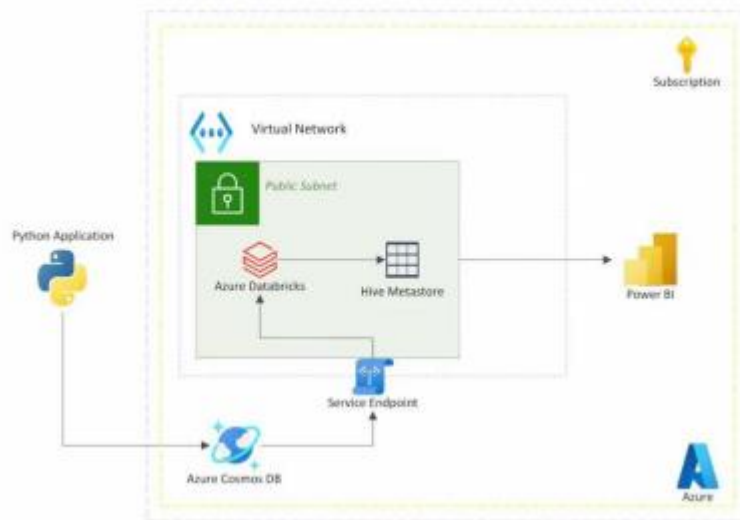


2. (3 puntos) Basado en la arquitectura planteada en la pregunta 1:



Mediante un estudio de presupuesto del Chief Data Officer, se te recomienda modificar la arquitectura de arriba sin utilizar Databricks ni Hive Metastore, sustituir los servicios anteriores por servicios nativos propios de Azure. Desplegar la nueva arquitectura con tu propuesta, deben verse al igual que en la pregunta 1 los plots en Power BI. Realizar la entrega en un documento pdf (o enlace a tu repo de GitHub) donde se vean las capturas de cada paso.

Flujo para Conexión y Análisis de Datos con Azure Synapse Analytics

1. Ingesta de Datos (Azure Cosmos DB)

- Configura tu base de datos en Azure Cosmos DB con la API de MongoDB.
- Activa **Azure Synapse Link** en el portal de Azure para habilitar un contenedor analítico asociado a tus datos transaccionales.

2. Conexión desde Azure Synapse Analytics

- Configura un Workspace de Azure Synapse Analytics en tu suscripción.
- Conecta el Workspace de Synapse con la cuenta de Cosmos DB habilitada con Synapse Link.
- Los contenedores analíticos estarán disponibles como tablas virtuales para consulta.

3. Consulta y Transformación de Datos (Synapse SQL)

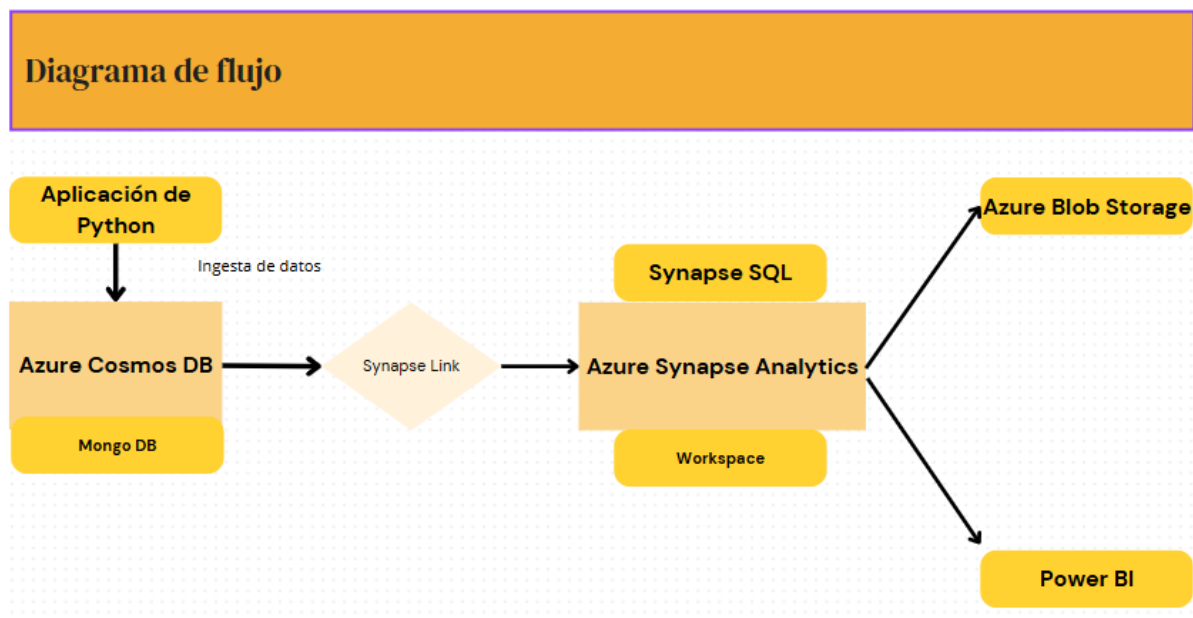
- Usar **SQL Serverless** en Synapse para escribir consultas SQL sobre los contenedores analíticos.
- Realiza transformaciones de datos, agregaciones y combinaciones con otros orígenes si es necesario.

