Microsoft Azure portal interface. At the top, the search bar contains the text "Factoría de datos". Below the search bar, a dropdown menu lists various services, with "Factorías de datos" highlighted. To the left of the search bar, there is a button labeled "Crear un recurso". Below the search bar, there is a section titled "Recursos" with a sub-section "Factorías de datos" containing a table of existing data factories. The table has columns for "Nombre" and "Tipo". One entry is visible: "Ins01" with the type "Factoría de datos (V2)".

Luego se abre la ventana de selección de los recursos Data Factory existentes, para crear uno nuevo, presionar el botón “Crear”

Creación data factory

Para crear el Nuevo Data Factory, se selecciona “Revisar y Crear” y “Crear”, después de configurar los tres datos necesarios (nombre, región y versión), además de elegir un grupo de recursos o carpeta, la cual puede ser una nueva o una ya existente.

Nota: Puede que este proceso se demore algunos minutos.

[Inicio](#) > [Factorías de datos](#) >

Crear Data Factory

One-click to create data factory with sample pipeline and datasets. [Try it](#)

Detalles del proyecto

Seleccione la suscripción para administrar recursos implementados y los costes. Use los grupos de recursos como carpetas para organizar y administrar todos los recursos.

Suscripción * ⓘ

Azure subscription 1

Grupo de recursos * ⓘ

g01

[Crear nuevo](#)

Detalles de la instancia

Nombre * ⓘ

IDF01

Región * ⓘ

East US

Versión * ⓘ

V2

[Revisar y crear](#)

[< Anterior](#)

[Siguiente: Configuración de Git >](#)

Creación data factory

- Para revisar el recurso Data Factory creado, se debe seleccionar el recurso de la lista:

Factorías de datos

explodat (explodat.onmicrosoft.com)

Crear Administrar vista Actualizar Exportar a CSV Abrir consulta | Asignar etiquetas

Filtrar por cual... es igual a **todo** Grupo de recursos es igual a **todo** Ubicación es igual a **todo** Agregar filtro

Sin agrupar Vista de lista

Tipo ↑↓	Suscripción ↑↓	Grupo de recursos ↑↓	Ubicación ↑↓	
Factoría de datos (V2)	Azure subscription 1	g01	East US	...
Factoría de datos (V2)	Azure subscription 1	g01	Central US	...

IDF01

Nombre ↑

IDF01 Ver

Ins01 Detalles del recurso

Tipo

Recursos

- Para revisar el recurso Data Factory creado , se debe seleccionar el recurso de la lista:

Factorías de datos  ... ✕

explodat (explodat.onmicrosoft.com)

 Crear  Administrar vista  Actualizar  Exportar a CSV  Abrir consulta |  Asignar etiquetas

Filtrar por cualc... es igual a **todo** Grupo de recursos es igual a **todo** ✕ Ubicación es igual a **todo** ✕  Agregar filtro

 **IDF01** 

Nombre ↑  Ver

 **IDF01** Detalles del recurso

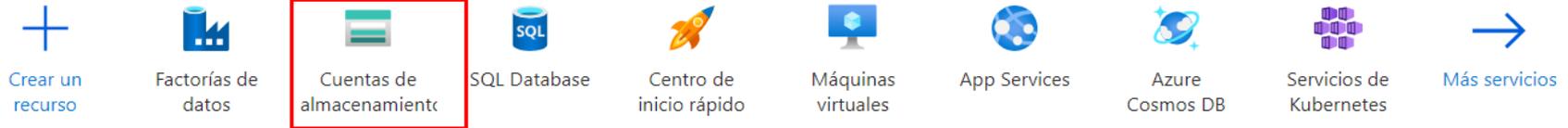
 **Ins01** Tipo

Tipo ↑↓	Suscripción ↑↓	Grupo de recursos ↑↓	Ubicación ↑↓	
Factoría de datos (V2)	Azure subscription 1	g01	East US	...
Factoría de datos (V2)	Azure subscription 1	g01	Central US	...

Cuenta de almacenamiento

Para cargar archivos locales desde nuestro computador, debemos crear una cuenta de almacenamiento, dentro de los servicios de azure, similar a como se creó el data factory

Servicios de Azure



Cuentas de almacenamiento ★ ⋮

UNIVERSIDAD ANDRES BELLO (uandresbelloedu.onmicrosoft.com)

+ Crear ↶ Restaurar ⚙ Administrar vista ↻ Actualizar ↓ Exportar a CSV 🔗 Abrir consulta | 🏷 Asignar etiquetas 🗑 Eliminar

Filtrar por cualquier ca...

Suscripción es igual a **todo**

Grupo de recursos es igual a **todo** ✕

Ubicación es igual a **todo** ✕

+🔍 Agregar filtro

Cuenta de almacenamiento

- Seleccionamos grupo de recursos, región, rendimiento (estándar o premium), y le asignamos nombre a la storage account
- Similar a cuando creamos el data factory, finalizamos con “revisar y crear”

✔ Se completó la implementación

Crear una cuenta de almacenamiento

Datos básicos Opciones avanzadas Redes Protección de datos Cifrado Etiquetas Revisar

Detalles del proyecto

Seleccione la suscripción en la que se creará la nueva cuenta de almacenamiento. Elija un grupo de recursos nuevo o uno ya existente para organizar y administrar la cuenta de almacenamiento junto con otros recursos.

Suscripción * Azure for Students

Grupo de recursos * g01

[Crear nuevo](#)

Detalles de la instancia

Nombre de la cuenta de almacenamiento ejemplobci1

Región (US) East US 2

Implementar en una zona perimetral

Estándar: Opción recomendada para la mayoría de los escenarios (cuenta de uso general v2)

Prémium: Se recomienda para escenarios que requieren una latencia baja.

Almacenamiento con redundancia geográfica (GRS)

Habilite el acceso de lectura a los datos en el caso de que la región no esté disponible.

Cuenta de almacenamiento

The screenshot shows the Azure portal interface for a storage account named 'ejemplobc1'. The left sidebar contains navigation options like 'Información general', 'Registro de actividad', and 'Explorador de almacenamiento'. The main area displays account details under 'Essentials', such as 'Grupo de recursos (mover) : g01' and 'Ubicación : eastus2'. A red box highlights the 'Cargar' button at the top. Another red box highlights the upload area with the text 'Arrastrar y colocar archivos aquí' and 'Buscar archivos'. A third red box highlights the 'Cargar' button at the bottom of the interface.

