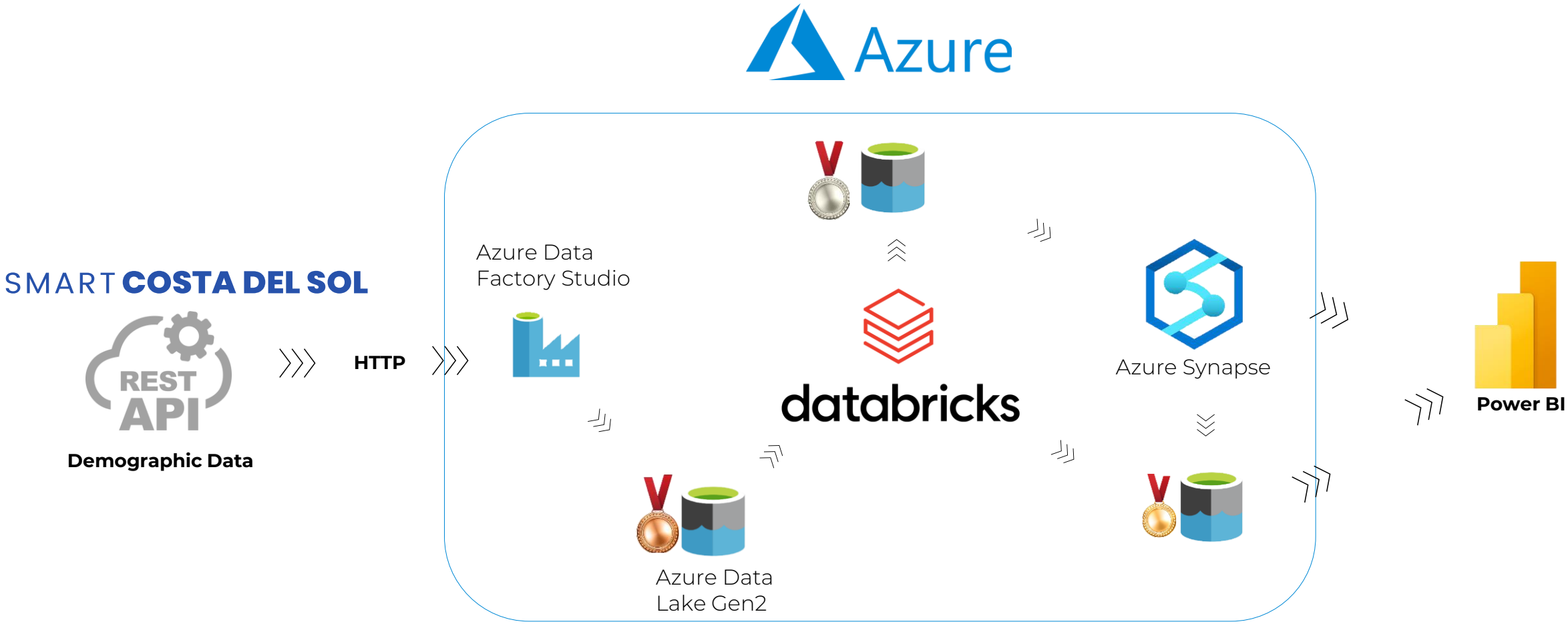
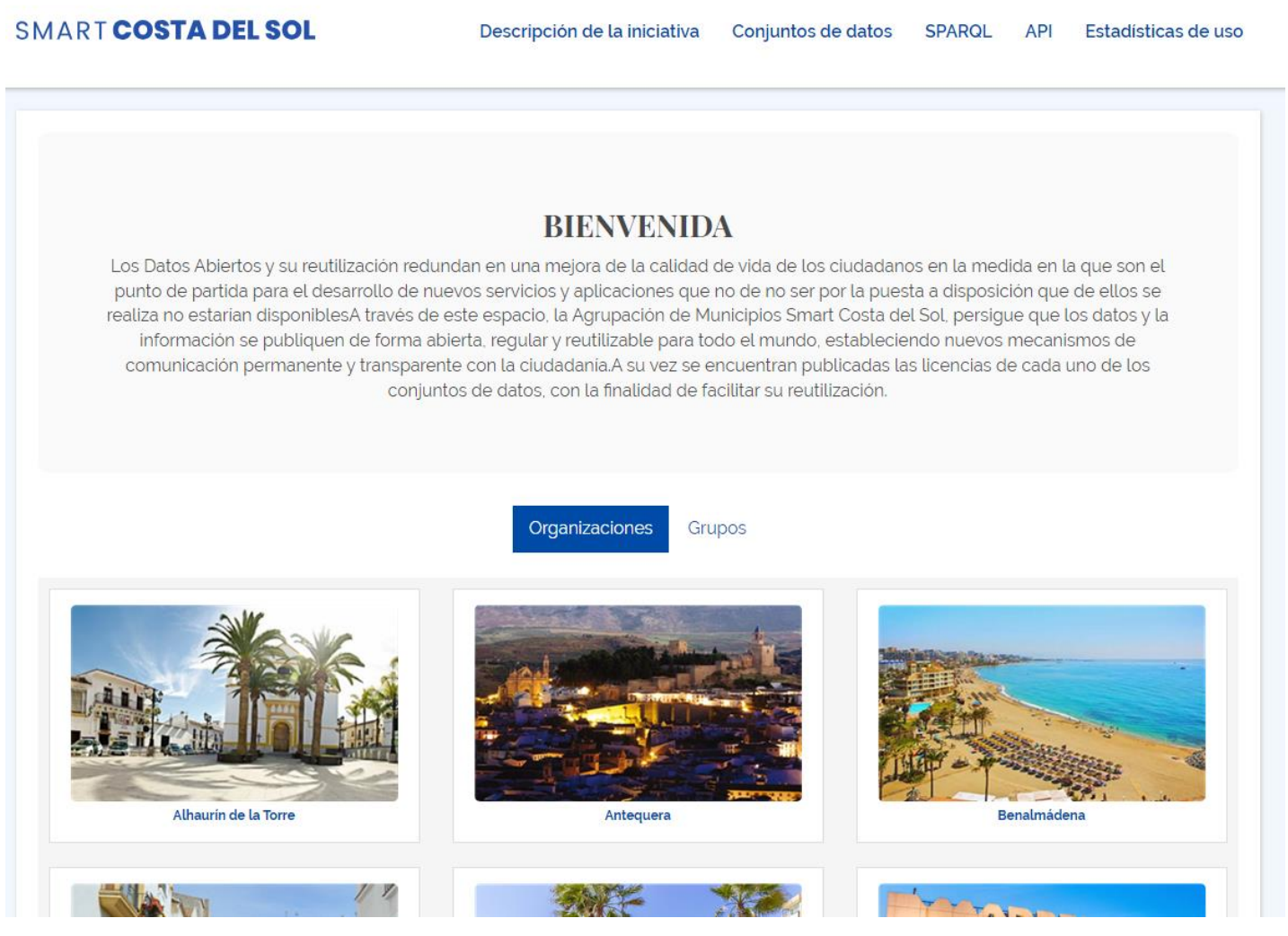
# Azure Databricks End-to-End Data Engineering Pipeline

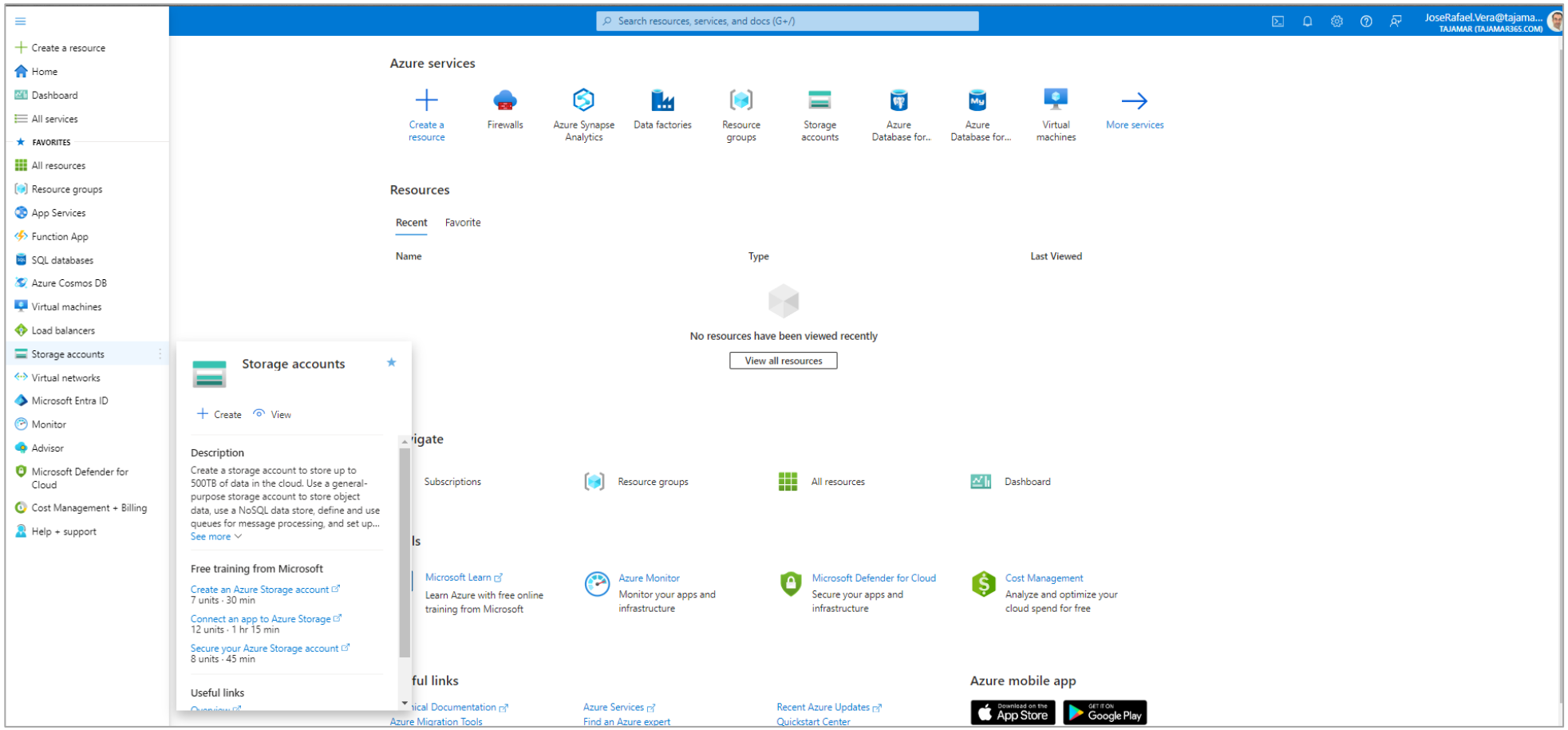


- API data source – Smart Costa del Sol – Demographic Data https://catalogo.smartcostadelsol.es/

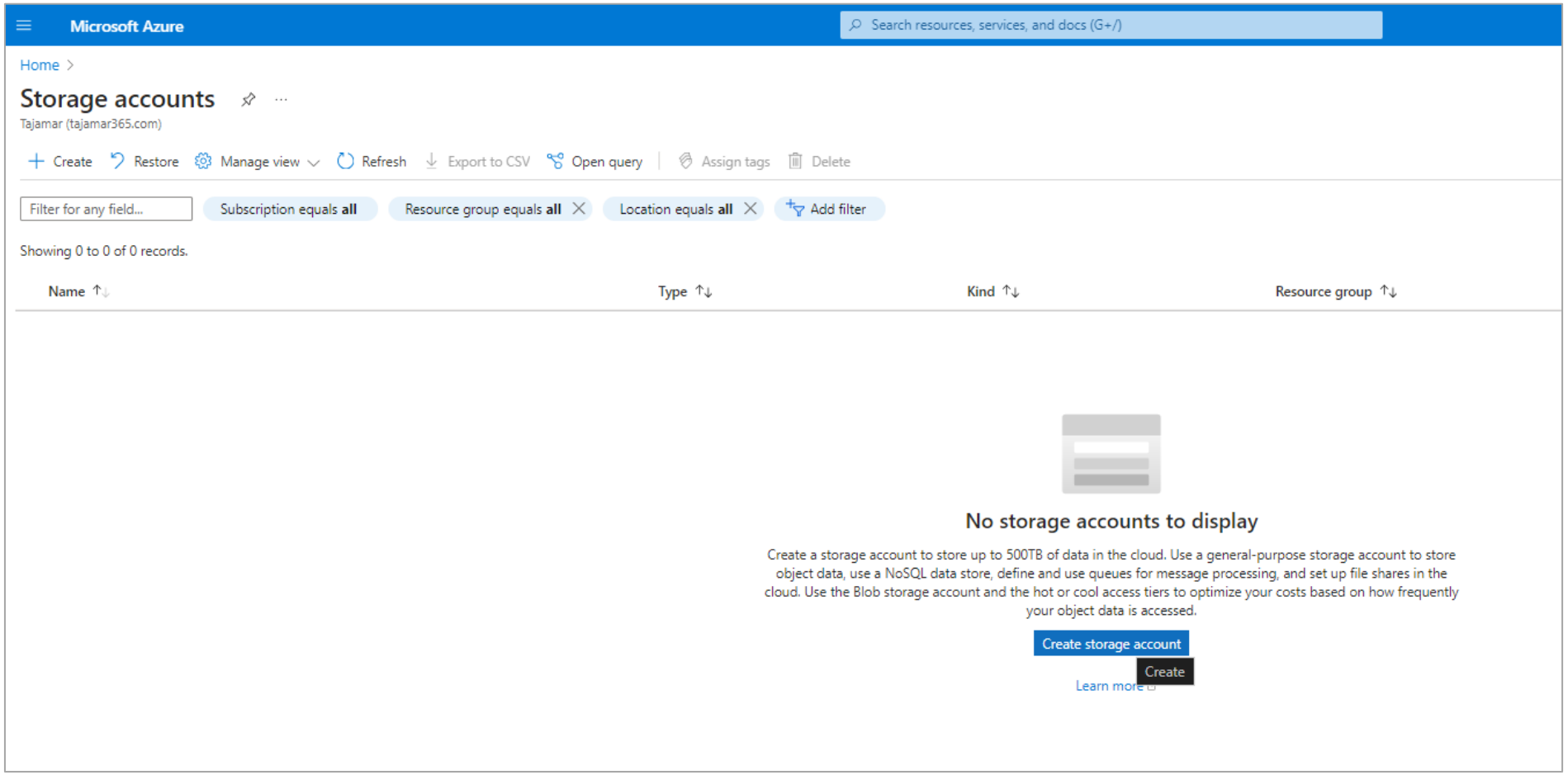


- API light instructions
- These are endpoints for listing packages
- I’ve explored them with postman (SEE POSTMAN COLLECTION IN MY REPOSITORY)
- Once located the package, take the endpoint url to get the data
    - Personally I find this API could be improved.

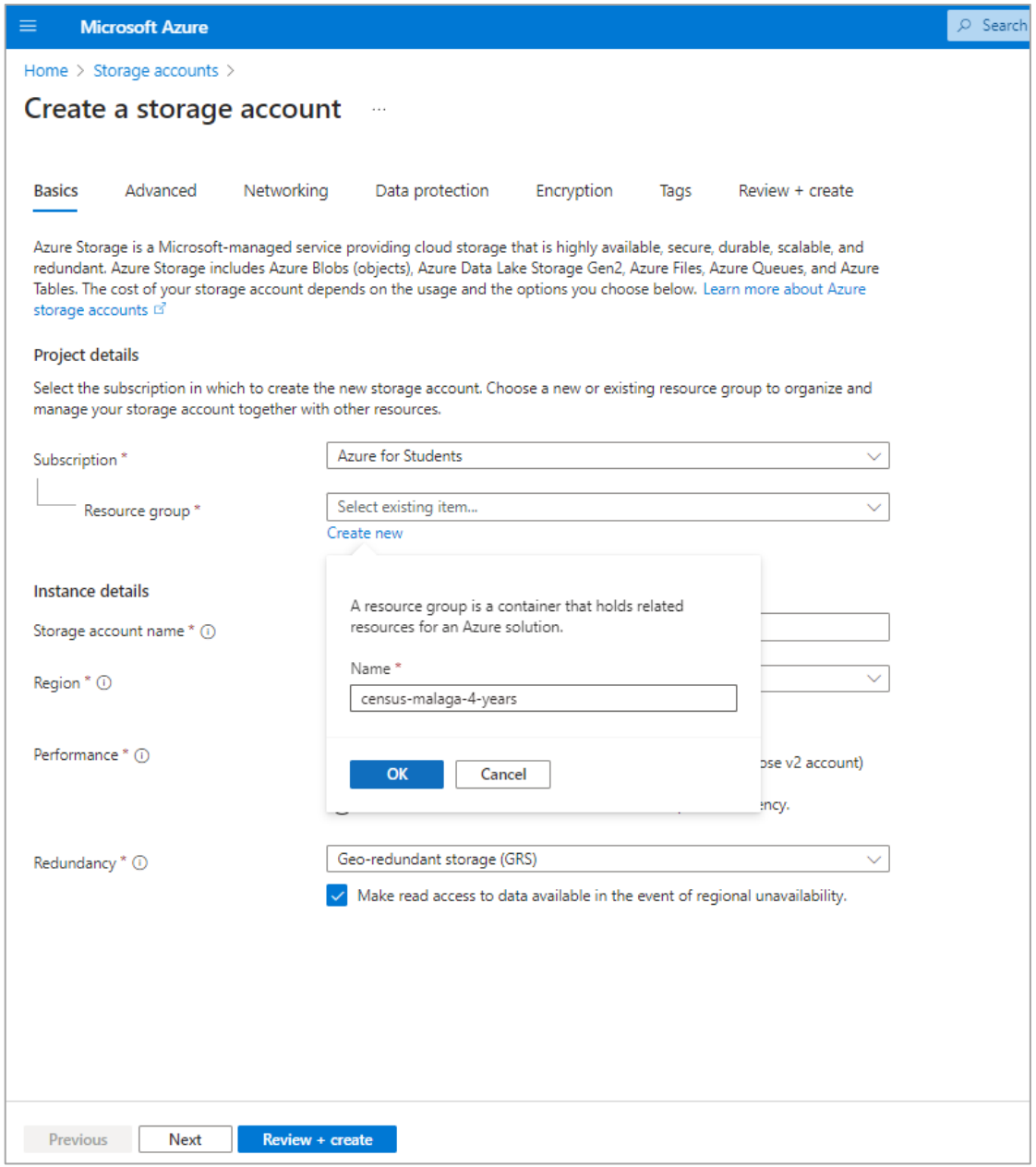
- Create a Storage Account



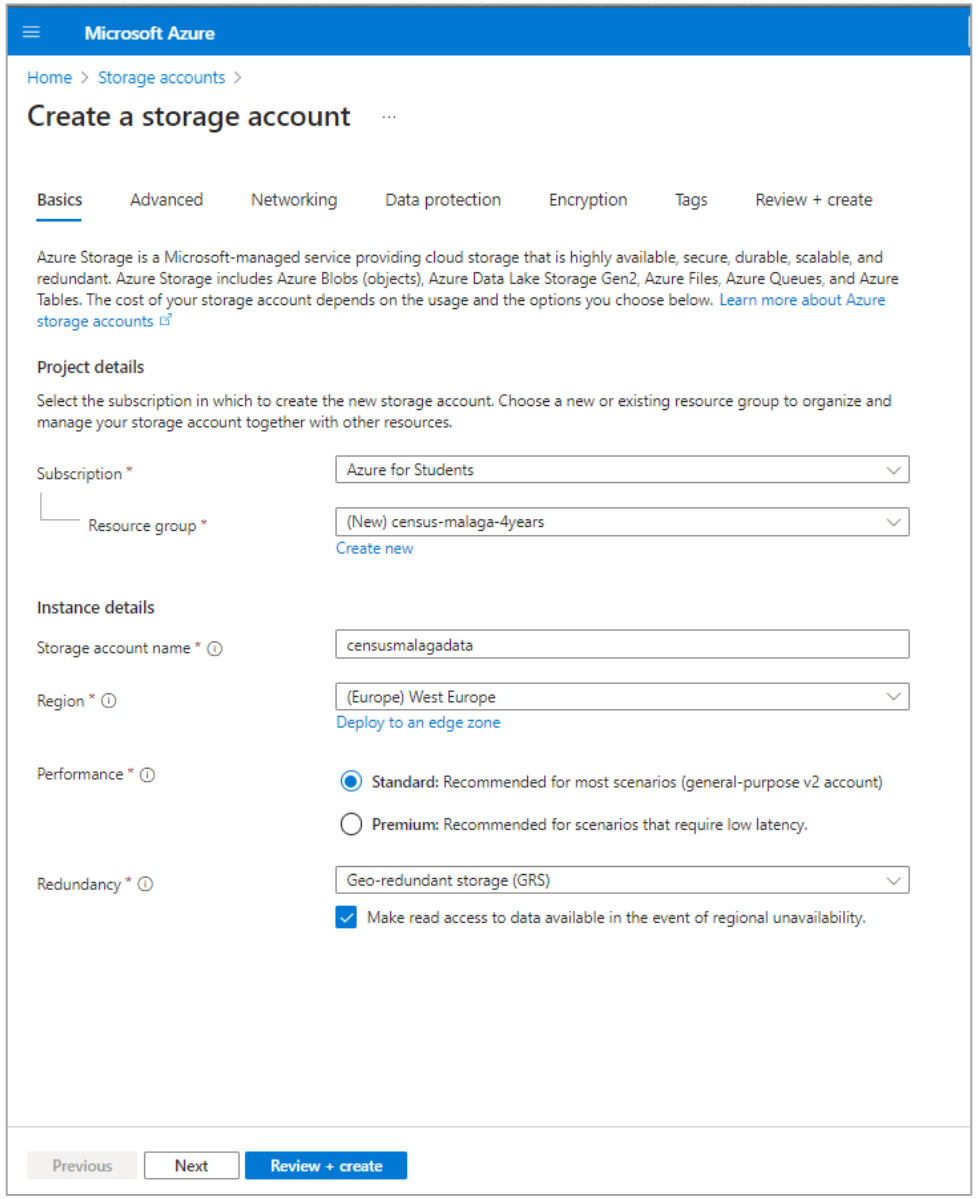
- Create a Storage Account



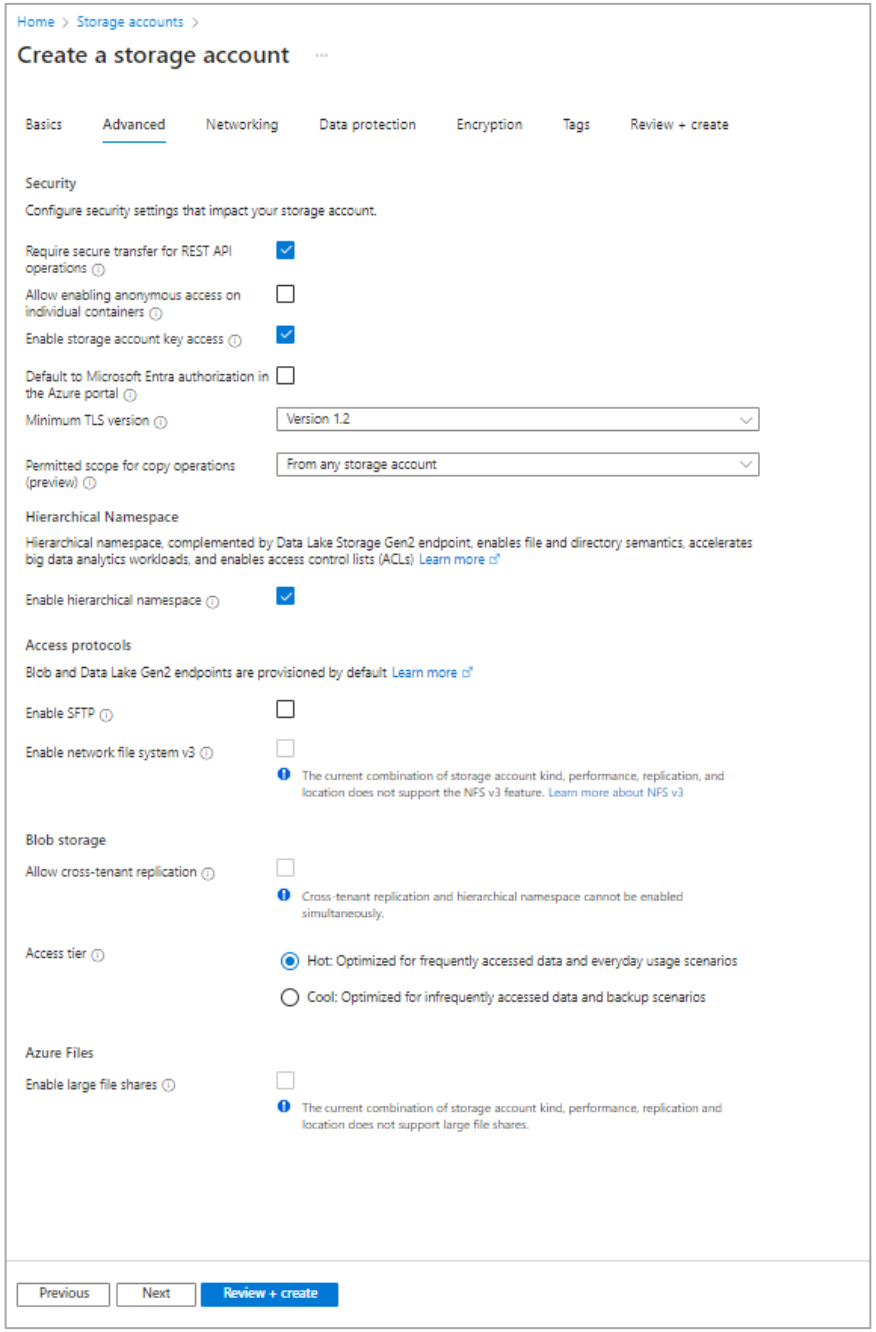
- Give your Free subscription.
- Create a Resource group. Just give it a name.



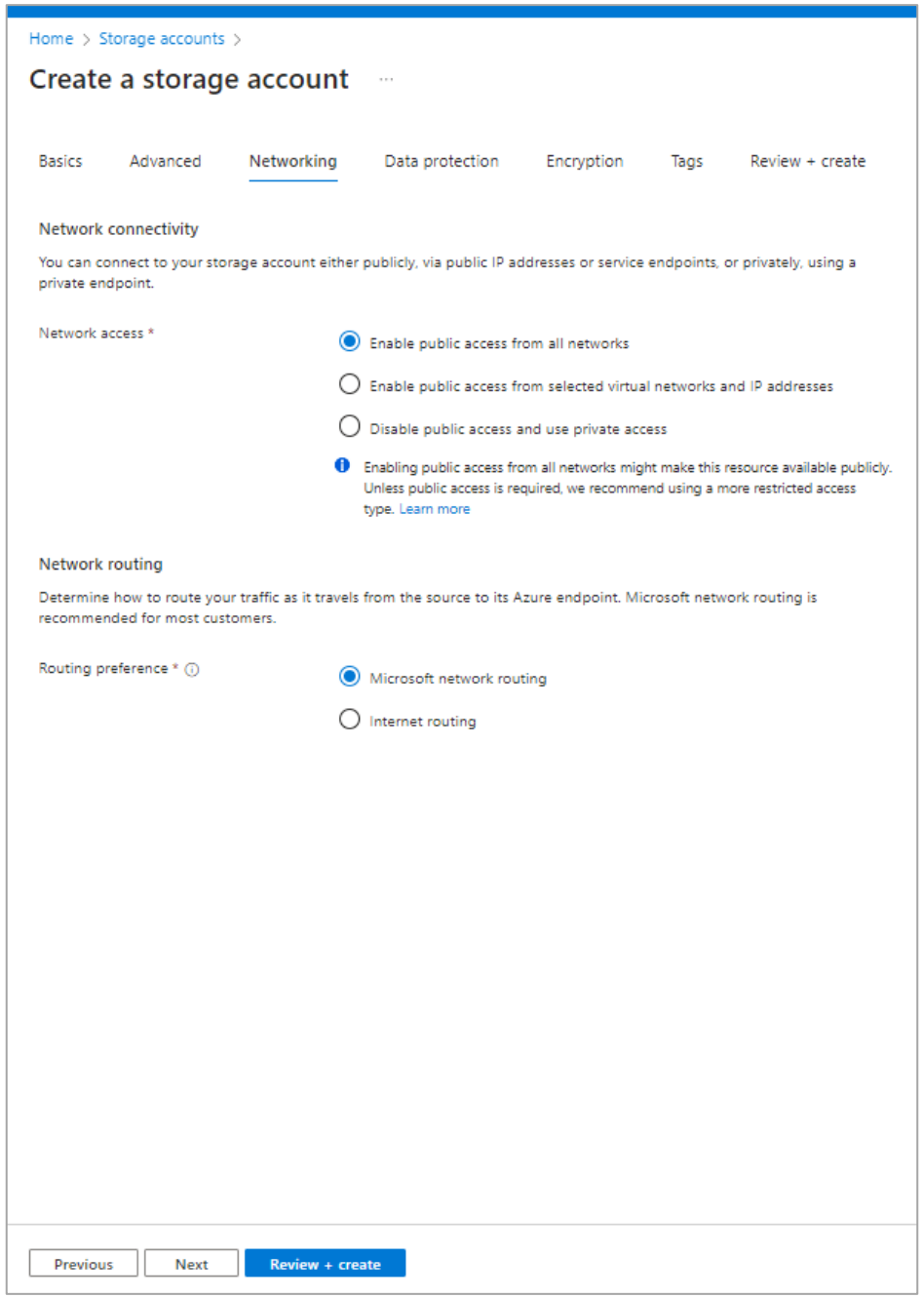
- Give the storage account a name
- Region: I highly recommend southEast Asia. It has more services and computing engine available
- Next



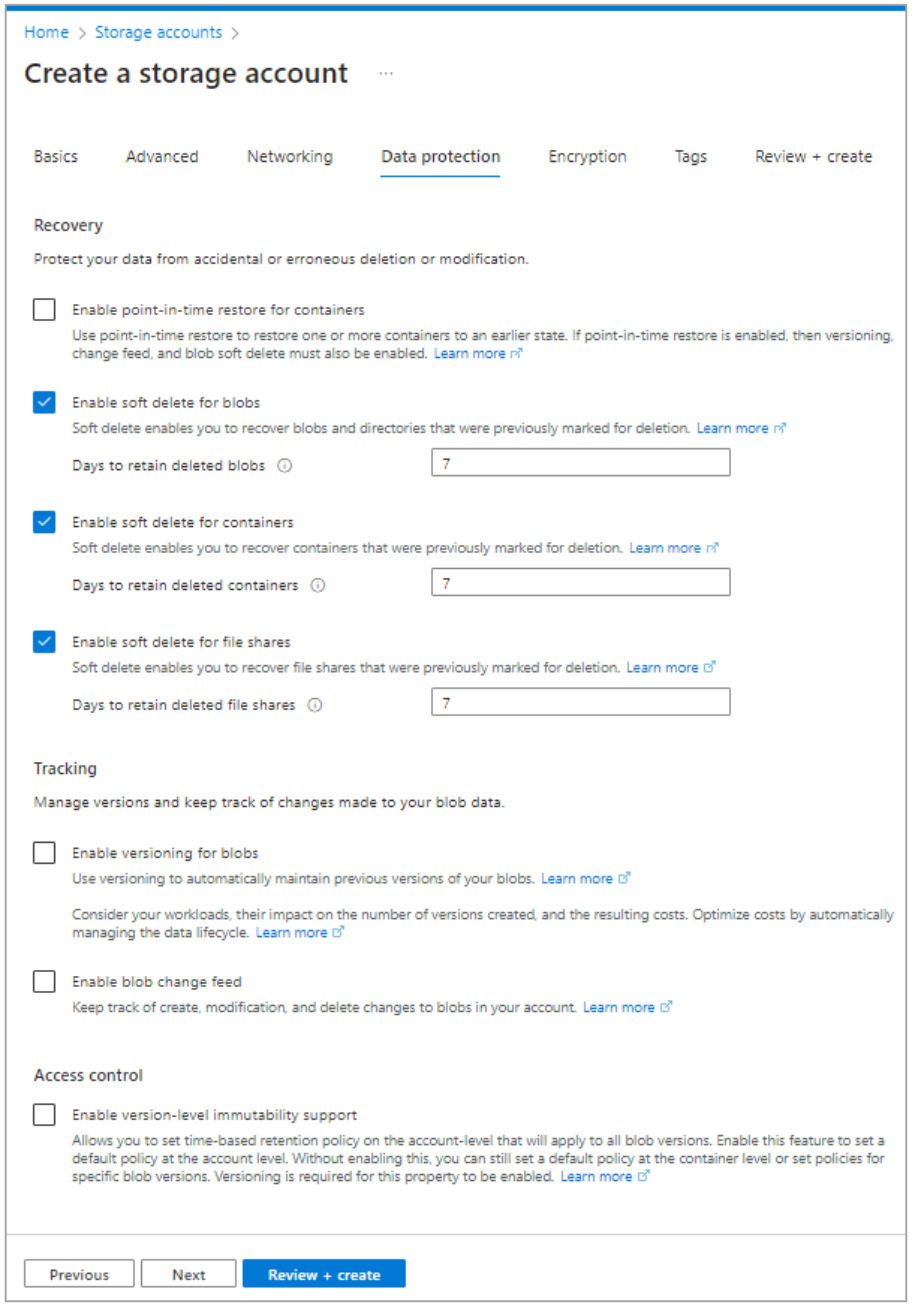
- Next


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
Version 1.2
Permitted scope for copy operations (preview)
From any storage account
Hierarchical Namespace
Hierarchical namespace, complemented by Data Lake Storage Gen2 endpoint, enables file and directory semantics, accelerates big data analytics workloads, and enables access control lists (ACLs) Learn more
Enable hierarchical namespace
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

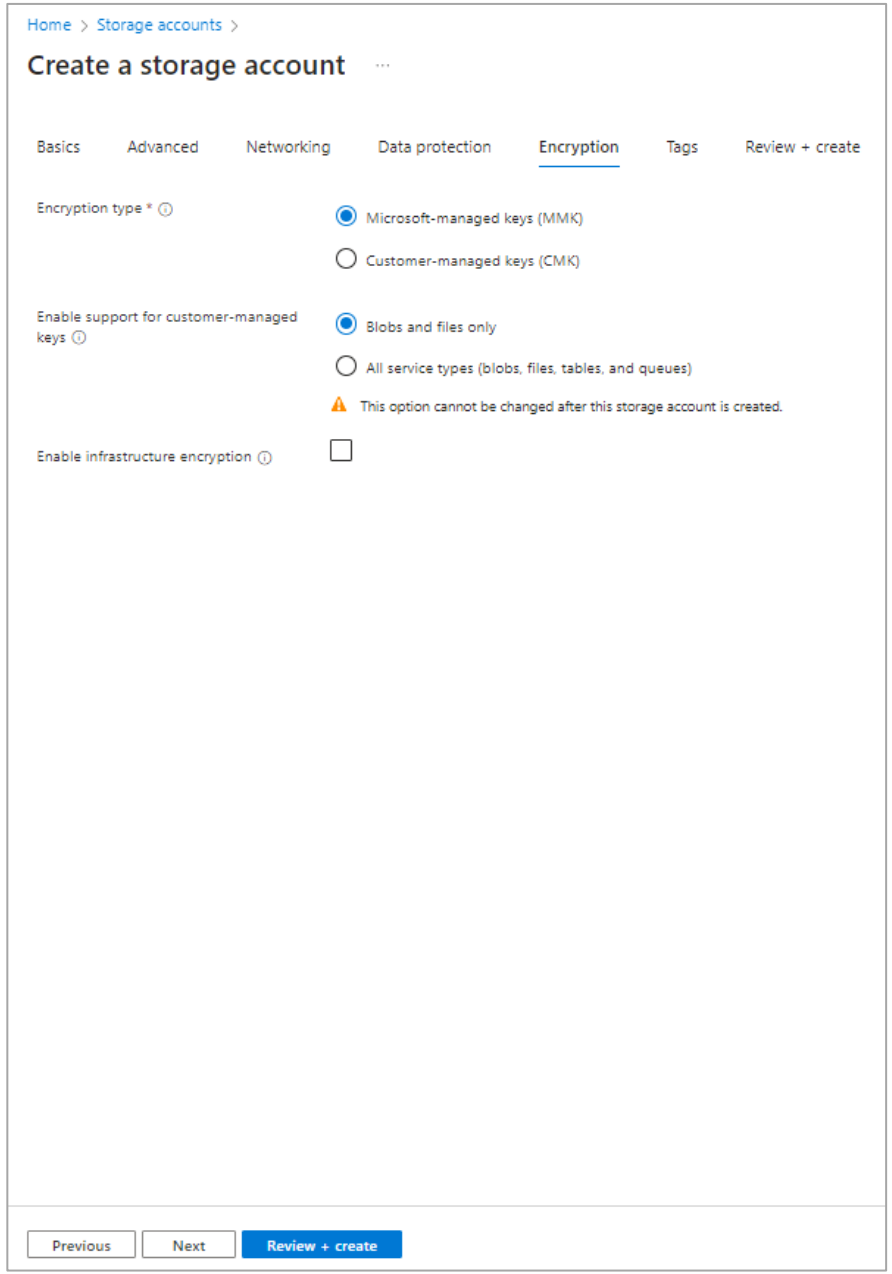
- Next


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

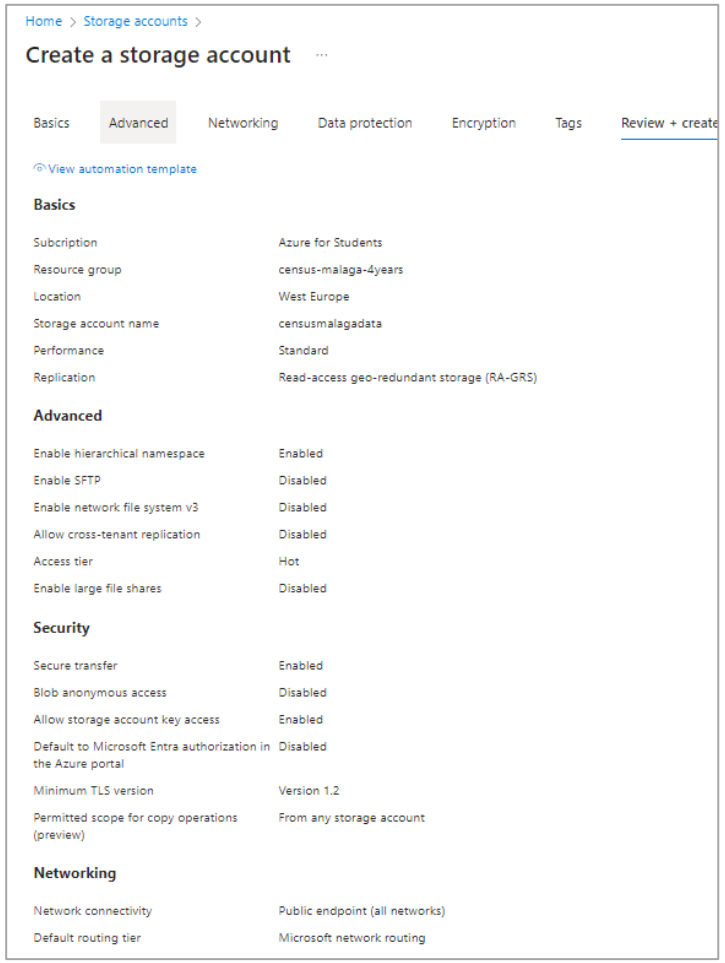
- Next

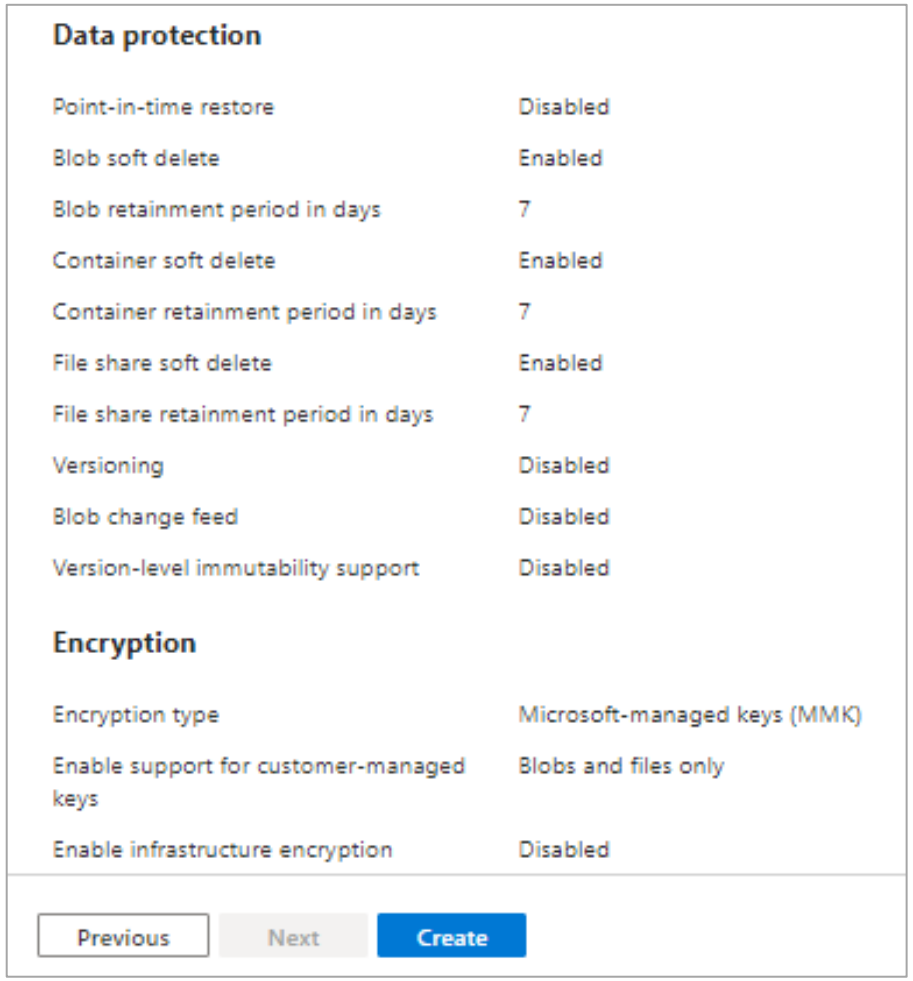

Home > Storage accounts >
Create a storage account
Basics
Advanced
Networking
Data protection
Encryption
Tags
Review + create
Recovery
Protect your data from accidental or erroneous deletion or modification.
Enable point-in-time restore for containers
Use point-in-time restore to restore one or more 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
7
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
7
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
7
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

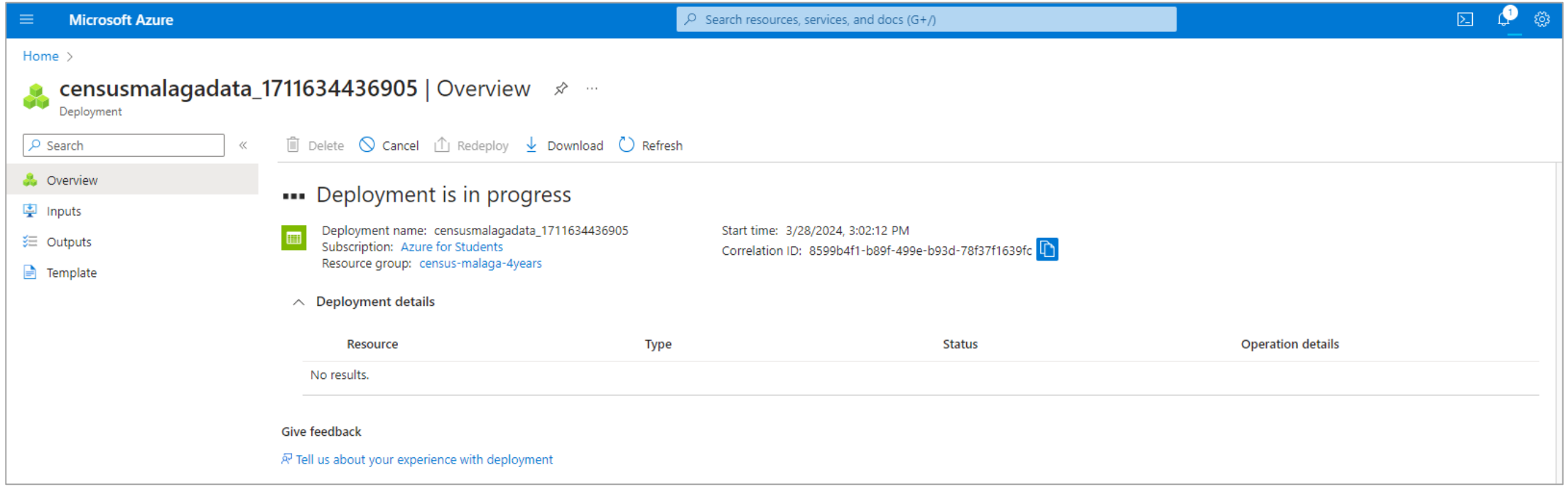

Home > Storage accounts >
Create a storage account
Basics
Advanced
Networking
Data protection
Encryption
Tags
Review + create
Encryption type *
Microsoft-managed keys (MMK)
Customer-managed keys (CMK)
Enable support for customer-managed keys
Blobs and files only
All service types (blobs, files, tables, 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