4. Visualización y Análisis

- Publica los resultados en Power BI para crear reportes interactivos.

- b. Opcionalmente, podemos exportar los datos transformados a un almacenamiento de datos (Data Lake o Blob Storage).

Diagrama del flujo:



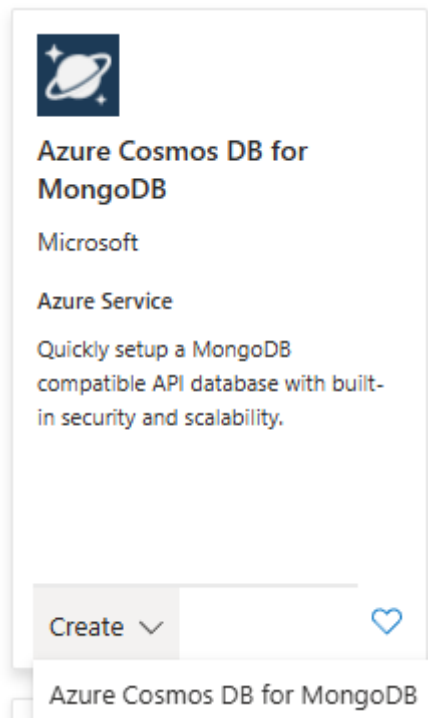
Guía paso a paso para Conectar Azure Cosmos DB con Synapse Analytics

1. Crear un Grupo de Recursos

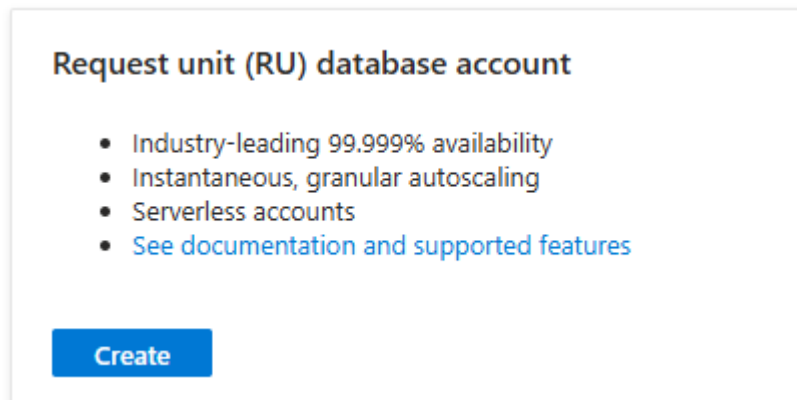
- a. Creamos un grupo de recursos llamado myResourceGroup.
- b. Región: France Central (EU).

2. Crear una cuenta de Azure Cosmos DB con MongoDB

- a. En el portal de Azure, selecciona **Create a resource** y busca **Azure Cosmos DB for MongoDB**.



- b. Selecciona **Azure Cosmos DB for MongoDB API** y haz clic en **Create**.
c. Selecciona **Request unit (RU) database account**.



- d. Configura la cuenta de la siguiente manera:

i. En **Basics**.

Instance Details

Account Name *

cosmosdbwithmongodb

Configure availability zone settings for your account. You cannot change these settings once

Availability Zones ⓘ

☐ Enable ☒ Disable

Location * ⓘ

(Europe) France Central

Available locations are determined by your subscription. If you cannot select your desired location, please [Click here for more details on how to create a new subscription](#).

Capacity mode ⓘ

☐ Provisioned throughput ☒ Serverless

[Learn more about capacity mode](#)

Version

7.0

- ii. En **Global Distribution**: selecciona en ambas: **Disable**.
- iii. En **Networking** selecciona: **All networks**.
- iv. En **Backup Policy** selecciona **Periodic** y **Locally-redundant backup storage**.

Backup policy ⓘ

- ☒ Periodic
Backup is taken at periodic interval based on your configuration
- ☐ Continuous (7 days)
Provides backup window of 7 days / 168 hours and you can restore to any point in time. This option is available for free.
- ☐ Continuous (30 days)
Provides backup window of 30 days / 720 hours and you can restore to any point in time. This option has cost impact.

Backup interval ⓘ

240

60-1440

Minute(s)

Backup retention ⓘ

8

8-720

Hours(s)

