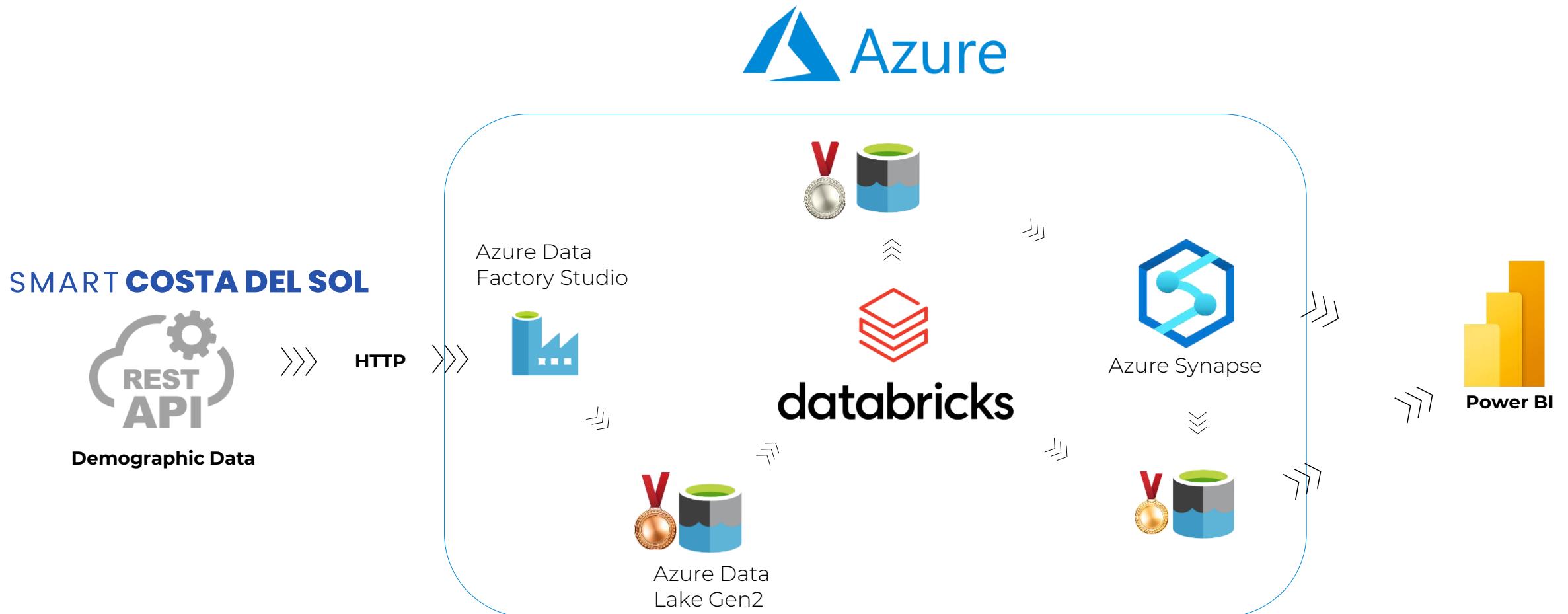


Azure Databricks End-to-End Data Engineering Pipeline



- API data source – Smart Costa del Sol – Demographic Data
<https://catalogo.smartcostadelsol.es/>

SMART COSTA DEL SOL Descripción de la iniciativa Conjuntos de datos SPARQL API Estadísticas de uso

BIENVENIDA

Los Datos Abiertos y su reutilización redundan en una mejora de la calidad de vida de los ciudadanos en la medida en la que son el punto de partida para el desarrollo de nuevos servicios y aplicaciones que no de no ser por la puesta a disposición que de ellos se realiza no estarian disponiblesA través de este espacio, la Agrupación de Municipios Smart Costa del Sol, persigue que los datos y la información se publiquen de forma abierta, regular y reutilizable para todo el mundo, estableciendo nuevos mecanismos de comunicación permanente y transparente con la ciudadanía.A su vez se encuentran publicadas las licencias de cada uno de los conjuntos de datos, con la finalidad de facilitar su reutilización.

[Organizaciones](#) [Grupos](#)



Alhaurin de la Torre



Antequera



Benalmádena







- API light instructions
- These are endpoints for listing packages
- I've explored them with postman (SEE POSTMAN COLLECTION IN MY REPOSITORY)
- Once located the package, take the endpoint url to get the data
 - Personally I find this API could be improved.

SMART COSTA DEL SOL

[Descripción de la iniciativa](#) [Conjuntos de datos](#) [SPARQL](#) [API](#) [Estadísticas de uso](#)

Open Data Portal API

El presente Portal de Datos Abiertos está basado en CKAN que es una aplicación desarrollada bajo el paradigma de software libre y que ofrece una solución para el almacenamiento de los datos abiertos y facilitar su distribución independientemente del formato en que se encuentren con el objetivo de que la reutilización de datos sea un hecho.

Este software presenta una API muy potente que permite el acceso a los diferentes datos almacenados y facilitando en gran medida la integración de aplicaciones.

En los enlaces expuestos a continuación se enumeran los principales recursos expuestos por dicha API para el acceso a los datos con la información necesaria y detalladamente especificada para su correcto consumo:

Obtener listas con formato JSON de los conjuntos de datos de un sitio, grupos u otros objetos CKAN:

https://catalogo.smartcostadelsoles/api/3/action/package_list

https://catalogo.smartcostadelsoles/api/3/action/group_list

https://catalogo.smartcostadelsoles/api/3/action/tag_list

Obtener una representación JSON completa de un dataset, un recurso u otro objeto:

https://catalogo.smartcostadelsoles/api/3/action/package_show

https://catalogo.smartcostadelsoles/api/3/action/tag_show?id=turismo

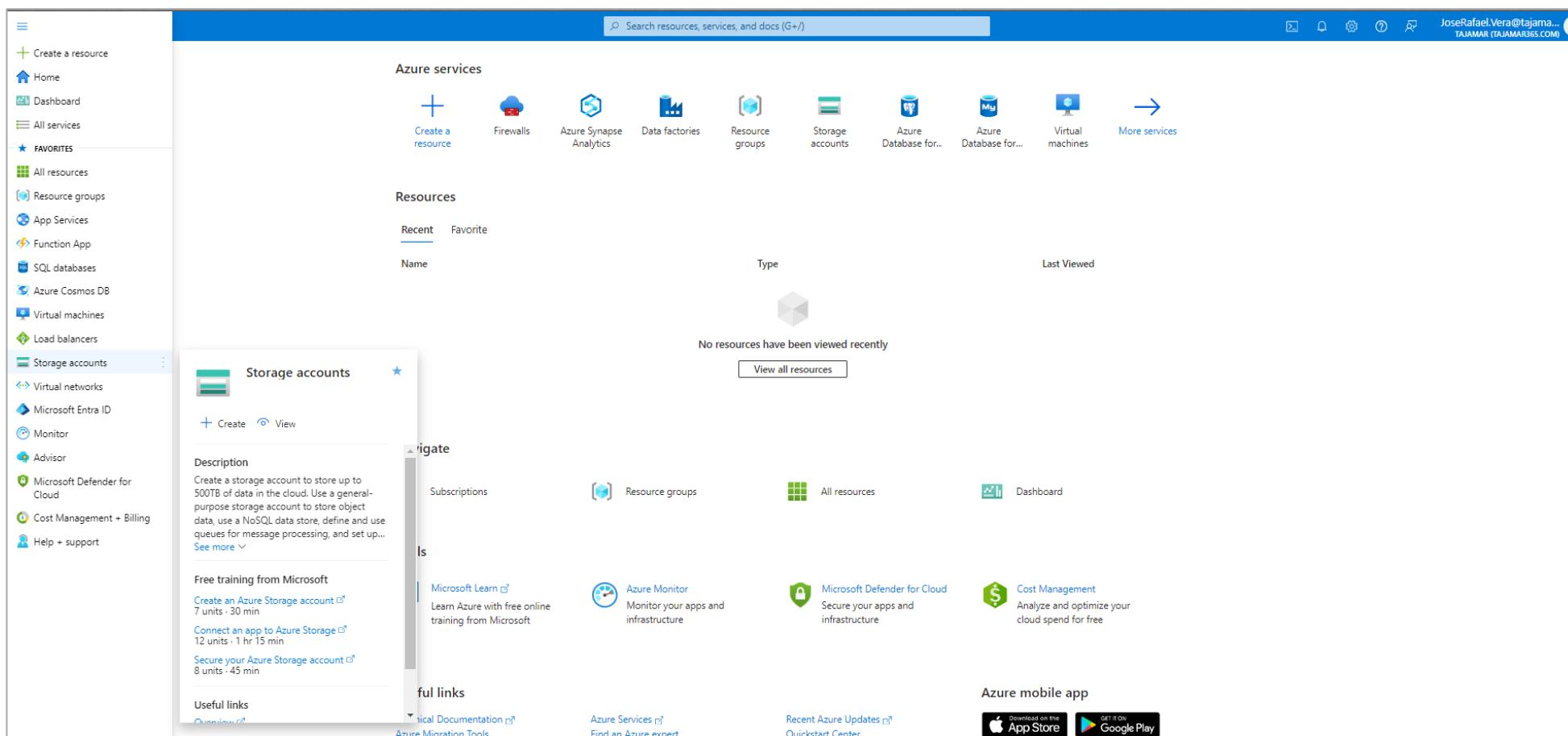
https://catalogo.smartcostadelsoles/api/3/action/group_show?id=comercio

Buscar paquetes o recursos que coincidan con una consulta:

https://catalogo.smartcostadelsoles/api/3/action/package_search?q=mercado

[https://catalogo.smartcostadelsoles/api/3/action/resource_search?query=name\\$](https://catalogo.smartcostadelsoles/api/3/action/resource_search?query=name$)

- Create a Storage Account



- Create a Storage Account

Microsoft Azure

Search resources, services, and docs (G+/)

Home > Storage accounts

Tajamar (tajamar365.com)

+ Create ⌂ Restore ⚙ Manage view ⏪ Refresh ⏴ Export to CSV ⚡ Open query | ⚡ Assign tags 🗑 Delete

Filter for any field... Subscription equals all Resource group equals all Location equals all Add filter

Showing 0 to 0 of 0 records.

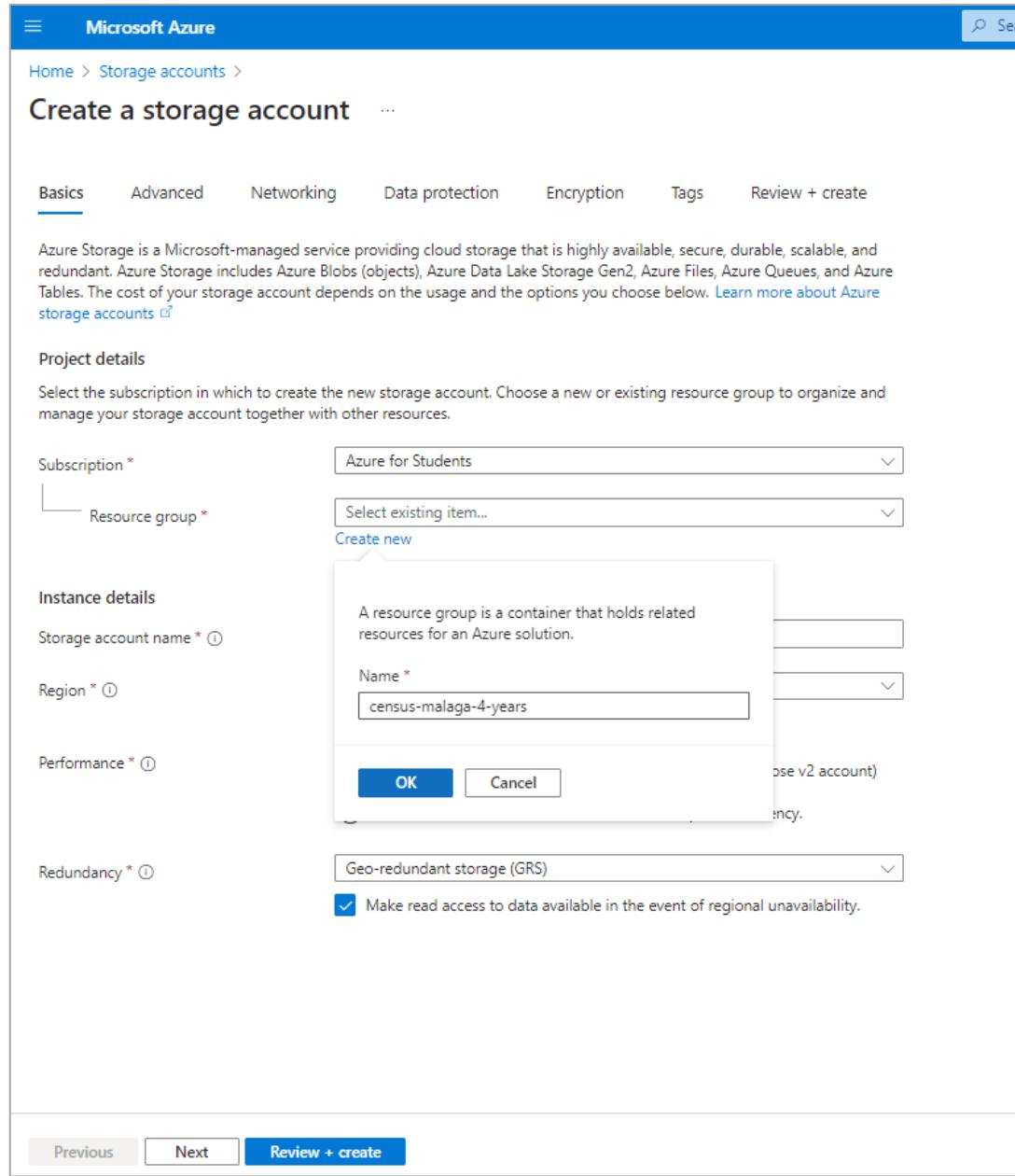
Name ↑↓	Type ↑↓	Kind ↑↓	Resource group ↑↓
---------	---------	---------	-------------------

No storage accounts to display

Create a storage account to store up to 500TB of data in the cloud. Use a general-purpose storage account to store object data, use a NoSQL data store, define and use queues for message processing, and set up file shares in the cloud. Use the Blob storage account and the hot or cool access tiers to optimize your costs based on how frequently your object data is accessed.

[Create storage account](#) [Create](#) [Learn more](#)

- Give your Free subscription.
- Create a Resource group. Just give it a name.



- Give the storage account a name
- Region: I highly recommend southEast Asia. It has more services and computing engine available
- Next

Microsoft Azure

Home > Storage accounts > Create a storage account ...

Basics Advanced Networking Data protection Encryption Tags Review + create

Azure Storage is a Microsoft-managed service providing cloud storage that is highly available, secure, durable, scalable, and redundant. Azure Storage includes Azure Blobs (objects), Azure Data Lake Storage Gen2, Azure Files, Azure Queues, and Azure Tables. The cost of your storage account depends on the usage and the options you choose below. [Learn more about Azure storage accounts](#)

Project details

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 *	Azure for Students
Resource group *	(New) census-malaga-4years
Create new	

Instance details

Storage account name * ⓘ	censusmalagadata
Region * ⓘ	(Europe) West Europe
Deploy to an edge zone	
Performance * ⓘ	<input checked="" type="radio"/> Standard: Recommended for most scenarios (general-purpose v2 account) <input type="radio"/> Premium: Recommended for scenarios that require low latency.
Redundancy * ⓘ	Geo-redundant storage (GRS)
<input checked="" type="checkbox"/> Make read access to data available in the event of regional unavailability.	

Previous [Next](#) [Review + create](#)

- Next

Home > Storage accounts >

Create a storage account ...

Basics Advanced Networking Data protection Encryption Tags Review + create

Security

Configure security settings that impact your storage account.

Require secure transfer for REST API operations

Allow enabling anonymous access on individual containers

Enable storage account key access

Default to Microsoft Entra authorization in the Azure portal

Minimum TLS version

Permitted scope for copy operations (preview)

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

Access protocols

Blob and Data Lake Gen2 endpoints are provisioned by default [Learn more](#)

Enable SFTP

Enable network file system v3

● The current combination of storage account kind, performance, replication, and location does not support the NFS v3 feature. [Learn more about NFS v3](#)

Blob storage

Allow cross-tenant replication

● Cross-tenant replication and hierarchical namespace cannot be enabled simultaneously.

Access tier Hot: Optimized for frequently accessed data and everyday usage scenarios
 Cool: Optimized for infrequently accessed data and backup scenarios

Azure Files

Enable large file shares

● The current combination of storage account kind, performance, replication and location does not support large file shares.

[Previous](#) [Next](#) [Review + create](#)

- Next

Home > Storage accounts >

Create a storage account

Basics Advanced Networking Data protection Encryption Tags Review + create

Network connectivity

You can connect to your storage account either publicly, via public IP addresses or service endpoints, or privately, using a private endpoint.

Network access *

Enable public access from all networks
 Enable public access from selected virtual networks and IP addresses
 Disable public access and use private access

ⓘ Enabling public access from all networks might make this resource available publicly. Unless public access is required, we recommend using a more restricted access type. [Learn more](#)

Network routing

Determine how to route your traffic as it travels from the source to its Azure endpoint. Microsoft network routing is recommended for most customers.

Routing preference * ⓘ

Microsoft network routing
 Internet routing

Previous Next Review + create

- Next

Home > Storage accounts >

Create a storage account ...

Basics Advanced Networking Data protection Encryption Tags Review + create

Recovery

Protect your data from accidental or erroneous deletion or modification.

Enable point-in-time restore for containers
Use point-in-time restore to restore one or more containers to an earlier state. If point-in-time restore is enabled, then versioning, change feed, and blob soft delete must also be enabled. [Learn more ↗](#)

Enable soft delete for blobs
Soft delete enables you to recover blobs and directories that were previously marked for deletion. [Learn more ↗](#)
Days to retain deleted blobs

Enable soft delete for containers
Soft delete enables you to recover containers that were previously marked for deletion. [Learn more ↗](#)
Days to retain deleted containers

Enable soft delete for file shares
Soft delete enables you to recover file shares that were previously marked for deletion. [Learn more ↗](#)
Days to retain deleted file shares

Tracking

Manage versions and keep track of changes made to your blob data.

Enable versioning for blobs
Use versioning to automatically maintain previous versions of your blobs. [Learn more ↗](#)
Consider your workloads, their impact on the number of versions created, and the resulting costs. Optimize costs by automatically managing the data lifecycle. [Learn more ↗](#)

Enable blob change feed
Keep track of create, modification, and delete changes to blobs in your account. [Learn more ↗](#)

Access control

Enable version-level immutability support
Allows you to set time-based retention policy on the account-level that will apply to all blob versions. Enable this feature to set a default policy at the account level. Without enabling this, you can still set a default policy at the container level or set policies for specific blob versions. Versioning is required for this property to be enabled. [Learn more ↗](#)

[Previous](#) [Next](#) [Review + create](#)

- Next

Home > Storage accounts >

Create a storage account ...

Basics Advanced Networking Data protection **Encryption** Tags Review + create

Encryption type * ⓘ Microsoft-managed keys (MMK) Customer-managed keys (CMK)

Enable support for customer-managed keys ⓘ Blobs and files only All service types (blobs, files, tables, and queues)
⚠ This option cannot be changed after this storage account is created.

Enable infrastructure encryption ⓘ

[Previous](#) [Next](#) [Review + create](#)

- Next
- Review and create
- Create
- All stay by default

Home > Storage accounts >

Create a storage account ...

[View automation template](#)

Basics		Advanced	Networking	Data protection	Encryption	Tags	Review + create
Subscription	Azure for Students						
Resource group	census-malaga-4years						
Location	West Europe						
Storage account name	censusmalagadata						
Performance	Standard						
Replication	Read-access geo-redundant storage (RA-GRS)						
Advanced							
Enable hierarchical namespace	Enabled						
Enable SFTP	Disabled						
Enable network file system v3	Disabled						
Allow cross-tenant replication	Disabled						
Access tier	Hot						
Enable large file shares	Disabled						
Security							
Secure transfer	Enabled						
Blob anonymous access	Disabled						
Allow storage account key access	Enabled						
Default to Microsoft Entra authorization in the Azure portal	Disabled						
Minimum TLS version	Version 1.2						
Permitted scope for copy operations (preview)	From any storage account						
Networking							
Network connectivity	Public endpoint (all networks)						
Default routing tier	Microsoft network routing						

Data protection

Point-in-time restore	Disabled
Blob soft delete	Enabled
Blob retention period in days	7
Container soft delete	Enabled
Container retention period in days	7
File share soft delete	Enabled
File share retention period in days	7
Versioning	Disabled
Blob change feed	Disabled
Version-level immutability support	Disabled
Encryption	
Encryption type	Microsoft-managed keys (MMK)
Enable support for customer-managed keys	Blobs and files only
Enable infrastructure encryption	Disabled

[Previous](#) [Next](#) [Create](#)

- Wait until deployment finishes

The screenshot shows the Microsoft Azure Deployment Overview page for a deployment named "censusmalagadata_1711634436905". The deployment status is "Deployment is in progress". The deployment details include:

- Deployment name: censusmalagadata_1711634436905
- Subscription: Azure for Students
- Resource group: census-malaga-4years

The deployment started at 3/28/2024, 3:02:12 PM with a Correlation ID of 8599b4f1-b89f-499e-b93d-78f37f1639fc.

The "Deployment details" section shows a table with columns: Resource, Type, Status, and Operation details. The table displays "No results."

Feedback options at the bottom include "Give feedback" and "Tell us about your experience with deployment".

- Deployment completed
- Go to resource

Microsoft Azure Search resources, services, and docs (G+)

Home > **censusmalagadata_1711634436905 | Overview** ⚡ ...

 Deployment

Search <> Delete Cancel Redeploy Download Refresh

 Overview  Inputs  Outputs  Template

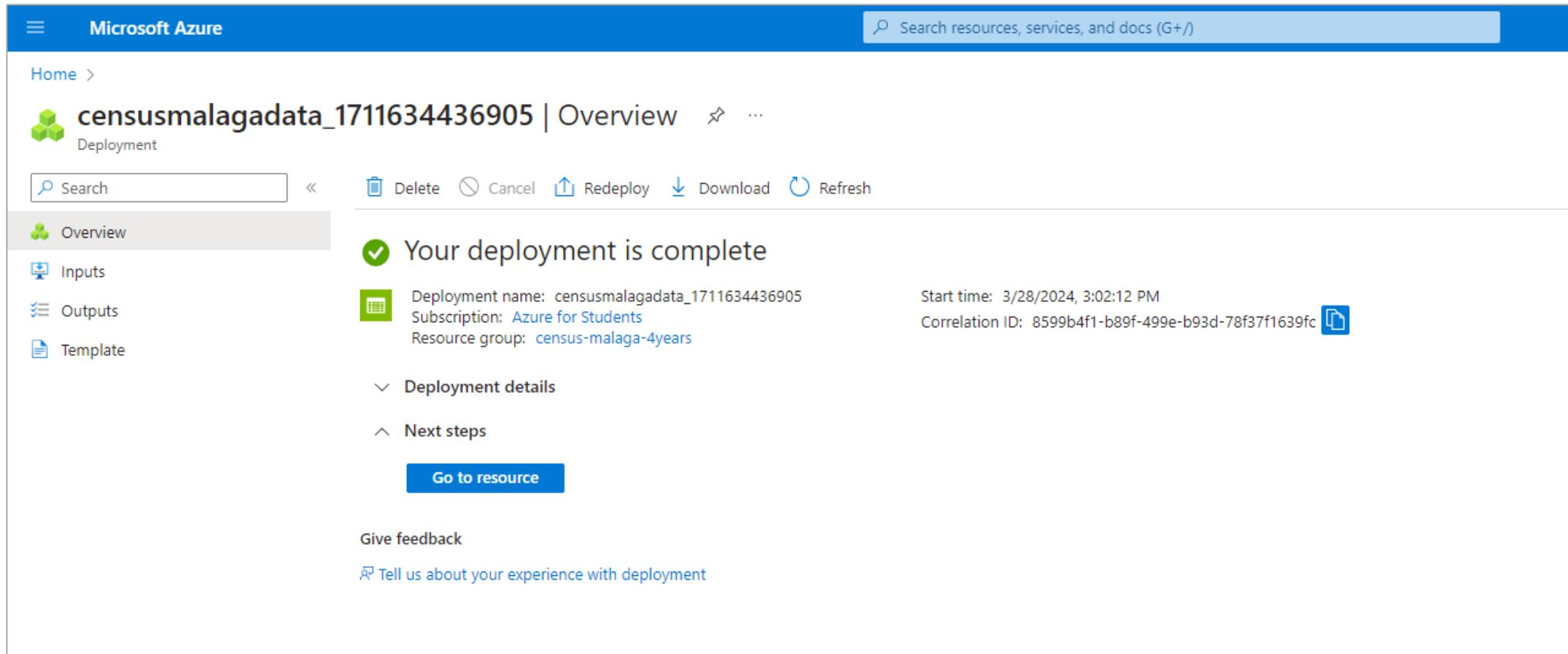
 Your deployment is complete

Deployment name: [censusmalagadata_1711634436905](#) Start time: [3/28/2024, 3:02:12 PM](#)
Subscription: [Azure for Students](#) Correlation ID: [8599b4f1-b89f-499e-b93d-78f37f1639fc](#) 

[Deployment details](#) [Next steps](#)

[Go to resource](#)

Give feedback  Tell us about your experience with deployment 



- Go to containers
- Create a container which will storage our data
- Just give it a name

The screenshot shows the Microsoft Azure portal interface for a storage account named 'censusmalagadata'. The left sidebar contains navigation links for Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data migration, Events, Storage browser, Data storage (Containers, File shares, Queues, Tables), Security + networking (Networking, Access keys, Shared access signature, Encryption, Microsoft Defender for Cloud), and Data management (Storage tasks (preview), Redundancy). The main content area displays the 'Containers' blade, which lists one container named '\$logs' with a last modified date of 3/28/2024, 3:02:38 PM and an anonymous access level of Private. A search bar at the top right allows searching by prefix. On the right side, a 'New container' dialog is open, prompting for a name ('census-malaga-data'), which is highlighted with a red box. The 'Anonymous access level' dropdown is set to 'Private (no anonymous access)'. A note in the dialog states: 'The access level is set to private because anonymous access is disabled on this storage account.' At the bottom right of the dialog are 'Create' and 'Give feedback' buttons.

Name	Last modified	Anonymous access level
\$logs	3/28/2024, 3:02:38 PM	Private

- Inside the container create the three directories

The screenshot shows the Microsoft Azure Storage Container Overview page for the 'census-malaga-data' container. The container name is displayed at the top left, and the location is listed as 'census-malaga-data'. The 'Overview' tab is selected in the left sidebar. The main area displays three directory entries: 'bronze', 'gold', and 'silver', each represented by a yellow folder icon.

Name	Modified
bronze	
gold	
silver	

- Search for Data Factories

The screenshot shows the Microsoft Azure Storage Explorer interface. On the left, there's a sidebar with options like Home, Overview, Diagnose and solve problems, Access Control (IAM), and Settings. The main area displays a container named "census-malaga-data". It includes a search bar, upload and add directory buttons, refresh and rename icons, and a note about authentication and location. A search bar at the top right contains the text "data". Below it, a results pane shows categories: All, Services (99+), Marketplace (20), and More (4). The "Services" section is expanded, showing items like Data factories, Reservations, Azure Database for MySQL servers, and SQL databases. The "Marketplace" section is collapsed.

- Create new Data Factory

Microsoft Azure Search resources, services, and docs (G+) JoseRafael.Vera@tajamar.com
TAJAMAR (TAJAMAR365.COM)

Home > Data factories ...

Tajamar (tajamar365.com)

+ Create Manage view Refresh Export to CSV Open query | Assign tags

Filter for any field... Subscription equals all Type equals all Resource group equals all Location equals all Add filter

Showing 0 to 0 of 0 records. No grouping List view

Name ↑↓	Type ↑↓	Subscription ↑↓	Resource group ↑↓	Location ↑↓
---------	---------	-----------------	-------------------	-------------

 No data factories to display

Try changing or clearing your filters.

[Create data factory](#) [Learn more](#)

- Select your Free subscription
- Select your existing resource group
- Give the instance a name
- Select region southEast Asia. It works better than west Europe

Microsoft Azure

Home > Data factories >

Create Data Factory

Basics Git configuration Networking Advanced Tags Review + create

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

Project details

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

Subscription * [?](#) Azure for Students

Resource group * [?](#) census-malaga-4years [Create new](#)

Instance details

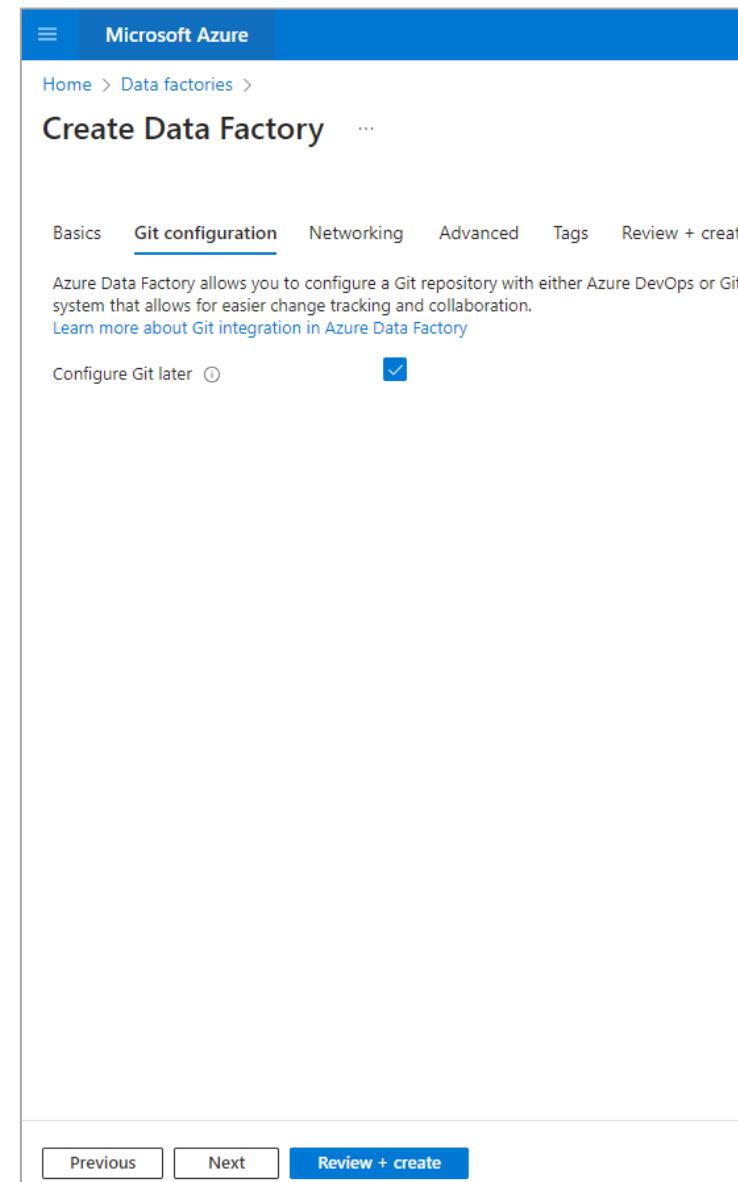
Name * [?](#) census-malaga-df

Region * [?](#) West Europe

Version * [?](#) V2

Previous Next [Review + create](#)

- Configure git later
- next



- Next

Microsoft Azure

Home > Data factories >

Create Data Factory

Basics Git configuration Networking Advanced Tags Review + create

Managed virtual network

Choose whether you want the default AutoResolveIntegrationRuntime to be provisioned on demand inside an ADF-managed virtual network. If this setting is disabled, after the data factory is created, you can still choose whether to provision explicitly created Azure integration runtime inside an ADF-managed virtual network.