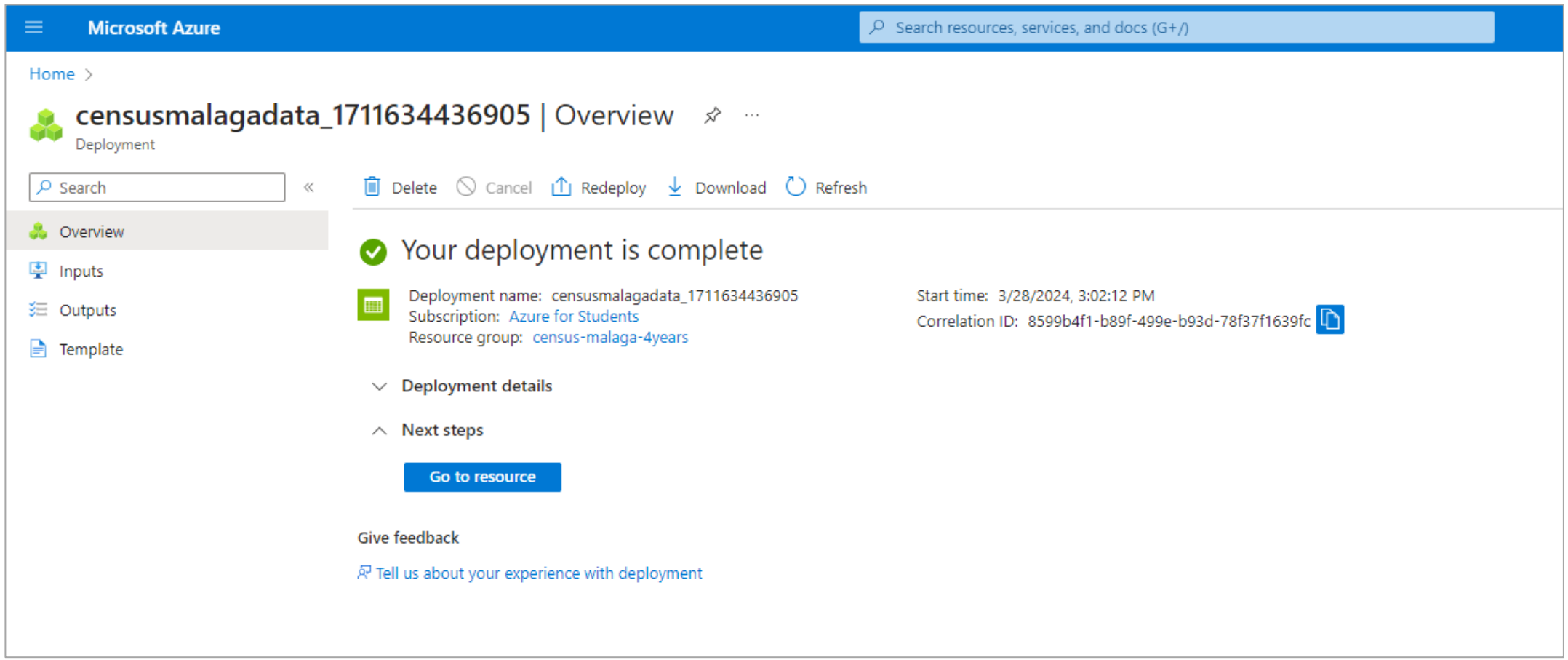


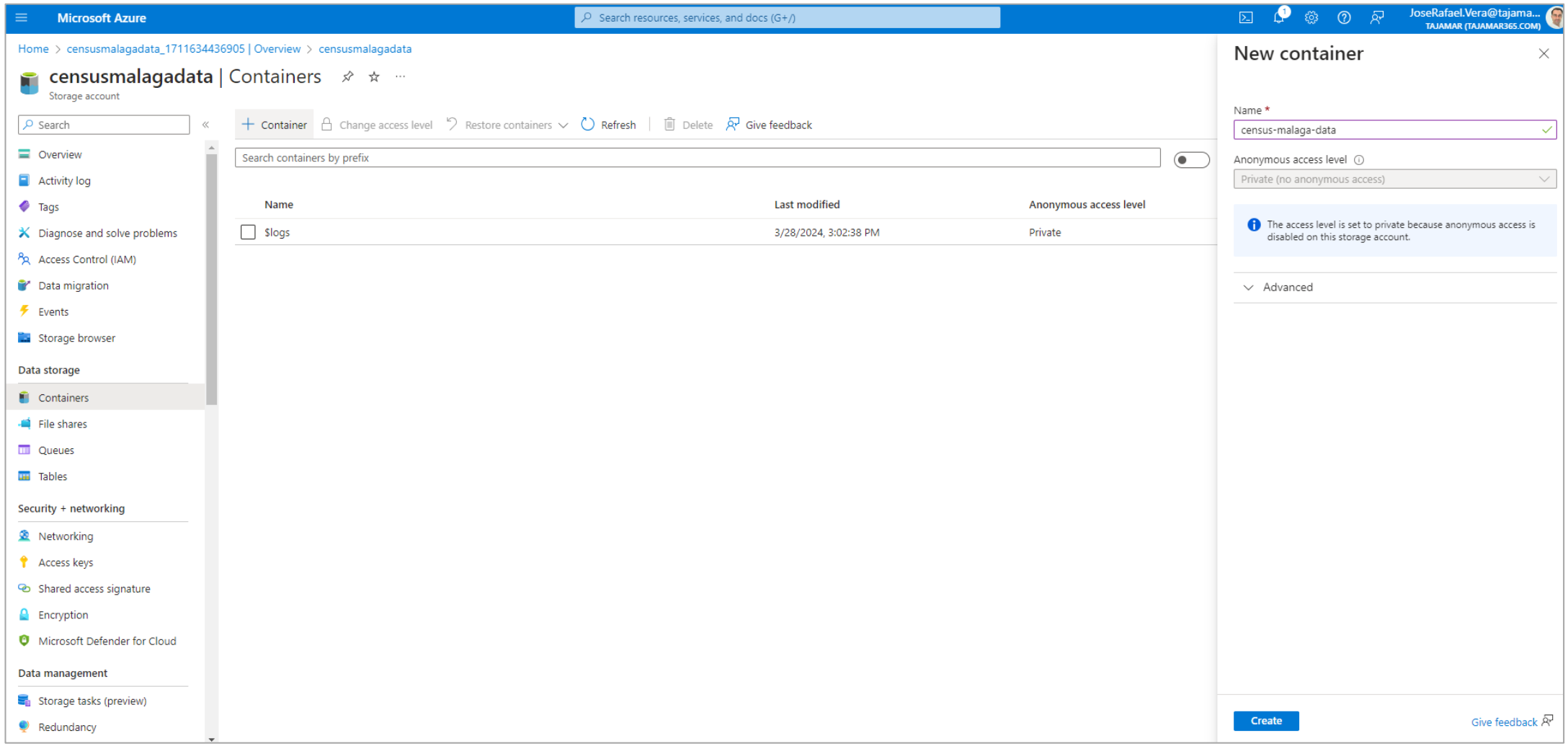
- Wait until deployment finishes



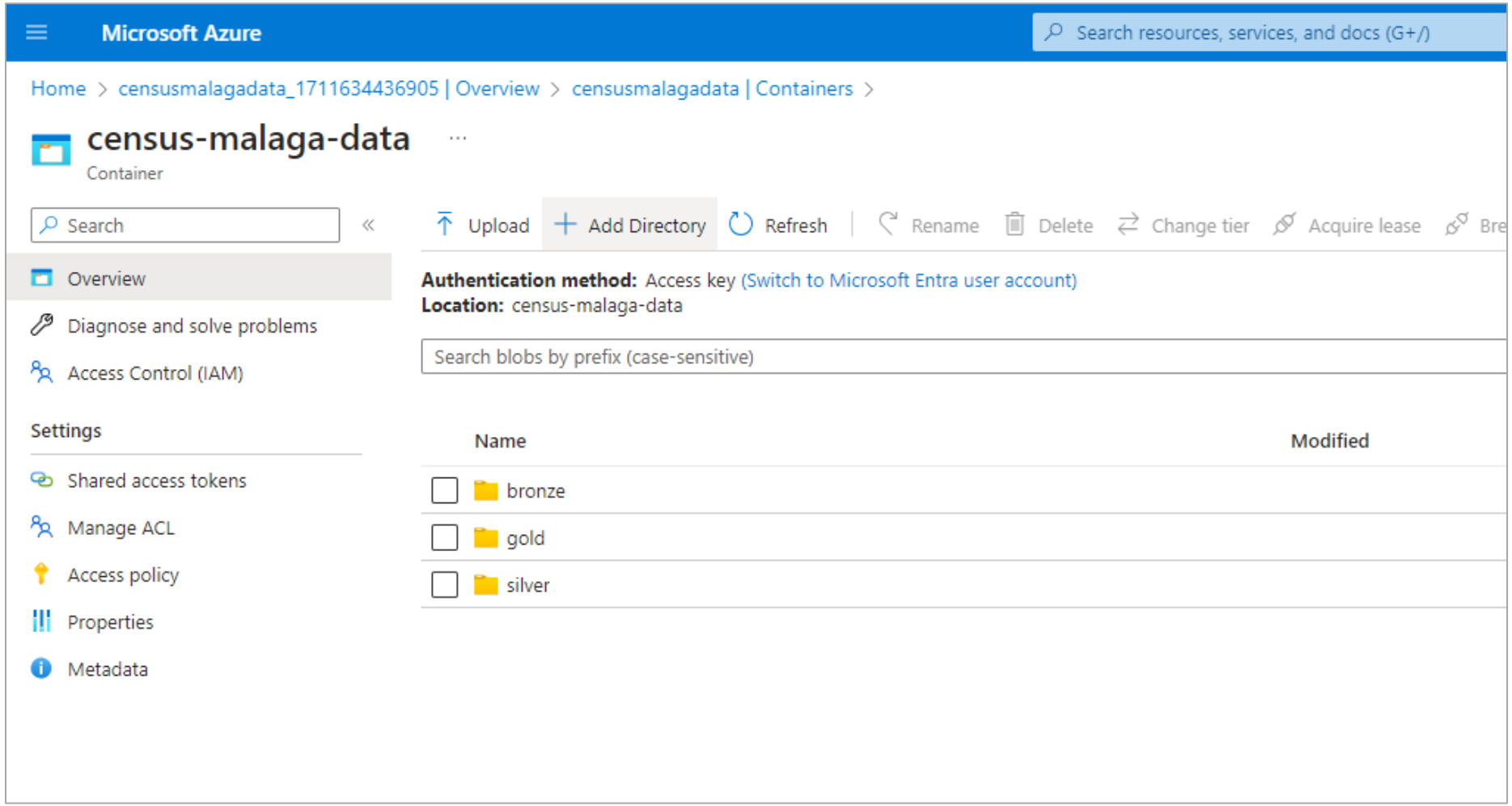
- Deployment completed
- Go to resource



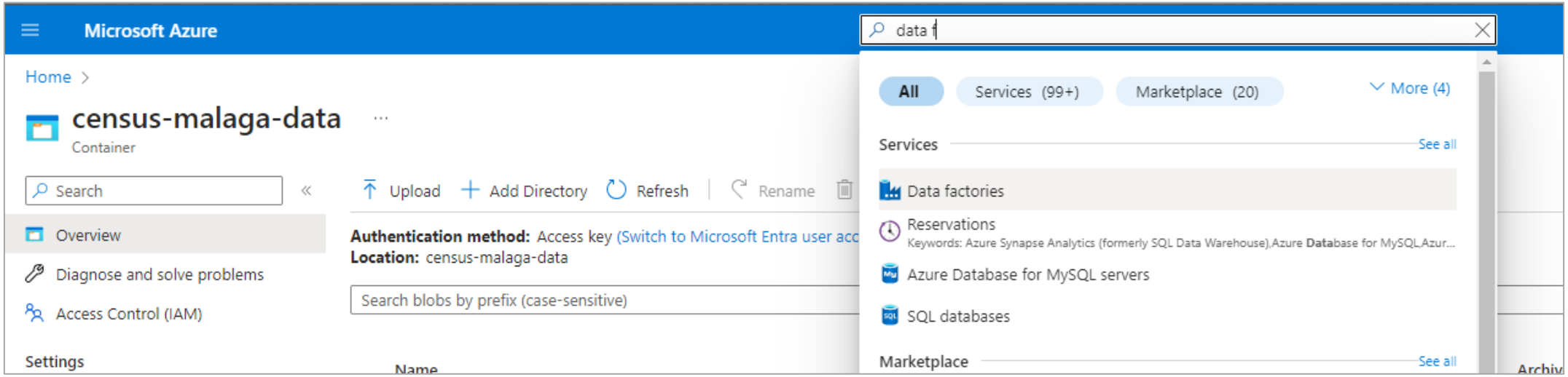
- Go to containers
- Create a container which will storage our data
- Just give it a name



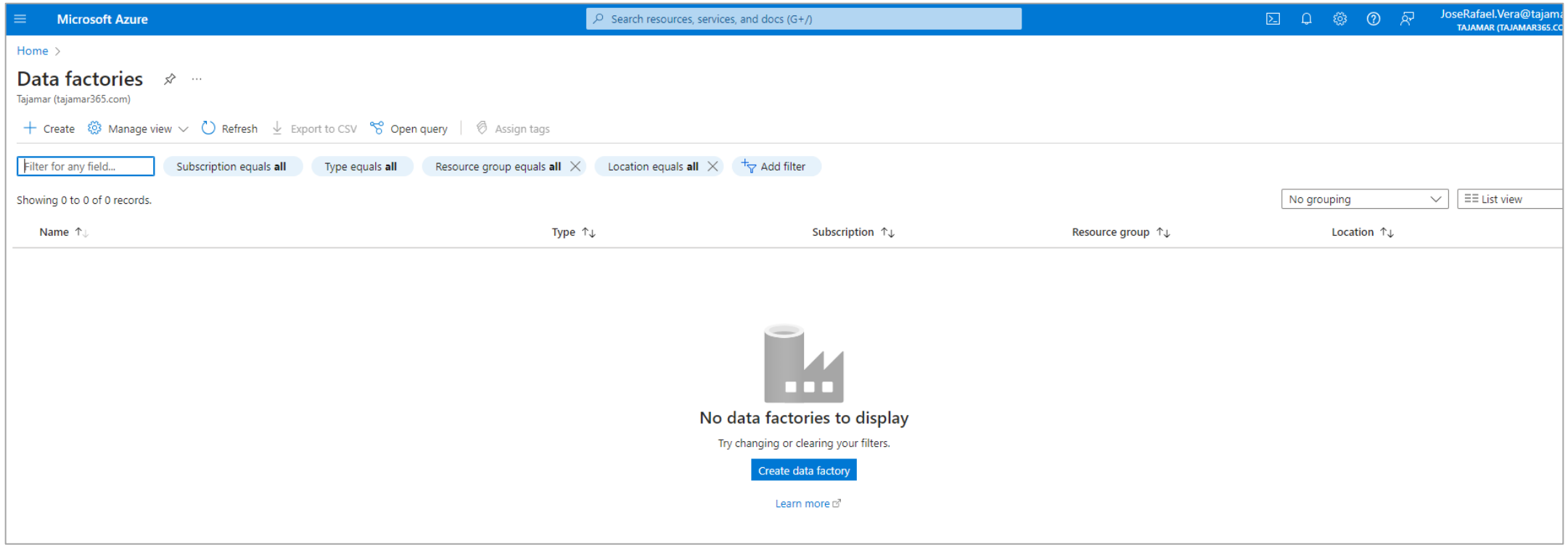
- Inside the container create the three directories



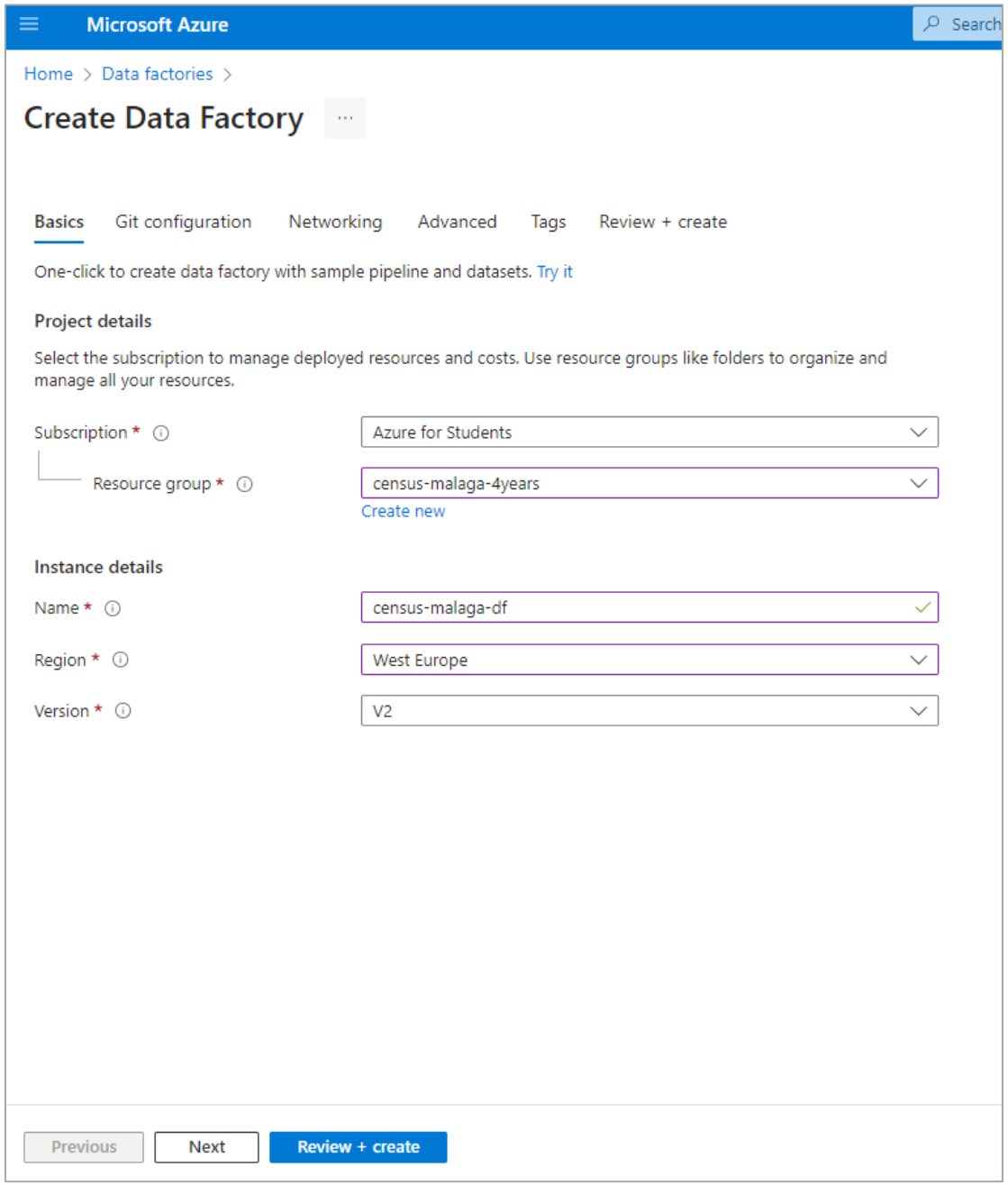
- Search for Data Factories



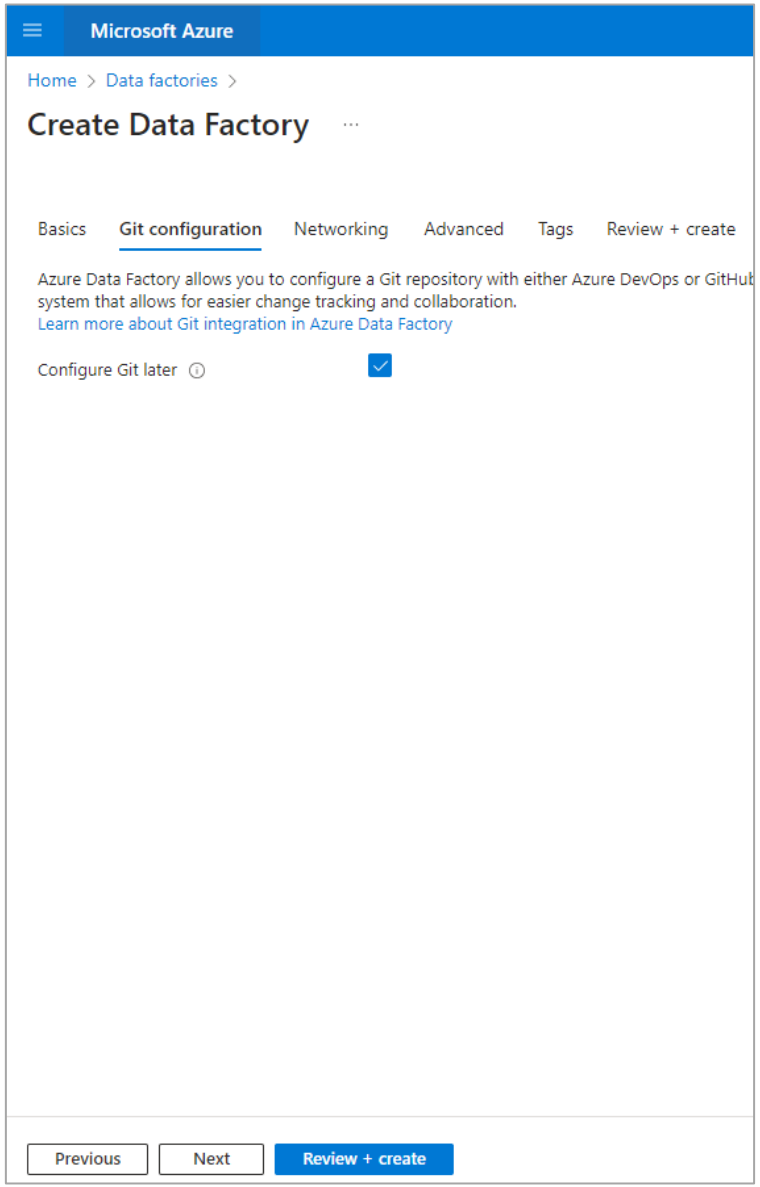
- Create new Data Factory



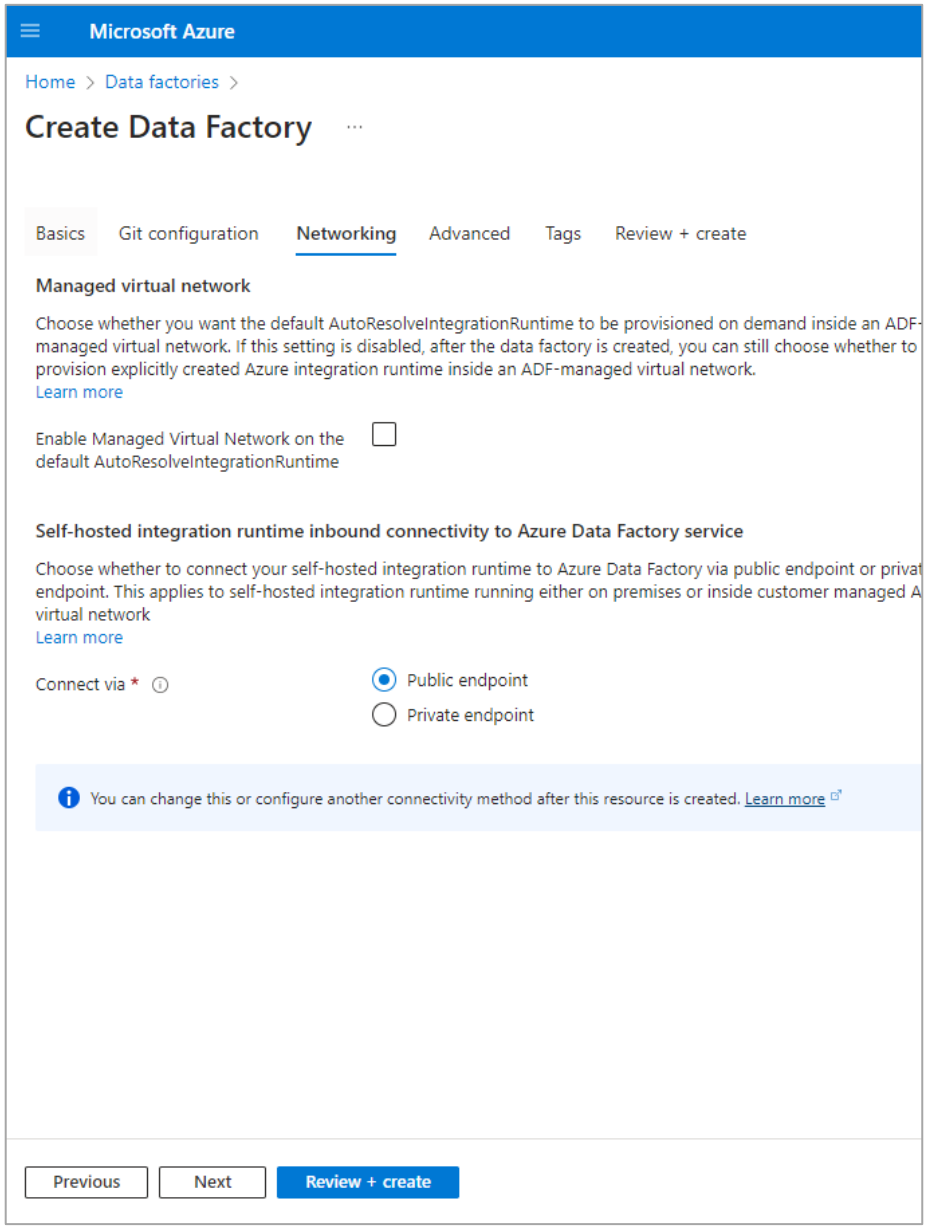
- Select your Free subscription
- Select your existing resource group
- Give the instance a name
- Select region southEast Asia. It works better than west Europe



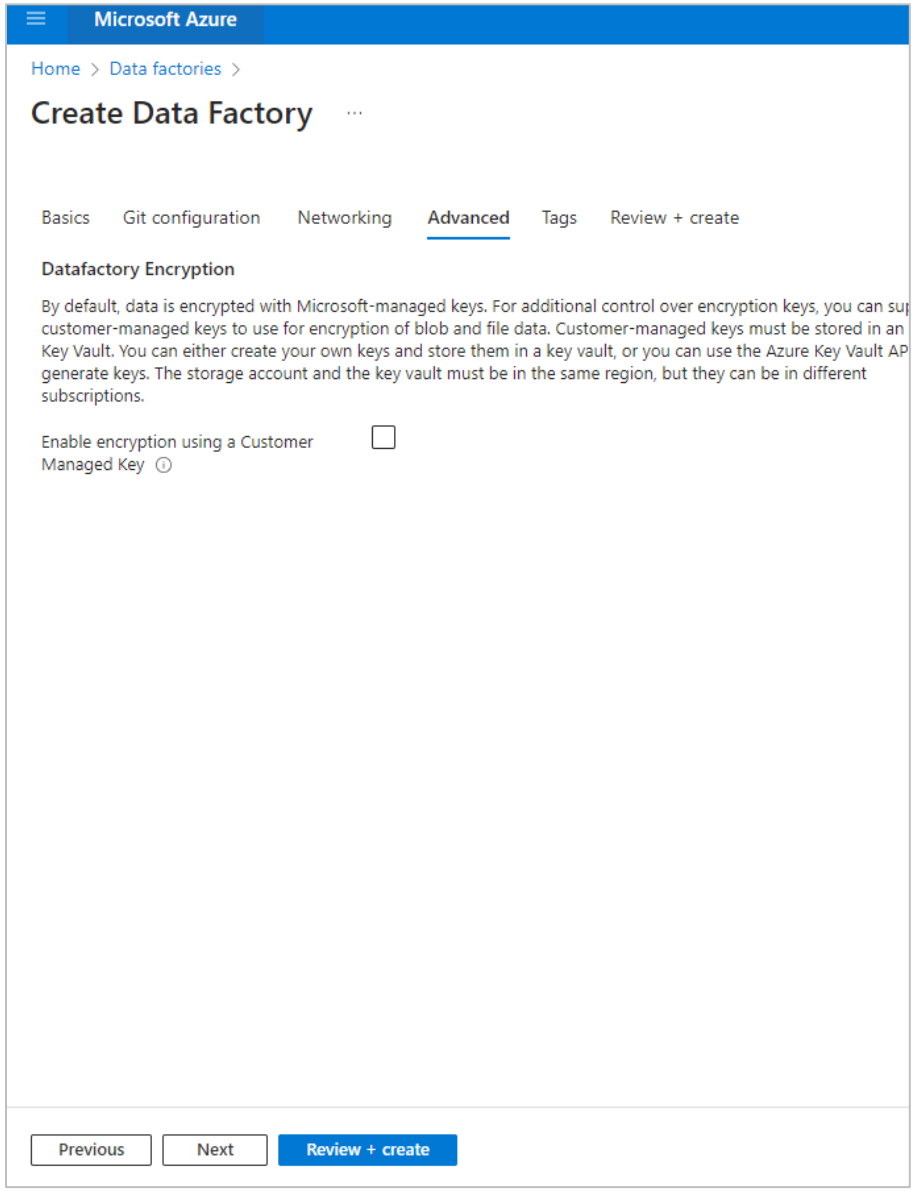
- Configure git later
- next

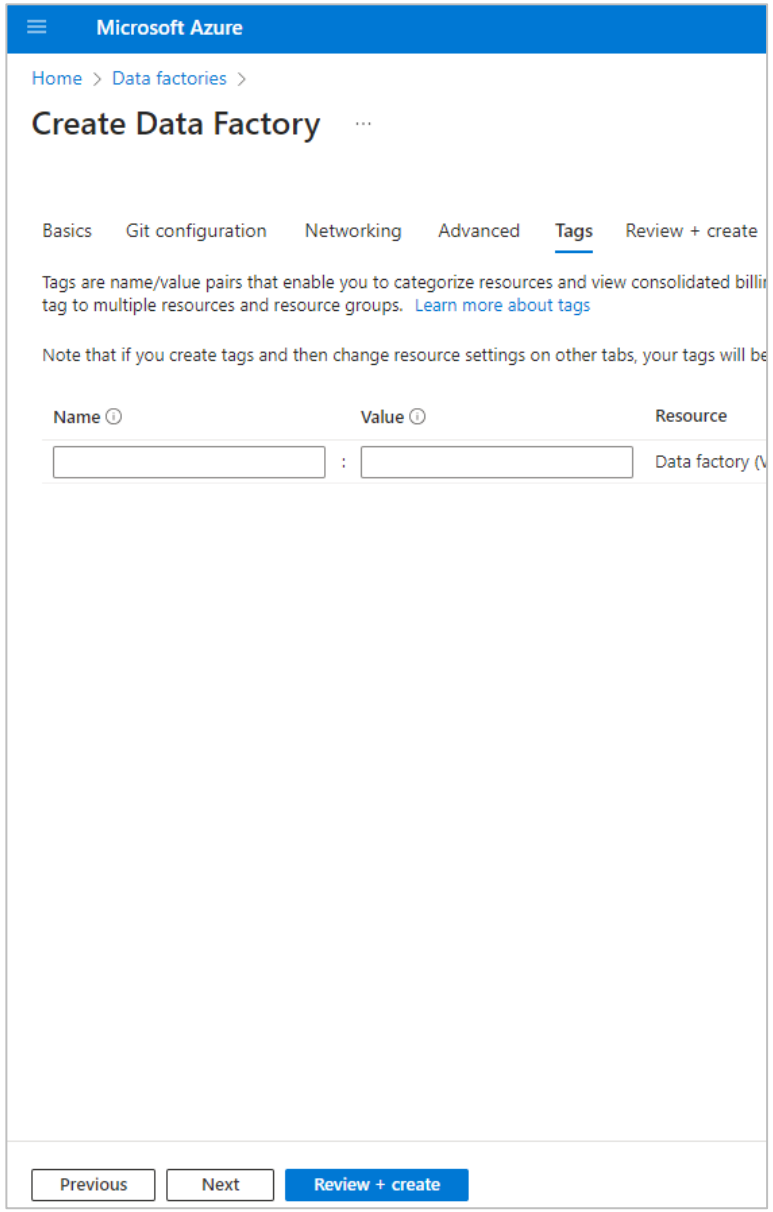


- Next

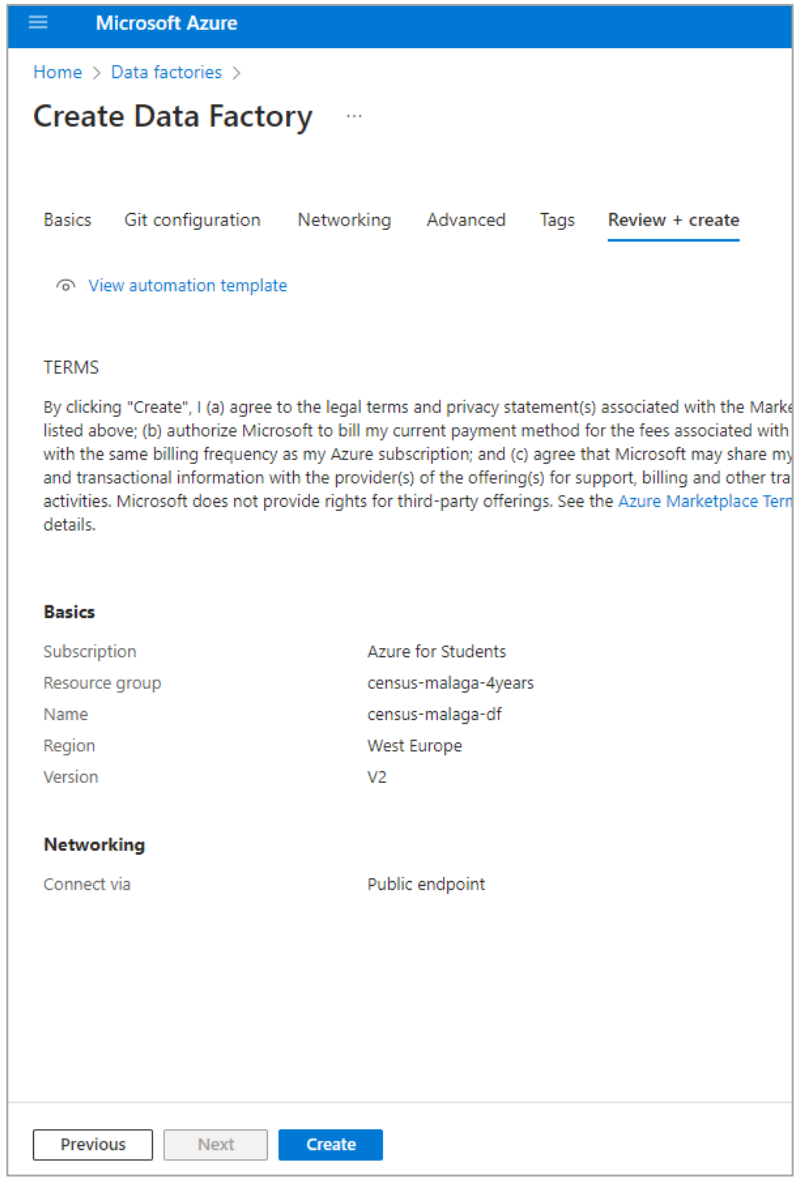

Microsoft Azure
Home > Data factories >
Create Data Factory
Basics
Git configuration
Networking
Advanced
Tags
Review + create
Managed virtual network
Choose whether you want the default AutoResolveIntegrationRuntime to be provisioned on demand inside an ADF-managed virtual network. If this setting is 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 privat 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

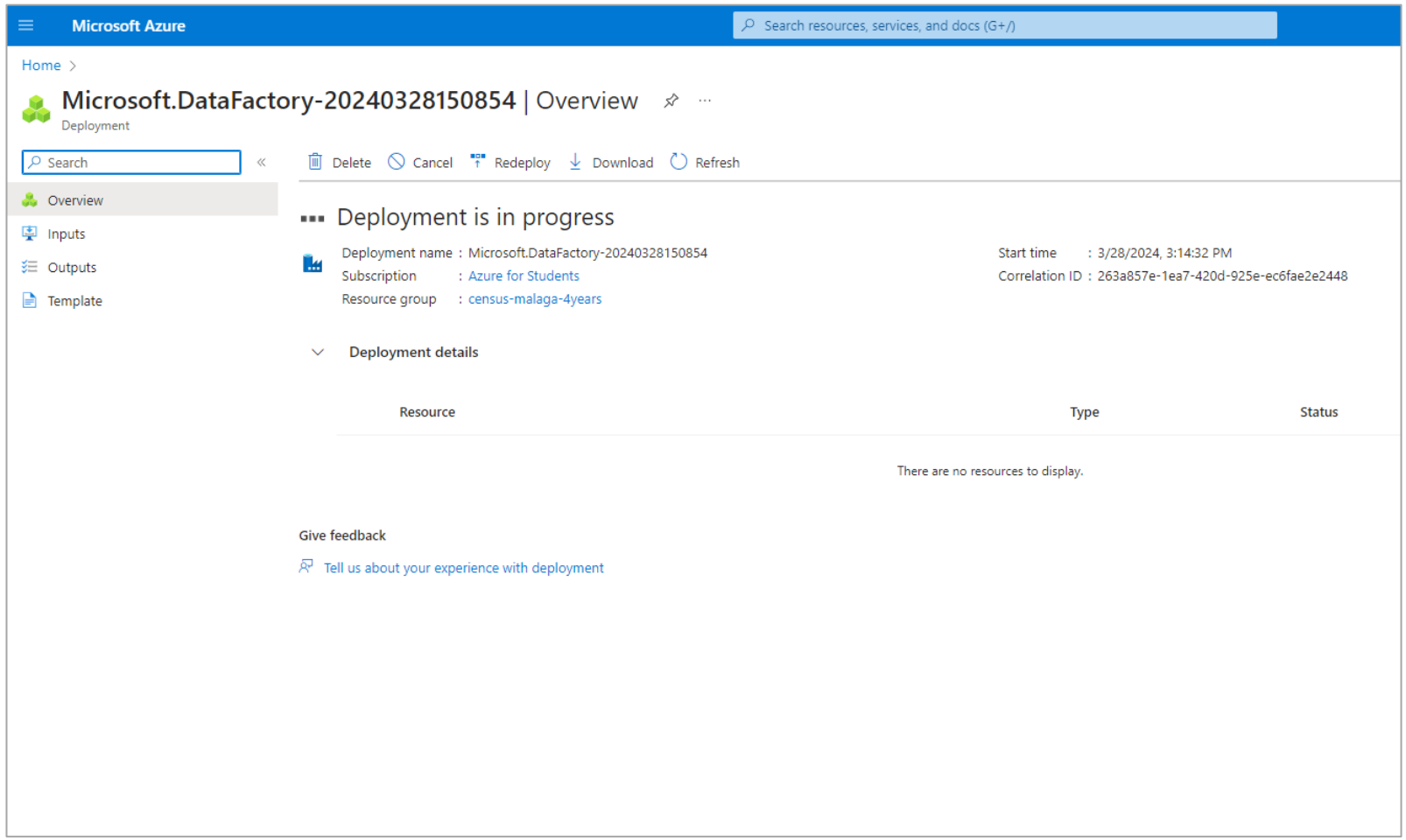

Microsoft Azure
Home > Data factories >
Create Data Factory
Basics
Git configuration
Networking
Advanced
Tags
Review + create
Datafactory Encryption
By default, data is encrypted with Microsoft-managed keys. For additional control over encryption keys, you can sup
customer-managed keys to use for encryption of blob and file data. Customer-managed keys must be stored in an
Key Vault. You can either create your own keys and store them in a key vault, or you can use the Azure Key Vault AP
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 billi
tag to 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
Name
Value
Resource
Data factory (
Previous
Next
Review + create

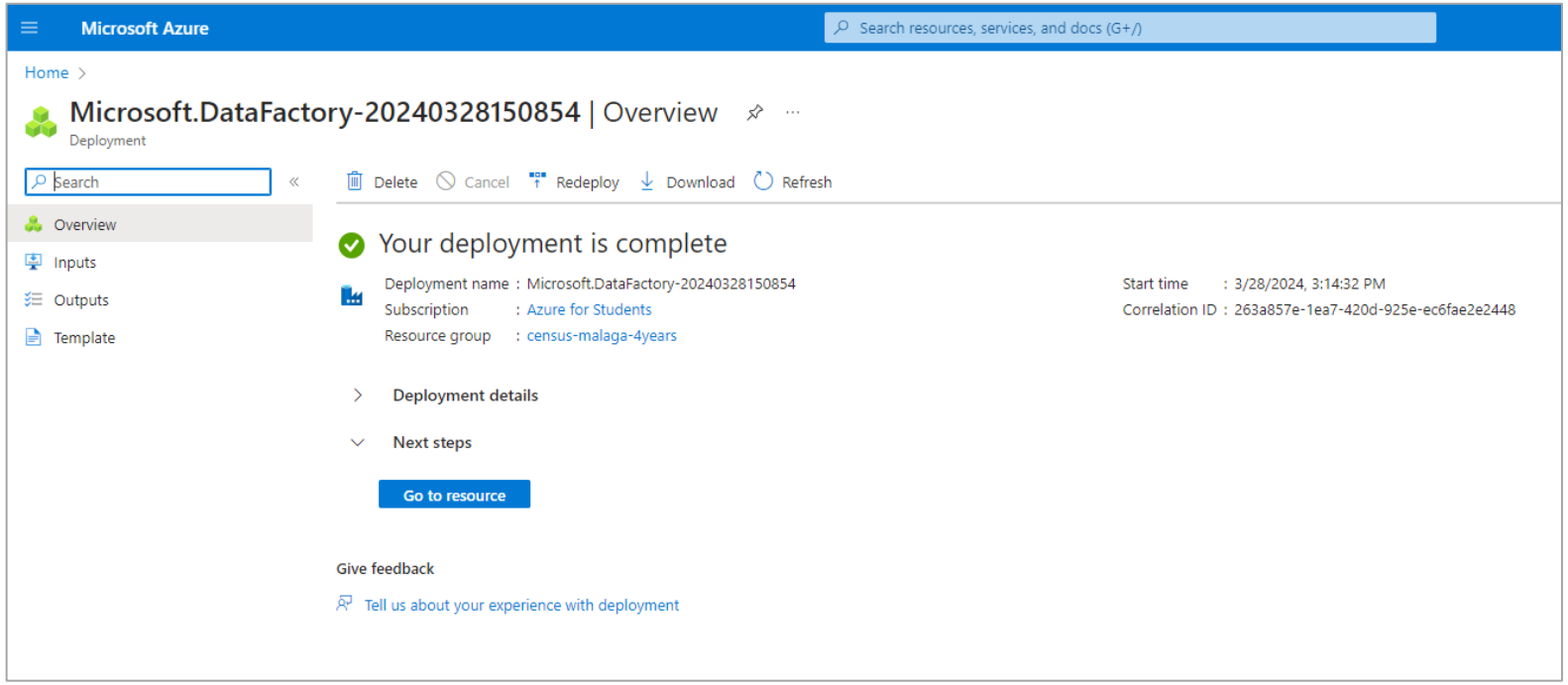
- Create


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 Marke
listed above; (b) authorize Microsoft to bill my current payment method for the fees associated with
with the same billing frequency as my Azure subscription; and (c) agree that Microsoft may share my
and transactional information with the provider(s) of the offering(s) for support, billing and other tra
activities. Microsoft does not provide rights for third-party offerings. See the Azure Marketplace Tern
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

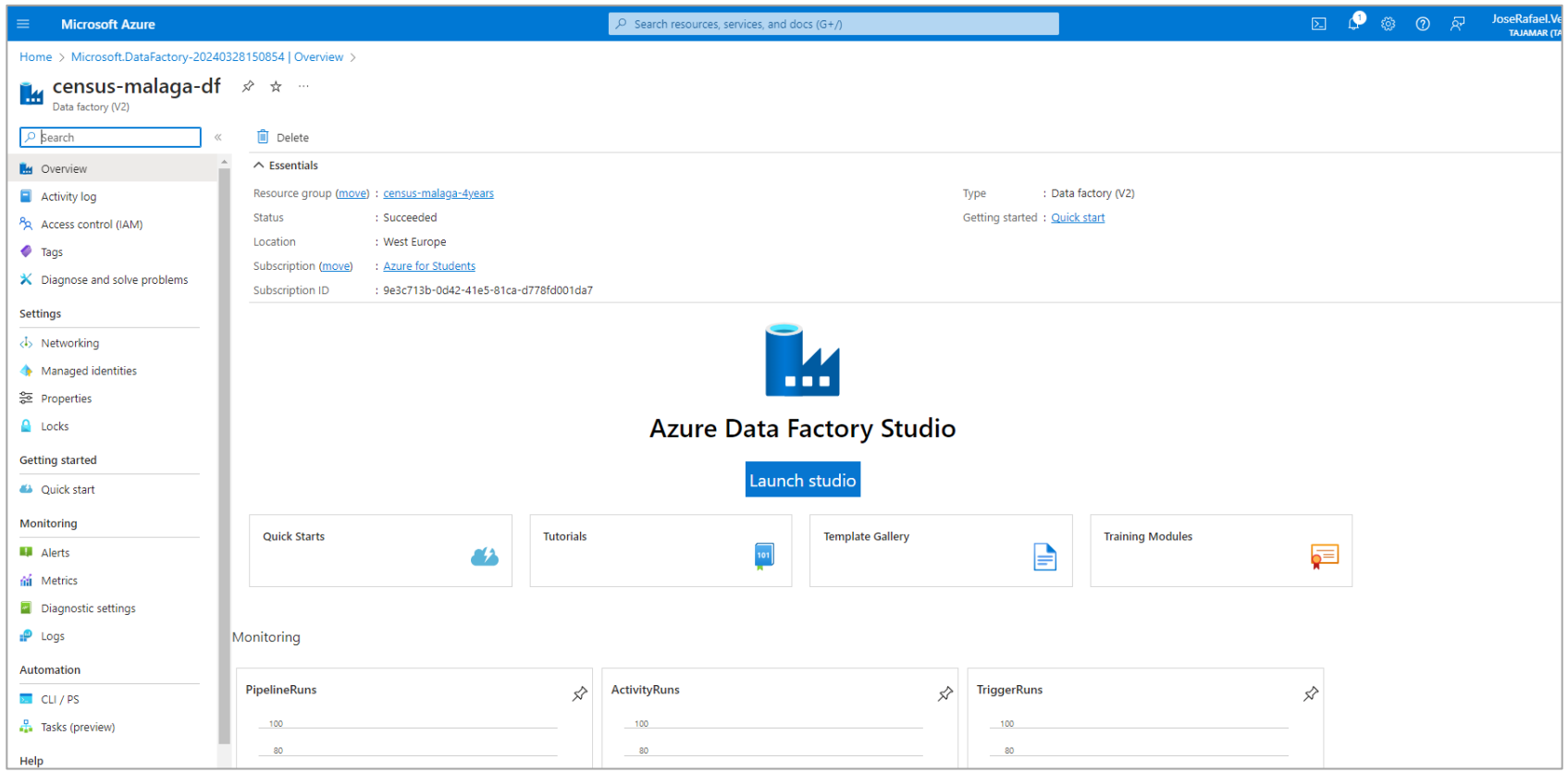
- Wait until data factory deployment is completed



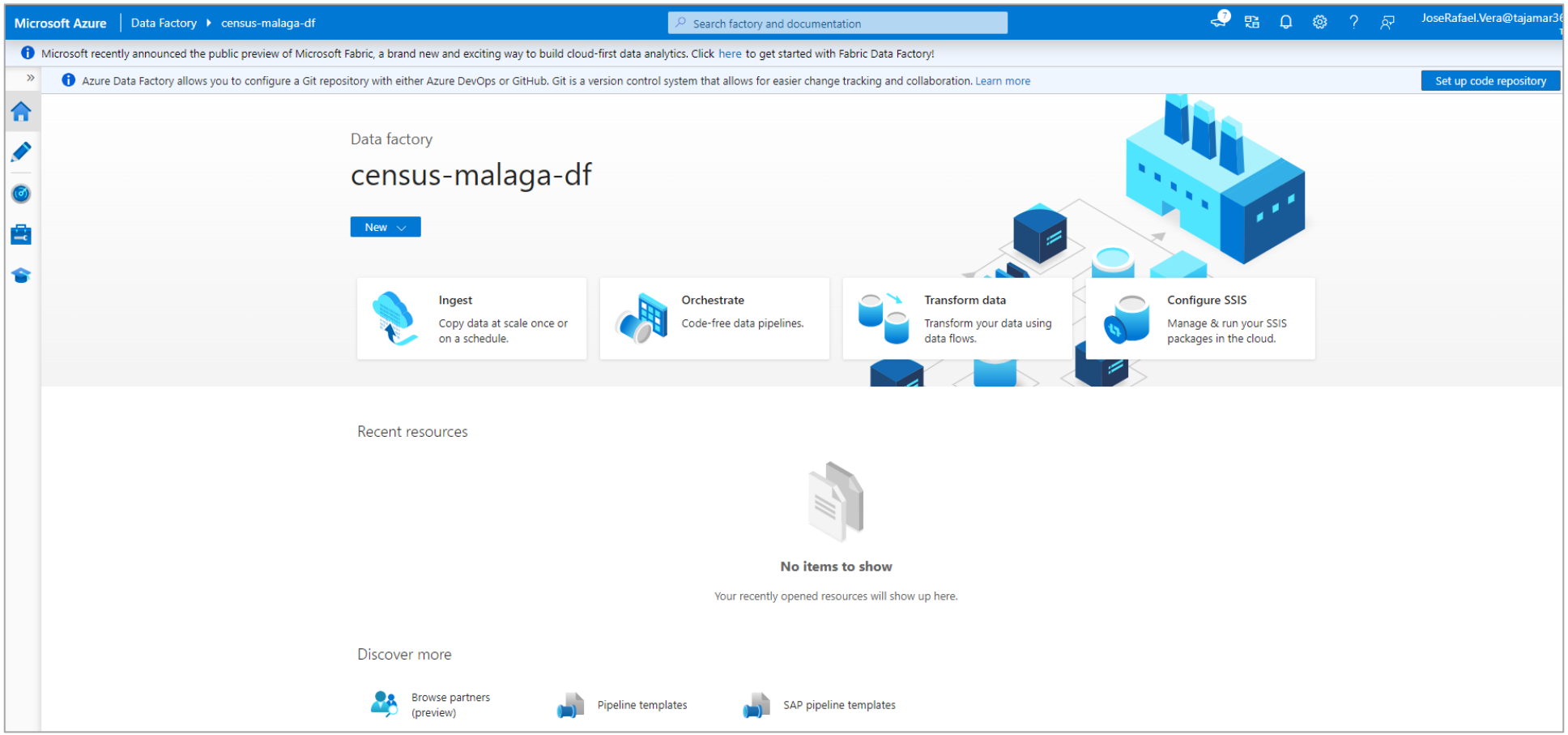
- Data Factory instance successfully deployed
- Go to resource



