

# GUÍA: TALLER MODERN DATA WAREHOUSE

Axayacatl Valenzuela Faddul

MICROSOFT

## INDICE

Ingresando al portal de administración de Azure .....	2
Creando el grupo de recursos .....	2
Azure Data Lake Storage .....	3
Crear contenedor.....	6
Subir archivos .....	7
Generar SAS .....	7
Databricks .....	9
Crear Workspace .....	9
Crear cluster .....	10
Importar el código .....	11
Synapse.....	12
Crear el workspace de Synapse.....	12
Crear SQL Pool .....	13
Importar el script .....	13
Azure Data Factory.....	15
1.    Creando Data Factory .....	15
2.    Creando el Data Flow .....	16
Creación del source.....	16
Creación del destino (sink).....	18
3.    Creando el Pipeline .....	19
Completando el Data Flow .....	20

## INGRESANDO AL PORTAL DE ADMINISTRACIÓN DE AZURE

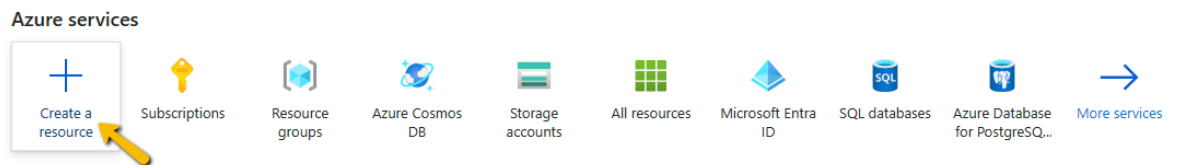
El primer paso para nuestro taller es ingresar al portal de administración de azure, donde crearemos los componentes de nuestra arquitectura.

1. Ingresar a la página <https://portal.azure.com>
2. Ingresar con las credenciales provistas

## CREANDO EL GRUPO DE RECURSOS

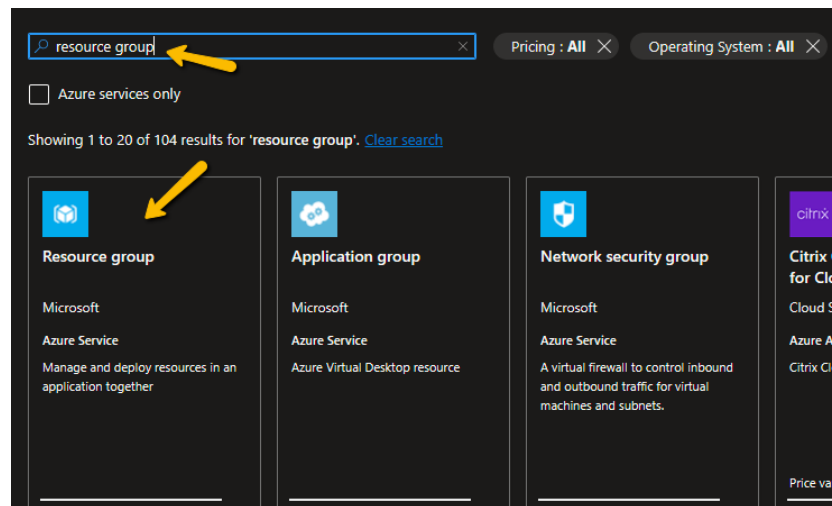
Una vez dentro del portal de azure, procederemos a crear nuestro grupo de recursos

1. Dar click en el botón “Crear Recurso”



### Resources

2. En el recuadro de búsqueda teclear “Resource Group” y seleccionar la opción “Resource Group” de Microsoft, Azure Service



3. En la pantalla de “Básicos”
  - a. El nombre del recurso debe ser “tusiniciales\_mdwh\_workshop”
  - b. Region: East US

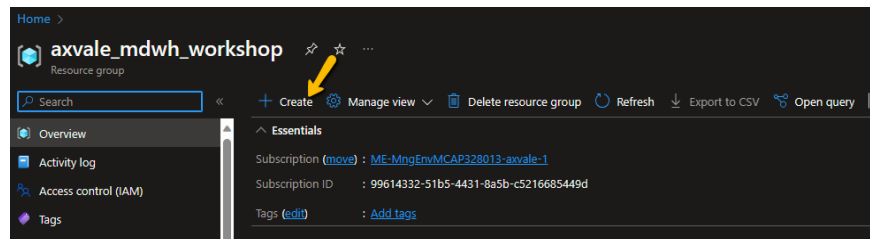
The screenshot shows the 'Basics' tab of the Azure portal. At the top, there are tabs for 'Basics', 'Tags', and 'Review + create'. Below the tabs is a description of a 'Resource group'. Under 'Project details', the 'Subscription' is set to 'ME-MngEnvMCAP328013-axvale-1' and the 'Resource group' is 'axvale\_mdwh\_workshop', which is highlighted with a blue border and a green checkmark. Under 'Resource details', the 'Region' is set to '(US) East US'.

4. Clic en el botón “Review + create”
5. Finalmente, clic en “Create”

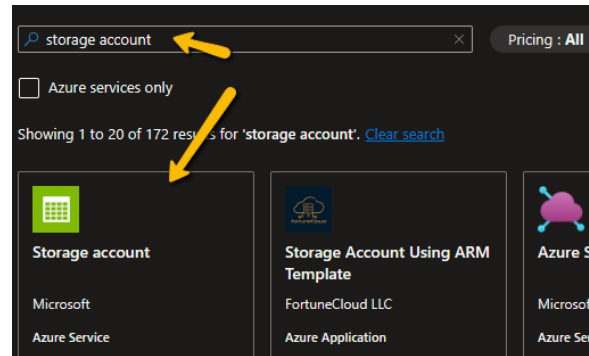
## AZURE DATA LAKE STORAGE

Con el grupo de recursos creado, es tiempo de crear nuestro almacenamiento.

