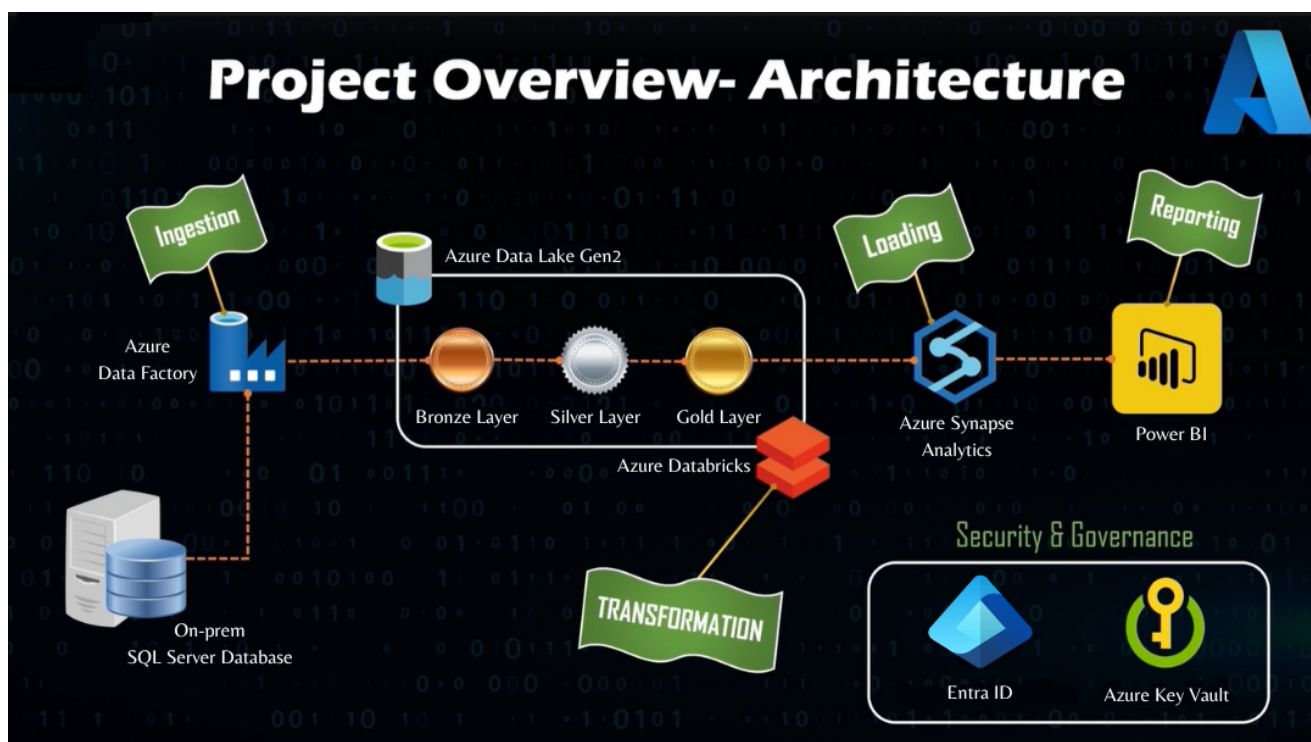


End-to-End-Data-Engineering-Project

Notion Webpage: <https://prickle-orangutan-450.notion.site/End-to-End-Data-Engineering-Project-15873b21d05080c3acaedb35bfb57fbb>



The proposed solution aims to address the challenge of migrating an on-premises SQL database to the Azure cloud, transforming the data, and leveraging it to gain insights using visualization tools.

To simulate the data, we used the AdventureWorks 2017 ([link](#)) sample database, which is initially stored locally on SQL Server Management Studio.

INGESTION



Azure Data Factory

This is a Microsoft cloud service that enables the integration and orchestration of data from multiple sources. It simplifies the creation of ETL (Extract, Transform, Load) or ELT workflows, allowing the movement and transformation of data between cloud and on-premises services. It is ideal for preparing data for analysis in tools such as Azure Synapse Analytics, Power BI, or storing it in Data Lakes.

Using Azure Data Factory, we created a pipeline where the locally hosted tables are transferred to the cloud into a Data Lake.



Azure Data Lake Gen 2

This is a cloud-based storage service optimized for massive-scale data, ideal for big data analytics. It combines the scalability and cost-efficiency of Azure Blob Storage with a hierarchical file system, making data organization and management easier. It is perfect for storing structured, semi-structured, and unstructured data and integrates seamlessly with tools like Azure Databricks, Synapse Analytics, and Azure Data Factory.

In our case, we store the data in an Azure Data Lake Gen 2 and used Databricks to transform the data.

TRANSFORMATION



Azure Databricks

This is a data analytics service built on Apache Spark, optimized for Azure. It enables real-time processing, transformation, and analysis of large volumes of data while combining machine learning, artificial intelligence, and big data capabilities. It provides a collaborative environment for data scientists, data engineers, and analysts, integrating with services like Azure Data Lake, Data Factory, Synapse Analytics, and Power BI.

Using Databricks, the data is transformed across the three layers:



Bronze Layer (Raw Layer):

Contains raw, unprocessed data as it is received from the sources. This data may include errors, duplicates, or unstructured formats.



Silver Layer (Clean Layer):

Stores clean and transformed data. This layer involves processes like validation, deduplication, error correction, and format standardization. The data here is ready for exploratory analysis or integration into other tables.

For each loaded table, it is checked whether any column contains the word "Date" or "date." If a date column is found, it is converted from its UTC format to a standard yyyy-MM-dd format. This is done using the **from_utc_timestamp** function (to handle the timezone) and the **date_format** function (to apply the desired format).



Gold Layer (Aggregated Layer):

Contains data optimized for final consumption, including analytics, reporting, or machine learning. This layer includes aggregations, calculations, metrics, and highly modeled tables.

For each table, the column names are retrieved and a transformation is applied to convert the column names from CamelCase format (e.g., ColumnName) to snake_case format (e.g., column_name).

This renaming is achieved through a loop that iterates over the characters of each column name and uses a concatenation approach to add underscores between uppercase letters. The **lstrip("_")** expression ensures that no leading underscore is present in the name.

LOADING



Azure Synapse Analytics

Is a Microsoft cloud analytics service that unifies enterprise data storage and big data analytics. It provides capabilities to query and analyze data at scale using SQL models (dedicated or serverless) and massive data processing with Apache Spark.

Azure Synapse Analytics allows the integration of data from various sources, such as Azure Data Lake and databases, making it ideal for real-time analytics, creating Power BI dashboards, and advanced modeling. Its flexibility makes it essential for implementing modern data warehouse architectures and integrated analytics.

It serves as a bridge to connect the data from the Gold layer with the visualization tool Power BI.

REPORTING



Power BI

Is a Microsoft business analytics tool that enables users to connect, transform, and visualize data in interactive reports and dashboards. It integrates with various data sources, both on-premises and in the cloud, such as Excel, SQL databases, Azure, and web services.

It is ideal for making data-driven decisions, thanks to its ease of use, advanced modeling and analytics capabilities, and its integration with tools like Azure Synapse Analytics and Microsoft Teams.

SECURITY AND GOVERNANCE



Entra ID

It is the primary service for identity management and access to cloud applications, providing authentication, authorization, and identity management for both internal and external users. In addition to these services, **Entra ID** integrates capabilities for access protection and permission management through advanced security policies.

Microsoft **Entra** also includes other security solutions, such as **Entra Permissions Management** and **Entra Verified ID**, for identity verification and data protection.

It is used to manage the credentials and access permissions for each member of our project.



Azure Key Vault

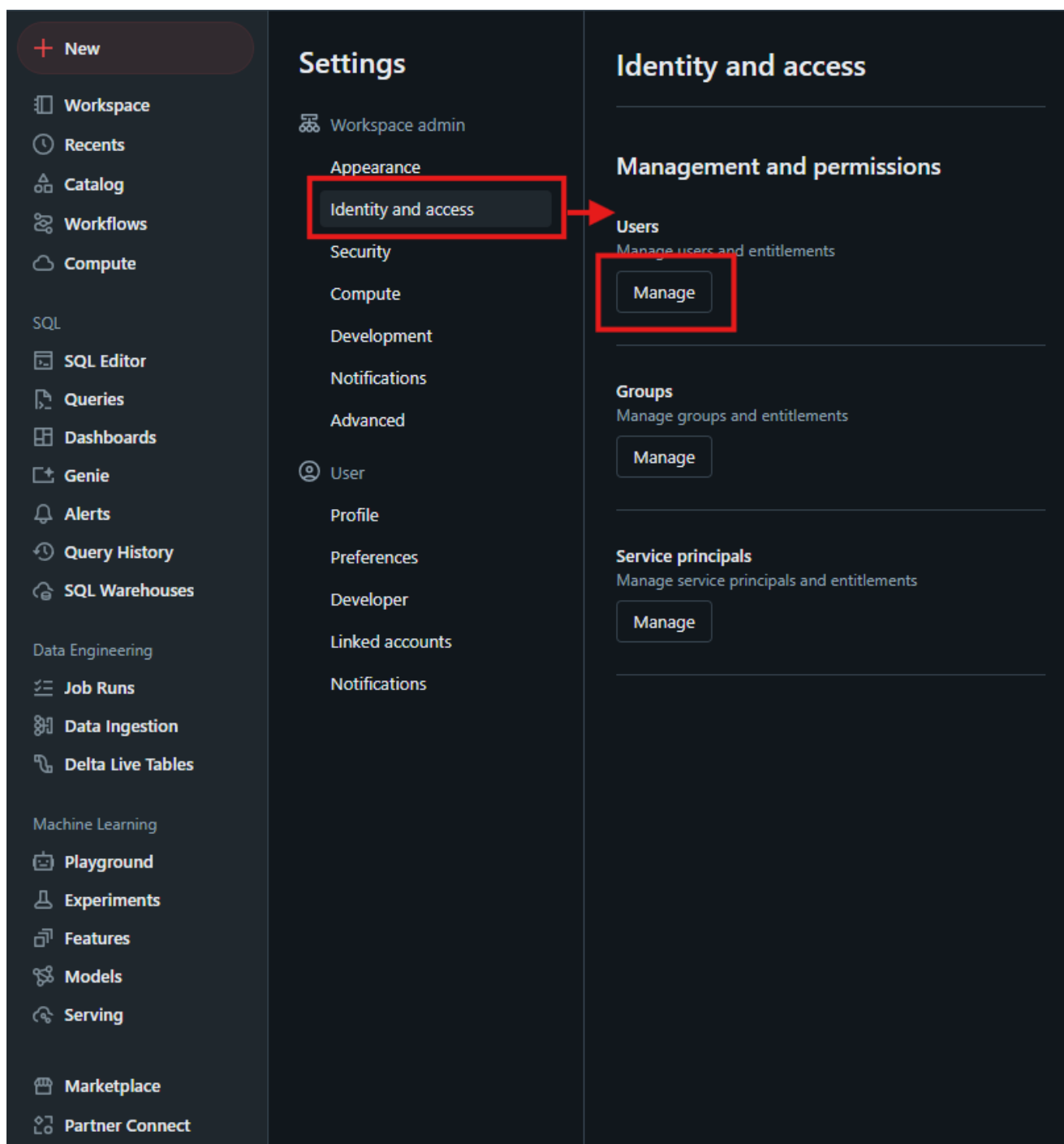
It is a Microsoft Azure cloud service that helps protect secrets, encryption keys, certificates, and other sensitive data. It allows organizations to manage access to these resources securely, maintaining control over who can access critical information.

With **Azure Key Vault**, developers and IT administrators can securely store and access secrets such as passwords, API keys, and SSL certificates. Additionally, it easily integrates with other Azure services, such as Azure Active Directory (Entra ID), to enforce access policies and ensure that only authorized users or applications can access these items.