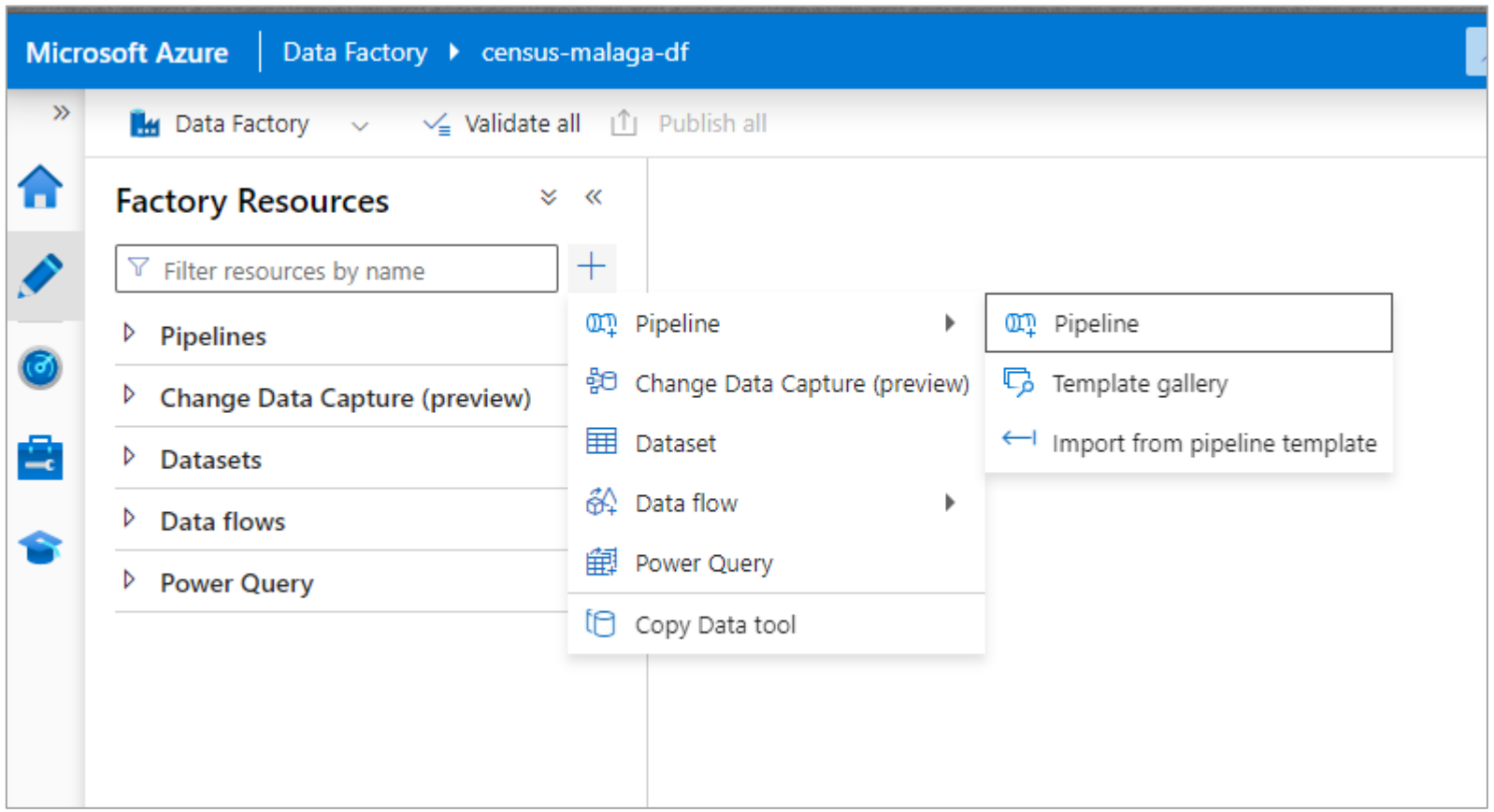
- Launch Data Factory Studio



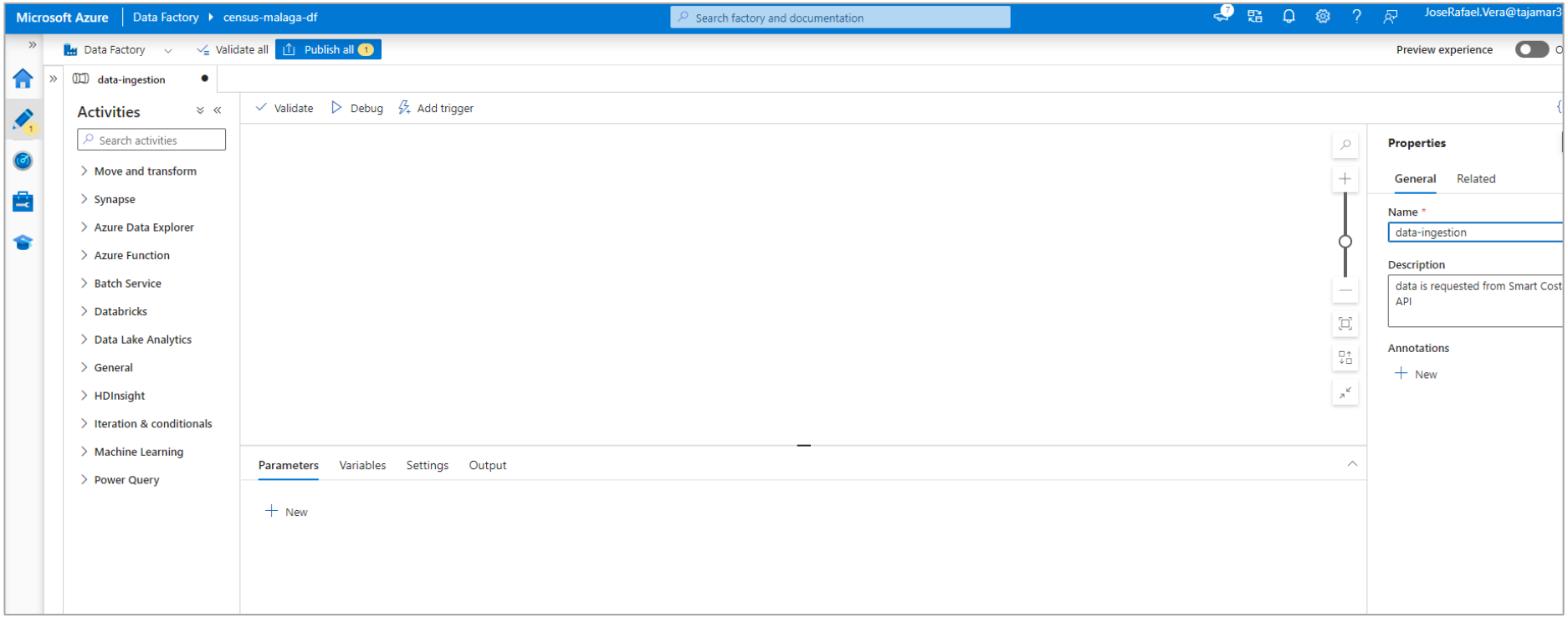
- Click on pencil



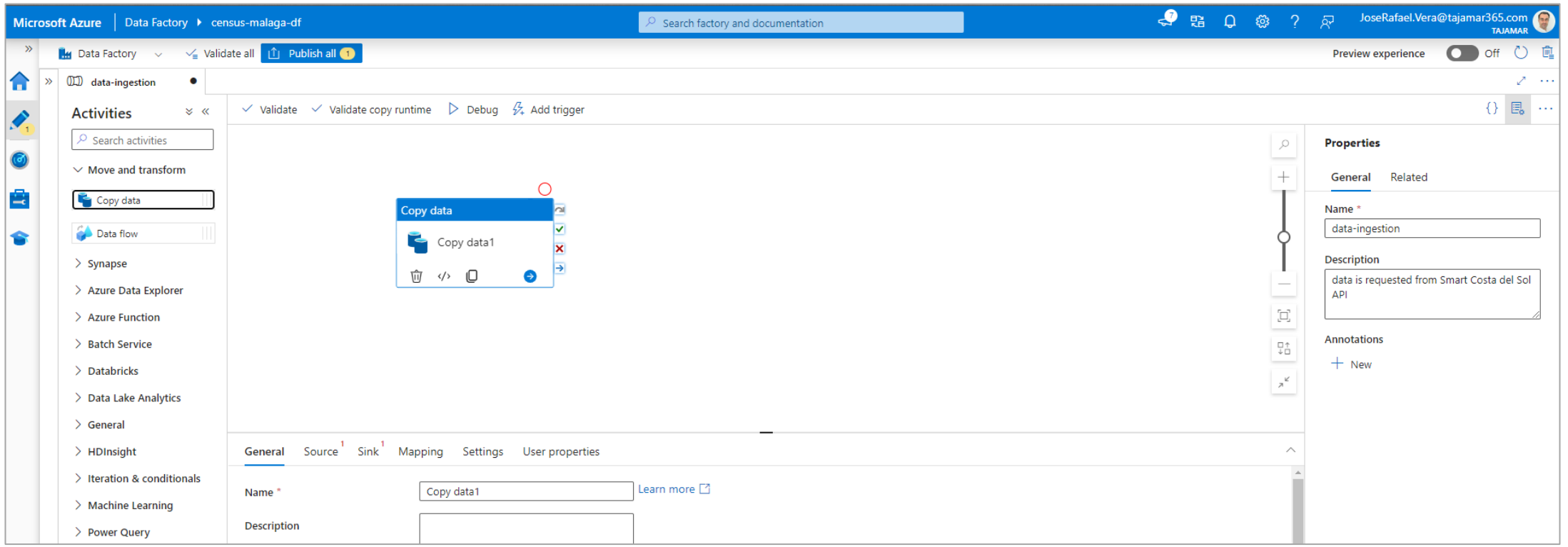
- Click on +
- Select pipeline > pipeline



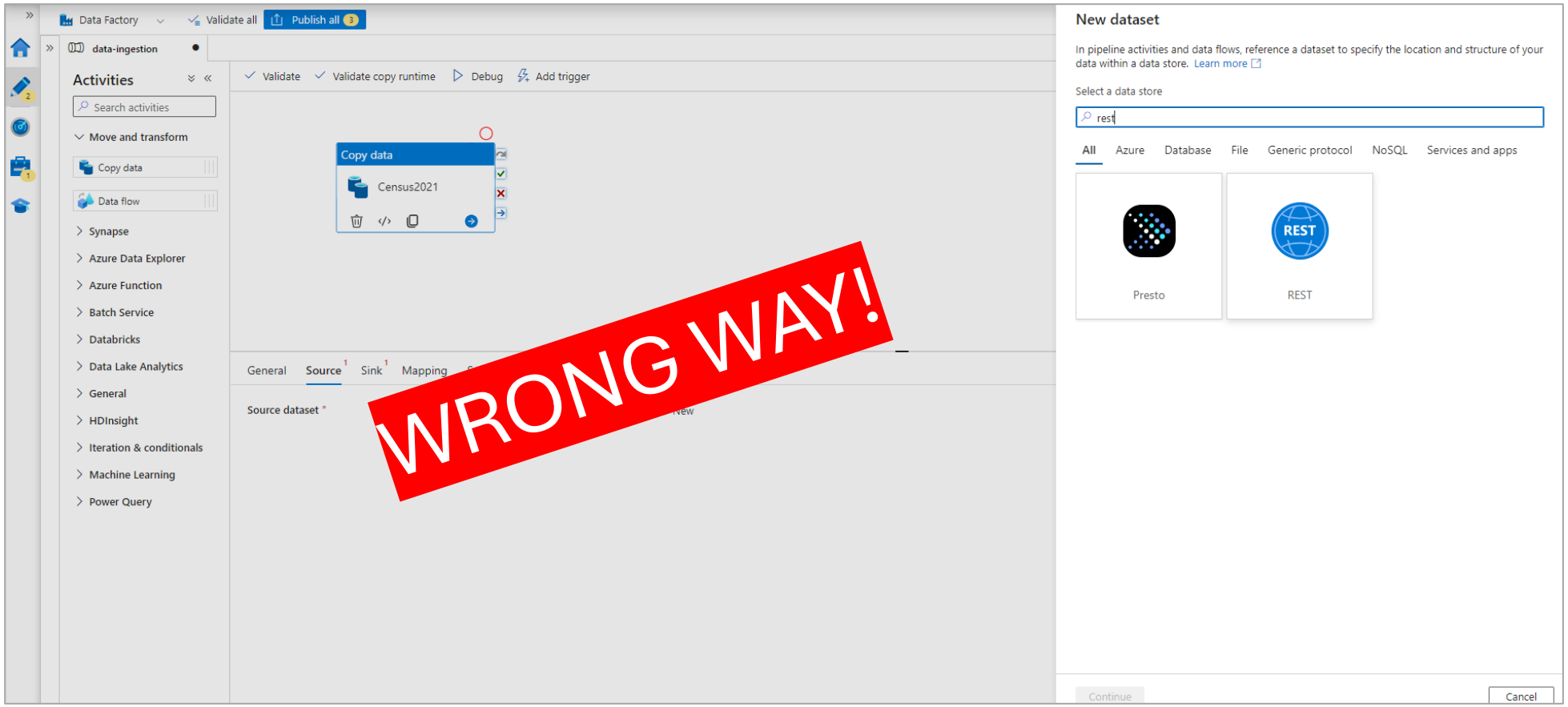
- Console to start building pipelines
- Give the pipeline a name



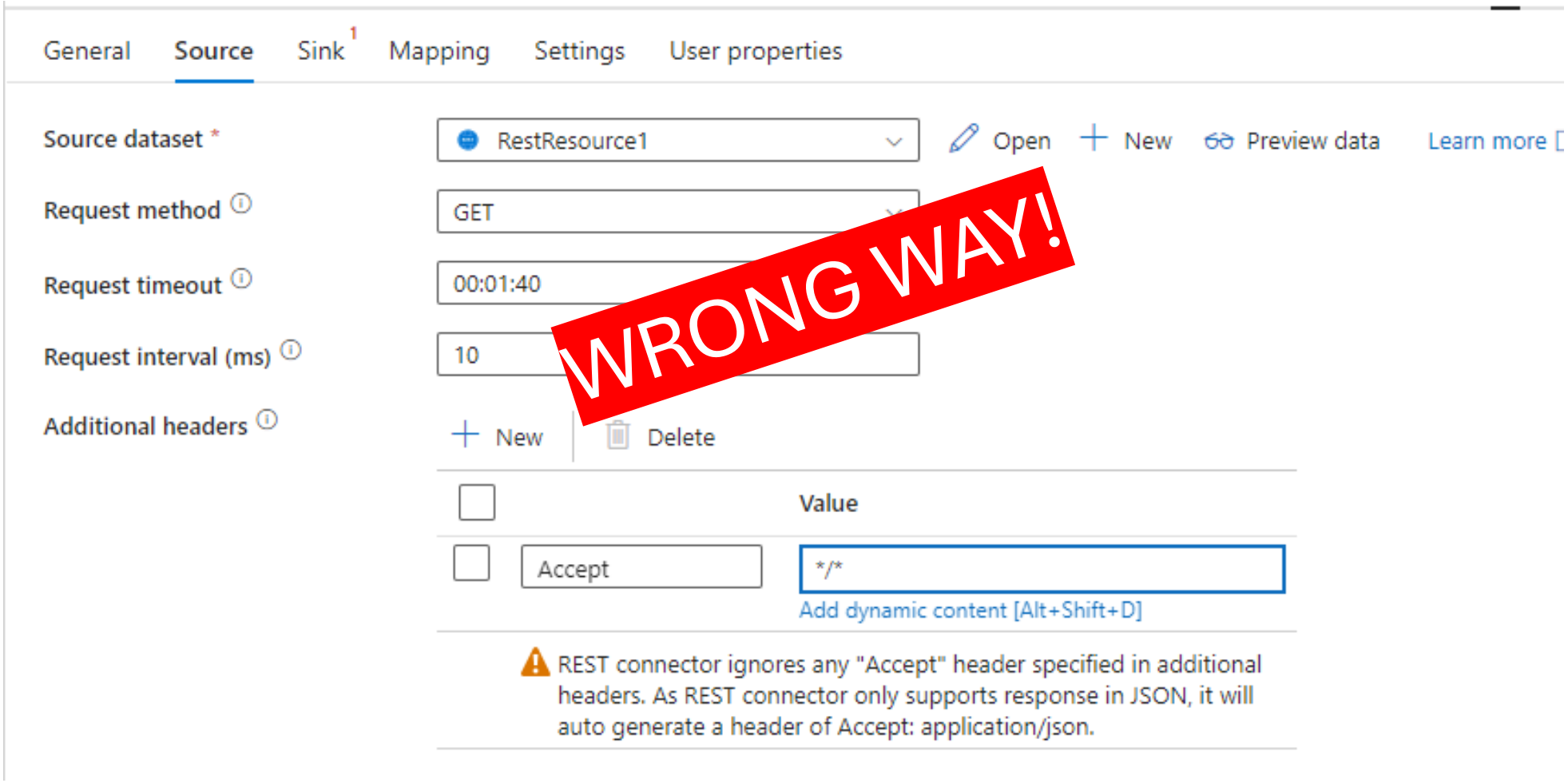
- Click the dropdown “Move and Transform”
- Drag and drop “copy data” to the central panel
- Give this step a name



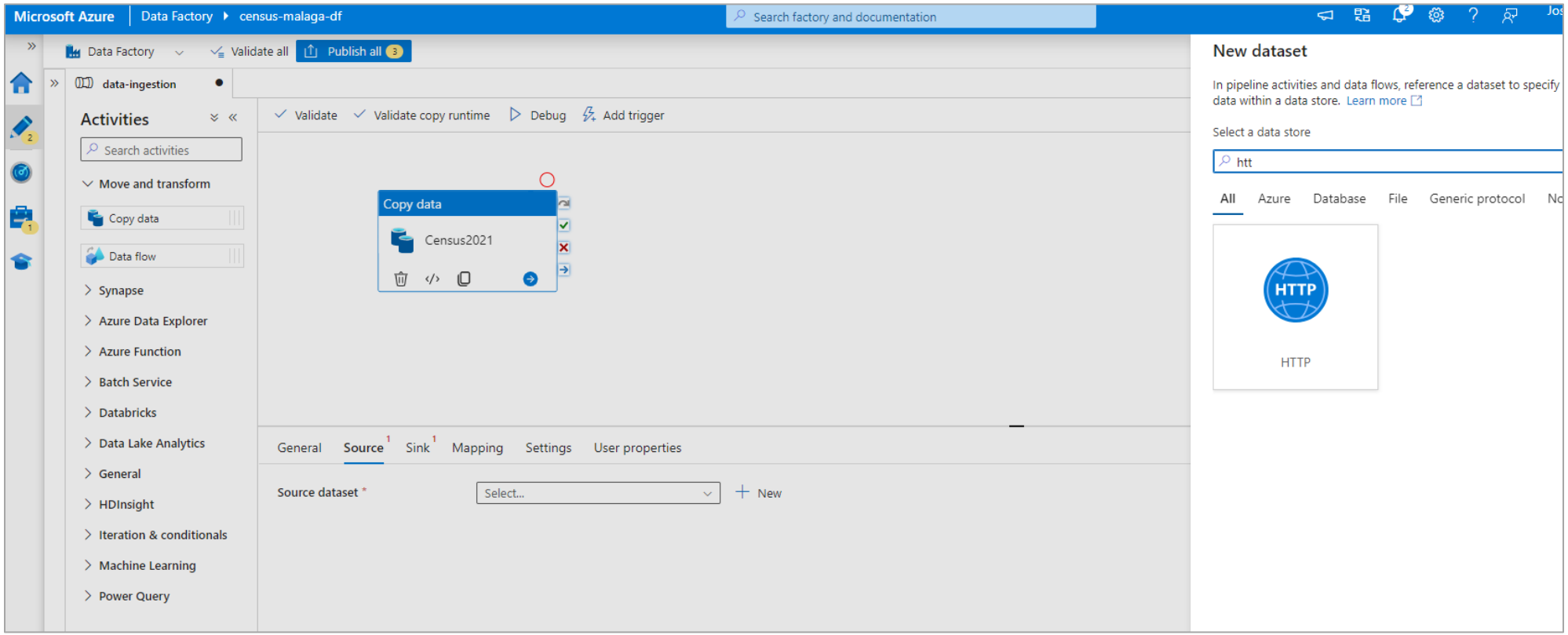
- The data source should not be REST, as such Azure service expect only JSON type response.
- The data served from smart Costa del Sol API is as .csv format.
- So we need to select a data source type as HTTP



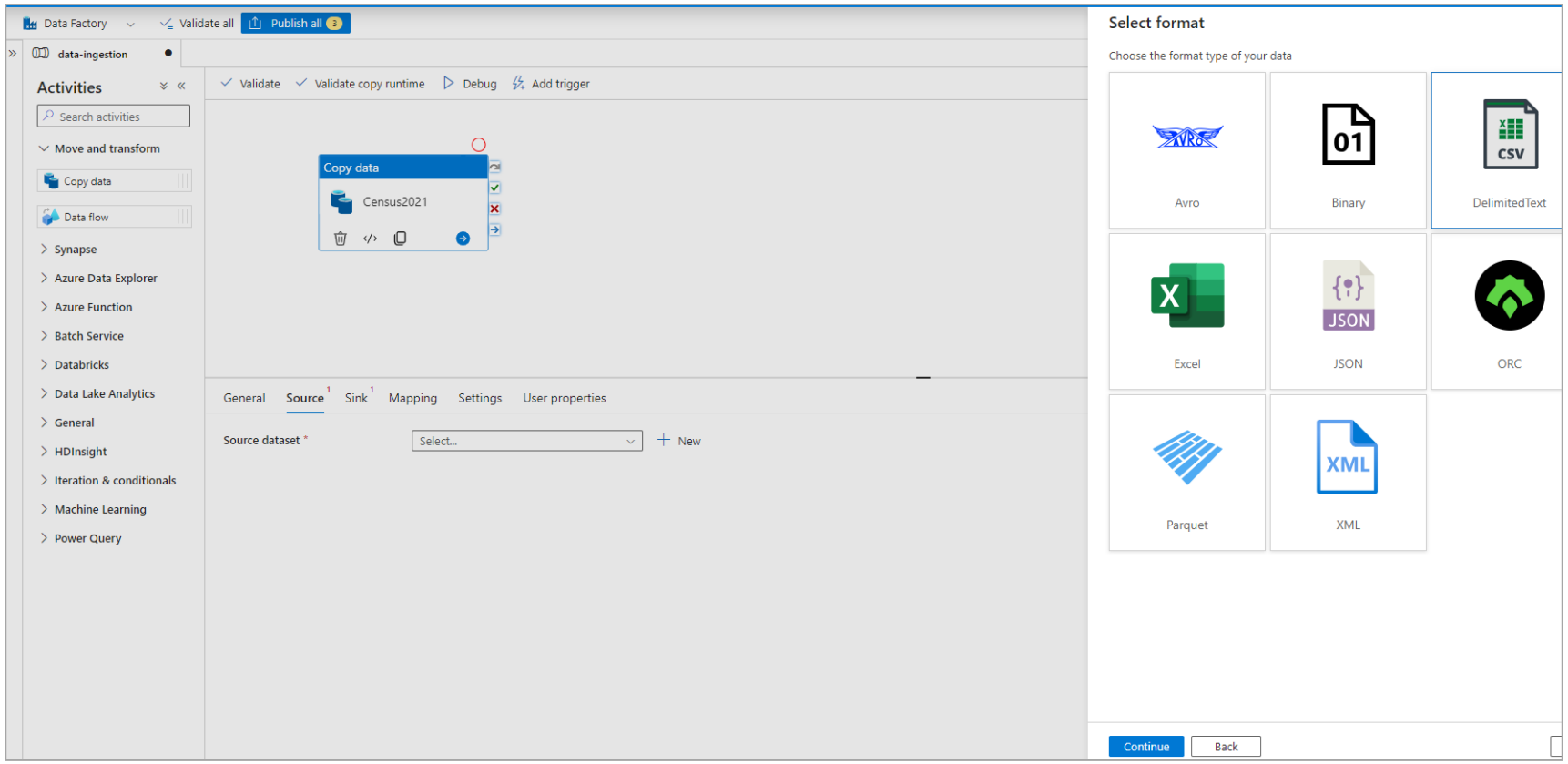
- Explanation below



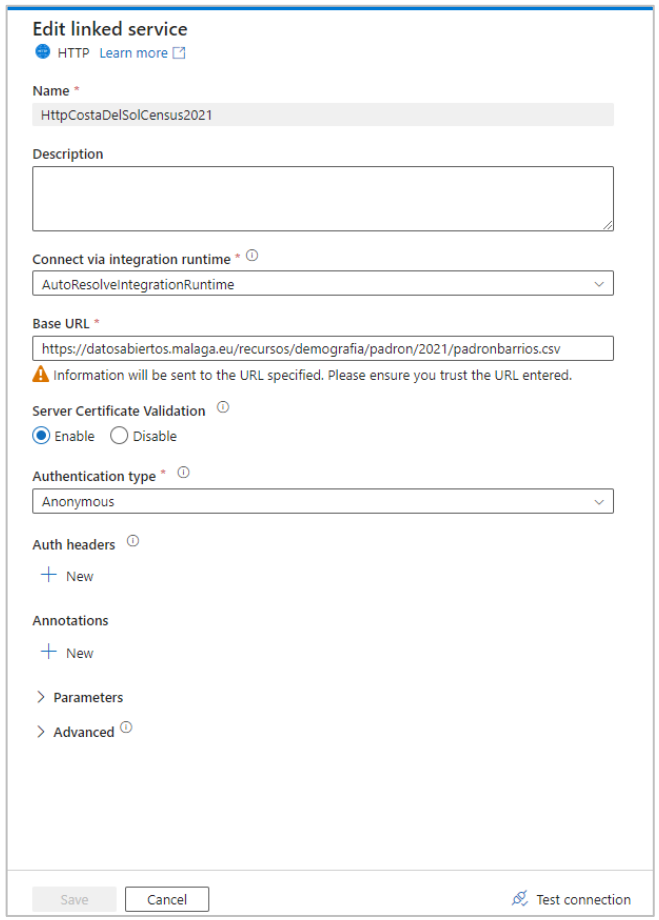
- Selecting the copy data, click on Source below
- Choose the data source as HTTP (if response is csv type)

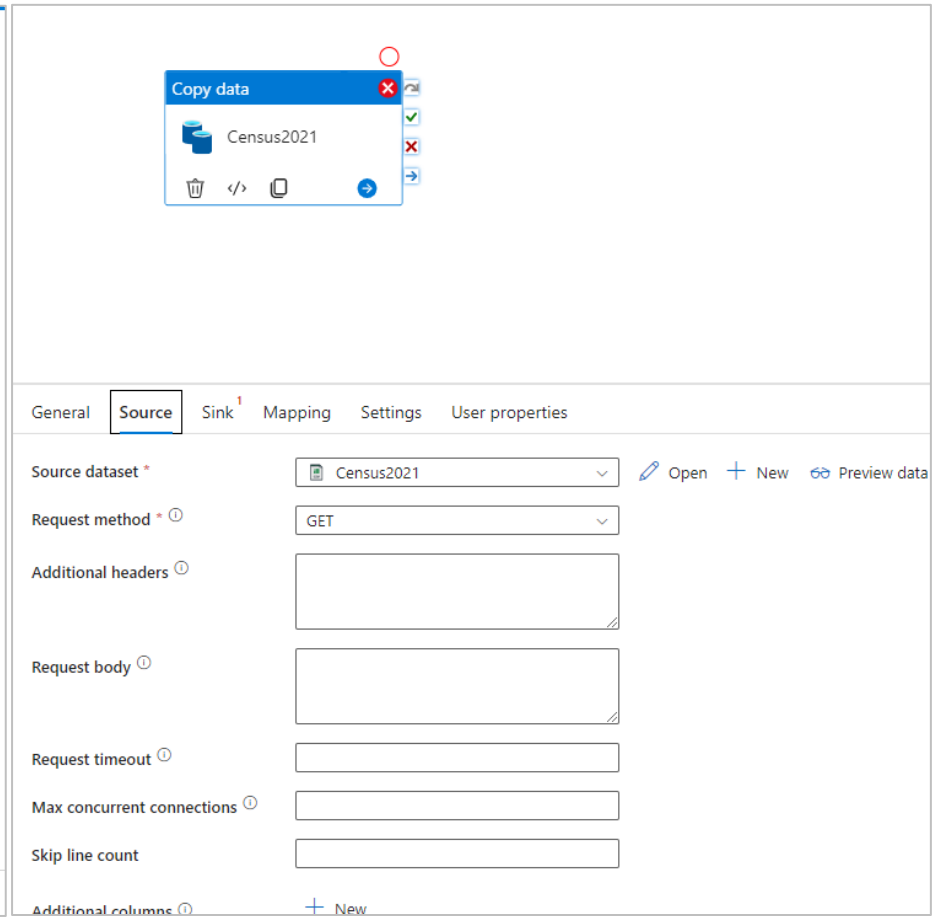


- Select de input format type. In this case .csv
- continue

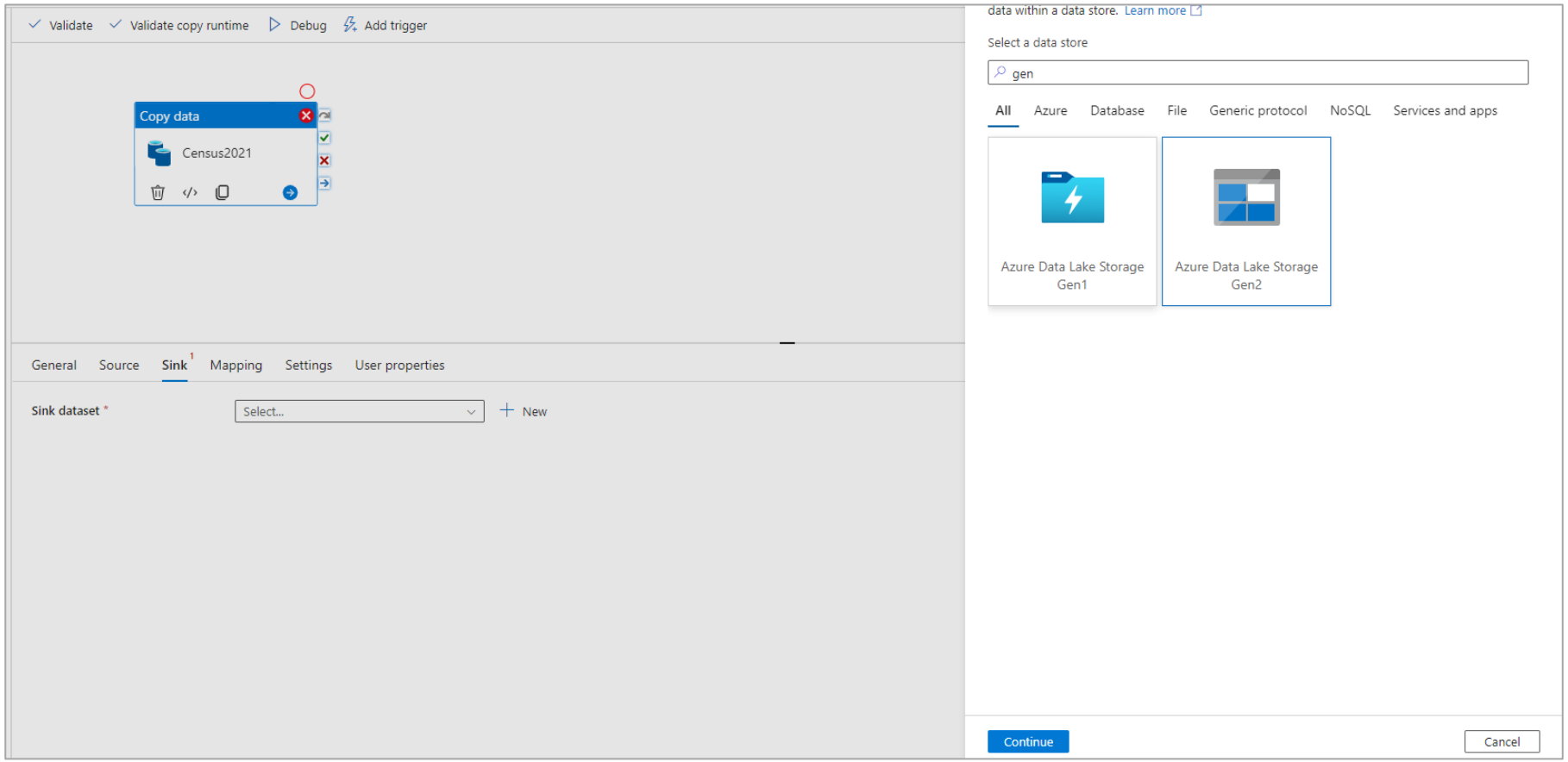


- Paste the base URL where Data Factory will request the Data. (in my repository you find all the URLs)
- Authentication “Anonymous” . In production env you may probably set an authn type
- Save

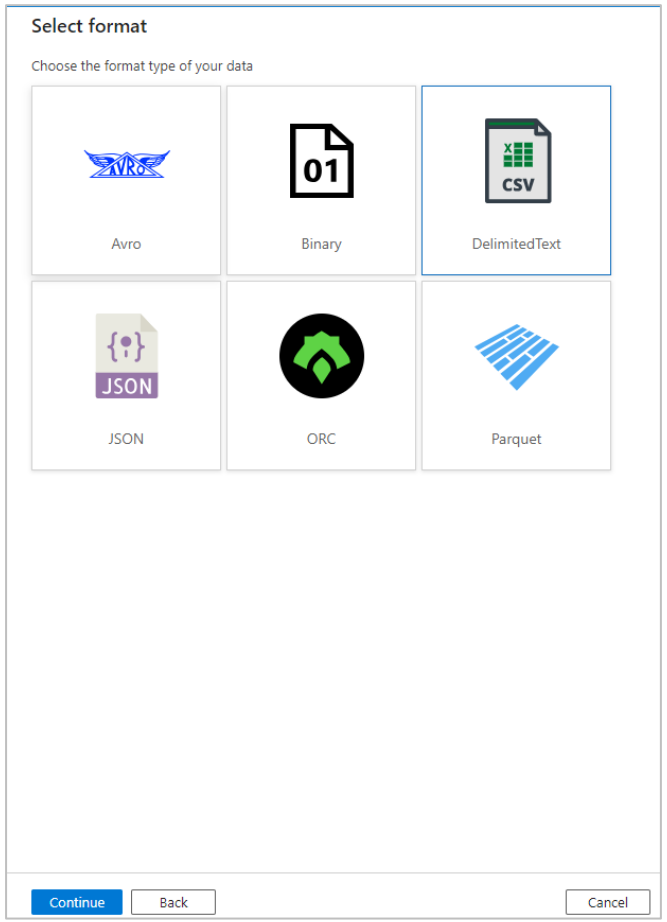


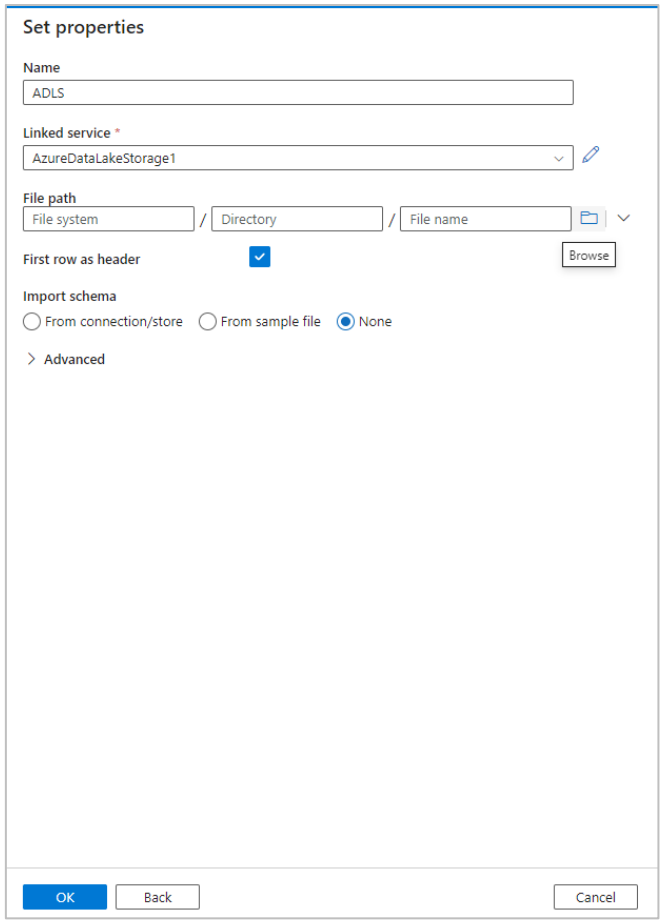
- Click on “Sink” to define where in Resource group will be the data persisted
- Select ADLSGen2
- Continue

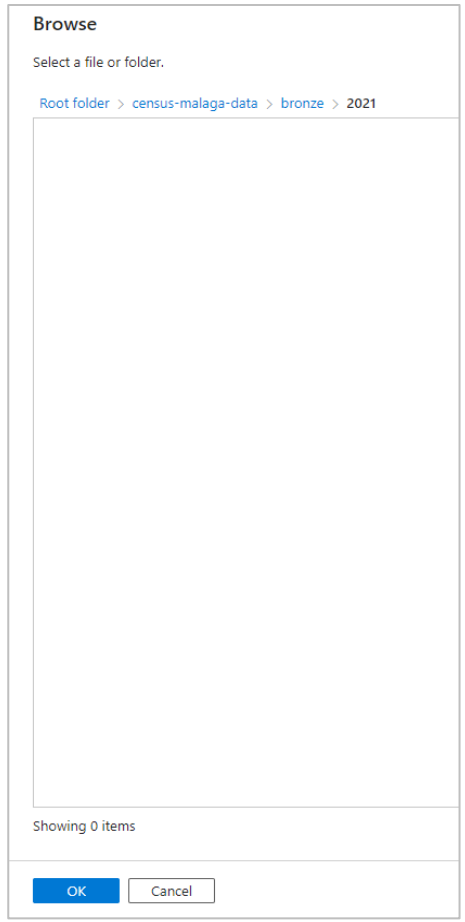


- Select the output file type as .csv
- continue

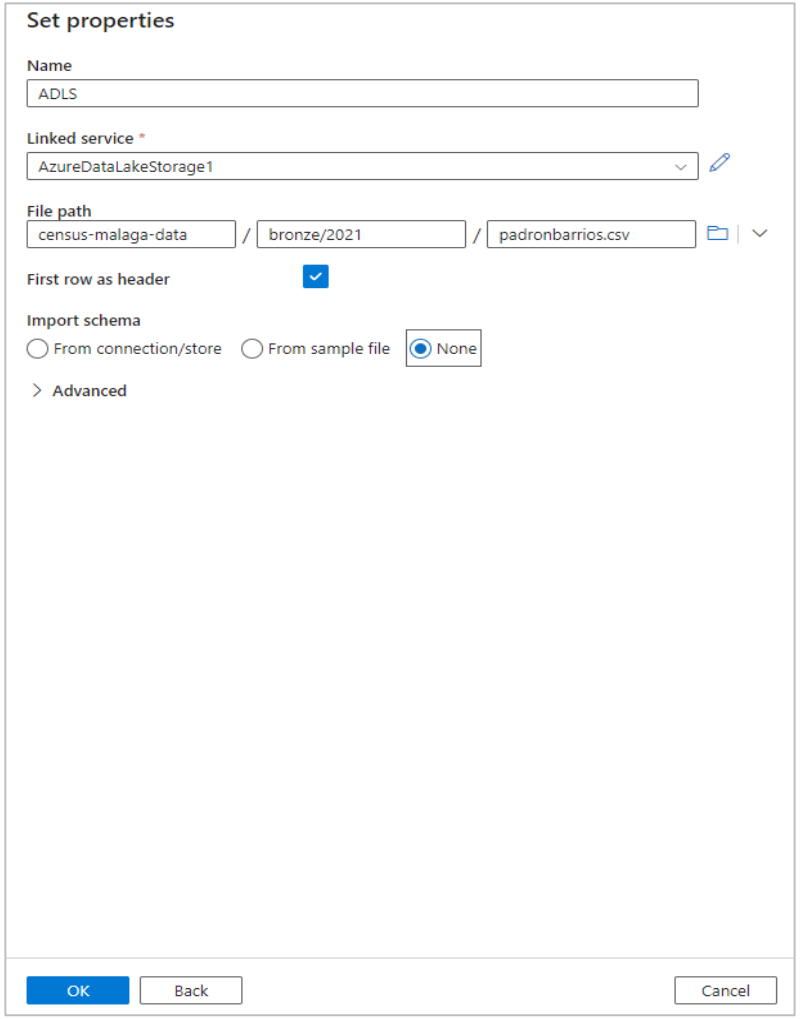



- Give a Name to this sink
- Linked service- what appears there
- Browse the folder into your ADLS container you've created before
- First row as header
- Import schema "None"

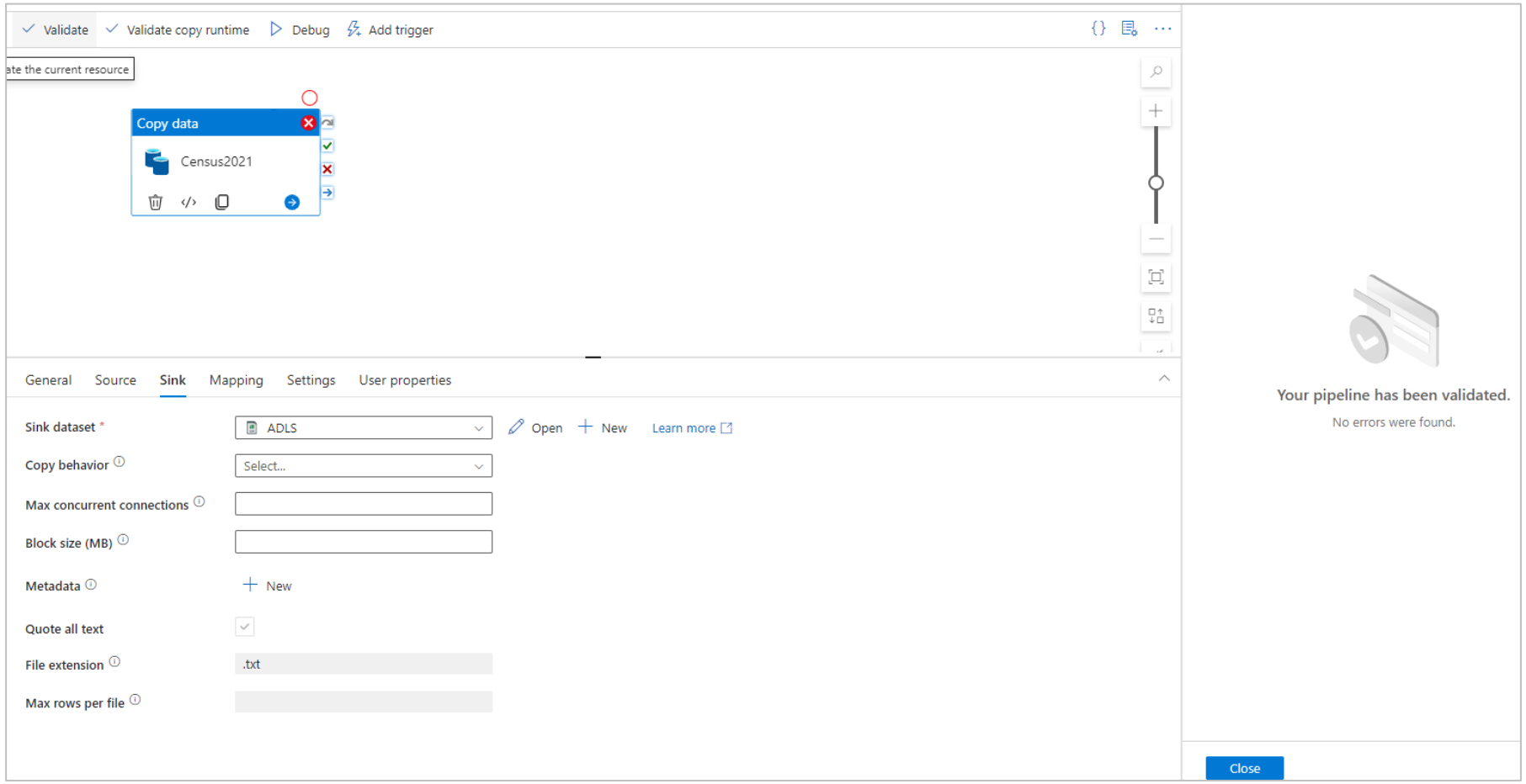


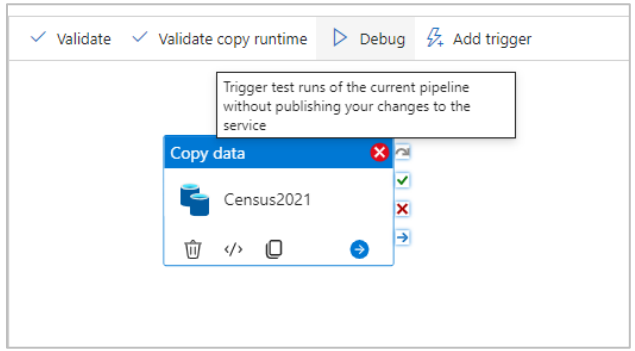
- Give the file a name with extension
- ok

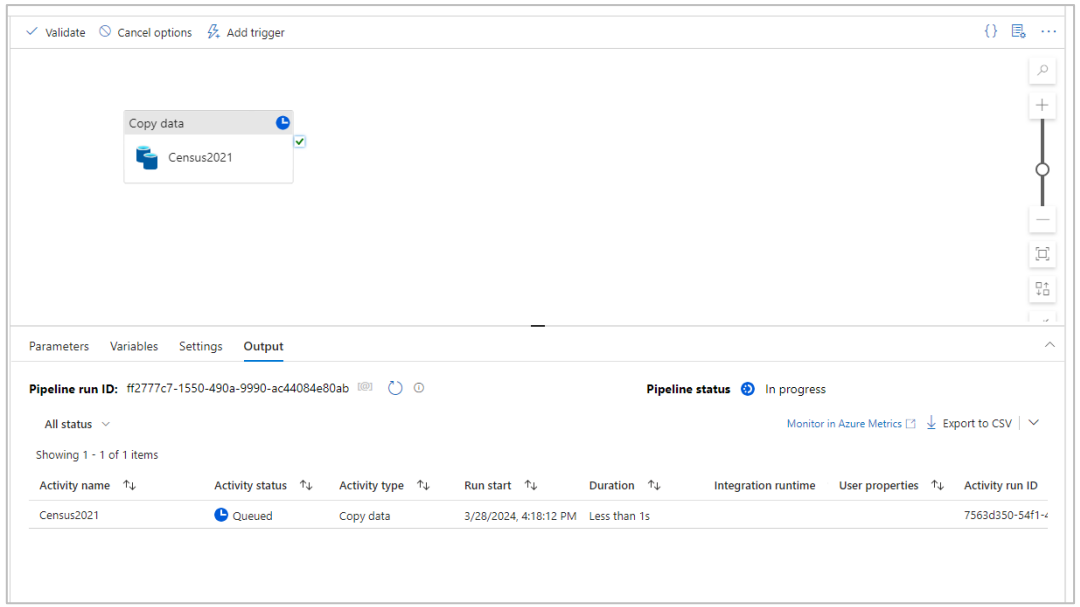


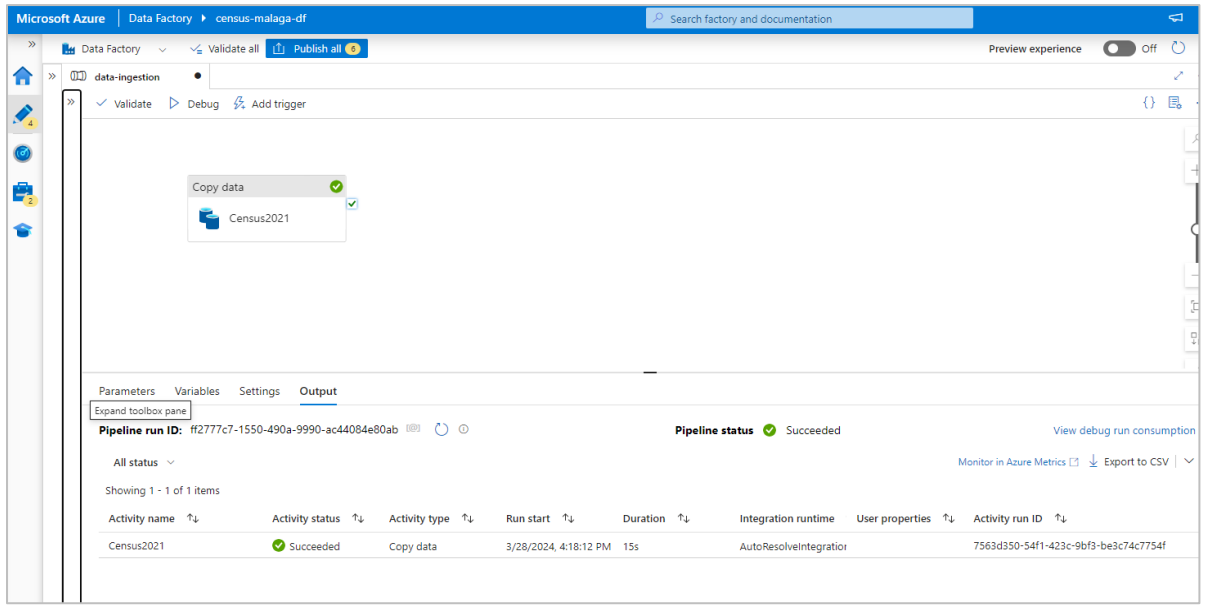
- First step configured: source and sink
- Click validate
- Click publish all