1. Estando en el grupo de recursos, clic en el botón “Create”



2. En el recuadro de búsqueda, teclear “storage account”. Seleccionar la opción Storage Account, Microsoft, Azure Service y clic en “Create”



3. En la página de creación seleccionar las siguientes opciones
  - a. Resource Group: El grupo de recursos que se creó para el workshop
  - b. Storage Account name: tusiniciales**mdwhadls**
    - i. Ejemplo: axvale**mdwhadls**
  - c. Region: EastUS
  - d. Performance: Standard
  - e. Redundancy: Locally-redundant storage (LRS)

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription \*

Resource group \*   
[Create new](#)

**Instance details**

Storage account name ⓘ \*

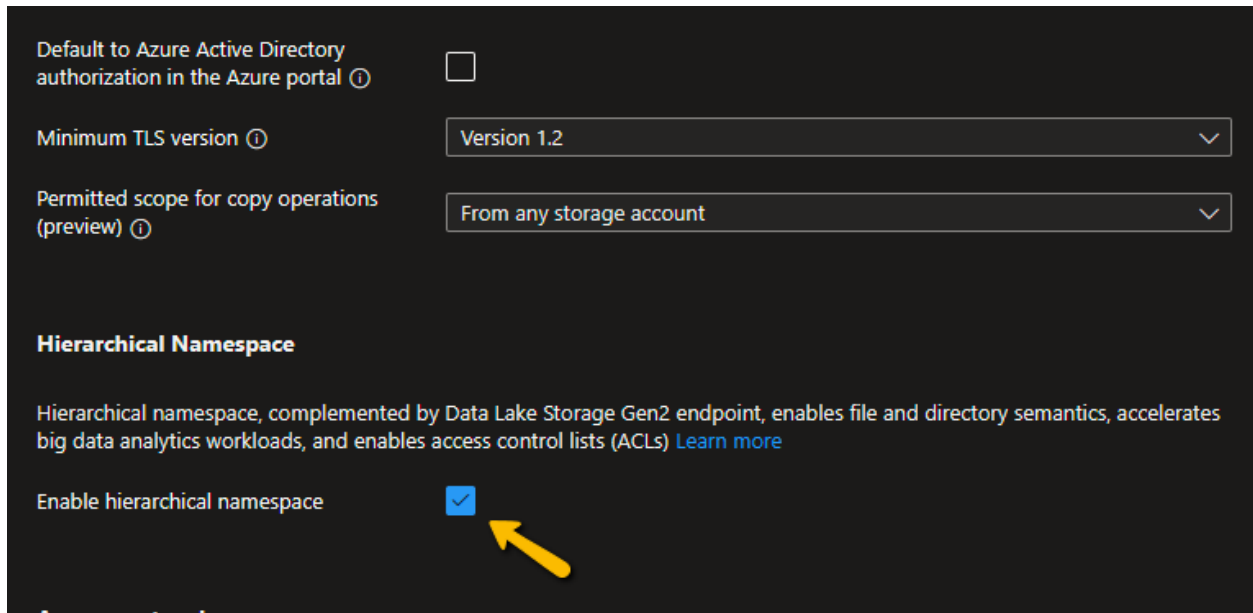
Region ⓘ \*   
[Deploy to an edge zone](#)

Performance ⓘ \*  
☒ **Standard:** Recommended for most scenarios (general-purpose v2 account)  
☐ **Premium:** Recommended for scenarios that require low latency.

Redundancy ⓘ \*

4. Clic en el boton “Next: Advanced>”

5. Estando en la pestaña “Advanced” habilitar la casilla “Hierarchical NameSpace”




Default to Azure Active Directory authorization in the Azure portal ⓘ ☐

Minimum TLS version ⓘ Version 1.2 ▼

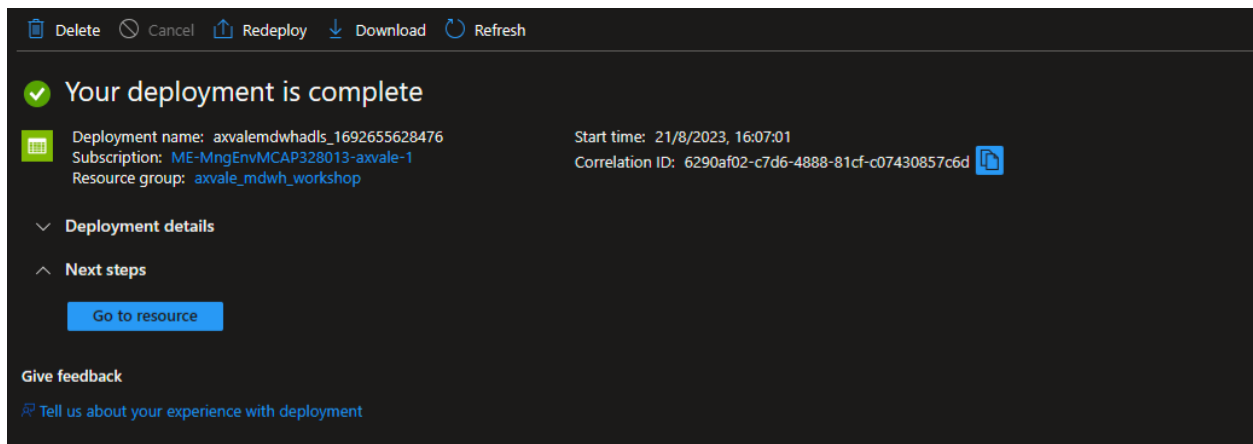
Permitted scope for copy operations (preview) ⓘ From any storage account ▼

### Hierarchical Namespace

Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) [Learn more](#)

Enable hierarchical namespace ☒ 


6. Clic en el botón “Review + Create”
7. Clic en el botón “Create”
8. Esperar a que el proceso de creación termine y dar clic en el botón “Go to Resource”



Delete Cancel Redeploy Download Refresh

✓ Your deployment is complete

Deployment name: axvalemmdwhadls\_1692655628476  
Subscription: ME-MngEnvMCAP328013-axvale-1  
Resource group: axvale\_mdwh\_workshop

Start time: 21/8/2023, 16:07:01  
Correlation ID: 6290af02-c7d6-4888-81cf-c07430857c6d 

Deployment details

Next steps

[Go to resource](#)