Copies of data retained

2

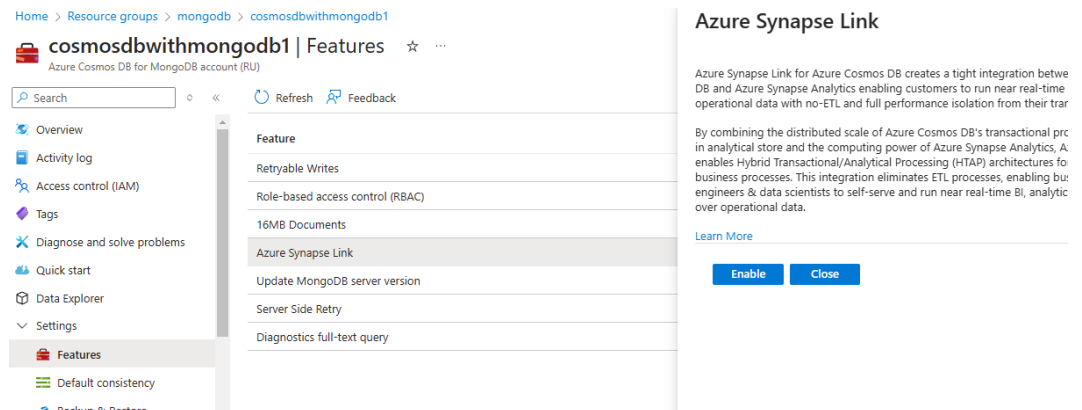
Backup storage redundancy *

- ☐ Geo-redundant backup storage
- ☐ Zone-redundant backup storage
- ☒ Locally-redundant backup storage

- v. En **Encryption** selecciona **Service-managed key**.
- vi. Haz clic en **Review + Create** y luego en **Create**.

3. Crear una base de datos y una colección en CosmosDB

- a. Ve a tu recurso de Cosmos DB en el portal.
- b. En el menú lateral izquierdo, selecciona **Settings -> Features**.
- i. Habilita Azure Synapse Link pulsando **Enable**.



- ii. Esto permitirá sincronizar los datos en contenedores analíticos optimizados para consulta.
- c. En el menú lateral izquierdo, selecciona **Data Explorer**.
 - i. Haz clic en **New Database** y ponle un nombre: ej.(**testingcosmosdb01**).
 - ii. Haz clic en **OK**.
- d. A continuación, haz clic en **New Collection** y asigna un nombre a tu colección: ej.(**cosmosdbcollection**).
 - i. En **Database name** selecciona **Use existing** y selecciona la que se ha creado anteriormente.
 - ii. Selecciona para **Sharding: Unsharded (20GB limit)**.
 - iii. Selecciona **On** para **Analytical Store**.
 - iv. Haz clic en **OK** para crear la colección.

The screenshot shows the 'New Collection' form in the Azure Cosmos DB portal. The form has the following fields and options:

- Database name**: A dropdown menu with 'testingcosmosdb01' selected. Above it are radio buttons for 'Create new' and 'Use existing' (which is selected).
- Collection id**: A text input field containing 'cosmosdbcollection'.
- Sharding**: Radio buttons for 'Unsharded (20GB limit)' (selected) and 'Sharded'.
- Analytical store**: Radio buttons for 'On' (selected) and 'Off'.

4. Crear un Servicio de Azure Synapse Analytics

a. Crea un Workspace de Synapse Analytics

- i. En el portal, selecciona **Crear un recurso** y busca **Azure Synapse Analytics**.
- ii. Haz clic en **Workspace de Synapse Analytics**.



b. Rellena la Información de Configuración

- i. **Grupo de recursos:** Selecciona uno existente.
- ii. **Nombre del Workspace:** Elige un nombre único para tu Workspace. Ej. (**synapseanalyticsforcosmosdb**)
- iii. **Ubicación:** Selecciona la misma región en la que está tu instancia de Azure Cosmos DB (esto mejora el rendimiento).
- iv. **Cuenta de Data Lake Storage Gen2:**
 1. Crea un nuevo Data Lake Storage Gen2 y elige un nombre. Ej. (**datalakestoragecosmosdb**).
 2. Pon nombre a sistema de ficheros, ej. **File system name: filesystem.**
 3. Synapse requiere una cuenta de almacenamiento para trabajar con los datos.

Subscription * ⓘ Azure for Students

Resource group * ⓘ myResourceGroup
[Create new](#)

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 * synapseworkspacemongodb ✓

Region * France Central

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

Account name * ⓘ (New) datalakestoragecosmosdb
[Create new](#)

File system name * (New) filesystem
[Create new](#)

4. En Security selecciona **Use only Microsoft Entra ID authentication.**

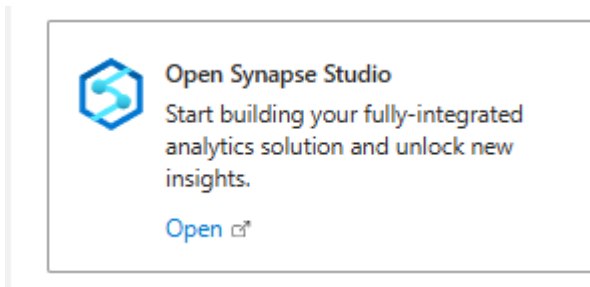
Authentication
Choose the authentication method for access to workspace resources such as SQL pools. The authentication method can be changed later on. [Learn more](#) ⓘ

Authentication method ⓘ ☐ Use both local and Microsoft Entra ID authentication
☒ Use only Microsoft Entra ID authentication
☒ Local authentication will be disabled for resources inside the workspace such as SQL pools.

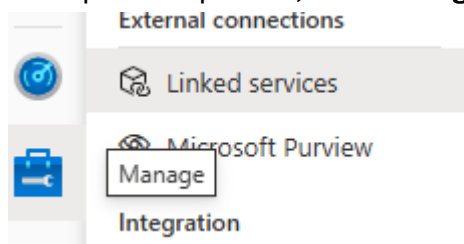
- v. Haz clic en **Revisar y crear**. Una vez que la validación sea exitosa, haz clic en **Crear**.