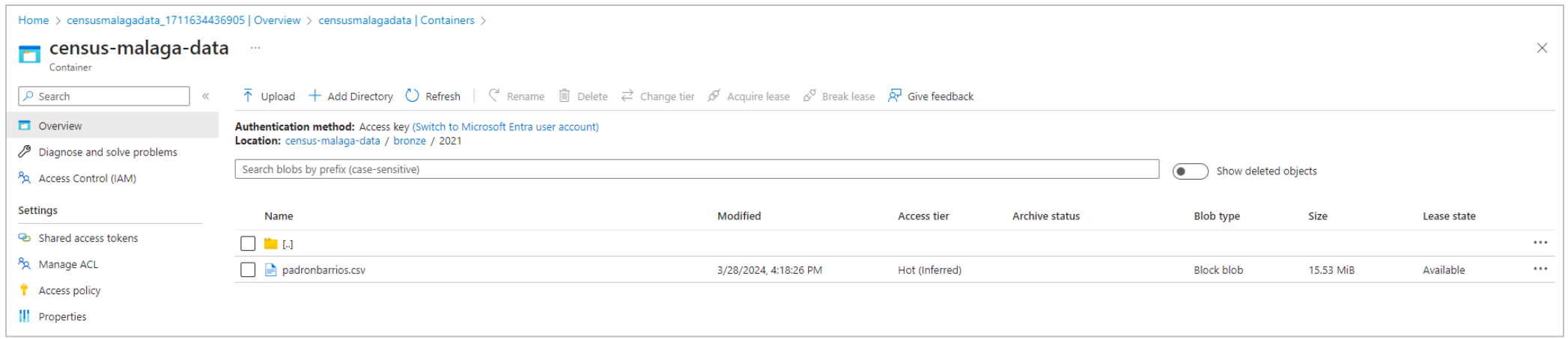
- Click debug to run the pipeline
- See below how pipeline is queued and running. Calling to the csv through Http
- After a while, the process succeded
- Check if the .csv file landed into bronze directory



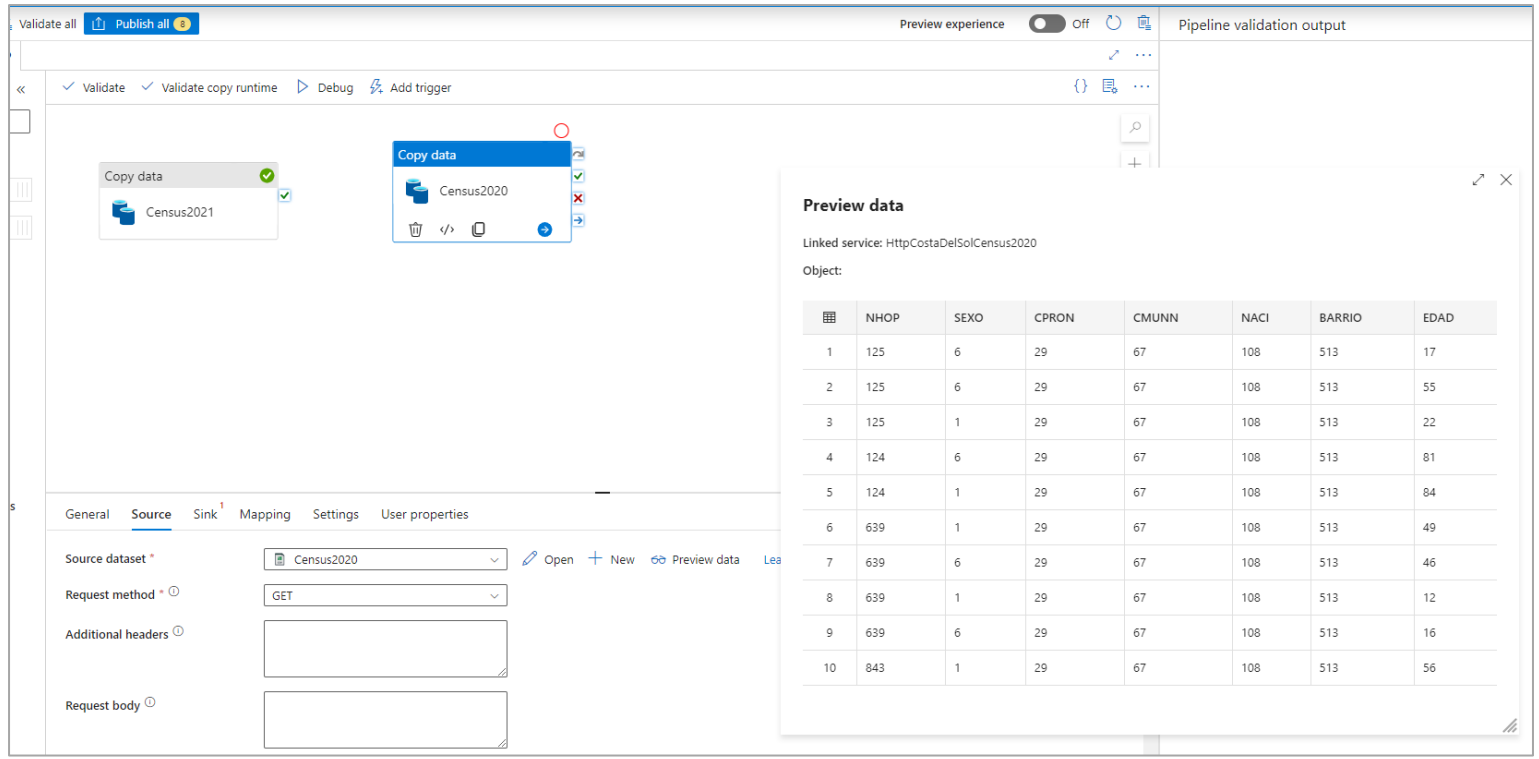



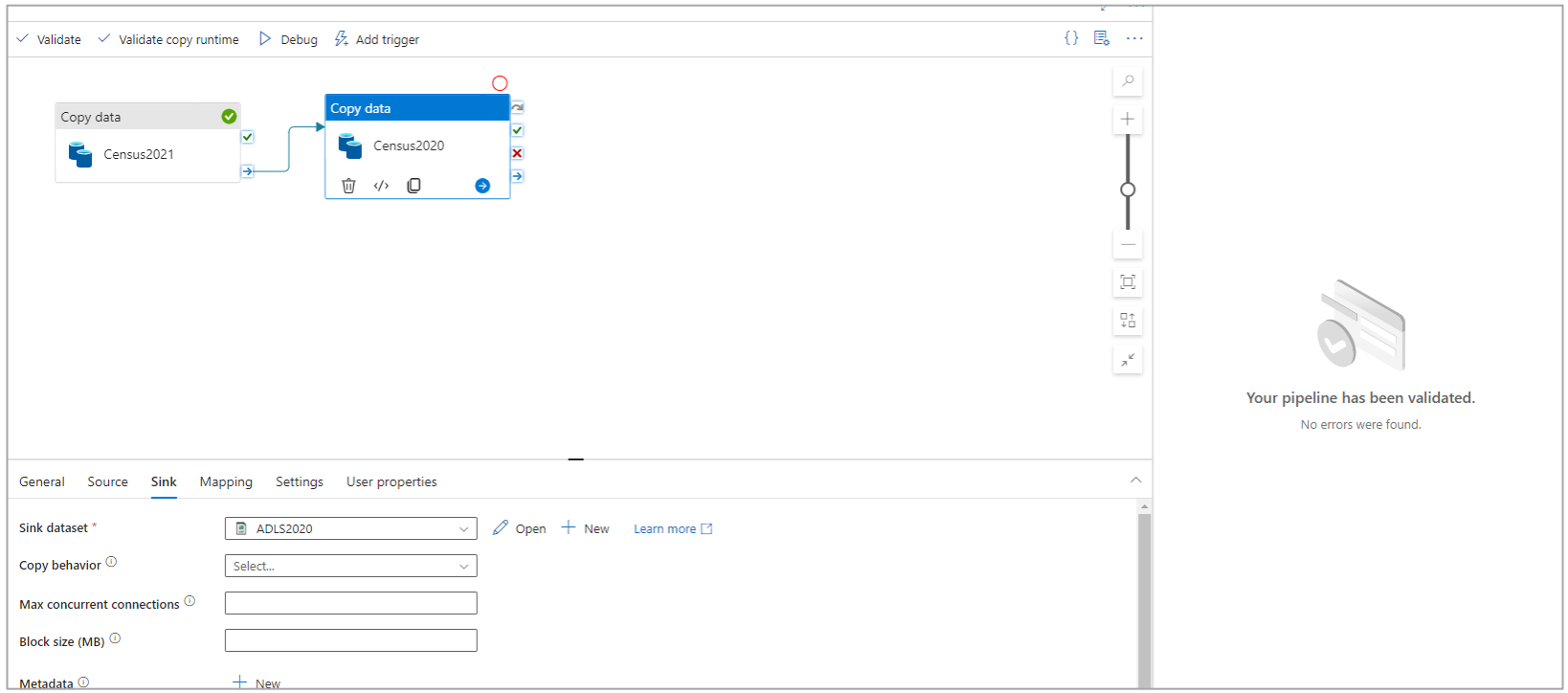
- There it is. First ingestion completed.
- Let’s move forward and create more copy data and concatenate the step into a pipeline to extract and persist all the .csv files



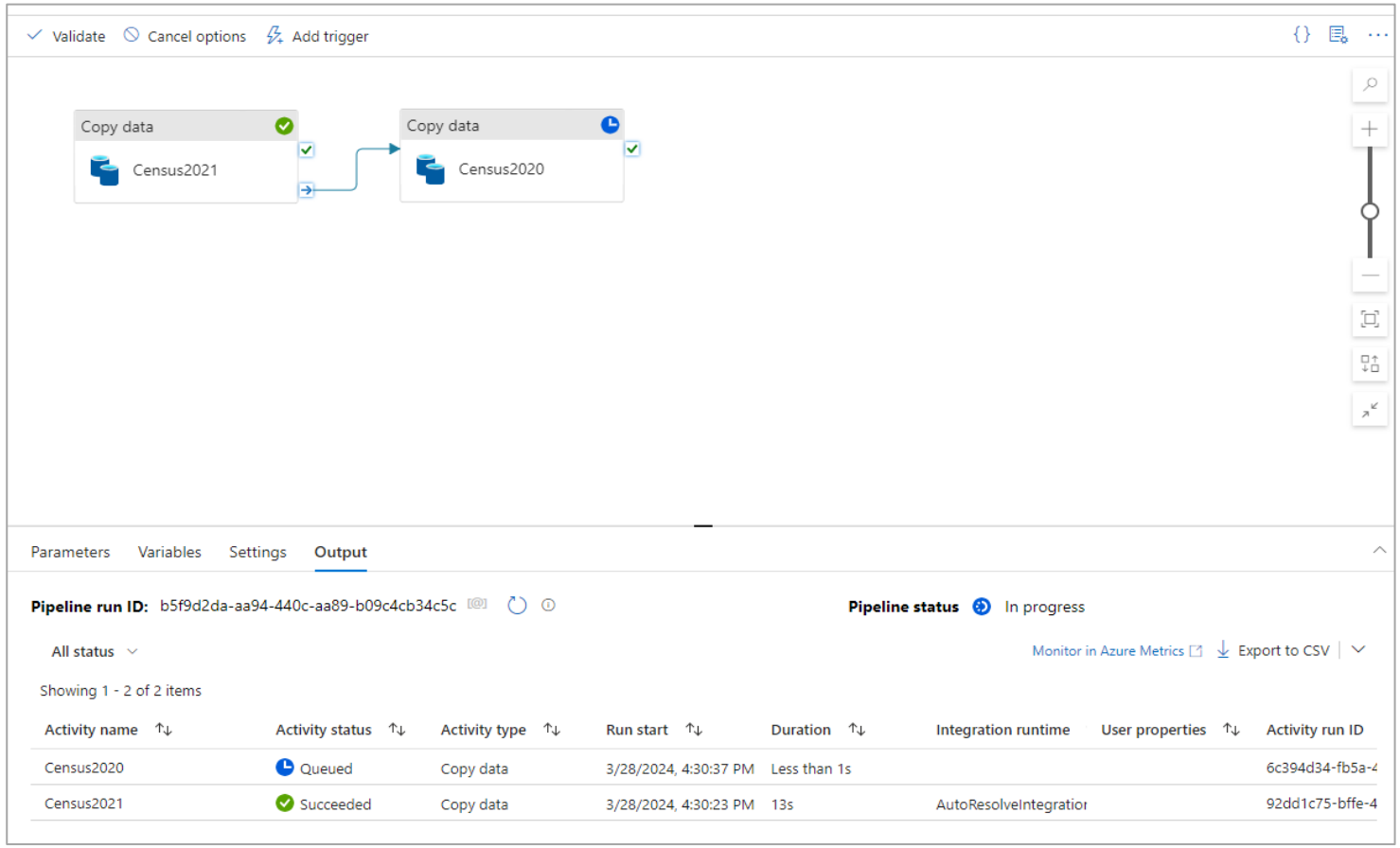
- You can preview the requested and persisted data clicking below on “Preview Data” (glasses)



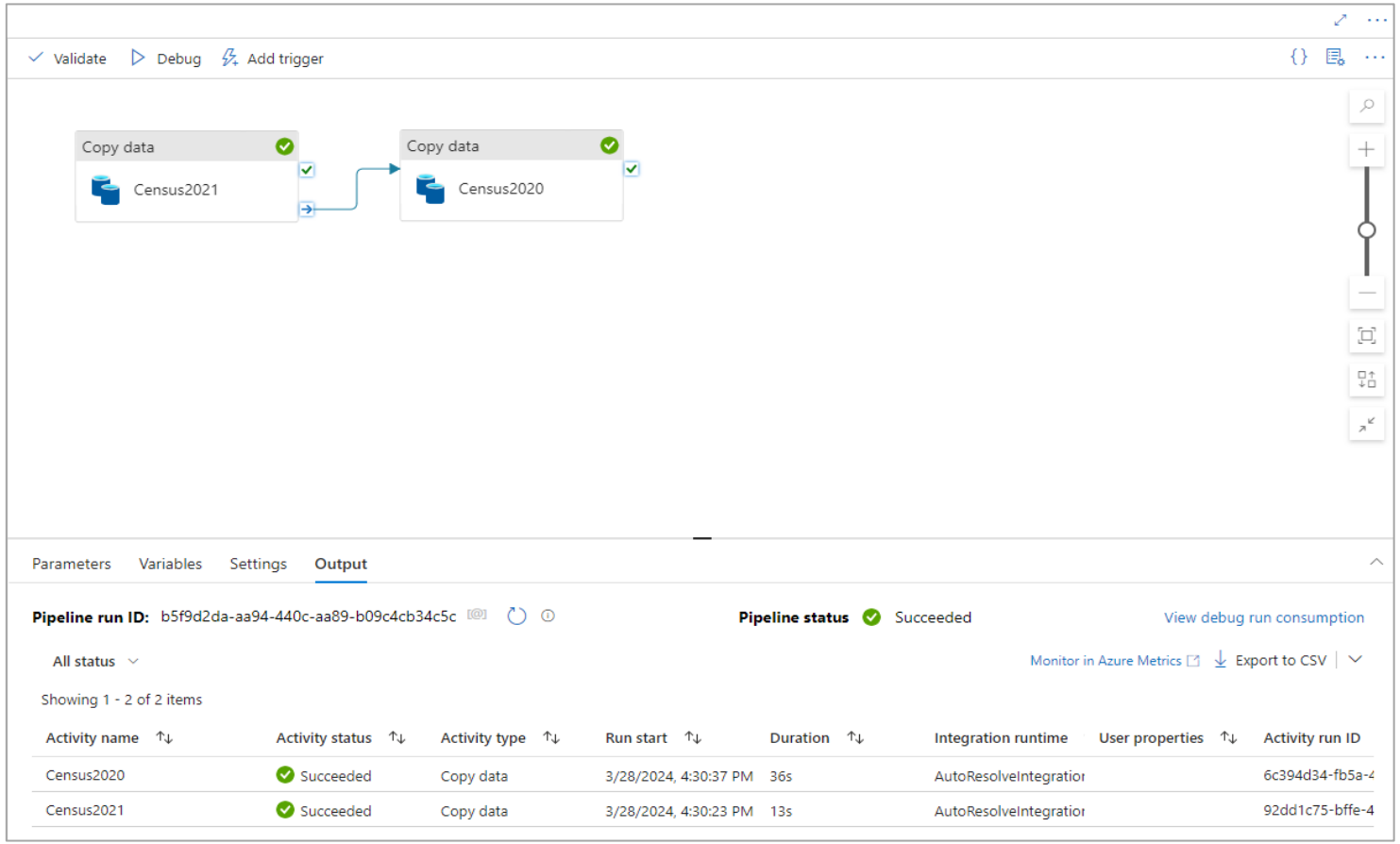
- Drag and drop the blue arrow from the first step to the second.
- This means that after first step completion, will continue with the next one
- Do the same with the other steps



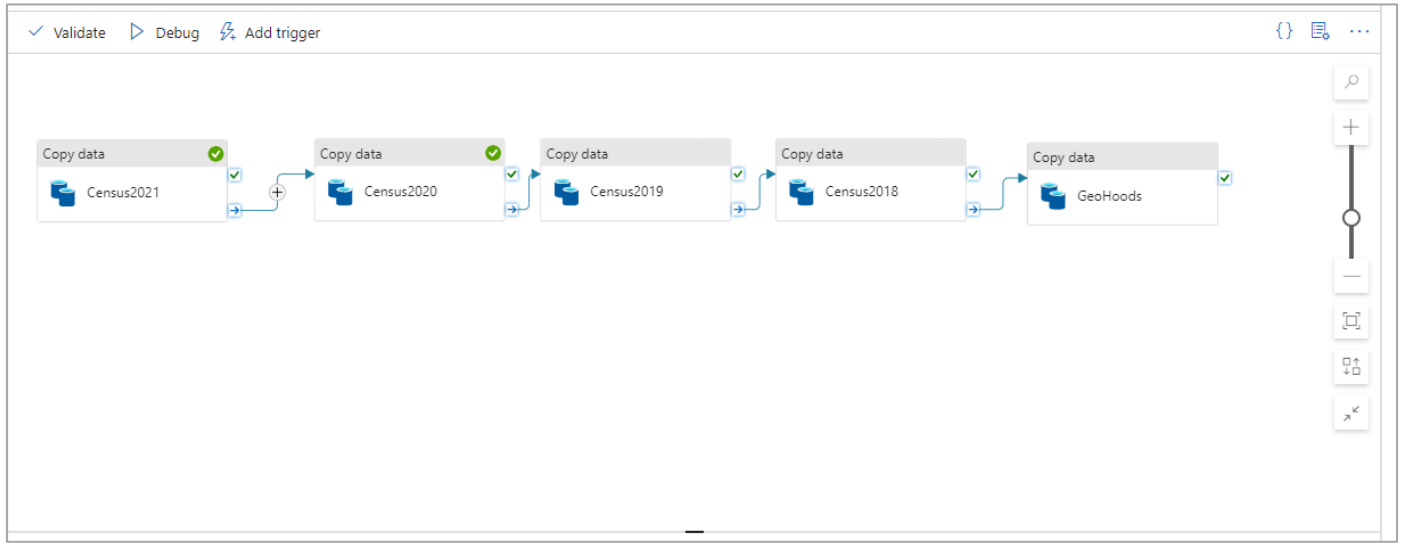
- Validate
- Debug
- See how the pipeline succeeds step by step
- See the .csv in ADLS container

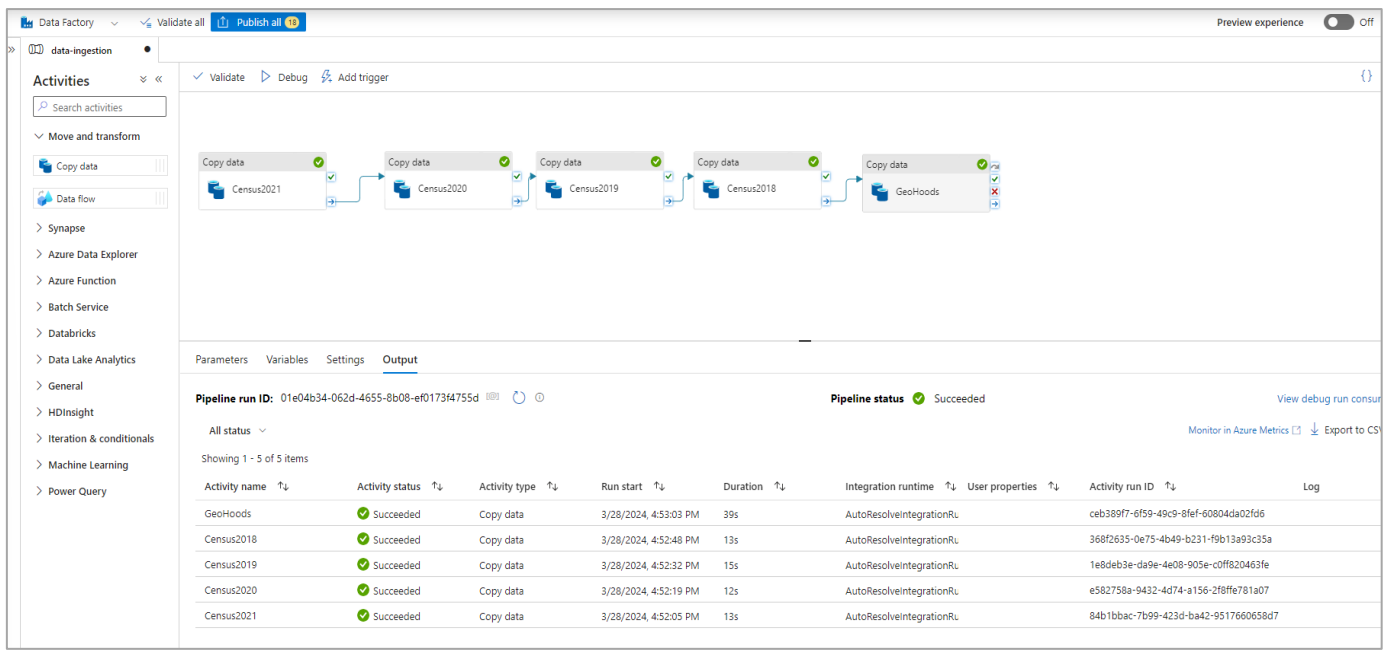


- The two-step pipeline succeeded

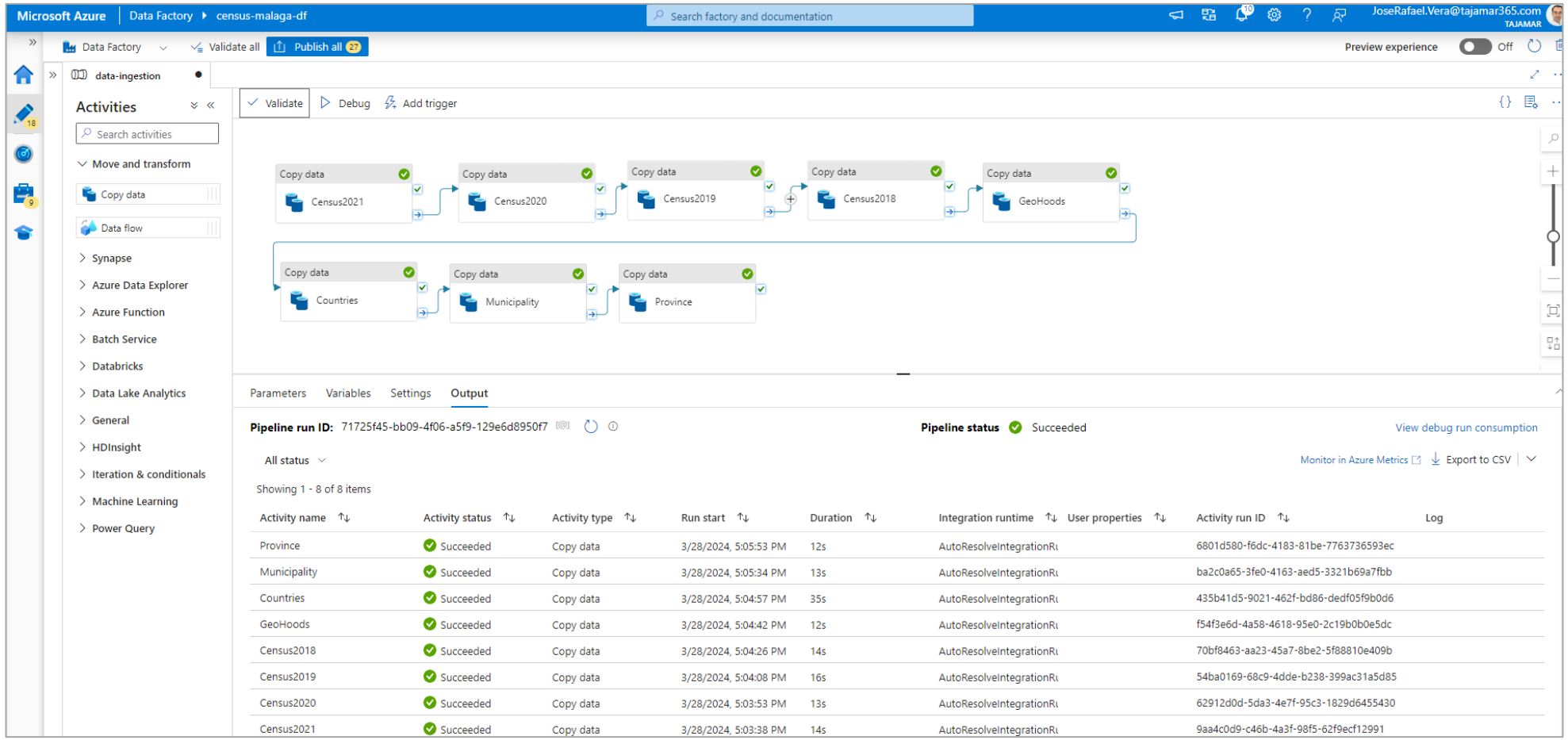


- Add so many steps as .csv files you need to get into your ADLS container in Bronze folder



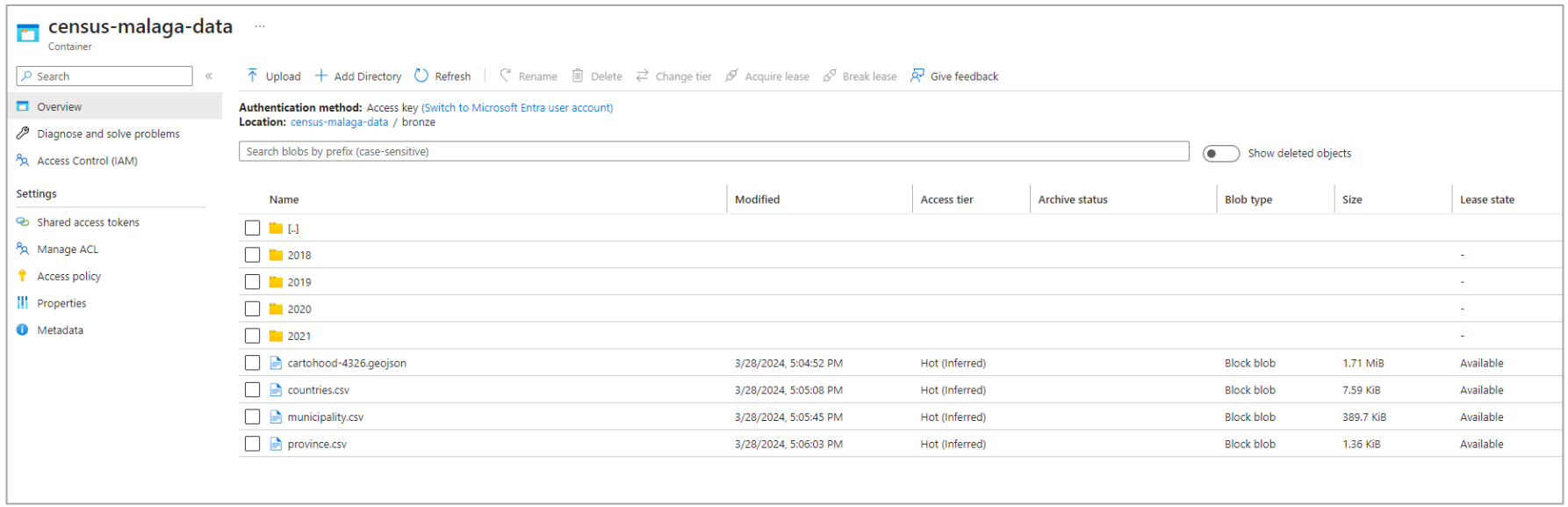



- My complete succeeded pipeline

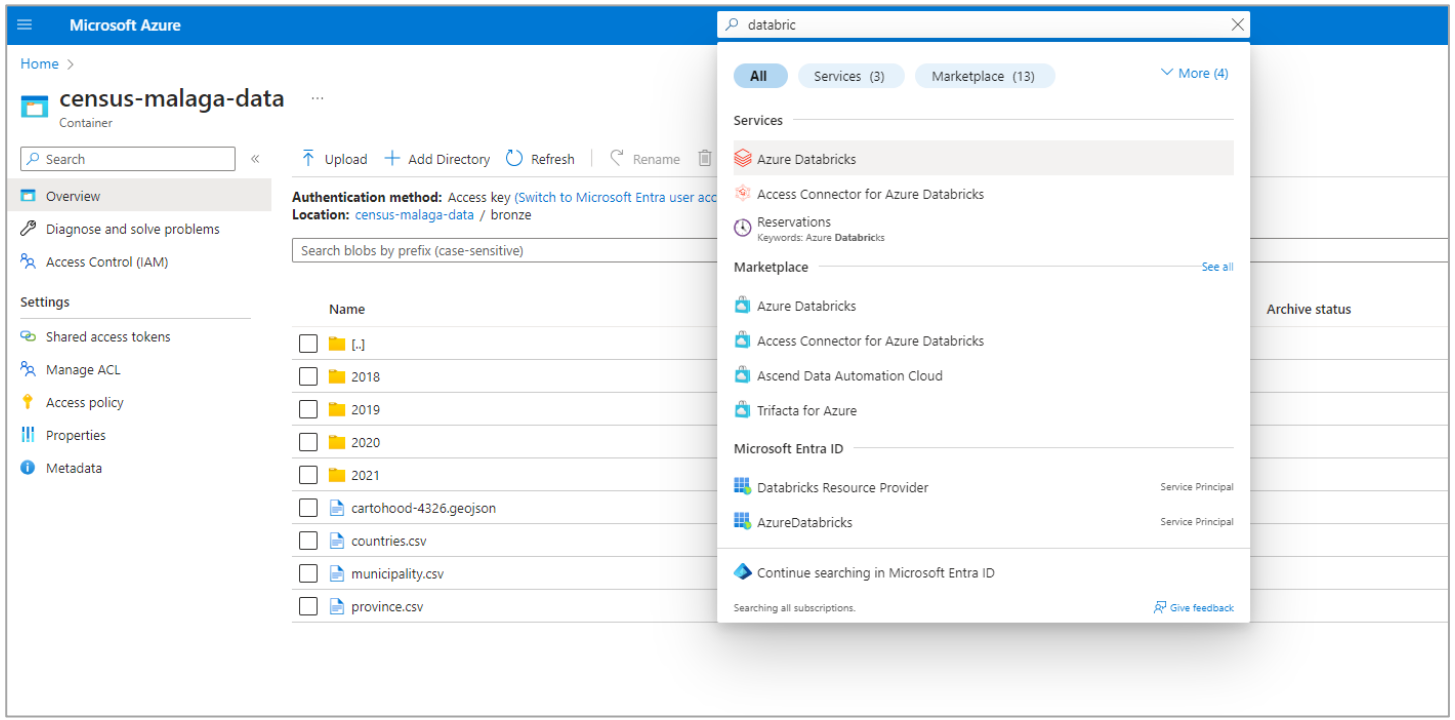



| Activity name | Activity status | Activity type | Run start | Duration | Integration runtime | User properties | Activity run ID | Log |
|---|---|---|---|---|---|---|---|---|
| Province | Succeeded | Copy data | 3/28/2024, 5:05:53 PM | 12s | AutoResolveIntegrationRu | | 6801d580-f6dc-4183-81be-7763736593ec | |
| Municipality | Succeeded | Copy data | 3/28/2024, 5:05:34 PM | 13s | AutoResolveIntegrationRu | | ba2c0a65-3fe0-4163-aed5-3321b69a7fbb | |
| Countries | Succeeded | Copy data | 3/28/2024, 5:04:57 PM | 35s | AutoResolveIntegrationRu | | 435b41d5-9021-462f-bd86-dedf05f9b0d6 | |
| GeoHoods | Succeeded | Copy data | 3/28/2024, 5:04:42 PM | 12s | AutoResolveIntegrationRu | | f54f3e6d-4a58-4618-95e0-2c19b0b0e5dc | |
| Census2018 | Succeeded | Copy data | 3/28/2024, 5:04:26 PM | 14s | AutoResolveIntegrationRu | | 70bf8463-aa23-45a7-8be2-5f88810e409b | |
| Census2019 | Succeeded | Copy data | 3/28/2024, 5:04:08 PM | 16s | AutoResolveIntegrationRu | | 54ba0169-68c9-4dde-b238-399ac31a5d85 | |
| Census2020 | Succeeded | Copy data | 3/28/2024, 5:03:53 PM | 13s | AutoResolveIntegrationRu | | 62912d0d-5da3-4e7f-95c3-1829d6455430 | |
| Census2021 | Succeeded | Copy data | 3/28/2024, 5:03:38 PM | 14s | AutoResolveIntegrationRu | | 9aa4c0d9-c46b-4a3f-98f5-62f9ecf12991 | |

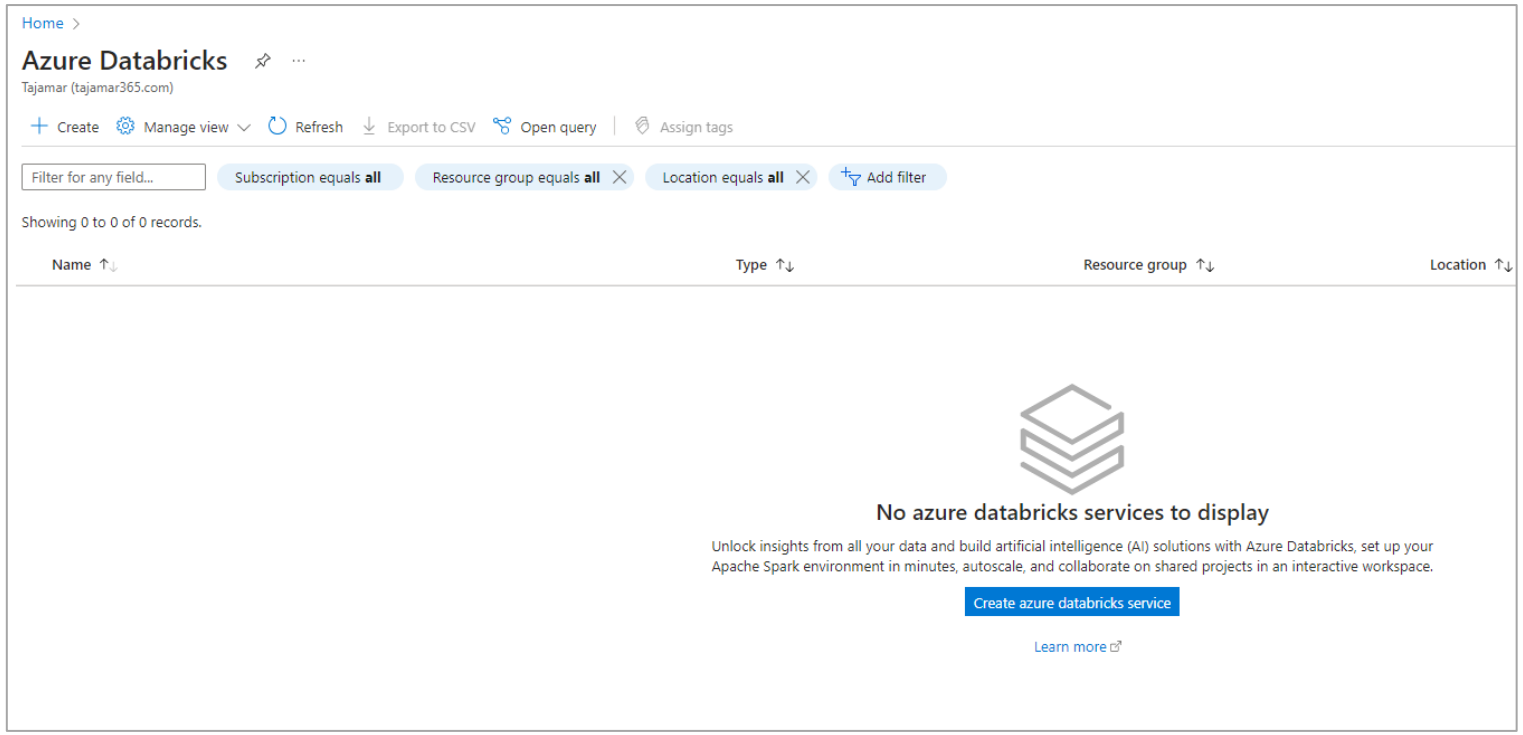
- And all the .csv files saved into Bronze Layer



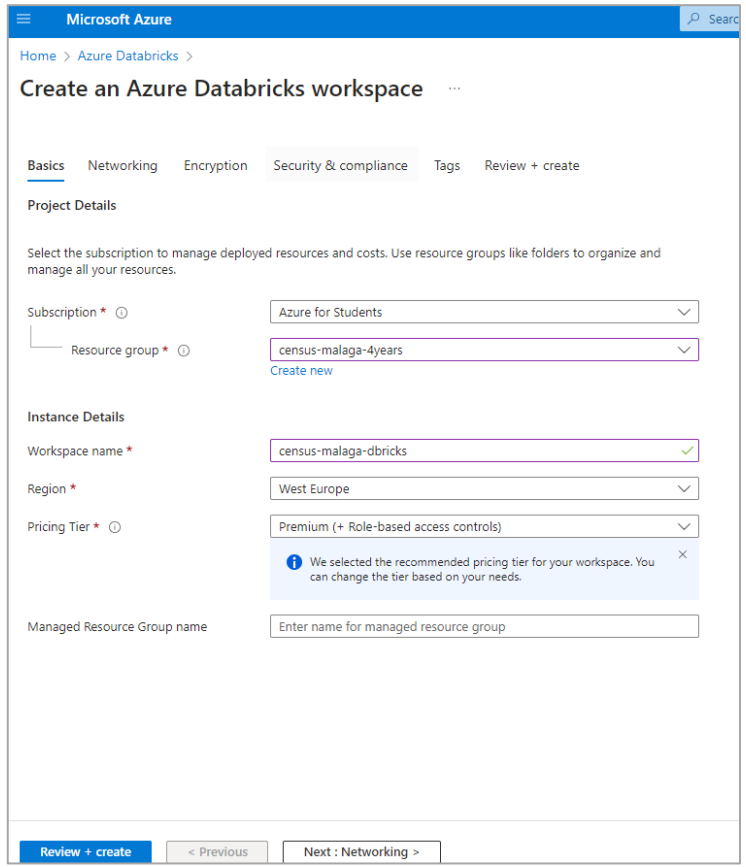
- Search for Databricks

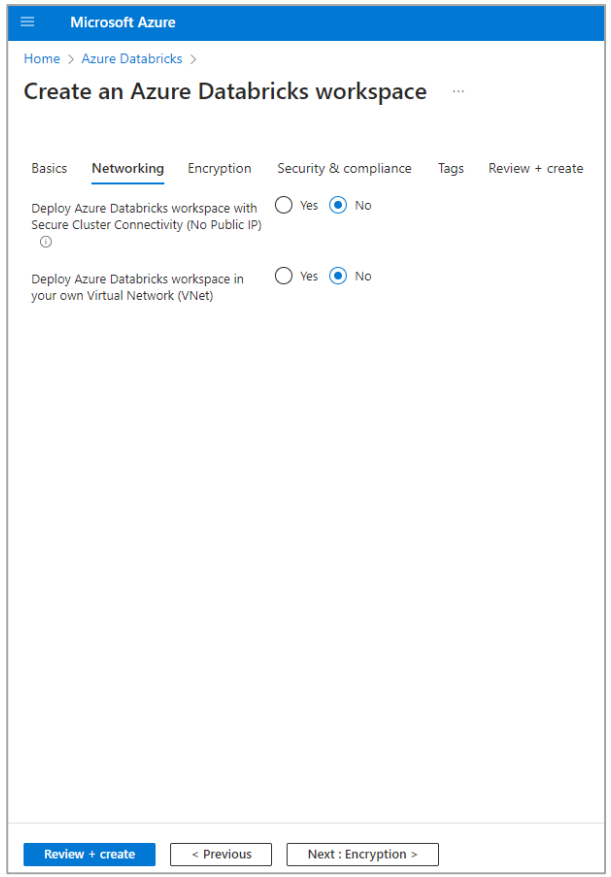


- Create a new Databricks service

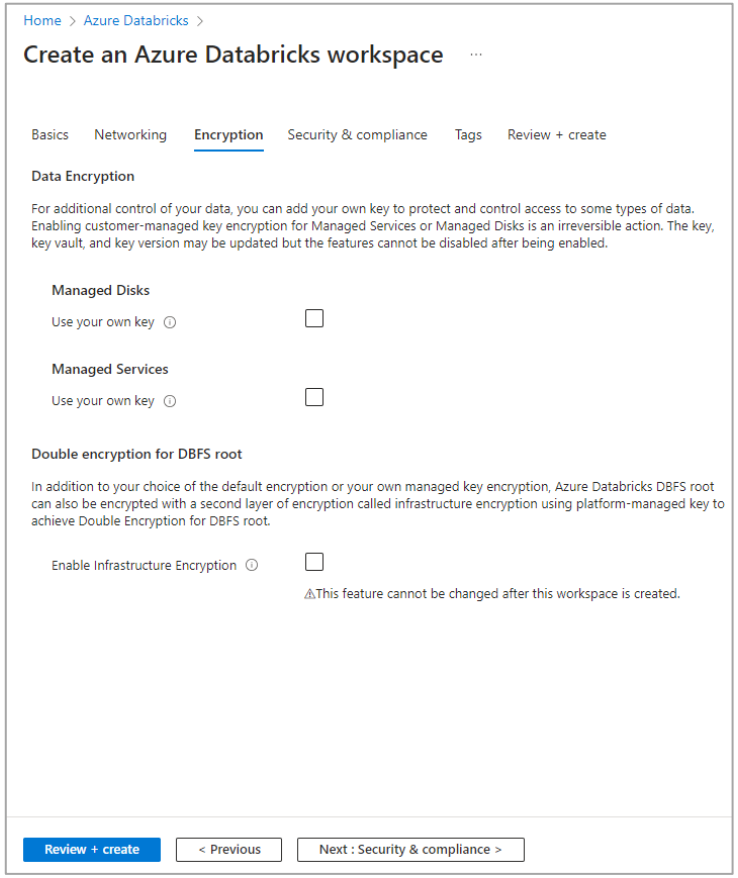


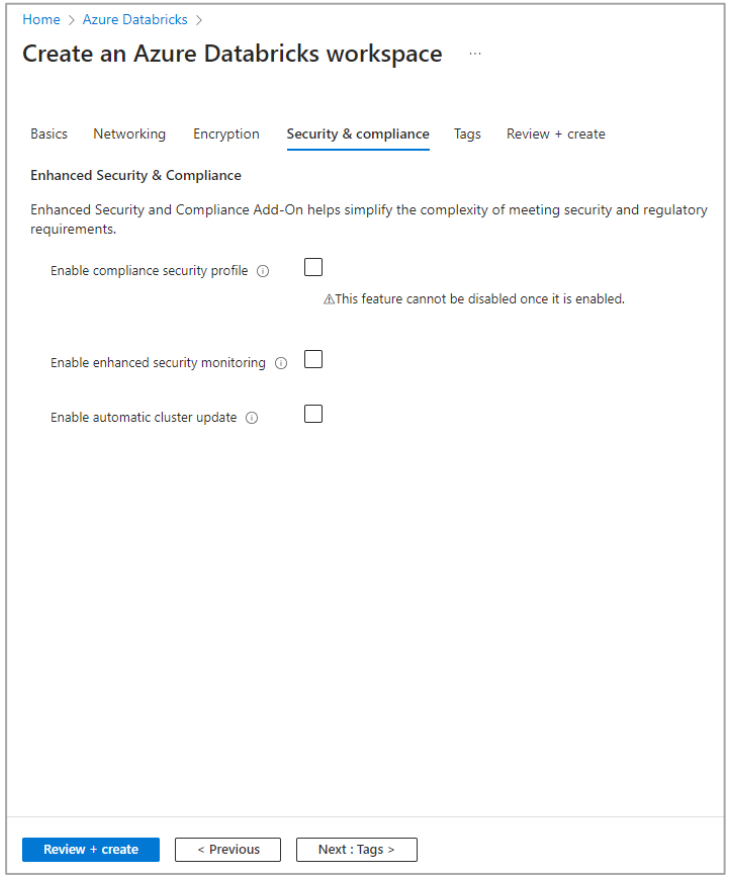
- Select your free subscription
- Select your preexisting Resource group
- Give Databrick service a name
- Select a Region. SouthEast Asia recommended
- Pricing Tier: “premium” > you won’t be charged. You are under your subscription and this Tier ables full-service access.
- next