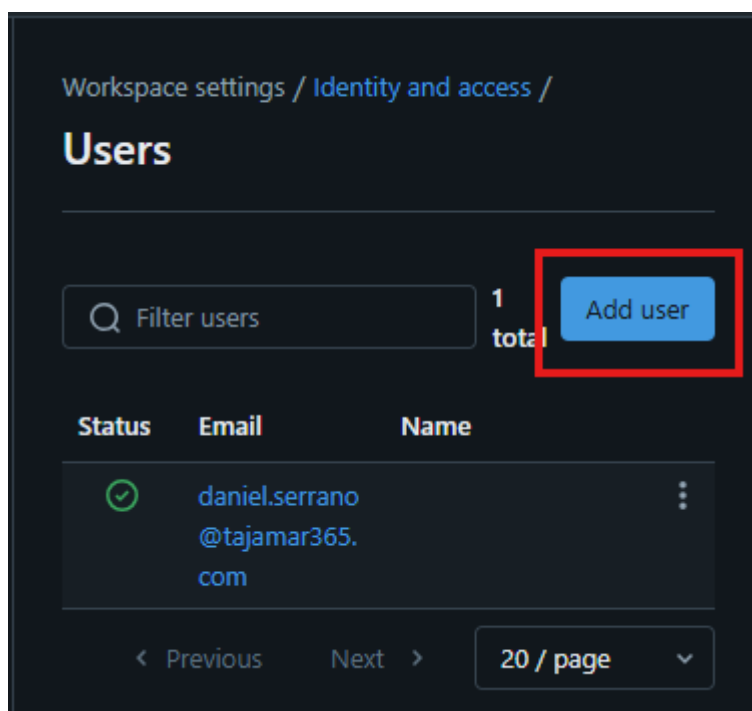
We store the credentials (primarily passwords) to enhance the security of all the services involved in this project.

Shared Databricks workspace

First, we need to create and access our workspace. Once inside the workspace, go to **Settings > Users:**



Add user:



Once added, we can configure each user by clicking on the :

Edit user

Email

daniel.garciavalencia@tajamar365.com

Groups

Not a member of any groups

Remove user

Entitlements

Admin access

Can manage this workspace and its users, groups, resources, and settings

Off ☐

Workspace access

Can access data engineering and ML environments

On ☒

Databricks SQL access

Can access SQL environment

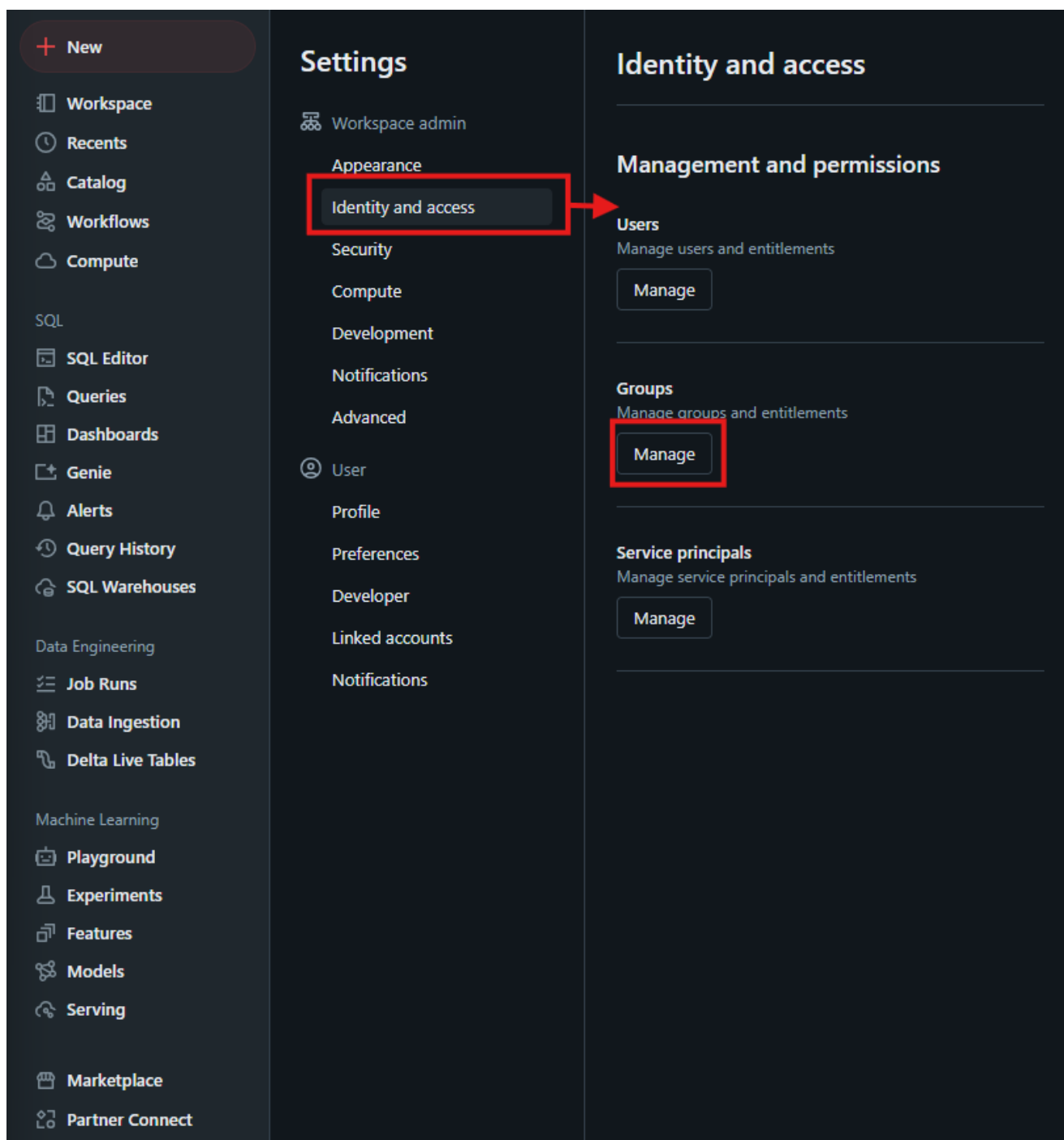
On ☒

Unrestricted cluster creation

Can create clusters; when disabled, the user is restricted to access granted by cluster policies

On ☒

Another useful management configuration is the “**Groups**” setting, where you can create workgroups and assign roles to each group in order to manage user groups effectively:



Just like with the “users” settings, we can create a new group and edit each of the existing ones by clicking on the ⋮

Groups

Filter groups

2
total

Add group

Name	Members	Source
admins	1	System ⓘ ⋮
users	4	System ⓘ ⋮

< Previous

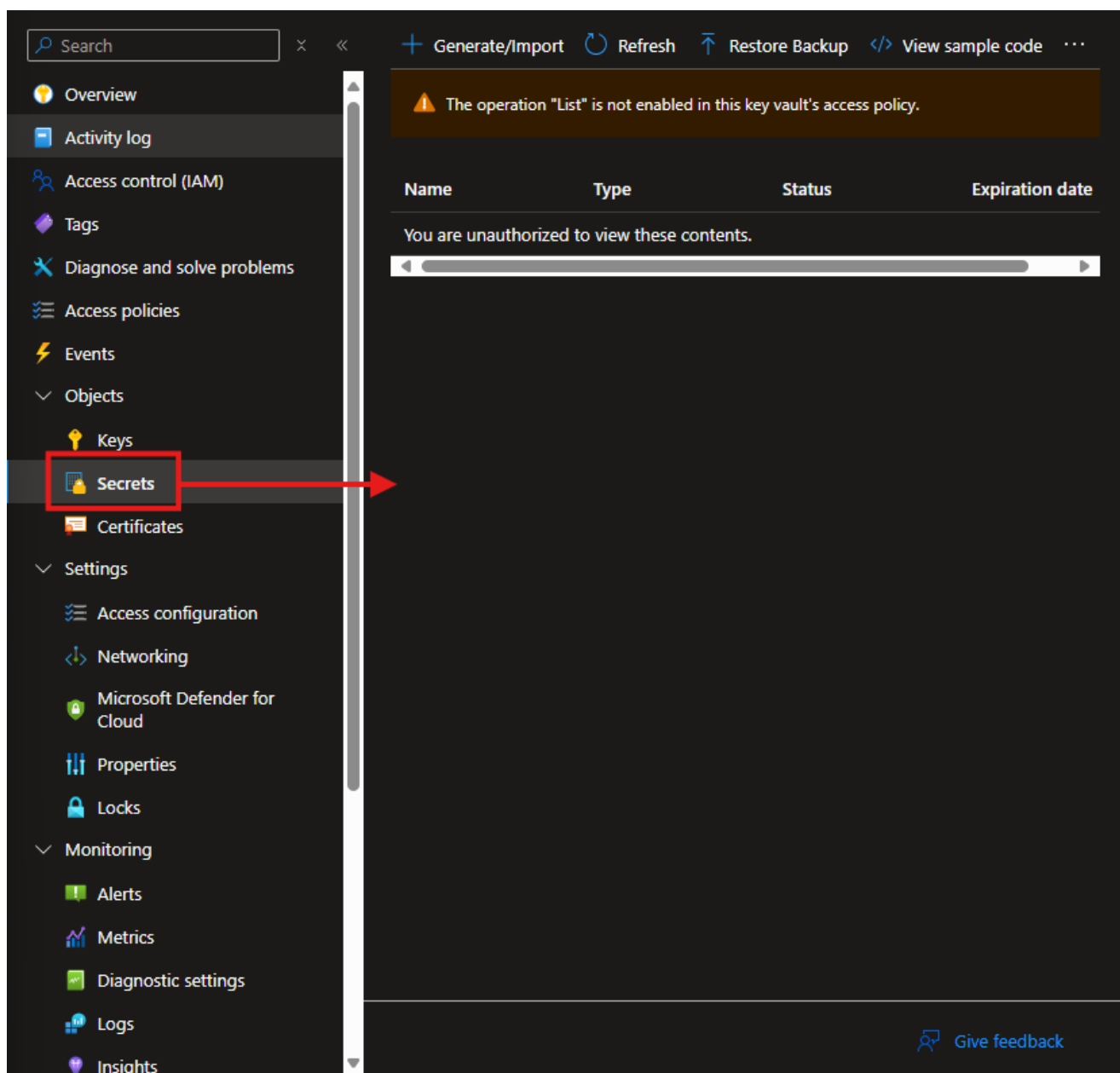
Next >

20 / page



Databricks token in Key Vault

We access the "Secrets" setting of our Key Vault:



We create a new secret, give it a name and copy the "Secret Value" :

Create a secret

Upload options: Manual

Name * *ⓘ*: databricks-token ✓

Secret value * *ⓘ*: Enter the secret.

Content type (optional):

Set activation date *ⓘ*: ☐

Set expiration date *ⓘ*: ☐

Enabled: Yes No

Tags: 0 tags

Where can we find it? We need to go to our Databricks workspace, launch it, and then go to **User > Settings**. Once there, navigate to **Developer > Access Tokens** :

The screenshot shows the Databricks Settings interface. On the left is a sidebar with two main sections: 'Settings' and 'User'. Under 'Settings', there are links for 'Workspace admin', 'Appearance', 'Identity and access', 'Security', 'Compute', 'Development', 'Notifications', and 'Advanced'. Under 'User', there are links for 'User', 'Profile', 'Preferences', 'Developer' (which is highlighted with a red box and a red arrow pointing to the right), 'Linked accounts', and 'Notifications'. The main content area on the right is titled 'Developer' and has the subtitle 'Manage your development settings'. It contains three sections: 'Access tokens' with a 'Manage' button (highlighted with a red box), 'SQL query snippets' with an 'SQL Editor' button, and 'Editor settings'. The 'Editor settings' section includes 'General', 'Spark tips' (with a toggle switch set to 'On'), 'Databricks Advisor' (with a toggle switch set to 'On'), and 'Automatically launch and attach to clusters'.

Settings

- Workspace admin
- Appearance
- Identity and access
- Security
- Compute
- Development
- Notifications
- Advanced

User

- User
- Profile
- Preferences
- Developer**
- Linked accounts
- Notifications

Developer

Manage your development settings

Access tokens

Set up secure authentication to Databricks API using access tokens

Manage

SQL query snippets

Configure SQL query snippets. Note: SQL query snippets will be moving to the SQL editor dropdown menu.

SQL Editor

Editor settings

General

Spark tips

Enrich notebook error stack traces by displaying high-level "error hints" which explain otherwise-confusing errors.