[Learn more](#)

Enable Managed Virtual Network on the

default AutoResolveIntegrationRuntime

Self-hosted integration runtime inbound connectivity to Azure Data Factory service

Choose whether to connect your self-hosted integration runtime to Azure Data Factory via public endpoint or private endpoint. This applies to self-hosted integration runtime running either on premises or inside customer managed A virtual network

[Learn more](#)

Connect via * ⓘ

Public endpoint
 Private endpoint

ⓘ You can change this or configure another connectivity method after this resource is created. [Learn more](#)

Previous Next Review + create

- Next
- Next

Microsoft Azure

Home > Data factories >

Create Data Factory

Basics Git configuration Networking Advanced **Tags** Review + create

Datafactory Encryption

By default, data is encrypted with Microsoft-managed keys. For additional control over encryption keys, you can support customer-managed keys to use for encryption of blob and file data. Customer-managed keys must be stored in an Azure Key Vault. You can either create your own keys and store them in a key vault, or you can use the Azure Key Vault API to generate keys. The storage account and the key vault must be in the same region, but they can be in different subscriptions.

Enable encryption using a Customer Managed Key

[Previous](#) [Next](#) [Review + create](#)

Microsoft Azure

Home > Data factories >

Create Data Factory

Basics Git configuration Networking Advanced **Tags** Review + create

Tags are name/value pairs that enable you to categorize resources and view consolidated billing information for a tag across multiple resources and resource groups. [Learn more about tags](#)

Note that if you create tags and then change resource settings on other tabs, your tags will be updated.

Name ⓘ	Value ⓘ	Resource
<input type="text"/>	<input type="text"/> :	Data factory (V)

[Previous](#) [Next](#) [Review + create](#)

- Create

Microsoft Azure

Home > Data factories >

Create Data Factory

Basics Git configuration Networking Advanced Tags Review + create

View automation template

TERMS

By clicking "Create", I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s); (b) authorize Microsoft to bill my current payment method for the fees associated with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my transactional information with the provider(s) of the offering(s) for support, billing and other related activities. Microsoft does not provide rights for third-party offerings. See the [Azure Marketplace Terms of Use](#) for details.

Basics

Subscription	Azure for Students
Resource group	census-malaga-4years
Name	census-malaga-df
Region	West Europe
Version	V2

Networking

Connect via	Public endpoint
-------------	-----------------

Previous Next Create

- Wait until data factory deployment is completed

The screenshot shows the Microsoft Azure Data Factory Overview page for a deployment named "Microsoft.DataFactory-20240328150854". The deployment status is "Deployment is in progress". Deployment details include:

- Deployment name: Microsoft.DataFactory-20240328150854
- Subscription: Azure for Students
- Resource group: census-malaga-4years
- Start time: 3/28/2024, 3:14:32 PM
- Correlation ID: 263a857e-1ea7-420d-925e-ec6fae2e2448

The "Deployment details" section is collapsed. A table below shows no resources displayed.

Resource	Type	Status
There are no resources to display.		

Feedback options are available at the bottom:

- Give feedback
- Tell us about your experience with deployment

- Data Factory instance successfully deployed
- Go to resource

The screenshot shows the Microsoft Azure portal interface for a Data Factory deployment. The top navigation bar includes the Microsoft Azure logo, a search bar, and a 'Search resources, services, and docs (G+ /)' bar. Below the navigation is a breadcrumb trail: Home > Microsoft.DataFactory-20240328150854 | Overview. On the left, there's a sidebar with icons for Deployment, Overview, Inputs, Outputs, and Template. The main content area displays a success message: 'Your deployment is complete' with a green checkmark icon. It provides deployment details: Deployment name: Microsoft.DataFactory-20240328150854, Subscription: Azure for Students, Resource group: census-malaga-4years. It also shows the start time (3/28/2024, 3:14:32 PM) and Correlation ID (263a857e-1ea7-420d-925e-ec6fae2e2448). Below the details are sections for 'Deployment details' (with a 'View' link) and 'Next steps' (with a 'Go to resource' button). At the bottom, there are links for 'Give feedback' and 'Tell us about your experience with deployment'.

Microsoft Azure

Home > Microsoft.DataFactory-20240328150854 | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Overview

Inputs

Outputs

Template

Your deployment is complete

Deployment name : Microsoft.DataFactory-20240328150854

Subscription : Azure for Students

Resource group : census-malaga-4years

Start time : 3/28/2024, 3:14:32 PM

Correlation ID : 263a857e-1ea7-420d-925e-ec6fae2e2448

> Deployment details

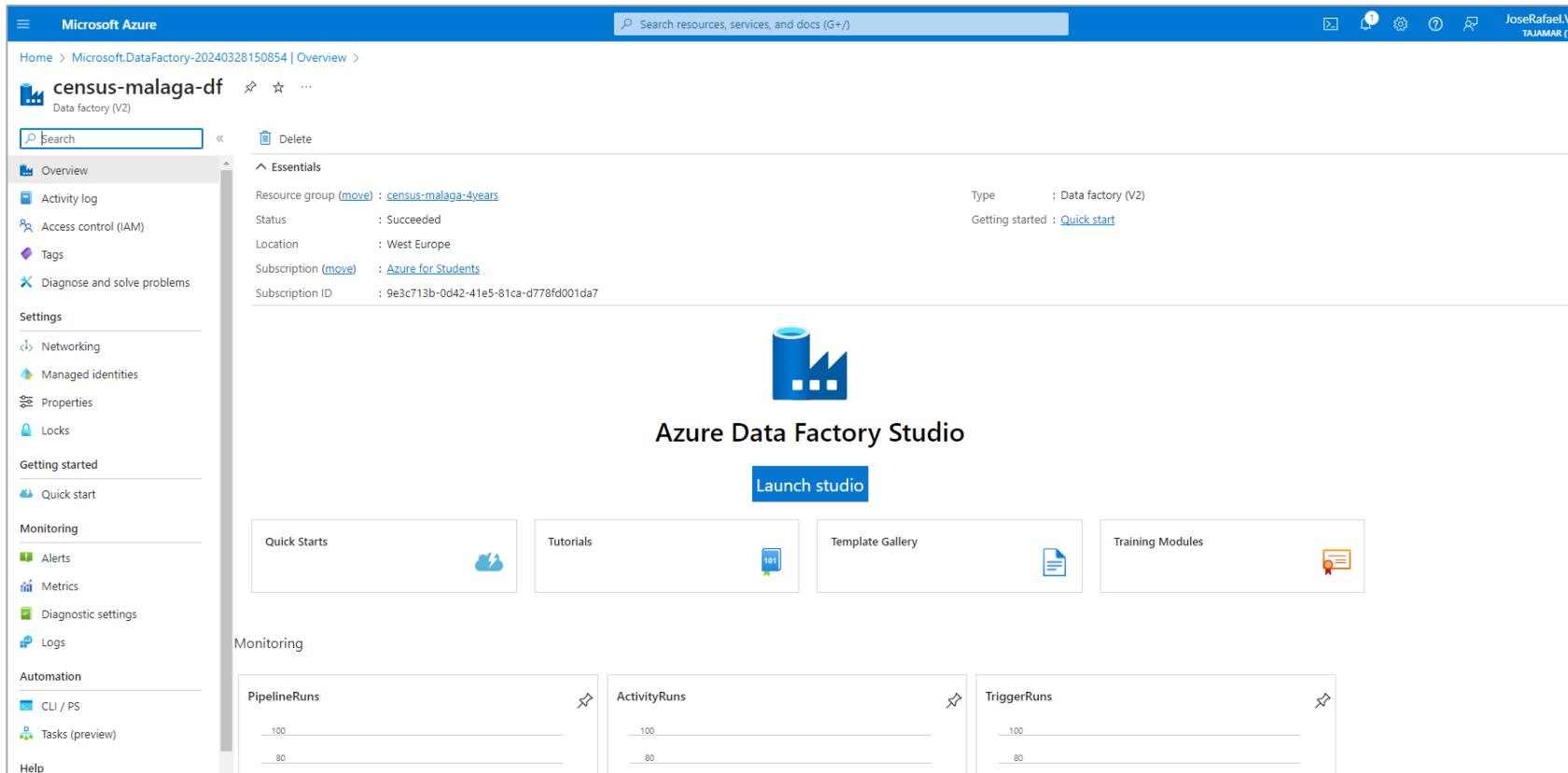
▽ Next steps

Go to resource

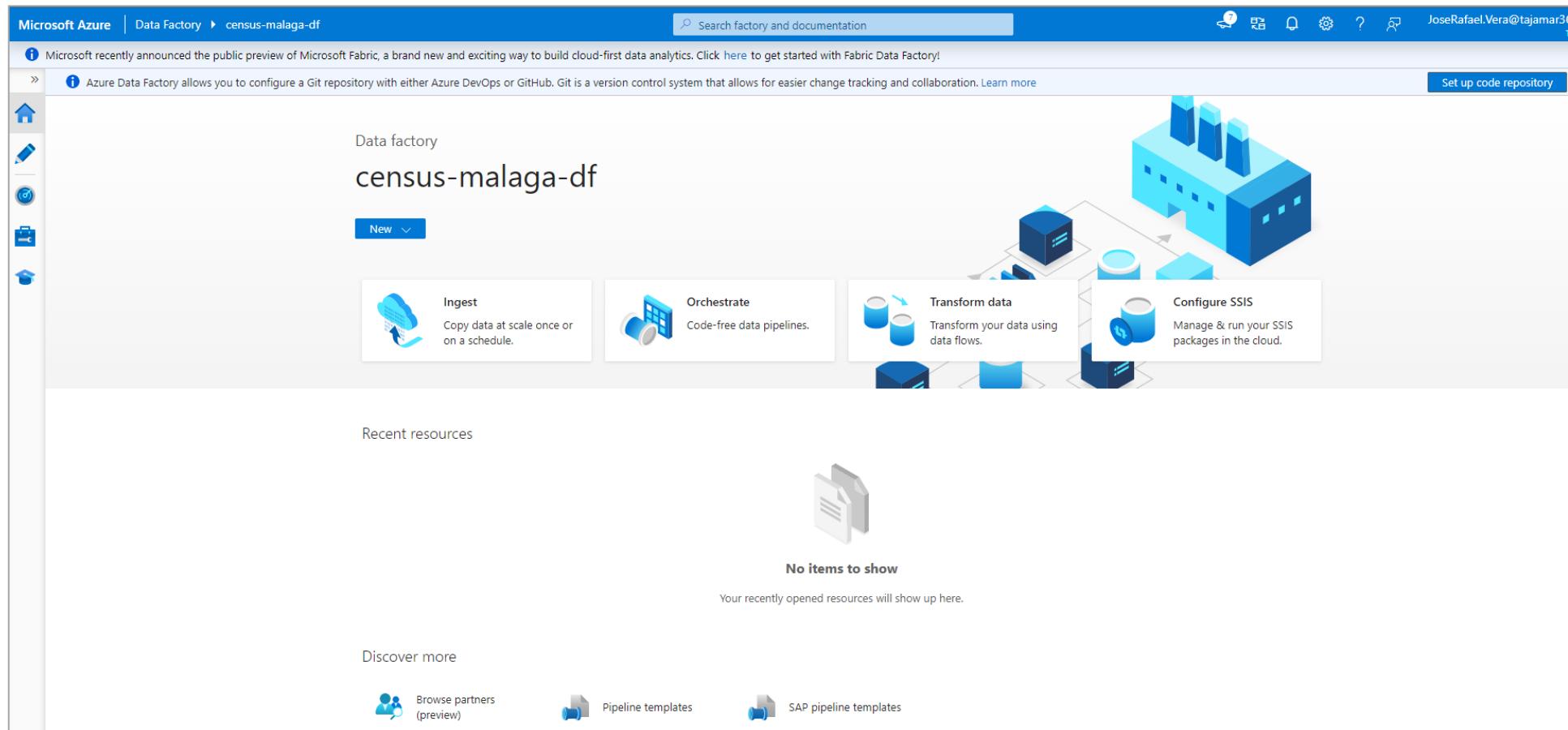
Give feedback

Tell us about your experience with deployment

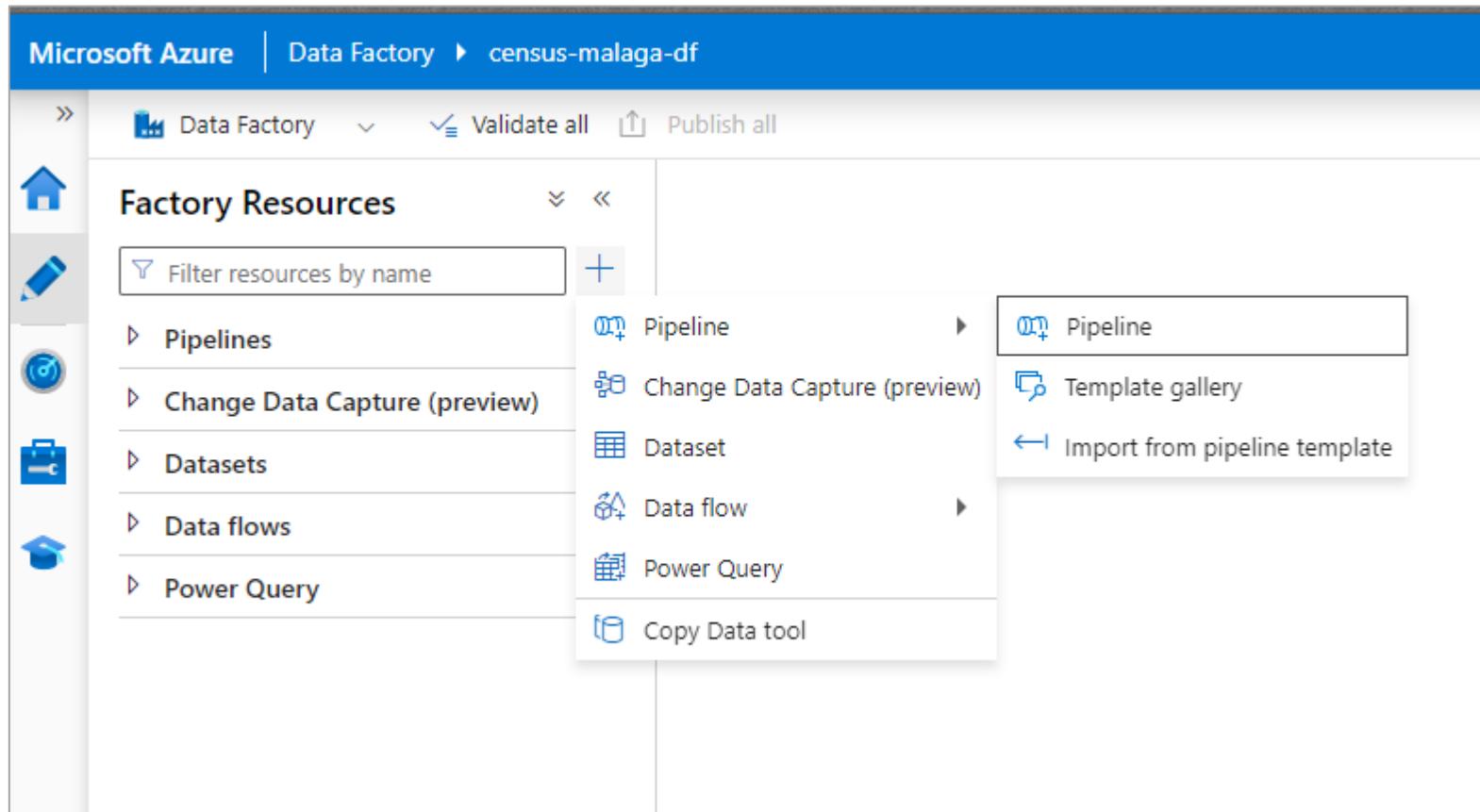
- Launch Data Factory Studio



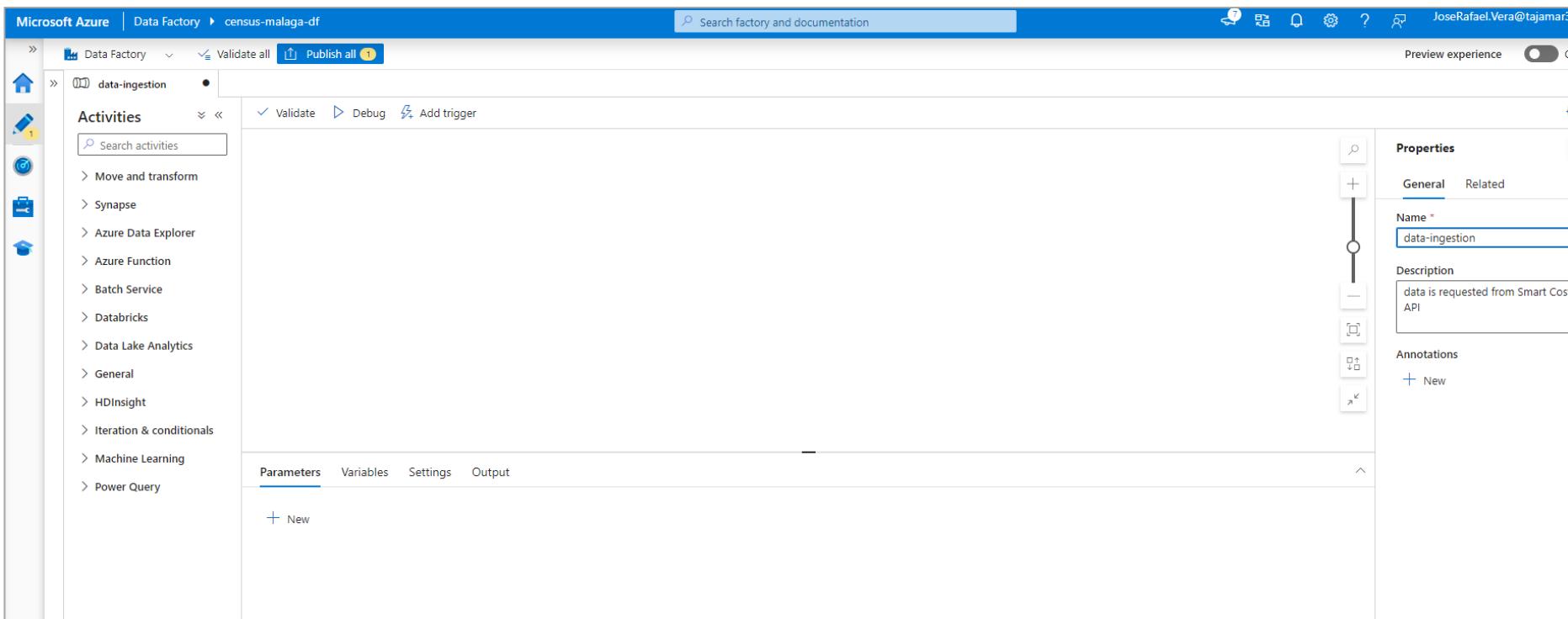
- Click on pencil



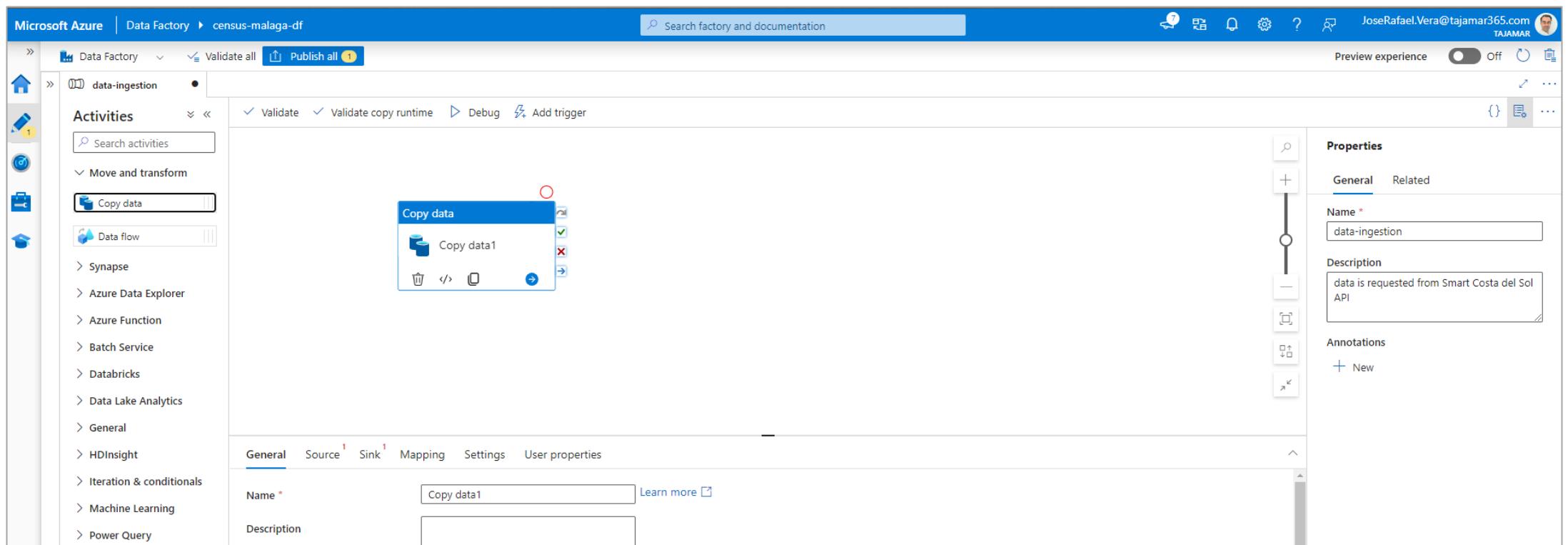
- Click on +
- Select pipeline > pipeline



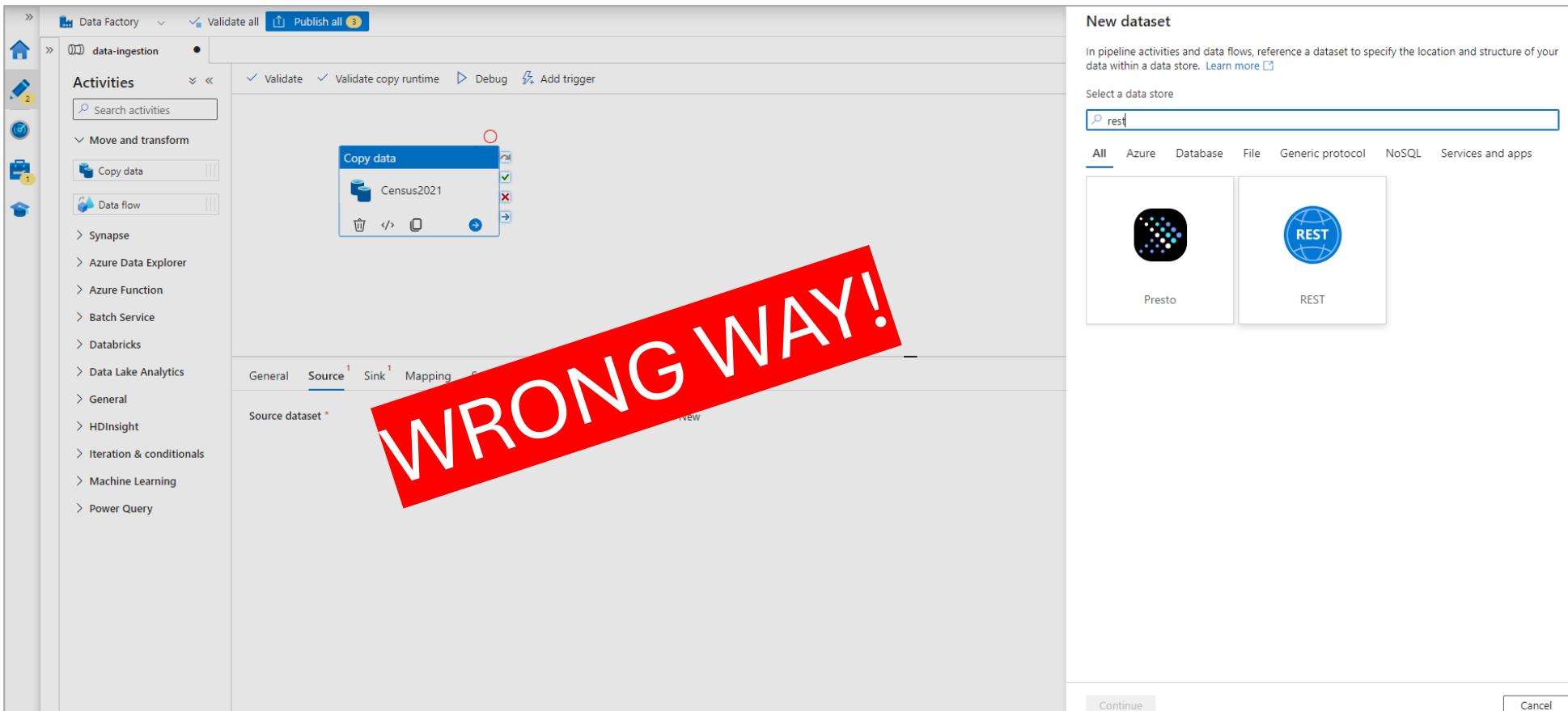
- Console to start building pipelines
- Give the pipeline a name



- Click the dropdown “Move and Transform”
- Drag and drop “copy data” to the central panel
- Give this step a name



- The data source should not be REST, as such Azure service expect only JSON type response.
- The data served from smart Costa del Sol API is as .csv format.
- So we need to select a data source type as HTTP



- Explanation below

The screenshot shows the 'Source' tab of a REST connector configuration. It includes fields for 'Source dataset' (RestResource1), 'Request method' (GET), 'Request timeout' (00:01:40), 'Request interval (ms)' (10), and 'Additional headers'. A red diagonal watermark 'WRONG WAY!' is overlaid across the middle of the screen. Below the headers, there is a table with columns for 'Value' and 'Header name'. One row shows 'Accept' with the value '*/*'. A warning message at the bottom states: '⚠️ REST connector ignores any "Accept" header specified in additional headers. As REST connector only supports response in JSON, it will auto generate a header of Accept: application/json.'