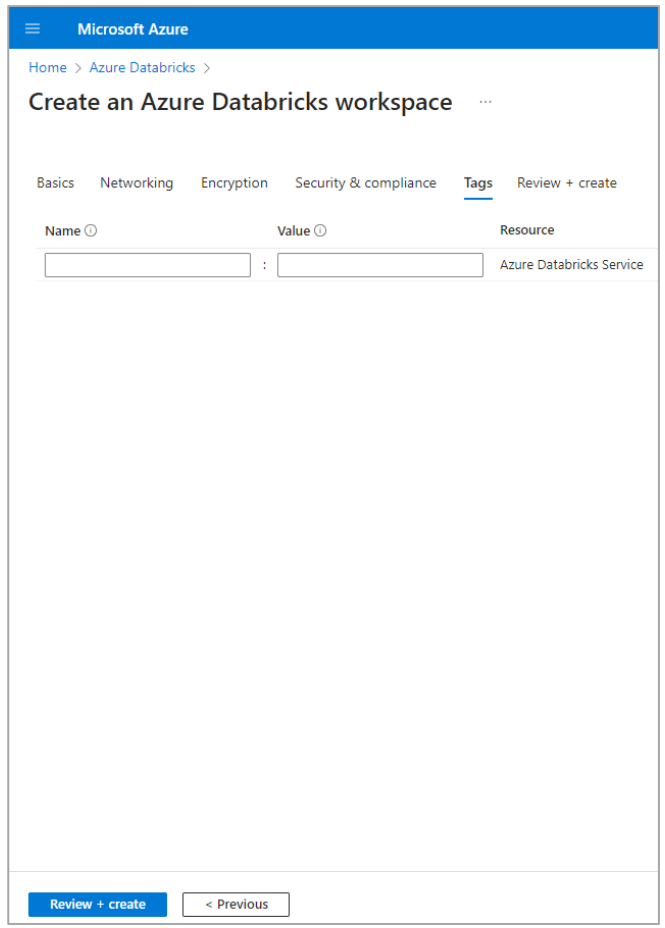



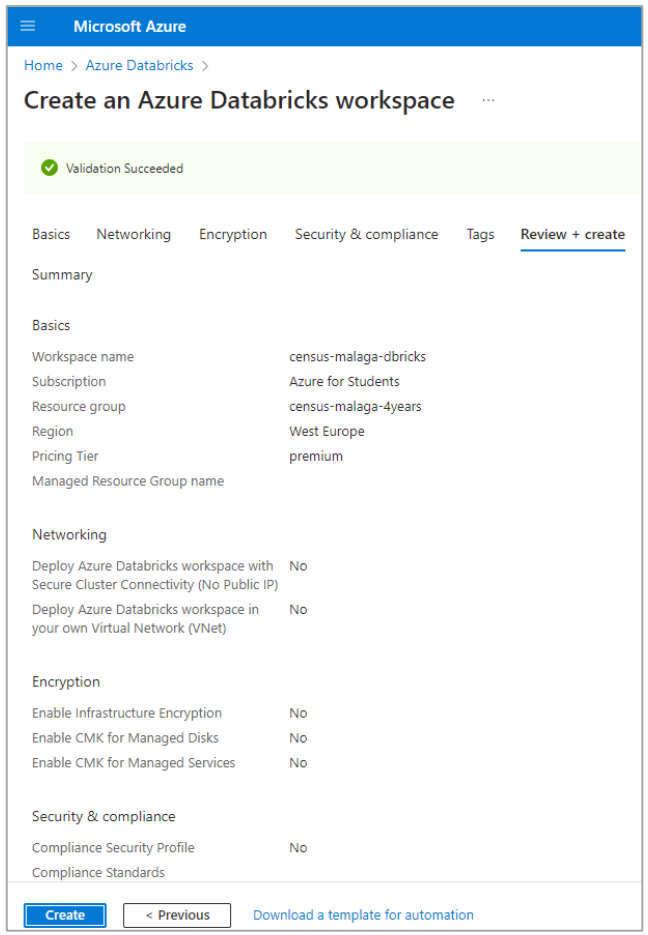
- Next
- Next


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 >

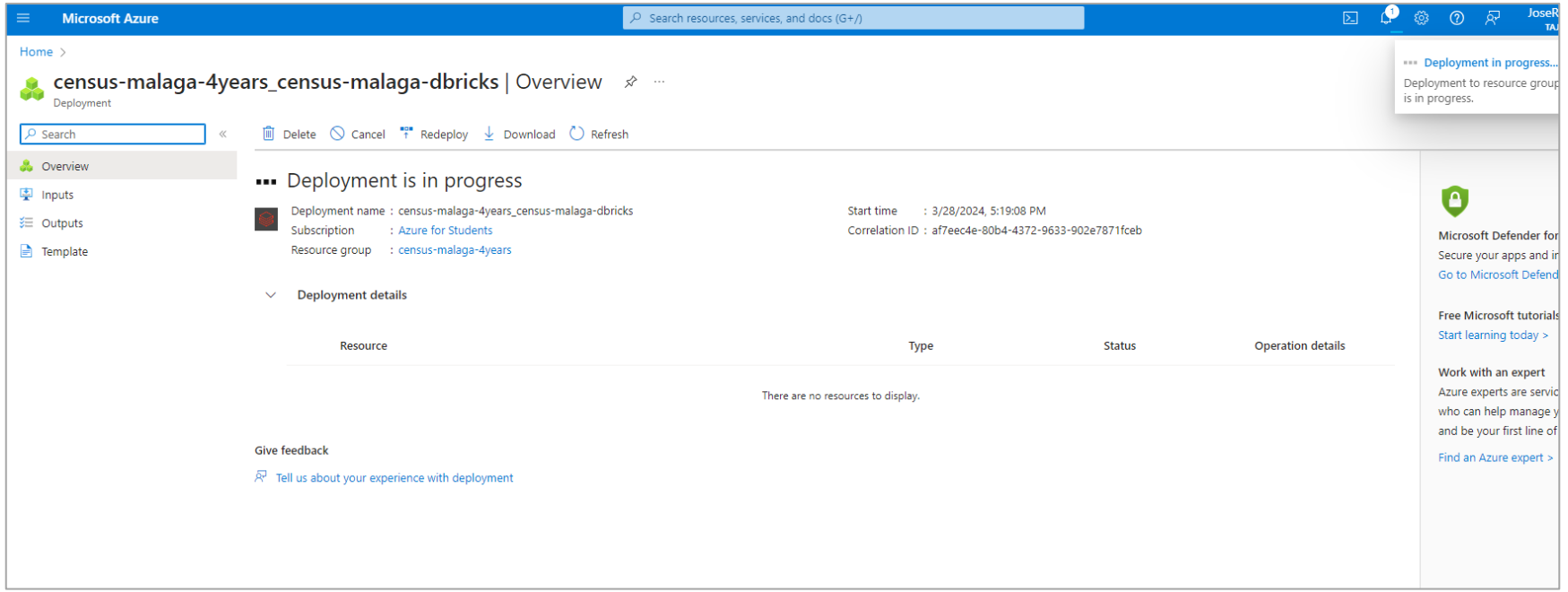

Home > Azure Databricks >
Create an Azure Databricks workspace
Basics Networking Encryption Security & compliance Tags Review + create
Enhanced Security & Compliance
Enhanced Security and Compliance Add-On helps simplify the complexity of meeting security and regulatory requirements.
Enable compliance 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

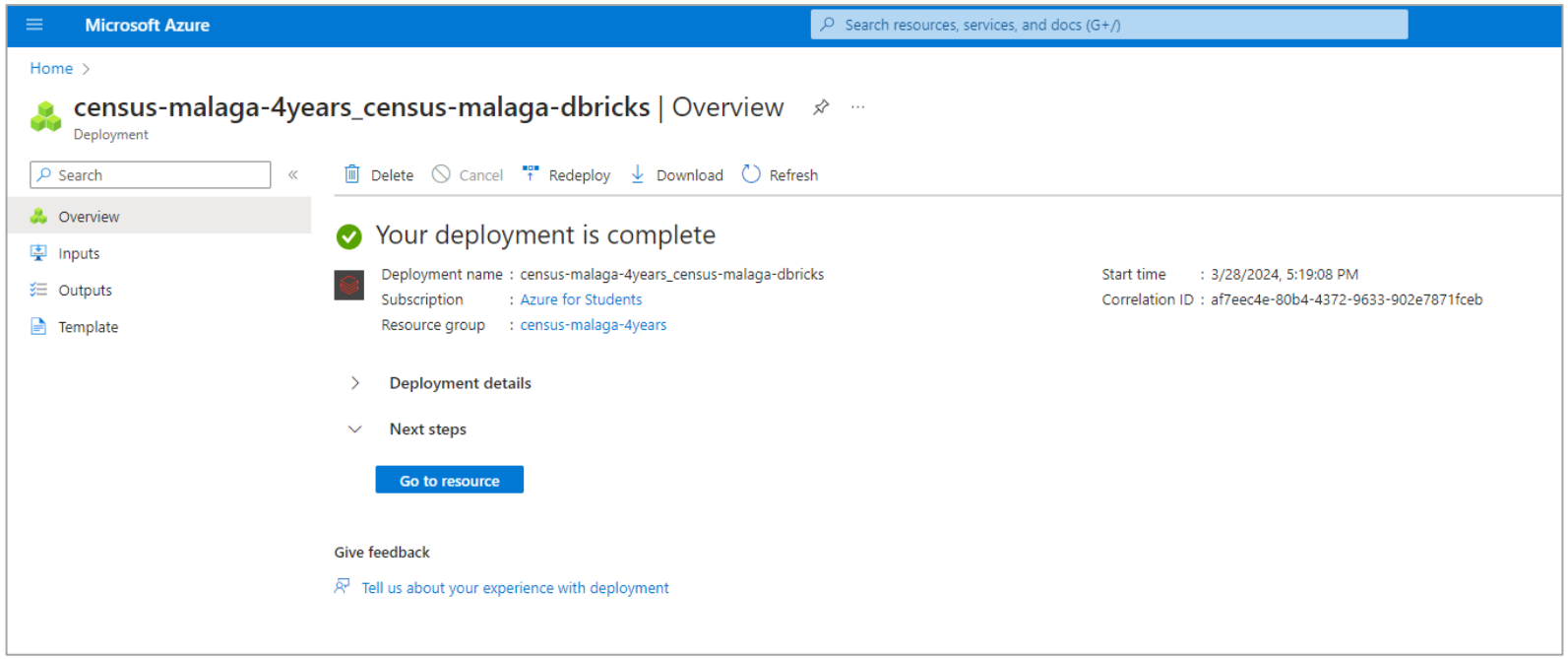

Microsoft Azure
Home > Azure Databricks >
Create an Azure Databricks workspace
Basics Networking Encryption Security & compliance Tags Review + create
Name
Value
Resource
Azure Databricks Service
Review + create
< Previous


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

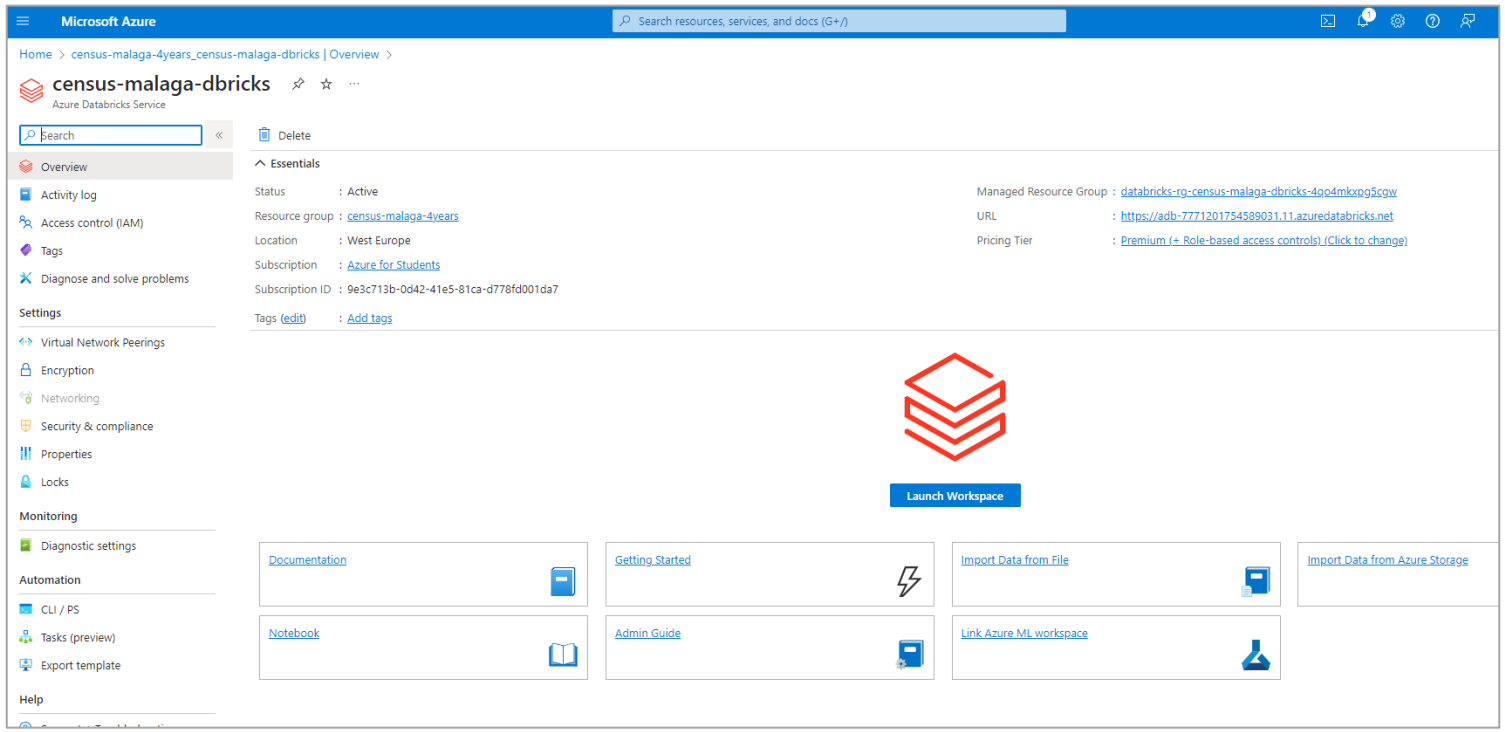
- Wait until deployed



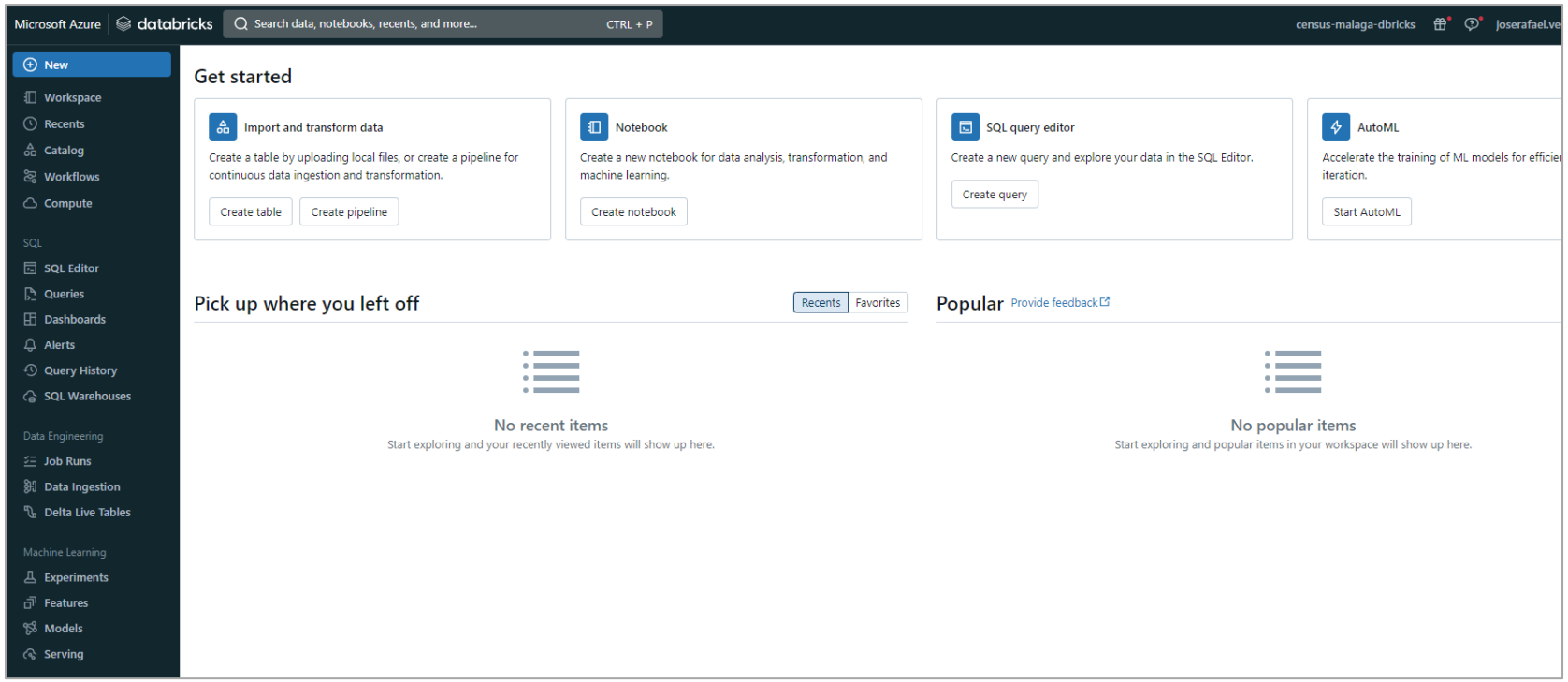
- Deployed
- Go to resource



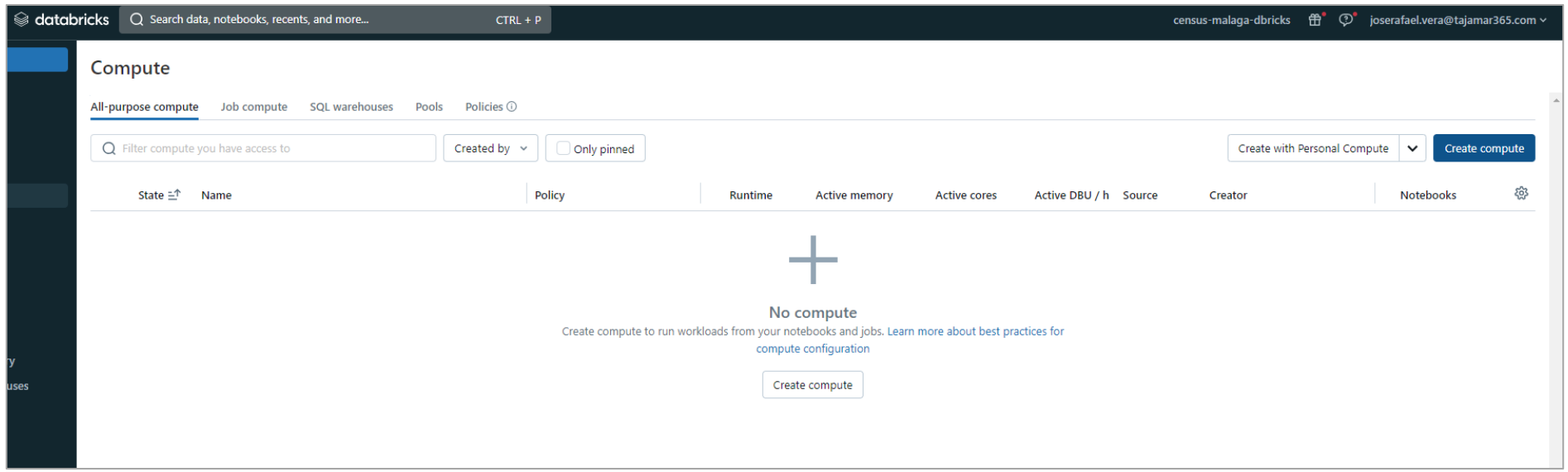
- Launch workspace



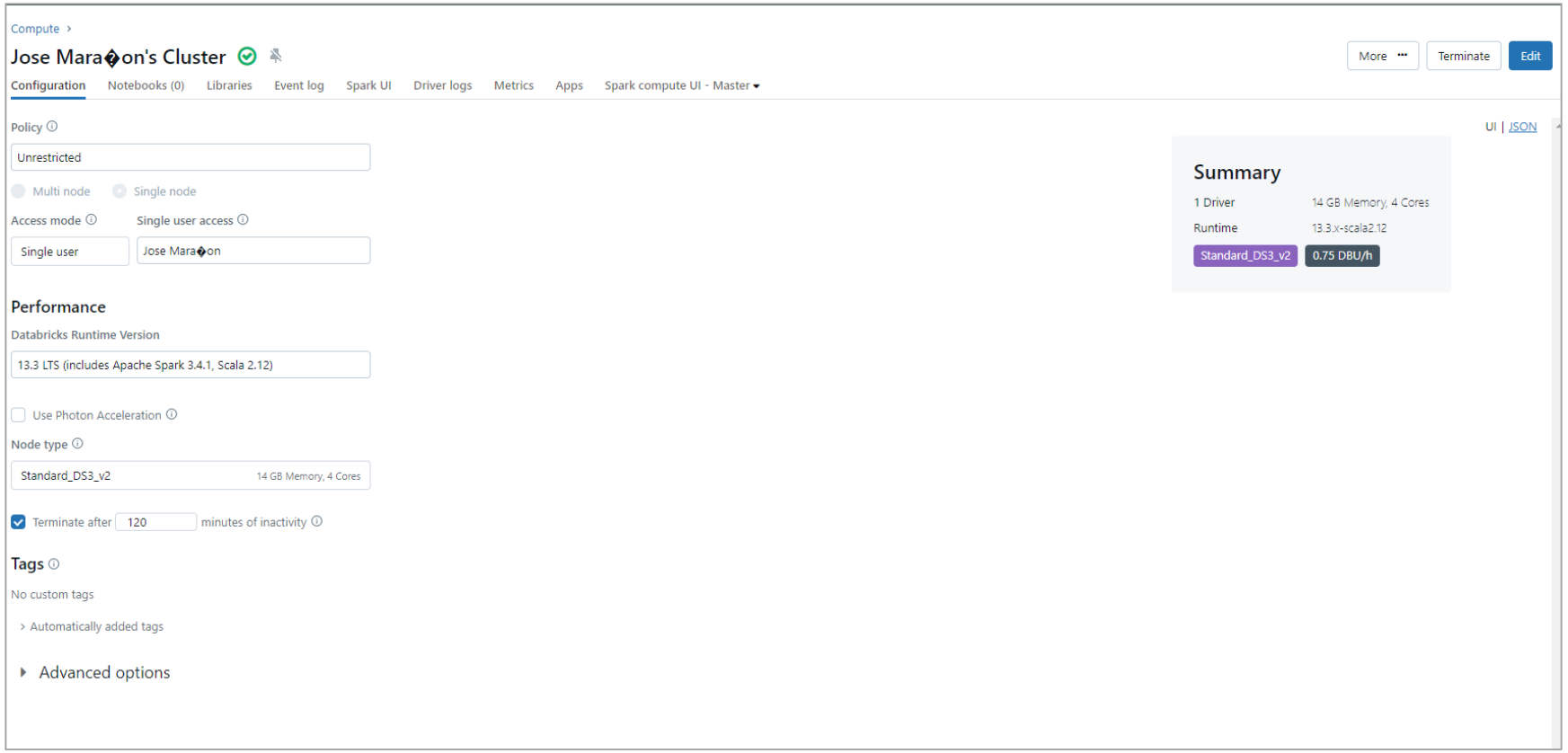
- You are in Databrick workspace already
- Click on “compute” to create the Spark Cluster



- Create compute

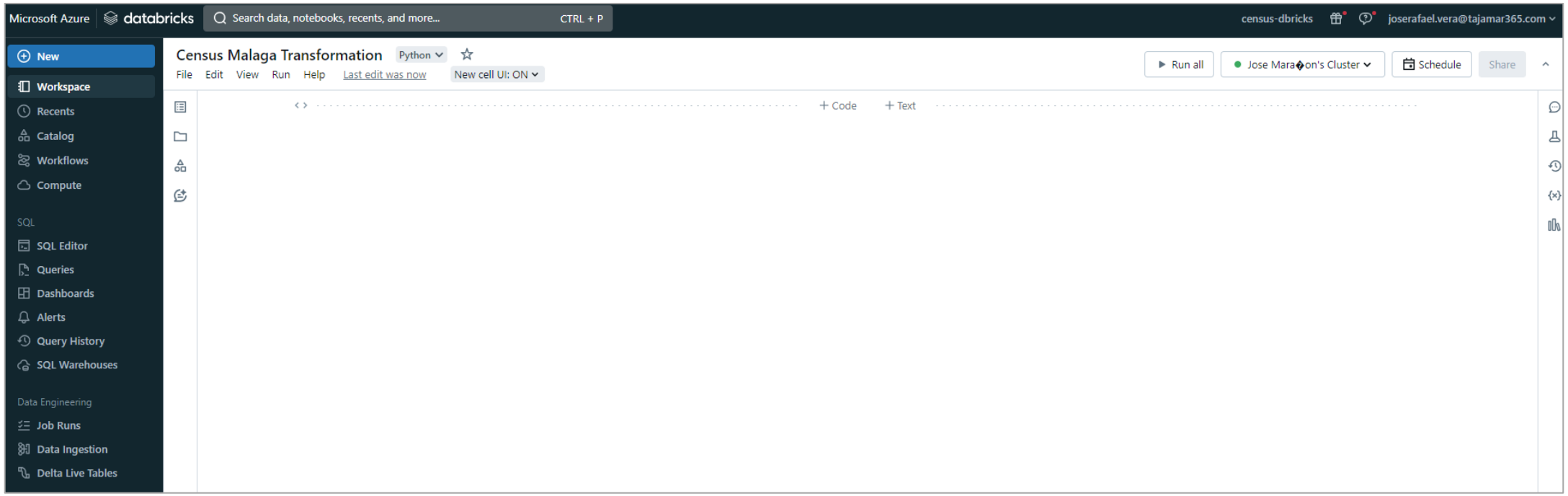


- In my case, for this demo I used a single node but you can choose more nodes if needed
- The Runtime version I left it by default
- Node type the cheapest
- Region: SoutheEast Asia, otherwise you will not have many node type available (VM types indeed).
- It takes some time until the cluster is up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

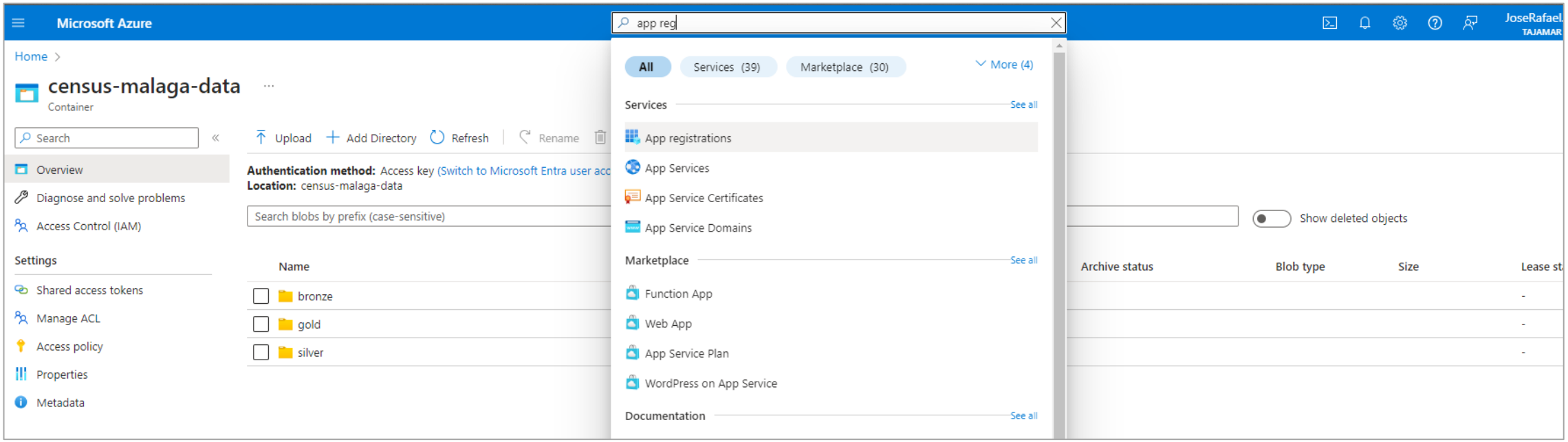


- Create a new Notebook

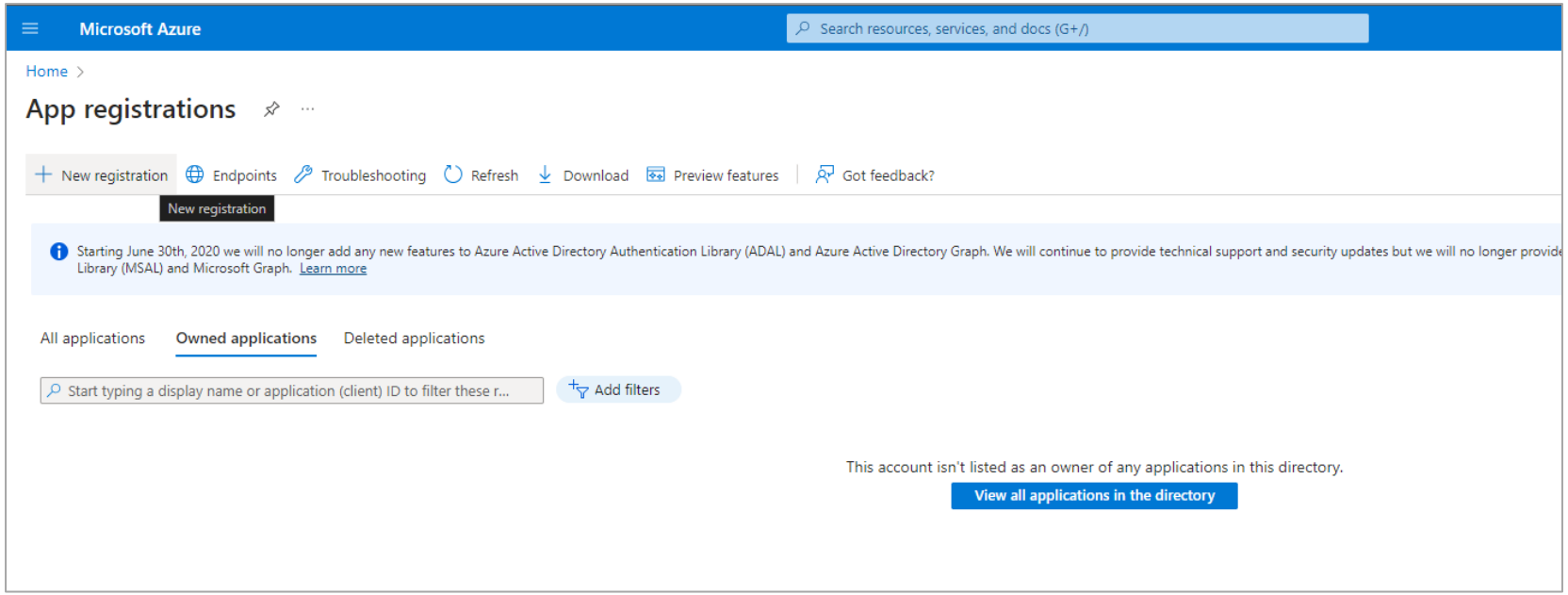
- The notebook ready to be written
- In order to get access from Databricks to our data in ADLS (Storage Account) it is necessary to link Databicks to an APP Registration which offer the needed credentials using OAuth2.0 protocol for Azure AD and thus grant authn and authz for Databricks to the Storage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



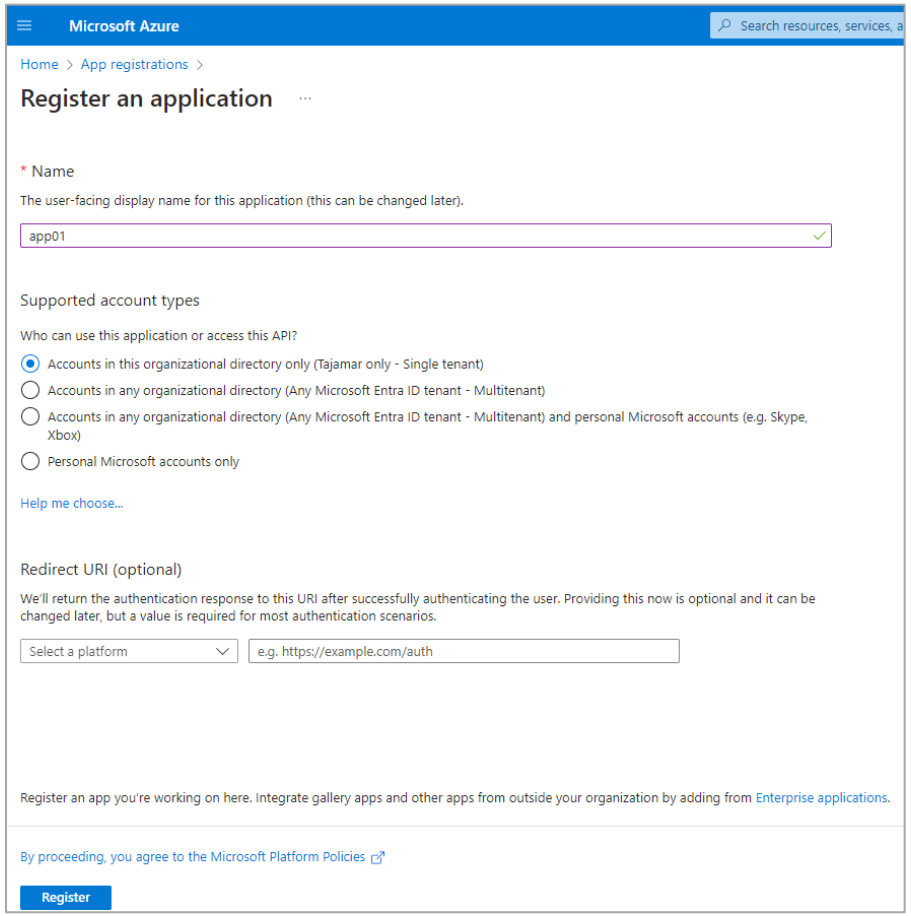
- Search for App registrations



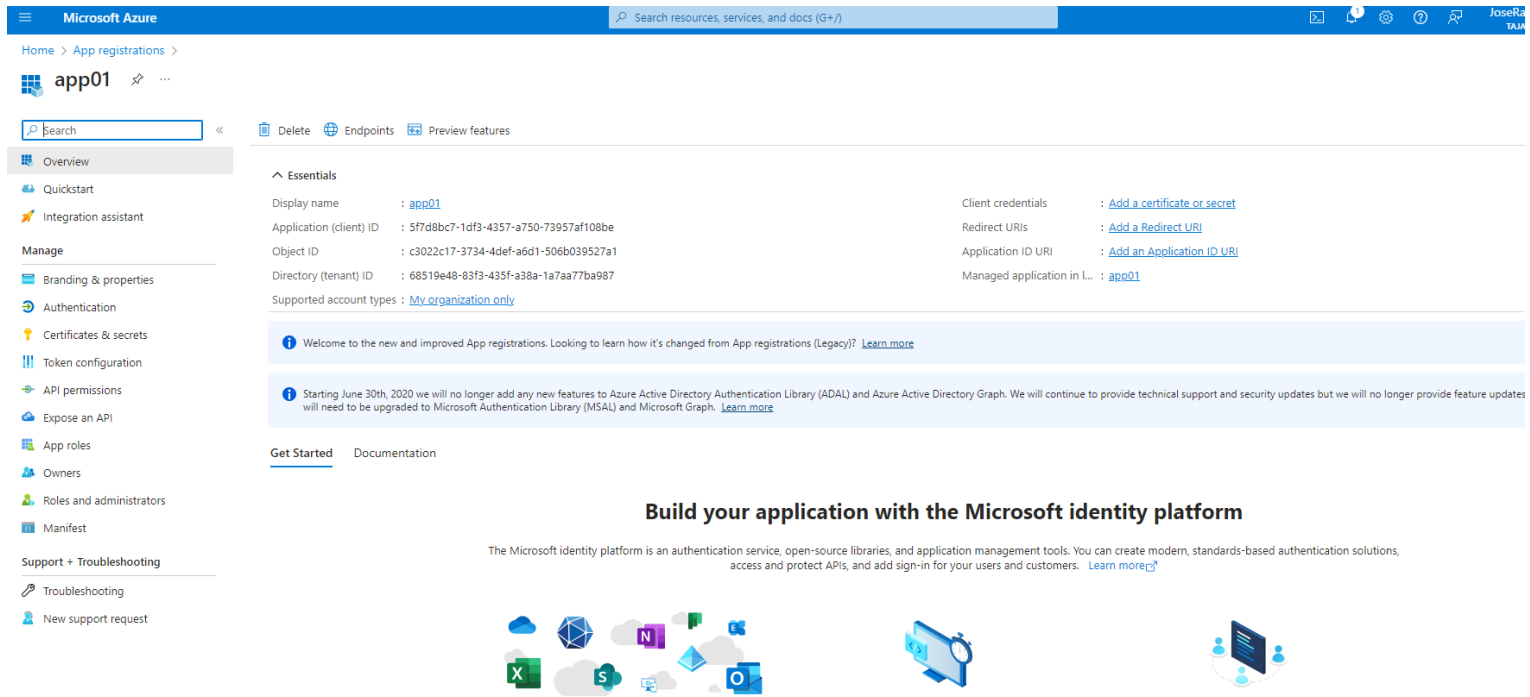
- Create New Registration



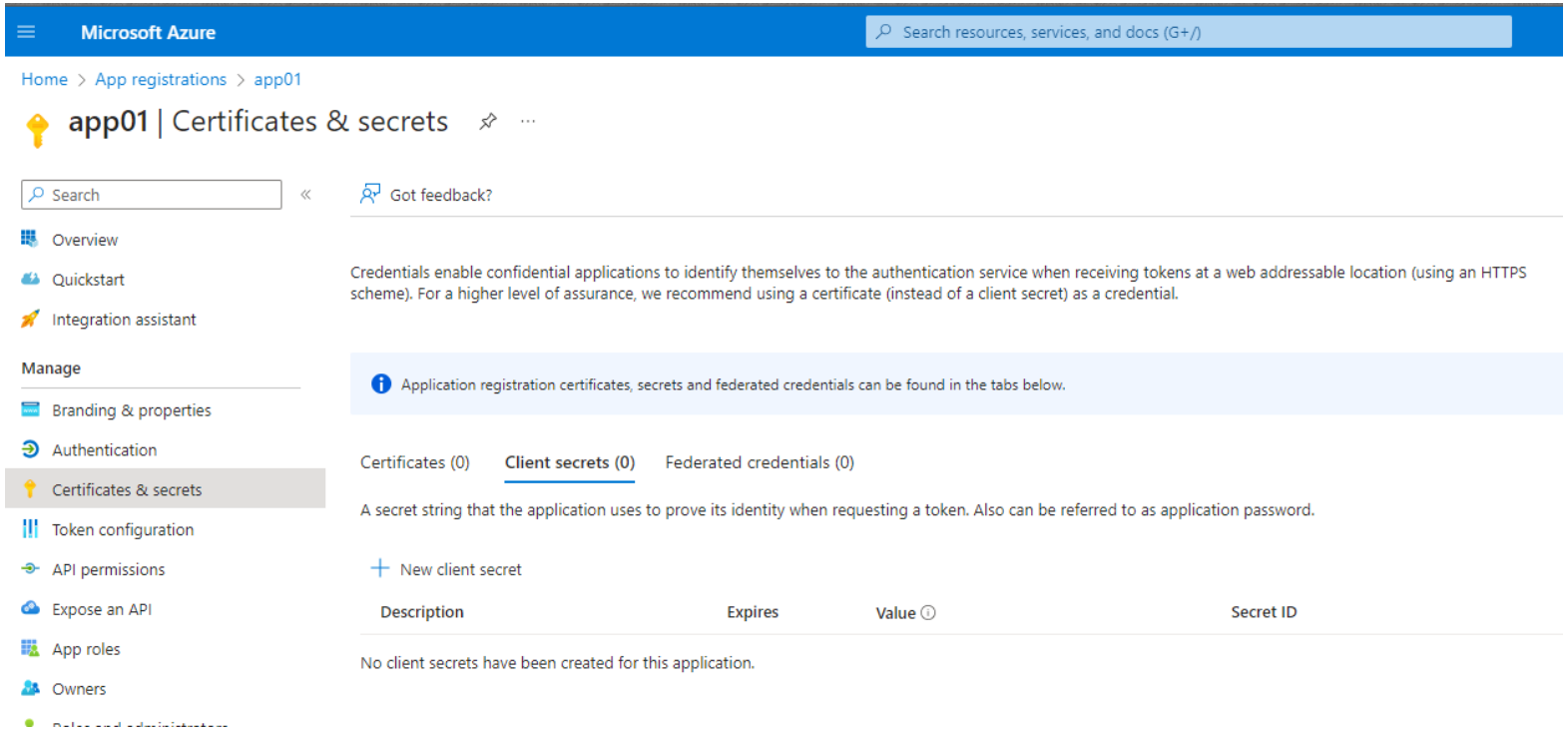
- Give the app registration a name of your choice
- Access to this API: first option
- Register



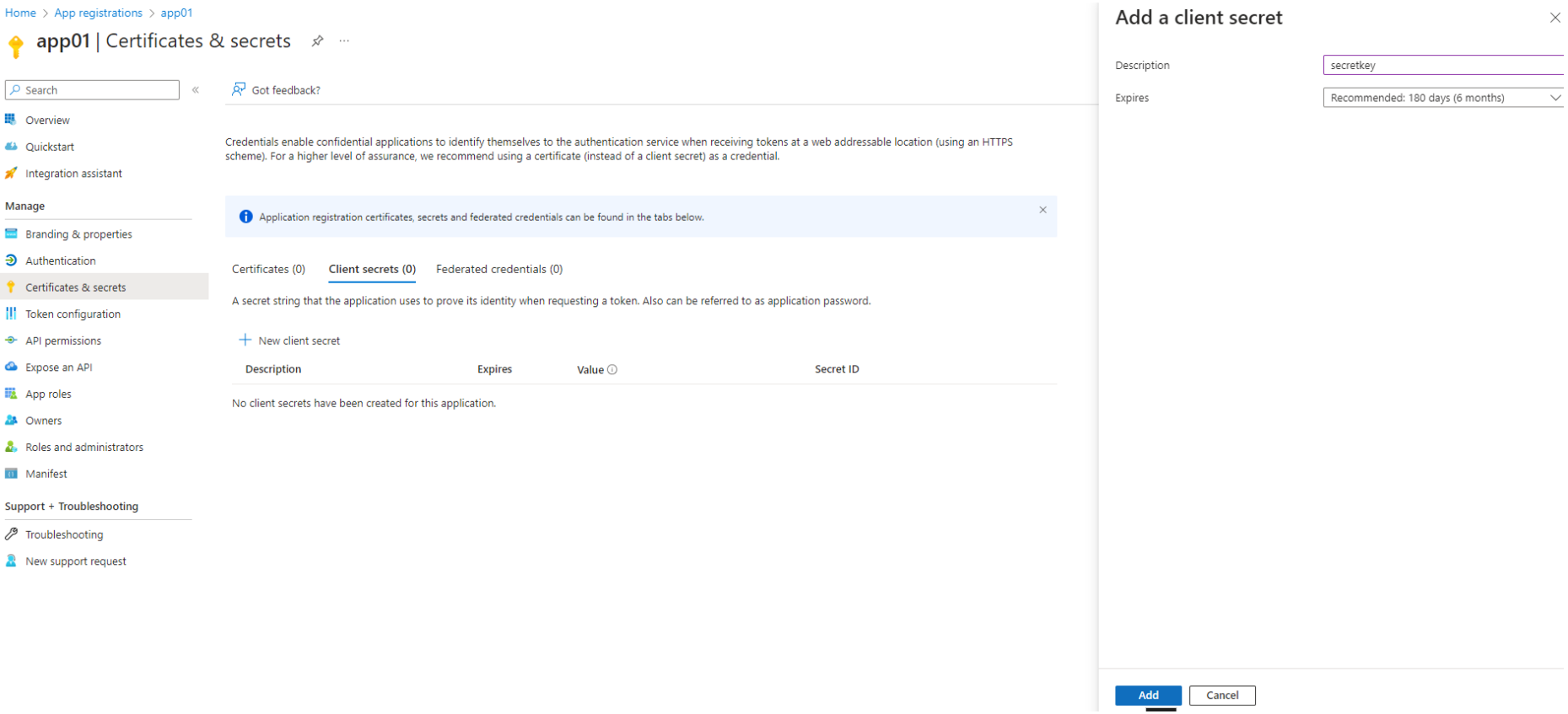
- App registered successfully
- Click on “ Certificates & secrets”



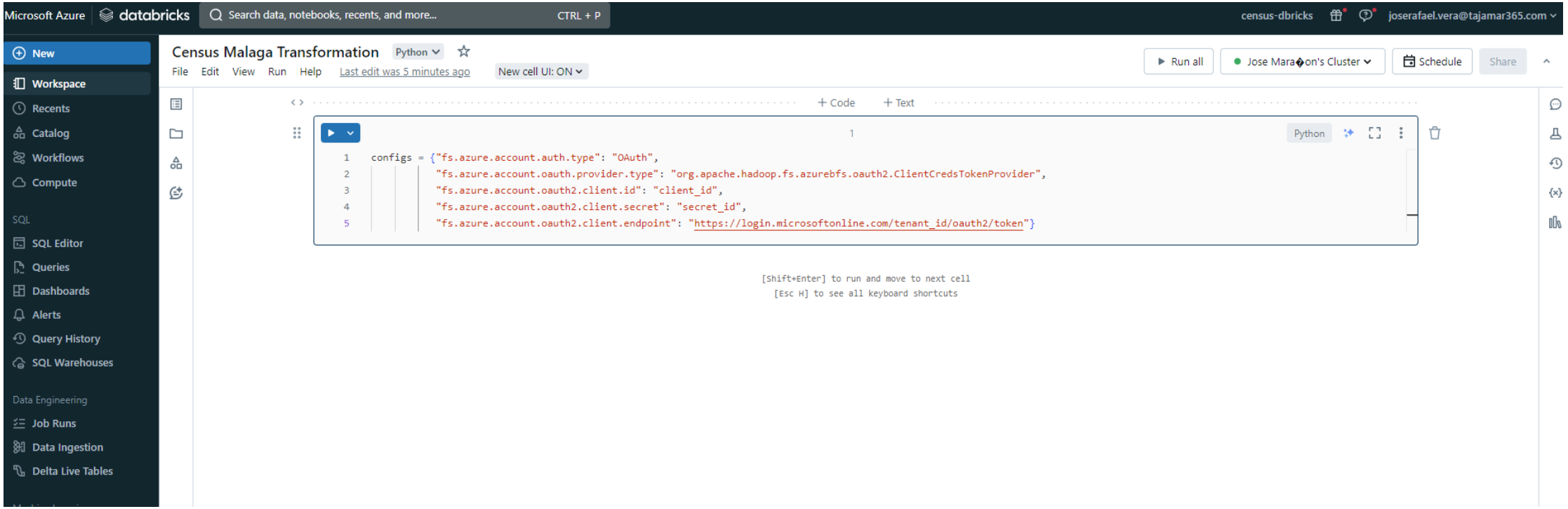
- Navigate to “client secrets”
- Click on “New client secret”



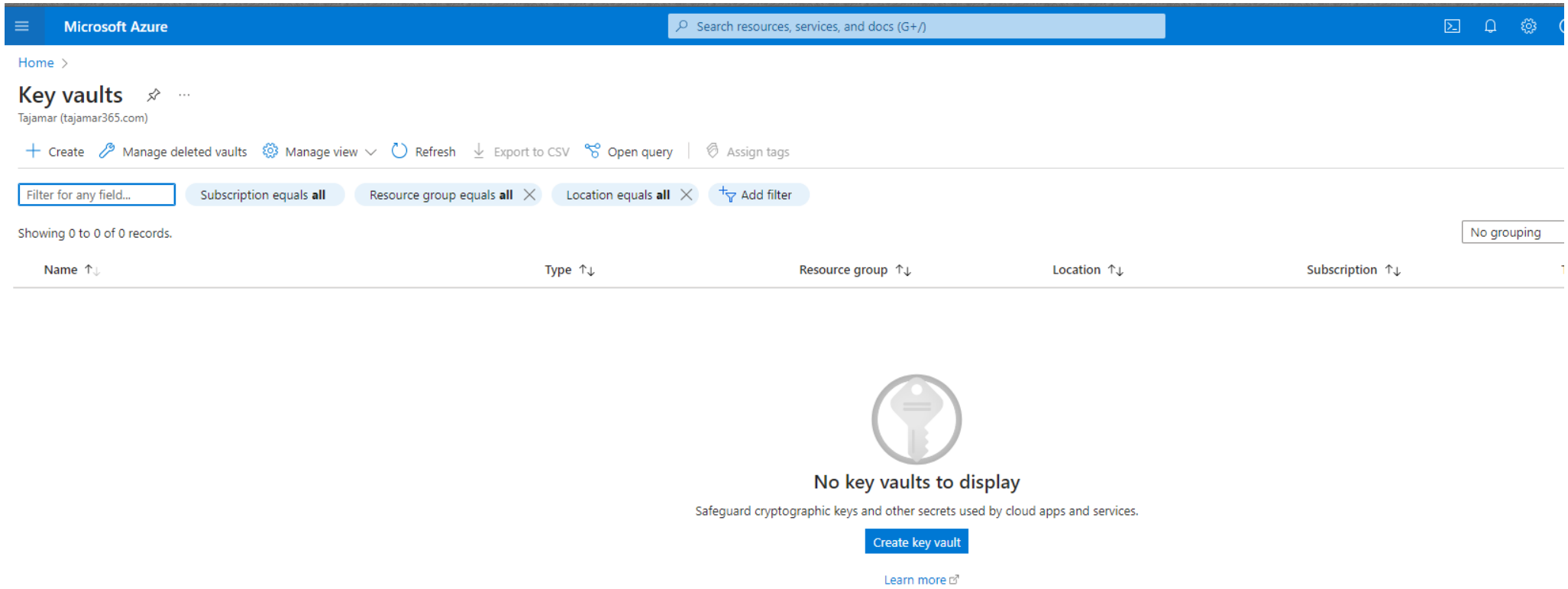
- Give the secretkey a name of your choice
- Expiration by default
- Add