On

Databricks Advisor

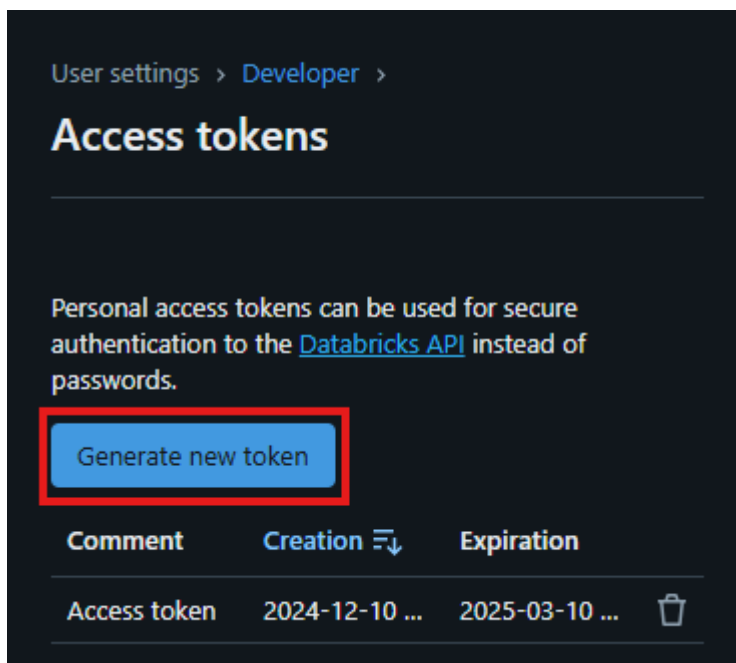
Databricks Advisor displays context-aware optimization hints in the notebooks. [Learn more.](#)

On

Automatically launch and attach to clusters

When running commands in Notebooks, automatically launch and attach to clusters without prompting.

Generate a new token and, very importantly, make sure to copy it, as this is the only time we can view its content:



Back to Key Vault, we copy the value:

Creating the on-premise SQL server database

As a prerequisite we need to have SQL Server, SQL Server Management Studio 20 and the AdventureWorks2017 bak file in our computer:

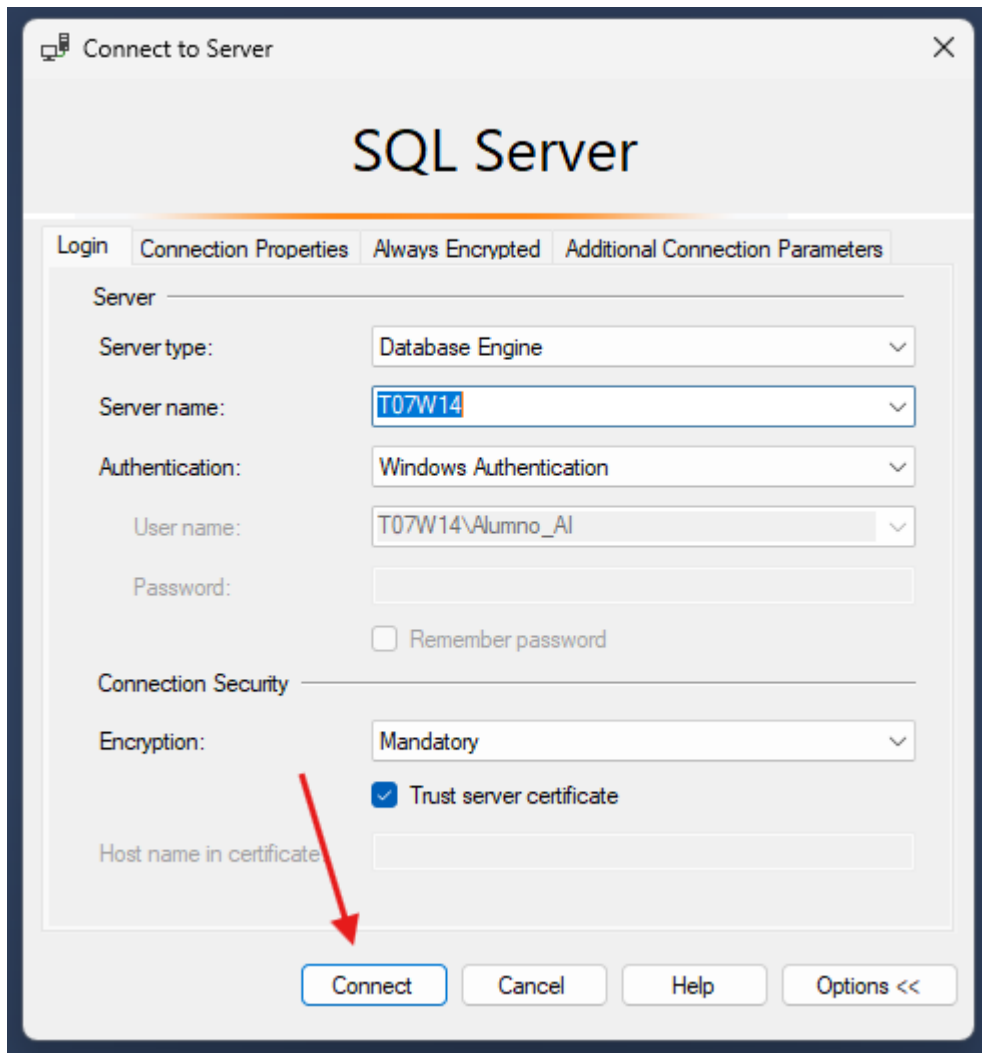
- [SQL Server](#)
- [SQL Server Management Studio 20](#)

- [AdventureWorksLT2017.bak](#)

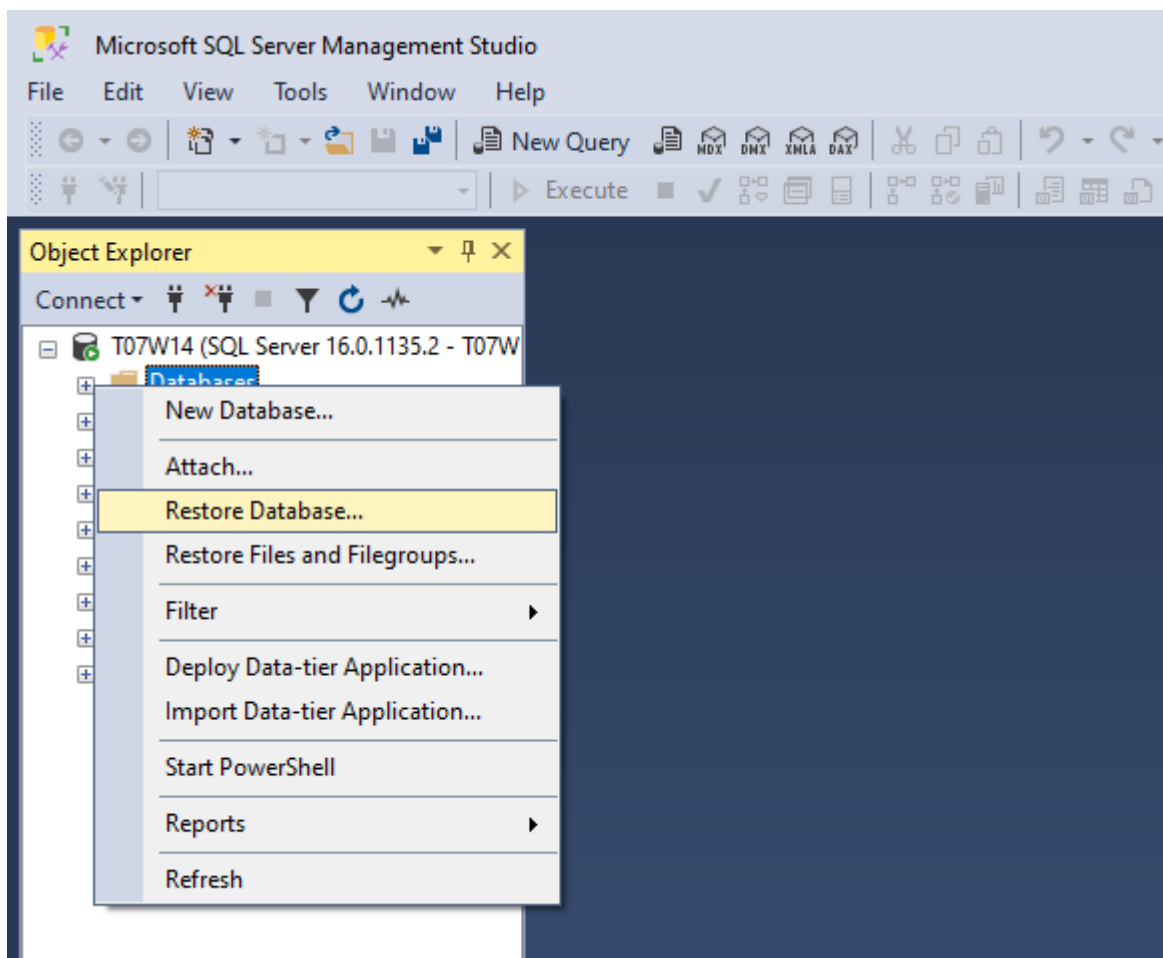
Once we have everything, the AdventureWorksLT2017.bak file has to be placed in the next directory:

`C:\Program Files\Microsoft SQL Server\MSSQL16.MSSQLSERVER\MSSQL\Backup`

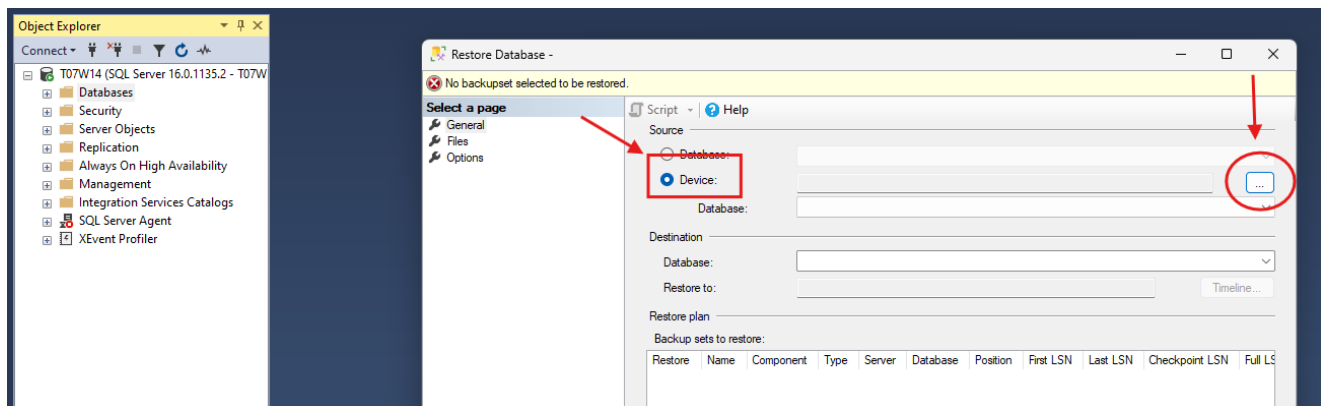
1. Open SQL Server Management Studio 20



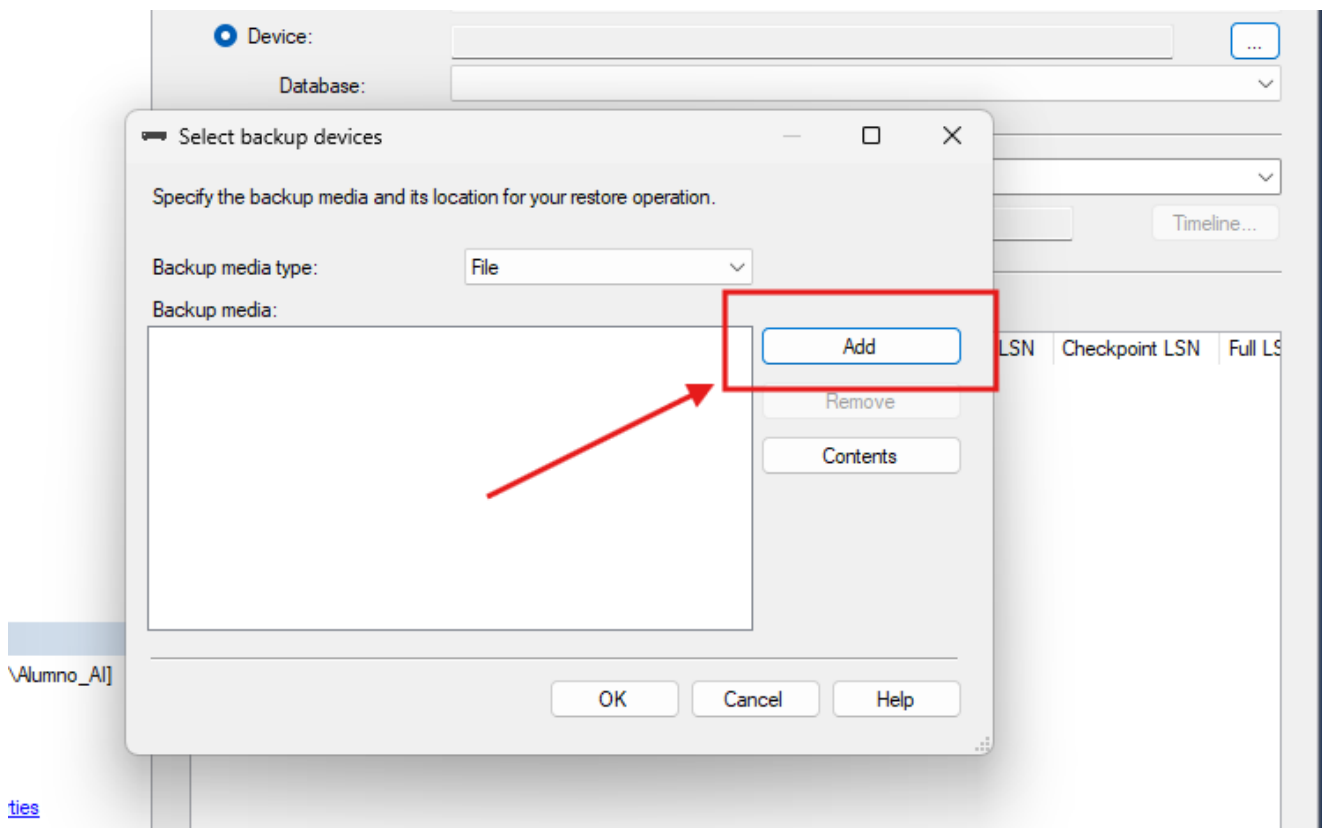
1. When it connects, we have to right clic on "databases" and then "Restore Database"



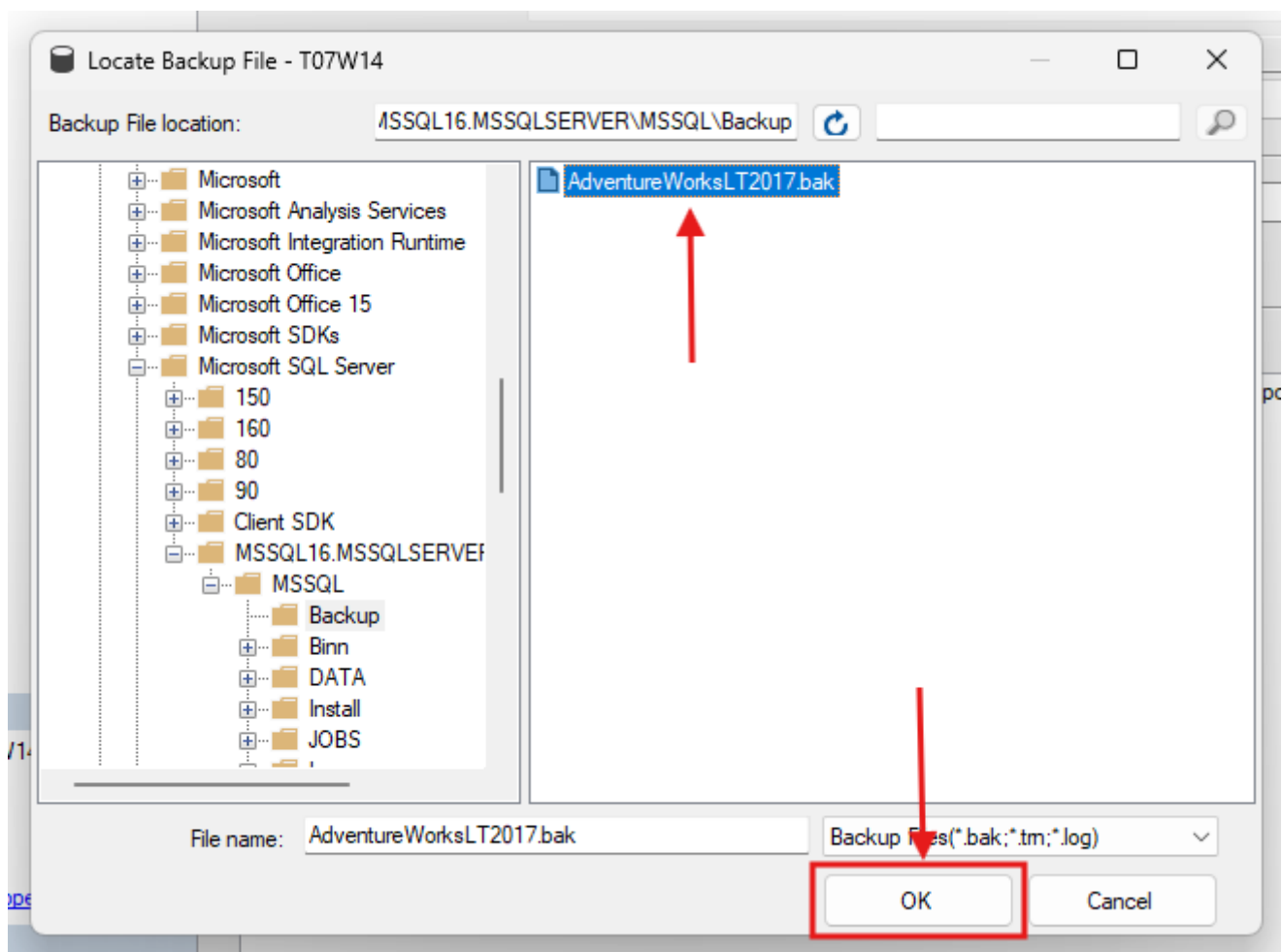
1. Choose device and clic the three dots on the right