5. Configura Azure Synapse Workspace

- a. Accede a tu **Synapse Workspace** desde el portal de Azure.
- b. Abre **Synapse Studio** (desde la pestaña **Open Synapse Studio** en tu recurso).



- c. En el panel izquierdo, ve a **Manage > Linked Services**:



- d. Haz clic en **+ New** y selecciona **Azure Cosmos DB for MongoDB**.
- e. Completa los campos necesarios:
 - i. **Connection name:** **CosmosDbMongoDb**
 - ii. Connect via: **AutoResolveIntegrationRuntime**
 - iii. Account selection method: **From Azure subscription**
 - 1. **Azure Subscription:** Selecciona tu subscripcion de Azure.
 - 2. **Azure Cosmos DB account name:** Selecciona tu recurso de Cosmos DB.
 - 3. Database name: **testingcosmosdb01**
 - iv. **Key:** Usa la clave de tu Cosmos DB para autenticar.

- f. Guarda los cambios.

Name *

CosmosDbMongoDb

Description

Connect via integration runtime * ⓘ

✓ AutoResolveIntegrationRuntime

Connection string Azure Key Vault

Account selection method ⓘ

☒ From Azure subscription ☐ Enter manually

Azure subscription ⓘ

Azure for Students (3ebaad0a-82a1-4238-b694-99ee24cefa37)

Azure Cosmos DB account name * ⓘ

cosmosdbdwithmongodb

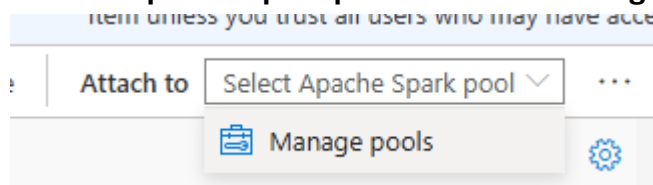
Database name *

testingcosmosdb01

Activar Windows

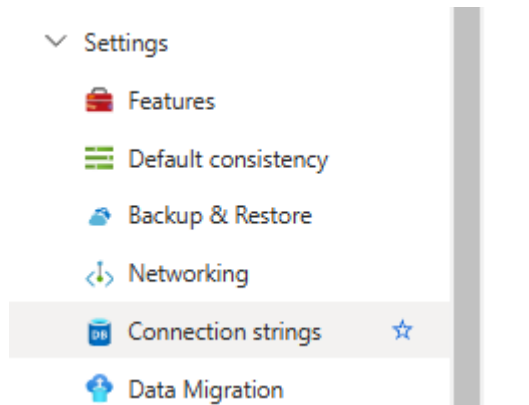
6. Ingesta de datos desde Synapse Studio

- Accede a tu **Synapse Workspace** desde el portal de Azure.
- Abre un nuevo notebook en el que ejecutaremos el siguiente [script](#).
- Crea un **Apache Spark pool**. Accede a **Manage pools** para ello.



- Selecciona **New Apache Spark pool**
 - Configura el Apache Spark pool
 - Name: apachesparkpool**
 - Number of nodes: 3**
- Antes de ejecutar el [script](#) realiza unos cambios en él.
 - Accede a tu cuenta de Azure Cosmos DB.

- ii. En el menú izquierdo selecciona **Settings -> Connection strings**



- iii. Copia la PRIMARY CONNECTION STRING.

PRIMARY CONNECTION STRING
`mongodb://cosmosdbdwithmongodb:KjgISZBY1m50Dn0vcqh2UPkfz7UtgMKn1T5EWZY8InvV6WtyfRp7q2Txgq4VrQW8EqHhGEFUn1eACDbFVys8w==@cosmosdbdwithmongodb.mongo.co...`

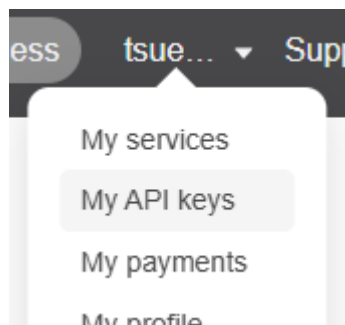
- iv. Cambia la CONNECTION STRING por la de tu MongoDB copiada anteriormente.

- v. Cambia también el nombre de la colección y de la base de datos.

```
# Configuración de Cosmos DB
DB_NAME = "testingcosmosdb01" # Nombre de tu base de datos
COLLECTION_NAME = "cosmosdbcollection" # Nombre de tu colección
CONNECTION = "mongodb://cosmosdbdwithmongodb:KjgISZBY1m50Dn0vcqh2UPkfz7UtgMKn1T5EWZY8InvV6WtyfRp7q2Txgq4VrQW8EqHhGEFUn1eACDbFVys8w==@cosmosdbdwithmongodb.mongo.co..."
```

- vi. Cambia la API KEY por la tuya de la API [Open Weather](#).