Abrimos el storage account recién creado, seleccionamos cargar y arrastramos o seleccionamos el archivo local que queremos tener en la nube

Contenedor

- Dentro de la cuenta de almacenamiento se debe crear un contenedor donde se cargará el archivo, o seleccionar un contenedor ya existente
- Una cuenta de almacenamiento puede tener n contenedores, y cada contenedor puede tener m archivos

Cargar blob



Nuevo contenedor

Nombre

Nivel de acceso anónimo ⓘ

Privada (sin acceso anónimo) ▾

ⓘ El acceso anónimo a este contenedor se está bloqueando porque el acceso anónimo está deshabilitado en esta cuenta de almacenamiento.

Aceptar Cancelar

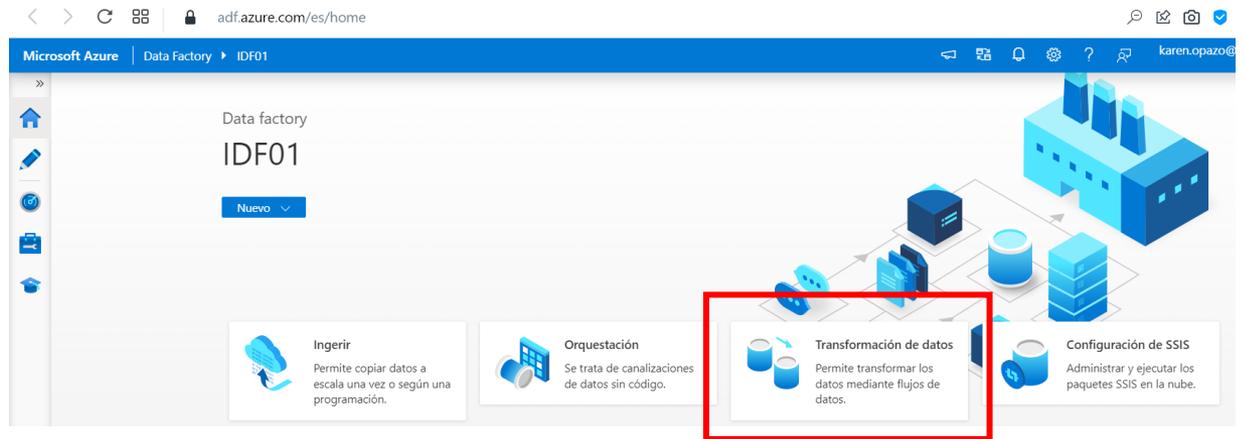
Vista desde DF

- Teniendo ambos recursos creados, podemos ingresar al data factory
- Una vez dentro del recurso Data Factory, se muestra un menú con su información (metadatos) y un botón de ingreso a la herramienta “Studio”, en la cual se realizarán los desarrollos

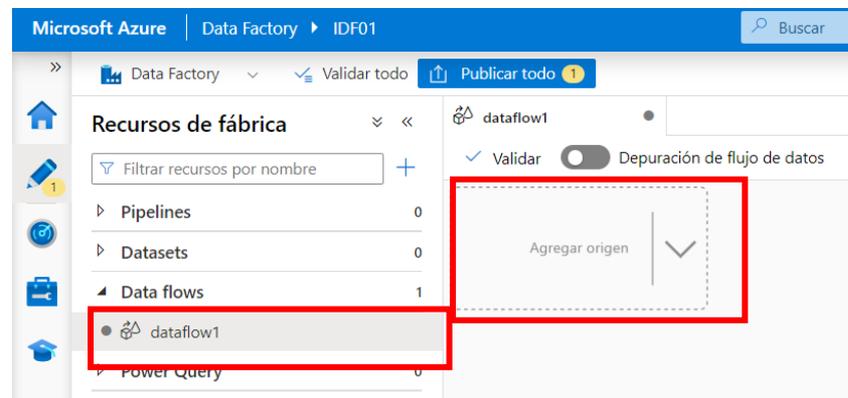
The screenshot shows the Microsoft Azure portal interface. At the top, there's a search bar and navigation icons. The main content area displays the 'Factorías de datos' (Data Factories) resource page for 'IDF01'. The left sidebar shows a list of data factories, with 'IDF01' selected. The main panel shows the 'Inicio rápido' (Quickstart) section, which is expanded to show various options. The 'Iniciar Studio' button is highlighted with a red box. The right sidebar shows the 'Información general' (General information) section, including details like 'Grupo de recursos' (Resource group), 'Estado' (Status), 'Ubicación' (Location), and 'Suscripción' (Subscription).

Transformación de datos

Una vez dentro de Studio, se realizará una “Transformación de Datos”, por lo cual se debe seleccionar el botón con el mismo nombre.

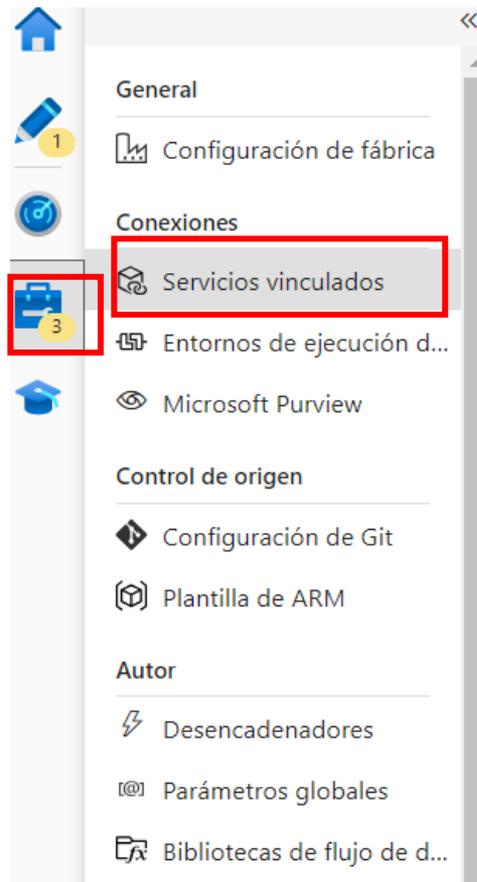


Se muestra la pantalla en donde se creará el flujo de datos de la transformación, el cual es llamado por defecto “dataflow1”. El primer paso es agregar un origen de datos con el botón “Agregar origen”, el cual debe ser configurado correctamente para poder agregar los siguientes elementos del flujo de datos.