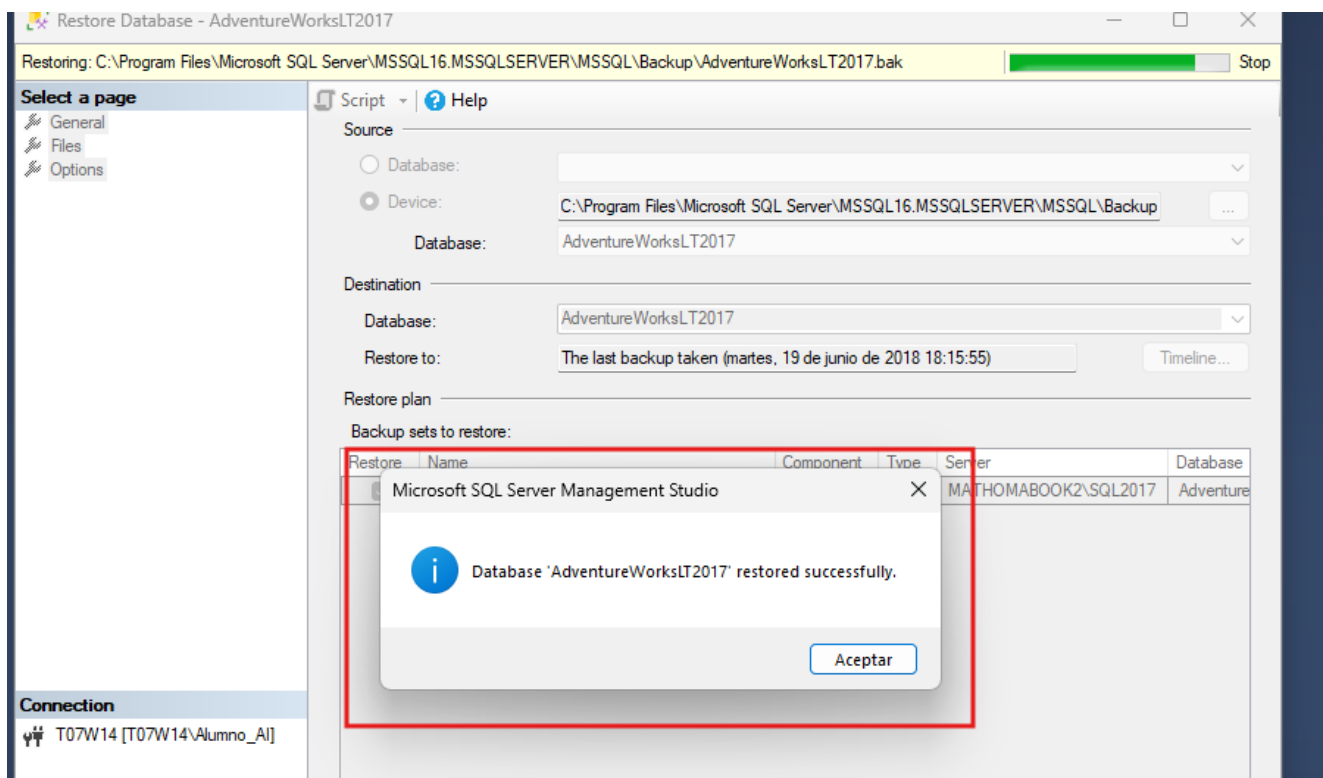


1. A new window will show in which we're selecting "Add"



1. Finally, we select the .bak file and every "Ok" button until this window shows up:

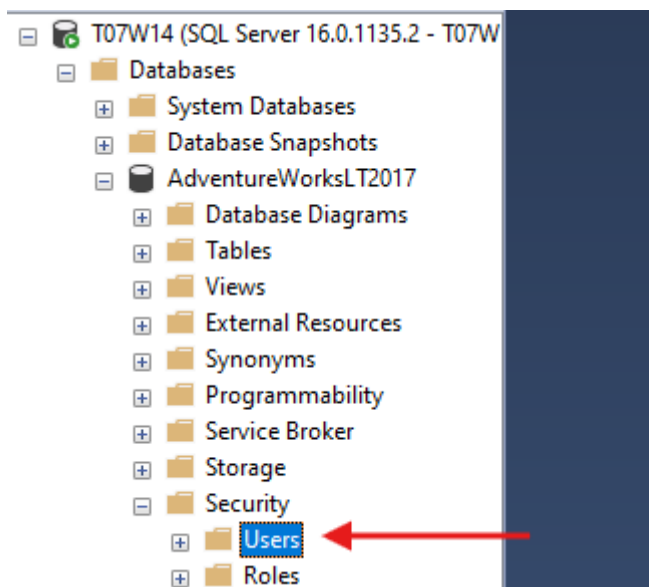




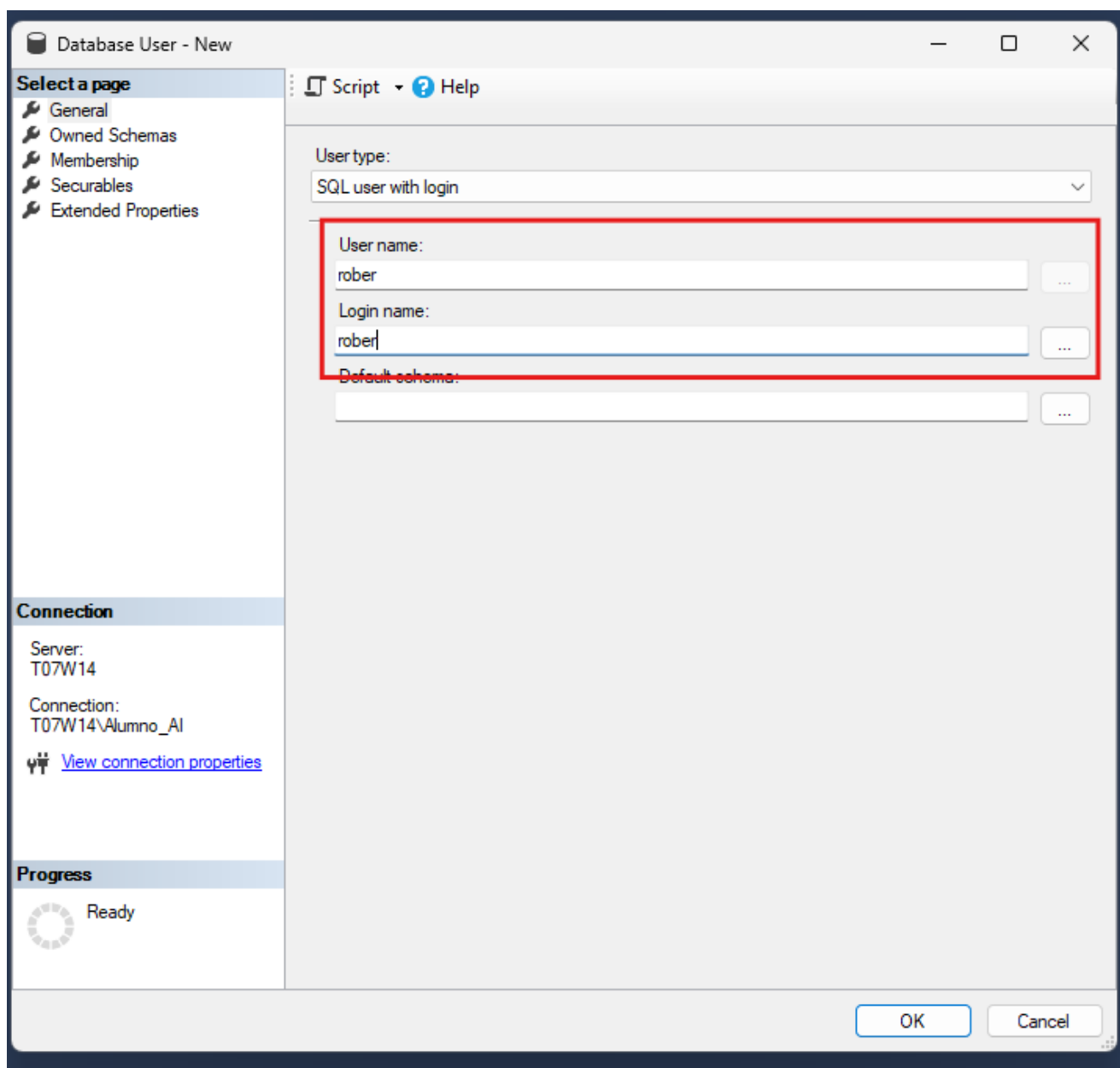
Configure SQL Server for remote access:

To make sure we can connect through Azure to our local database this are the settings that need to be configured:

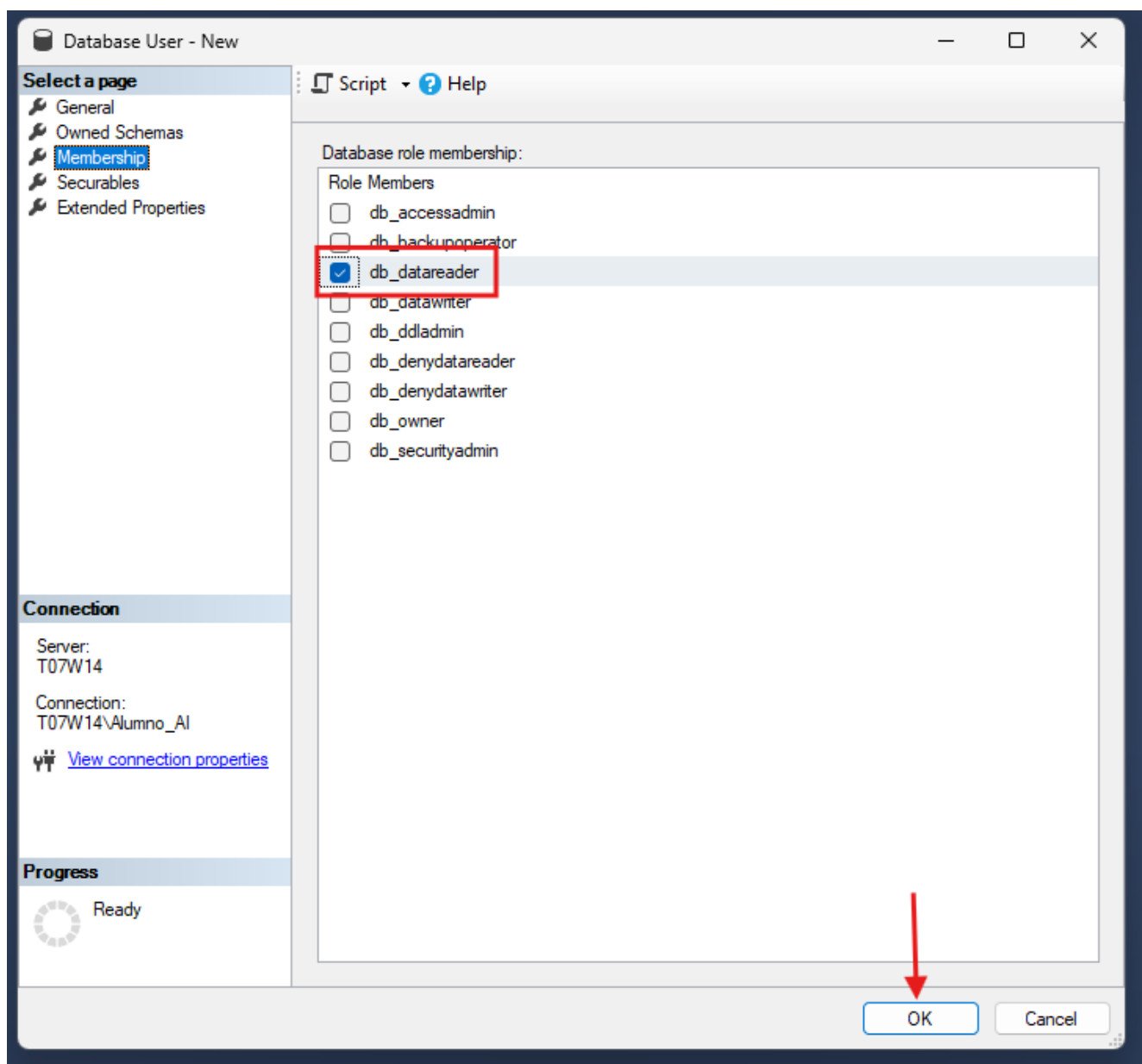
1. Create new user, under "Security" > "Users", right clic and "New User":



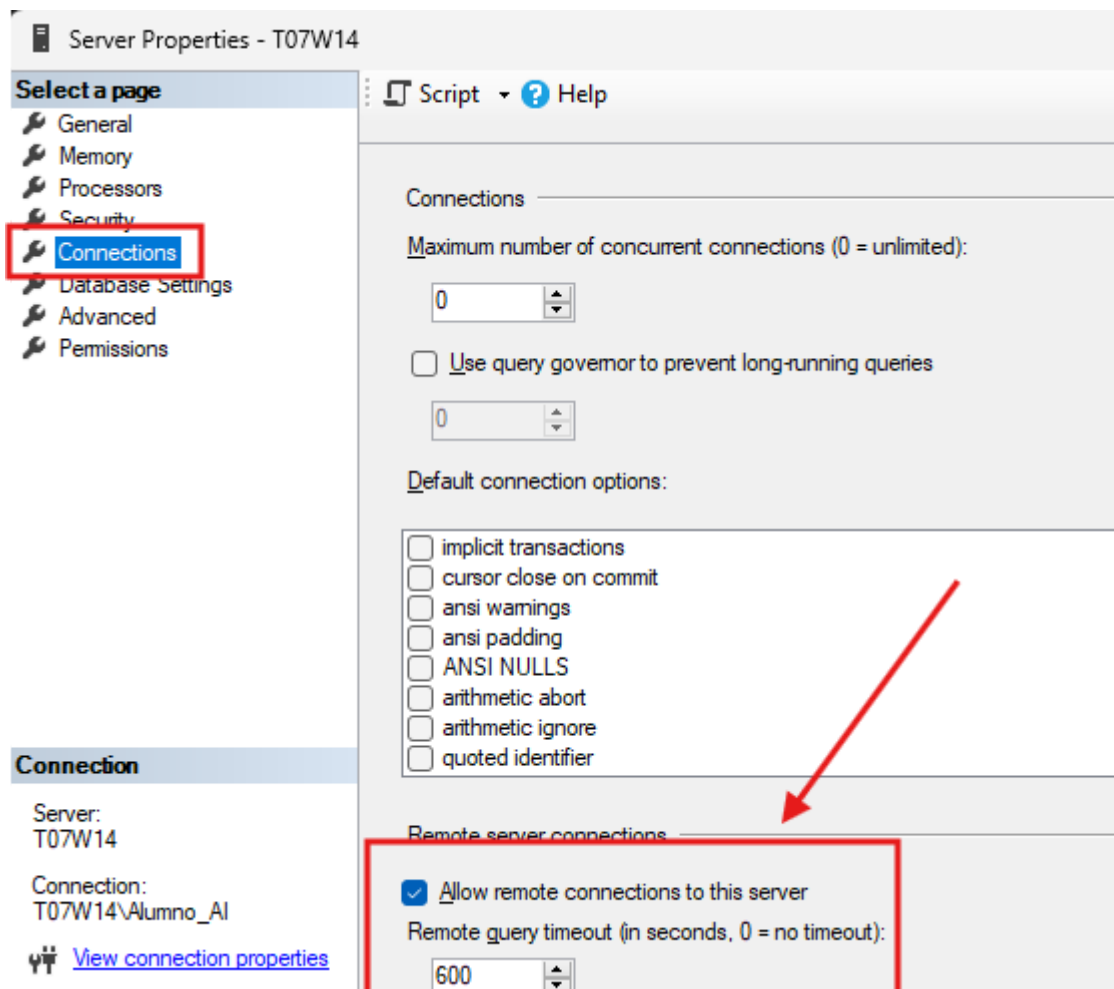
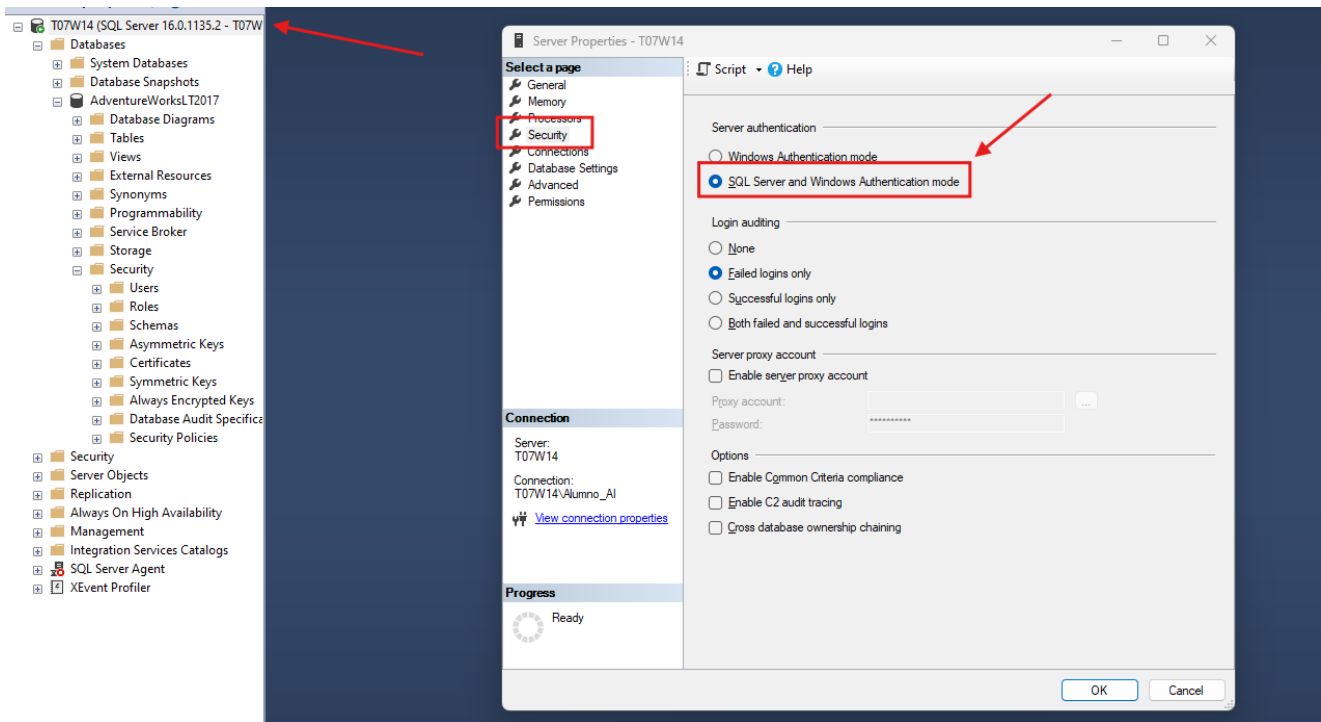
1. Fill the fields...