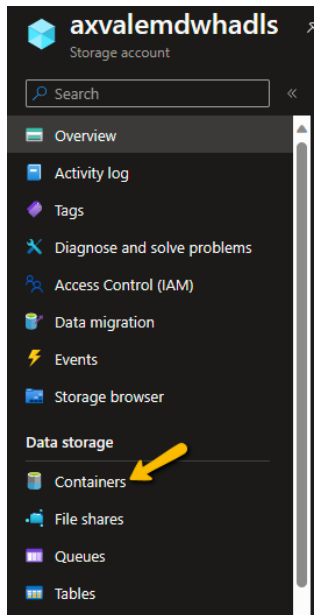
Give feedback

[Tell us about your experience with deployment](#)

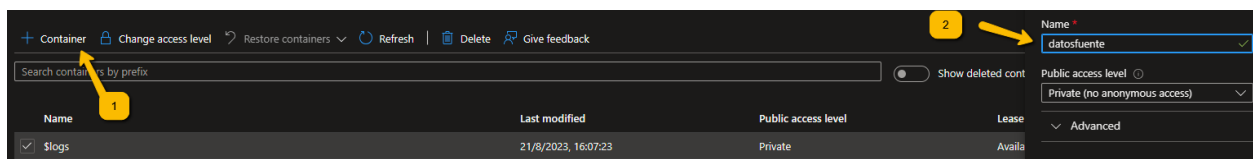
## CREAR CONTENEDOR

Estando en la página de administración de la cuenta de almacenamiento

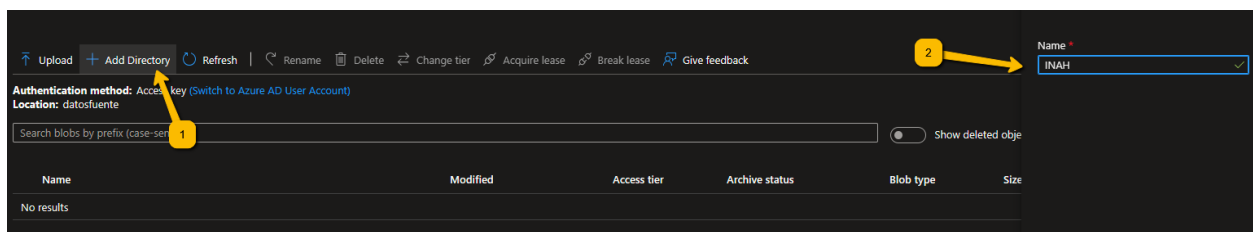
1. Clic en el botón “Containers”, ubicado en el menú de extrema izquierda



2. En el menú del centro
  - a. Clic en el botón “+ Container”
  - b. Nombre: datosfuente
  - c. Public Access level: Private
  - d. Clic en el botón “Create”



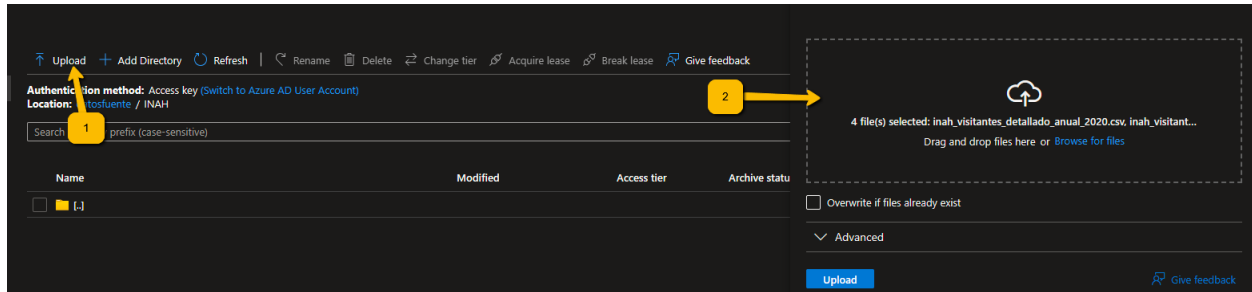
3. Una vez creado el contenedor
  - a. Seleccionar el contenedor
  - b. Clic en el botón “+Add Directory”



## SUBIR ARCHIVOS

En este paso vamos a subir los archivos de datos al storage Account. Lo primero que tienes que hacer es ubicar la carpeta donde clonaste el repositorio de github

1. Regresa al portal de Azure y en la carpeta que acabamos de crear, da clic en el botón Upload.
2. Da clic en “Browse files” y selecciona los cuatro archivos que están en la carpeta Data/INAH/inah\_visitantes en el repositorio clonado
3. Da clic en el botón Upload

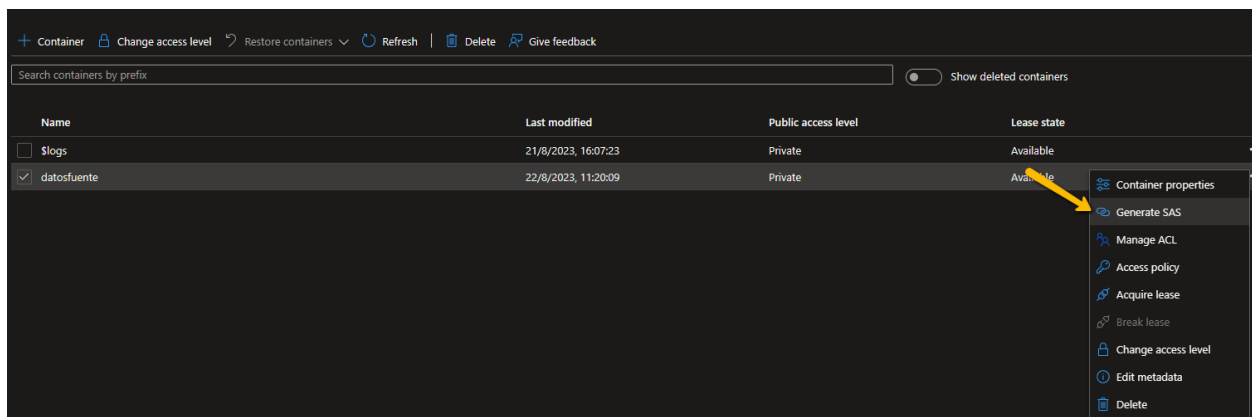


## GENERAR SAS

Ahora que has subido los archivos, es necesario otorgar acceso a la carpeta. Para eso utilizaremos una forma de autenticación llamada SAS (Shared Access Sign).