General Source Sink¹ Mapping Settings User properties

Source dataset *

Request method ⓘ

Request timeout ⓘ

Request interval (ms) ⓘ

Additional headers ⓘ

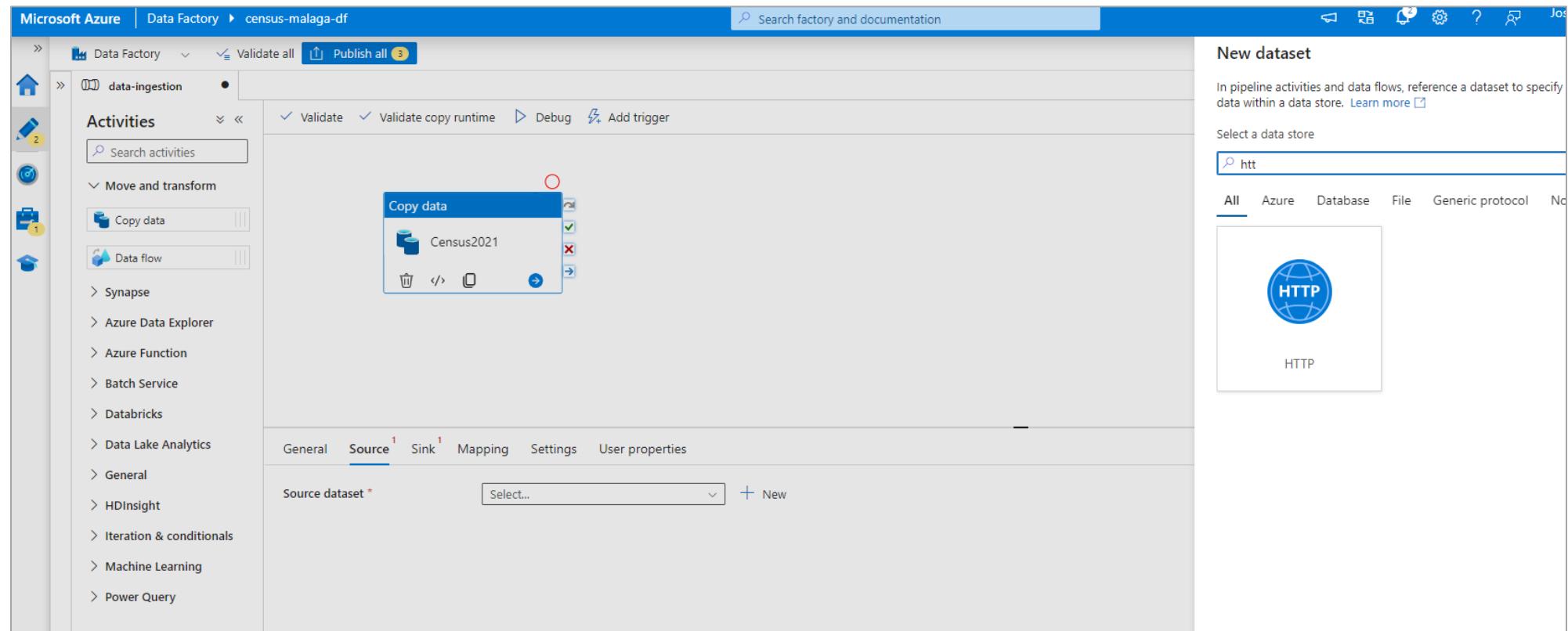
Value

Accept */*

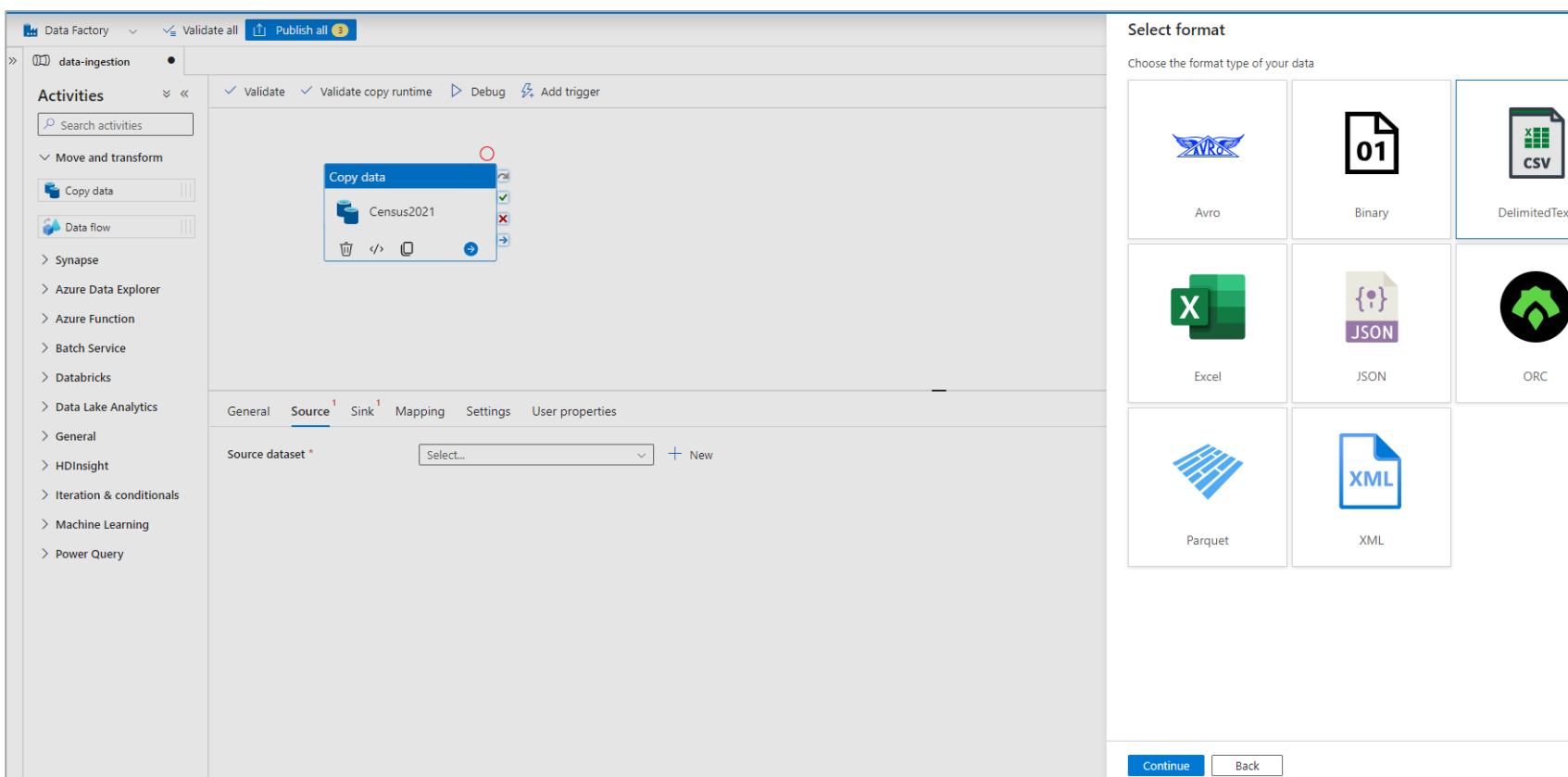
Add dynamic content [Alt+Shift+D]

⚠️ REST connector ignores any "Accept" header specified in additional headers. As REST connector only supports response in JSON, it will auto generate a header of Accept: application/json.

- Selecting the copy data, click on Source below
- Choose the data source as HTTP (if response is csv type)



- Select de input format type. In this case .csv
- continue



- Paste the base URL where Data Factory will request the Data. (in my repository you find all the URLs)
- Authentication “Anonymous”. In production env you may probably set an authn type
- Save

Edit linked service
HTTP Learn more

Name *
HttpCostaDelSolCensus2021

Description

Connect via integration runtime * ⓘ
AutoResolveIntegrationRuntime

Base URL *
https://datosabiertos.malaga.eu/recursos/demografia/padron/2021/padronbarrios.csv
⚠ Information will be sent to the URL specified. Please ensure you trust the URL entered.

Server Certificate Validation ⓘ
Enable Disable

Authentication type * ⓘ
Anonymous

Auth headers ⓘ
+ New

Annotations
+ New

> Parameters

> Advanced ⓘ

Save Cancel Test connection

Copy data

Census2021

General Source Sink¹ Mapping Settings User properties

Source dataset * Census2021 Open + New Preview data

Request method * GET

Additional headers ⓘ

Request body ⓘ

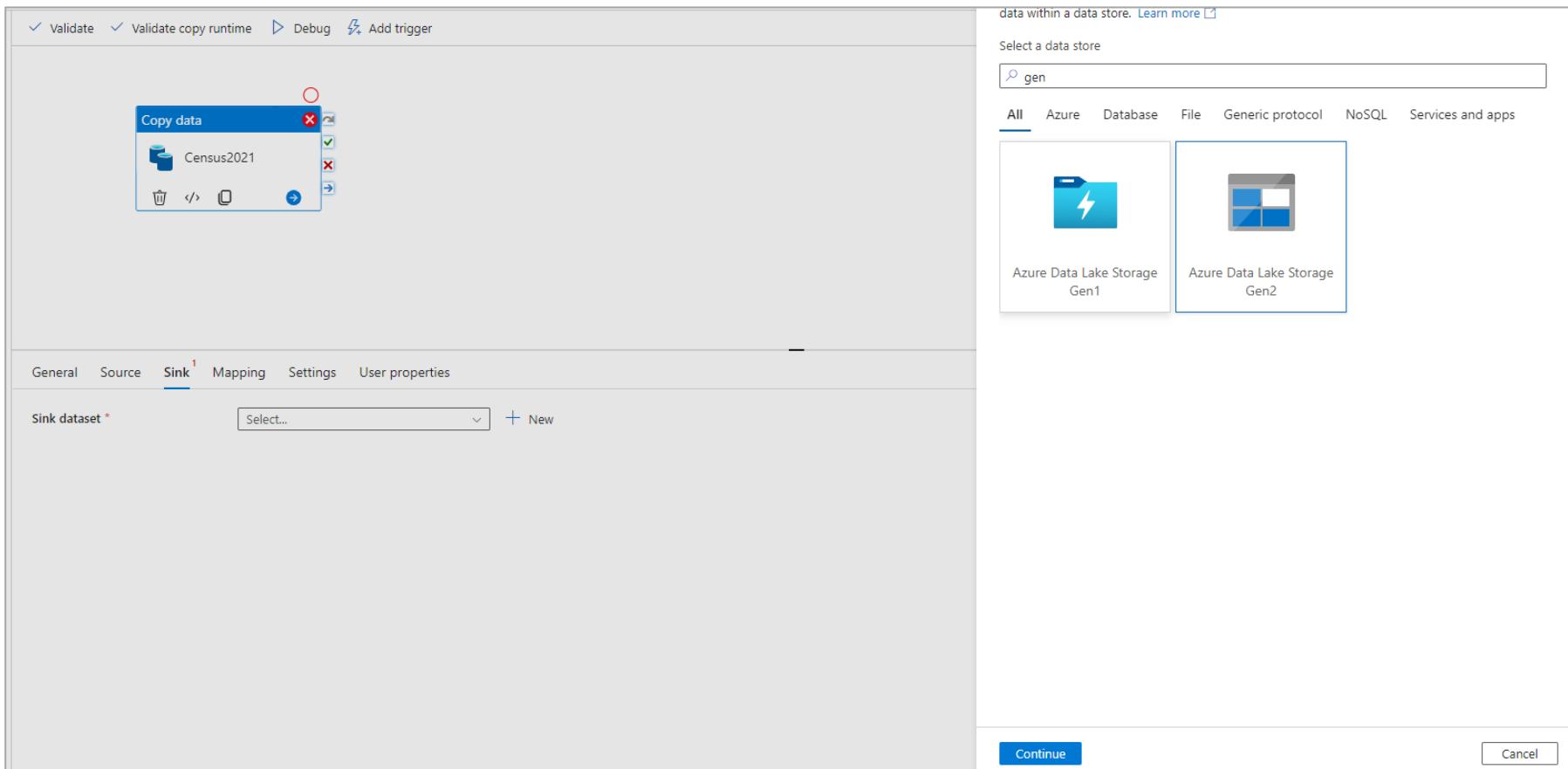
Request timeout ⓘ

Max concurrent connections ⓘ

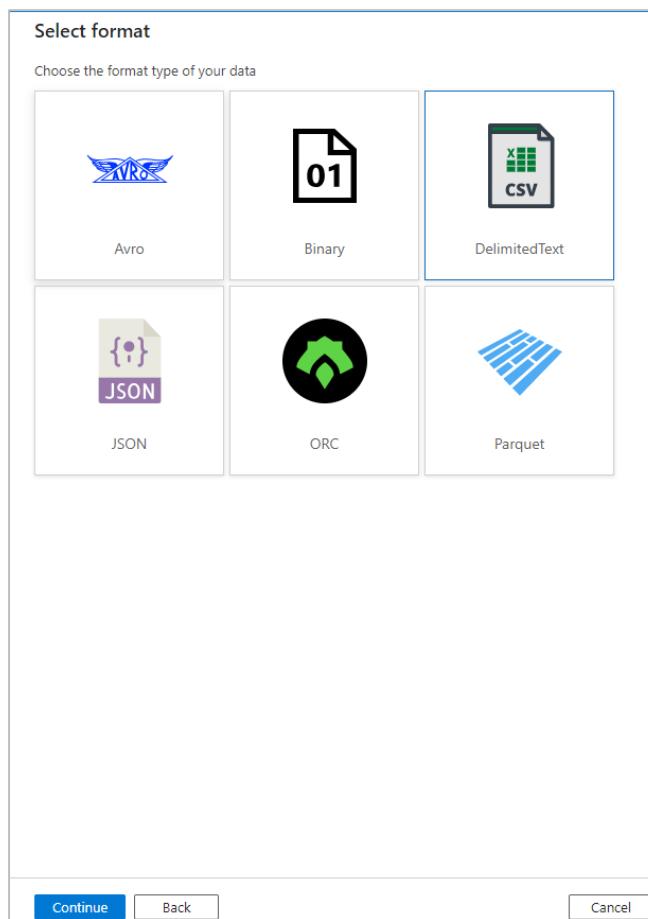
Skip line count

Additional columns ⓘ + New

- Click on “Sink” to define where in Resource group will be the data persisted
- Select ADLSGen2
- Continue



- Select the output file type as .csv
- continue



- Give a Name to this sink
- Linked service- what appears there
- Browse the folder into your ADLS container you've created before
- First row as header
- Import schema “None”

Set properties

Name

Linked service * [Edit](#)

File path / / [Browse](#)

First row as header [Browse](#)

Import schema From connection/store From sample file None

[Advanced](#)

OK Back Cancel

Browse

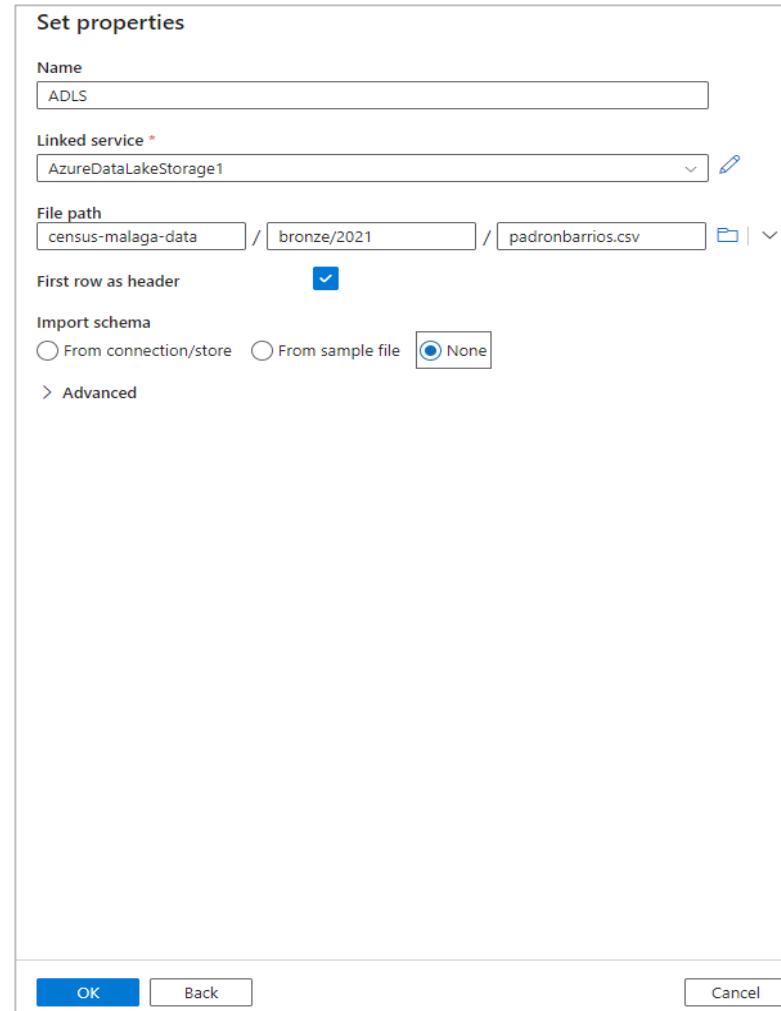
Select a file or folder.

Root folder > census-malaga-data > bronze > 2021

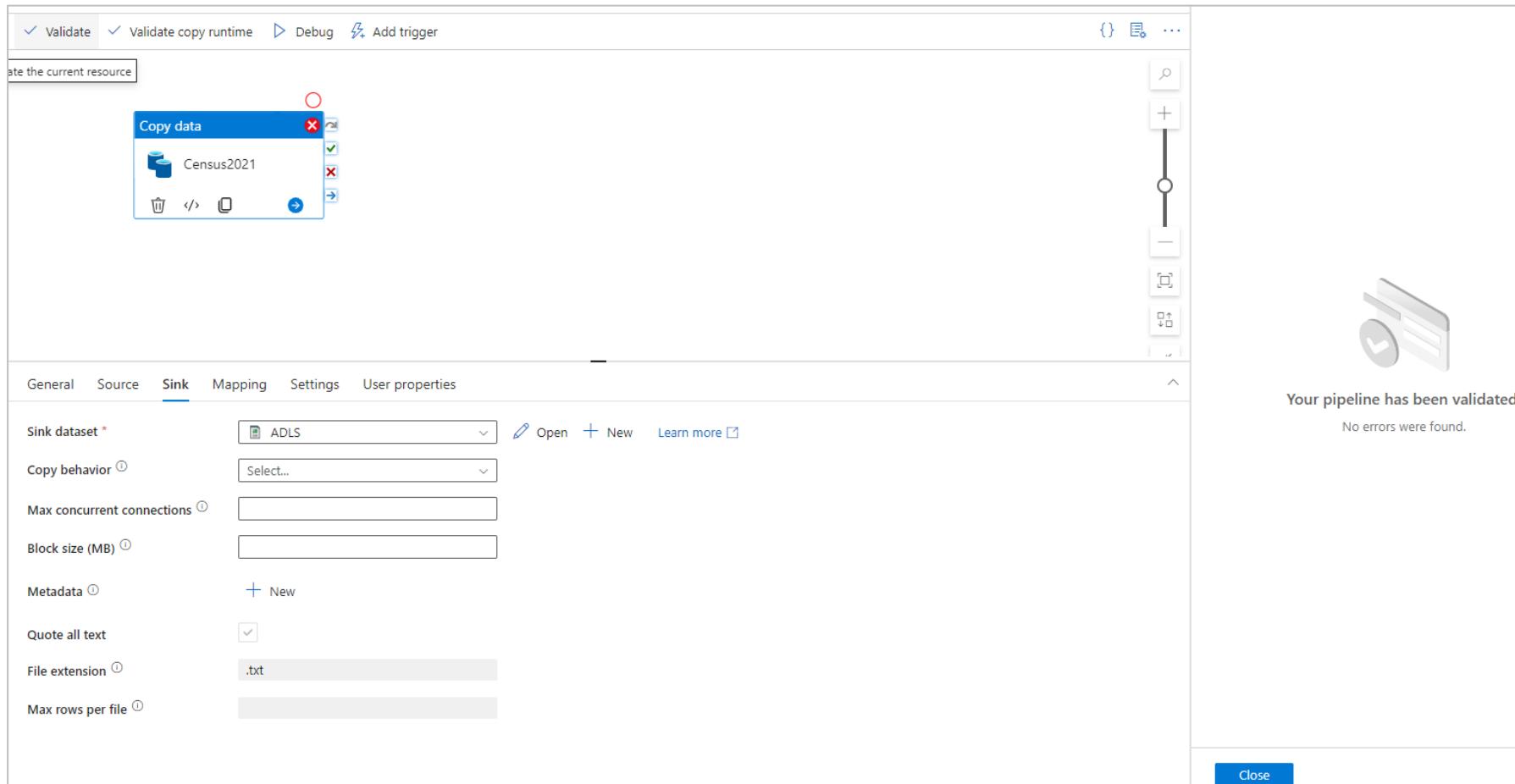
Showing 0 items

OK Cancel

- Give the file a name with extension
- ok



- First step configured: source and sink
- Click validate
- Click publish all



- Click debug to run the pipeline
- See below how pipeline is queued and running. Calling to the csv through Http
- After a while, the process succeeded
- Check if the .csv file landed into bronze directory

Validate Cancel options Add trigger

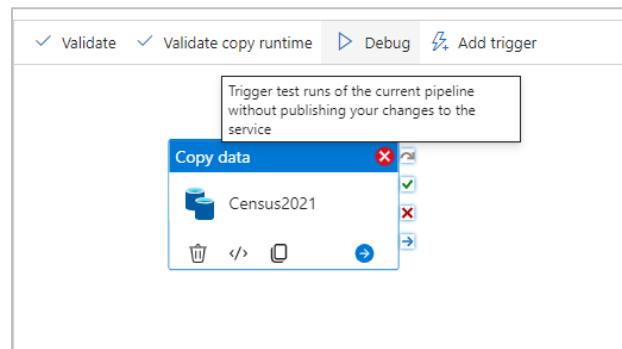
Copy data
Census2021

Parameters Variables Settings Output

Pipeline run ID: ff2777c7-1550-490a-9990-ac44084e80ab

Pipeline status In progress

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Census2021	Queued	Copy data	3/28/2024, 4:18:12 PM	Less than 1s			7563d350-54f1-4



Microsoft Azure | Data Factory > census-malaga-df Search factory and documentation

Validate ▶ Debug ⚡ Add trigger

Copy data
Census2021

Parameters Variables Settings Output

Pipeline run ID: ff2777c7-1550-490a-9990-ac44084e80ab

Pipeline status Succeeded

Activity name	Activity status	Activity type	Run start	Duration	Integration runtime	User properties	Activity run ID
Census2021	Succeeded	Copy data	3/28/2024, 4:18:12 PM	15s	AutoResolveIntegratio		7563d350-54f1-423c-9bf3-be3c74c7754f

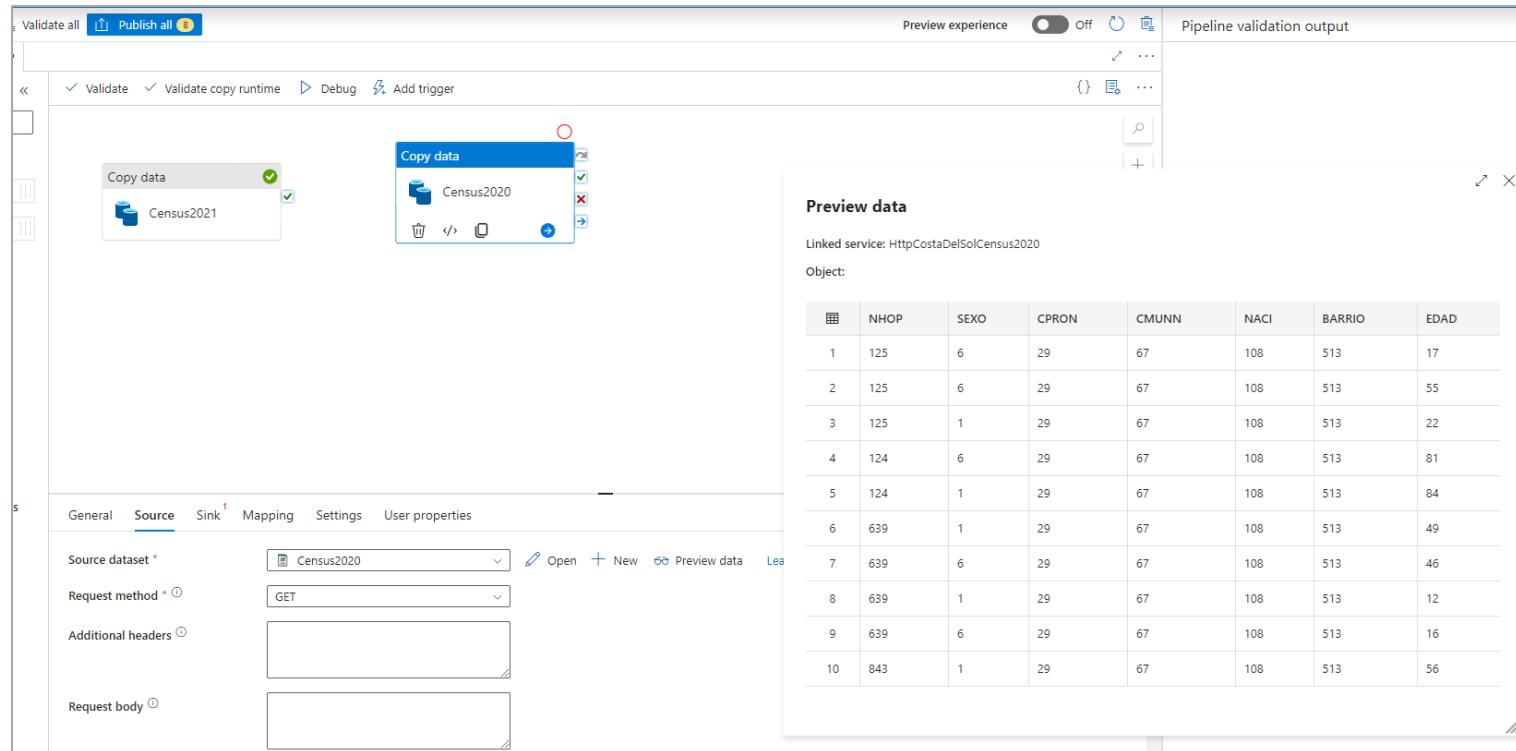
- There it is. First ingestion completed.
- Let's move forward and create more copy data and concatenate the step into a pipeline to extract and persist all the .csv files

The screenshot shows the Azure Storage Blob Container Overview page for the container 'census-malaga-data'. The left sidebar contains navigation links: Home, censusmalagadata_1711634436905 | Overview, censusmalagadata | Containers, and census-malaga-data (Container). The main content area has a breadcrumb path: Home > censusmalagadata_1711634436905 | Overview > censusmalagadata | Containers > census-malaga-data. The top navigation bar includes buttons for Search, Upload, Add Directory, Refresh, Rename, Delete, Change tier, Acquire lease, Break lease, and Give feedback. Below the navigation is a section for Authentication method: Access key (Switch to Microsoft Entra user account) and Location: census-malaga-data / bronze / 2021. A search bar for blobs by prefix (case-sensitive) and a 'Show deleted objects' toggle are also present. The main table lists the contents of the container:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]	3/28/2024, 4:18:26 PM	Hot (Inferred)		Block blob	15.53 MiB	Available
padronbarrios.csv						

The left sidebar also includes sections for Settings (Shared access tokens, Manage ACL, Access policy, Properties), Diagnosis (Diagnose and solve problems), and Access Control (IAM).

- You can preview the requested and persisted data clicking below on “Preview Data” (glasses)



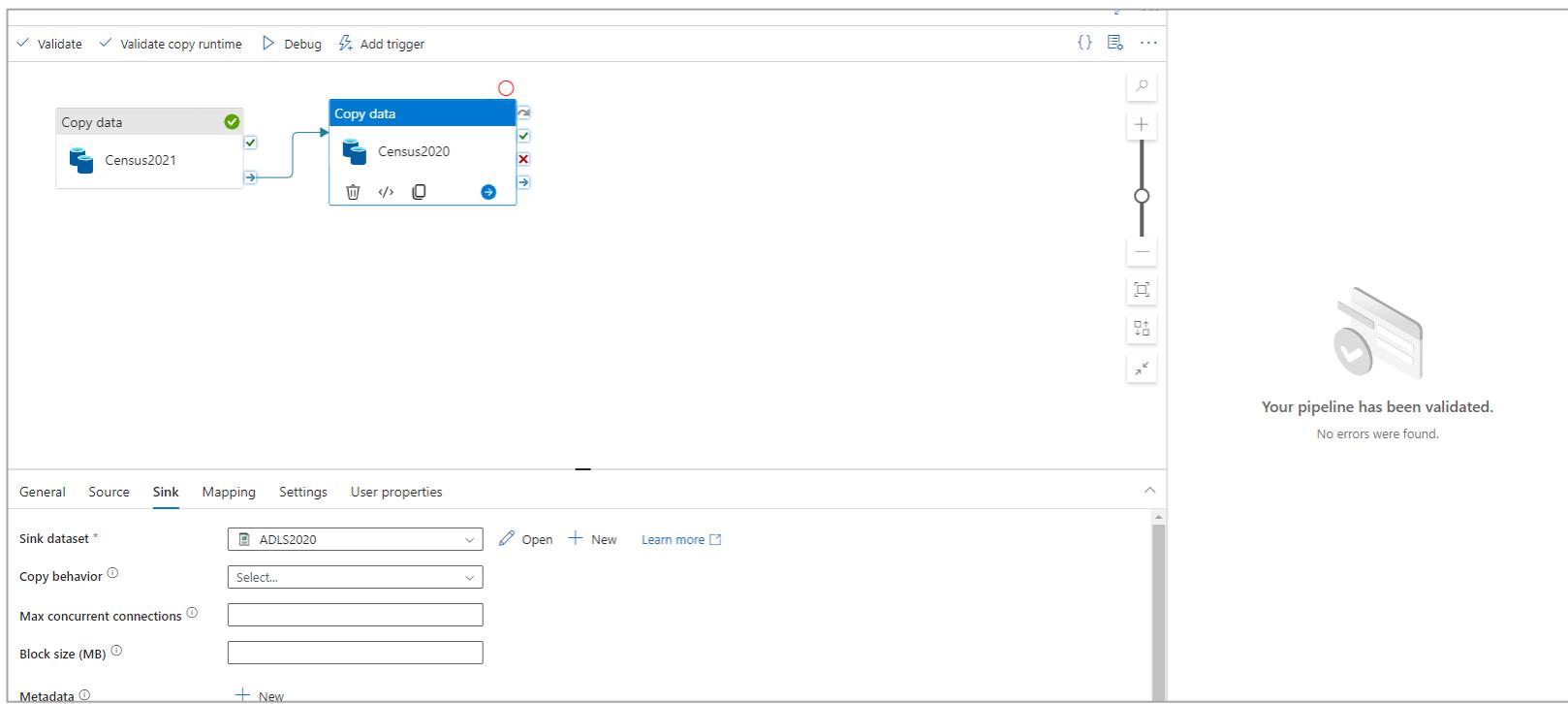
The screenshot shows the Azure Data Factory interface for configuring a copy activity. On the left, there's a main toolbar with 'Validate all' and 'Publish all'. Below it are buttons for 'Validate', 'Validate copy runtime', 'Debug', and 'Add trigger'. A 'Copy data' step is selected, showing a green checkmark and a 'Census2021' dataset icon.

The main configuration pane has tabs for 'General', 'Source' (which is selected), 'Sink', 'Mapping', 'Settings', and 'User properties'. Under 'Source', the 'Source dataset' is set to 'Census2020'. Other source settings include 'Request method' (set to 'GET') and an empty 'Additional headers' and 'Request body' section.