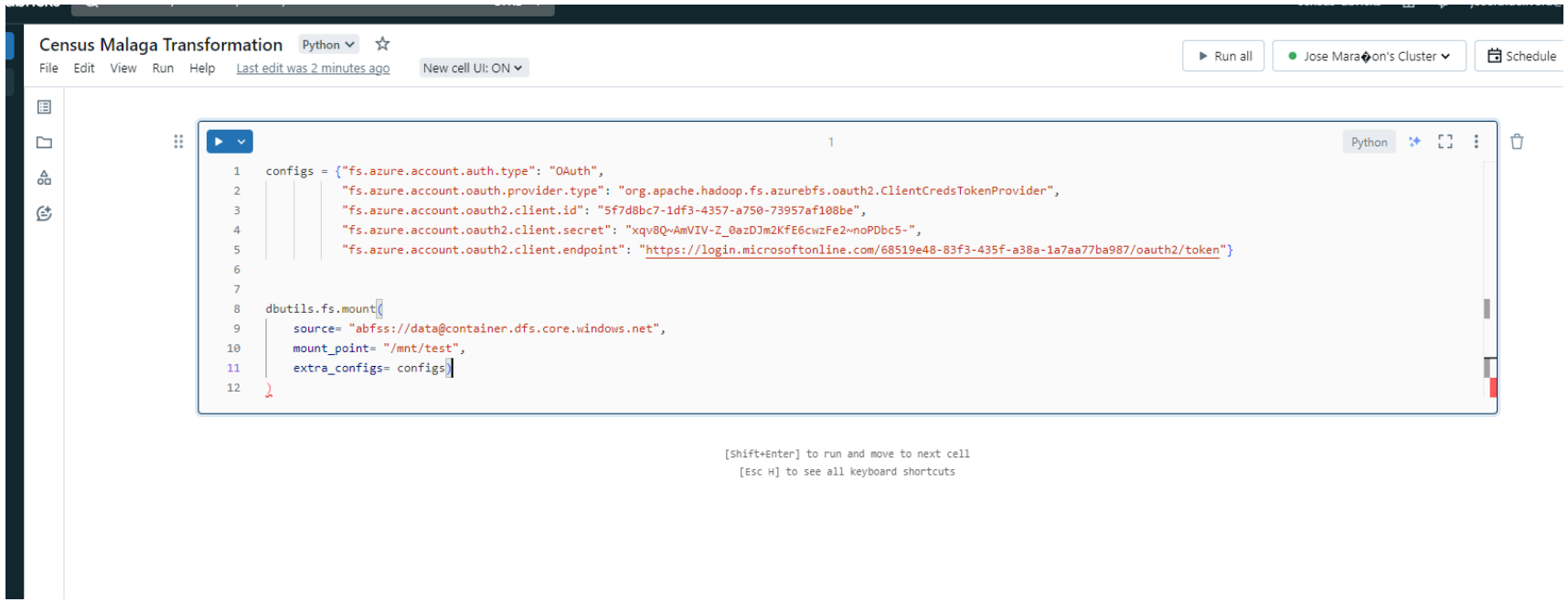
- The secrets created in the step before is copy pasted here
- WARNING! Bad practice. Never expose your credentials within your code. Handle your secrets properly with “Key Vault”



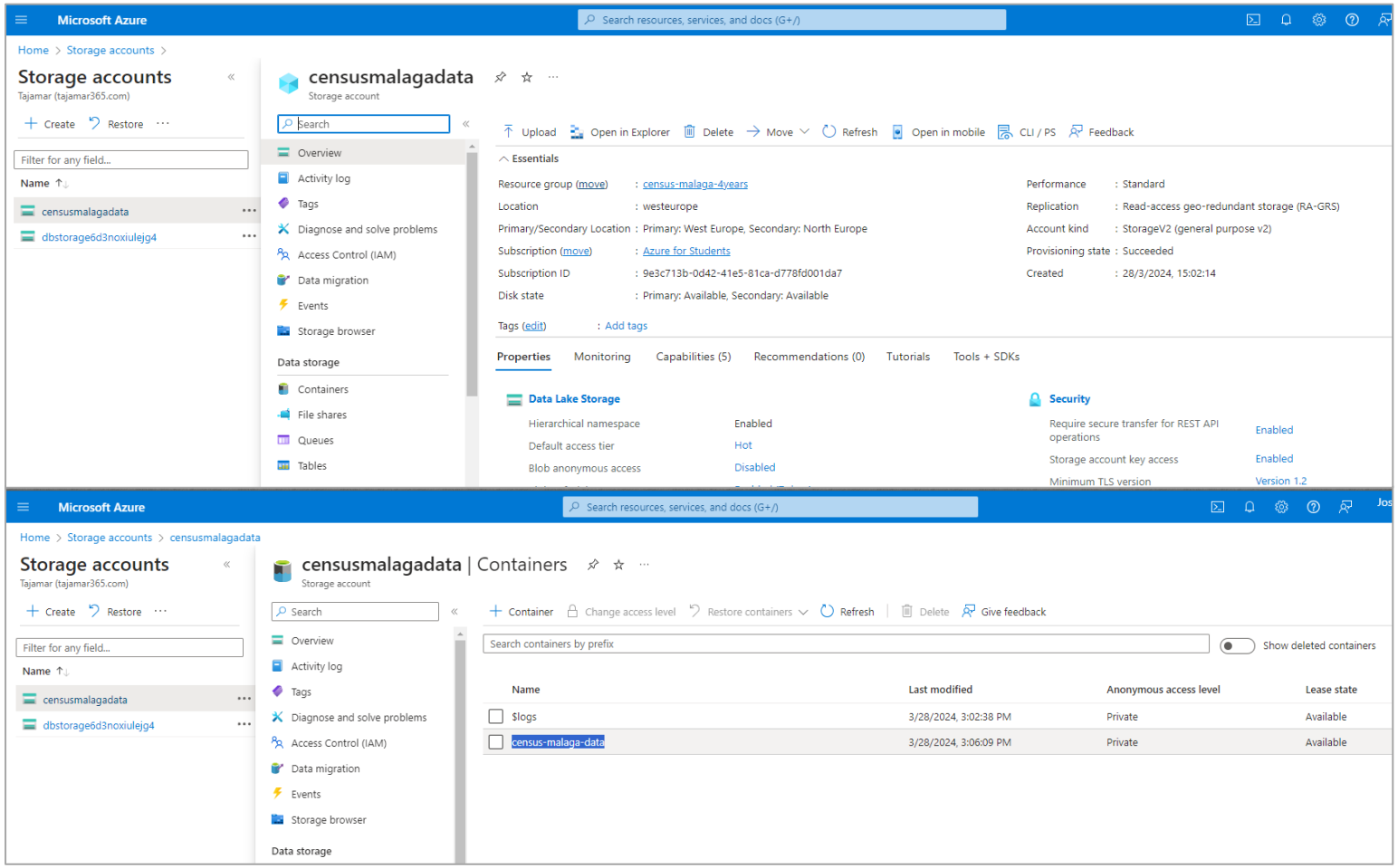
- Search for “Key Vault”
- Create key vault
- I didn’t implement this step as the Secrets used in this project were deleted after completion. All the services as well.
- Keep this Key vault in mind for Production environment



- 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



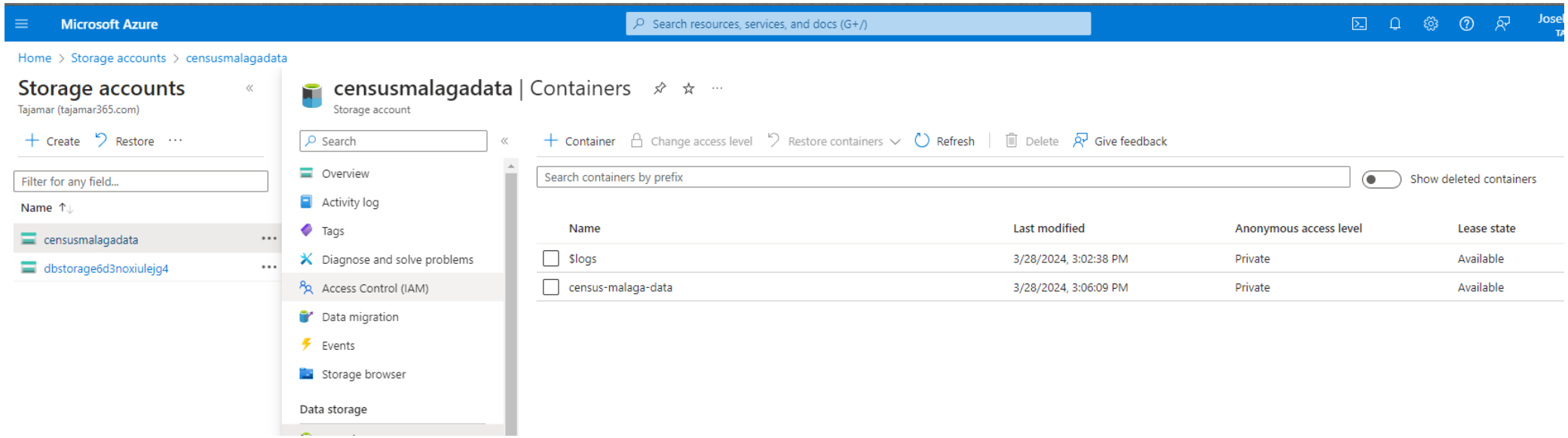
- Go to your Storage account and copy the storage account Name and paste it in Databricks
- Go to containers and copy the container name and paste it in Databricks



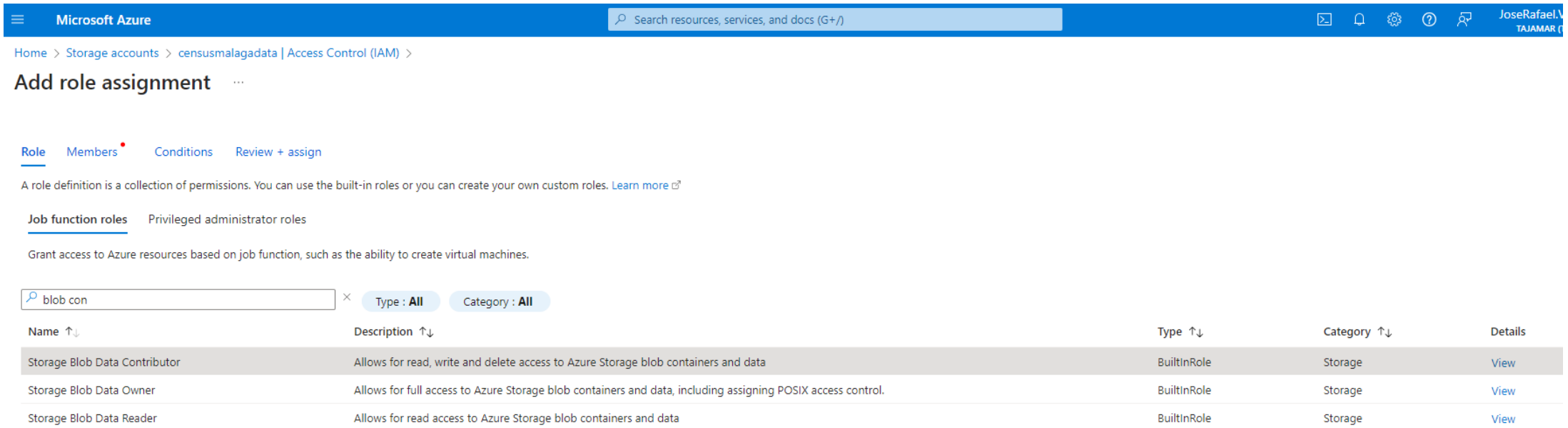
- There you go! You have all the needed information to:
  - Connect Databricks access to the Storage account
  - Mount a directory into Databricks based on the existing container

- Give your mount point a name
- Before run it, you need to set access role permission to the Storage Account. See next step

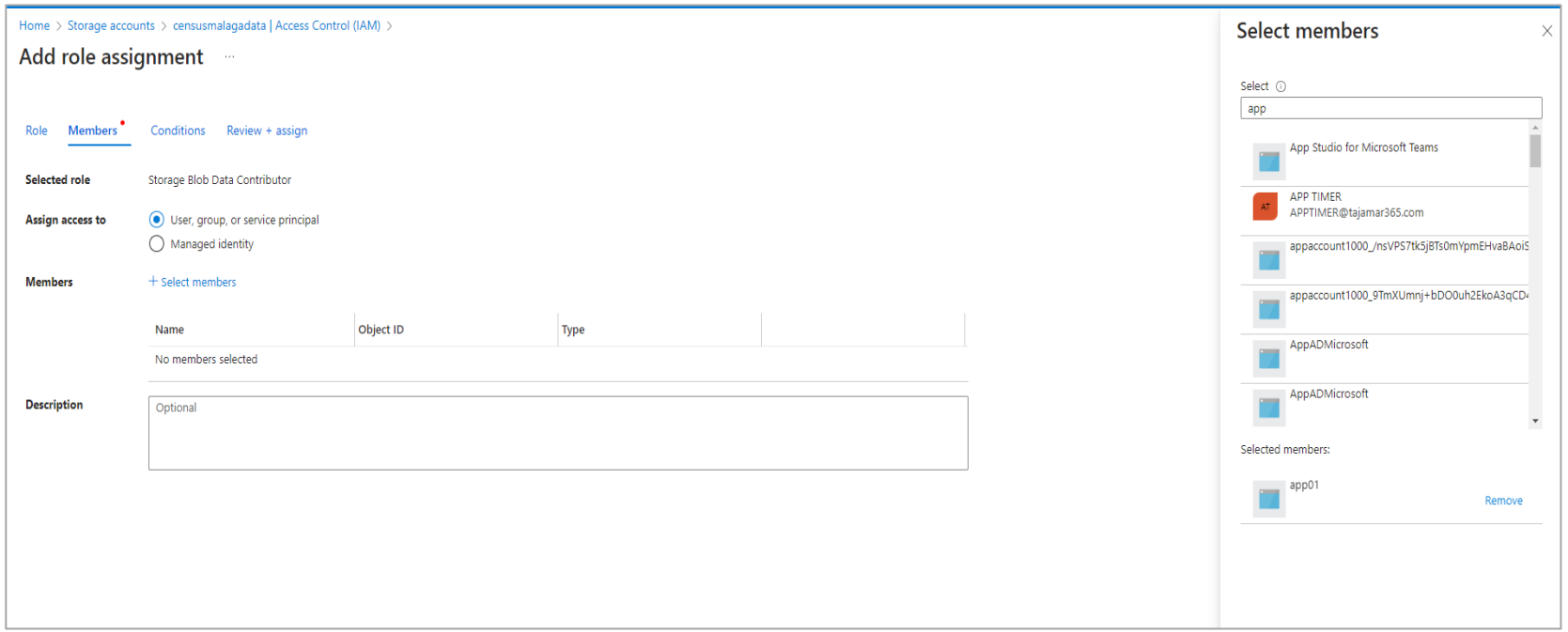
- Go to Storage account and click on “Access Control IAM”

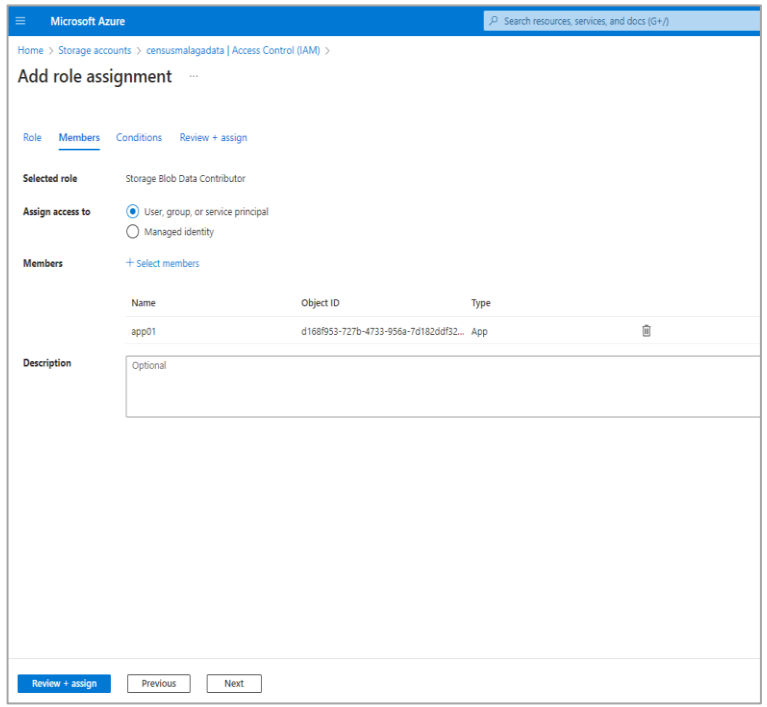


- In IAM click on Add role assignment
- Navigate to Members

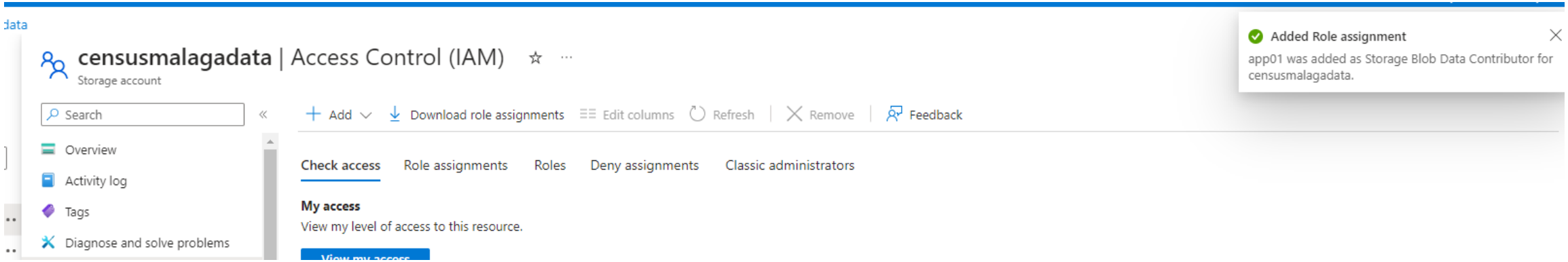


- In Members, “Assign access to”
  - Select user, group or service principal
- On the right “Select members”, search for the Application Registration you made before and select it
- Review + assign

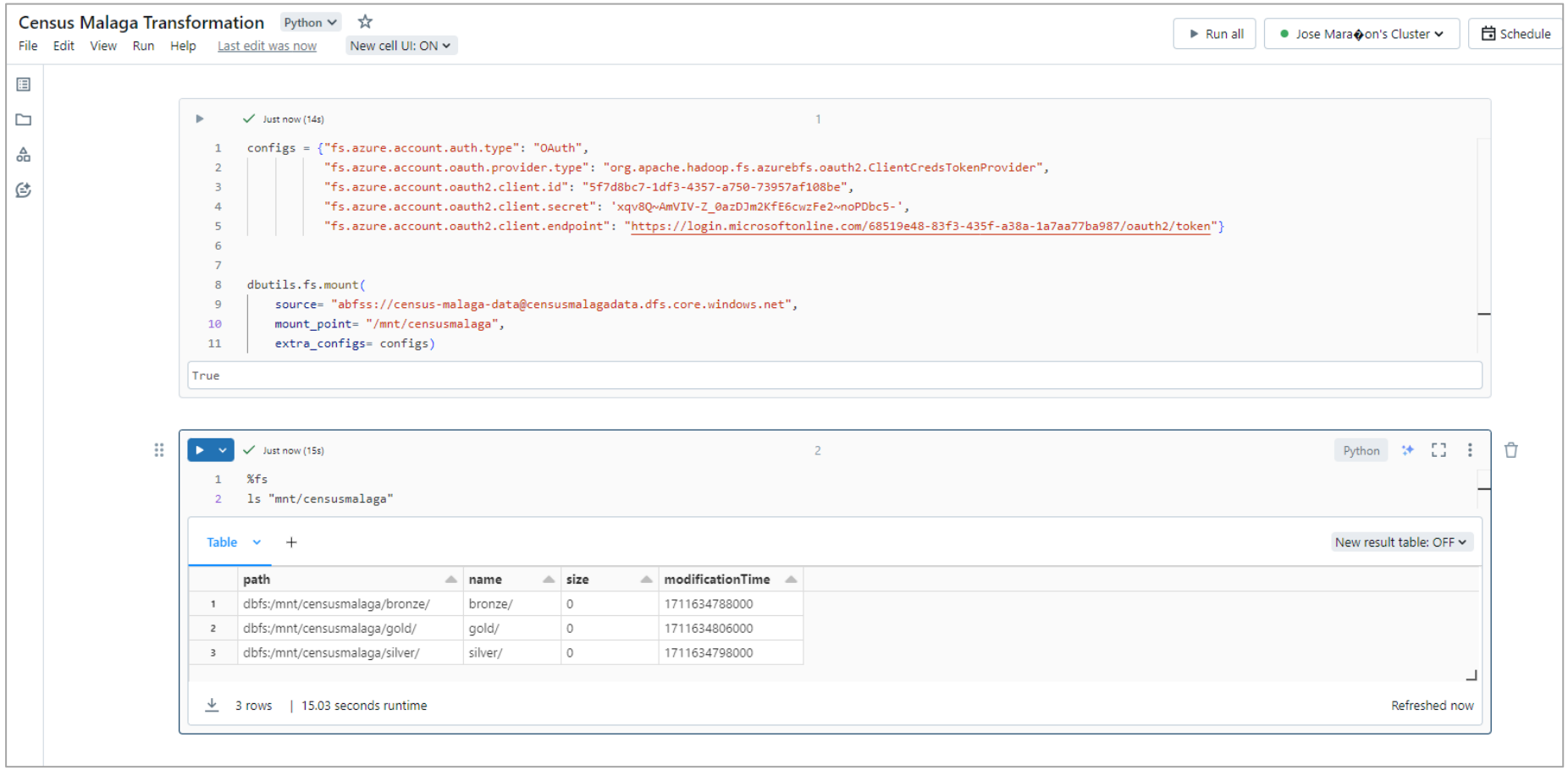


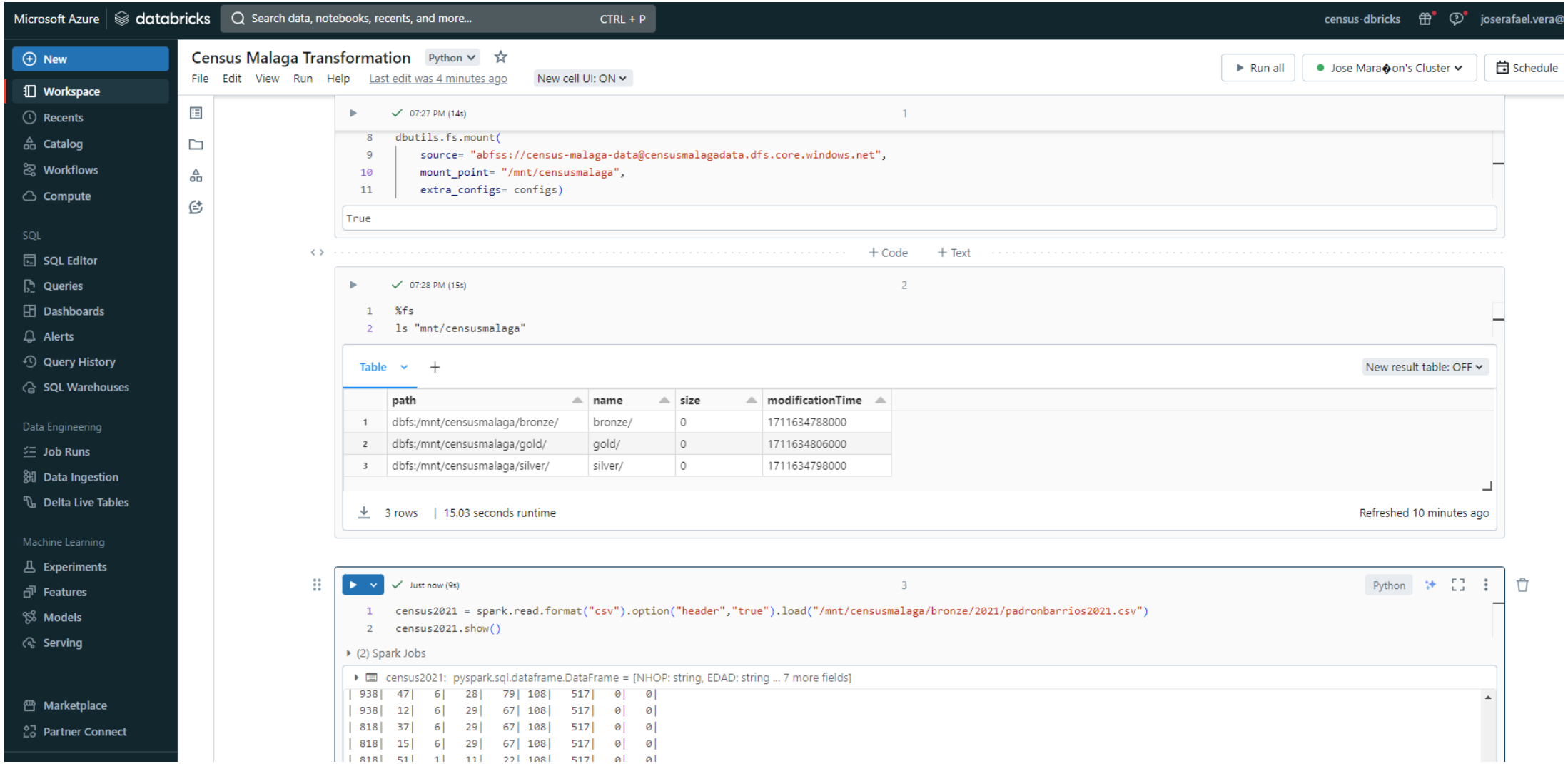
- The new role assignment is successfully added



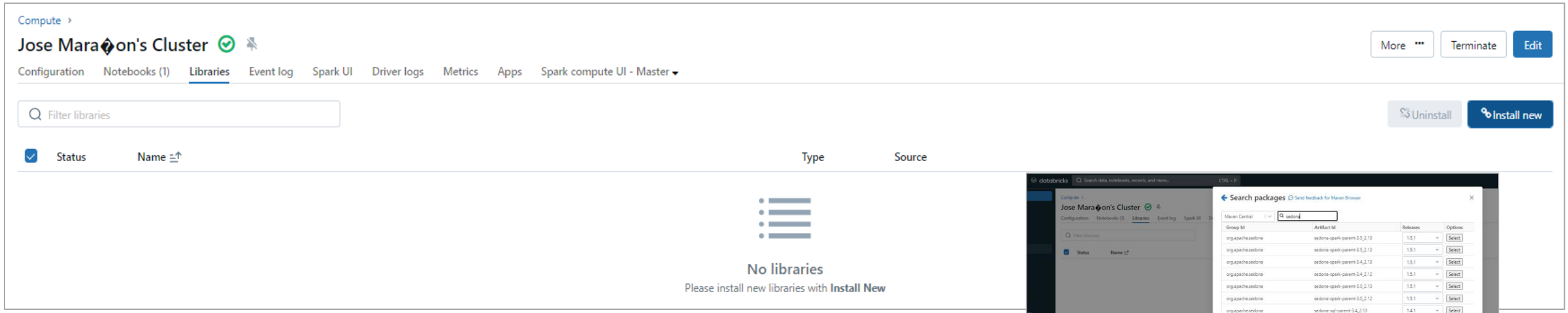
- Run the first cell of the Databricks notebook: you should get a “True” output
- Run the second cell to list all the available directories

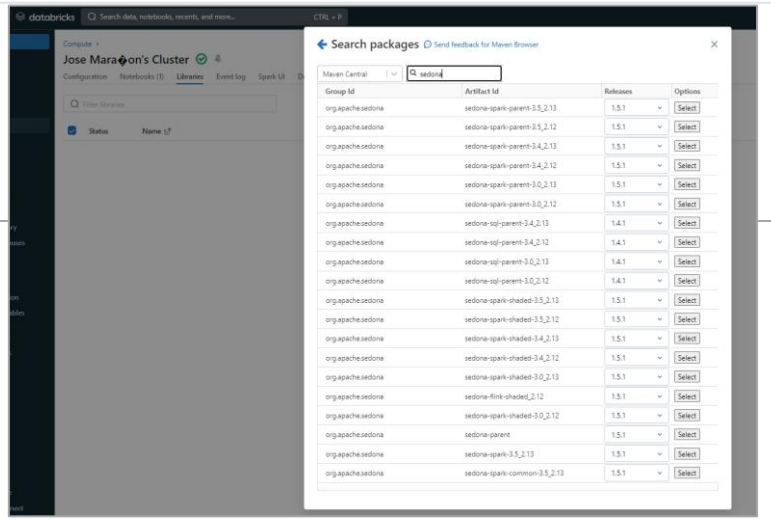


- Start loading the csv files
- Continue with the transformations. Let's coding. **SEE THE NOTEBOOK IN MY REPOSITORY**

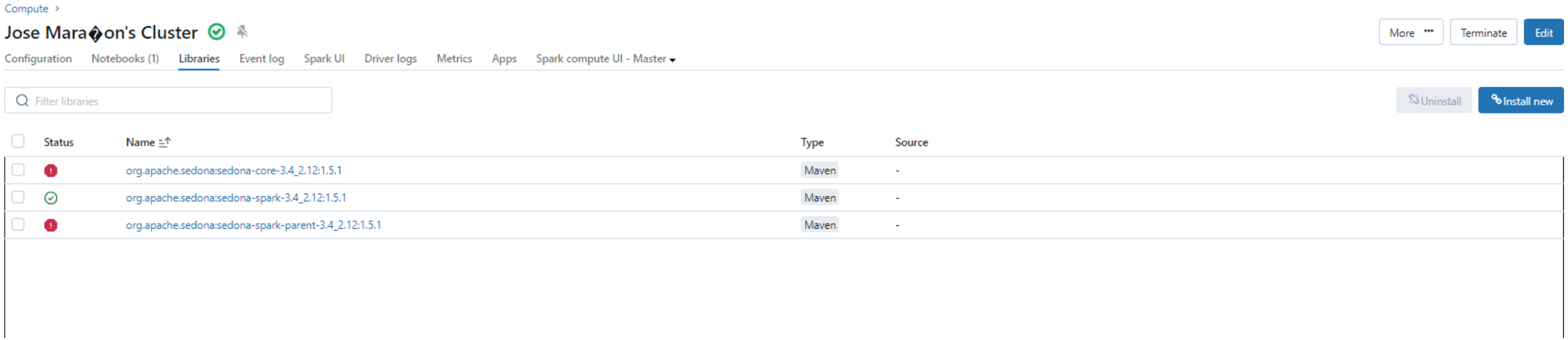


- As I managed cartographic data I tried to use one of the most used spark libraries: Apache SEDONA
- Unfortunately, after importing the library successfully, the engine was not able to find de Sedona module, so I decided not to get deeper in processing geodata. I obtained centroid, area and polygon 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

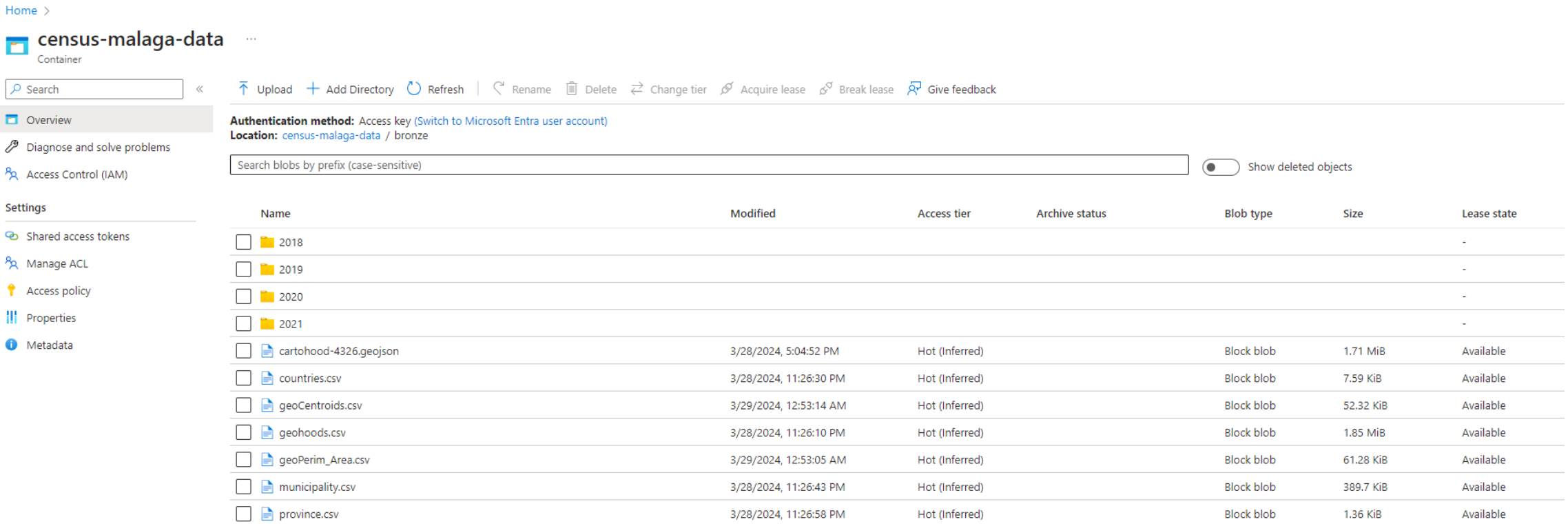


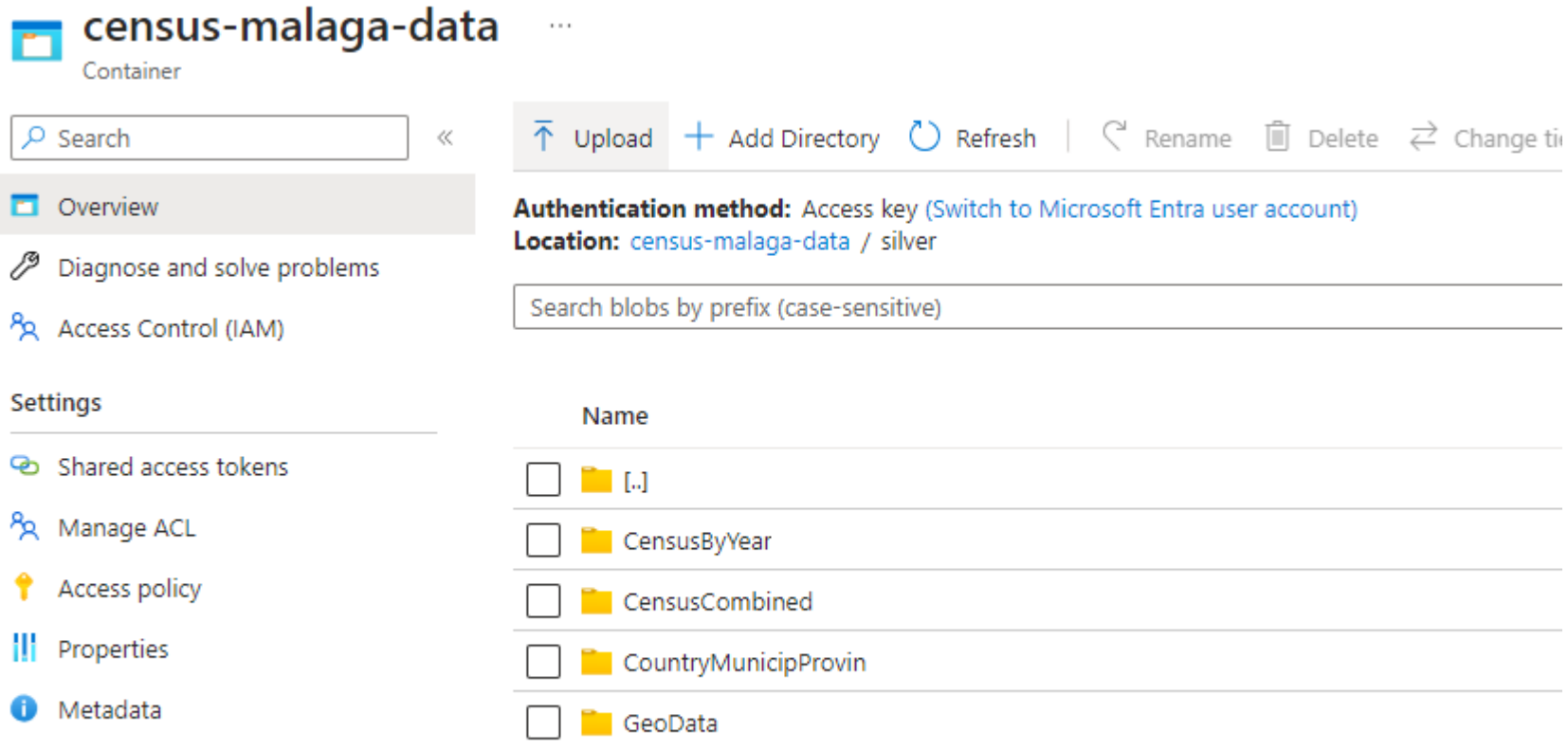
- Once installed you will see something like this
- Note there are two failed library installations



- Databricks available Bronze file



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

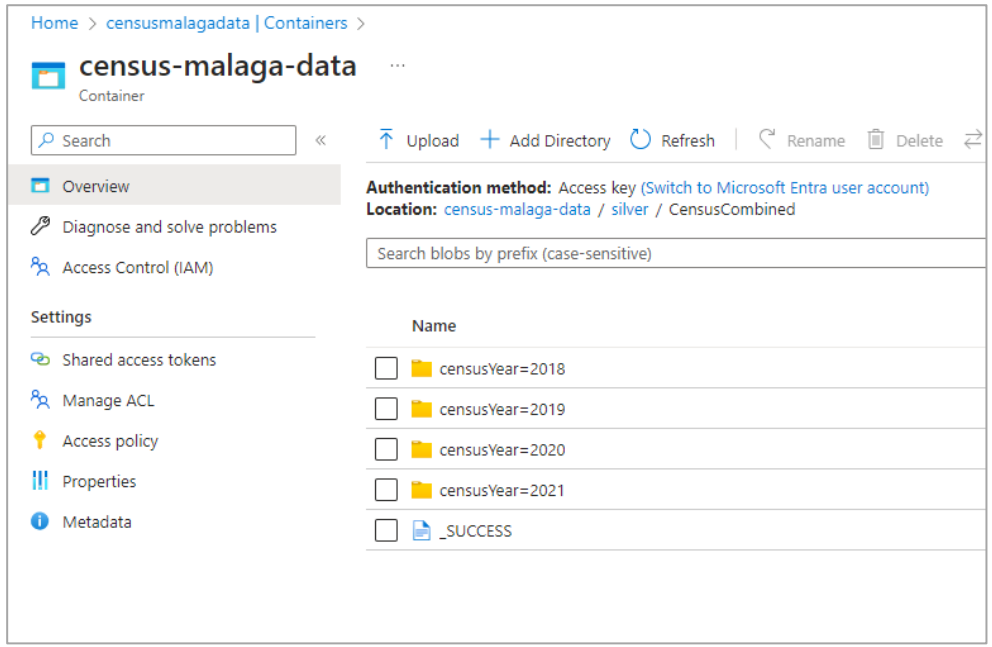


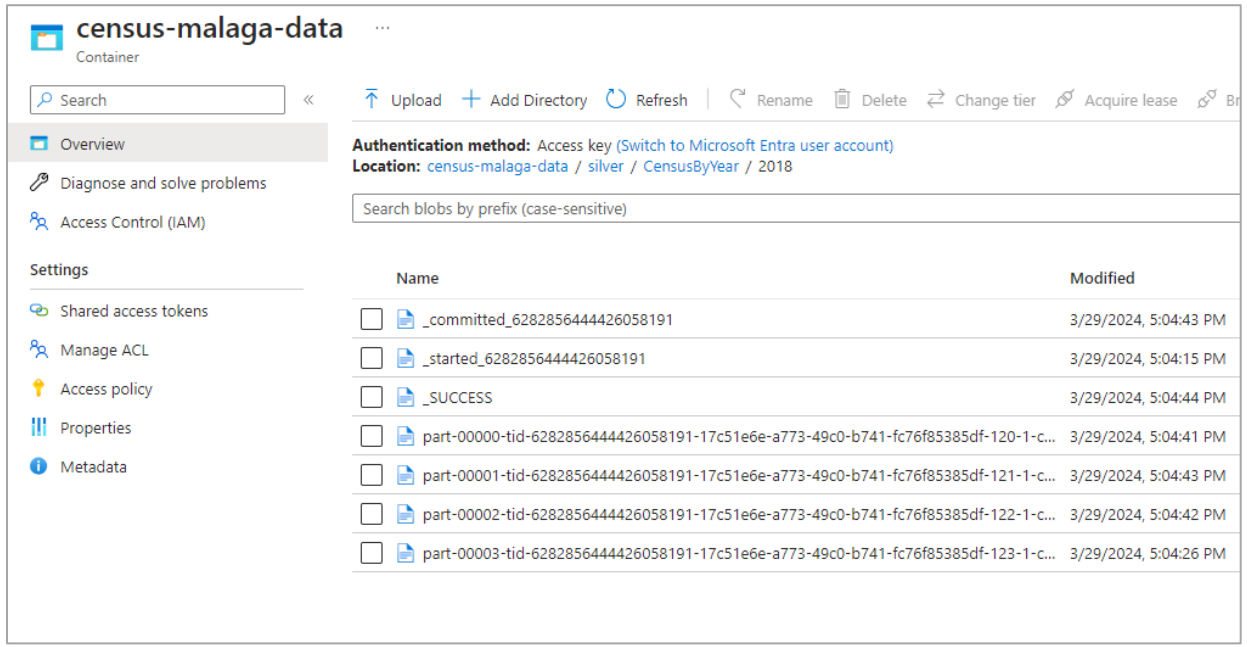
- This Databricks notebook cell save all the transformations into the Silver layer
- Check the complete notebook in my Github Repository

- **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

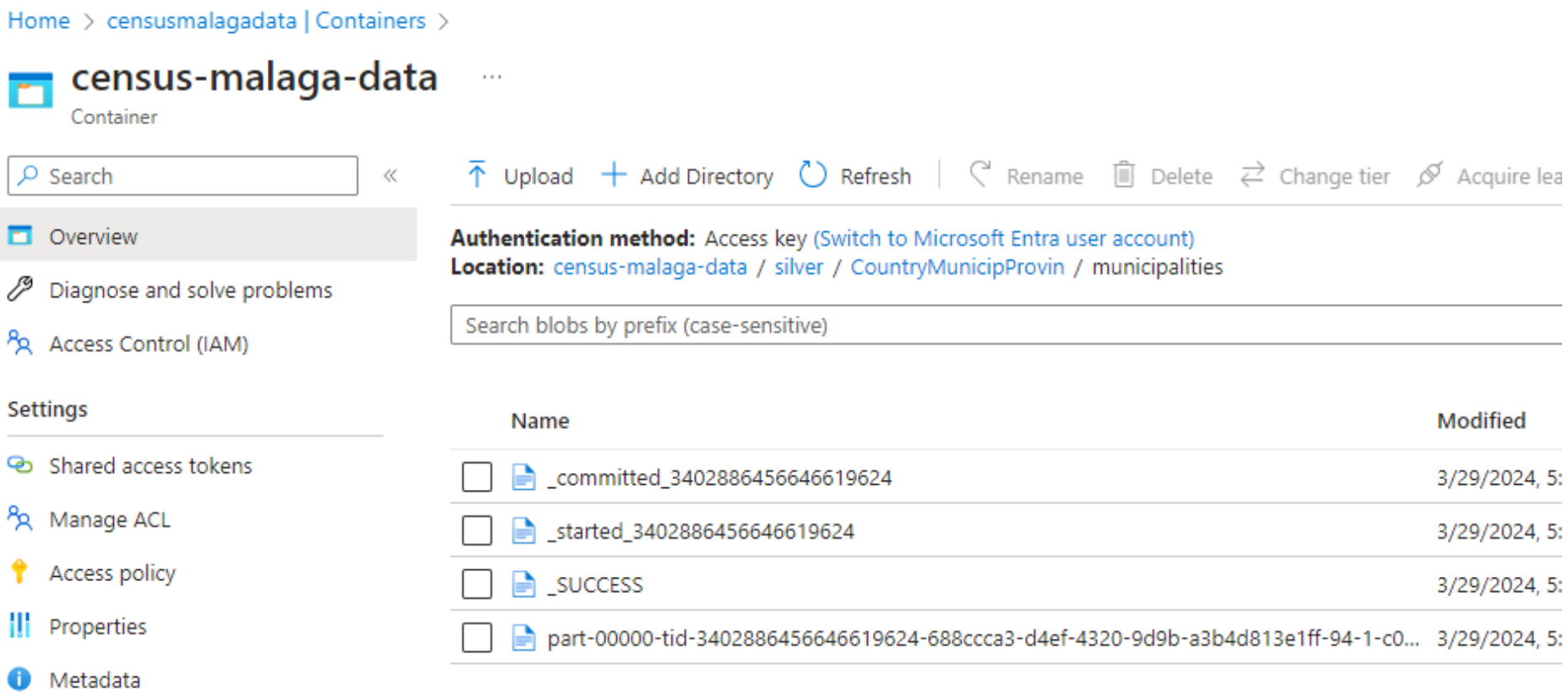
## 4. SAVING INTO SILVER LAYER

- 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

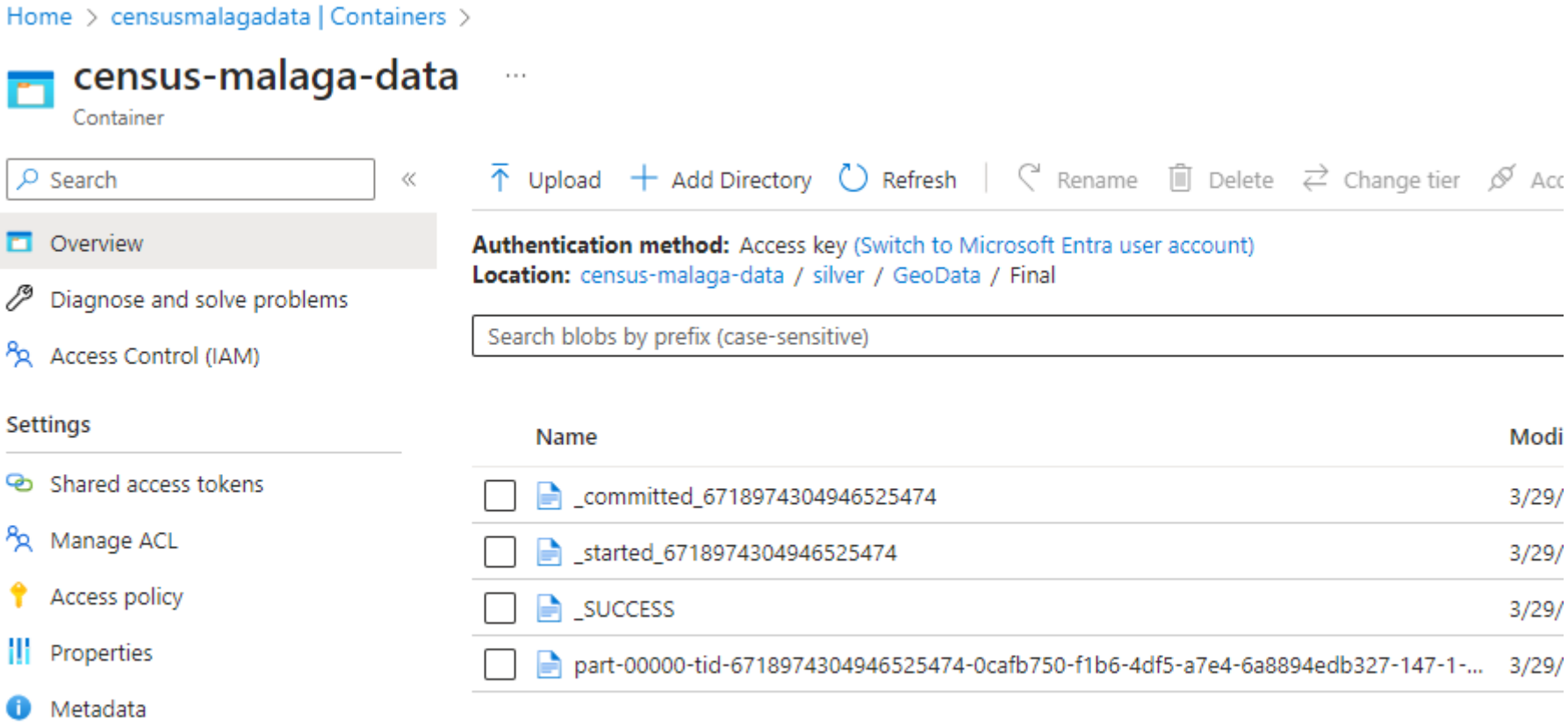


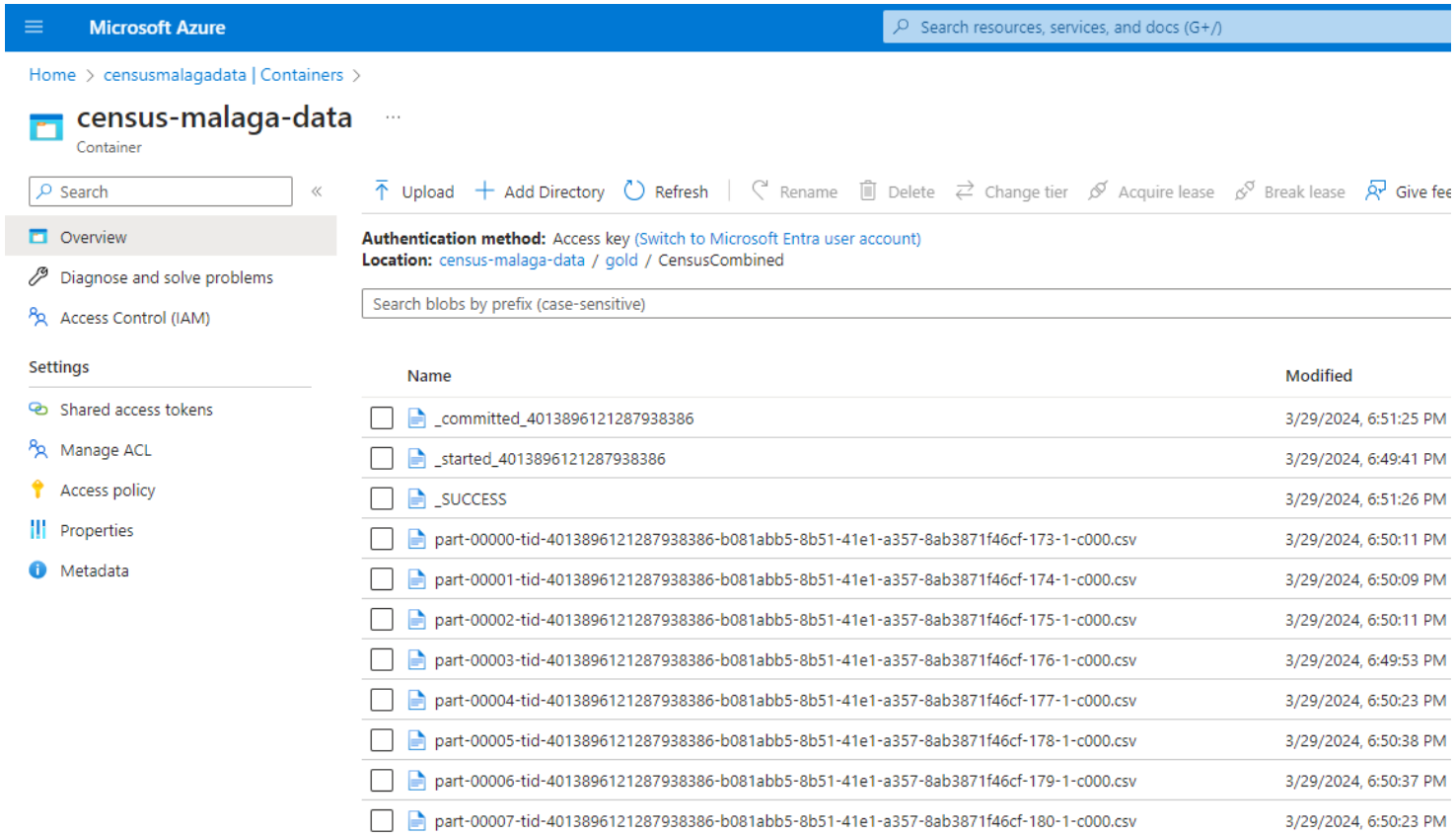
- The other transformed data was persisted correctly inside Silver layer



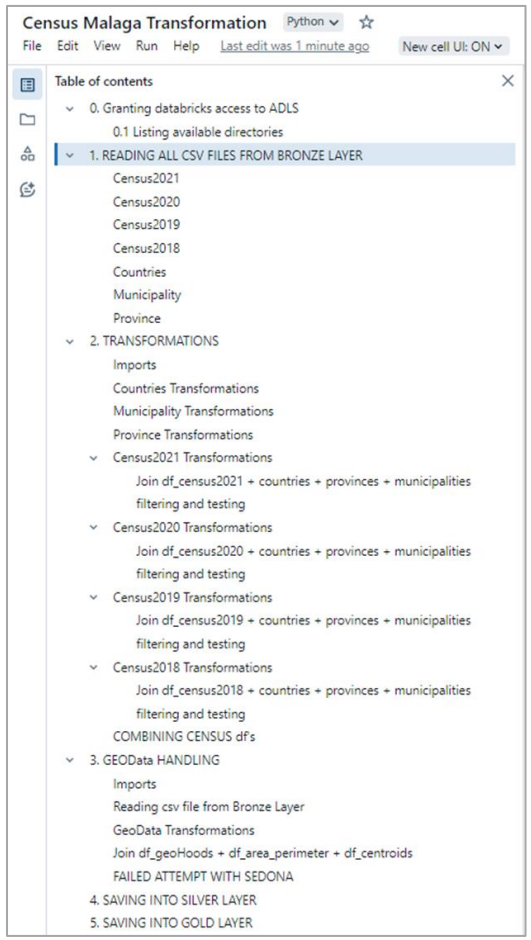
- In my case, I decided to persist the GOLD data directly from Databricks as I didn’t make relevant changes from Silver data.
- For the next steps with Azure Synapse, I just used it to create tables and make them available for PowerBI.