SAS es una forma rápida de otorgar acceso a un usuario, pero no es la más segura. La utilizaremos únicamente para este taller.

1. Ubica el contenedor datos fuente y da clic en el botón con los tres puntos (...) en el extremo derecho
2. Ahora da clic en el botón Generate SAS





3. En la pantalla de generación de SAS, seleccionas las siguientes opciones:
  - a. Signing method: Account key
  - b. Permissions:
    - i. Read
    - ii. Create
    - iii. Write
    - iv. List
    - v. Execute
  - c. Clic en “Generate SAS”. Guarda la llave generada, la ocuparemos después

## Generate SAS

A shared access signature (SAS) is a URI that grants restricted access to an Azure Storage container. Use it when you want to grant access to storage account resources for a specific time range without sharing your storage account key. [Learn more about creating an account SAS](#)

**Signing method**

☒ Account key ☐ User delegation key

**Signing key** ⓘ

Key 1

**Stored access policy**

None

**Permissions** \* ⓘ

2 selected

- ☒ Read
- ☐ Add
- ☐ Create
- ☐ Write
- ☐ Delete
- ☒ List
- ☐ Move
- ☐ Execute
- ☐ Ownership
- ☐ Permissions

11:24:47

Mexico City, Monterrey

19:24:47

Mexico City, Monterrey

or 168.1.5.65-168.1....

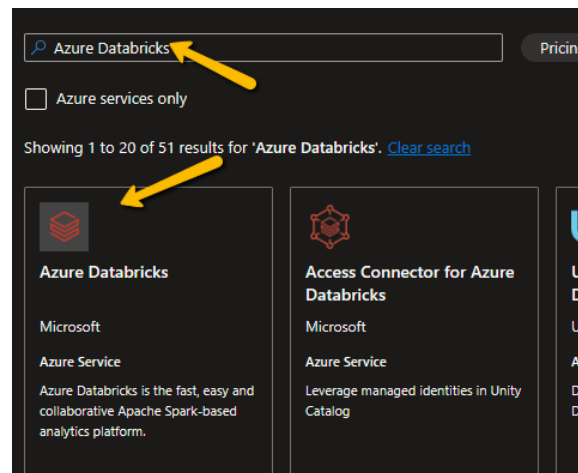
HTTPS and HTTP

**Generate SAS token and URL**

## DATABRICKS

### CREAR WORKSPACE

1. Regresa al grupo de recursos y da clic en el botón “+ Create”
2. En recuadro de búsqueda escribe: Databricks
3. De las opciones disponibles selecciona la primera



4. En la pantalla de creación selecciona las siguientes opciones:
  - a. Workspace name: <tusiniciales>\_databricks
  - b. Region: East US
  - c. Pricing Tier: Trial
  - d. Clic en “Review + Create”
  - e. Después de la validación, clic en “Create”


Home > axvale\_mdwh\_workshop > Marketplace > Azure Databricks >


### Create an Azure Databricks workspace

Basics Networking Encryption Tags Review + create


**Project Details**


Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.



Subscription \*  ME-MngEnvMCAP328013-axvale-1

Resource group \*  axvale\_mdwh\_workshop  
[Create new](#)

**Instance Details**

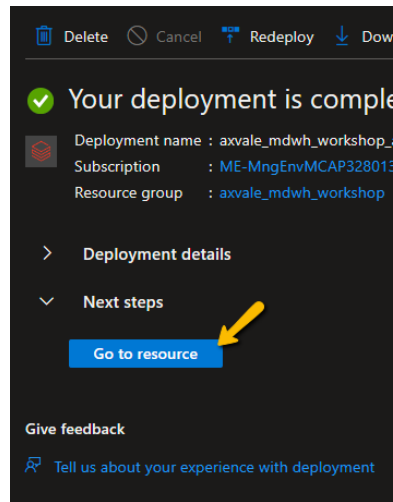
Workspace name \* axvale\_mdwh\_dbricks 

Region \* East US 

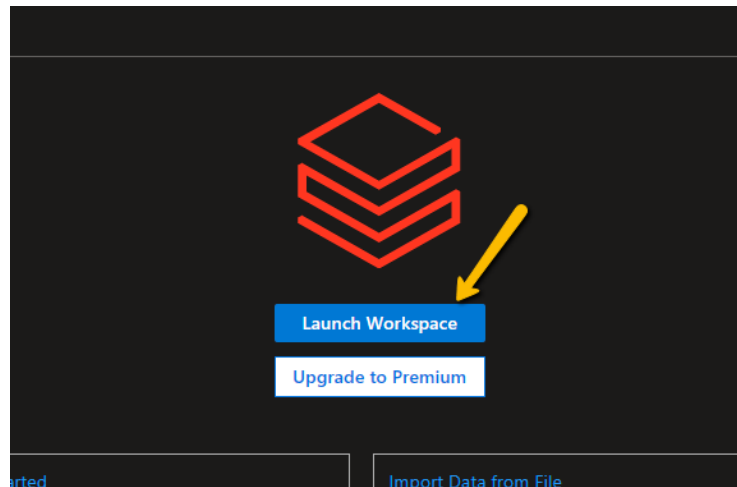
Pricing Tier \*  Trial (Premium - 14-Days Free DBUs) 

Managed Resource Group name Enter name for managed resource group

- f. Cuando el proceso de creación termine, da clic en “Go to Resource”

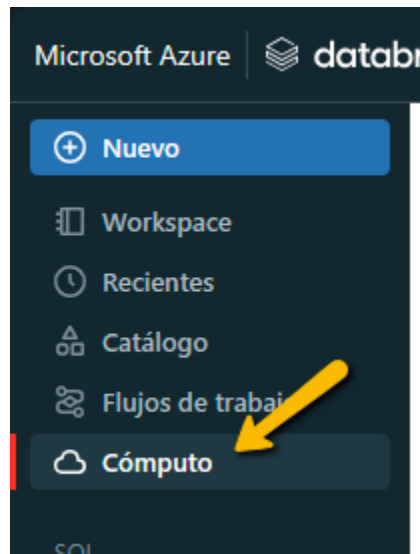


- g. En la pantalla del workspace da clic en “Launch Workspace”