La encontrarás en tu perfil -> My API Key




- vii. Regístrate si no tienes cuenta.
- viii. Descarga la dependencia de pymongo que se necesitará en el script.

```
1 # Descargar las dependencias necesarias
2 !pip install pymongo
```

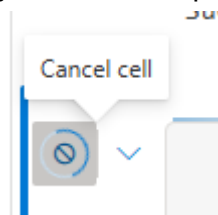
[5] ✓ 13 sec - Command executed in 13 sec 51 ms by tsuenkit.lui on 3:58:31 AM, 12/07/24

- ix. Por último, ejecuta el script. Podrás cambiar le nombre de la ciudad por la ciudad deseada.

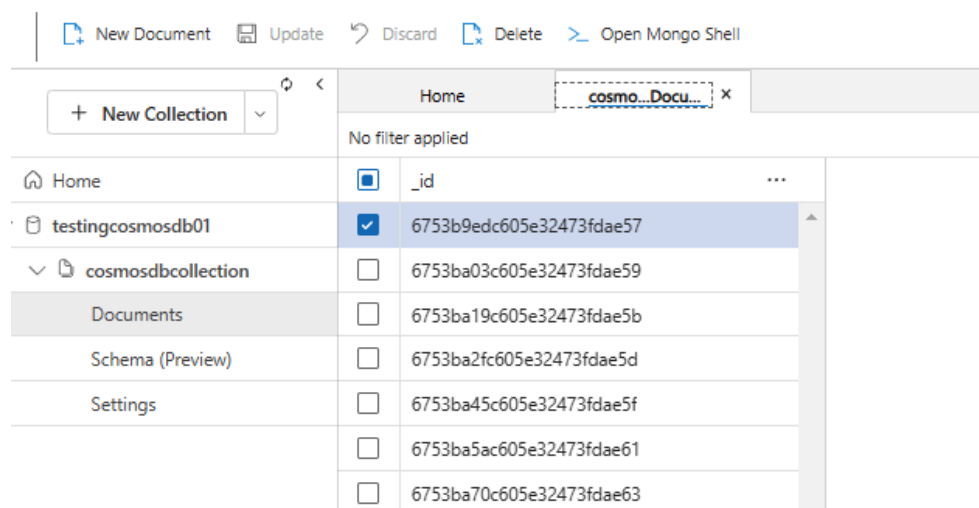


```
1 from pymongo import MongoClient
2 import requests
3 import json
4 import time
5
6 # Configuración de OpenWeather API
7 API_KEY = "2fb0077b4ae9ae01842d4d5b0ea387d6"
8 CITY = "Toronto" # Cambia por la ciudad deseada
9 WEATHER_URL = f"http://api.openweathermap.org/data/2.5/weather?q={CITY}&appid={API_KEY}"
10
11 # Configuración de Cosmos DB
12 DB_NAME = "testingcosmosdb01" # Nombre de la base de datos
13 COLLECTION_NAME = "cosmosdbcollection" # Nombre de la colección
14 CONNECTION = "mongodb://cosmosdbwithmongodb:password@localhost:27020"
15
```

- x. Vemos como recoge datos cada 20 segundos. Podemos para la ejecución siempre que deseemos, haz click en Cancel cell.



- xi. Podemos comprobar en nuestra base de datos de CosmosDB que recibe datos.



The screenshot shows the MongoDB Compass interface. At the top, there are buttons for "New Document", "Update", "Discard", "Delete", and "Open Mongo Shell". Below these, there is a "New Collection" button. The main area shows a list of documents in the "cosmosdbcollection" collection. The first document is selected, and its details are shown in the right pane.

Document	_id
6753b9edc605e32473fdae57	6753b9edc605e32473fdae57
6753ba03c605e32473fdae59	6753ba03c605e32473fdae59
6753ba19c605e32473fdae5b	6753ba19c605e32473fdae5b
6753ba2fc605e32473fdae5d	6753ba2fc605e32473fdae5d
6753ba45c605e32473fdae5f	6753ba45c605e32473fdae5f
6753ba5ac605e32473fdae61	6753ba5ac605e32473fdae61
6753ba70c605e32473fdae63	6753ba70c605e32473fdae63

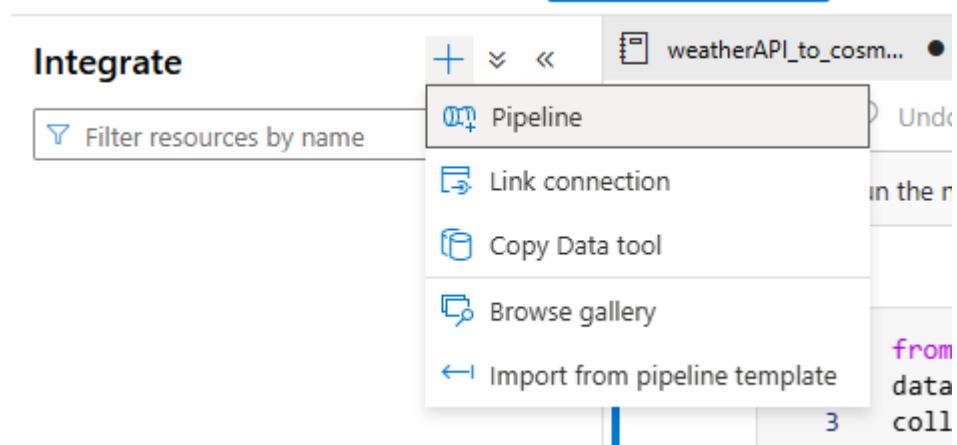
7. Consulta datos en Synapse Studio usando Pipeline

a. Configura un Linked Service a Cosmos DB:

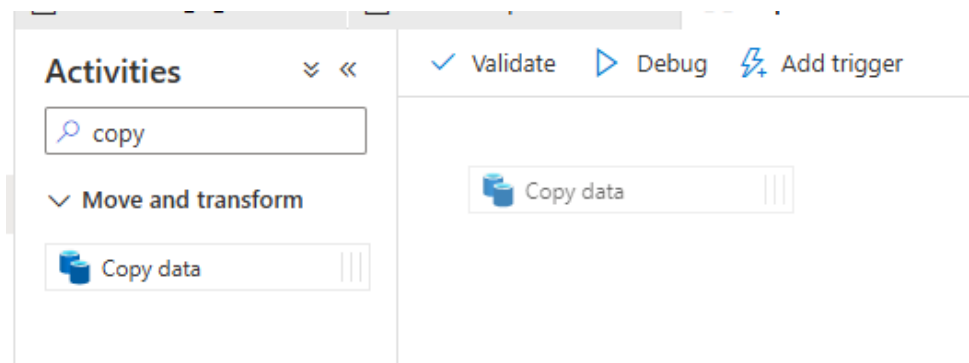
- i. En el panel izquierdo de Synapse Studio, haz clic en **Manage > Linked Services**.
- ii. Crea un nuevo Linked Service para **Cosmos DB (MongoDB API)** y proporciona la cadena de conexión.

b. Crea un Pipeline:

- i. En el panel izquierdo de Synapse Studio, haz clic en **Integrate > Pipeline**.



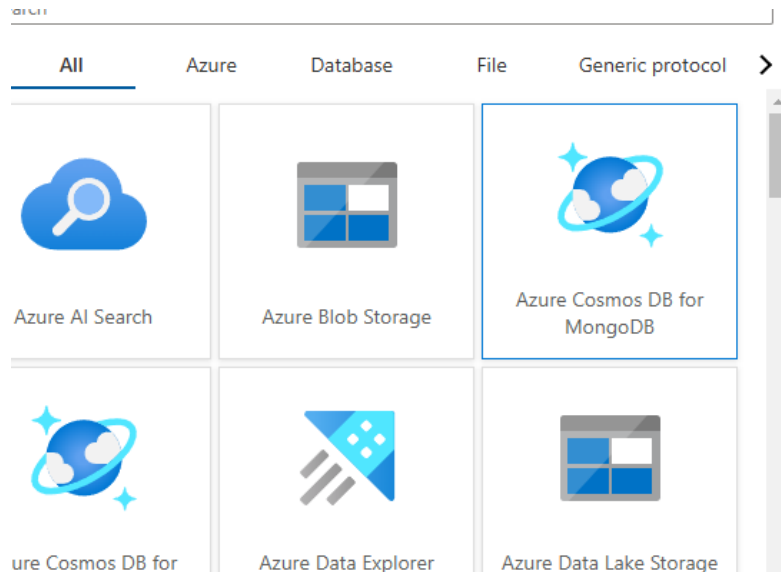
- ii. Dentro de la nueva canalización, arrastra y suelta la actividad **Copy Data** desde la paleta de actividades a la superficie de diseño.



- iii. Configura la actividad de copia:

1. **Source (Origen):** Selecciona la conexión a CosmosDB que configuraste anteriormente. Elige la base de datos y la colección que deseas copiar.
 - a. Haz clic en New

b. Selecciona Azure Cosmos DB for MongoDB



c. Pon un nombre

d. Selecciona el **Linked Service** creado

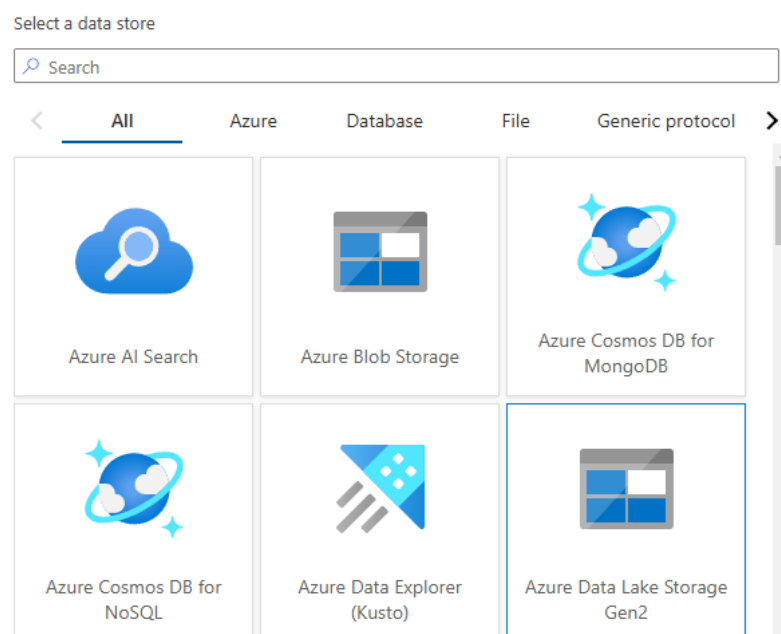
e. Selecciona la colección que queremos copiar

f. Haz clic en OK

2. **Sink (Destino):** Configura el destino donde se copiarán los datos, en un **Data Lake Storage**.