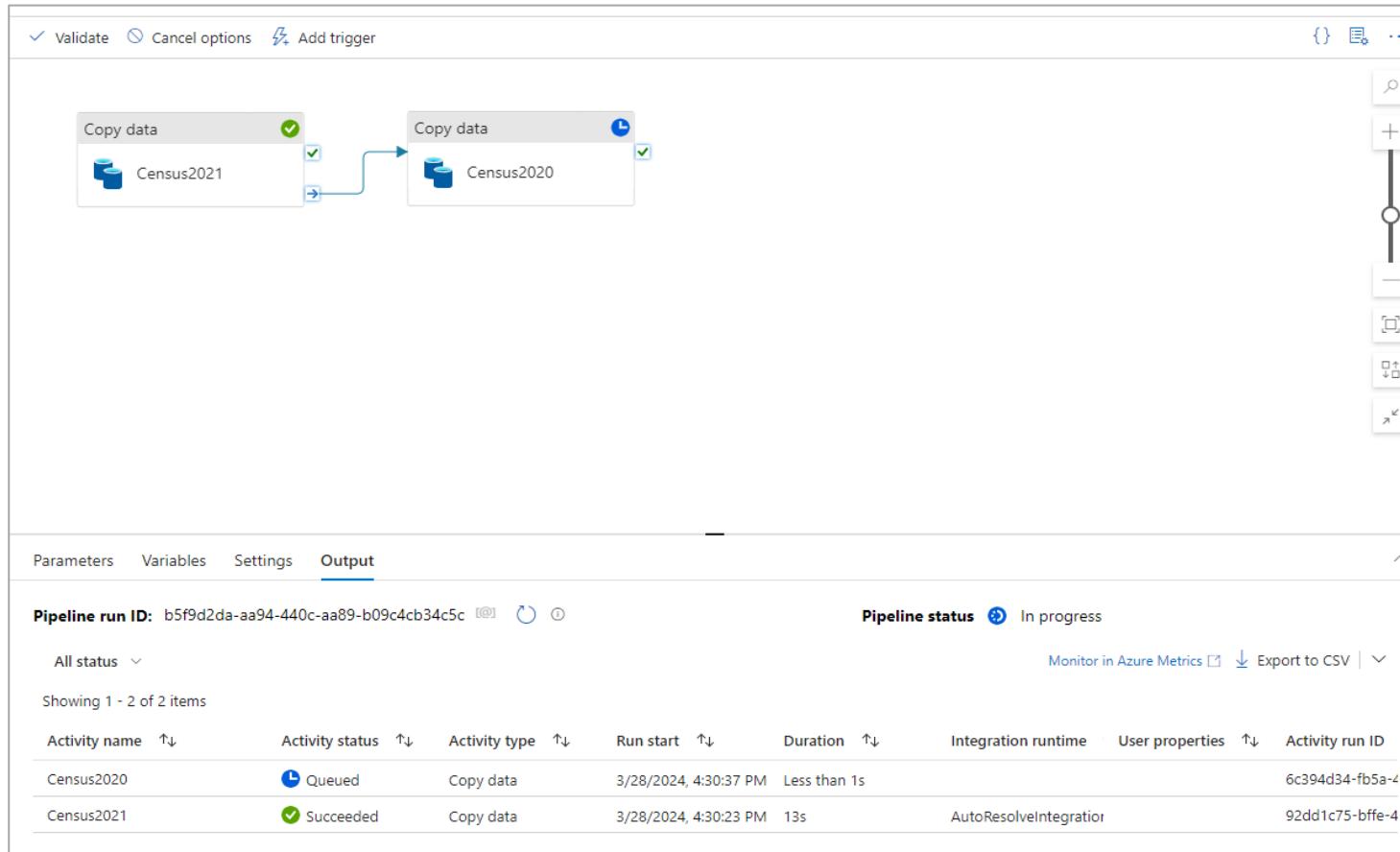
To the right, a 'Preview experience' toggle is set to 'off'. A 'Pipeline validation output' pane is visible above a 'Preview data' pane. The 'Preview data' pane shows a table titled 'Object:' with 10 rows of sample data from the Census2020 dataset. The columns are labeled: NHOP, SEXO, CPRON, CMUNN, NACI, BARRIO, and EDAD.

	NHOP	SEXO	CPRON	CMUNN	NACI	BARRIO	EDAD
1	125	6	29	67	108	513	17
2	125	6	29	67	108	513	55
3	125	1	29	67	108	513	22
4	124	6	29	67	108	513	81
5	124	1	29	67	108	513	84
6	639	1	29	67	108	513	49
7	639	6	29	67	108	513	46
8	639	1	29	67	108	513	12
9	639	6	29	67	108	513	16
10	843	1	29	67	108	513	56

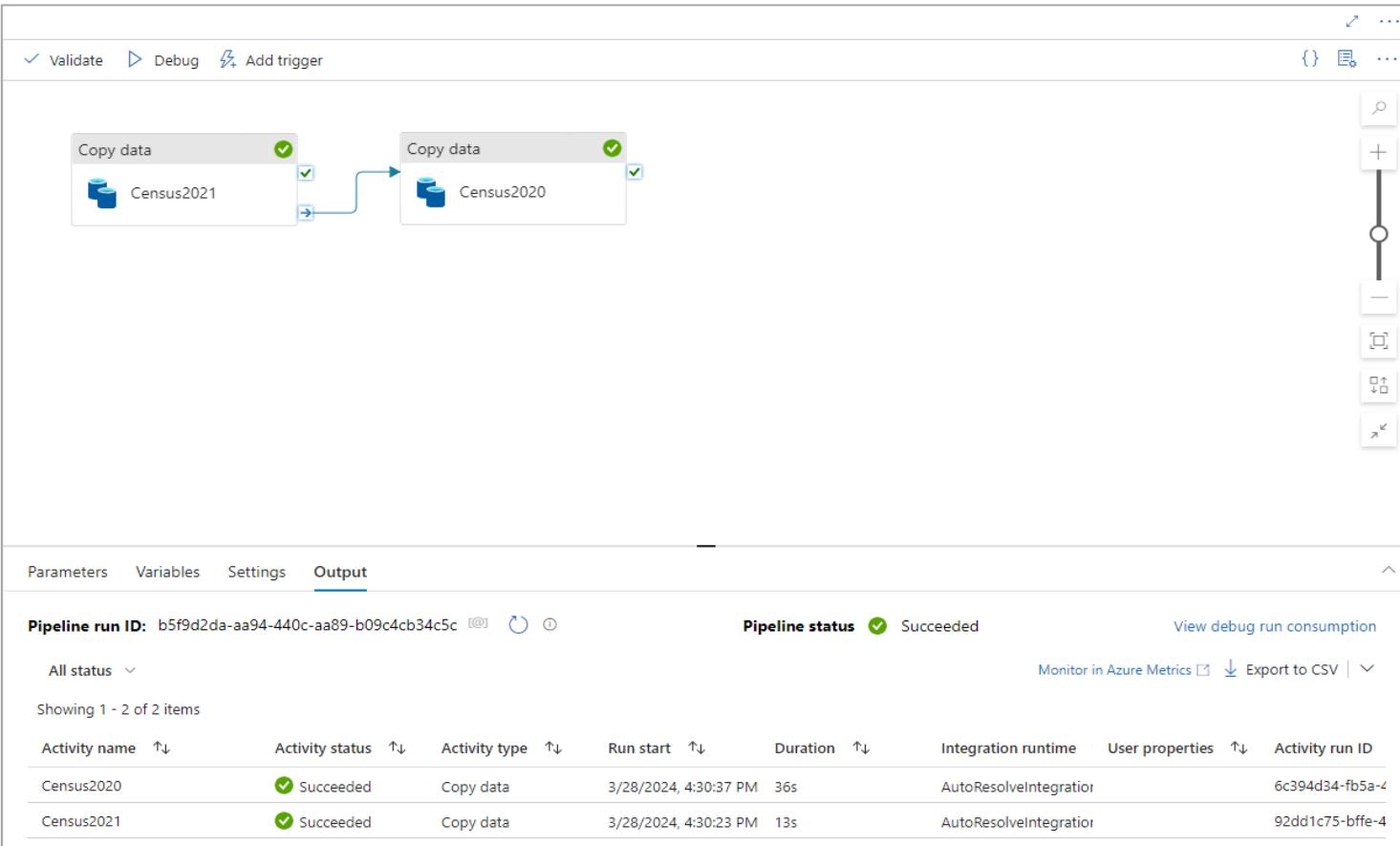
- Drag and drop the blue arrow from the first step to the second.
- This means that after first step completion, will continue with the next one
- Do the same with the other steps



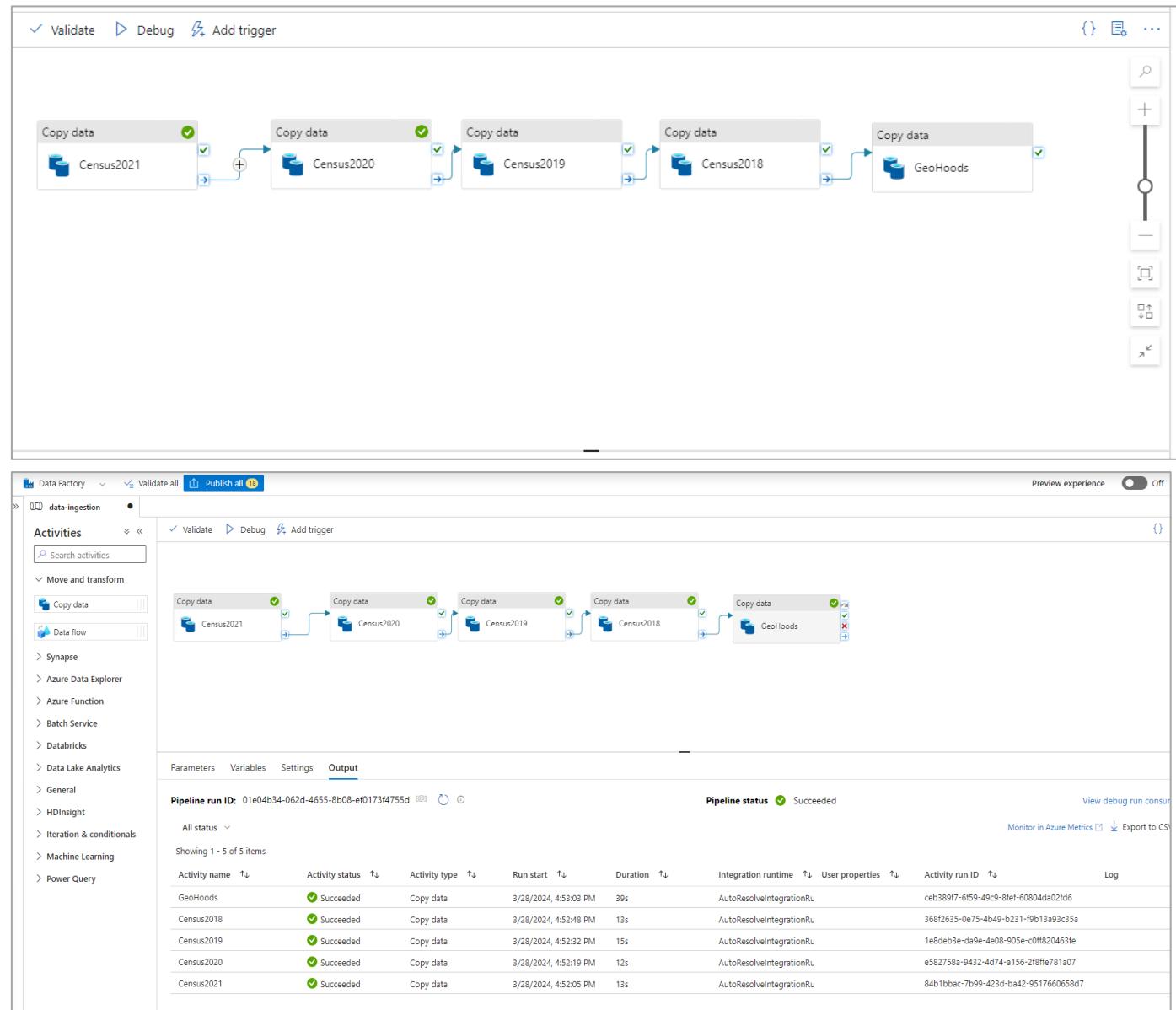
- Validate
- Debug
- See how the pipeline succeeds step by step
- See the .csv in ADLS container



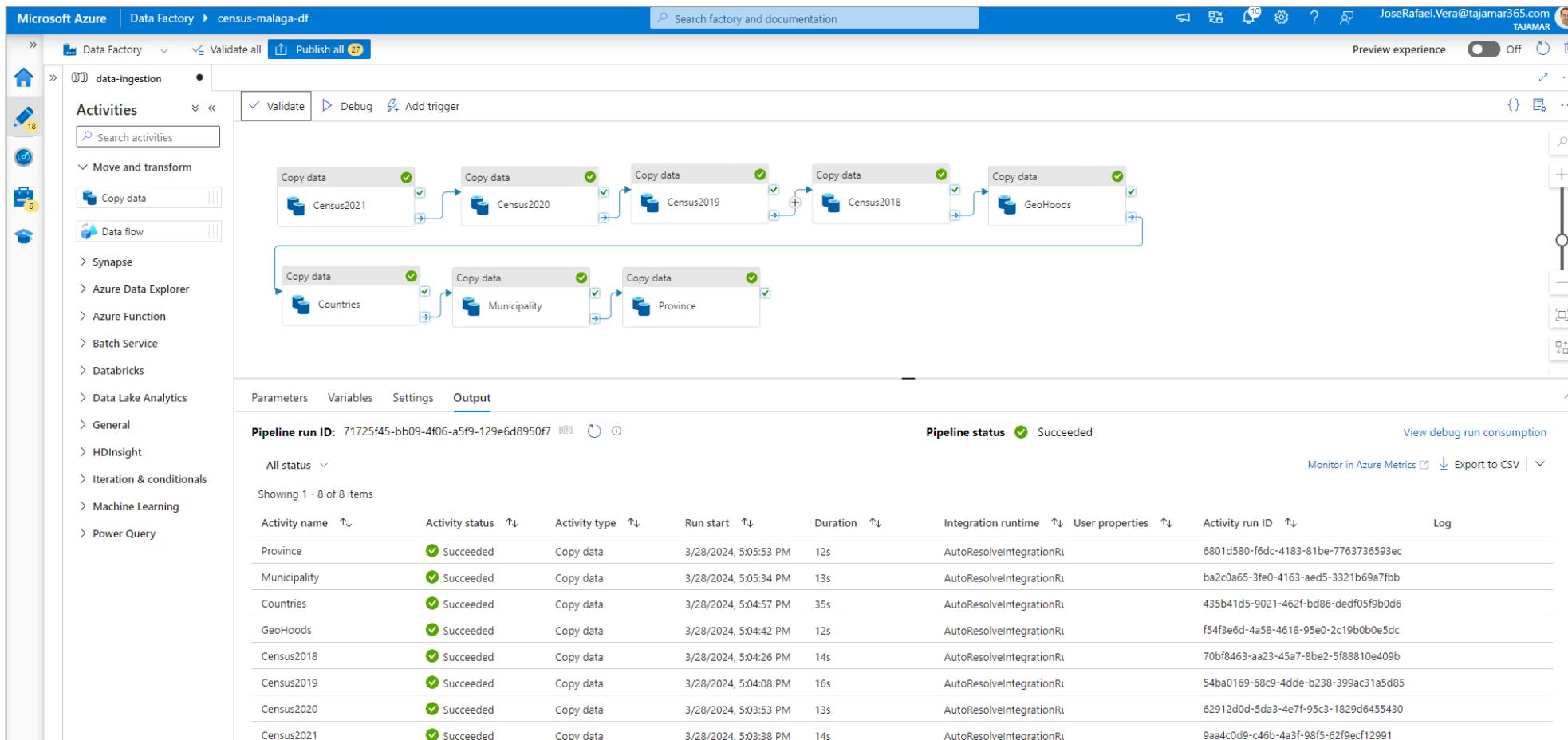
- The two-step pipeline succeeded



- Add so many steps as .csv files you need to get into your ADLS container in Bronze folder



- My complete succeeded pipeline



- And all the .csv files saved into Bronze Layer

census-malaga-data ...

Container

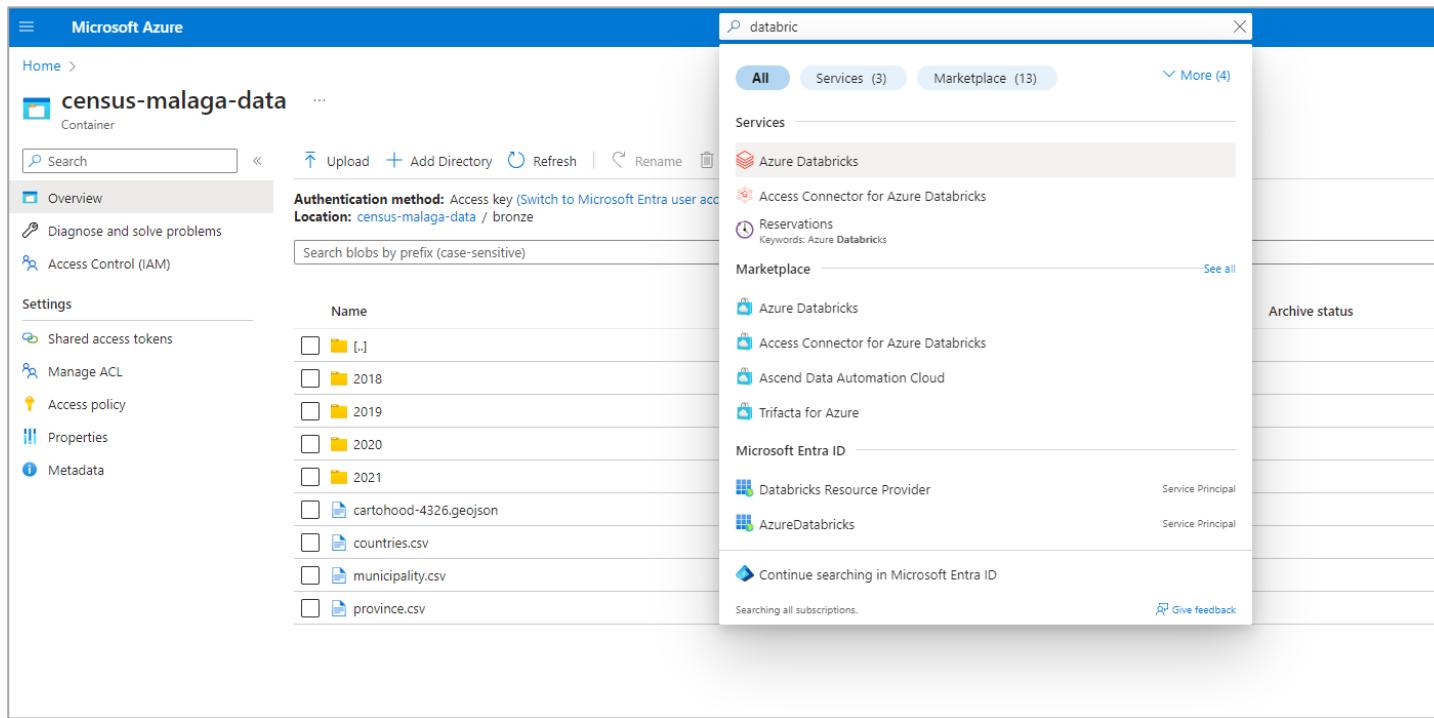
Search Upload Refresh | Rename Delete Change tier Acquire lease Break lease Give feedback

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: census-malaga-data / bronze

Search blobs by prefix (case-sensitive) Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[...]					-	
2018					-	
2019					-	
2020					-	
2021					-	
cartohood-4326.geojson	3/28/2024, 5:04:52 PM	Hot (Inferred)		Block blob	1.71 MiB	Available
countries.csv	3/28/2024, 5:05:08 PM	Hot (Inferred)		Block blob	7.59 KiB	Available
municipality.csv	3/28/2024, 5:05:45 PM	Hot (Inferred)		Block blob	389.7 KiB	Available
province.csv	3/28/2024, 5:06:03 PM	Hot (Inferred)		Block blob	1.36 KiB	Available

- Search for Databricks



- Create a new Databricks service

Home >

Azure Databricks

Tajamar (tajamar365.com)

+ Create Manage view Refresh Export to CSV Open query Assign tags

Filter for any field... Subscription equals all Resource group equals all Location equals all Add filter

Showing 0 to 0 of 0 records.

Name ↑↓	Type ↑↓	Resource group ↑↓	Location ↑↓
---------	---------	-------------------	-------------

No azure databricks services to display

Unlock insights from all your data and build artificial intelligence (AI) solutions with Azure Databricks, set up your Apache Spark environment in minutes, autoscale, and collaborate on shared projects in an interactive workspace.

Create azure databricks service

Learn more ↗



- Select your free subscription
- Select your preexisting Resource group
- Give Databrick service a name
- Select a Region. SouthEast Asia recommended
- Pricing Tier: “premium” > you won’t be charged. You are under your subscription and this Tier ables full-service access.
- next

Microsoft Azure

Create an Azure Databricks workspace

Basics Networking Encryption Security & compliance 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 * Azure for Students

Resource group * census-malaga-4years

Instance Details

Workspace name *

Region *

Pricing Tier * Premium (+ Role-based access controls)

We selected the recommended pricing tier for your workspace. You can change the tier based on your needs.

Managed Resource Group name

Review + create < Previous Next : Networking >

Microsoft Azure

Create an Azure Databricks workspace

Basics Networking Encryption Security & compliance Tags Review + create

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP) Yes No

Deploy Azure Databricks workspace in your own Virtual Network (VNet) Yes No

Review + create < Previous Next : Encryption >

- Next
- Next

Home > Azure Databricks >
Create an Azure Databricks workspace ...

Basics Networking **Encryption** Security & compliance Tags Review + create

Data Encryption

For additional control of your data, you can add your own key to protect and control access to some types of data. Enabling customer-managed key encryption for Managed Services or Managed Disks is an irreversible action. The key, key vault, and key version may be updated but the features cannot be disabled after being enabled.

Managed Disks

Use your own key

Managed Services

Use your own key

Double encryption for DBFS root

In addition to your choice of the default encryption or your own managed key encryption, Azure Databricks DBFS root can also be encrypted with a second layer of encryption called infrastructure encryption using platform-managed key to achieve Double Encryption for DBFS root.

Enable Infrastructure Encryption
⚠This feature cannot be changed after this workspace is created.

Review + create < Previous Next : Security & compliance >

Home > Azure Databricks >
Create an Azure Databricks workspace ...

Basics Networking **Encryption** **Security & compliance** Tags Review + create

Enhanced Security & Compliance

Enhanced Security and Compliance Add-On helps simplify the complexity of meeting security and regulatory requirements.

Enable compliance security profile
⚠This feature cannot be disabled once it is enabled.

Enable enhanced security monitoring

Enable automatic cluster update

Review + create < Previous Next : Tags >

- Next
- Create

Microsoft Azure

Home > Azure Databricks >

Create an Azure Databricks workspace

Basics Networking Encryption Security & compliance Tags Review + create

Name	Value	Resource
<input type="text" value="census-malaga-dbricks"/>	:	Azure Databricks Service

Review + create < Previous

Microsoft Azure

Home > Azure Databricks >

Create an Azure Databricks workspace

Validation Succeeded

Basics Networking Encryption Security & compliance Tags Review + create

Summary

Basics

Workspace name	census-malaga-dbricks
Subscription	Azure for Students
Resource group	census-malaga-4years
Region	West Europe
Pricing Tier	premium
Managed Resource Group name	

Networking

Deploy Azure Databricks workspace with Secure Cluster Connectivity (No Public IP)	No
Deploy Azure Databricks workspace in your own Virtual Network (VNet)	No

Encryption

Enable Infrastructure Encryption	No
Enable CMK for Managed Disks	No
Enable CMK for Managed Services	No

Security & compliance

Compliance Security Profile	No
Compliance Standards	

Create < Previous Download a template for automation

- Wait until deployed

The screenshot shows the Microsoft Azure Deployment Overview page for a deployment named "census-malaga-4years_census-malaga-dbricks". The deployment status is "Deployment is in progress". The deployment details table shows no resources listed. A tooltip indicates "Deployment to resource group is in progress". The page includes a search bar, navigation links, and various Azure services and tutorials.

Microsoft Azure

census-malaga-4years_census-malaga-dbricks | Overview

Deployment

Search

Delete Cancel Redeploy Download Refresh

Overview

Deployment is in progress

Deployment name : census-malaga-4years_census-malaga-dbricks
Subscription : Azure for Students
Resource group : census-malaga-4years

Start time : 3/28/2024, 5:19:08 PM
Correlation ID : af7eec4e-80b4-4372-9633-902e7871fceb

Deployment details

Resource	Type	Status	Operation details
There are no resources to display.			

Give feedback

Tell us about your experience with deployment

Deployment in progress...
Deployment to resource group is in progress.

Microsoft Defender for
Secure your apps and infrastructure
Go to Microsoft Defender

Free Microsoft tutorials
Start learning today >

Work with an expert
Azure experts are service professionals who can help manage your Azure environment and be your first line of defense
Find an Azure expert >

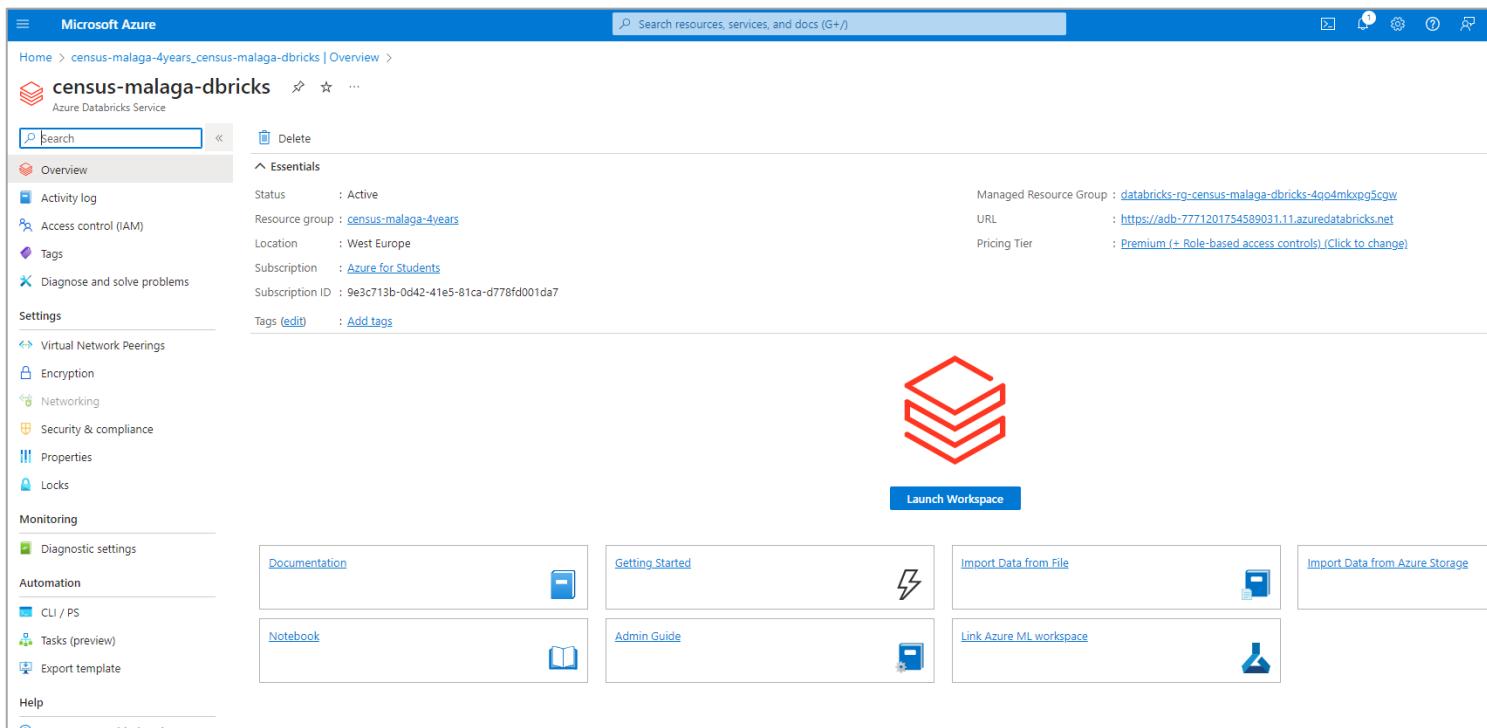
- Deployed
- Go to resource

The screenshot shows the Microsoft Azure Deployment Overview page for a deployment named "census-malaga-4years_census-malaga-dbricks". The deployment is marked as complete with a green checkmark icon. Key details listed include:

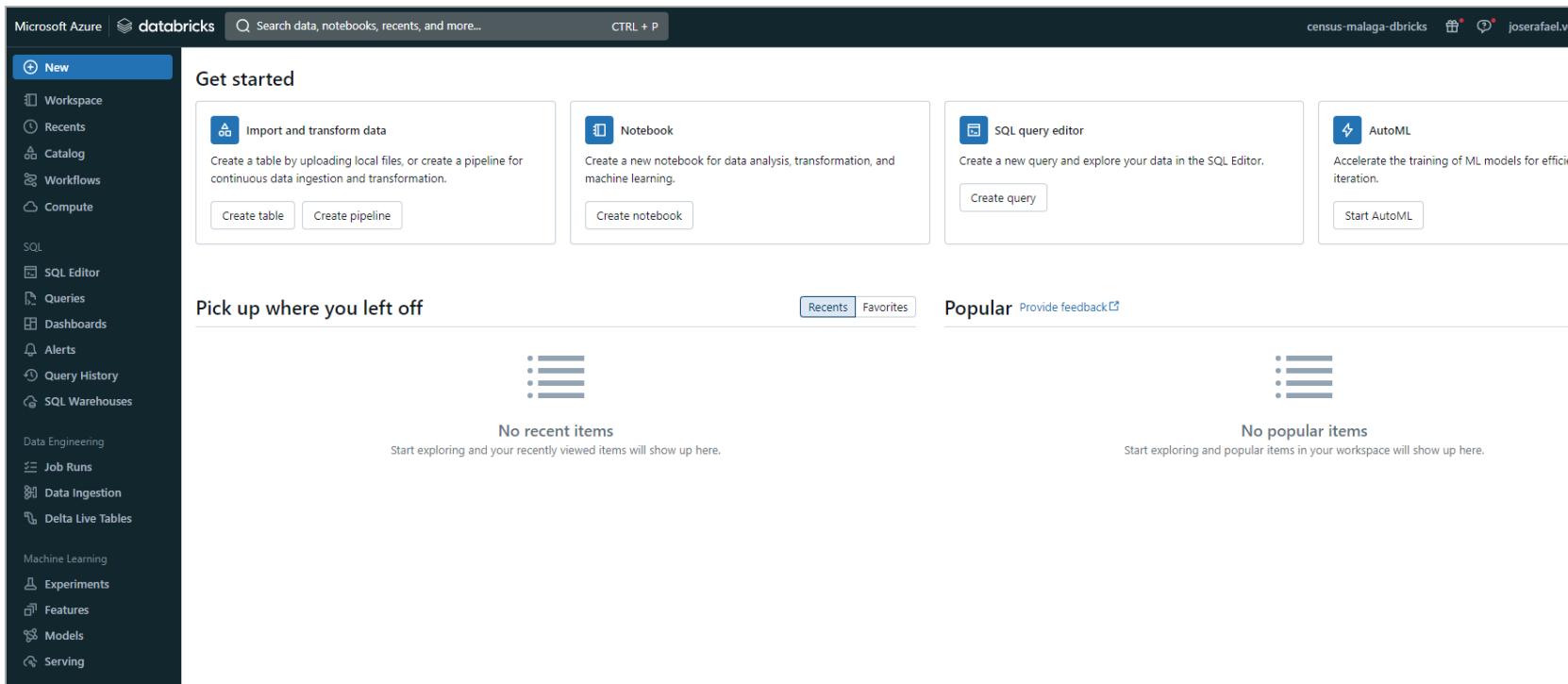
- Deployment name: census-malaga-4years_census-malaga-dbricks
- Subscription: Azure for Students
- Resource group: census-malaga-4years
- Start time: 3/28/2024, 5:19:08 PM
- Correlation ID: af7eec4e-80b4-4372-9633-902e7871fceb

Navigation links on the left include Home, Overview (selected), Inputs, Outputs, and Template. A "Go to resource" button is located at the bottom of the main content area.