## CREAR CLUSTER

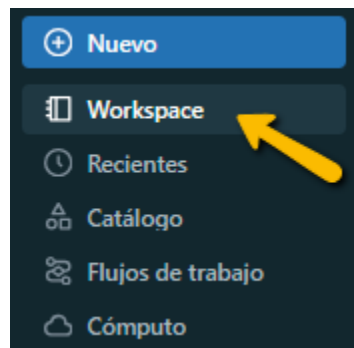
1. Estando en workspace de Databricks, clic en el botón “Cómputo” ubicado en el mené de extrema derecha



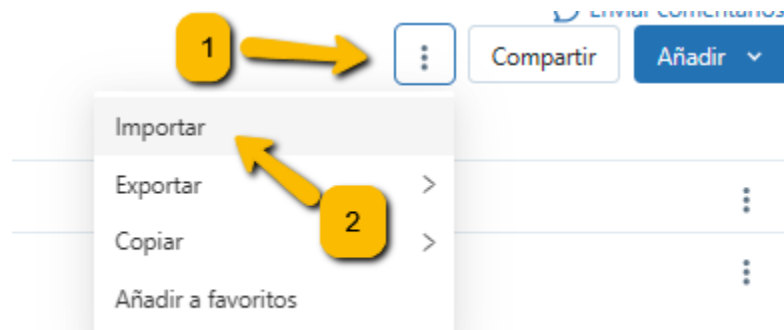
2. En la esquina derecha dar clic en “Crear Computo”
3. En la pantalla de creación de cluster seleccionar las siguientes opciones
  - a. Directriz: Sin restricciones
  - b. Multi-nodo
  - c. Runtime: 13.3 LTS
  - d. Deshabilitar aceleración Photon
  - e. Tipo de worker: Standard\_D4a\_V4
  - f. Workers:2
  - g. Tipo de driver: El mismo que el worker
  - h. Deshabilitar auto expansión
  - i. Habilitar terminar después de 120 minutos de actividad
4. Clic en crear/Importar el código

## IMPORTAR EL CÓDIGO

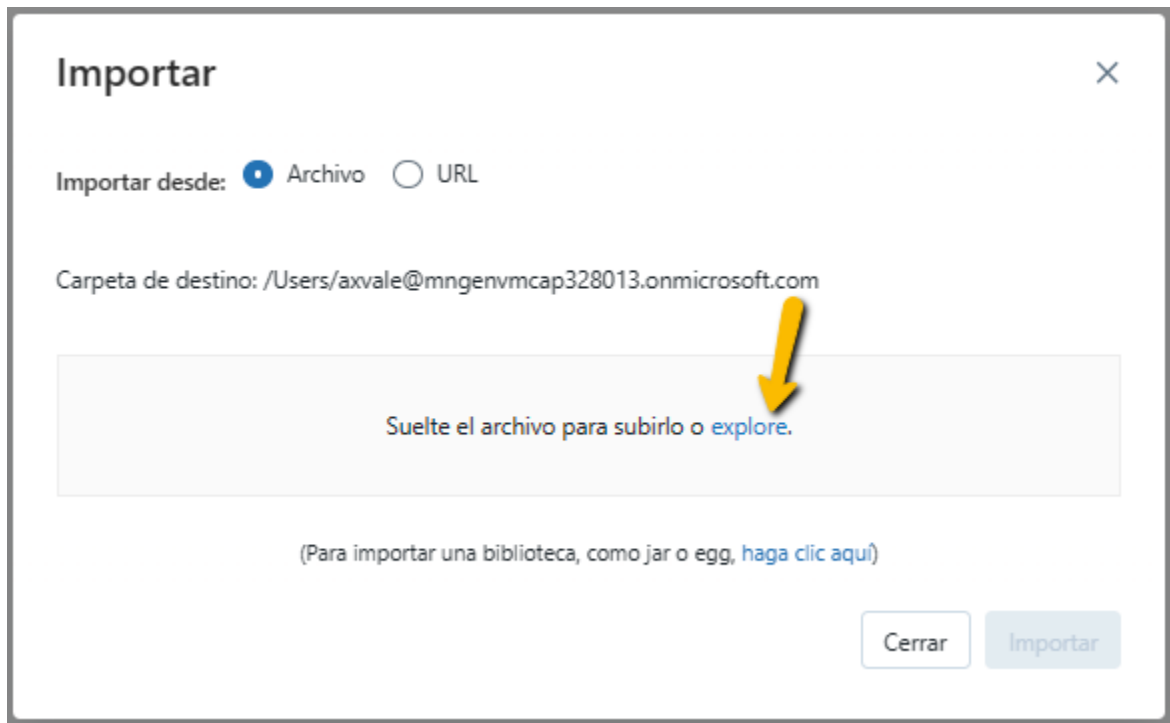
1. En el menú de extrema izquierda clic en “Workspace”



2. En el explorador del workspace, expande las carpetas Workspace>Users hasta encontrar la carpeta de tu usuario
3. Selecciona la carpeta de tu usuario y da clic en el botón de los tres puntos y después selecciona importar



4. En la ventana emergente da clic en explorar y selecciona el archivo en la carpeta Databricks/MDWH\_Workshop.dbc



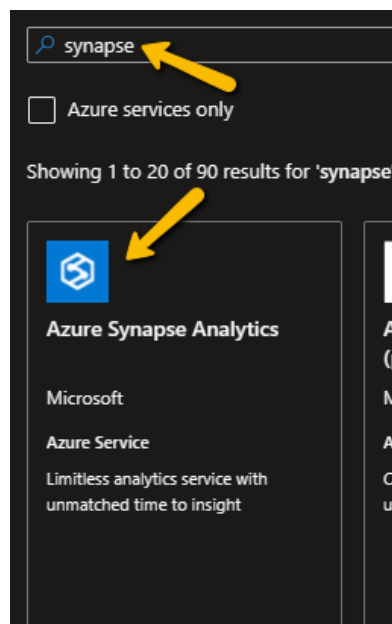
5. Clic en importar

## SYNAPSE

Después de ejecutar el proceso de Databricks es tiempo de crear nuestra capa de consumo