a. Haz clic en New

b. Selecciona Azure Data Lake Storage Gen 2



c. Selecciona el tipo de formato de tus datos: **JSON**

d. Pon un nombre

e. Selecciona el **Linked Service** creado

Set properties

Name

Json1

Linked service *

synapseworkspacemongodb-WorkspaceDefaultStorage

Connect via integration runtime * ⓘ

✓ AutoResolveIntegrationRuntime

File path

File system

/ Directory

/ File name

Import schema

☐ From connection/store ☐ From sample file ☒ None

> Advanced

f. Si no, crea otro Data Lake Storage

New linked service

Azure Data Lake Storage Gen2 [Learn more](#)

Choose a name for your linked service. This name cannot be updated later.

Name *

AzureDataLakeStorage1

Description

Connect via integration runtime * ⓘ

✓ AutoResolveIntegrationRuntime

Authentication type

Account key

Account selection method ⓘ

☒ From Azure subscription ☐ Enter manually

Azure subscription ⓘ

Azure for Students (3ebaad0a-82a1-4238-b694-99ee24cefa37)

Storage account name *

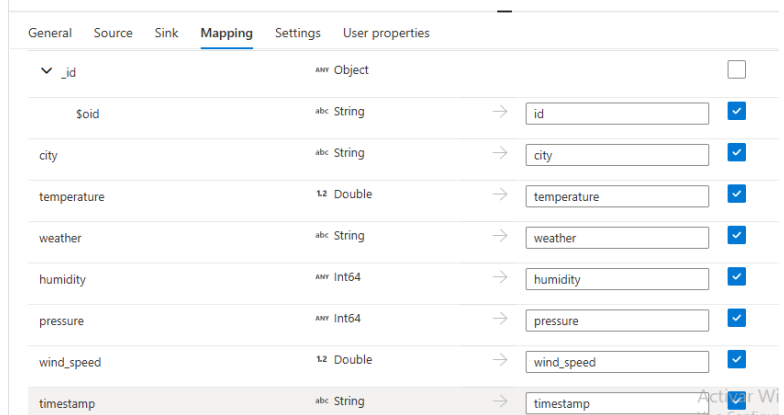
datalakestoragecosmosdb

Activar Windows

g. Haz clic en OK

3. Configura el **esquema de mapeo**

- a. En la pestaña **Mapping**, pon el nombre de las columnas del Data Lake Storage Gen 2



General	Source	Sink	Mapping	Settings	User properties
▼	_id	ANY Object			<input type="checkbox"/>
	\$_id	abc String	→	id	<input checked="" type="checkbox"/>
	city	abc String	→	city	<input checked="" type="checkbox"/>
	temperature	12 Double	→	temperature	<input checked="" type="checkbox"/>
	weather	abc String	→	weather	<input checked="" type="checkbox"/>
	humidity	ANY Int64	→	humidity	<input checked="" type="checkbox"/>
	pressure	ANY Int64	→	pressure	<input checked="" type="checkbox"/>
	wind_speed	12 Double	→	wind_speed	<input checked="" type="checkbox"/>
	timestamp	abc String	→	timestamp	<input checked="" type="checkbox"/>

- iv. Publica y ejecuta la canalización

1. Haz clic en **Publish All** para guardar y aplicar la canalización, haz clic en **Publish**.

Publish all

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

Pending changes (4)

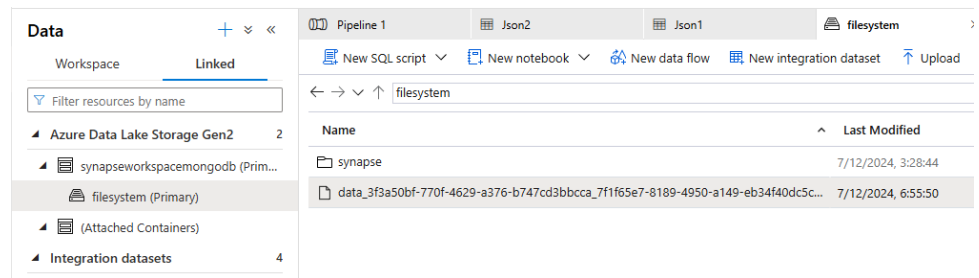
NAME	CHANGE	EXISTING
▼ Pipelines		
Pipeline 1	(New)	-
▼ Datasets		
CosmosDbMongoDbColle...	(New)	-
Json1	(New)	-
CosmosDbMongoDbColle...	(New)	-

2. Ejecuta la canalización haciendo clic en el botón **Add trigger** y seleccionando **Trigger Now**

c. Visualiza datos desde Azure Synapse Studio

- En Synapse Studio, en el menú de la izquierda, selecciona **Data**.
- Luego, haz clic en **Linked** para ver las conexiones de almacenamiento.
- Encuentra tu **Data Lake Storage** vinculado, expándelo y selecciona el contenedor donde deberían estar los datos.

iv. Verás los archivos o carpetas que hayas cargado.