- Launch workspace



- You are in Databrick workspace already
- Click on “compute” to create the Spark Cluster



- Create compute

The screenshot shows the Databricks Compute interface. At the top, there's a navigation bar with the Databricks logo, a search bar, and user information. Below the header, the title "Compute" is displayed, followed by a tab menu: "All-purpose compute" (which is selected), "Job compute", "SQL warehouses", "Pools", and "Policies". A search bar and filter options ("Created by", "Only pinned") are located above the main table. To the right, there are buttons for "Create with Personal Compute" and "Create compute". The main area features a table with columns: State, Name, Policy, Runtime, Active memory, Active cores, Active DBU / h, Source, Creator, Notebooks, and a settings gear icon. A large plus sign (+) is centered in the middle of the table area, indicating that no compute resources have been created yet. Below the plus sign, the text "No compute" is displayed, along with a sub-instruction: "Create compute to run workloads from your notebooks and jobs. Learn more about best practices for compute configuration". A prominent "Create compute" button is positioned at the bottom of this section.

- In my case, for this demo I used a single node but you can choose more nodes if needed
- The Runtime version I left it by default
- Node type the cheapest
- Region: SoutheEast Asia, otherwise you will not have many node type available (VM types indeed).
- It takes some time until the cluster is up and running. Be patient

The screenshot shows the Databricks Compute interface for configuring a cluster named "Jose Maranon's Cluster".

Configuration Tab:

- Policy:** Unrestricted
- Access mode:** Single user access (Single user, Jose Maranon)

Performance Tab:

- Databricks Runtime Version: 13.3 LTS (includes Apache Spark 3.4.1, Scala 2.12)
- Use Photon Acceleration
- Node type:** Standard_DS3_v2 (14 GB Memory, 4 Cores)
- Terminate after 120 minutes of inactivity

Tags Tab:

- No custom tags
- > Automatically added tags
- > Advanced options

Summary Panel:

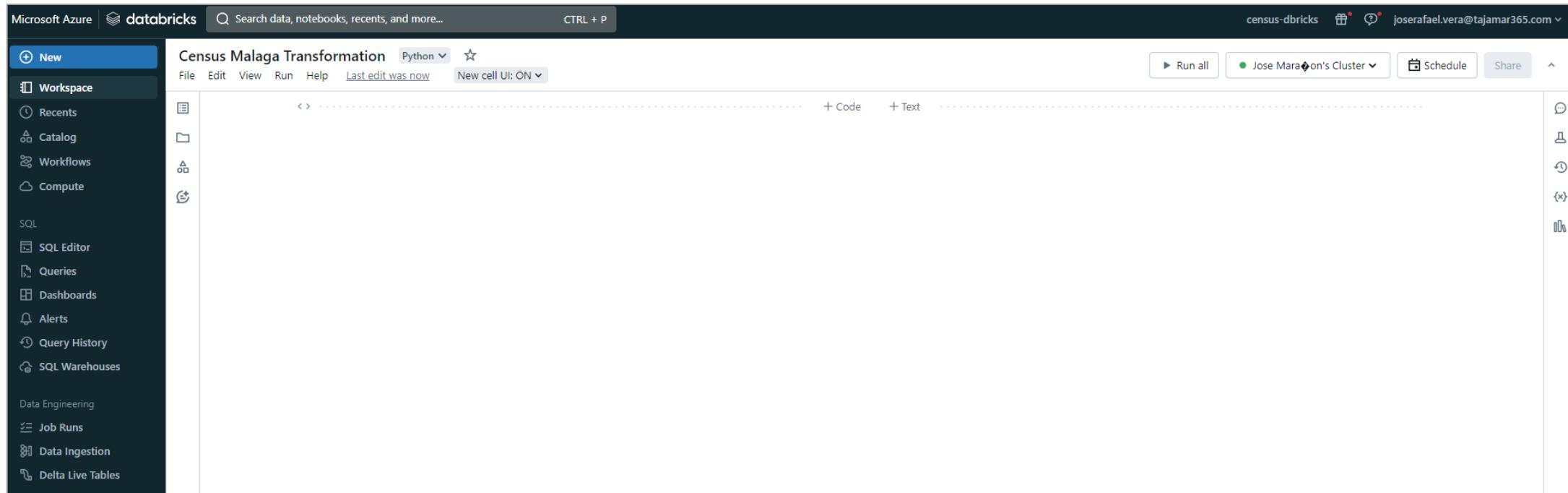
Summary	
1 Driver	14 GB Memory, 4 Cores
Runtime	13.3.x-scala2.12
Standard_DS3_v2	0.75 DBU/h

UI | JSON

- Create a new Notebook



- The notebook ready to be written
- In order to get access from Databricks to our data in ADLS (Storage Account) it is necessary to link Databricks to an APP Registration which offer the needed credentials using OAuth2.0 protocol for Azure AD and thus grant authn and authz for Databricks to the Storage Account



- Search for App registrations

The screenshot shows the Microsoft Azure Storage Explorer interface. On the left, there's a sidebar for a container named "census-malaga-data". The main area displays a search bar with the query "app reg". Below the search bar, there are tabs for "All", "Services (39)", and "Marketplace (30)". The "Services" tab is selected, showing a list of services including "App registrations", "App Services", "App Service Certificates", "App Service Domains", "Function App", "Web App", "App Service Plan", and "WordPress on App Service". A "See all" link is visible next to each category. On the right side, there's a table header for "Archive status", "Blob type", "Size", and "Lease st". The table body is currently empty.

- Create New Registration

The screenshot shows the Microsoft Azure portal interface for managing app registrations. The top navigation bar is blue with the Microsoft Azure logo and a search bar. Below it, the breadcrumb navigation shows 'Home > App registrations'. The main content area has a header with several buttons: '+ New registration' (highlighted in grey), 'Endpoints', 'Troubleshooting', 'Refresh', 'Download', 'Preview features', and 'Got feedback?'. A sub-header 'New registration' is centered above a message box. The message box contains a blue information icon and text: 'Starting June 30th, 2020 we will no longer add any new features to Azure Active Directory Authentication Library (ADAL) and Azure Active Directory Graph. We will continue to provide technical support and security updates but we will no longer provide Library (MSAL) and Microsoft Graph.' followed by a link 'Learn more'. Below the message box are three tabs: 'All applications', 'Owned applications' (underlined in blue), and 'Deleted applications'. To the right of these tabs is a search bar with placeholder text 'Start typing a display name or application (client) ID to filter these r...' and a 'Add filters' button. At the bottom of the page, a message states 'This account isn't listed as an owner of any applications in this directory.' and a blue button 'View all applications in the directory'.

- Give the app registration a name of your choice
- Access to this API: first option
- Register

Microsoft Azure

Home > App registrations > Register an application

Name
The user-facing display name for this application (this can be changed later).

app01

Supported account types

Who can use this application or access this API?

Accounts in this organizational directory only (Tajamar only - Single tenant)

Accounts in any organizational directory (Any Microsoft Entra ID tenant - Multitenant)

Accounts in any organizational directory (Any Microsoft Entra ID tenant - Multitenant) and personal Microsoft accounts (e.g. Skype, Xbox)

Personal Microsoft accounts only

[Help me choose...](#)

Redirect URI (optional)
We'll return the authentication response to this URI after successfully authenticating the user. Providing this now is optional and it can be changed later, but a value is required for most authentication scenarios.

Select a platform e.g. https://example.com/auth

Register an app you're working on here. Integrate gallery apps and other apps from outside your organization by adding from [Enterprise applications](#).

By proceeding, you agree to the Microsoft Platform Policies [\[link\]](#)

Register

- App registered successfully
- Click on “ Certificates & secrets ”

Microsoft Azure

Home > App registrations >

app01

Search Delete Endpoints Preview features

Overview Quickstart Integration assistant

Manage Branding & properties Authentication Certificates & secrets Token configuration API permissions Expose an API App roles Owners Roles and administrators Manifest

Support + Troubleshooting Troubleshooting New support request

Essentials

Display name : app01	Client credentials : Add a certificate or secret
Application (client) ID : 5f7d8bc7-1df3-4357-a750-73957af108be	Redirect URIs : Add a Redirect URI
Object ID : c3022c17-3734-4def-a6d1-506b039527a1	Application ID URI : Add an Application ID URI
Directory (tenant) ID : 68519e48-83f3-435f-a38a-1a7aa77ba987	Managed application in ... : app01
Supported account types : My organization only	

Welcome to the new and improved App registrations. Looking to learn how it's changed from App registrations (Legacy)? [Learn more](#)

Starting June 30th, 2020 we will no longer add any new features to Azure Active Directory Authentication Library (ADAL) and Azure Active Directory Graph. We will continue to provide technical support and security updates but we will no longer provide feature updates will need to be upgraded to Microsoft Authentication Library (MSAL) and Microsoft Graph. [Learn more](#)

Get Started Documentation

Build your application with the Microsoft identity platform

The Microsoft identity platform is an authentication service, open-source libraries, and application management tools. You can create modern, standards-based authentication solutions, access and protect APIs, and add sign-in for your users and customers. [Learn more](#)

- Navigate to “client secrets”
- Click on “New client secret”

The screenshot shows the Microsoft Azure portal interface for managing application registrations. The top navigation bar includes the Microsoft Azure logo, a search bar, and a 'Search resources, services, and docs (G+ /)' field. Below the navigation is a breadcrumb trail: Home > App registrations > app01. The main title is 'app01 | Certificates & secrets' with a key icon.

The left sidebar, titled 'Manage', lists several options: Overview, Quickstart, Integration assistant, Branding & properties, Authentication, Certificates & secrets (which is selected and highlighted in grey), Token configuration, API permissions, Expose an API, App roles, Owners, and Data and administrators.

The right pane starts with a note: 'Credentials enable confidential applications to identify themselves to the authentication service when receiving tokens at a web addressable location (using an HTTPS scheme). For a higher level of assurance, we recommend using a certificate (instead of a client secret) as a credential.' Below this is a callout box stating: 'Application registration certificates, secrets and federated credentials can be found in the tabs below.'

Three tabs are present: 'Certificates (0)', 'Client secrets (0)', and 'Federated credentials (0)'. The 'Client secrets (0)' tab is currently selected. A sub-note under this tab explains: 'A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password.' A 'New client secret' button is available for creating a new secret.

A table header is shown for managing client secrets, with columns: Description, Expires, Value ⓘ, and Secret ID. The table body states: 'No client secrets have been created for this application.'

- Give the secretkey a name of your choice
- Expiration by default
- Add

Home > App registrations > app01

app01 | Certificates & secrets

Search Got feedback?

Credentials enable confidential applications to identify themselves to the authentication service when receiving tokens at a web addressable location (using an HTTPS scheme). For a higher level of assurance, we recommend using a certificate (instead of a client secret) as a credential.

Manage

- Branding & properties
- Authentication
- Certificates & secrets** (selected)
- Token configuration
- API permissions
- Expose an API
- App roles
- Owners
- Roles and administrators
- Manifest

Support + Troubleshooting

- Troubleshooting
- New support request

Application registration certificates, secrets and federated credentials can be found in the tabs below.

Certificates (0) Client secrets (0) Federated credentials (0)

A secret string that the application uses to prove its identity when requesting a token. Also can be referred to as application password.

+ New client secret

Description	Expires	Value ⓘ	Secret ID
No client secrets have been created for this application.			

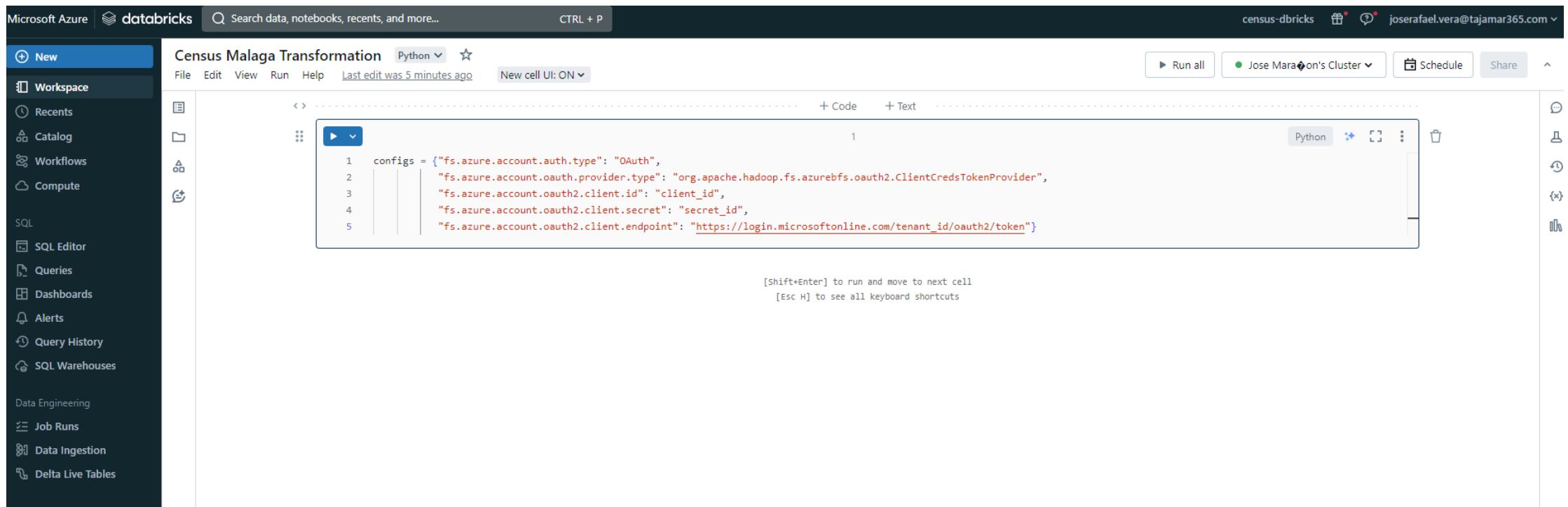
Add a client secret

Description: secretkey

Expires: Recommended: 180 days (6 months)

Add Cancel

- The secrets created in the step before is copy pasted here
- **WARNING!** Bad practice. Never expose your credentials within your code. Handle your secrets properly with “Key Vault”



The screenshot shows a Databricks workspace interface. On the left, there's a sidebar with various navigation options like 'New', 'Workspace', 'Recents', 'Catalog', 'Workflows', 'Compute', 'SQL', 'SQL Editor', 'Queries', 'Dashboards', 'Alerts', 'Query History', 'SQL Warehouses', 'Data Engineering', 'Job Runs', 'Data Ingestion', and 'Delta Live Tables'. The main area is titled 'Census Malaga Transformation' and is set to 'Python'. A single code cell contains the following Python code:

```
1 configs = {"fs.azure.account.auth.type": "OAuth",
2             "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
3             "fs.azure.account.oauth2.client.id": "client_id",
4             "fs.azure.account.oauth2.client.secret": "secret_id",
5             "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/tenant_id/oauth2/token"}
```

Below the code cell, there are instructions: '[Shift+Enter] to run and move to next cell' and '[Esc H] to see all keyboard shortcuts'. At the top right, there are buttons for 'Run all', 'Jose Maranon's Cluster', 'Schedule', and 'Share'. The top bar also includes a search bar, a 'CTRL + P' keybinding, and user information.

- Search for “Key Vault”
- Create key vault
- I didn’t implement this step as the Secrets used in this project were deleted after completion. All the services as well.
- **Keep this Key vault in mind for Production environment**

Microsoft Azure

Search resources, services, and docs (G+/)

Home >

Key vaults

Tajamar (tajamar365.com)

+ Create Manage deleted vaults Manage view Refresh Export to CSV Open query Assign tags

Filter for any field... Subscription equals all Resource group equals all Location equals all Add filter

Showing 0 to 0 of 0 records. No grouping

Name ↑↓	Type ↑↓	Resource group ↑↓	Location ↑↓	Subscription ↑↓
---------	---------	-------------------	-------------	-----------------

No key vaults to display

Safeguard cryptographic keys and other secrets used by cloud apps and services.

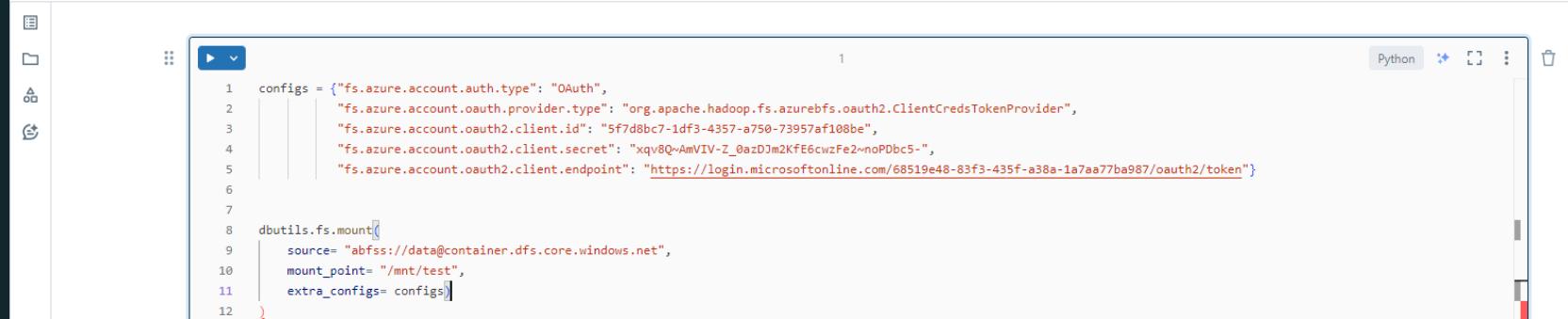
Create key vault Learn more ↗



- Paste client_id and client_secret and tenant_id. **REMINDER** this is a bad practice. Just use Key Vault
- Create a mount point in databricks from the Storage account container. See next step
- In dbutils.fs.mount:
 - Source = “abfss://[Container name]@[storage Account name].dfs.core.windows.net”

Census Malaga Transformation Python 

File Edit View Run Help Last edit was 2 minutes ago New cell UI: ON   



```

1 configs = {"fs.azure.account.auth.type": "OAuth",
2             "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
3             "fs.azure.account.oauth2.client.id": "5f7d8bc7-1df3-4357-a750-73957af108be",
4             "fs.azure.account.oauth2.client.secret": "xqv8QwAmVIV-Z_0zDm2KfE6cwzfe2noPDbc5-",
5             "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/68519e48-83f3-435f-a38a-1a7aa77ba987/oauth2/token"}
6
7
8 dbutils.fs.mount(
9   source= "abfss://data@container.dfs.core.windows.net",
10  mount_point= "/mnt/test",
11  extra_configs= config)
12

```

[Shift+Enter] to run and move to next cell
[Esc H] to see all keyboard shortcuts

- Go to your Storage account and copy the storage account Name and paste it in Databricks
- Go to containers and copy the container name and paste it in Databricks

The image consists of two vertically stacked screenshots from the Microsoft Azure portal.

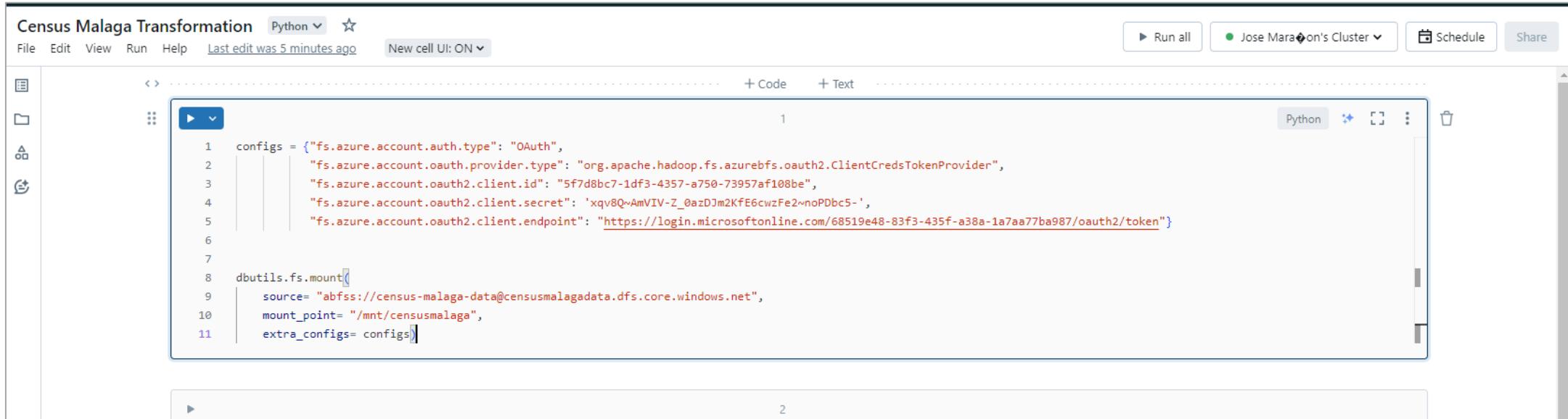
Screenshot 1: Storage account Overview

This screenshot shows the 'Storage accounts' blade for the 'censusmalagadata' account. The left sidebar lists options like 'Create', 'Restore', and 'Data storage'. The main area displays the account's details, including its name, resource group ('census-malaga-4years'), location ('westeurope'), and subscription ('Azure for Students'). It also shows disk state, provisioning state ('Succeeded'), and creation date ('28/3/2024, 15:02:14'). The 'Properties' tab is selected, showing 'Data Lake Storage' settings (Hierarchical namespace: Enabled, Default access tier: Hot, Blob anonymous access: Disabled) and 'Security' settings (Require secure transfer for REST API operations: Enabled, Storage account key access: Enabled, Minimum TLS version: Version 1.2).

Screenshot 2: Container List

This screenshot shows the 'Containers' blade for the same storage account. The left sidebar is identical. The main area lists the existing containers: 'Slogs' and 'census-malaga-data'. Both containers were created on 3/28/2024 at different times (3:02:38 PM and 3:06:09 PM respectively) and are set to 'Private' with 'Available' lease states.

- There you go! You have all the needed information to:
 - Connect Databricks access to the Storage account
 - Mount a directory into Databricks based on the existing container
- Give your mount point a name
- Before run it, you need to set access role permission to the Storage Account. See next step



The screenshot shows a Databricks notebook interface with the following details:

- Title:** Census Malaga Transformation
- Languages:** Python
- Last edit:** 5 minutes ago
- Cluster:** Jose Maranon's Cluster
- UI Mode:** ON
- Code Cell:**

```

1 configs = {"fs.azure.account.auth.type": "OAuth",
2             "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
3             "fs.azure.account.oauth2.client.id": "5f7d8bc7-1df3-4357-a750-73957af108be",
4             "fs.azure.account.oauth2.client.secret": 'xqv8Q~AmVIV-Z_0azDJm2KfE6cwzFe2~noPDbc5-',
5             "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/68519e48-83f3-435f-a38a-1a7aa77ba987/oauth2/token"}
6
7
8 dbutils.fs.mount(
9   source= "abfss://census-malaga-data@censusmalagadata.dfs.core.windows.net",
10  mount_point= "/mnt/censusmalaga",
11  extra_configs= configs)

```

- Go to Storage account and click on “Access Control IAM”

The screenshot shows the Microsoft Azure Storage accounts interface. The top navigation bar includes the Microsoft Azure logo, a search bar, and user profile information. Below the navigation bar, the breadcrumb path indicates the current location: Home > Storage accounts > censusmalagadata. The main title is "Storage accounts" with a sub-title "censusmalagadata | Containers". On the left, there is a sidebar with options like "Create", "Restore", and a "Filter for any field..." input. The main content area displays a list of containers under the heading "Containers". The table has columns for "Name", "Last modified", "Anonymous access level", and "Lease state". Two containers are listed: "\$logs" (last modified 3/28/2024, 3:02:38 PM, Private, Available) and "census-malaga-data" (last modified 3/28/2024, 3:06:09 PM, Private, Available). A "Search containers by prefix" input field and a "Show deleted containers" toggle switch are also present.

Name	Last modified	Anonymous access level	Lease state
\$logs	3/28/2024, 3:02:38 PM	Private	Available
census-malaga-data	3/28/2024, 3:06:09 PM	Private	Available

- In IAM click on Add role assignment
- Navigate to Members

Microsoft Azure Search resources, services, and docs (G+) JoseRafael.V
TAJAMAR

Home > Storage accounts > censusmalagadata | Access Control (IAM) >

Add role assignment

[Role](#) [Members *](#) [Conditions](#) [Review + assign](#)

A role definition is a collection of permissions. You can use the built-in roles or you can create your own custom roles. [Learn more](#)

[Job function roles](#) [Privileged administrator roles](#)

Grant access to Azure resources based on job function, such as the ability to create virtual machines.

blob con		Type : All	Category : All	Type ↑	Category ↑	Details
Name ↑↓	Description ↑↓					
Storage Blob Data Contributor	Allows for read, write and delete access to Azure Storage blob containers and data			BuiltinRole	Storage	View
Storage Blob Data Owner	Allows for full access to Azure Storage blob containers and data, including assigning POSIX access control.			BuiltinRole	Storage	View
Storage Blob Data Reader	Allows for read access to Azure Storage blob containers and data			BuiltinRole	Storage	View

- In Members, “Assign access to”
 - Select user, group or service principal
- On the right “Select members”, search for the Application Registration you made before and select it
- Review + assign

Home > Storage accounts > censusmalagadata | Access Control (IAM) >

Add role assignment ...

Role	Members*	Conditions	Review + assign
Selected role	Storage Blob Data Contributor		
Assign access to	<input checked="" type="radio"/> User, group, or service principal <input type="radio"/> Managed identity		
Members	+ Select members		
Name	Object ID	Type	
No members selected			
Description	Optional		

Select members

Select ⓘ app

- App Studio for Microsoft Teams
- APP TIMER APPTIMER@tajamar365.com
- appaccount1000_/_nsVP57tk5jBTs0mYpmEHvaBAoiS
- appaccount1000_9TmXUmnj+bDO0uh2EkoA3qCD+
- AppADMicrosoft
- AppADMicrosoft

Selected members:

- app01

Remove

Microsoft Azure

Home > Storage accounts > censusmalagadata | Access Control (IAM) >

Add role assignment ...

Role	Members	Conditions	Review + assign
Selected role	Storage Blob Data Contributor		
Assign access to	<input checked="" type="radio"/> User, group, or service principal <input type="radio"/> Managed identity		
Members	+ Select members		
Name	Object ID	Type	
app01	d168f953-727b-4733-956a-7d182dd32...	App	
Description	Optional		

Review + assign Previous Next

- The new role assignment is successfully added

The screenshot shows the Azure portal interface for managing access control (IAM) for a storage account named 'censusmalagadata'. The left sidebar includes options like Overview, Activity log, Tags, and Diagnose and solve problems. The main content area displays a table with columns for Role, Principal, and Start Date. A success message box is visible in the top right corner.

data

censusmalagadata | Access Control (IAM) ★ ...

Storage account

Search Add Download role assignments Edit columns Refresh Remove Feedback

Overview Activity log Tags Diagnose and solve problems

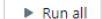
Check access Role assignments Roles Deny assignments Classic administrators

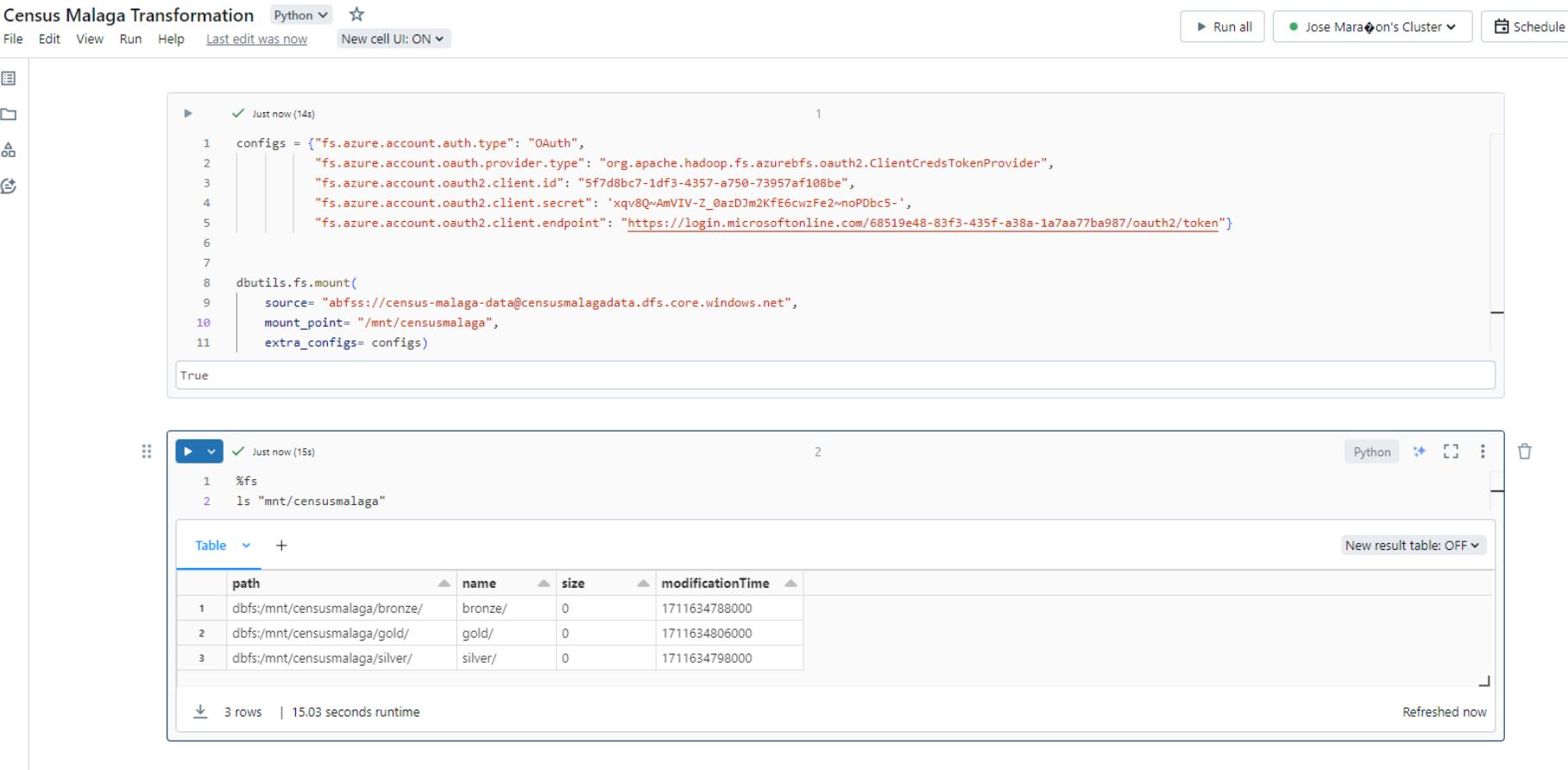
My access View my level of access to this resource. View my access

Added Role assignment
app01 was added as Storage Blob Data Contributor for censusmalagadata.