1. Check db_datareader under the Membership tab and clic "OK"



1. Also ensure that server authentication is on "SQL Server and Windows Authentication mode" and the server allows remote connections. So, to check this we have to right clic on the server:

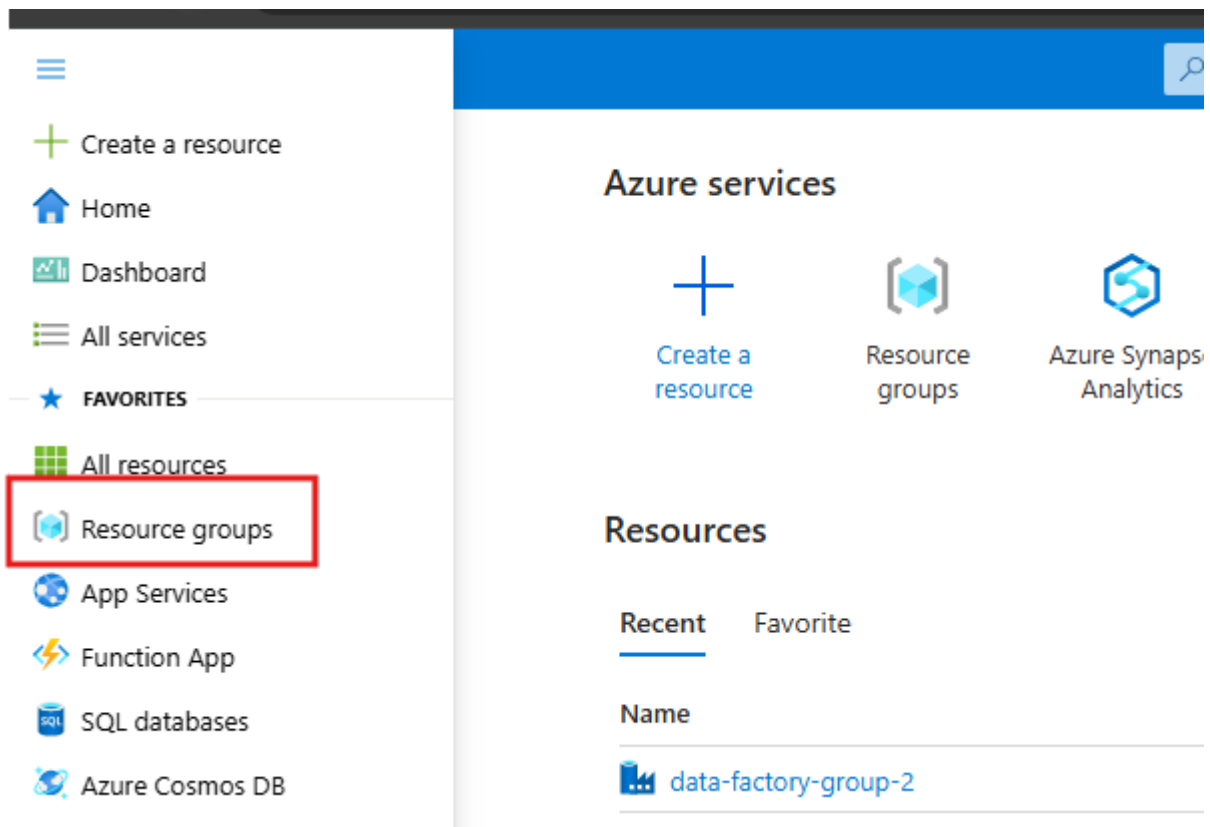


With this steps we have almost everything configured on the local part.

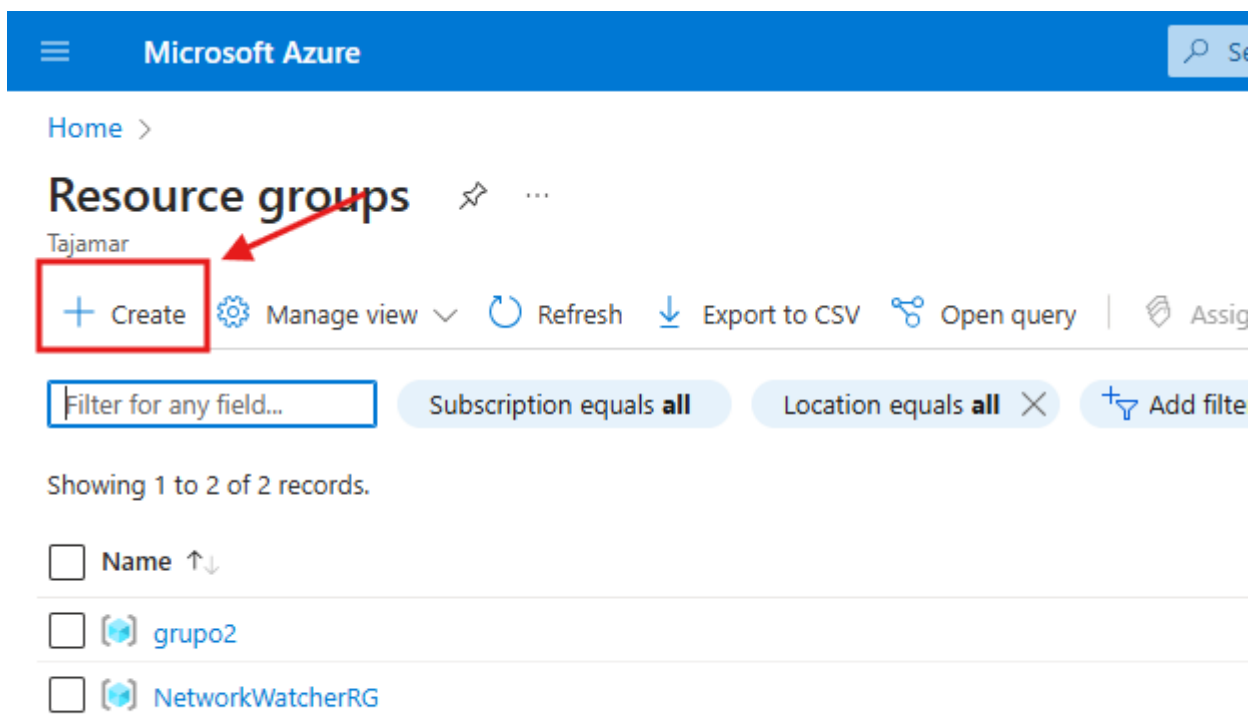
Create Resource Group




1. Once in Azure, we click on "Resource groups":



1. Inside Resource groups, we create a new one:



1. We give the resource group a name (I'm calling ours "ddmrs-proyecto-1"), and the region in which the group is going to be located.

 Microsoft Azure Search resources, services, and documentation

[Home](#) > [Resource groups](#) >

Create a resource group

Basics

Tags

Review + create

Resource group - A container that holds related resources for an Azure solution. The resource group can include all the resources for the solution, or only those resources that you want to manage as a group. You decide how you want to allocate resources to resource groups based on what makes the most sense for your organization. [Learn more](#)

Project details

Subscription * ⓘ

Azure for Students (5c25d654-8aea-4cad-9bbf-b122d958985d) ▼

Resource group * ⓘ

ddmrs-proyecto-1 ✓

Resource details

Region * ⓘ

(Europe) France Central ▼

1. I created no tags here.

[Home](#) > [Resource groups](#) >

Create a resource group ...

[Basics](#) [Tags](#) [Review + create](#)

Apply tags to your Azure resources to logically organize them by categories. A tag consists of a key (name) and a value. Tag names are case-insensitive and tag values are case-sensitive. [Learn more](#)

Name	Value	Resource
<input type="text"/>	: <input type="text"/>	Resource group

1. And finally, we review and create it.

[Home](#) > [Resource groups](#) >

Create a resource group ...



Validation passed.

[Basics](#) [Tags](#) [Review + create](#)

Basics

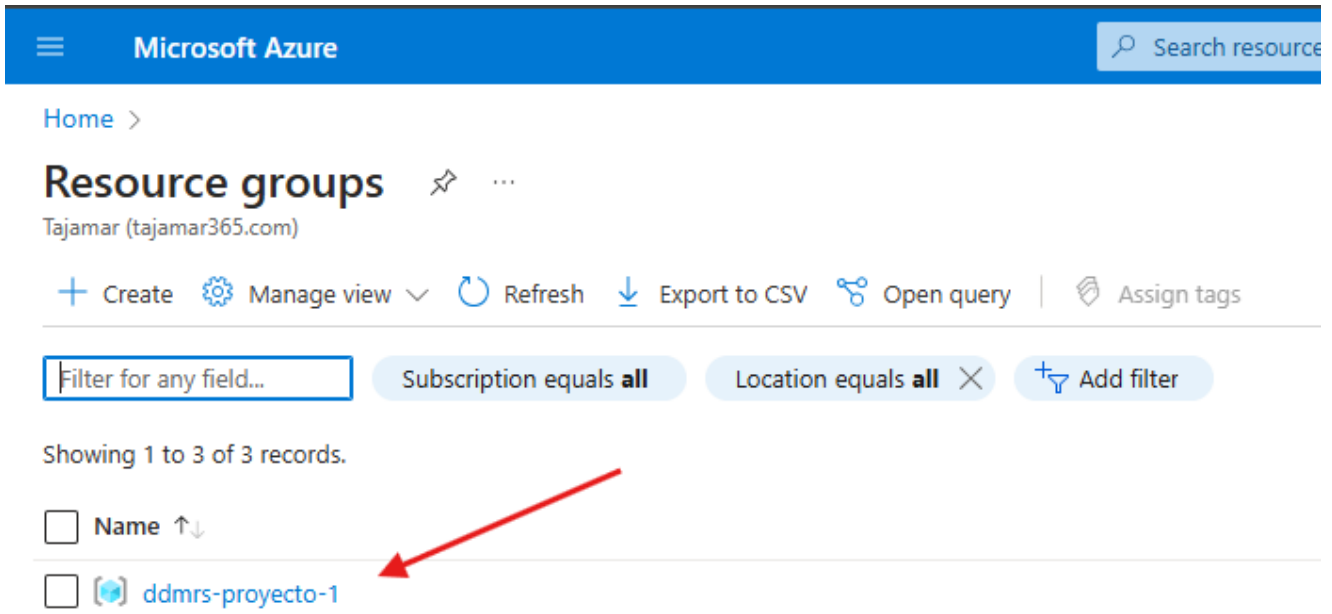
Subscription	Azure for Students
Resource group	ddmrs-proyecto-1
Region	France Central