Visualización con Power BI

1. Conectar Power BI a Synapse Analytics

- Abre Power BI Desktop.
- Selecciona **Obtener datos > Azure Datalake Storage Gen 2**.
- Proporciona la dirección URL de tu Azure Data Lake Storage Gen 2.
- En **Azure Datalake Storage Gen 2**, haz clic derecho en tu carpeta y selecciona **Properties** y copia la URL:

URL

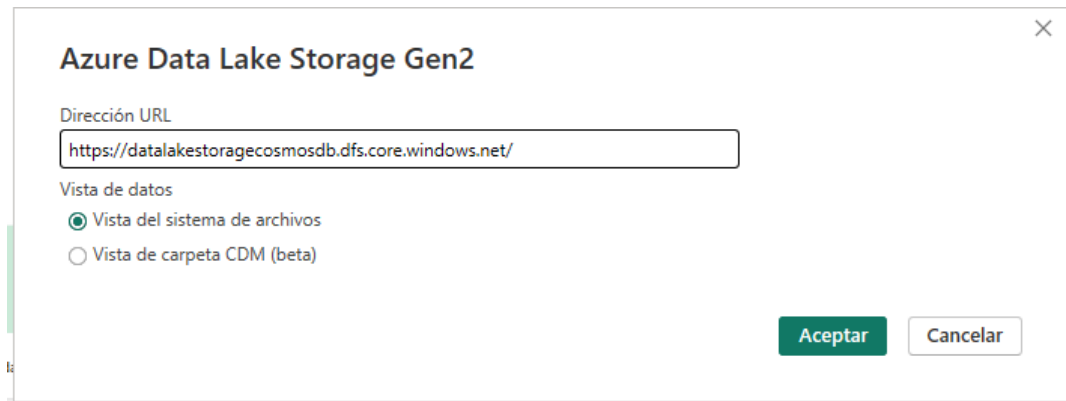
<https://datalakestorage...>

Directorio:

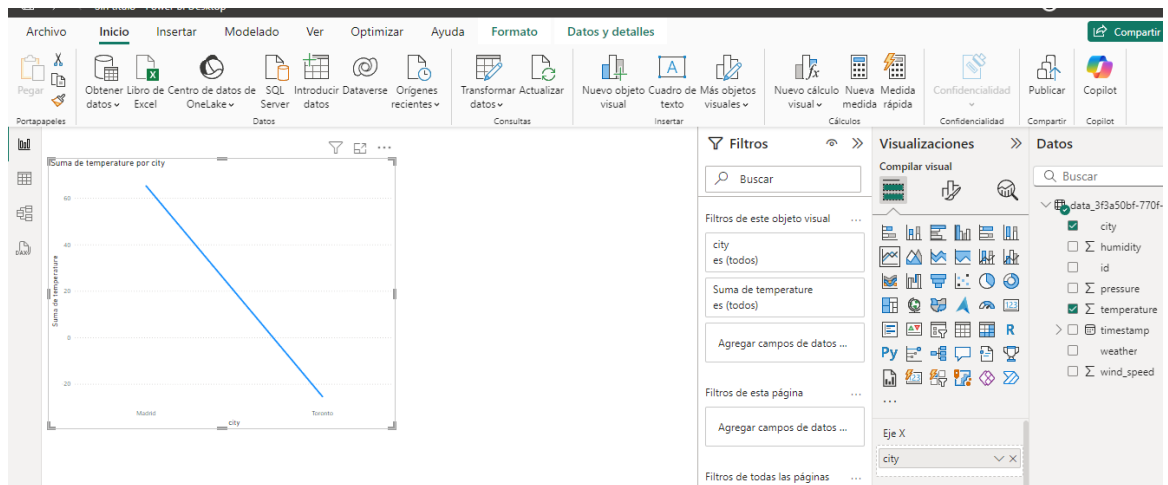
<https://datalakestoragecosmosdb.dfs.core.windows.net/filesystem/synapse/workspaces/synapseworkspacemongodb/warehouse/>

Archivo:

https://datalakestoragecosmosdb.dfs.core.windows.net/filesystem/synapse/workspaces/synapseworkspacemongodb/warehouse/data_0d697cf6-6f98-4c57-90af-1d59495d9fa7_1e390447-60c0-426b-ae66-b87d189aa937.json



- e. Selecciona Transformar y cambia los el archivo de binario a json.
2. Alternativa: Descargarse el **JSON** de Azure Data Lake Gen 2 y abrirlo con **Power BI**.



3. Crear Dashboards

- a. Una vez conectados los datos, diseña tus visualizaciones y dashboards en Power BI.
- b. Publica el reporte en el servicio de Power BI para compartirlo con otros usuarios.

Flujo utilizado

CosmosDB (MongoDB) -> Azure Synapse Analytics -> Azure Data Lake Gen 2-> Power BI

Cuadro de flujo de datos