- Run the first cell of the Databricks notebook: you should get a “True” output
- Run the second cell to list all the available directories

Census Malaga Transformation Python 

File Edit View Run Help Last edit was now New cell UI: ON  Jose Maranon's Cluster 



```

1 configs = {"fs.azure.account.auth.type": "OAuth",
2   "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
3   "fs.azure.account.oauth2.client.id": "5f7d8bc7-1df3-4357-a750-73957af108be",
4   "fs.azure.account.oauth2.client.secret": "xqv8Q~AmVIV-Z_0azDjm2KfE6cwzFe2~noPDbc5-",
5   "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/68519e48-83f3-435f-a38a-1a7aa77ba987/oauth2/token"}
6
7
8 dbutils.fs.mount(
9   source= "abfss://census-malaga-data@censusmalagadata.dfs.core.windows.net",
10  mount_point= "/mnt/censusmalaga",
11  extra_configs= configs)

```

True


```

1 %fs
2 ls "/mnt/censusmalaga"

```

	path	name	size	modificationTime
1	dbfs:/mnt/censusmalaga/bronze/	bronze/	0	1711634788000
2	dbfs:/mnt/censusmalaga/gold/	gold/	0	1711634806000
3	dbfs:/mnt/censusmalaga/silver/	silver/	0	1711634798000

3 rows | 15.03 seconds runtime

Refreshed now

- Start loading the csv files
- Continue with the transformations. Let's coding. **SEE THE NOTEBOOK IN MY REPOSITORY**

Microsoft Azure |  **databricks** Search data, notebooks, recents, and more... CTRL + P census-dbricks Jose Marañon's Cluster Schedule

Census Malaga Transformation Python Last edit was 4 minutes ago New cell UI: ON

+ New Workspace Recents Catalog Workflows Compute SQL SQL Editor Queries Dashboards Alerts Query History SQL Warehouses Data Engineering Job Runs Data Ingestion Delta Live Tables Machine Learning Experiments Features Models Serving Marketplace Partner Connect

Cell 1 07:27 PM (14s)

```
dbutils.fs.mount(
  source= "abfss://census-malaga-data@censusmalagadata.dfs.core.windows.net",
  mount_point= "/mnt/censusmalaga",
  extra_configs= configs)
```

True

Cell 2 07:28 PM (15s)

```
%fs
ls "/mnt/censusmalaga"
```

Table + New result table: OFF

	path	name	size	modificationTime
1	dbfs:/mnt/censusmalaga/bronze/	bronze/	0	1711634788000
2	dbfs:/mnt/censusmalaga/gold/	gold/	0	1711634806000
3	dbfs:/mnt/censusmalaga/silver/	silver/	0	1711634798000

3 rows | 15.03 seconds runtime Refreshed 10 minutes ago

Cell 3 Just now (9s)

```
census2021 = spark.read.format("csv").option("header","true").load("/mnt/censusmalaga/bronze/2021/padronbarrios2021.csv")
census2021.show()
```

(2) Spark Jobs

census2021: pyspark.sql.dataframe.DataFrame = [NHOP: string, EDAD: string ... 7 more fields]

938	47	6	28	79	108	517	0	0
938	12	6	29	67	108	517	0	0
818	37	6	29	67	108	517	0	0
818	15	6	29	67	108	517	0	0
818	51	11	111	221	1081	5171	01	01

- As I managed cartographic data I tried to use one of the most used spark libraries: Apache SEDONA
- Unfortunately, after importing the library successfully, the engine was not able to find de Sedona module, so I decided not to get deeper in processing geodata. I obtained centroid, area and polygon perimeter from other opensource tool as ArcSig and loaded this information into the Bronze directory.
- Just you want to know how to import libraries, click on “compute”
- Navigate to Libraries and install new
- Search the library whether in PiPy or Maven repository and click on select.
- After installing you need to restart the cluster

The screenshot shows the Databricks Compute UI interface. At the top, there's a navigation bar with tabs for Configuration, Notebooks (1), Libraries (selected), Event log, Spark UI, Driver logs, Metrics, Apps, and Spark compute UI - Master. Below the navigation bar, there's a search bar labeled "Filter libraries" and a button labeled "Install new". The main area displays a table with columns for Status, Name, Type, and Source. A message "No libraries" is centered in the table area, with a note below it: "Please install new libraries with [Install New](#)". On the right side, a modal window titled "Search packages" is open, showing a list of artifacts from the Maven Central repository. The search term "sedona" is entered in the search bar. The table in the modal lists Group Id, Artifact Id, Releases, and Options for various Sedona artifacts, such as sedona-parent, sedona-spark-parent, and sedona-sql-parent, across different versions like 1.5.1, 1.4.1, and 3.0.2.

Status	Name	Type	Source

No libraries
Please install new libraries with [Install New](#)

Group Id	Artifact Id	Releases	Options
org.apache.sedona	sedona-parent-3.3.2.13	1.5.1	Select
org.apache.sedona	sedona-spark-parent-3.3.2.12	1.5.1	Select
org.apache.sedona	sedona-spark-parent-3.4.2.13	1.5.1	Select
org.apache.sedona	sedona-spark-parent-3.4.2.12	1.5.1	Select
org.apache.sedona	sedona-spark-parent-3.0.2.13	1.5.1	Select
org.apache.sedona	sedona-spark-parent-3.0.2.12	1.5.1	Select
org.apache.sedona	sedona-sql-parent-3.4.2.13	1.4.1	Select
org.apache.sedona	sedona-sql-parent-3.4.2.12	1.4.1	Select
org.apache.sedona	sedona-sql-parent-3.0.2.13	1.4.1	Select
org.apache.sedona	sedona-sql-parent-3.0.2.12	1.4.1	Select
org.apache.sedona	sedona-spark-shaded-3.3.2.13	1.5.1	Select
org.apache.sedona	sedona-spark-shaded-3.3.2.12	1.5.1	Select
org.apache.sedona	sedona-spark-shaded-3.4.2.13	1.5.1	Select
org.apache.sedona	sedona-spark-shaded-3.4.2.12	1.5.1	Select
org.apache.sedona	sedona-spark-shaded-3.0.2.13	1.5.1	Select
org.apache.sedona	sedona-spark-shaded-3.0.2.12	1.5.1	Select
org.apache.sedona	sedona-parent	1.5.1	Select
org.apache.sedona	sedona-parent-3.3.2.13	1.5.1	Select
org.apache.sedona	sedona-spark-common-3.5.2.13	1.5.1	Select

- Once installed you will see something like this
- Note there are two failed library installations

Compute >

Jose Marañon's Cluster ✓ !

Configuration Notebooks (1) Libraries Event log Spark UI Driver logs Metrics Apps Spark compute UI - Master

<input type="checkbox"/>	Status	Name	Type	Source
<input type="checkbox"/>	!	org.apache.sedona:sedona-core-3.4_2.12:1.5.1	Maven	-
<input type="checkbox"/>	✓	org.apache.sedona:sedona-spark-3.4_2.12:1.5.1	Maven	-
<input type="checkbox"/>	!	org.apache.sedona:sedona-spark-parent-3.4_2.12:1.5.1	Maven	-

- Databricks available Bronze file

Home >

 **census-malaga-data** ...

Container

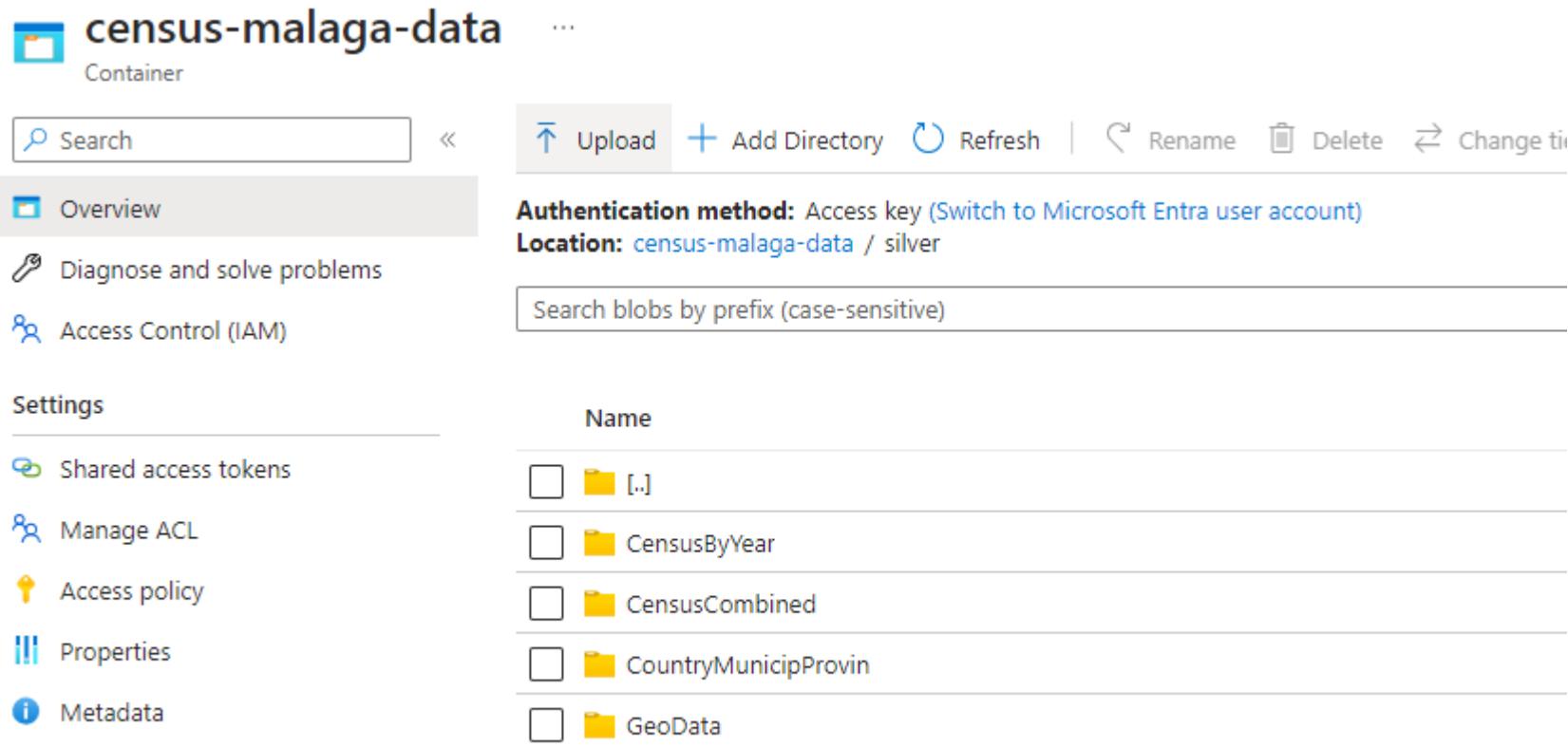
Upload Add Directory Refresh | Rename Delete Change tier Acquire lease Break lease Give feedback

Authentication method: Access key ([Switch to Microsoft Entra user account](#))
Location: [census-malaga-data](#) / bronze

Show deleted objects

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
<input type="checkbox"/>  2018					-	
<input type="checkbox"/>  2019					-	
<input type="checkbox"/>  2020					-	
<input type="checkbox"/>  2021					-	
<input type="checkbox"/>  cartohood-4326.geojson	3/28/2024, 5:04:52 PM	Hot (Inferred)		Block blob	1.71 MiB	Available
<input type="checkbox"/>  countries.csv	3/28/2024, 11:26:30 PM	Hot (Inferred)		Block blob	7.59 KiB	Available
<input type="checkbox"/>  geoCentroids.csv	3/29/2024, 12:53:14 AM	Hot (Inferred)		Block blob	52.32 KiB	Available
<input type="checkbox"/>  geohoods.csv	3/28/2024, 11:26:10 PM	Hot (Inferred)		Block blob	1.85 MiB	Available
<input type="checkbox"/>  geoPerim_Area.csv	3/29/2024, 12:53:05 AM	Hot (Inferred)		Block blob	61.28 KiB	Available
<input type="checkbox"/>  municipality.csv	3/28/2024, 11:26:43 PM	Hot (Inferred)		Block blob	389.7 KiB	Available
<input type="checkbox"/>  province.csv	3/28/2024, 11:26:58 PM	Hot (Inferred)		Block blob	1.36 KiB	Available

- Databricks transformed data is persisted into Silver layer in the ADLS container.



census-malaga-data ...

Container

Search <>

Upload Add Directory Refresh Rename Delete Change tier

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: [census-malaga-data](#) / silver

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Name

	Name
<input type="checkbox"/>	[..]
<input type="checkbox"/>	CensusByYear
<input type="checkbox"/>	CensusCombined
<input type="checkbox"/>	CountryMunicipProvin
<input type="checkbox"/>	GeoData

- This Databricks notebook cell save all the transformations into the Silver layer
- Check the complete notebook in my Github Repository
- **Disclaimer:** writing the file with `.partitionBy(censusYear)` is a good idea to keep all file partitions organized, although it became a bit messy as I wanted to access the data from Azure Synapse.
- I recommend to use the approach of line 10. All the file partitions will be inside the same directory and easy parsed by Azure Synapse

4. SAVING INTO SILVER LAYER



The screenshot shows a Databricks notebook cell with the following details:

- Timestamp:** 5 minutes ago (5m)
- Cell ID:** 76
- Languages:** Python
- Code Content:**

```

1
2 df_countries.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CountryMunicipProvin/countries")
3 df_municipalities.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CountryMunicipProvin/municipalities")
4 df_provinces.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CountryMunicipProvin/provinces")
5
6 df_full_census2021.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CensusByYear/2021")
7 df_full_census2020.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CensusByYear/2020")
8 df_full_census2019.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CensusByYear/2019")
9 df_full_census2018.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CensusByYear/2018")
10 #df_census_combined.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/CensusCombined")
11
12 df_census_combined.write.mode("overwrite").partitionBy("censusYear").option("header", "true").csv("/mnt/censusmalaga/silver/CensusCombined")
13
14
15 df_geoDetail.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/GeoData/Detail")
16 df_geoFinal.write.mode("overwrite").option("header", "true").csv("/mnt/censusmalaga/silver/GeoData/Final")
17
18
19

```
- Spark Jobs:** (28) Spark Jobs

- See the folder path, an additional directory was created for each year where are the file partitions
- It didn't work well for me. So I delete censusByYear and run the code of line 10 show before

Home > censusmalagadata | Containers >

census-malaga-data

Container

Search

Upload Add Directory Refresh Rename Delete

Overview

Authentication method: Access key ([Switch to Microsoft Entra user account](#))
Location: census-malagadata / silver / CensusCombined

Search blobs by prefix (case-sensitive)

Name
<input type="checkbox"/>  censusYear=2018
<input type="checkbox"/>  censusYear=2019
<input type="checkbox"/>  censusYear=2020
<input type="checkbox"/>  censusYear=2021
<input type="checkbox"/>  _SUCCESS

Settings

Shared access tokens
Manage ACL
Access policy
Properties
Metadata

census-malaga-data

Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Br

Overview

Authentication method: Access key ([Switch to Microsoft Entra user account](#))
Location: census-malagadata / silver / CensusByYear / 2018

Search blobs by prefix (case-sensitive)

Name	Modified
<input type="checkbox"/>  _committed_6282856444426058191	3/29/2024, 5:04:43 PM
<input type="checkbox"/>  _started_6282856444426058191	3/29/2024, 5:04:15 PM
<input type="checkbox"/>  _SUCCESS	3/29/2024, 5:04:44 PM
<input type="checkbox"/>  part-00000-tid-6282856444426058191-17c51e6e-a773-49c0-b741-fc76f85385df-120-1-c...	3/29/2024, 5:04:41 PM
<input type="checkbox"/>  part-00001-tid-6282856444426058191-17c51e6e-a773-49c0-b741-fc76f85385df-121-1-c...	3/29/2024, 5:04:43 PM
<input type="checkbox"/>  part-00002-tid-6282856444426058191-17c51e6e-a773-49c0-b741-fc76f85385df-122-1-c...	3/29/2024, 5:04:42 PM
<input type="checkbox"/>  part-00003-tid-6282856444426058191-17c51e6e-a773-49c0-b741-fc76f85385df-123-1-c...	3/29/2024, 5:04:26 PM

- The other transformed data was persisted correctly inside Silver layer

Home > censusmalagadata | Containers >

census-malaga-data

Container

Search

Upload Add Directory Refresh Rename Delete Change tier Acquire lease

Authentication method: Access key ([Switch to Microsoft Entra user account](#))
Location: [census-malaga-data](#) / [silver](#) / [CountryMunicipProvin](#) / [municipalities](#)

Search blobs by prefix (case-sensitive)

Name	Modified
<input type="checkbox"/> _committed_3402886456646619624	3/29/2024, 5:
<input type="checkbox"/> _started_3402886456646619624	3/29/2024, 5:
<input type="checkbox"/> _SUCCESS	3/29/2024, 5:
<input type="checkbox"/> part-00000-tid-3402886456646619624-688ccca3-d4ef-4320-9d9b-a3b4d813e1ff-94-1-c0...	3/29/2024, 5:

Overview
Diagnose and solve problems
Access Control (IAM)

Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata

- In my case, I decided to persist the GOLD data directly from Databricks as I didn't make relevant changes from Silver data.
- For the next steps with Azure Synapse, I just used it to create tables and make them available for PowerBI.

Home > [censusmalagadata](#) | Containers >

census-malaga-data

Container

Search < Upload Add Directory Refresh | Rename Delete Change tier Acc

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: [census-malaga-data](#) / [silver](#) / [GeoData](#) / [Final](#)

Search blobs by prefix (case-sensitive)

Name	Modified
_committed_6718974304946525474	3/29/
_started_6718974304946525474	3/29/
_SUCCESS	3/29/
part-00000-tid-6718974304946525474-0cafb750-f1b6-4df5-a7e4-6a8894edb327-147-1...	3/29/

Overview Diagnose and solve problems Access Control (IAM)

Settings

- Shared access tokens
- Manage ACL
- Access policy
- Properties
- Metadata

- The main csv file partitions with all the demographic information saved into the GOLD layer

Microsoft Azure

Home > censusmalagadata | Containers >

census-malaga-data Container

Search Overview Diagnose and solve problems Access Control (IAM)

Upload Add Directory Refresh Rename Delete Change tier Acquire lease Break lease Give fees

Authentication method: Access key ([Switch to Microsoft Entra user account](#))
Location: [census-malaga-data](#) / [gold](#) / [CensusCombined](#)

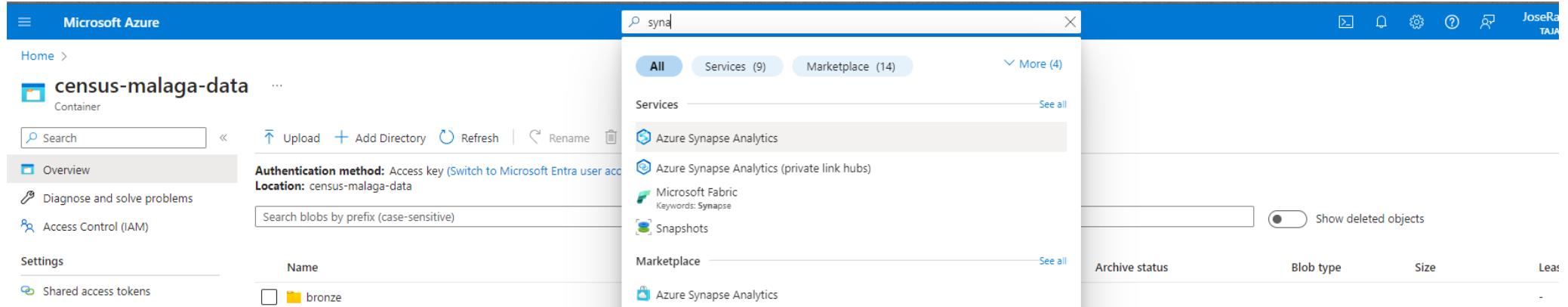
Search blobs by prefix (case-sensitive)

Name	Modified
<input type="checkbox"/> _committed_4013896121287938386	3/29/2024, 6:51:25 PM
<input type="checkbox"/> _started_4013896121287938386	3/29/2024, 6:49:41 PM
<input type="checkbox"/> _SUCCESS	3/29/2024, 6:51:26 PM
<input type="checkbox"/> part-00000-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-173-1-c000.csv	3/29/2024, 6:50:11 PM
<input type="checkbox"/> part-00001-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-174-1-c000.csv	3/29/2024, 6:50:09 PM
<input type="checkbox"/> part-00002-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-175-1-c000.csv	3/29/2024, 6:50:11 PM
<input type="checkbox"/> part-00003-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-176-1-c000.csv	3/29/2024, 6:49:53 PM
<input type="checkbox"/> part-00004-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-177-1-c000.csv	3/29/2024, 6:50:23 PM
<input type="checkbox"/> part-00005-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-178-1-c000.csv	3/29/2024, 6:50:38 PM
<input type="checkbox"/> part-00006-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-179-1-c000.csv	3/29/2024, 6:50:37 PM
<input type="checkbox"/> part-00007-tid-4013896121287938386-b081abb5-8b51-41e1-a357-8ab3871f46cf-180-1-c000.csv	3/29/2024, 6:50:23 PM

- Notebook table of contents



- Search for Azure Synapse Analytics



- Create an Azure Synapse workspace

Microsoft Azure

Search resources, services, and docs (G+)

Home > Azure Synapse Analytics

Tajamar (tajamar365.com)

+ Create Manage view Refresh Export to CSV Open query Assign tags

Create Filter for any resource Subscription equals all Resource group equals all Location equals all Add filter

Showing 0 to 0 of 0 records.

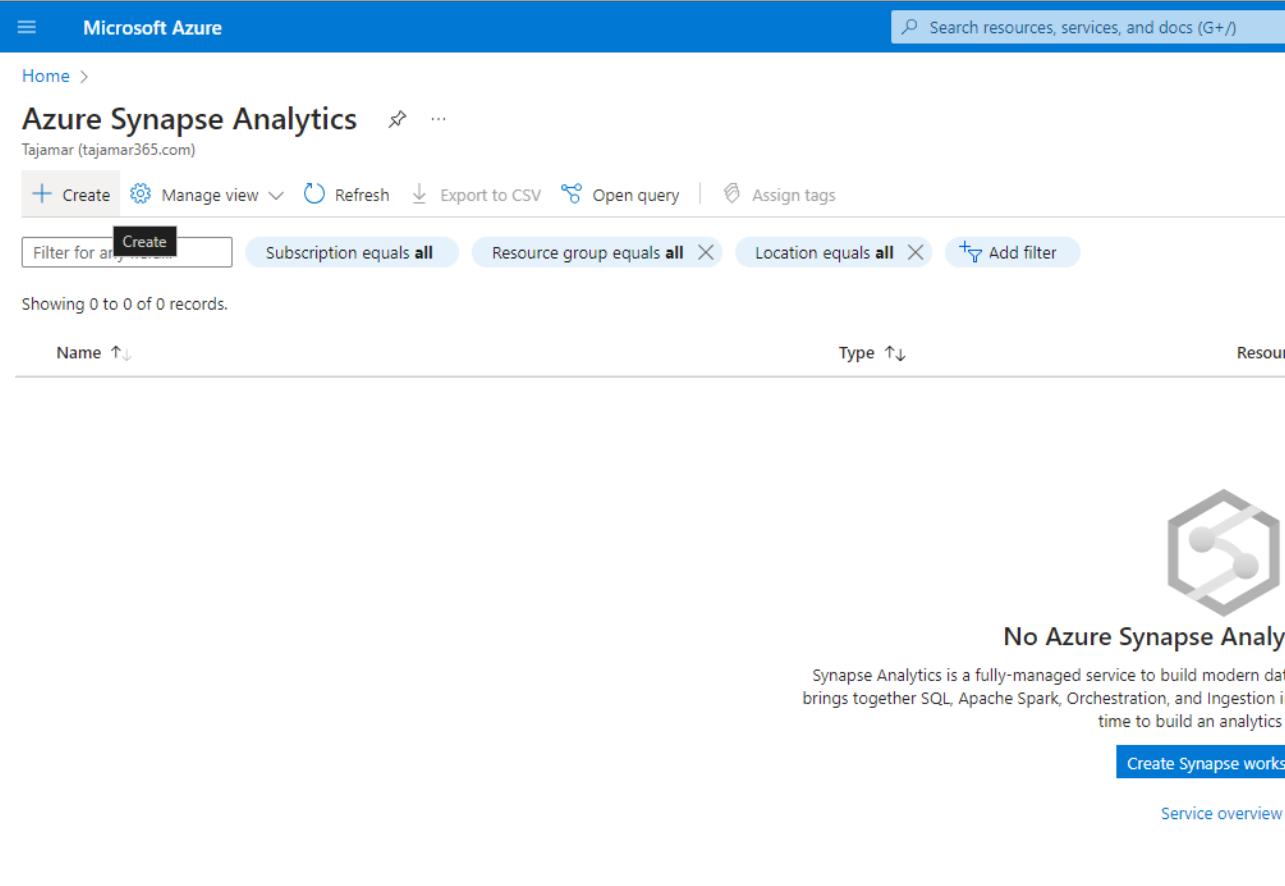
Name ↑↓	Type ↑↓	Resource
---------	---------	----------

No Azure Synapse Analytics

Synapse Analytics is a fully-managed service to build modern data warehouses and analytics pipelines. It brings together SQL, Apache Spark, Orchestration, and Ingestion into one service, so you can spend less time to build an analytics solution.

Create Synapse workspace

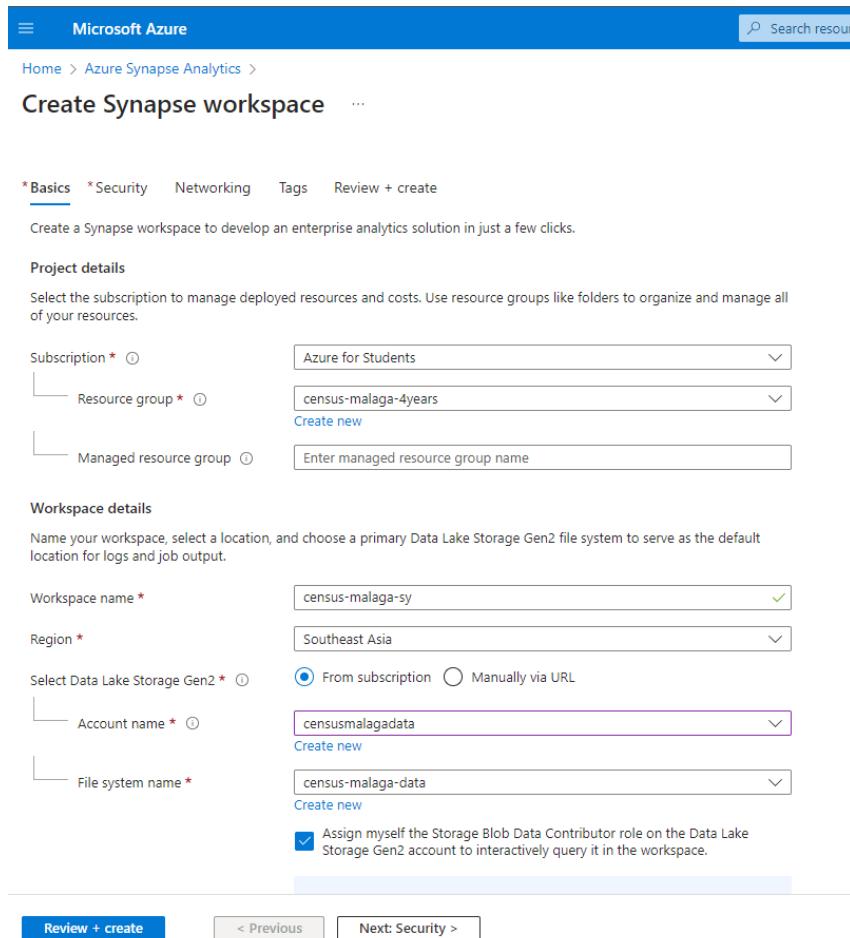
Service overview



- Select your free subscription
- Select the existing resource group

- Give the synapse workspace a name
- Region: southEast Asia
- Select Data Lake Storage Gen2: from subscription
 - Accountname: select the existing one
 - File system name: appear by default as you select the account name

- NEXT



The screenshot shows the 'Create Synapse workspace' page in the Microsoft Azure portal. The top navigation bar includes 'Microsoft Azure', a search bar, and a 'Search resources' icon. Below the navigation, the breadcrumb trail shows 'Home > Azure Synapse Analytics > Create Synapse workspace'. The main title is 'Create Synapse workspace ...'. The page is divided into sections: 'Project details' and 'Workspace details'. In the 'Project details' section, there are fields for 'Subscription' (set to 'Azure for Students'), 'Resource group' (set to 'census-malaga-4years'), and 'Managed resource group' (with a placeholder 'Enter managed resource group name'). In the 'Workspace details' section, the 'Workspace name' is 'census-malaga-sy' (with a green checkmark), the 'Region' is 'Southeast Asia', and the 'Select Data Lake Storage Gen2' option is set to 'From subscription'. Under 'From subscription', the 'Account name' is 'censusalagadata' and the 'File system name' is 'census-malaga-data'. A checkbox at the bottom is checked, stating: 'Assign myself the Storage Blob Data Contributor role on the Data Lake Storage Gen2 account to interactively query it in the workspace.' Navigation buttons at the bottom include 'Review + create', '< Previous', and 'Next: Security >'.