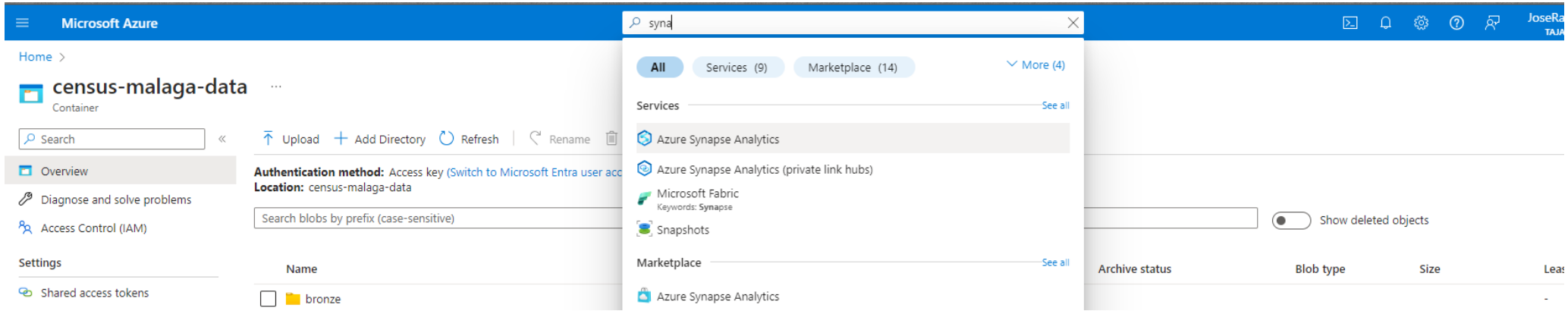
- The main csv file partitions with all the demographic information saved into the GOLD layer



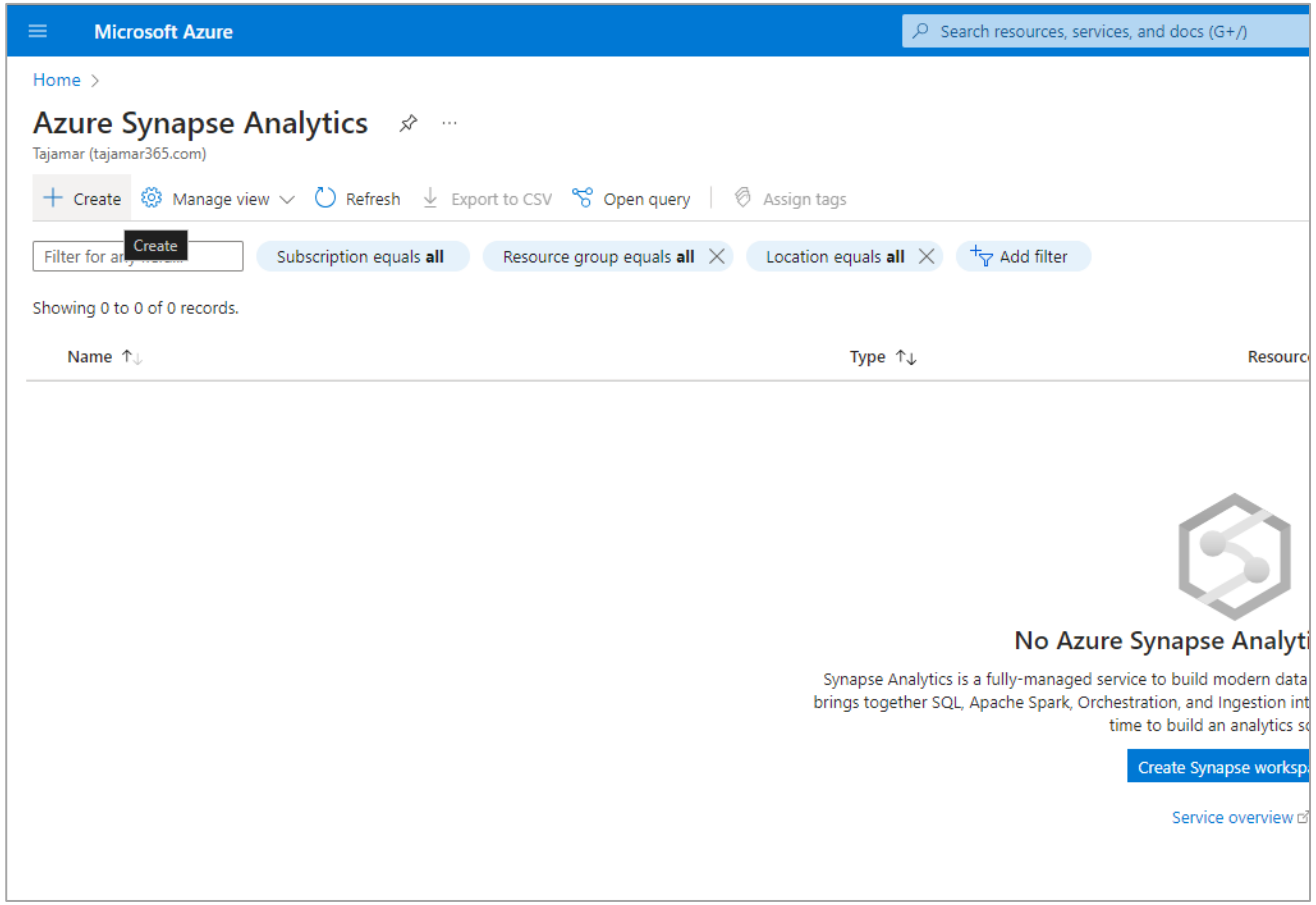
- Notebook table of contents



- Search for Azure Synapse Analytics



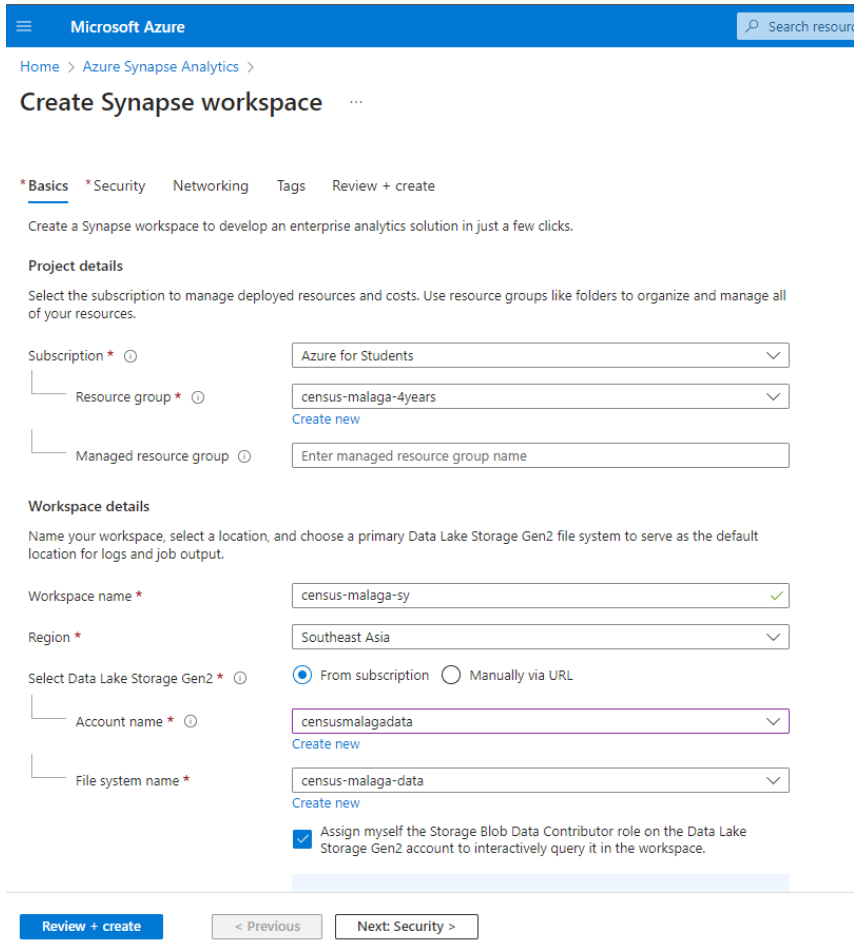
- Create an Azure Synapse workspace



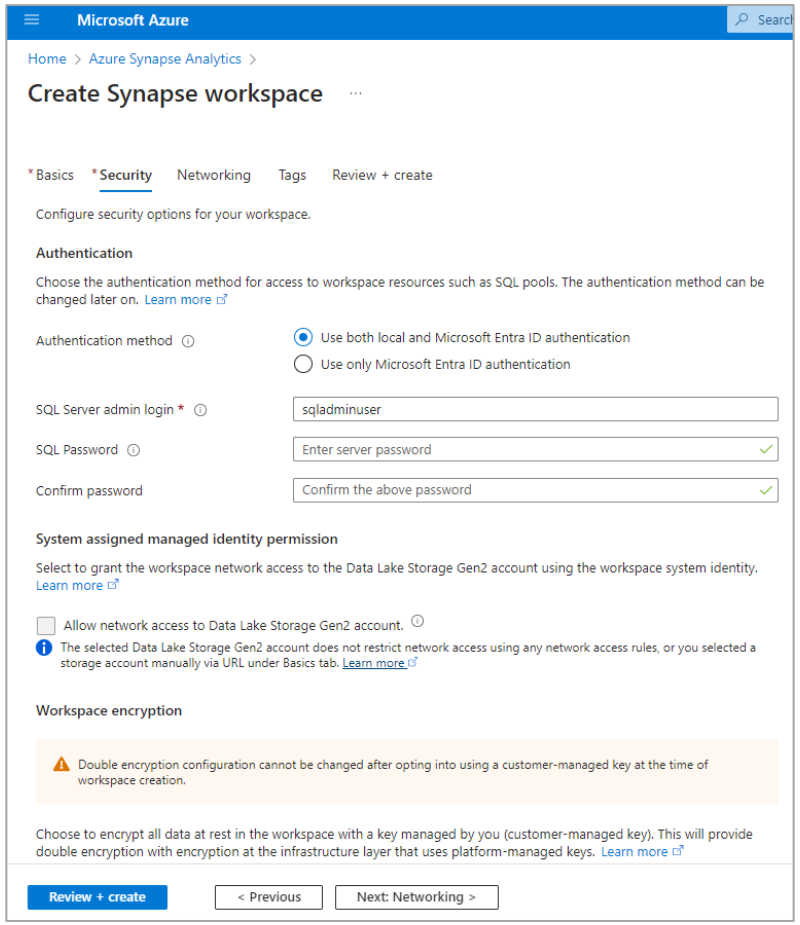
- Select your free subscription
- Select the existing resource group

- 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

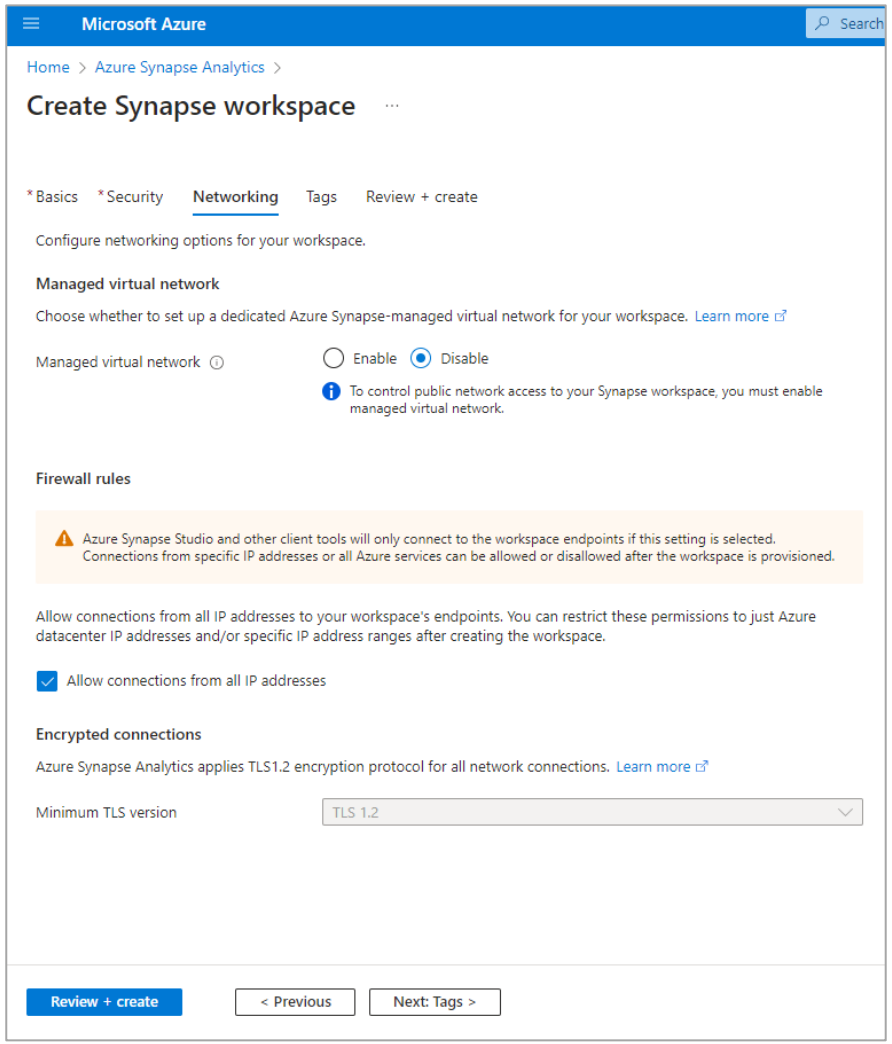
- NEXT



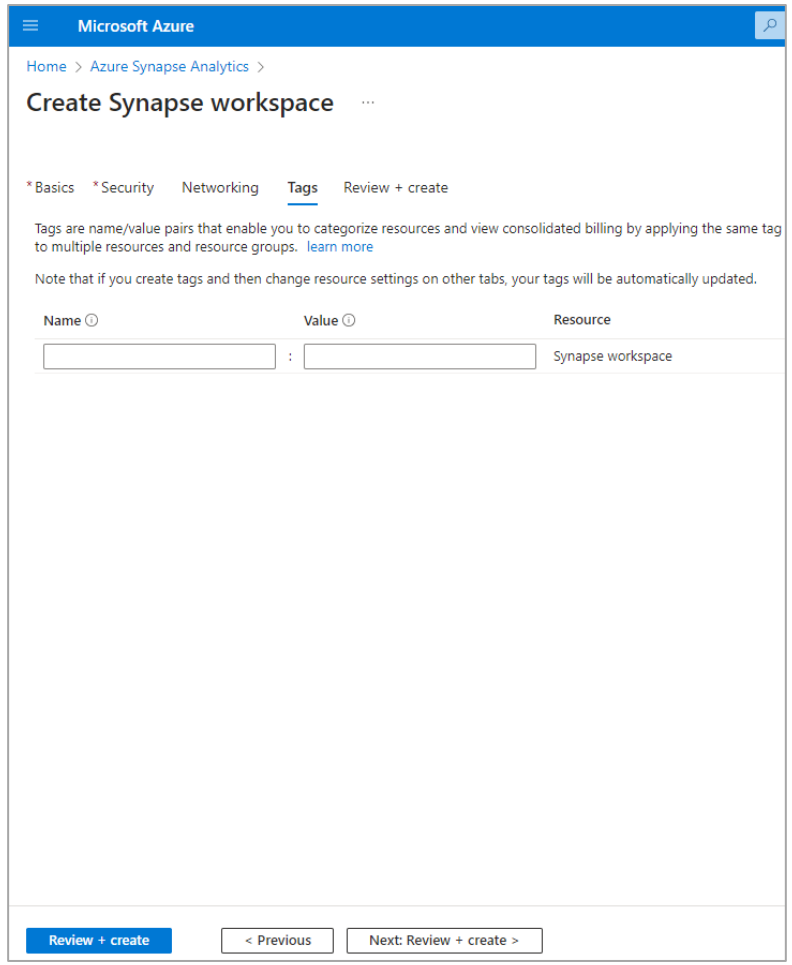
- Next


Microsoft Azure
Search
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 *
sqladminuser
SQL Password
Enter server password
Confirm password
Confirm the above 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 >

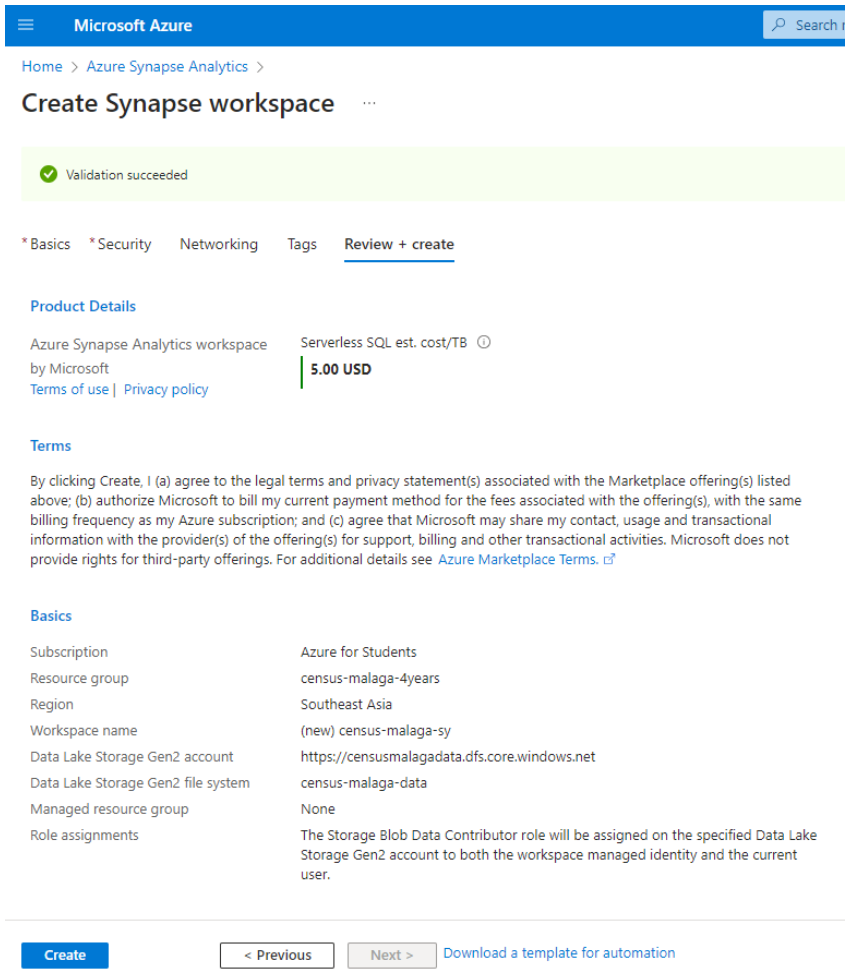
- Next


Microsoft Azure
Search
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
Managed virtual network
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
TLS 1.2
Review + create
< Previous
Next: Tags >

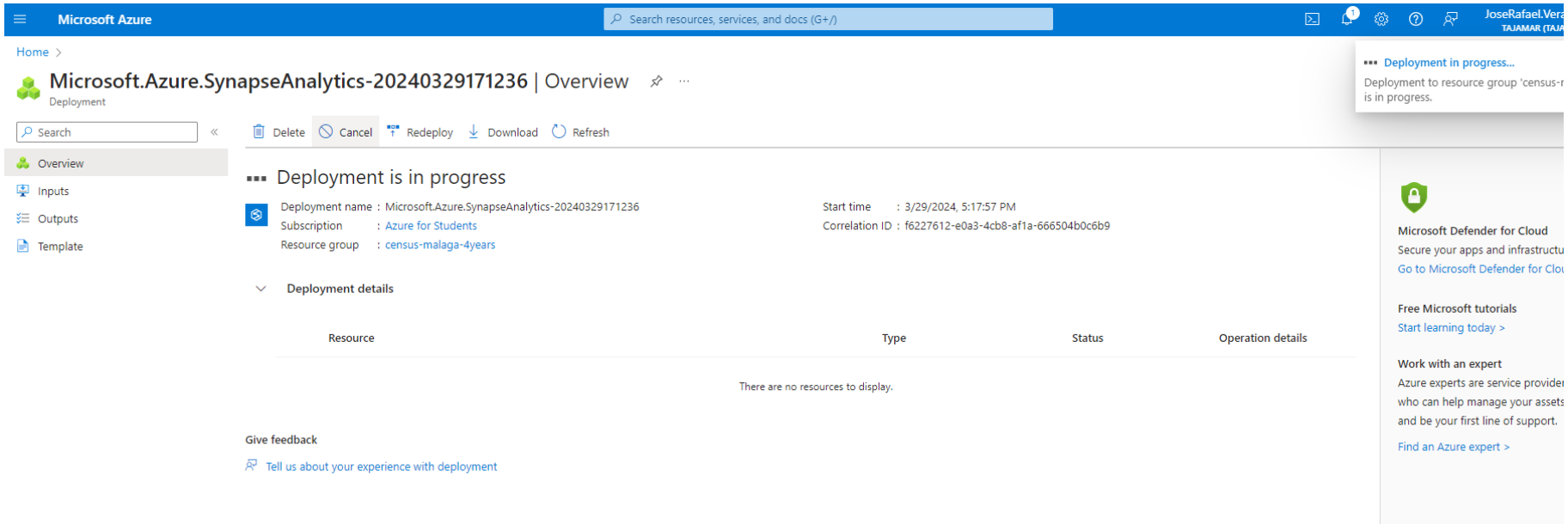
- Next


Microsoft Azure
Home > Azure Synapse Analytics >
Create Synapse workspace
*Basics *Security Networking Tags Review + create
Tags are name/value pairs that enable you to categorize resources and view consolidated billing by applying the same tag to multiple resources and resource groups. learn more
Note that if you create tags and then change resource settings on other tabs, your tags will be automatically updated.
Name
Value
Resource
:
Synapse workspace
Review + create
< Previous
Next: Review + create >

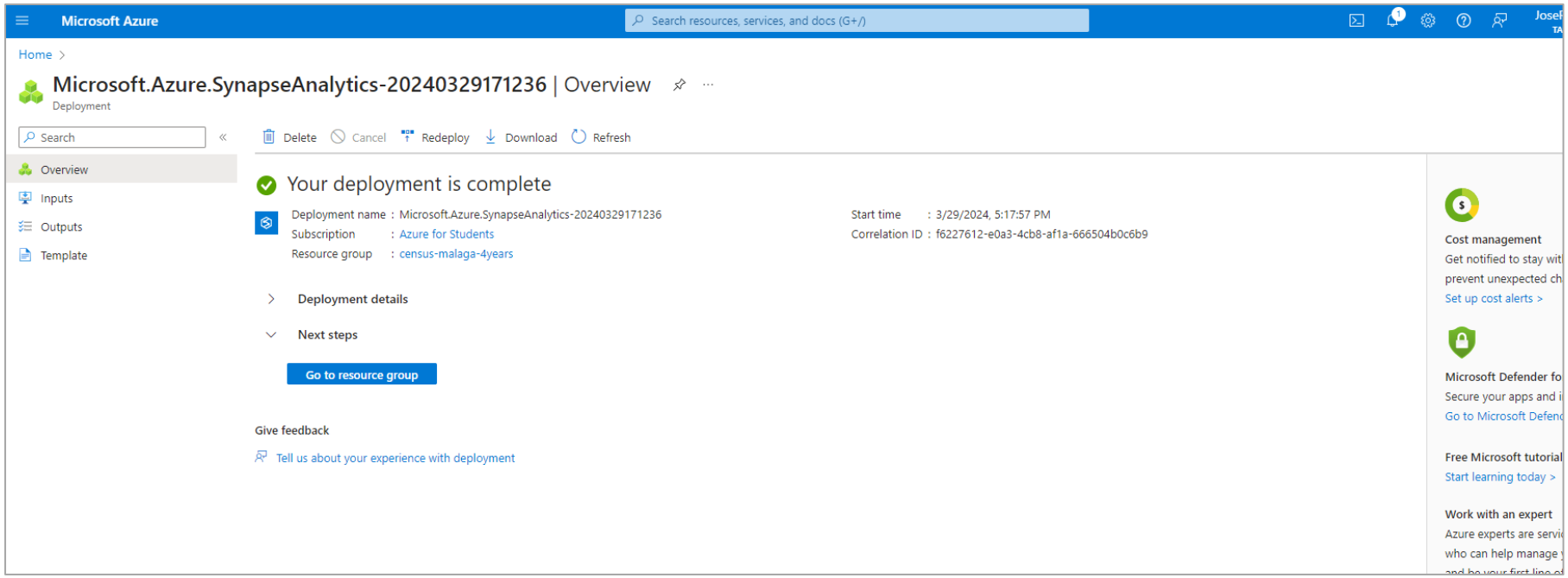
- Create


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
Terms of use | Privacy policy
Serverless SQL est. cost/TB
5.00 USD
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

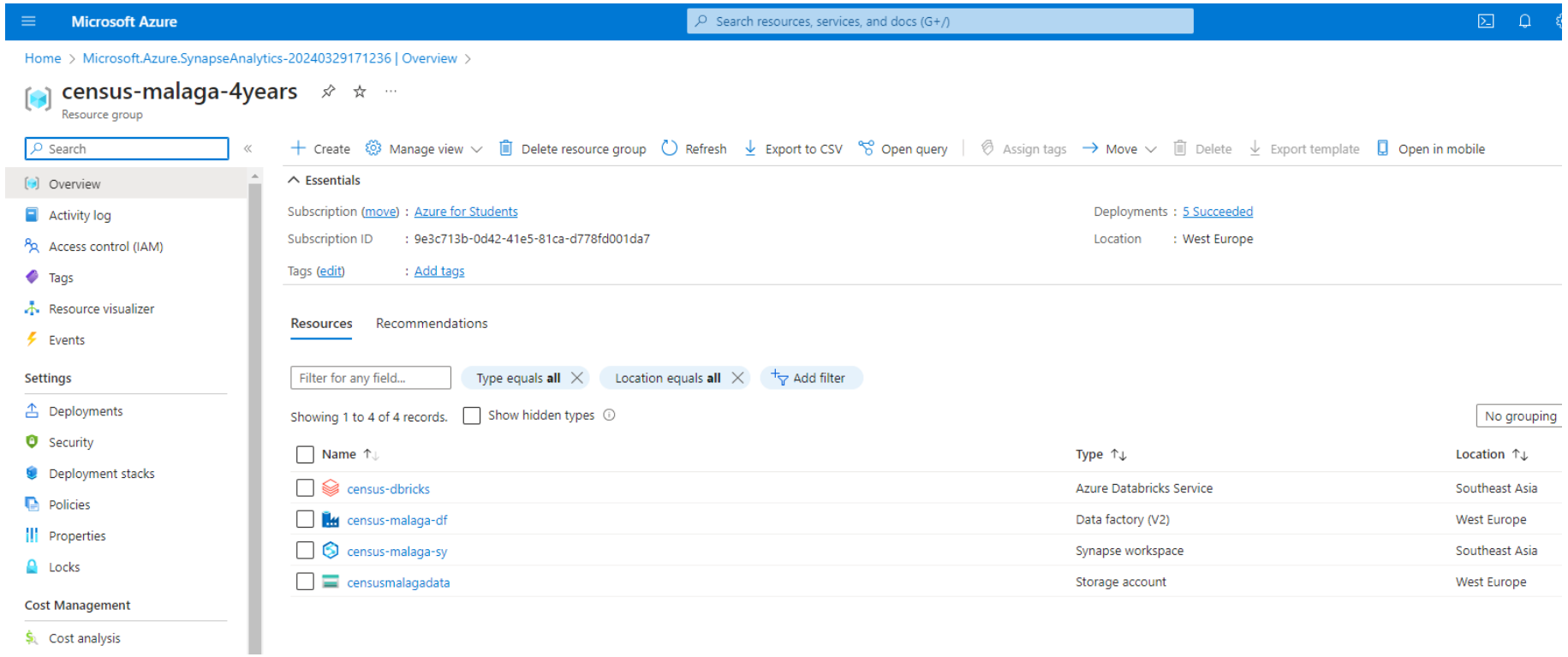
- Wait until Synapse workspace deployment



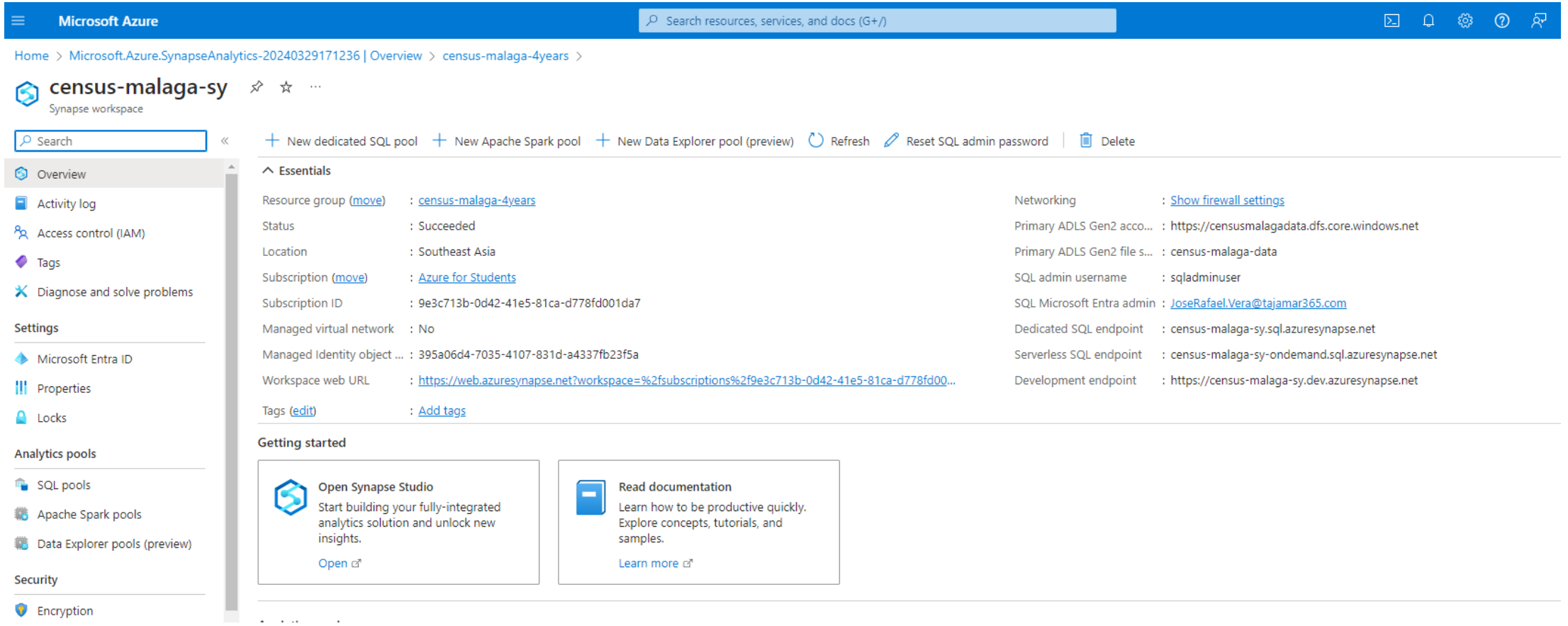
- Synapse workspace deployed
- Go to resource group



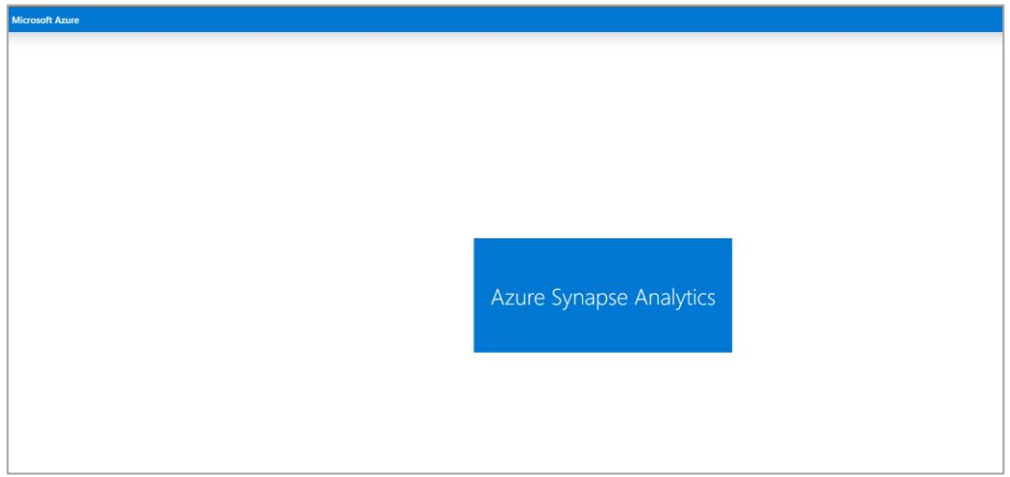
- Click on the synapse workspace

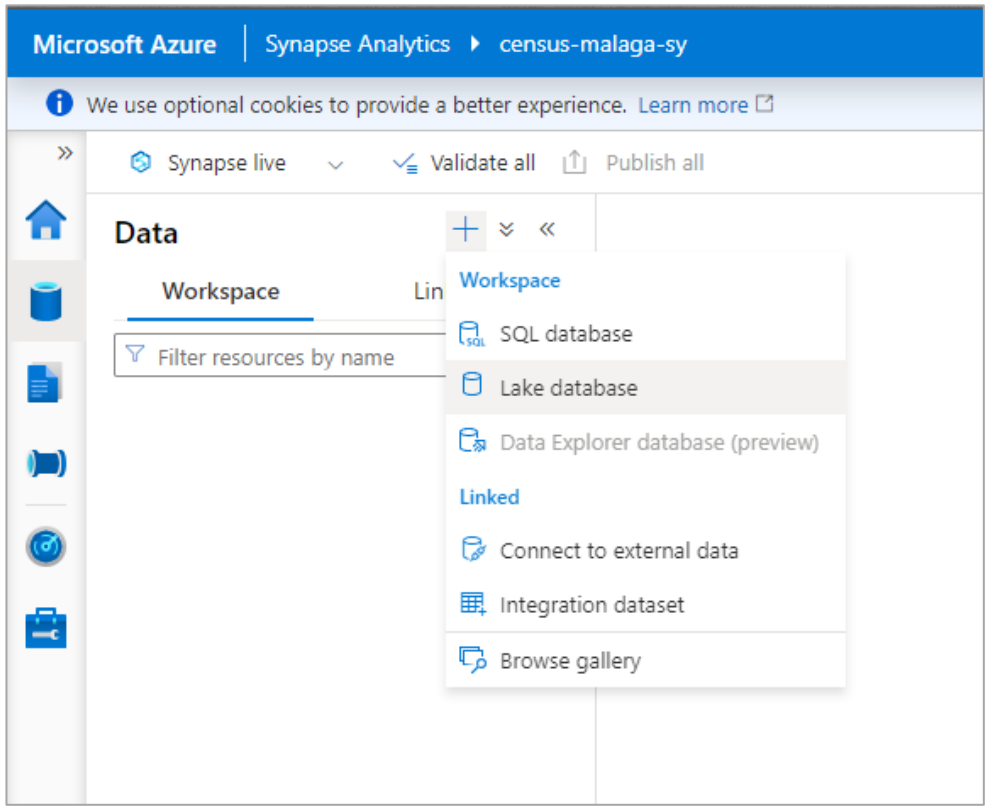


- Azure synapse workspace console
- Open Synapse Studio

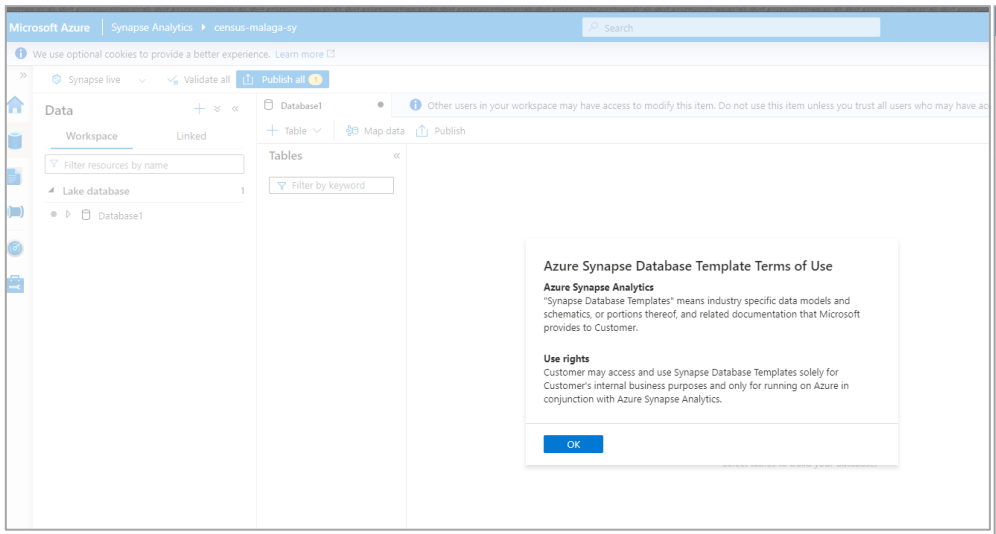


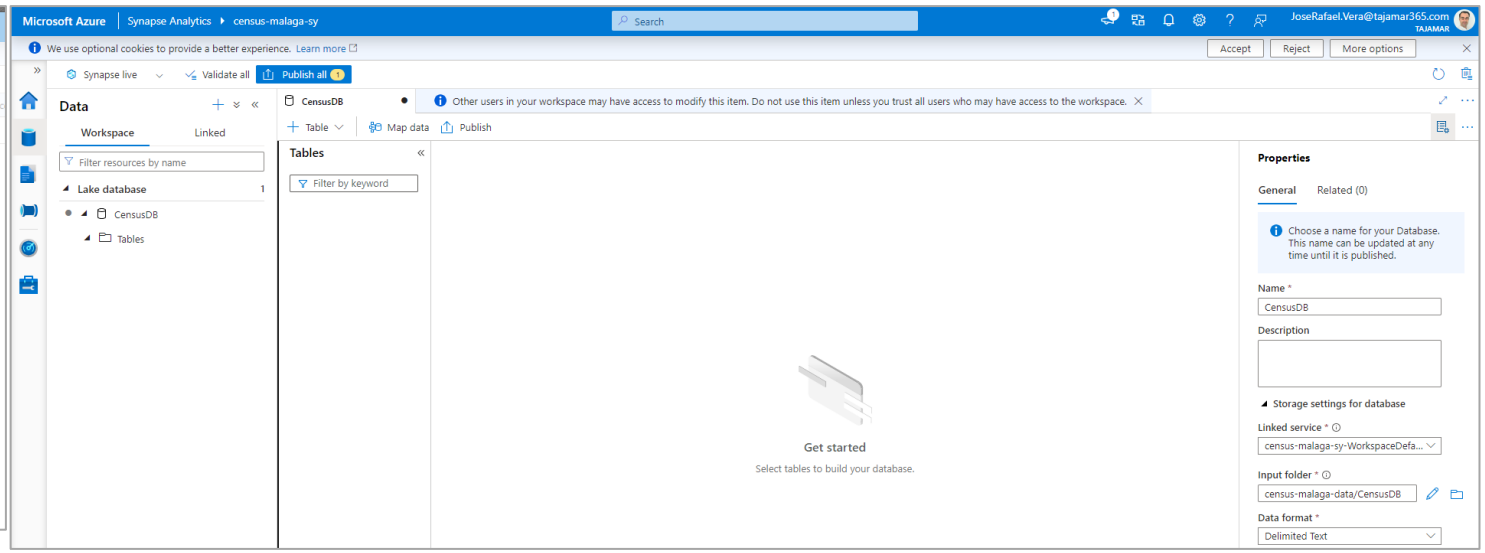
- Azure synapse studio
- Click on Data
- Click on + and “Lake database”