Servicio vinculado

Para poder acceder a los archivos cargados, debemos vincular la cuenta de almacenamiento creada



Filtrar por nombre

Anotaciones : Cualquiera



No hay servicio vinculado para mostrar

Si esperaba ver resultados, pruebe a cambiar los filtros o cree una nueva servicios vinculados.

Crear servicio vinculado

Servicio vinculado

Nuevo servicio vinculado

Almacén de datos Compute

storage

Todo Azure Database File Generic protocol NoSQL Services and apps



Azure File Storage



Azure Data Lake Storage Gen2



Azure Data Lake Storage Gen2



Oracle Cloud Storage (S3 API)

Para este ejemplo, vinculamos un data lake storage gen2

Seleccionamos un nombre y un método de selección de la cuenta, en donde seleccionamos por suscripción de Azure y elegimos la cuenta de almacenamiento creada al inicio

Nuevo servicio vinculado

Azure Data Lake Storage Gen2 [Más información](#)

Nombre *

AzureDataLakeStorage1

Descripción

Conectar mediante Integration Runtime * ⓘ

AutoResolveIntegrationRuntime

Tipo de autenticación

Clave de cuenta

Método de selección de cuenta ⓘ

From Azure subscription Enter manually

Suscripción de Azure ⓘ

Seleccionar todo

Nombre de cuenta de almacenamiento *

Prueba de conexión ⓘ

Al servicio vinculado A la ruta de acceso de archivo

Servicio vinculado

Una vez vinculado el servicio, aparece un mensaje con un ticket verde y se muestra el servicio desde la pestaña de “administrar”

En esta pestaña también podemos ver todos los procesos relacionados a este servicio vinculado

Servicios vinculados

El servicio vinculado define la información de conexión a un almacén de datos o un proceso. [Más información](#) 

+ Nuevo

Filtrar por nombre

Anotaciones : **Cualquiera**

Mostrando elementos del 1 al 1 de un total de 1

Nombre ↑↓	Tipo ↑↓	Relacionado ↑↓	Anotaciones ↑↓
 AzureDataLakeStorage1	Azure Data Lake Storage Gen2	0	

Configuración del origen

En “Configuración de Origen” se debe asignar un nombre y usaremos la opción de “Conjunto de Datos”. Este conjunto puede estar previamente en Data Factory o podemos crear uno nuevo.

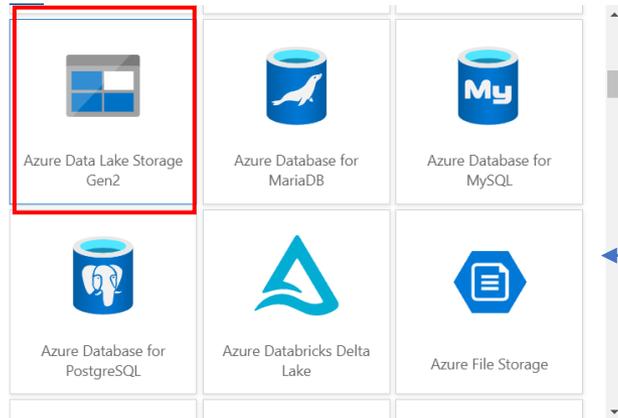
New dataset

In pipeline activities and data flows, reference a dataset to specify the location and structure of your data within a data store. [Learn more](#)

Select a data store

Search

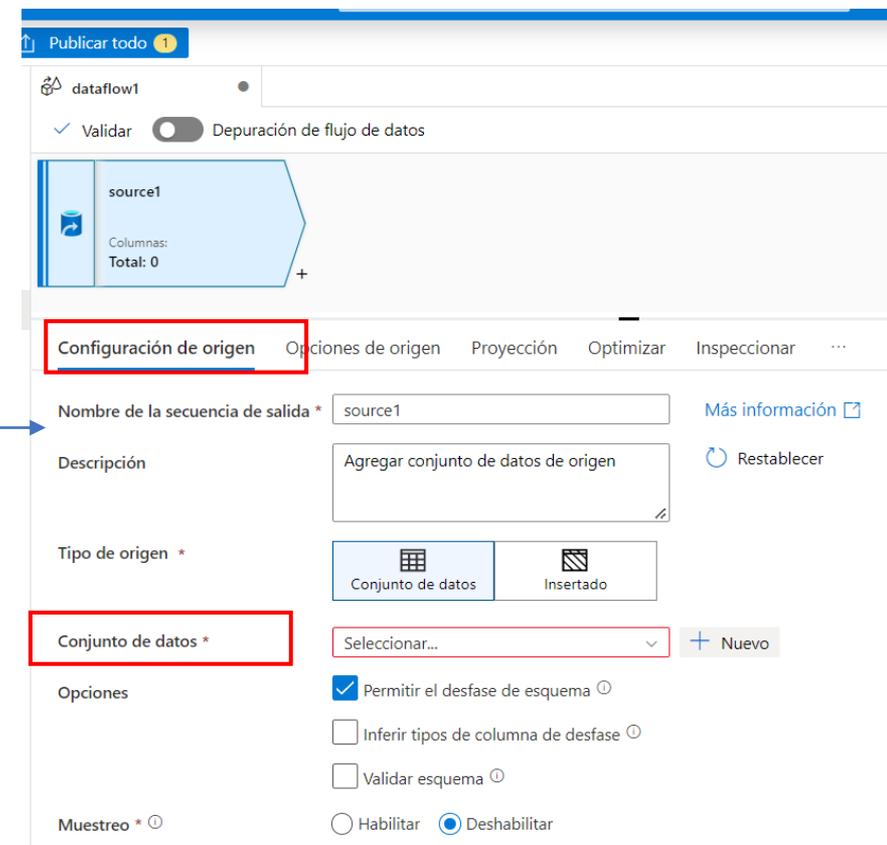
All Azure Database File Generic protocol NoSQL Services and apps



Continue

Cancel

CONTACTO@LITUUS.CL



Para este ejercicio se utilizará como conjunto de datos un archivo Excel llamado “Financial Sample” que está ubicado en una Storage de Azure: “Azure Data Lake Storage Gen2”. Presionar el botón “Continuar”

Nuevo conjunto de datos

1.