- Next

Microsoft Azure

Home > Azure Synapse Analytics >

Create Synapse workspace

* Basics * **Security** Networking Tags Review + create

Configure security options for your workspace.

Authentication

Choose the authentication method for access to workspace resources such as SQL pools. The authentication method can be changed later on. [Learn more](#)

Authentication method Use both local and Microsoft Entra ID authentication Use only Microsoft Entra ID authentication

SQL Server admin login *

SQL Password ✓

Confirm password ✓

System assigned managed identity permission

Select to grant the workspace network access to the Data Lake Storage Gen2 account using the workspace system identity. [Learn more](#)

Allow network access to Data Lake Storage Gen2 account. ⓘ The selected Data Lake Storage Gen2 account does not restrict network access using any network access rules, or you selected a storage account manually via URL under Basics tab. [Learn more](#)

Workspace encryption

⚠ Double encryption configuration cannot be changed after opting into using a customer-managed key at the time of workspace creation.

Choose to encrypt all data at rest in the workspace with a key managed by you (customer-managed key). This will provide double encryption with encryption at the infrastructure layer that uses platform-managed keys. [Learn more](#)

Review + create [< Previous](#) [Next: Networking >](#)

- Next

Microsoft Azure

Home > Azure Synapse Analytics >

Create Synapse workspace

* Basics * Security Networking Tags Review + create

Configure networking options for your workspace.

Managed virtual network

Choose whether to set up a dedicated Azure Synapse-managed virtual network for your workspace. [Learn more](#)

Enable Disable

ⓘ To control public network access to your Synapse workspace, you must enable managed virtual network.

Firewall rules

⚠ Azure Synapse Studio and other client tools will only connect to the workspace endpoints if this setting is selected. Connections from specific IP addresses or all Azure services can be allowed or disallowed after the workspace is provisioned.

Allow connections from all IP addresses to your workspace's endpoints. You can restrict these permissions to just Azure datacenter IP addresses and/or specific IP address ranges after creating the workspace.

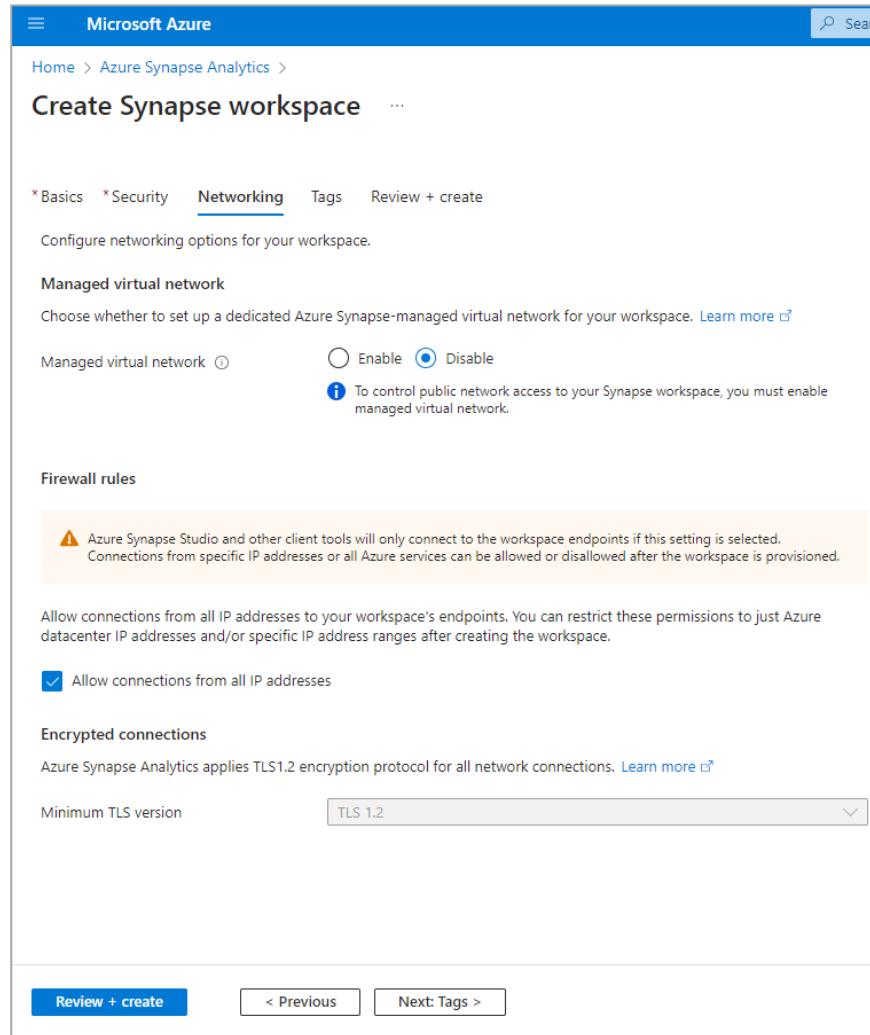
Allow connections from all IP addresses

Encrypted connections

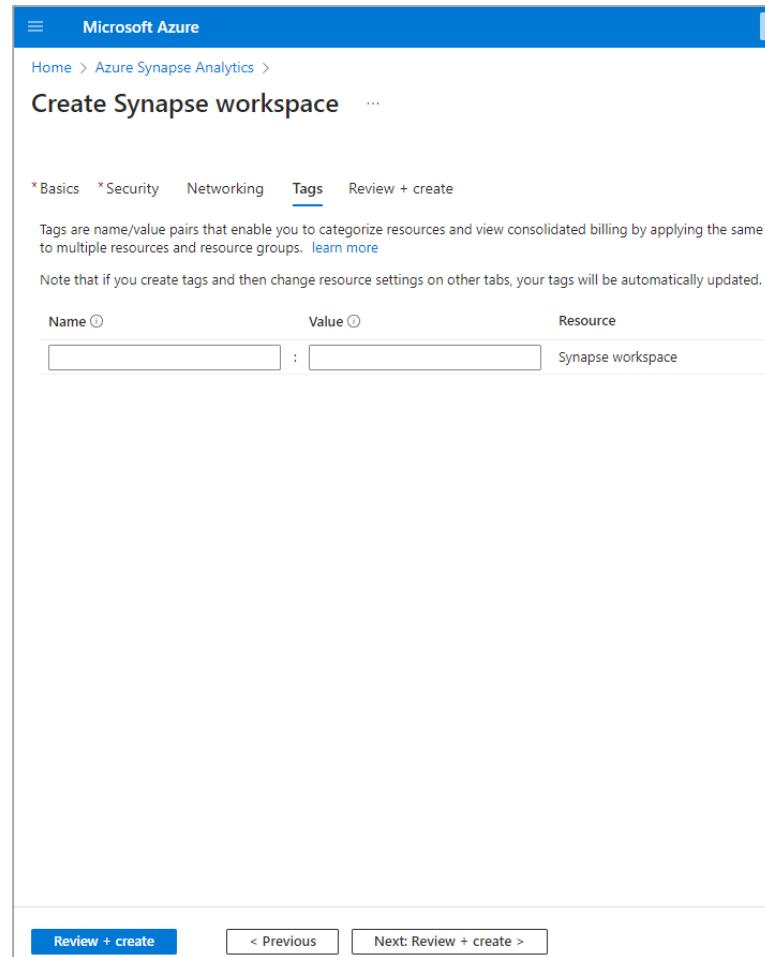
Azure Synapse Analytics applies TLS1.2 encryption protocol for all network connections. [Learn more](#)

Minimum TLS version:

[Review + create](#) [< Previous](#) [Next: Tags >](#)

The screenshot shows the 'Networking' tab of the 'Create Synapse workspace' wizard. It includes sections for 'Managed virtual network' (with 'Disable' selected), 'Firewall rules' (allowing connections from all IP addresses), and 'Encrypted connections' (set to TLS 1.2). Navigation buttons at the bottom include 'Review + create', '< Previous', and 'Next: Tags >'.

- Next



- Create

Microsoft Azure Search

Home > Azure Synapse Analytics >

Create Synapse workspace

Validation succeeded

* Basics * Security Networking Tags Review + create

Product Details

Azure Synapse Analytics workspace by Microsoft Serverless SQL est. cost/TB ⓘ **5.00 USD**

[Terms of use](#) | [Privacy policy](#)

Terms

By clicking Create, I (a) agree to the legal terms and privacy statement(s) associated with the Marketplace offering(s) listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with the offering(s), with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my contact, usage and transactional information with the provider(s) of the offering(s) for support, billing and other transactional activities. Microsoft does not provide rights for third-party offerings. For additional details see [Azure Marketplace Terms](#).

Basics

Subscription	Azure for Students
Resource group	census-malaga-4years
Region	Southeast Asia
Workspace name	(new) census-malaga-sy
Data Lake Storage Gen2 account	https://censusmalagadata.dfs.core.windows.net
Data Lake Storage Gen2 file system	census-malaga-data
Managed resource group	None
Role assignments	The Storage Blob Data Contributor role will be assigned on the specified Data Lake Storage Gen2 account to both the workspace managed identity and the current user.

Create < Previous Next > [Download a template for automation](#)

- Wait until Synapse workspace deployment

Microsoft Azure | Microsoft.Azure.SynapseAnalytics-20240329171236 | Overview

Deployment is in progress.

Deployment name : Microsoft.Azure.SynapseAnalytics-20240329171236

Subscription : Azure for Students

Resource group : census-malaga-4years

Start time : 3/29/2024, 5:17:57 PM

Correlation ID : f6227612-e0a3-4cb8-af1a-666504b0c6b9

Deployment details

Resource	Type	Status	Operation details
There are no resources to display.			

Give feedback

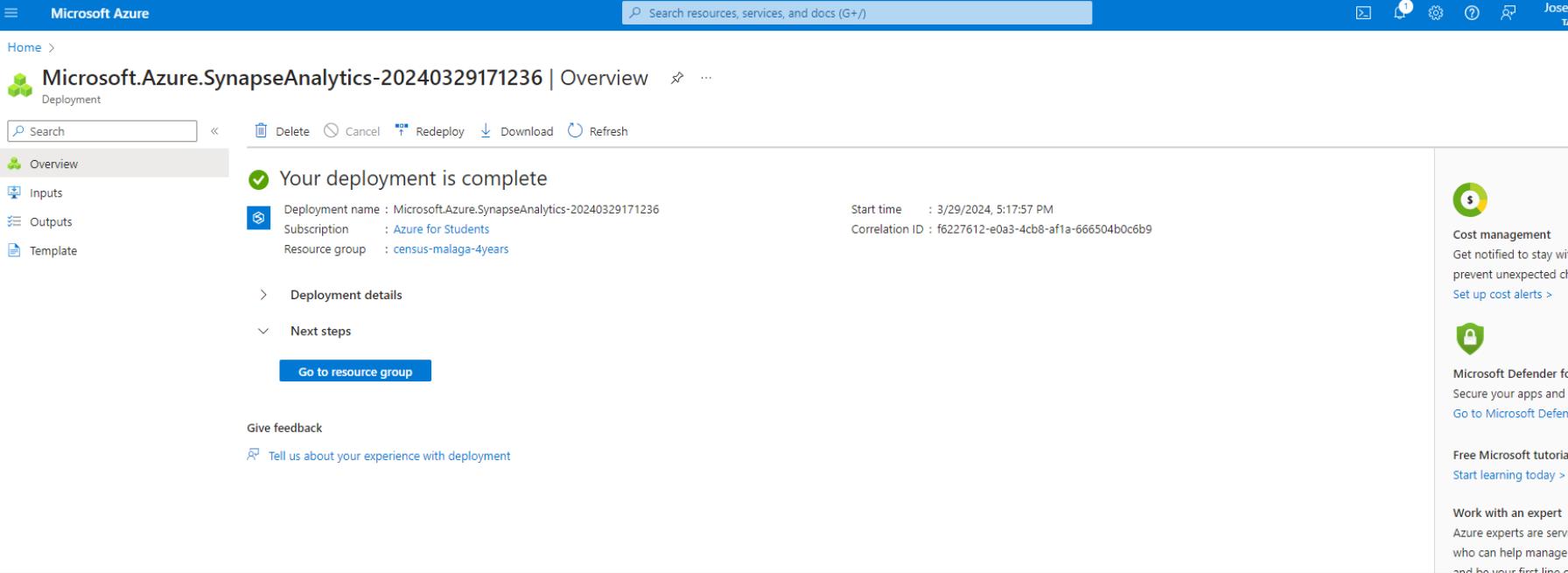
Tell us about your experience with deployment

Microsoft Defender for Cloud
Secure your apps and infrastructure
[Go to Microsoft Defender for Cloud](#)

Free Microsoft tutorials
[Start learning today >](#)

Work with an expert
Azure experts are service providers who can help manage your assets and be your first line of support.
[Find an Azure expert >](#)

- Synapse workspace deployed
- Go to resource group



The screenshot shows the Microsoft Azure Deployment Overview page for a Synapse workspace. The deployment is complete, with the name `Microsoft.Azure.SynapseAnalytics-20240329171236`, subscription `Azure for Students`, and resource group `census-malaga-4years`. The deployment started at 3/29/2024, 5:17:57 PM. The page includes links for deployment details, next steps, and a Go to resource group button. It also features a sidebar with links for cost management, Microsoft Defender, free tutorials, and working with experts.

Microsoft Azure

Microsoft.Azure.SynapseAnalytics-20240329171236 | Overview

Deployment

Search | Delete | Cancel | Redeploy | Download | Refresh

Overview

Your deployment is complete

Deployment name : Microsoft.Azure.SynapseAnalytics-20240329171236
Subscription : Azure for Students
Resource group : census-malaga-4years

Start time : 3/29/2024, 5:17:57 PM
Correlation ID : f6227612-e0a3-4cb8-af1a-666504b0c6b9

Deployment details

Next steps

Go to resource group

Give feedback

Tell us about your experience with deployment

Cost management

Get notified to stay within budget and prevent unexpected charges

Set up cost alerts >

Microsoft Defender for Cloud

Secure your apps and infrastructure

Go to Microsoft Defender

Free Microsoft tutorial

Start learning today >

Work with an expert

Azure experts are service providers who can help manage your cloud environment and be your first line of defense

- Click on the synapse workspace

Microsoft Azure

Search resources, services, and docs (G+)

Home > Microsoft.Azure.SynapseAnalytics-20240329171236 | Overview >

census-malaga-4years

Resource group

Essentials

Subscription (move) : Azure for Students
Subscription ID : 9e3c713b-0d42-41e5-81ca-d778fd001da7
Tags (edit) : Add tags

Deployments : 5 Succeeded
Location : West Europe

Resources Recommendations

Name	Type	Location
census-dbricks	Azure Databricks Service	Southeast Asia
census-malaga-df	Data factory (V2)	West Europe
census-malaga-sy	Synapse workspace	Southeast Asia
censusmalagadata	Storage account	West Europe

Filter for any field... Type equals all X Location equals all X + Add filter

Show hidden types ⓘ No grouping

Name ↑↓ Type ↑↓ Location ↑↓

- Azure synapse workspace console
- Open Synapse Studio

Microsoft Azure

Search resources, services, and docs (G+)

Home > Microsoft.Azure.SynapseAnalytics-20240329171236 | Overview > census-malaga-4years >

census-malaga-sy Synapse workspace

Search + New dedicated SQL pool + New Apache Spark pool + New Data Explorer pool (preview) Refresh Reset SQL admin password Delete

Essentials

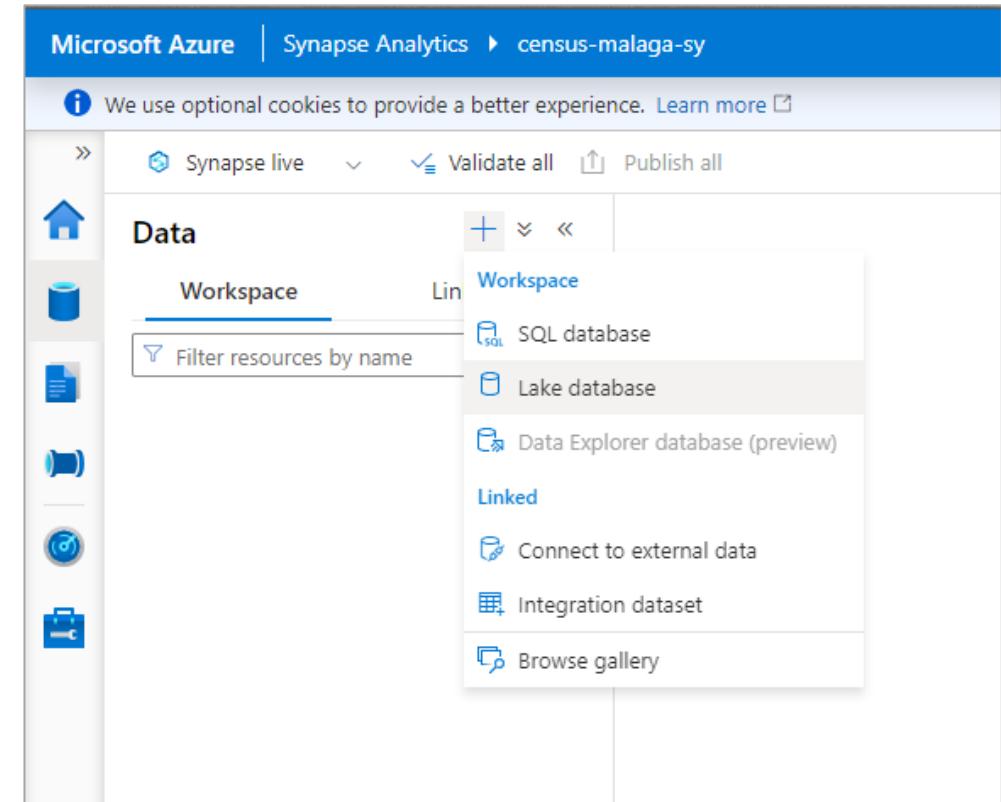
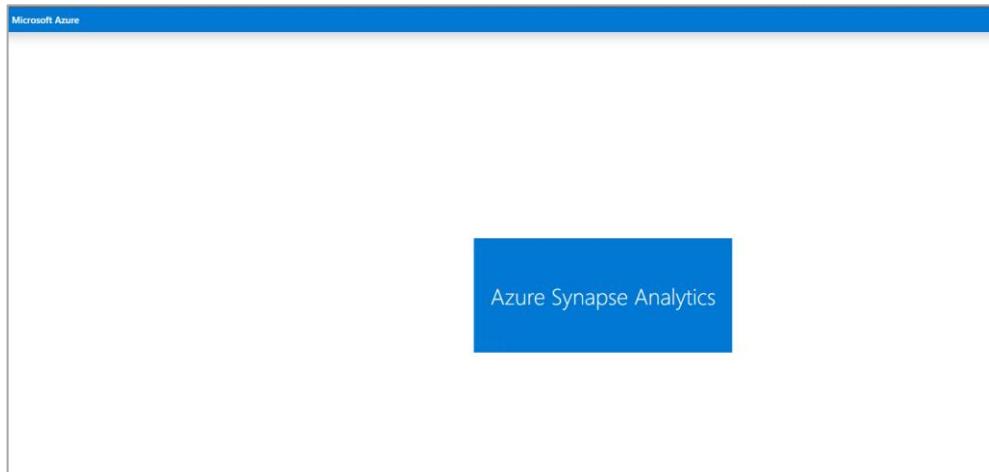
Resource group (move) :	census-malaga-4years	Networking :	Show firewall settings
Status :	Succeeded	Primary ADLS Gen2 acco...	https://censusmalagadata.dfs.core.windows.net
Location :	Southeast Asia	Primary ADLS Gen2 file s...	census-malaga-data
Subscription (move) :	Azure for Students	SQL admin username :	sqladminuser
Subscription ID	: 9e3c713b-0d42-41e5-81ca-d778fd001da7	SQL Microsoft Entra admin :	JoseRafael.Vera@tajamar365.com
Managed virtual network	: No	Dedicated SQL endpoint :	census-malaga-sy.sql.azuresynapse.net
Managed Identity object ...	: 395a06d4-7035-4107-831d-a4337fb23f5a	Serverless SQL endpoint :	census-malaga-sy-on-demand.sql.azuresynapse.net
Workspace web URL	: https://web.azuresynapse.net?workspace=%2bsubscriptions%2fe3c713b-0d42-41e5-81ca-d778fd00...	Development endpoint :	https://census-malaga-sy.dev.azuresynapse.net
Tags (edit)	: Add tags		

Getting started


Open Synapse Studio
 Start building your fully-integrated analytics solution and unlock new insights.
[Open ↗](#)


Read documentation
 Learn how to be productive quickly. Explore concepts, tutorials, and samples.
[Learn more ↗](#)

- Azure synapse studio
- Click on Data
- Click on + and “Lake database”



- Click ok
- You will see the console of the recent created Lake database
- On the right, give the database a name
- Tables directory is yet empty

Microsoft Azure | Synapse Analytics > census-malaga-sy

We use optional cookies to provide a better experience. Learn more [\[\]](#)

Synapse live Validate all Publish all [\[\]](#)

Data Workspace Linked

Filter resources by name

Lake database 1

Database1

+ Table Map data Publish

Tables Filter by keyword

Azure Synapse Database Template Terms of Use

Azure Synapse Analytics

"Synapse Database Templates" means industry specific data models and schematics, or portions thereof, and related documentation that Microsoft provides to Customer.

Use rights

Customer may access and use Synapse Database Templates solely for Customer's internal business purposes and only for running on Azure in conjunction with Azure Synapse Analytics.

OK

Microsoft Azure | Synapse Analytics > census-malaga-sy

We use optional cookies to provide a better experience. Learn more [\[\]](#)

Synapse live Validate all Publish all [\[\]](#)

Data Workspace Linked

Filter resources by name

CensusDB 1

Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace. [\[\]](#)

+ Table Map data Publish

Tables Filter by keyword

Get started

Select tables to build your database.

Properties

General Related (0)

Choose a name for your Database. This name can be updated at any time until it is published.

Name * CensusDB

Description

Storage settings for database

Linked service * census-malaga-sy-WorkspaceDefa...

Input folder * census-malaga-data/CensusDB

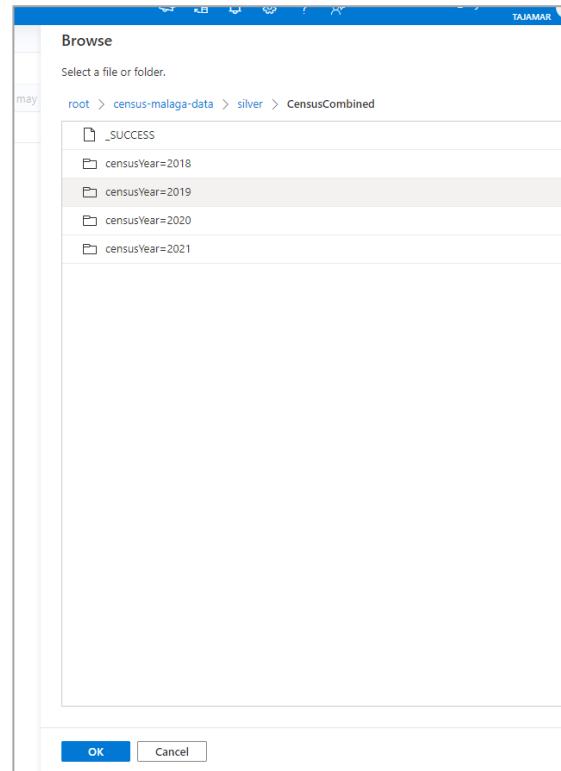
Data format * Delimited Text

- Let's create the tables from Silver data (in our ADLS container)
- Click on the + Table and select "From data lake"
- On the right:
 - Give the table a name
 - Linked service: select which appears as you click in the dropdown
 - Input file or folder: browse the directories to find the desired file

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. The top navigation bar displays 'Microsoft Azure | Synapse Analytics > census-malaga-sy'. Below the navigation bar, there are several icons: a house icon for Home, a blue cylinder for Databases, a document icon for Workspaces, and a target icon for Pipelines. The 'Data' section is currently selected, indicated by a blue underline. A 'Workspace' tab is active. On the left, there is a sidebar with a search bar labeled 'Filter resources by name' and a tree view under 'Lake database' showing 'CensusDB' and 'Tables'. In the main content area, there is a 'Table' button with a dropdown arrow. The dropdown menu is open, showing four options: 'Custom', 'From template', 'From data lake', and 'From data lake' (which is highlighted with a gray background). Above the dropdown, there is a 'Map data' button.

The screenshot shows the 'Create external table from data lake' dialog box. At the top, it says 'Create external table from data lake'. Below that, 'External table details' are described: 'Select the storage location where the files containing the data is staged. Currently Azure Data Lake Storage (ADLS) Gen2 and Azure Blob Storage are supported.' Under 'External table name', the value 'CensusCombined' is entered. Under 'Linked service', the value 'census-malaga-sy-WorkspaceDefaultStorage(censusmalagadata)' is selected. Under 'Input file or folder', there is a red error message: 'Select the file path'. At the bottom of the dialog are 'Continue' and 'Cancel' buttons.

- While browsing you are able to select a whole directory or a specific file
 - If the file is partitioned, then you should select the directory where all the partitions are.
 - Synapse will merge all the partitions together to build up a table
- The image below shows how I was selecting the CensusCombined folder, pretending to get all the partitions together as they were partitioned by Year (do you remember before?).
 - I tried this way, and the table was empty, so I loaded into Silver all the partitions of censuscombined together inside the same folder (remember line 10 of databricks notebook)
- ok



- As you clicked ok, you see the file path assilver/CensusCombined/**
 - That means that the input path includes all the files under this directory
- First row: infer column names must be checked
- Create

New external table

Source file format settings
Specify the format and layout of your data. [Learn more](#)

File path
census-malaga-data/silver/CensusCombined/**

Preview Data

File type
CSV

Field terminator ⓘ
Default (comma ,)
 Edit

First row
 Infer column names ⓘ

String delimiter ⓘ
Default (Empty string)
 Edit

Use default type ⓘ
Default type (true,false)

Max string length * ⓘ
4000

- Table added to the synapse workspace

The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The top navigation bar includes 'Microsoft Azure', 'Synapse Analytics', and 'census-malaga-sy'. The main area is titled 'CensusDB' under the 'Tables' section. A message at the top right states: 'Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace.' Below this, the 'Properties' pane is open, showing the 'General' tab with fields for 'Name' (set to 'CensusDB') and 'Description'. Under 'Storage settings for database', it shows a 'Linked service' dropdown set to 'census-malaga-sy-WorkspaceDefa...' and an 'Input folder' dropdown set to 'census-malaga-data/CensusDB'. The 'Data' sidebar on the left shows a tree structure with 'Lake database' expanded, showing 'CensusDB' and 'Tables'.

CensusDB

Tables

CensusCombined

- 121 SheetNumber
- 121 Age
- 121 Sex
- 121 MunicipalityCode
- 121 BirthCountry
- 121 DistrictNumber
- 121 YearOfBirth
- abc SexDescription
- 121 Control
- abc CountryName

Name *: CensusDB

Description: Enter a description

Linked service *: census-malaga-sy-WorkspaceDefa...

Input folder *: census-malaga-data/CensusDB

Data format *: Delimited Text

- Explore the table columns, data type and names. You can even start creating relationships between tables. I made this on POWERBI
- Click on validate all
- Click on publish all to save all the changes in the workspace

The screenshot shows the Microsoft Power BI workspace validation interface. At the top, there are buttons for "Synapse live", "Validate all" (with a count of 1), and "Publish all". Below this, the workspace navigation bar includes "CensusDB", "Table", "Map data", and "Publish". A warning message states: "Other users in your workspace may have access to modify this item. Do not use this item unless you trust all users who may have access to the workspace." The main area displays the "Tables" section with a single table named "CensusCombined". The table has the following columns:

Column Name	Type
SheetNumber	long
Age	long
Sex	string
MunicipalityCode	long
BirthCountry	long
DistrictNumber	long
YearOfBirth	long

On the right side, there is a "Workspace validation output" panel. It features a green checkmark icon and the message: "Your workspace has been validated. No errors were found."