### CREAR EL WORKSPACE DE SYNAPSE

1. Regresa al portal de azure y ubica el grupo de recursos con el que hemos trabajado. Da clic en crear
2. En la caja de busca escribe: "Synapse" y selecciona Azure Synapse Analytics



3. Dar clic en el botón “Crear”
4. En la página de creación selecciona las siguientes opciones:
  - a. Workspace name: tusinicialessynapse
  - b. Region: East US
  - c. Select Data Lake Storage Gen2: From subscription
  - d. Account name: Selecciona el ADLS que creamos para el taller
  - e. File system name: Clic en “Create New”
  - f. Pon el nombre que desees pj: “synapsefilesystem”. Clic en Ok
5. Clic en “Next: Security>”
  - a. Authentication method: Use both local and Microsoft Entra ID
  - b. SQL Server admin login: escribe tus iniciales o un nombre de usuario que recuerdes y una contraseña que recuerdes
  - c. Clic en review y crear
  - d. Clic en crear
  - e. Al terminar la creación, clic en ir al recurso

## CREAR SQL POOL

1. En la pantalla de administración de Synapse, clic en “New dedicated SQL pool”
2. En la pantalla de creación seleccionar las siguientes opciones
  - a. Pool name: tusinicialessqlpool
  - b. Performance Level: Selecciona la opción mas baja : DW100C

\* Basics Additional settings Tags Review + create

Create a dedicated SQL pool with your preferred configurations. Complete the Basics tab then go to Review + Create to provision with smart defaults, or visit each tab to customize. [Learn more](#)

**Dedicated SQL pool details**

Name your dedicated SQL pool and choose its initial settings.

Dedicated SQL pool name \* axvalesqlpool ✓

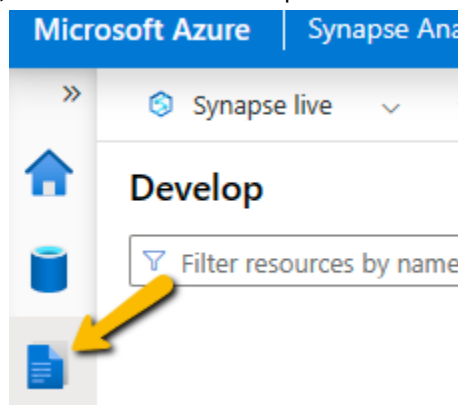
Performance level ⓘ DW100c

Estimated price ⓘ **Est. Cost Per Hour**  
1.51 USD  
[View pricing details](#)

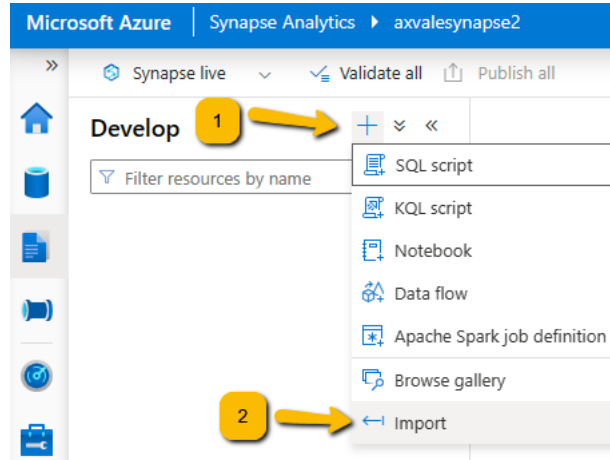
- c. Clic en “review + create”

## IMPORTAR EL SCRIPT

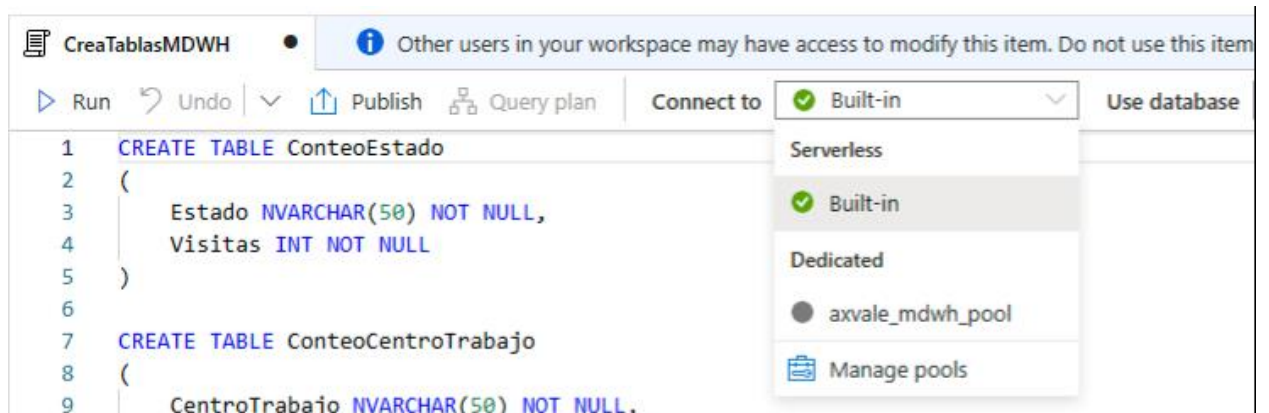
1. Ubica el wrkspc de synapse que acabamos de crear
2. Da clic en el botón “Open Synapse Studio”
3. Estando en el Synapse studio, clic en el botón “Develop”



4. En el menú “Develop”, clic en el botón “Agregar” y después en “Import”



5. Seleccionar el archivo SQL/CreaTablasMDWH.sql
6. Después de importar el archivo, seleccionar el SQL pool dedicado en “Connect to”