Indicamos el tipo de archivo del origen, en este caso Excel.

Seleccionar formato

Elija el tipo de formato de los datos.

Avro

DelimitedText

Excel

JSON

ORC

XML

Binario

Continuar

Atrás

Establecer propiedades

Nombre

Excel1

Servicio vinculado *

AzureBlobStorage1

Ruta de acceso del archivo

Contenedor / Directorio / Nombre de archivo

Desde la raíz

Desde la ruta de acceso especificada

Modo de hoja de cálculo

Nombre Index

Nombre de hoja *

Ninguno

Editar

Primera fila como encabezado

En “Servicio Vinculado” se selecciona la carpeta de almacenamiento y en “Ruta de acceso al archivo” se debe especificar con la opción “Desde ruta de acceso especificada”

Nuevo conjunto de datos

Seleccionamos
“Financial
Sample.xlsx” y se
presiona el botón
Aceptar

Examinar

Seleccione un archivo o una carpeta.

Carpeta raíz > newcontainer

- Financial Sample.csv
- Financial Sample.xlsx**

Mostrando elementos del 1 al 2 de un total de 2

Finalmente se elige la hoja a procesar del archivo Excel y se establece la primera fila como encabezado.

Establecer propiedades

Nombre
Excel1

Servicio vinculado *
AzureBlobStorage1

Ruta de acceso del archivo
newcontainer / Directorio / Financial Sample.xlsx

Modo de hoja de cálculo
 Nombre Index

Nombre de hoja *
Sheet1

Editar

Primera fila como encabezado

Importar esquema
 Desde una conexión o un almacén Desde un archivo de ejemplo
 Ninguno

> Avanzadas

Prueba de conexión

Si la configuración anterior se hizo de forma adecuada, la “Prueba de conexión” debe mostrar el mensaje “Conexión Correcta”. Para observar los datos del origen, se debe tener activada la opción “Depuración de flujo de datos”

dataflow1

✓ Validar Depuración de flujo de datos Configuración de depuración

source1

Columnas:
Total: 16

Configuración de origen Opciones de origen Proyección Optimizar Inspeccionar ...

Nombre de la secuencia de salida * source1 [Más información](#)

Descripción Importar datos de Excel1 [Restablecer](#)

Tipo de origen * Conjunto de datos Insertado

Conjunto de datos * Excel1 [Prueba de conexión](#) [Abierto](#) [Nuevo](#)

Opciones Permitir el desfase de esquema Inferir tipos de columna de desfase Validar esquema

Muestreo * Habilitar Deshabilitar

Conexión correcta

Vista previa de datos

En la pestaña “Vista previa de los Datos” se puede observar que el Origen posee 16 columnas, las cuales se pueden revisar a detalle.

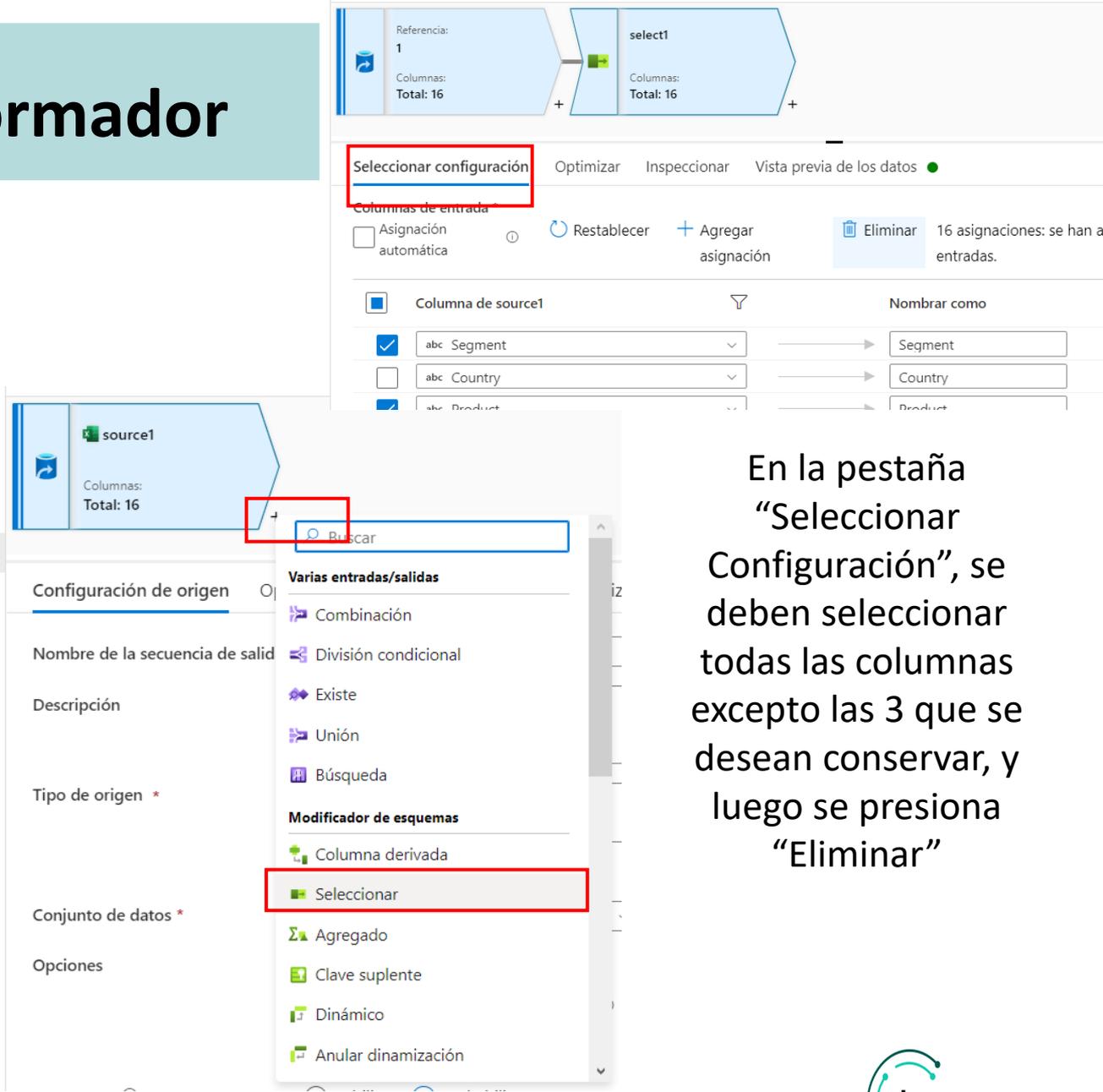
Segment	Country	Product	Discou...	Units S...	Manuf...	Sal
Govern...	Canada	Carretera	None	1618.5	3	20
Govern...	Germany	Carretera	None	1321.0	3	20
Midmar...	France	Carretera	None	2178.0	3	15
Midmar...	Germany	Carretera	None	888.0	3	15
Midmar...	Mexico	Carretera	None	2470.0	3	15
Govern...	Germany	Carretera	None	1513.0	3	350
Midmar...	Germany	Montana	None	921.0	5	15
Channe...	Canada	Montana	None	2518.0	5	12

En la pestaña “Proyección” se pueden hacer modificaciones generales a las columnas, por ejemplo, aplicar filtros o definir el tipo de datos.

Nombre de columna	Tipo	Formato
Segment	abc string	Especificar formato
Country	abc string	Especificar formato
Product	abc string	Especificar formato
Discount Band	abc string	Especificar formato
Units Sold	1,2 double	Especificar formato
Manufacturing Price	12s short	Especificar formato
Sale Price	12s short	Especificar formato
Gross Sales	1,2 double	Especificar formato
Discounts	1,2 double	Especificar formato
Sales	1,2 double	Especificar formato

Elemento transformador

Con el origen configurado, añadiremos un componente que realice una transformación. Para esto se debe presionar el signo más “+” que está en la esquina inferior derecha del origen, con esto se va a expandir un listado de elementos disponibles. Se elige “Modificador de Esquema” y “Seleccionar”, ya que se pretende seleccionar y renombrar 3 columnas de las 16 del archivo de origen, para crear una nueva tabla.



En la pestaña “Seleccionar Configuración”, se deben seleccionar todas las columnas excepto las 3 que se desean conservar, y luego se presiona “Eliminar”

Renombrar columnas

Una vez eliminadas las columnas que no se van a usar, se renombrarán las que, sí se van a conservar, desde “Columna de source1” hasta la columna “Nombrar como”

En la opción “Vista previa de los datos”, se pueden observar los datos antes de que agreguemos un segundo elemento transformador, el cual se va a encargar de **agrupar** los datos por País

The screenshot shows the configuration of a 'select' transformation in a data tool. The 'source1' is 'Importar datos de Excel1'. The 'Opciones' section has 'Omitir columnas de entrada duplicadas' and 'Omitir columnas de salida duplicadas' checked. The 'Columnas de entrada' section shows three columns: 'abc Country', '1.2 Sales', and '1.2 Profit'. The 'Nombrar como' section shows three output columns: 'País', 'Ventas', and 'Rentabilidad'. The 'Columna de source1' and the 'Nombrar como' section are highlighted with red boxes.

Seleccionar configuración Optimizar Inspeccionar **Vista previa de los datos**

Número de filas: + INSERTAR 100 * ACTUALIZACIÓN 0 x ELIMINAR 0 + ACTUALIZAR/INSERTAR (UPSERT) 0

Actualizar | Conversión | Modificar | Asignar desfasados | Estadísticas | Quitar

	País	Ventas	Rentabilidad
+	Canada	32370.0	16185.0
+	Germany	26420.0	13210.0
+	France	32670.0	10890.0
+	Germany	13320.0	4440.0
+	Mexico	37050.0	12350.0
+	Germany	529550.0	136170.0
+	Germany	13815.0	4605.0
+	Canada	30216.0	22662.0
+	France	37980.0	18990.0

Agrupar datos

Para agregar un segundo modificador de esquemas, en este caso un “Agregado”, se debe seleccionar del menú que se despliega al presionar el signo más (+)

The screenshot shows the data flow tool interface. At the top, there are tabs for 'Validar', 'Depuración de flujo de datos', and 'Configuración de depuración'. Below this, a flow diagram shows a sequence of steps: 'source1' (Importar datos de Excel1), 'select1' (Se está cambiando el nombre de source1 a select1 con las columnas "País, Ventas, Rentabilidad"), and 'aggregate1' (Columns: Total: 3). A red box highlights the plus sign (+) next to the 'select1' step. Below the flow diagram, a 'Modificador de esquemas' menu is open, showing options: 'Columna derivada', 'Seleccionar', 'Agregado' (highlighted with a red box), 'Clave suplente', 'Dinámico', and 'Anular dinamicación'. At the bottom, there is a 'Nombre de la secuencia de salida' field with the value 'select1'.

The screenshot shows the 'Configuración de agregado' settings panel. At the top, there are tabs for 'Configuración de agregado', 'Optimizar', 'Inspeccionar', and 'Vista previa de los datos'. The 'Configuración de agregado' tab is active. The settings are as follows: 'Nombre de la secuencia de salida' is 'aggregate1'; 'Descripción' is 'Agregación de datos por "País" que genera las columnas "Ventas por País, Rentabilidad por País"'; 'Secuencia entrante' is 'select1'. Below these settings, there are two buttons: 'Agrupar por' (highlighted with a red box) and 'Agregados'. At the bottom, there is a 'Columnas' field with the value 'abc País' (highlighted with a red box) and a 'Nombrar como' field with the value 'País'.

Luego en la pestaña “Agrupar por” se elige la columna por la cual se desean agrupar los datos, en este caso “País”

Configurar columnas

En la pestaña “Agregados” indicaremos las dos columnas que presentarán serán agrupados por País. Se deben especificar el nombre de las columnas y la expresión lógica de los datos agrupados, en este caso, será la suma de Ventas y la suma de Rentabilidad.

Configuración de agregado Optimizar Inspeccionar **Vista previa de los datos** ●

Número de filas + INSERTAR 5 * ACTUALIZACIÓN 0 x ELIMINAR 0 * ACTUALIZAI

Actualizar | 12l Conversión | Modificar | Asignar desfasados | Estadísticas X

↑↓	Pais	abc ↑↓	Ventas	12l ↑↓	Rentabilidad	12l ↑↓
+	Canada		24887613		3529194	
+	Germany		23505300		3680357	
+	France		24354130		3780987	
+	Mexico		20949312		2907490	
+	United States of Am...		25029787		2995506	

Agrupar por

Agregados

Agrupado por: Pais

+ Agregar Clonar Eliminar **Abrir generador de expresiones**

<input type="checkbox"/>	Columna		Expresión		
<input type="checkbox"/>	Ventas	□	sum(Ventas)	12l	+ [icon]
<input type="checkbox"/>	Rentabilidad	□	sum(Rentabilidad)	12l	+ [icon]

Una vez revisados los datos en la pestaña “Vista previa de los datos”, se configura el último elemento de este flujo de datos, que es el Destino o Receptor

Configurar sink

La última parte del flujo de datos es el elemento receptor “Destino”, para añadirlo se presiona el signo más “+” en la esquina inferior derecha del elemento “aggregate1”. Debe seleccionar “Receptor”

Validar Depuración de flujo de datos Configuración de depuración

source1 Importar datos de Excel1

select1 Se está cambiando el nombre de source1 a select1 con las columnas "Pais, Ventas, Rentabilidad".

aggregate1 Agregación de datos por "Pais" que genera las columnas "Ventas, Rentabilidad"

sin Columna Total

Receptor Configuración Errores Asignación Optimizar Inspeccionar Vista previa de los datos

Nombre de la secuencia de salida * sink1 Más información

Descripción Exportar datos a DelimitedText1 Restablecer

Secuencia entrante * aggregate1

Tipo de receptor *

Conjunto de datos * DelimitedText1 Prueba de conexión Abierto

Omitir recuento de líneas

Opciones Permitir el desfase de esquema Validar esquema

aggregate1

Columnas: Total: 3

Conversion

previa de los datos

ELIMINAR 0 ACT

Asignar desfasados Estadísticas

12l	Rentabilidad	12r
	3529194	
	3680357	
	3780987	
	2907490	
	2995506	

Formateadores

- Acoplar
- Analizar
- Stringify

Modificador de filas

- Filtrar
- Ordenar
- Modificar fila
- Declarar

Flowlets

- Flowlet

Destino

- Receptor

Se debe asignar un nombre y un nuevo conjunto de datos, de forma similar a la descrita en la diapositiva “Configurar Origen”, con la única diferencia de que el tipo de archivo en esta oportunidad será “csv”

Validar y publicar

1. Si la configuración del elemento receptor “sink1” fue bien hecha, la prueba de conexión debe mostrar el mensaje “Conexión correcta”.

2. Con los elementos del flujo de datos bien configurados, procederemos a validar el Dataflow con el botón “Validar”. Si no hay errores, aparecerá un ticket al lado del botón.

3. El último paso es publicar los elementos creados, con el botón “Publicar todo”.

The screenshot displays the Databricks Dataflow Builder interface. At the top, a blue button labeled "Publicar todo" with a yellow ticket icon is highlighted with a red box. Below it, a "Validar" button with a green checkmark and a "Depuración de flujo de datos" toggle are also highlighted with a red box. The main area shows a data pipeline with four stages: "source1" (Importar datos de Excel1), "select1" (Se está cambiando el nombre de source1 a select1 con las columnas "Pais, Ventas, Rentabilidad"), "aggregate1" (Agregación de datos por "Pais" que genera las columnas "Ventas, Rentabilidad"), and "sink1" (Columns: Total: 3). Below the pipeline, the configuration for the "sink1" stage is shown. The "Nombre de la secuencia de salida" is "sink1", the "Descripción" is "Exportar datos a DelimitedText1", and the "Secuencia entrante" is "aggregate1". The "Tipo de receptor" is set to "Conjunto de datos". The "Conjunto de datos" is "DelimitedText1", and the "Prueba de conexión" status is "Conexión correcta" (highlighted with a red box). Other options include "Permitir el desfase de esquema" (checked) and "Validar esquema" (unchecked).

Validar y publicar

Una vez publicado el flujo de datos, se puede descargar el archivo CSV creado directamente desde la **Vista previa de los datos**

Receptor Configuración Errores Asignación Optimizar Inspeccionar **Vista previa de los datos** ●

Número de filas **+ INSERTAR** N/A *** ACTUALIZACIÓN** N/A **× ELIMINAR** N/A **+ ACTUALIZAR/INSERTAR (UP)**

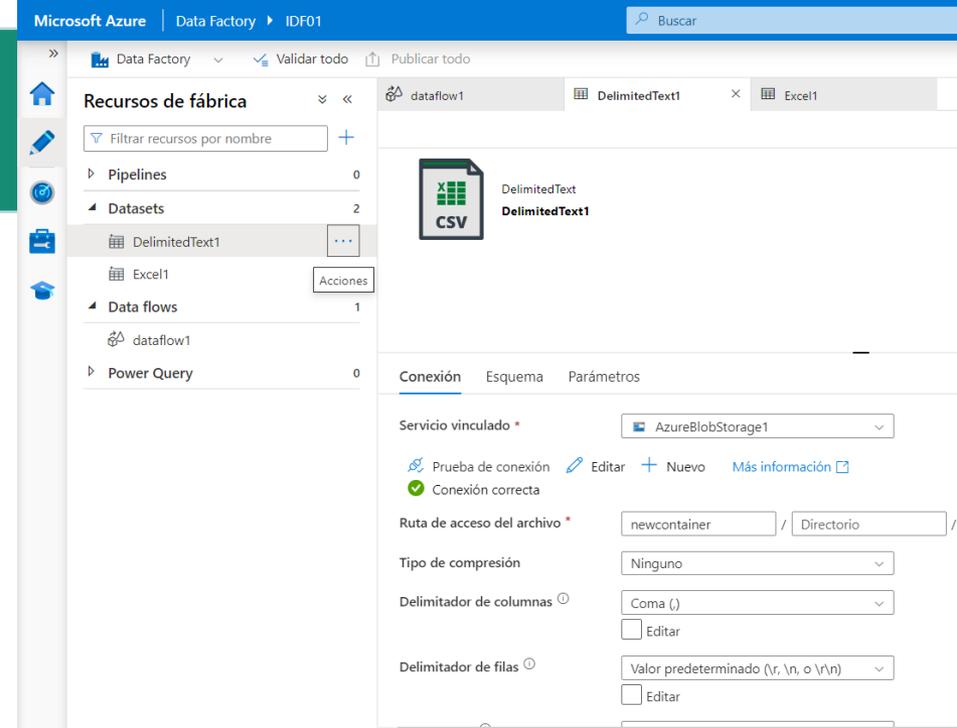
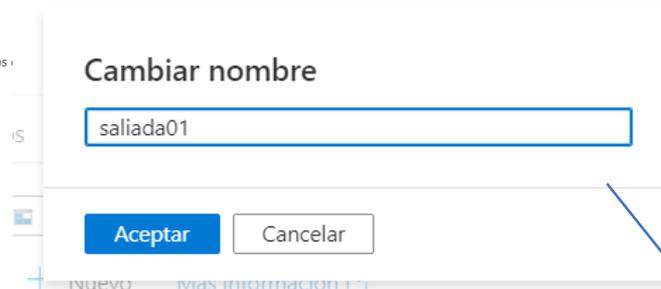
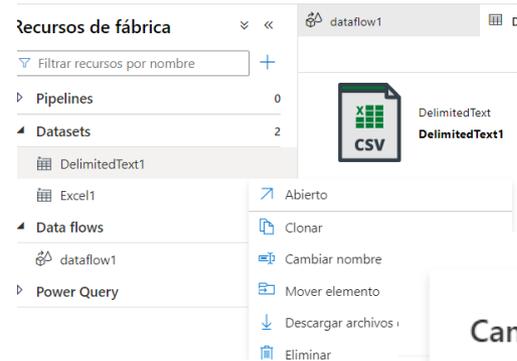
Exportar CSV: solo se exportarán las filas de la página actual

↓ Exportar a CSV | ▾

Pais	abc ↑↓	Ventas	abc ↑↓	Rentabilidad	abc ↑↓
Canada		24887613		3529194	
Germany		23505300		3680357	
France		24354130		3780987	
Mexico		20949312		2907490	
United States of Am...		25029787		2995506	

Revisar archivos de salida

Al revisar los recursos disponibles de Data Factory, se puede observar el nuevo Dataset de nombre "DelimitedText1", el cual se puede renombrar, clonar o mover sin problemas.



Pipeline

Como el dataflow NO es un orquestador, no corre al depurarlo ni con un trigger, por lo que hay que insertarlo dentro de un pipeline o canalización

Recursos de fábrica

Filtrar recursos por nombre

Canalizaciones

● pipeline1

Existen diversas actividades que se pueden ejecutar dentro de un pipeline, como copiar datos, anexar variables, hacer un lookup, entre muchas otras

Actividades

Buscar actividades

- > Mover y transformar
- > Synapse
- > Azure Data Explorer
- > Función de Azure
- > Servicio de Batch
- > Databricks
- > Data Lake Analytics
- > General
- > HDInsight
- > Iteración y condicionales
- > Machine Learning
- > Power Query

Pipeline

Para este ejemplo, ejecutaremos el dataflow en un pipeline

The screenshot displays the Azure Data Factory user interface. On the left, the 'Recursos de fábrica' (Factory Resources) pane shows a tree view with categories like 'Canalizaciones' (1), 'Captura de datos modificados (...)' (0), 'Conjuntos de datos' (3), and 'Flujos de datos' (1). The 'pipeline1' resource is selected. The main 'Actividades' (Activities) pane is open, showing a search bar and a list of activity types under 'Mover y transformar' (Move and transform). The 'Flujo de datos' (Data flow) activity is highlighted with a red box. To the right, the pipeline canvas shows the 'Flujo de datos' activity as a diamond-shaped node labeled 'Data flow1', also highlighted with a red box. The top of the interface includes tabs for 'dataflow1', 'DelimitedText1', and 'pipeline1', along with control buttons for 'Validar', 'Cancelar opciones', 'Desencadenador (1)', and 'Depuración' (Debugging).

Pipeline

Seleccionamos el dataflow (ya creado) que vamos a utilizar en el pipeline

General **Configuración** Parámetros ¹ Propiedades de usuario

Flujo de datos *  Abierto [+ Nuevo](#)

Ejecutar en (Azure IR) * ¹

Tamaño de proceso * ¹

> Avanzadas

Nivel de registro * ¹ Verbose Basic None

> Propiedades del receptor

> Ensayo ¹

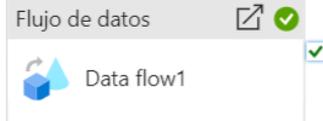
Validar y publicar

Similar al dataflow, se deben validar los cambios antes de poder publicarlos, teniendo activada la opción “debug” o depuración de flujo de datos

The screenshot displays the Azure Data Studio interface. At the top left, a blue button labeled 'Publicar todo' with a yellow notification badge containing the number '7' is highlighted with a red box. Below this, the 'Actividades' sidebar is visible, containing a search bar and a list of activities: 'Copiar datos' and 'Flujo de datos'. The main workspace shows a 'Flujo de datos' (Data flow) activity named 'Data flow1'. In the top right of the workspace, a toolbar contains several controls: a 'Validar' button with a checkmark icon, a 'Cancelar opciones' button with a circle-slash icon, a 'Desencadenador (1)' button with a lightning bolt icon, and a 'Depuración de flujo de datos' toggle switch which is currently turned on and accompanied by a green checkmark. The 'Validar' button is also highlighted with a red box.