- What you see when click on publish all
- Click on publish

Microsoft Azure | Synapse Analytics > census-malaga-sy

We use optional cookies to provide a better experience. Learn more

Synapse live ▾ Validate all Publishing 1

CensusDB

Tables

Filter by keyword

Others 1

CensusCombined

121. SheetNumber
121. Age
121. Sex
121. MunicipalityCode
121. BirthCountry
121. DistrictNumber
121. YearOfBirth

General Columns Relationships

Filter by keyword

	Column	PK	Null	Type	Length
<input type="checkbox"/>	DistrictNumber	<input type="checkbox"/>	<input checked="" type="checkbox"/>	long	
<input type="checkbox"/>	YearOfBirth	<input type="checkbox"/>	<input checked="" type="checkbox"/>	long	
<input type="checkbox"/>	SexDescription	<input type="checkbox"/>	<input checked="" type="checkbox"/>	string	8000
<input type="checkbox"/>	Control	<input type="checkbox"/>	<input checked="" type="checkbox"/>	long	
<input type="checkbox"/>	CountryName	<input type="checkbox"/>	<input checked="" type="checkbox"/>	string	8000
<input type="checkbox"/>	Province	<input type="checkbox"/>	<input checked="" type="checkbox"/>	string	8000
<input type="checkbox"/>	ProvinceCode	<input type="checkbox"/>	<input checked="" type="checkbox"/>	long	
<input type="checkbox"/>	ProvinceName	<input type="checkbox"/>	<input checked="" type="checkbox"/>	string	8000
<input type="checkbox"/>	MunicipalityName	<input type="checkbox"/>	<input checked="" type="checkbox"/>	string	8000

Partition column (0)

Publish all

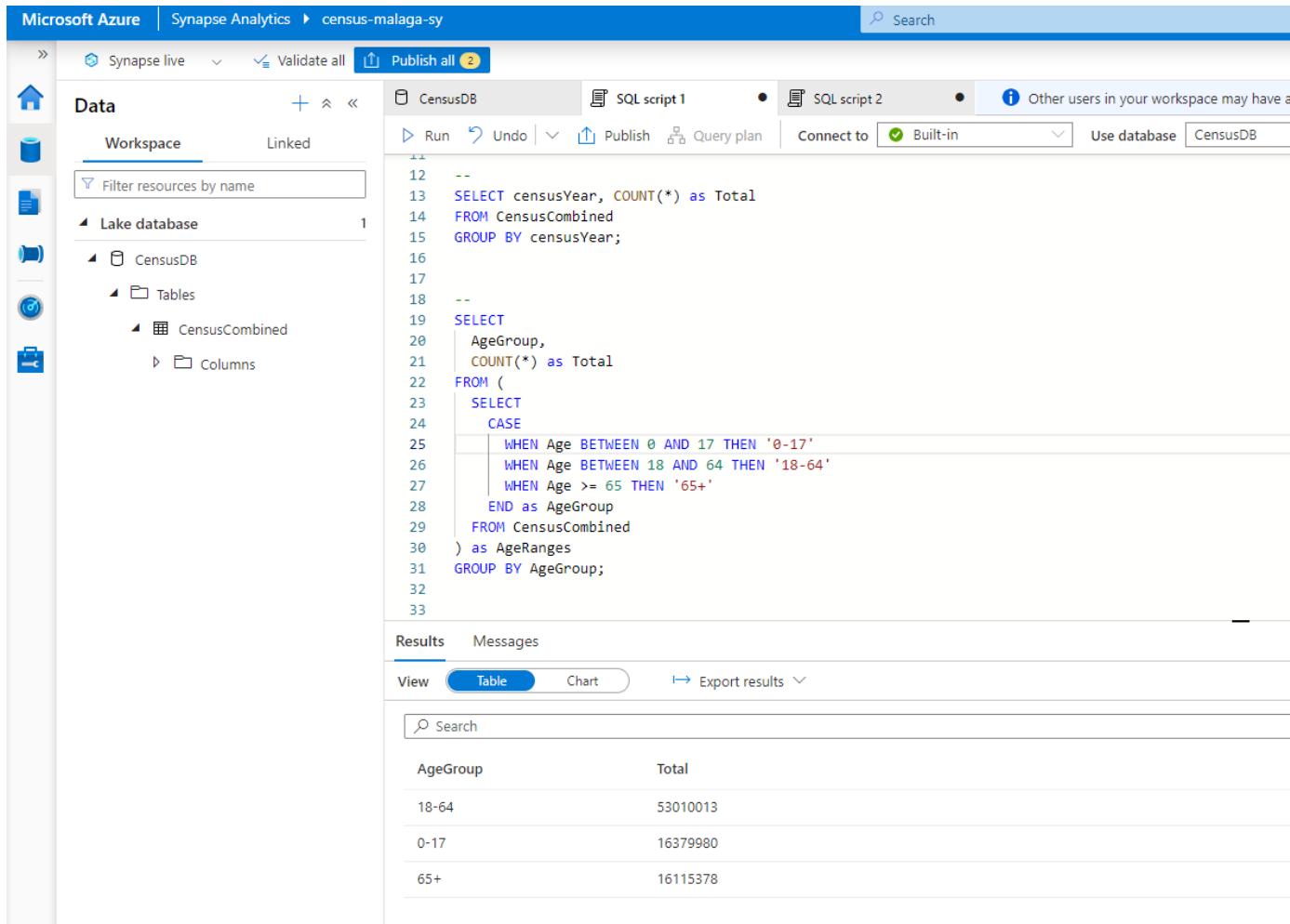
You are about to publish all pending changes to the live environment. [Learn](#)

Pending changes (1)

NAME	CHANGE	EXISTING
Database		
CensusDB	(New)	-

Publish Cancel

- Run queries against the table. **SEE MY REPOSITORY FOR QUERY EXAMPLES**



The screenshot shows the Microsoft Azure Synapse Analytics workspace interface. The left sidebar displays the Data section with a workspace named "census-malaga-sy". Under "Lake database", there is a "CensusDB" database containing a "CensusCombined" table and its columns. The main area shows a SQL script editor with the following code:

```

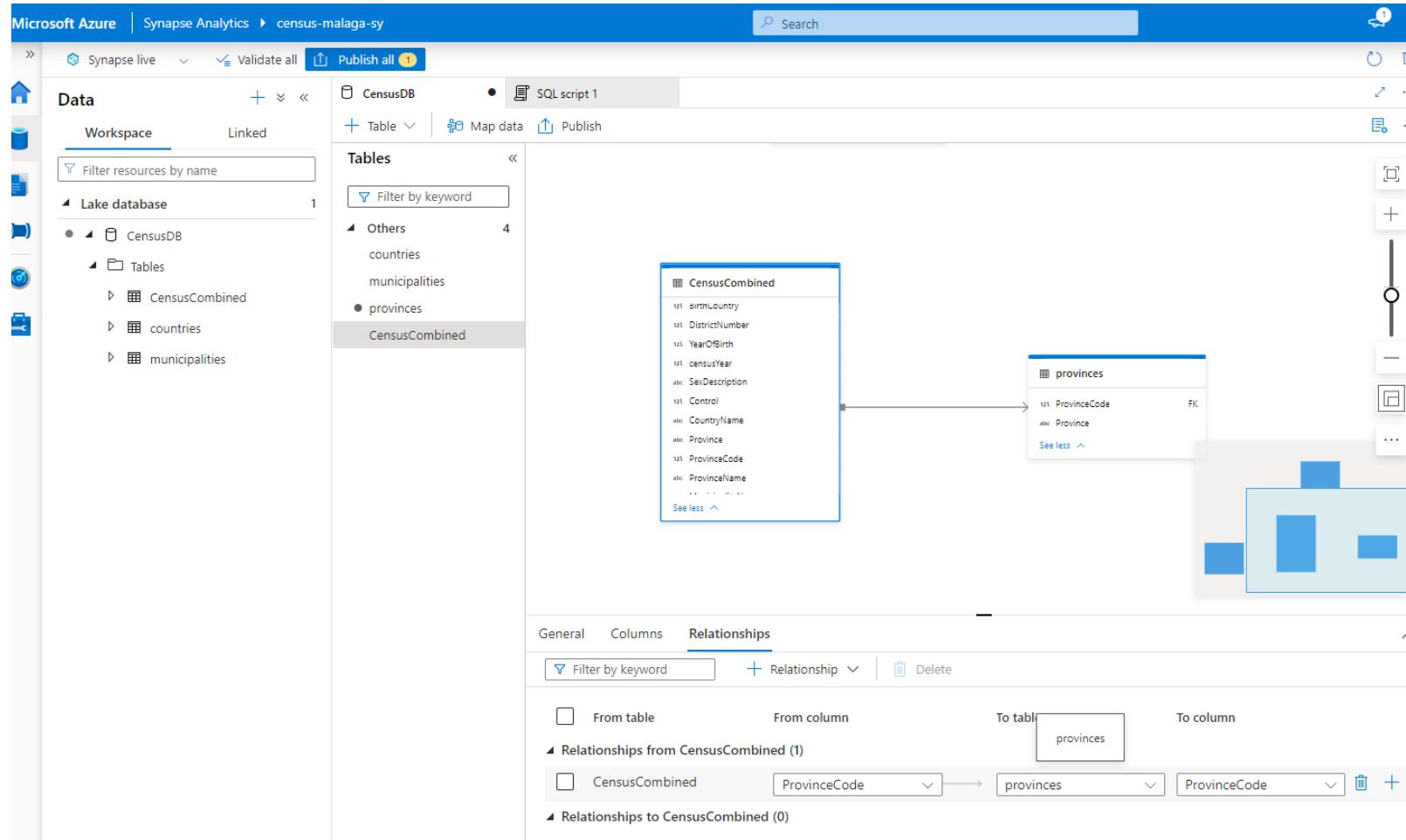
12  --
13  SELECT censusYear, COUNT(*) as Total
14  FROM CensusCombined
15  GROUP BY censusYear;
16
17  --
18  --
19  SELECT
20      AgeGroup,
21      COUNT(*) as Total
22  FROM (
23      SELECT
24          CASE
25              WHEN Age BETWEEN 0 AND 17 THEN '0-17'
26              WHEN Age BETWEEN 18 AND 64 THEN '18-64'
27              WHEN Age >= 65 THEN '65+'
28          END as AgeGroup
29      FROM CensusCombined
30  ) as AgeRanges
31  GROUP BY AgeGroup;
32
33

```

The "Results" tab is selected, showing the output of the query:

AgeGroup	Total
18-64	53010013
0-17	16379980
65+	16115378

- Just showing how to create relationships between tables

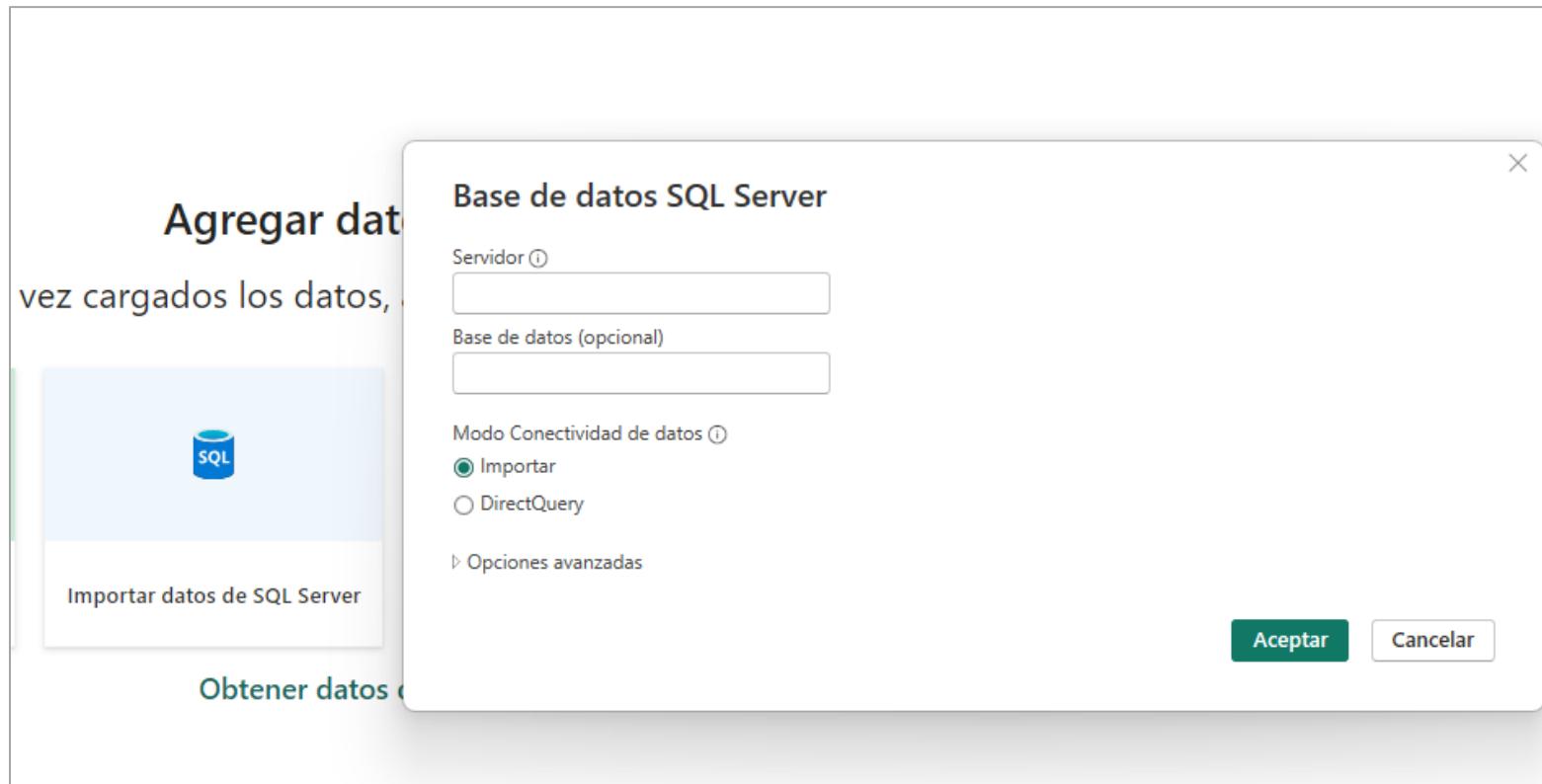


- All the tables created
- These tables will be used in PowerBI thanks to the Synapse connector

The screenshot shows the Microsoft Azure Synapse Analytics Data workspace interface. The left sidebar displays the 'Data' section under 'Workspace'. A tree view shows a 'Lake database' named 'CensusDB' containing several tables: 'CensusCombined', 'countries', 'municipalities', 'geoFinal', and 'provinces'. The 'Tables' section on the right lists these tables with their respective columns. The 'CensusCombined' table includes columns like SheetNumber, Age, Sex, MunicipalityCode, BirthCountry, DistrictNumber, YearOfBirth, censusYear, SexDescription, and Control. The 'countries' table includes CountryCode, Control, and CountryName. The 'municipalities' table includes ProvinceCode, ProvinceName, MunicipalityCode, and MunicipalityName. The 'geoFinal' table includes FeatureID, DistrictID, DistrictNumber, DistrictName, CommonDistrictName, Geometry, Area, Perimeter, X, and Y. The 'provinces' table includes ProvinceCode and Province.

Table	Columns
CensusCombined	SheetNumber, Age, Sex, MunicipalityCode, BirthCountry, DistrictNumber, YearOfBirth, censusYear, SexDescription, Control
countries	CountryCode, Control, CountryName
municipalities	ProvinceCode, ProvinceName, MunicipalityCode, MunicipalityName
geoFinal	FeatureID, DistrictID, DistrictNumber, DistrictName, CommonDistrictName, Geometry, Area, Perimeter, X, Y
provinces	ProvinceCode, Province

- Open PowerBI Desktop
- Create new report
- Get data from Azure Synapse analytics
- You will see this popup window below
- On the server blank paste the Dedicated SQL endpoint of your Synapse workspace
- On the database blank paste the database name on synapse workspace



- Go to Synapse workspace
- Copy the Dedicated SQL endpoint and paste it on the server blank

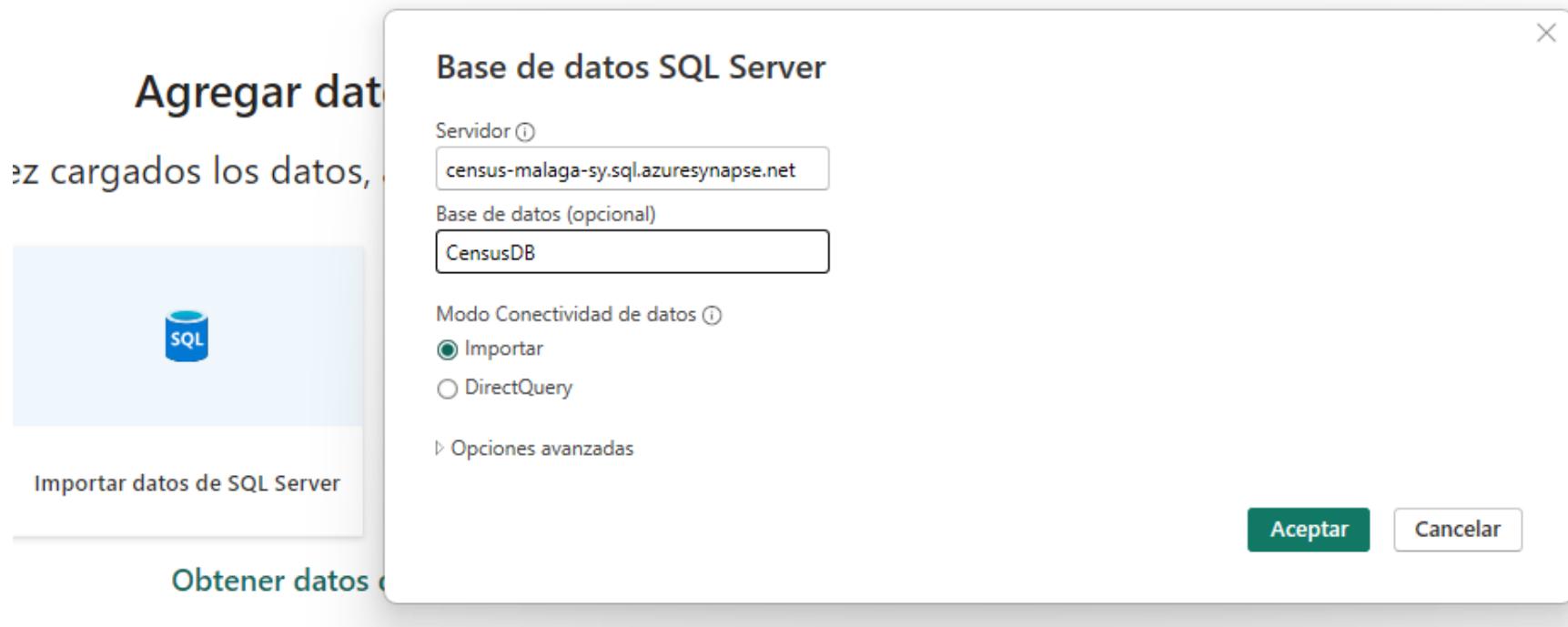
census-malaga-sy Synapse workspace

Search Activity log New dedicated SQL pool New Apache Spark pool New Data Explorer pool (preview) Refresh Reset SQL admin password Delete

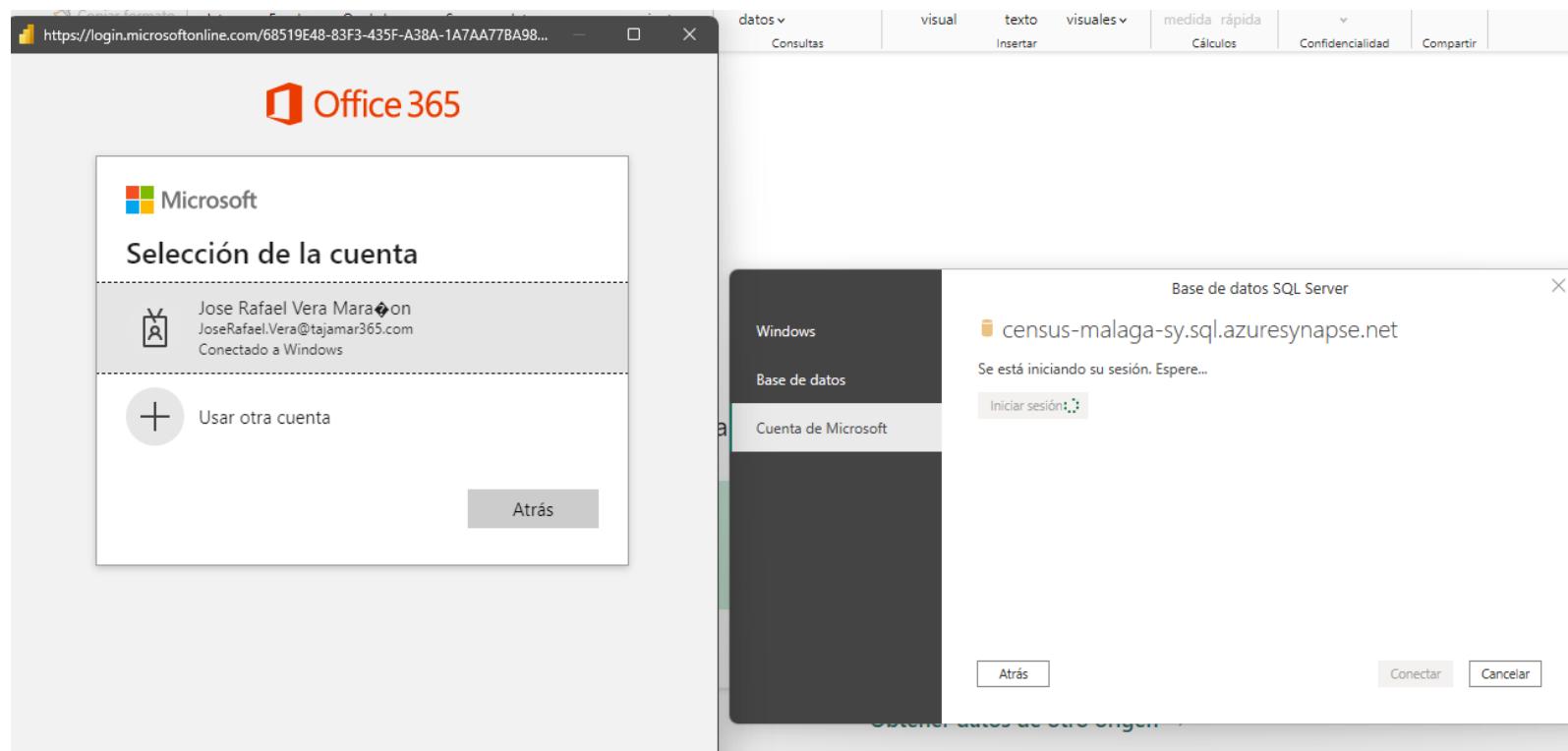
Essentials

Resource group (move)	:	census-malaga-4years	Networking	:	Show firewall settings
Status	:	Succeeded	Primary ADLS Gen2 acco...	:	https://censusmalagadata.dfs.core.windows.net
Location	:	Southeast Asia	Primary ADLS Gen2 file s...	:	census-malaga-data
Subscription (move)	:	Azure for Students	SQL admin username	:	sqladminuser
Subscription ID	:	9e3c713b-0d42-41e5-81ca-d778fd001da7	SQL Microsoft Entra admin	:	JoseRafael.Vera@tajamar365.co Copy to clipboard
Managed virtual network	:	No	Dedicated SQL endpoint	:	census-malaga-sy.sql.azureSynapse.net Copy
Managed Identity object ...	:	395a06d4-7035-4107-831d-a4337fb23f5a	Serverless SQL endpoint	:	census-malaga-sy-onDemand.sql.azureSynapse.net
Workspace web URL	:	https://web.azureSynapse.net?workspace=%2bsubscriptions%2f9e3c713b-0d42-41e5-81ca-d778fd00...	Development endpoint	:	https://census-malaga-sy.dev.azureSynapse.net

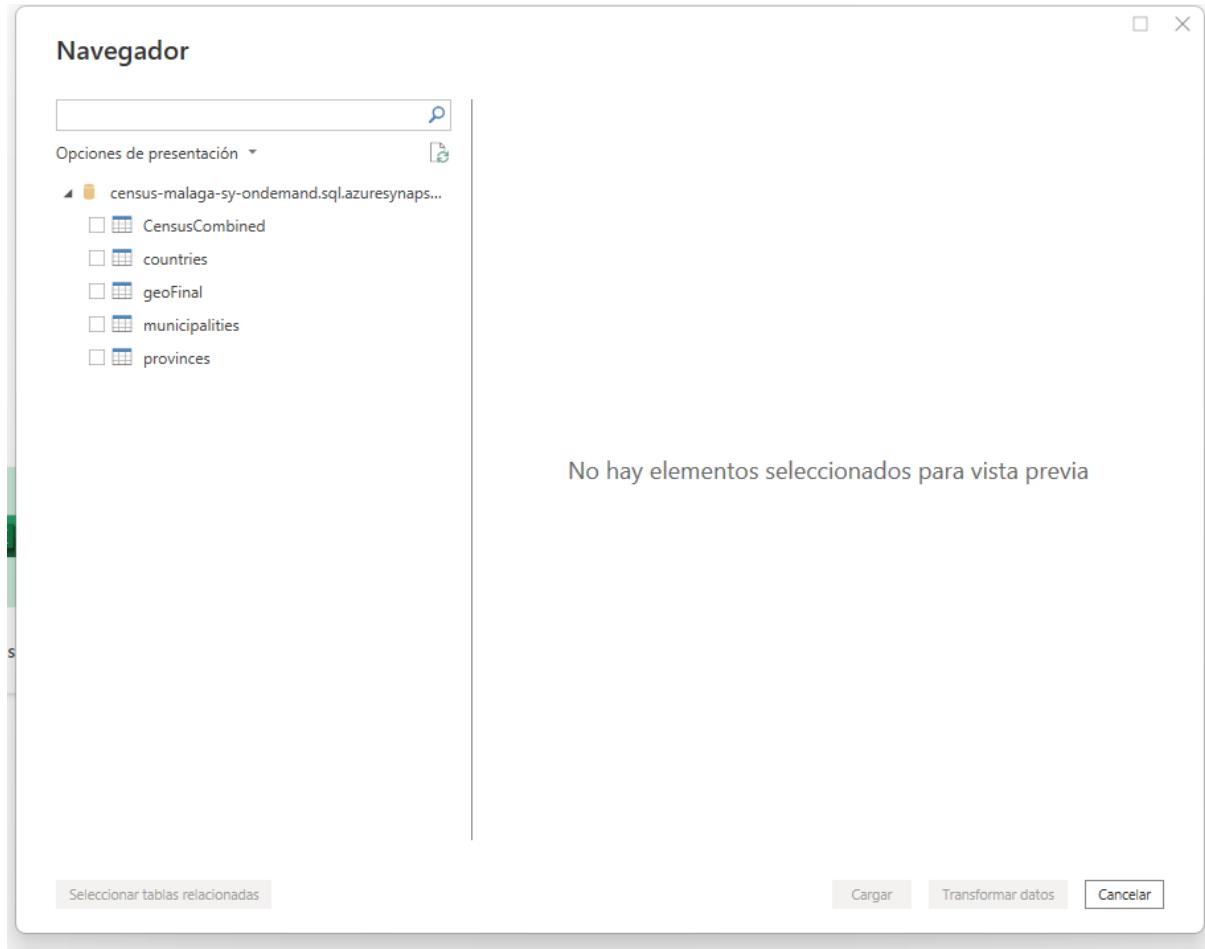
- You should see something like this
- Is up on you to select directquery (for realtime actualization) or Import (save a table snapshot into PBI)
- Accept



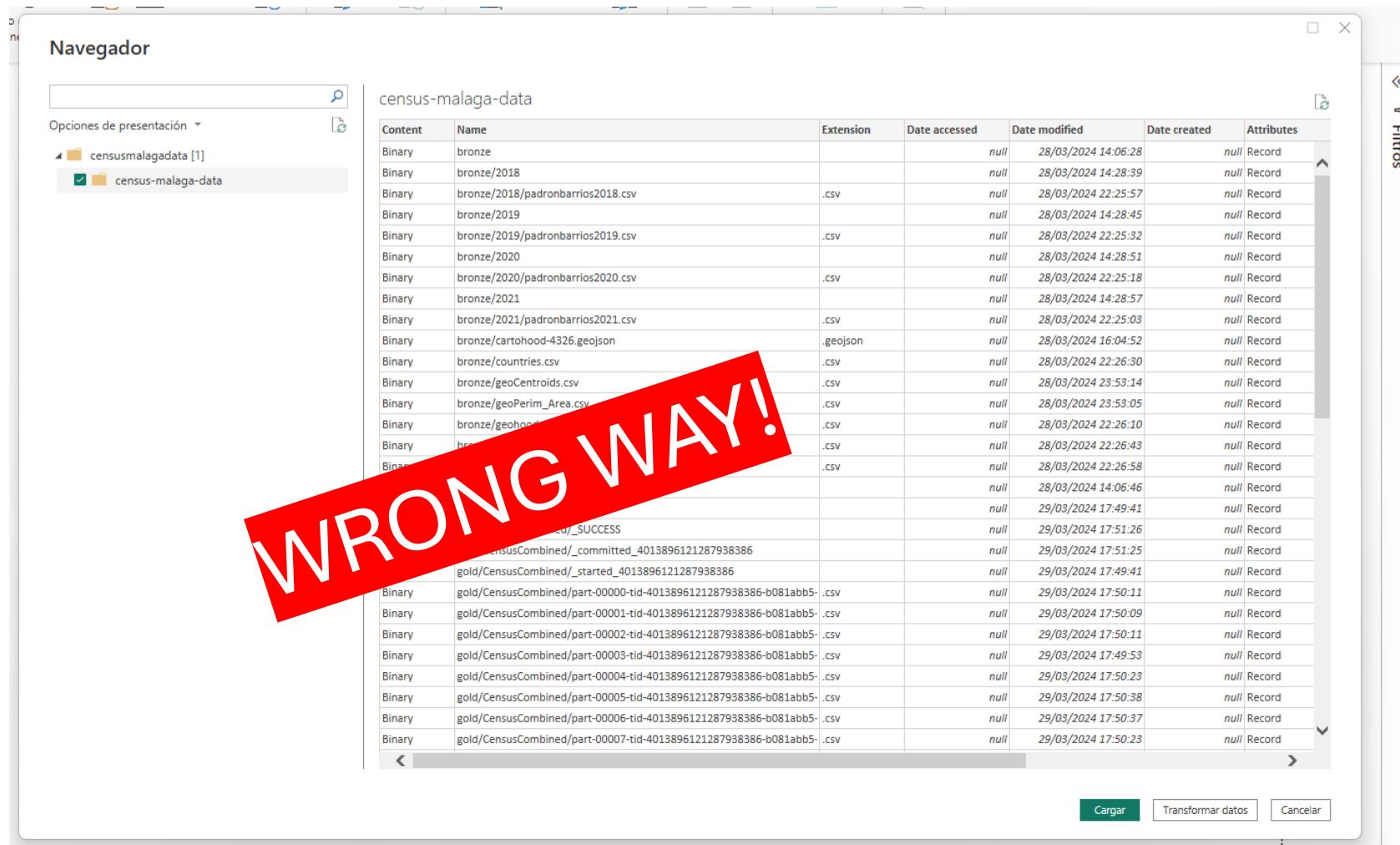
- After you need to be authenticated.
- I used the Microsoft account credentials



- Once authenticated you will be able to see the tables of your Synapse workspace
- Select the tables
- Load the tables
- Start creating visualizations



- I tried the Azure Data Lake Gen2 connector and blobStorage connector
- Those were useless for me as I didn't know how to select those csv and partitions from the Gold layer
- As you can see below once connected you get all the files



- DELETE ALL THE CREATED SERVICES TO AVOID BILLING OR CONSUME YOUR FREE CREDITS
- DELETE RESOURCE GROUP
- DELETE SYNAPSE WORKSPACE
- DELETE DATABRICKS WORKSPACE
- DELETE APP REGISTRATION
- DELETE STORAGE ACCOUNT
- DELETE EVERYTHING