Tags

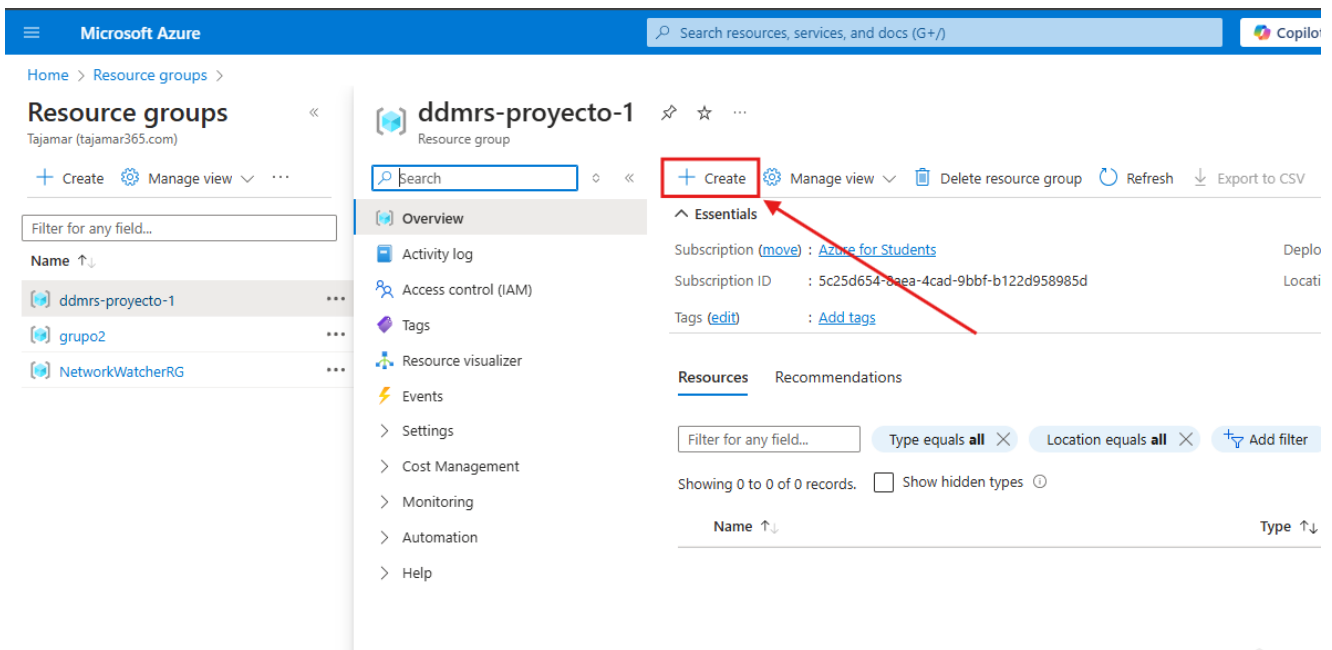
None

Create Azure Synapse Analytics and Datalake Storage Gen2

1. The first step is opening our resource group from the resource groups tab:



1. Once we select it, we click on "create":



1. Inside the marketplace, search for synapse analytics, and once found, create the resource:

Microsoft Azure

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

Home > Resource groups > ddmrs-proyecto-1 >

Marketplace

Get Started

Service Providers

Management

Private Marketplace

Private Offer Management

My Marketplace

Favorites

My solutions

Recently created

Private plans

Categories


Analytics (79)

Databases (35)

synapse

☐ Azure services only

Showing 1 to 20 of 106 results for 'synapse'. [Clear search](#)




Azure Synapse Analytics


Microsoft

Azure Service

Limitless analytics service with unmatched time to insight

Create 

Azure Synapse Analytics





Azure Synapse Analytics (private link hubs)

Microsoft

Azure Service

Connect to Azure Synapse Studio using private endpoints

Create 




Synapse Data Fabric


Spektra Systems LLC

SaaS

Synapse Data Fabric is a comprehensive data management platform that unifies disparate data sources

Software plan starts at less than **€0.001/month**

Subscribe 




Datometry by Azure Synapse

Datometry

SaaS

Run existing Teradata natively on Azure

Starts at **€8,465.608/month**

Subscribe 

1. We then configure the basics page as shown below, and also create the Data Lake Storage Gen2 and file system (Bronze) that we will use to store the data from the on-premise sql server.

Home > Resource groups > ddmrs-proyecto-1 > Marketplace >

Create Synapse workspace ...

Project details

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

Subscription * ⓘ Azure for Students (5c25d654-8aea-4cad-9bbf-b122d958985d) ✓

Resource group * ⓘ ddmrs-proyecto-1 ✓

Managed resource group ⓘ Enter managed resource group name

Workspace details

Name your workspace, select a location, and choose a primary Data Lake Storage Gen2 file system to serve as the default location for logs and job output.

Workspace name * ddmrs-synapse-analytics ✓

Region * France Central ✓

Select Data Lake Storage Gen2 * ⓘ ☒ From subscription ☐ Manually via URL

Account name * ⓘ Create new

File system name * Create new

i We will automatically grant the workspace identity data access to the specified Data Lake Storage Gen2 account using the Storage Blob Data

Select Data Lake Storage Gen2 * ⓘ ☒ From subscription ☐ Manually via URL

Account name * ⓘ Create new

File system name * Create new

Select Data Lake Storage Gen2 * ⓘ

☒ From subscription ☐ Manually via URL

Account name * ⓘ

File system name *

[Create new](#)

Data Lake Storage Gen2 account

Name *

ddmrsadlsgen2 ✓

OK

Cancel

Contact an **Owner** of the storage account and ask

Select Data Lake Storage Gen2 * ⓘ

☒ From subscription ☐ Manually via URL

Account name * ⓘ

(New) ddmrsadlsgen2

[Create new](#)

File system name *

[Create new](#)

Data Lake Storage Gen2 file system

Name *

bronze ✓

OK

Cancel

Studio

- Assign yourself and other users to the Stor

1. We are not modifying any of the next config pages, so we can skip directly to "Review and create"

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	ddmrs-proyecto-1
Region	France Central
Workspace name	(new) ddmrs-synapse-analytics
Data Lake Storage Gen2 account	(new) https://ddmrsadlsgen2.dfs.core.windows.net
Data Lake Storage Gen2 file system	(new) bronze
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](#)

Creating bronze-silver-gold hierarchy

1. Access the datalake storage account inside Azure portal
2. Go to the "Storage browser" tab
3. Click on "Blob containers"

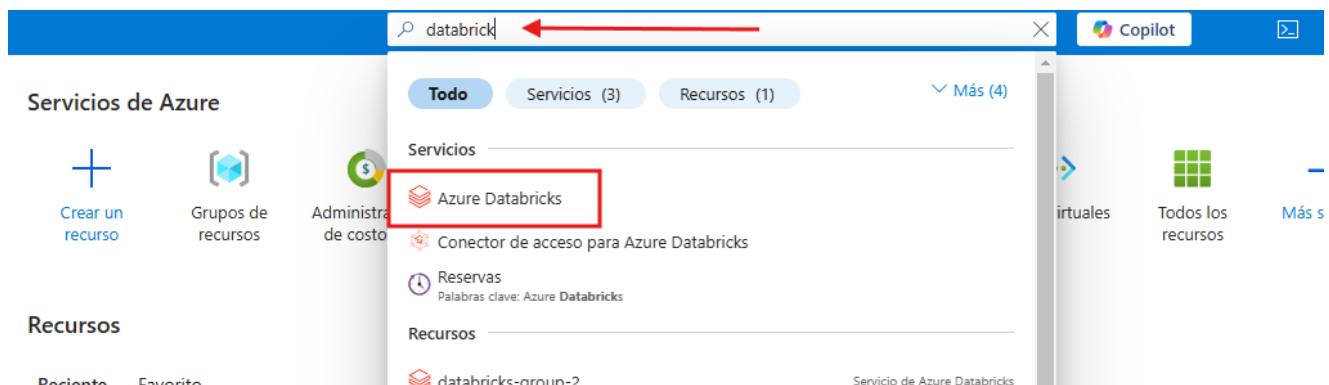
4. We should be able to see that bronze is already created. That was done while creating the Azure Synapse Analytics workspace
5. Click on "Add container" at the top of the screen
6. Write "silver" inside the "Name" text input, then click on "Create"
7. Repeat step 6 for "gold"

Create Azure Data Factory

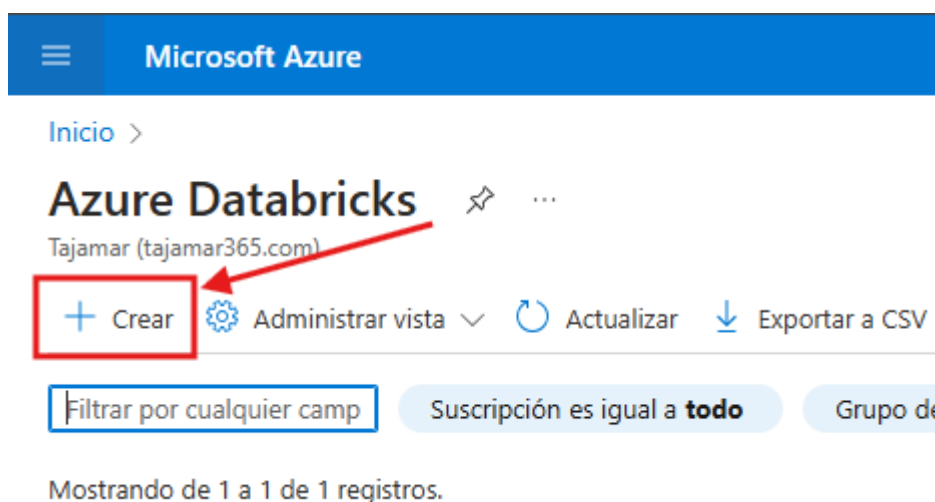
1. We look for Azure Databricks in the azure search bar and clic on "Azure Data Factory"
2. Then clic on "create"
3. In the "Basics" tab we fill the following information:
 1. Resource group → ddmrs-proyecto-1
 2. Name → ddmrs-data-factory
 3. Region → France Central
4. Finally, click on "Review + create"

Create Azure Databricks

1. We look for Azure Databricks in the azure search bar and click on "Azure Databricks"



1. Then click on "create"



1. We choose the resource group

Microsoft Azure

Buscar recursos, servicios y de

[Inicio](#) > [Azure Databricks](#) >

Creación de un área de trabajo de Azure Databricks ...

Datos básicos

Redes

Cifrado

Security & compliance

Etiquetas

Revisar y crear

Detalles del proyecto

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

Suscripción * ⓘ

Azure for Students (5c25d654-8aea-4cad-9bbf-b122d958985d) ▼

Grupo de recursos * ⓘ

ddmrs-proyecto-1 ▼

[Crear nuevo](#)

Detalles de instancia

Nombre del área de trabajo *

ddmrs-databricks ✓

Región *

West Europe ▼

Plan de tarifa * ⓘ

Premium (+ controles de acceso basados en roles) ▼

ⓘ Hemos seleccionado el plan de tarifa recomendado para su área de trabajo. Puede cambiar el nivel en función de sus necesidades.

×

Nombre del grupo de recursos administrados

Enter name for managed resource group

1. We don't change any of the following configurations, so we just need to review and create the resource:

[Inicio](#) > [Azure Databricks](#) >

Creación de un área de trabajo de Azure Databricks ...