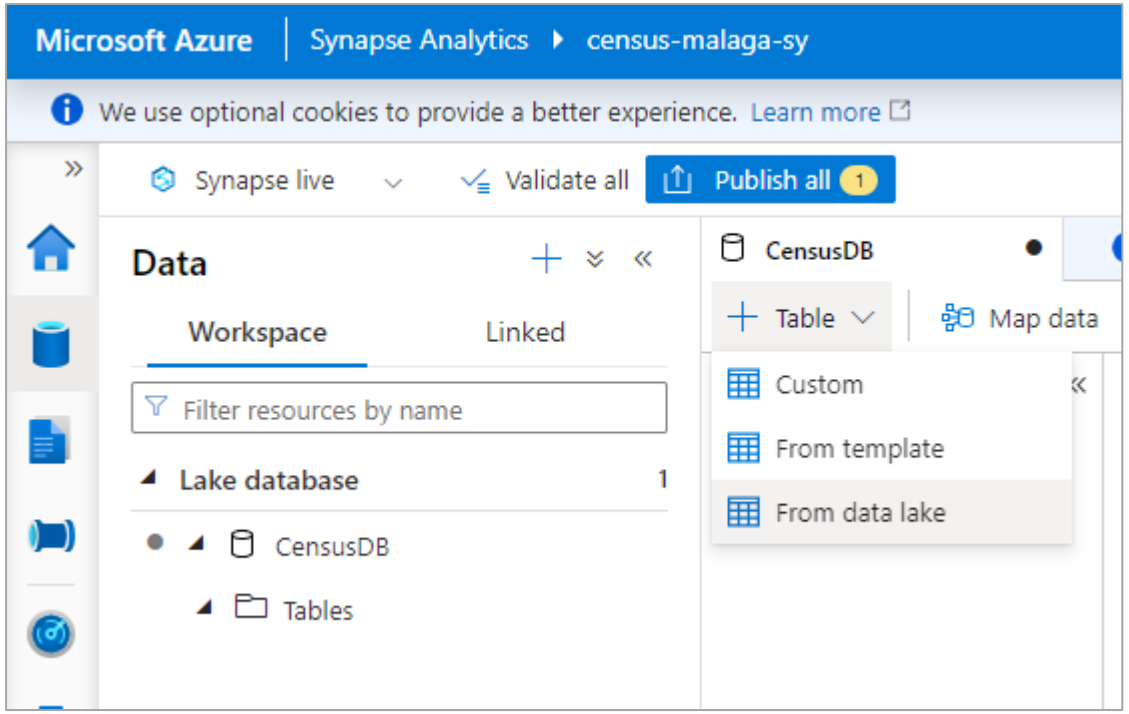



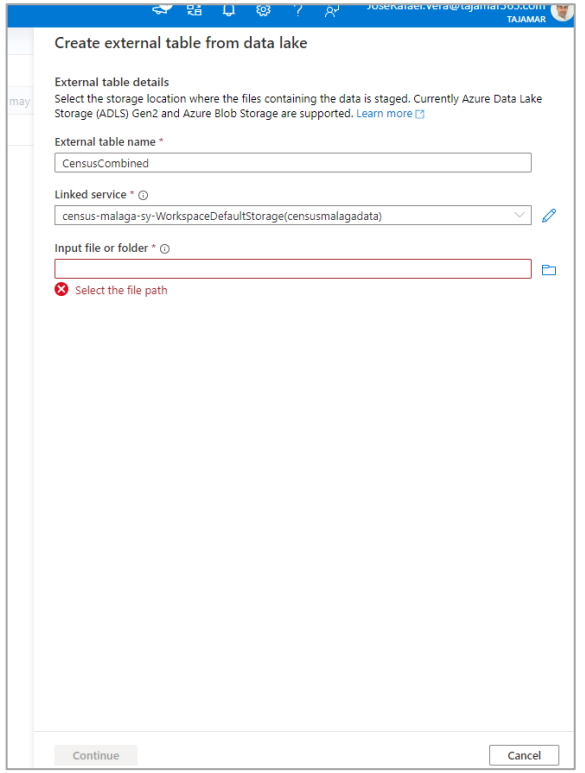
- Click ok
- You will see the console of the recent created Lake database
- On the right, give the database a name
- Tables directory is yet empty



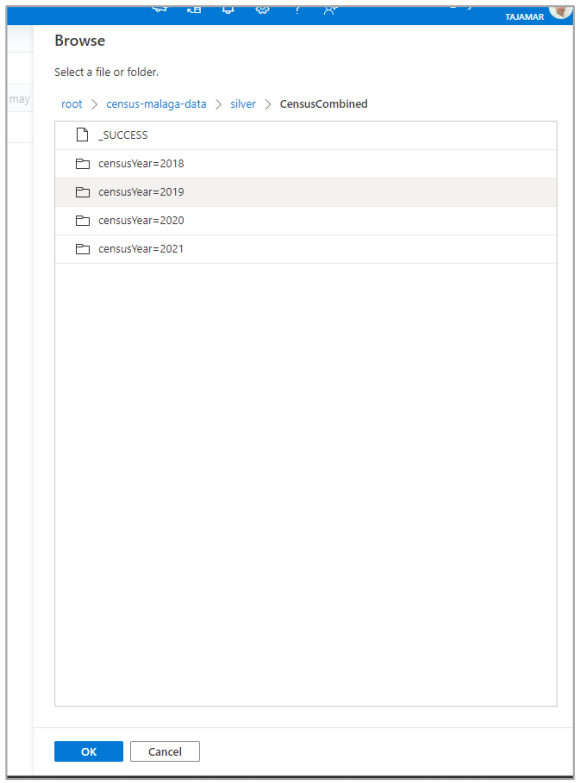



- Let’s create the tables from Silver data (in our ADLS container)
- Click on the + Table and select “From data lake”
- On the right:
  - Give the table a name
  - Linked service: select which appears as you click in the dropdown
  - Input file or folder: browse the directories to find the desired file

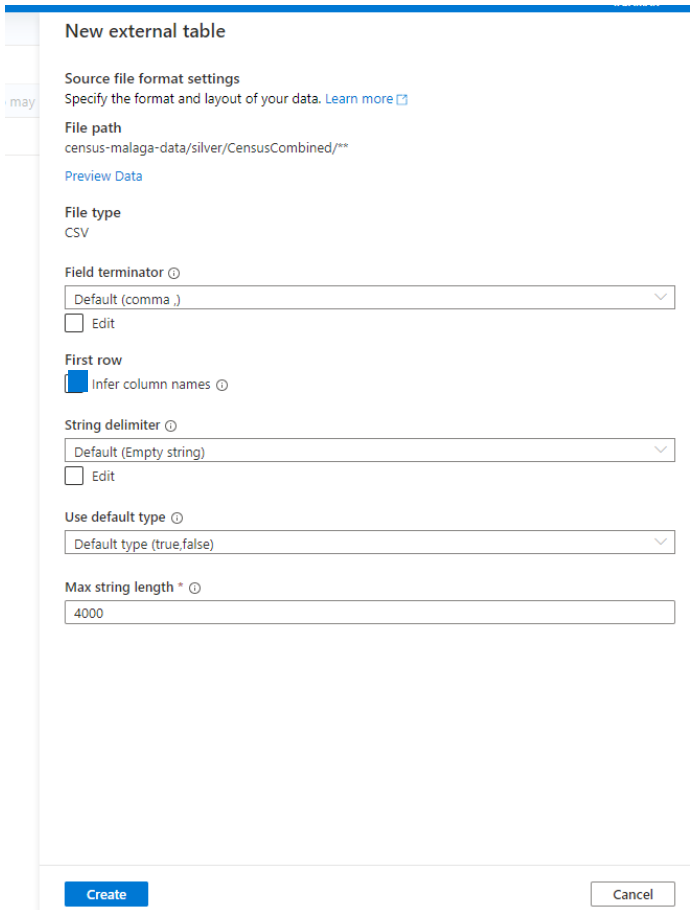


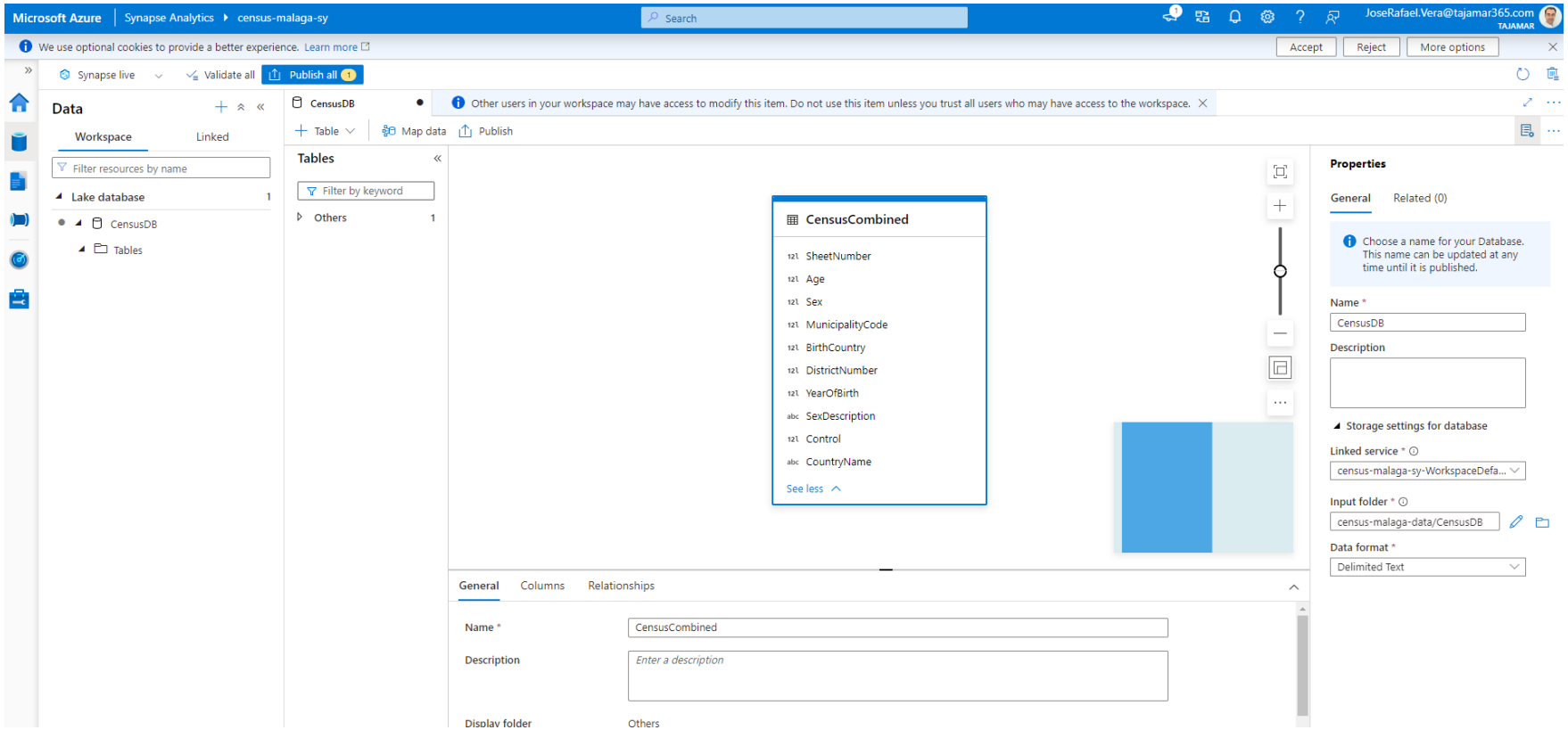
- While browsing you are able to select a whole directory or a specific file
  - If the file is partitioned, then you should select the directory where all the partitions are.
  - Synapse will merge all the partitions together to build up a table
- The image below shows how I was selecting the CensusCombined folder, pretending to get al 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



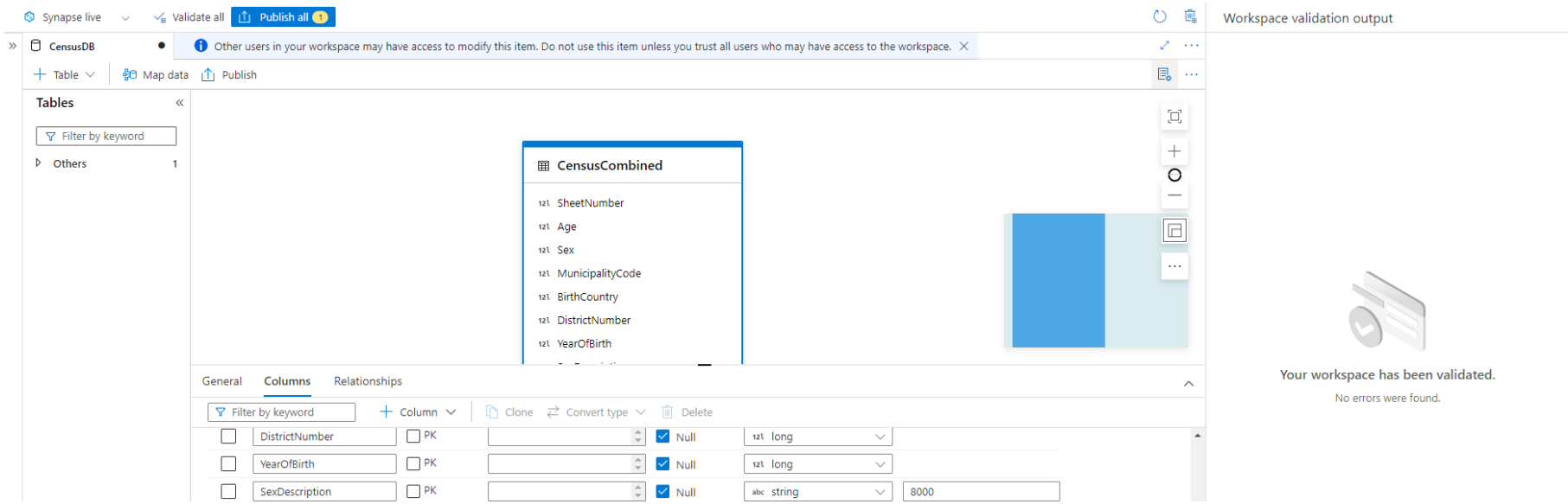
- As you clicked ok, you see the file path as ….silver/CensusCombined/**
  - That means that the input path includes all the files under this directory
- First row: infer column names must be checked
- Create



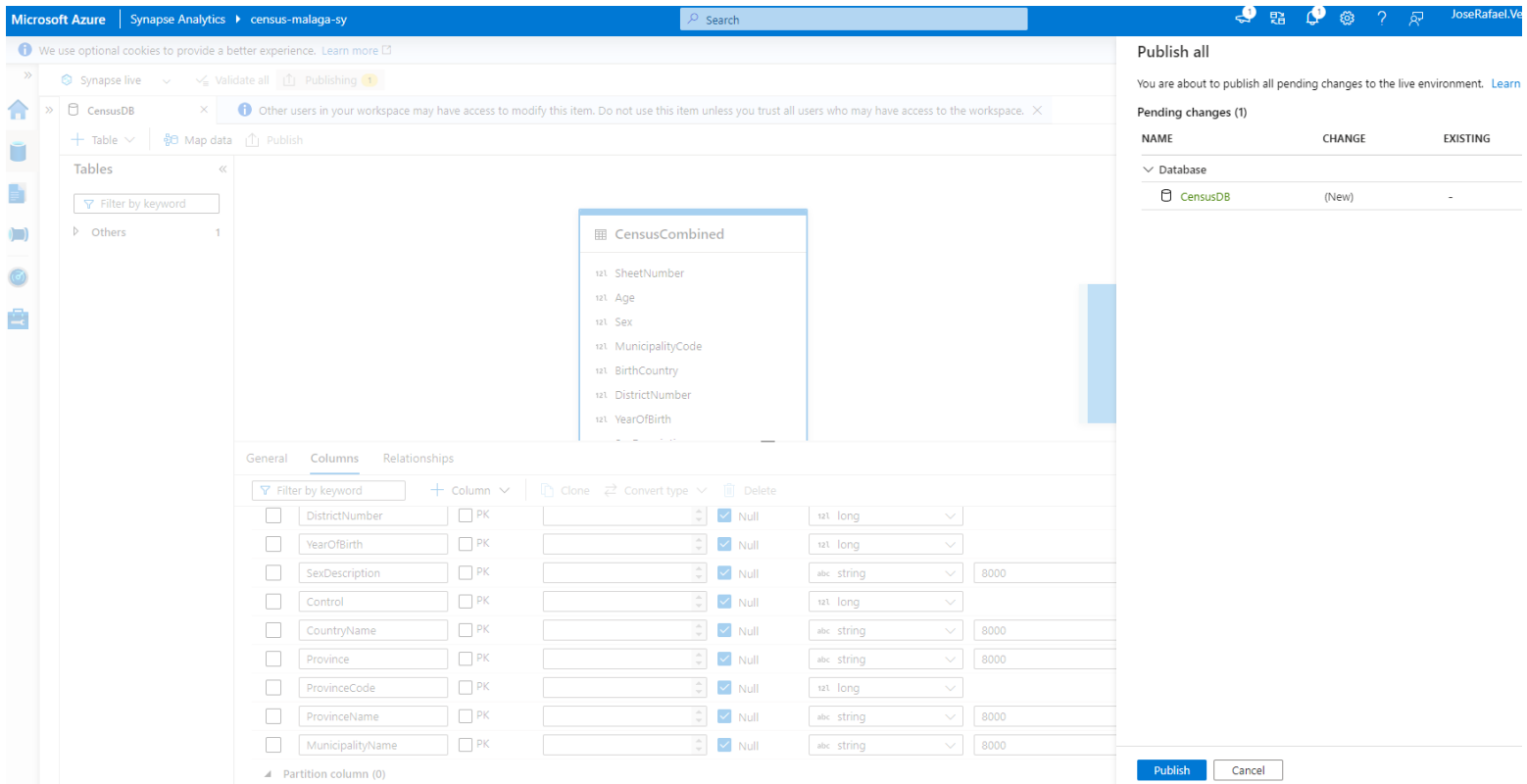
- Table added to the synapse workspace



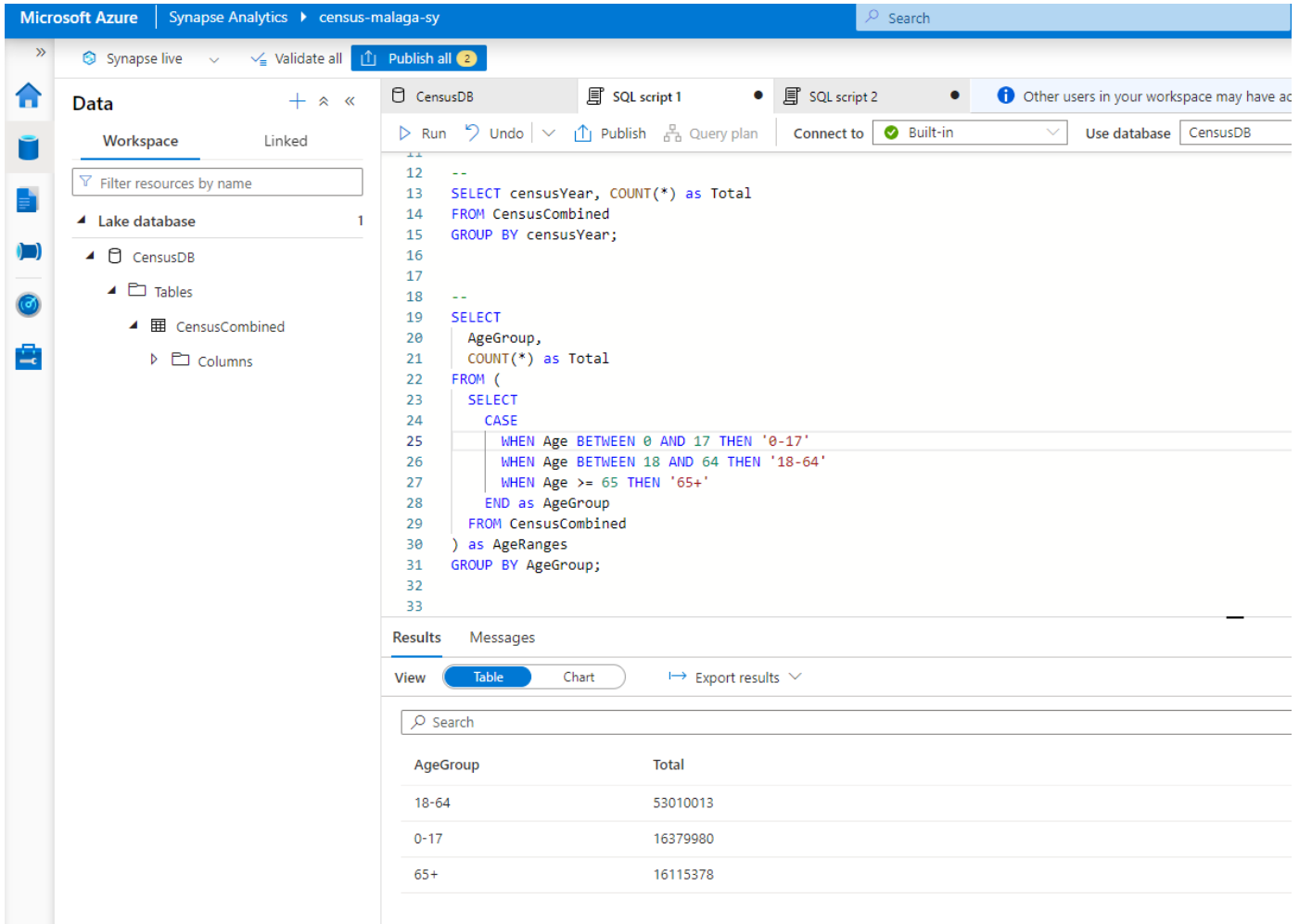
- Explore the table columns, data type and names. You can even start creating relationships between tables. I made this on POWERBI
- Click on validate all
- Click on publish all to save all the changes in the workspace



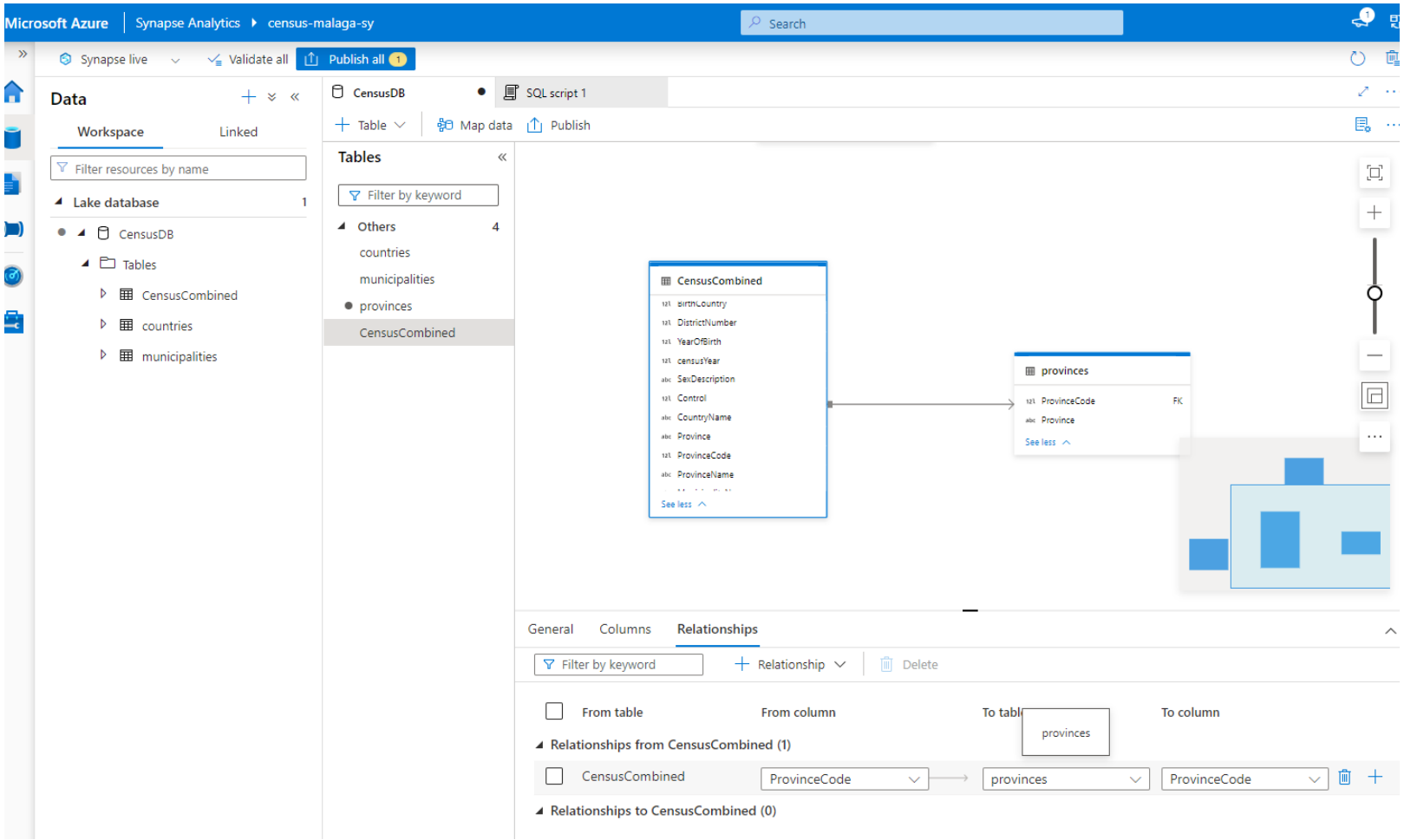
- What you see when click on publish all
- Click on publish



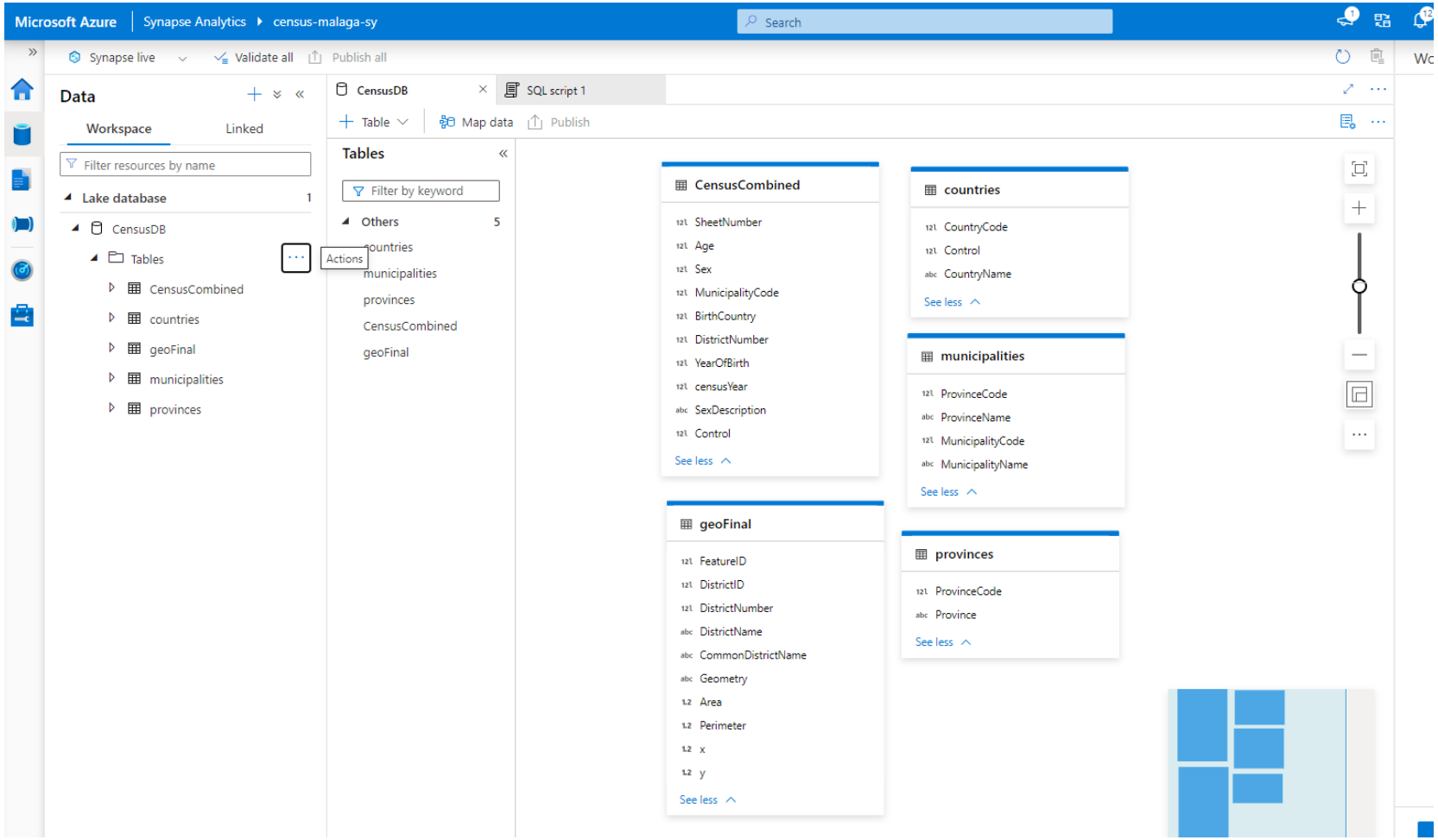
- Run queries against the table. **SEE MY REPOSITORY FOR QUERY EXAMPLES**



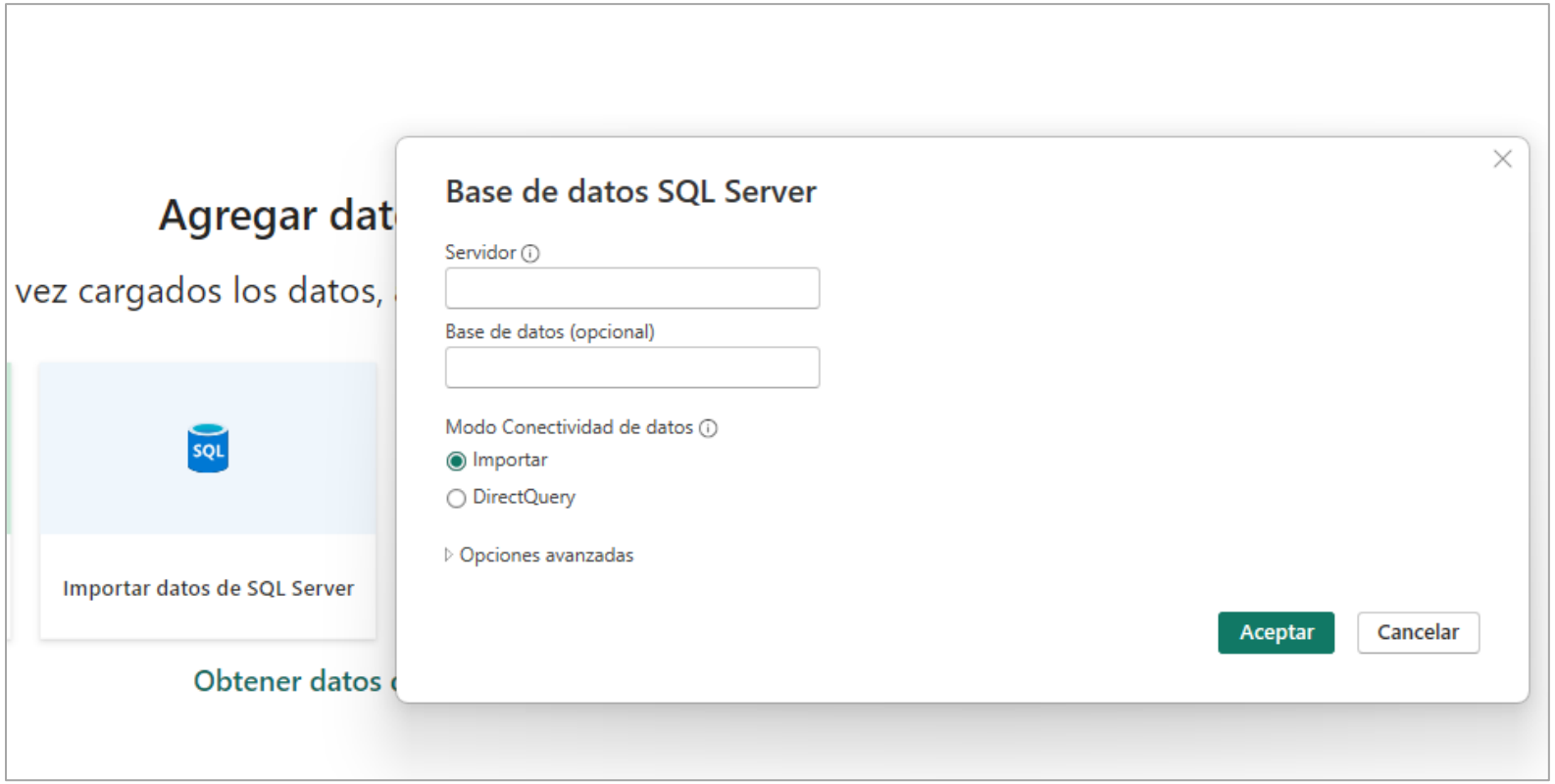
- Just showing how to create relationships between tables



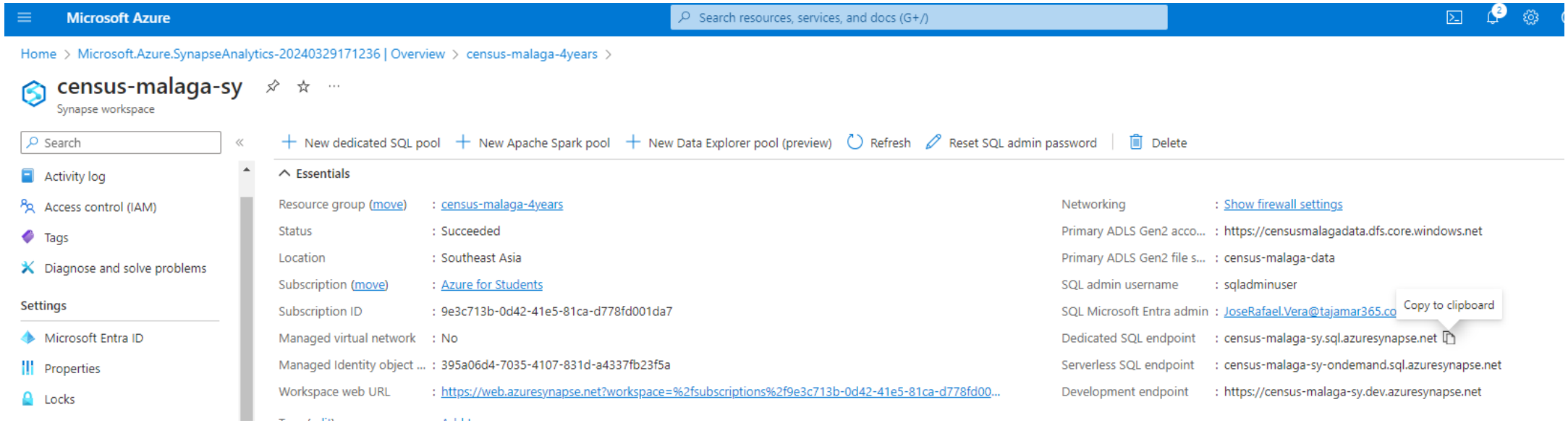
- All the tables created
- This tables will be used in PowerBI thanks to the Synapse connector



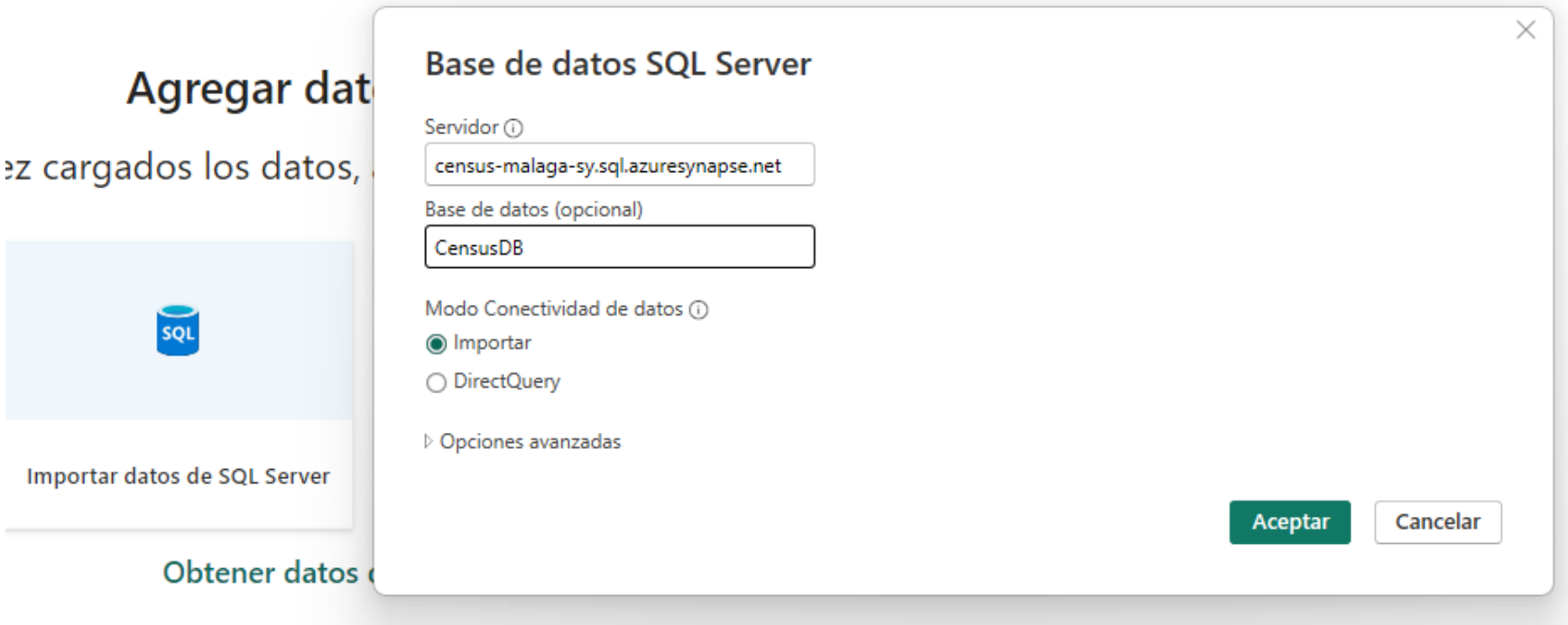
- Open PowerBI Desktop
- Create new report
- Get data from Azure Synapse analytics
- You will see this popup window below
- On the server blank paste the Dedicated SQL endpoint of your Synapse workspace
- On the database blank paste the database name on synapse workspace



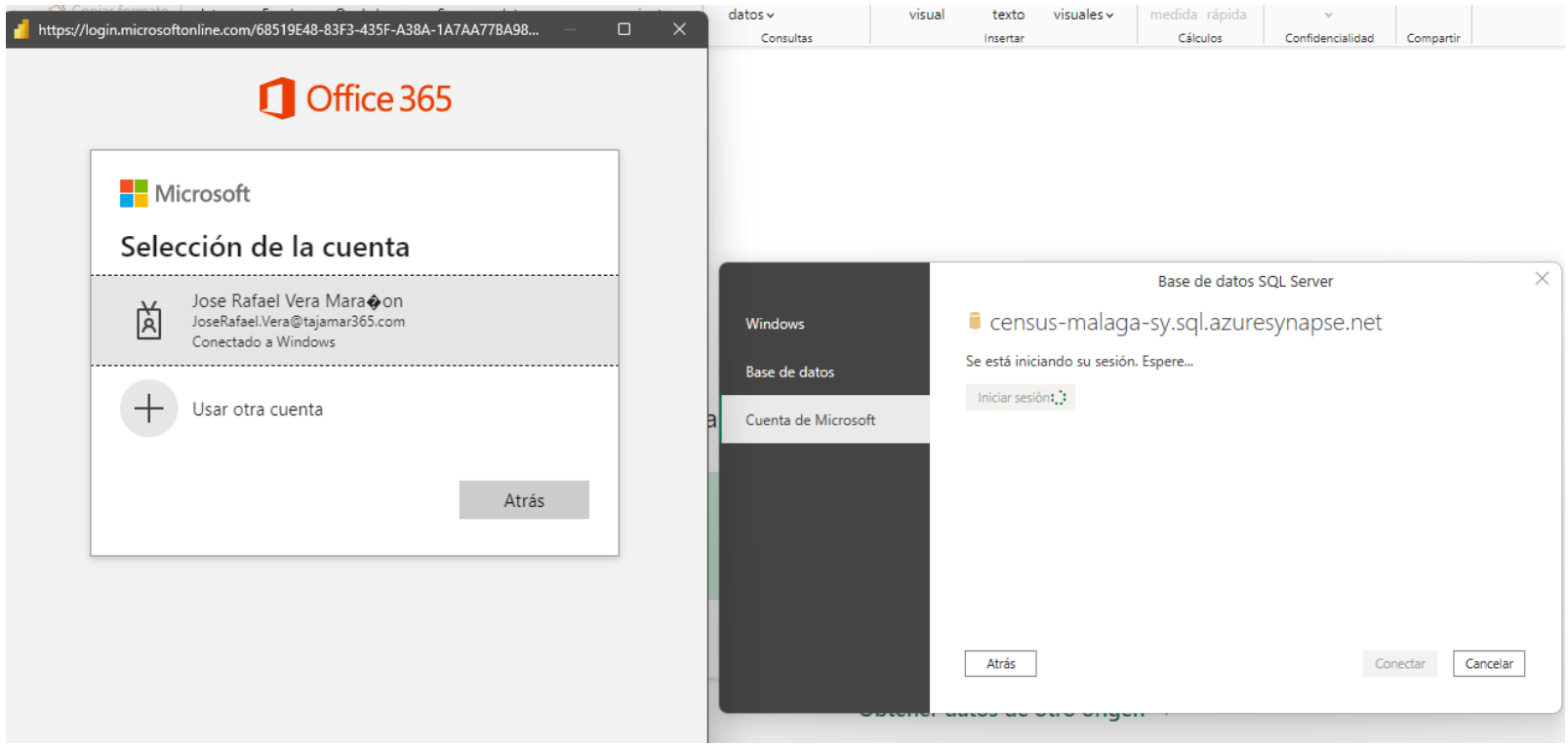
- Go to Synapse workspace
- Copy the Dedicated SQL endpoint and paste it on the server blank



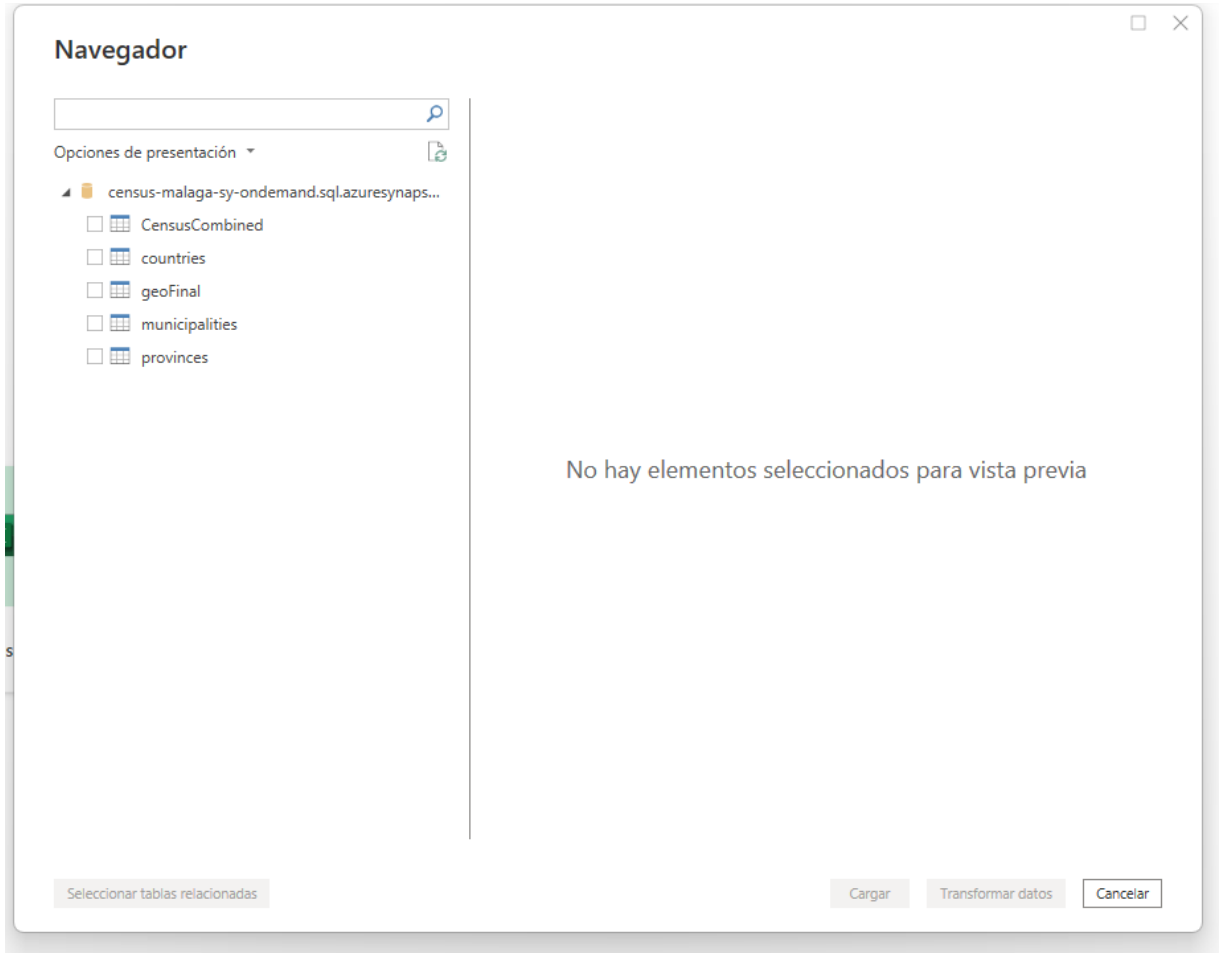
- You should see something like this
- Is up on you to select directquery (for realtime actualization) or Import (save a table snapshot into PBI)
- Accept



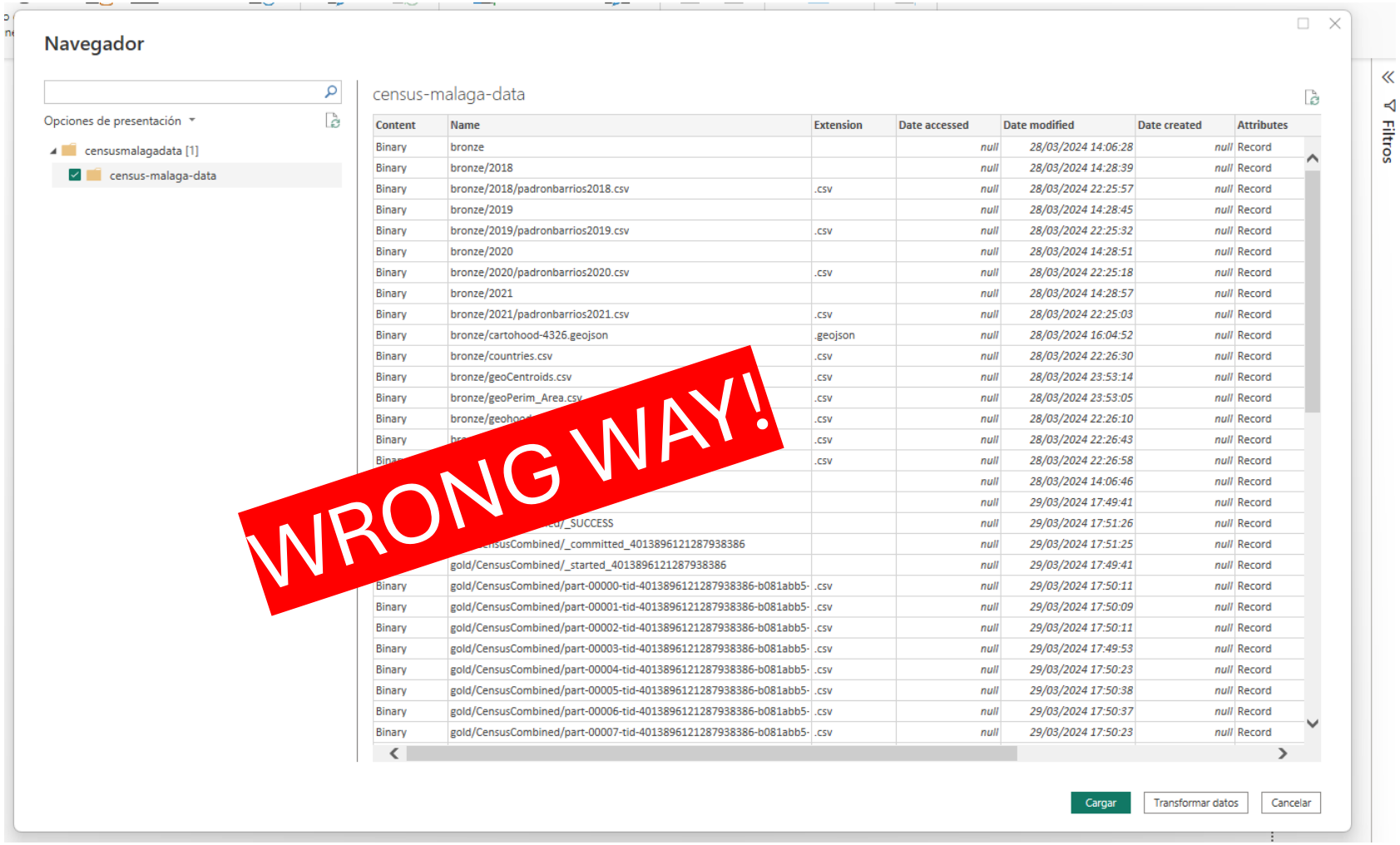
- After you need to be authenticated.
- I used the Microsoft account credentials



- 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