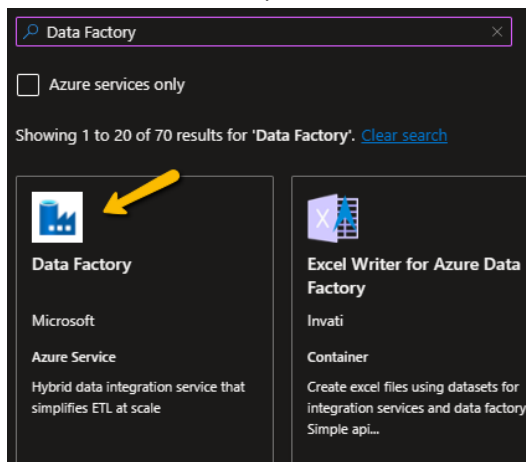
7. Ejecutar el código con el botón “Run”

## AZURE DATA FACTORY

Utilizaremos Azure Datafactory para trasdar el modelo oro creado en Delta hacia Synapse. Si bien esto se puede hacer directamente desde Databricks, Datafactory nos permitirá orquestar procesos complejos de manera más sencilla

### 1. CREANDO DATA FACTORY

1. Ubica el grupo de recursos del taller y da clic en crear
2. En el recuadro de búsqueda escribe “Data Factory” y elije la primera opción de los resultados



3. Da clic en “crear”
4. En la pantalla de creación ingresa la siguiente información:
  - a. **Nombre:** tusiniciales**DataFactory**
  - b. **Region:** EastUs
  - c. **Version:** V2

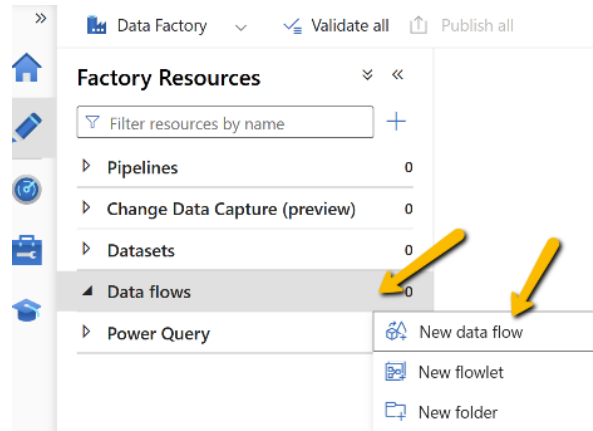
A screenshot of the 'Instance details' form in the Azure portal. The form has three rows. The first row is 'Name \*' with a value of 'avaleDataFactory' and a green checkmark icon. The second row is 'Region \*' with a value of 'East US' and a dropdown arrow icon. The third row is 'Version \*' with a value of 'V2' and a dropdown arrow icon. Each row has a small information icon (i) to its left.

5. Clic en “Review + create”
6. Clic en “Create”

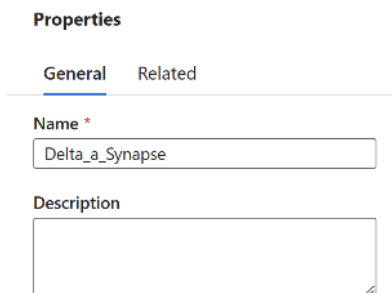


## 2. CREANDO EL DATA FLOW

1. Selecciona, en el grupo de recursos, la instancia de Data Factory que se acaba de crear
2. Da clic en el botón “Launch Studio”
3. Estando en el Studio da clic en el botón “Author” (Lápiz) ubicado en el menú de extrema derecha
4. En el menú del medio, da clic en el botón de los tres puntos de “Data Flow” y selecciona “New Data Flow”



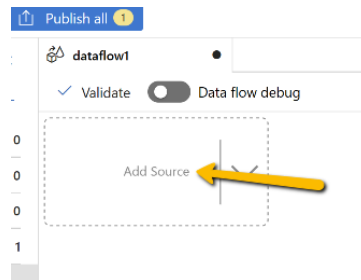
5. En la ventana de las propiedades del dataflow (extrema izquierda) nombra al data flow como “Delta\_a\_Synapse”



---

## CREACIÓN DEL SOURCE

6. En la sección central (Canvas o lienzo) da clic en el botón “Add Source”



7. En las opciones de la fuente (abajo del lienzo)
  - a. Nombra a la fuente como “ConteoAnioMes”
  - b. En source type : “InLine”
  - c. Inline data set type: Delta
  - d. Linked Service: Clic en “+ New”

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

Description  [Reset](#)

Source type \*

Dataset

Inline

Inline dataset type \*

Linked service \*  [+ New](#)

Sampling \* ☐ Enable ☒ Disable

#### CREACIÓN DEL LINKED SERVICE (SOURCE)

1. Selecciona Azure Data Lake Storage Gen 2, clic en continue
2. En las opciones de creación:
  - a. Nombre: Fuente\_ADLS
  - b. Account selection method: From Azure subscription
  - c. Azure Subscription: Selecciona la suscripción del taller
  - d. Storage Account Name: Selecciona la cuenta de storage que creamos en la primer sección
3. Clic en test connection
4. Si la conexión fue satisfactoria, clic en Create

Después de haber creado el linked Service es necesario especificar la ruta de donde se obtendrán los datos

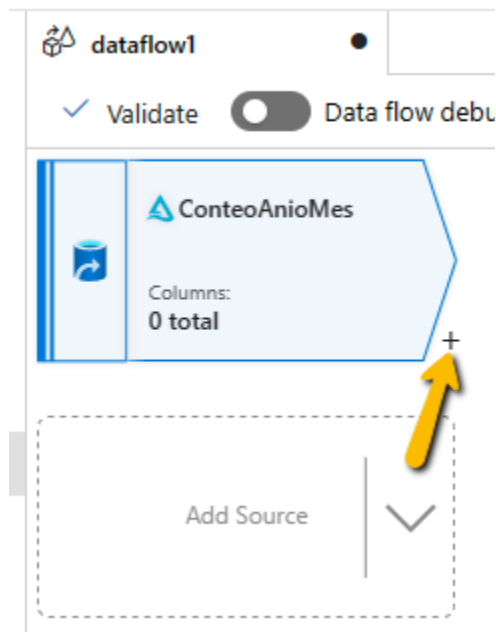
1. Selecciona la pestaña "Source Options" en la parte inferior de la pantalla
2. En folder path, selecciona Browse y navega hasta Root  
Folder>datosfuente>DB>INAH\_ORO>ConteoAnioMes
3. Clic "Ok"