Validación correcta

[Datos básicos](#) [Redes](#) [Cifrado](#) [Security & compliance](#) [Etiquetas](#) [Revisar y crear](#)

Resumen

Datos básicos

Nombre del área de trabajo	ddmrs-databricks
Suscripción	Azure for Students
Grupo de recursos	ddmrs-proyecto-1
Región	West Europe
Plan de tarifa	premium
Nombre del grupo de recursos administrados	

Redes

Implementar un área de trabajo de Azure Databricks con conectividad de clúster segura (sin IP pública)	No
Implementar área de trabajo de Azure Databricks en una red virtual (VNet) propia	No

Cifrado

Habilitar cifrado de infraestructura	No
Enable CMK for Managed Disks	No
Enable CMK for Managed Services	No

[Crear](#)[< Anterior](#)[Descargar una plantilla para la automatización](#)

Create Azure Key Vault

1. We look for Azure Databricks in the azure search bar and clic on "Azure Key Vaults"
2. Then clic on "create"
3. Afterwards fill the info like this:
 1. Key vault name → ddmrs-key-vault
 2. Region → France Central
 3. Pricing tier → Standard
4. Lastly change to the configuration tab and check the "Vault access policy option" and create the resource

Creating SQL Database User Login as a Secret in Azure Key Vault

Username

1. Access the key vault inside Azure portal
2. Go to the "Secrets" tab inside the "Objects" menu
3. Click on "Generate/Import"
4. Fill the information like this:
 1. Name → username
 2. Secret value → rober (Paste here your sql database username instead of "rober")
5. Click on "Create"

Password

1. Access the key vault inside Azure portal
2. Go to the "Secrets" tab inside the "Objects" menu
3. Click on "Generate/Import"
4. Fill the information like this:
 1. Name → password
 2. Secret value → (Paste here your sql database user password)
5. Click on "Create"

Connecting SQL Database to Azure Datafactory

1. Access the key vault inside Azure portal
2. Click on "Launch studio"
3. Click on the "Manage" tab at the left sidebar
4. Click on "Integration Runtimes"
5. Click on "New"
6. Select "Azure, Self-Hosted" and "continue"
7. Select "Self-Hosted"
8. Give it a name, we named it "SHIR" and click on "Create"
9. Click on "Express Setup" and install that in the computer with SQL Database

Creating the Pipelines for Copying Data from SQL Database to Azure Datalake Storage Gen2

1. From the Azure Data Factory studio click on the "Author" tab
2. Click on the "+" icon, then click on "pipeline"
3. In the Activity search bar write "Lookup", drag and drop the result to the right
4. In the bottom tabs select "settings"
5. Add a new source dataset
6. We search and select sql server
7. On the linked service section we select "onpremsqlserver" and click "ok"

Creating the Computing Unit in Azure Databricks

Creating the Notebooks in Azure Databricks

Creating the Connection Notebook

Creating the First Transformation Notebooks (bronze - silver)

Creating the Second Transformation Notebooks (silver - gold)

Connecting Azure Databricks to Data Factory

Creating a Data Factory Pipeline to Run Databricks Notebooks

Creating SQL Pool in Azure Synapse Analytics

Creating a Stored Procedure with Parameters that can Dynamically Create

Creating a New Link Service Connection to Connecto to the Serverless SQL Database

Creating the Link Service

Creating Azure Synapse Analytics Pipeline to Create a Delta View for Each Table

Connecting PowerBI to Azure Synapse Analytics

Creating Permissions For Synapse Analytics

Access control

Grant others access to this workspace by assigning roles to users, groups, and/or service principals. [Learn more](#)

[+ Add](#) [Refresh](#) [Remove access](#)

Type: **All** Role: **All** Scope: **All**

Showing 1 - 2 of 2 role assignments at all scopes in the workspace (1 user(s), 0 group(s), 1 service principal(s))

<input type="checkbox"/>	Name	Type	Role	Scope
<input type="checkbox"/>	Daniel García Valencia daniel.garciavalencia@tajamar365.com	User	Synapse Administrator	Workspace
<input type="checkbox"/>	synapse-analytics-group-2	Service principal	Synapse Administrator	Workspace

Add role assignment

Grant others access to this workspace by assigning roles to users, groups, and/or service principals.
[Learn more](#)

Scope * ⓘ

☒ Workspace ☐ Workspace item




Role * ⓘ

Synapse Administrator

Select user * ⓘ

 Search by name or email address

Selected user(s), group(s), or service principal(s)

	Sergio Simón Fernández sergio.simon@tajamar365.com	Remove
	Miguel Marañón Quero miguel.maranon@tajamar365.com	Remove
	Daniel Serrano Real daniel.serrano@tajamar365.com	Remove
	Roberto García Moreno roberto.garciamoreno@tajamar365.com	Remove

Apply

Cancel

KPI's

708,69 mil

Ingresos Totales

22,15 mil

Promedio Venta Pedido

Classic Vest, S

Producto Más Vendido

32

Cantidad de Pedidos

32

Clientes Activos

7,00

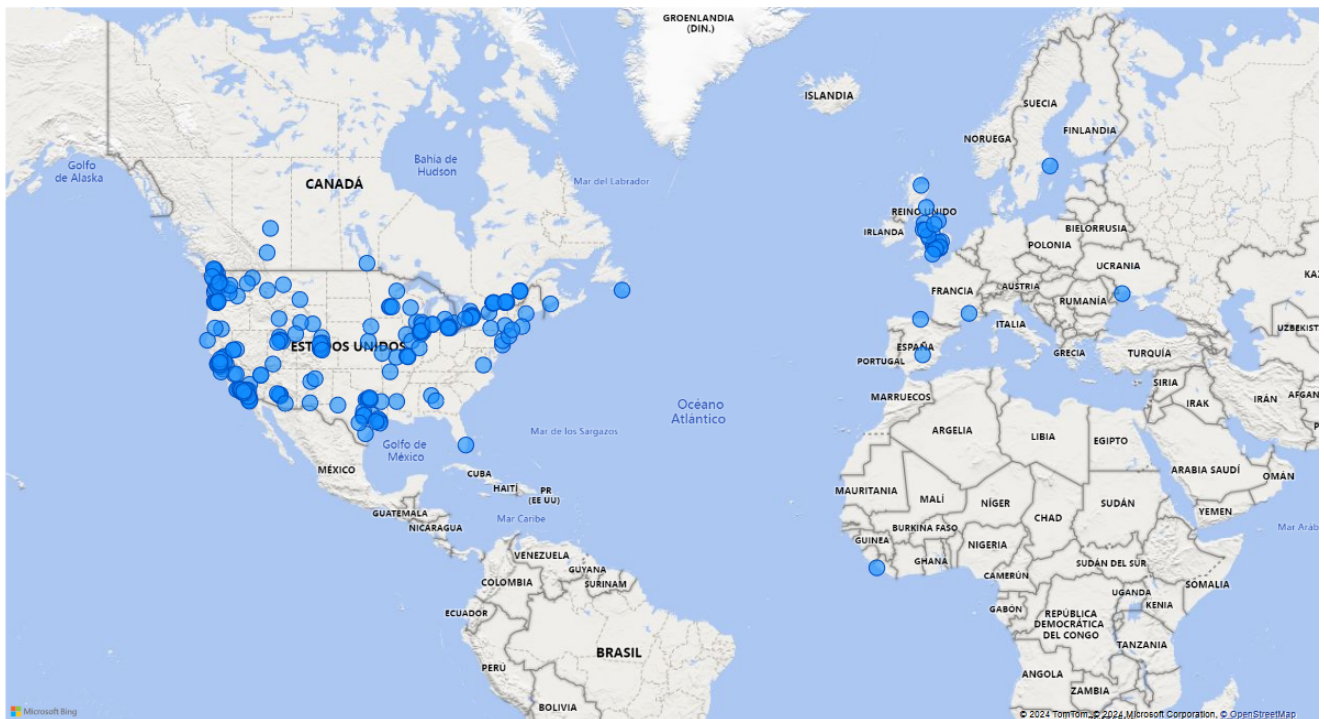
Tiempo Promedio Entrega

0,02

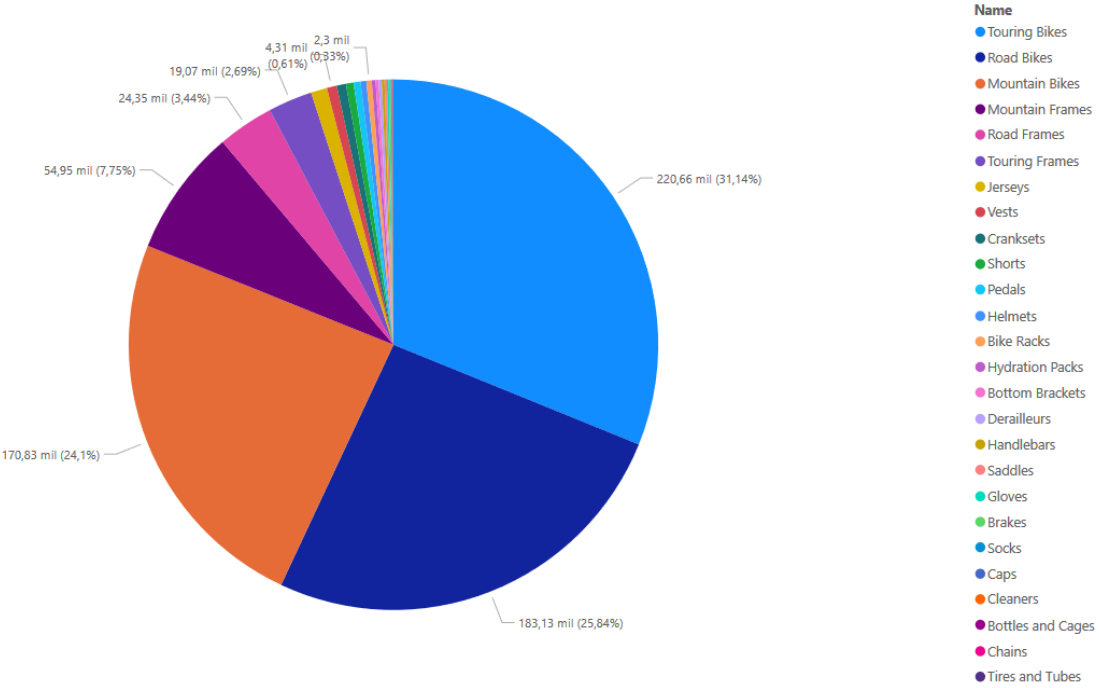
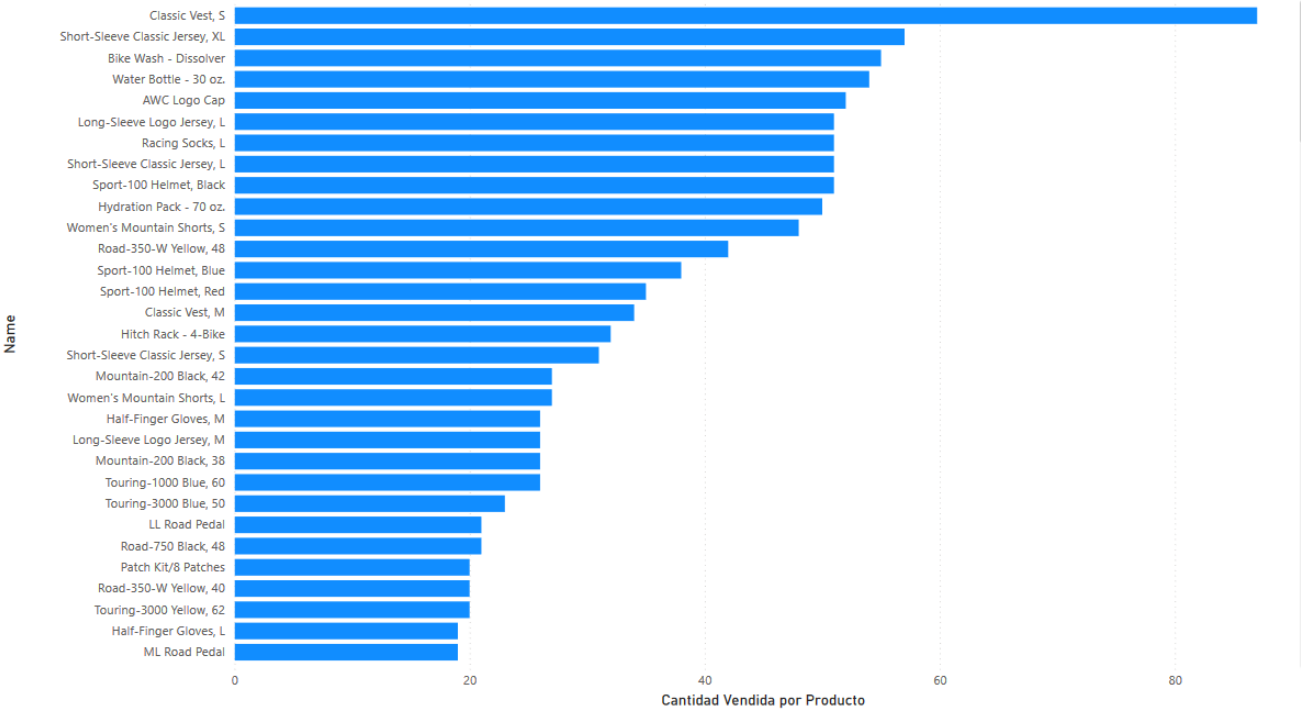
Tasa Descuento Promedio

22,15 mil

Ingresos promedio generados por cada cliente.



Cantidad Vendida por Producto por Name



Productos Vendidos por Categoría por Name