Depurar / debug

✓ Validar ▶ Depurar ▼ ⚡ Agregar desencadenador Depuración de flujo de datos ✓



Parámetros Variables Configuración **Salida**

Id. de ejecución de canalización: db7c8664-9d77-4ee0-b048-82f443cffc3a **Estado de la canalización** ✓ Correcto

All status ▼

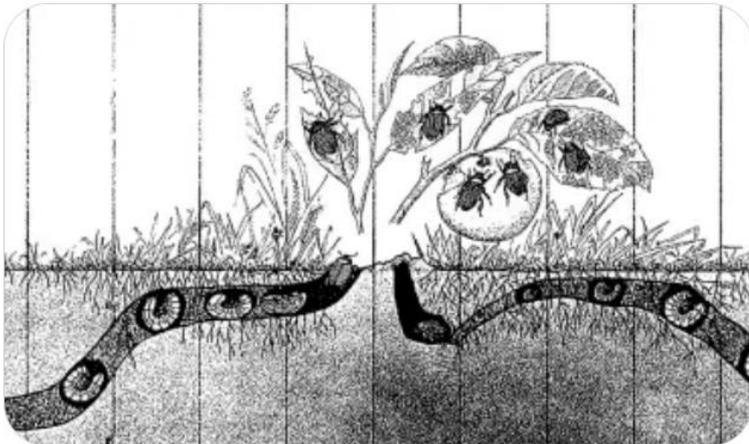
[Supervisión en métricas de Azure](#) [Exportar a CSV](#) | ▼

Mostrando elementos del 1 al 1 de un total de 1

Nombre de actividad	Estado de actividad	Tipo de actividad	Inicio de la ejecución	Duración	Entorno de ejecución
Data flow1	✓ Correcto	Flujo de datos	3/7/2024, 9:27:00 PM	16m 3s	debugpool-8Cores-Gei

Corremos el pipeline seleccionando depurar. Cuando se realice correctamente, aparecerá un log al final de la página con detalles de la actividad.

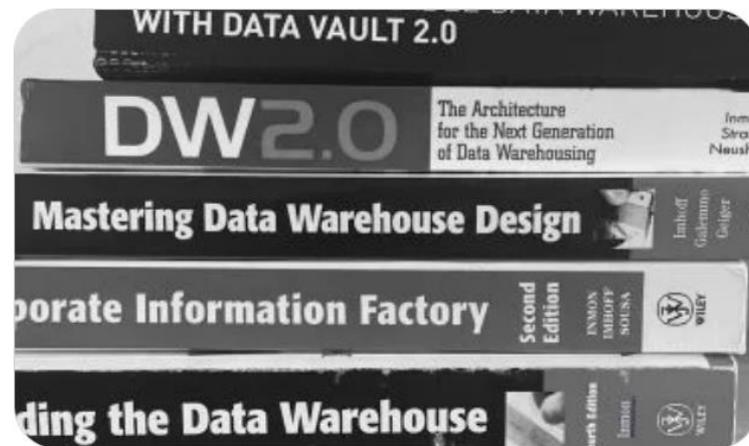
Más contenido



¿Qué es la visualización de datos cuantitativos y por qué es importante?

11/03/2021 Lenin González

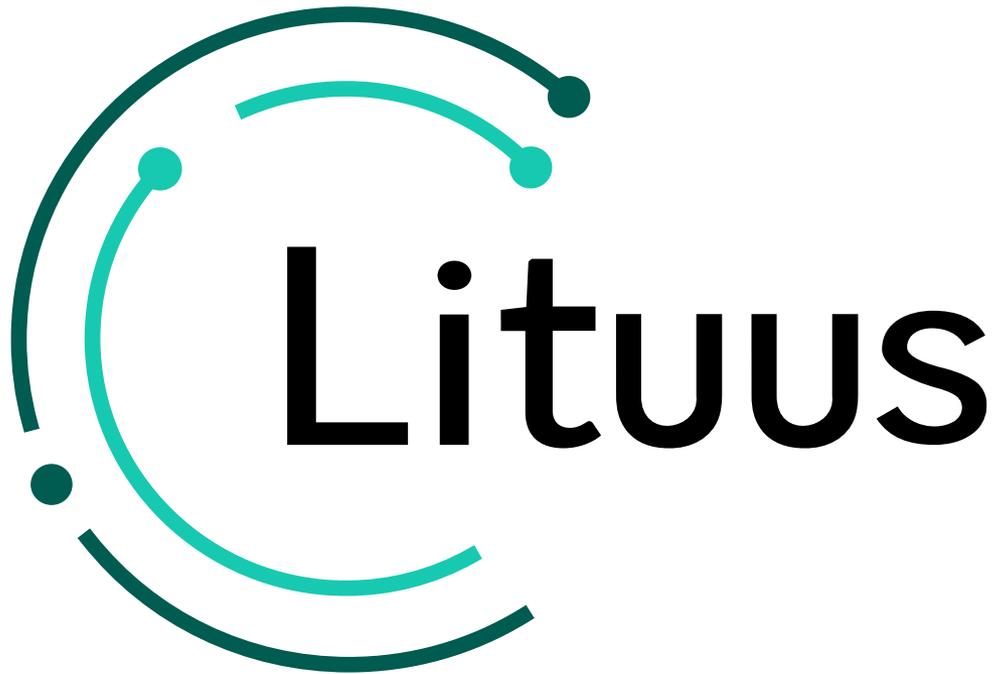
Intentaré convencerte de por qué debemos aprender visualización de datos cuantitativos. Verás ejemplos, un poco de historia y cuatro recomendaciones prácticas que puedes aplicar desde hoy mismo.



¿Qué es un Data Warehouse y Para Qué Sirve?

18/01/2021 Lenin González

En la medida que las empresas requieren ser más competitivas suelen incorporar Software que les permita gestionar sus procesos de negocios de forma tal que se vuelvan predecibles y optimizables. Cada proceso de negocio tiene su complejidad y vida propia, por lo que sin quererlo ni beberlo, las organizaciones comienzan adquirir más Software o Módulos encargados de administrar sus procesos de forma atómica, durable y concurrente. Suele suceder que dichos sistemas/módulos no usan las mismas tecnologías, puede que no sean del mismo proveedor o sus arquitecturas sean incompatibles. Es Aquí cuando un Data Warehouse puede aportar credibilidad en la toma de decisiones al integrar en una sola fuente de la verdad nuestros dato



Presentado por:

- Lenin González S.
- Director General, CEO
- +569 3072 9405
- lgonzalez@lituus.cl