---

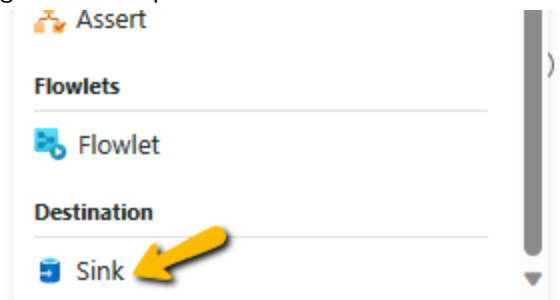
## CREACIÓN DEL DESTINO (SINK)

Ya que configuramos el origen de la información ahora es necesario configurar el destino final, en este caso tablas en Synapse

1. Regresa al Dataflow y da clic en el botón “+” ubicado en la esquina inferior derecha del origen “ConteoAnioMes”



2. En el menú desplegable elige la última opción “Sink”



3. Define los siguientes valores en la pestaña Sink
  - a. Output stream name: DestinoConteoAnioMes
  - b. DataSet : clic en “+ New”

---

## CREACIÓN DEL LINKED SERVICE (SINK)

4. En la selección de conector busca “Synapse” y elige “Azure Synapse Analytics”
5. Define el nombre como “SynapseConteoAnioMes”
6. Linked Service: Crear uno nuevo
  - a. Nombre: SynapseTaller
  - b. Account Selection method: From Azure subscription
  - c. Azure subscription: Elegir la suscripción que estás usando para el taller
  - d. Server name: Selecciona el servidor de synapse que hemos creado
  - e. Database name: El sql pool creado
  - f. Authentication: SQL Authentication
  - g. User name y password: Usa las credenciales de SQL admin
  - h. Clic en Test connection
  - i. Si la conexión fue satisfactoria , clic en créate

- De regreso en la pantalla del sink, elije la tabla “ConteoAnioMes”

Name  
SynapseConteoAnioMes

Linked service \*  
SynapseTaller

☒ Select from existing table ☐ New table

Table name  
dbo.ConteoAnioMes

☐ Edit

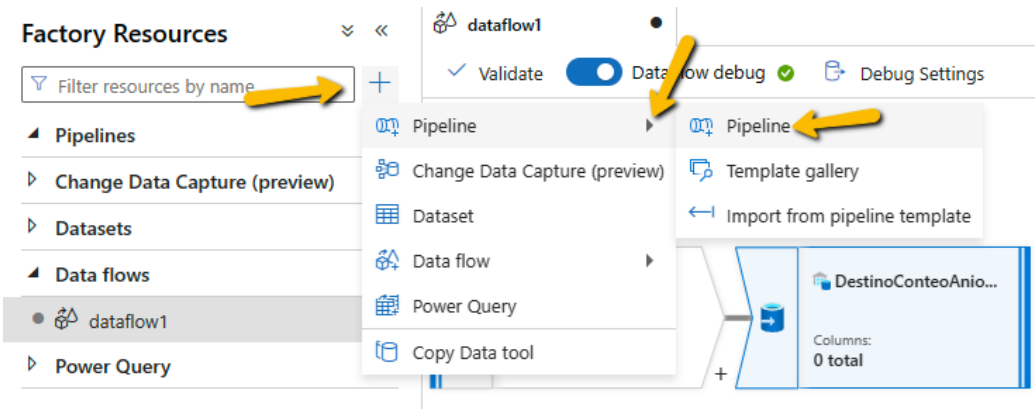
Import schema  
☒ From connection/store ☐ None

> Advanced

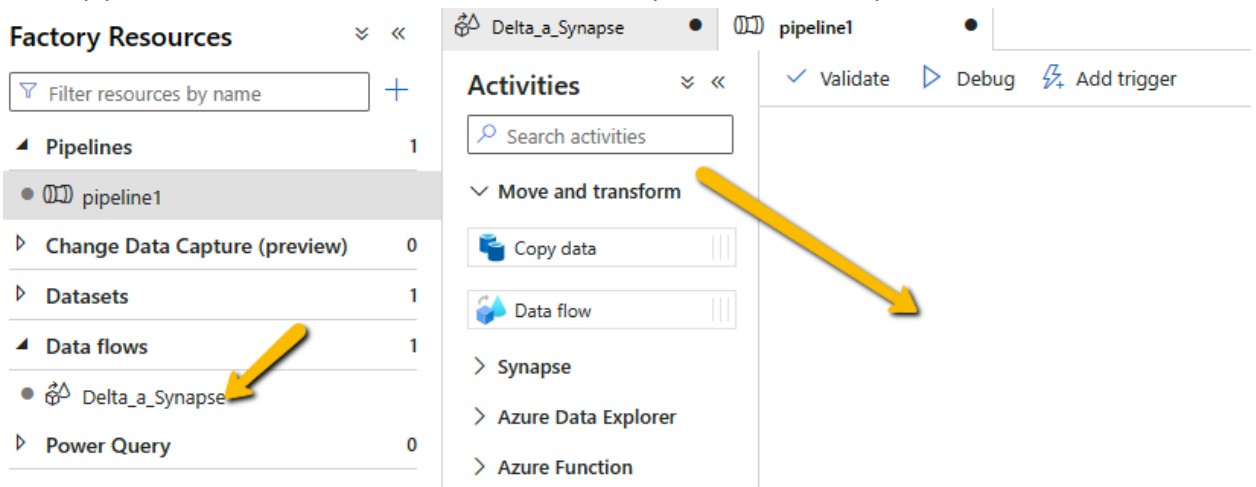
- Clic en Ok

### 3. CREANDO EL PIPELINE

- En el menú central del Studio, clic en “+” y después “pipeline”



- Con el pipeline creado, arrastra al canvas el data Flow que creamos en los pasos anteriores



- Clic en “Debug”
- Si todo funcionó correctamente, clic en publish

## COMPLETANDO EL DATA FLOW

Ahora repite los pasos de los segmentos 2 y 3 para agregar al data flow la fuente y el destino de las tablas ConteoEstado y ConteoNacionalidad.

Tendrá que verse algo como la siguiente imagen:

