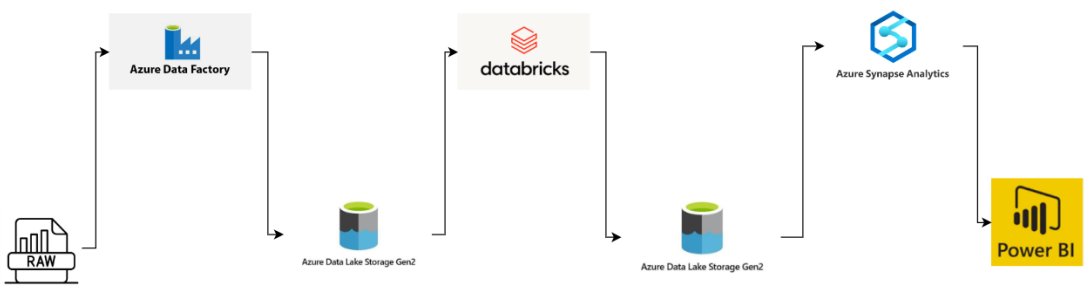


Azure End-to-End Data Pipeline

In this project I have tried to build an End-to End Data Pipeline using Microsoft Azure Services. The resources deployed in this project are

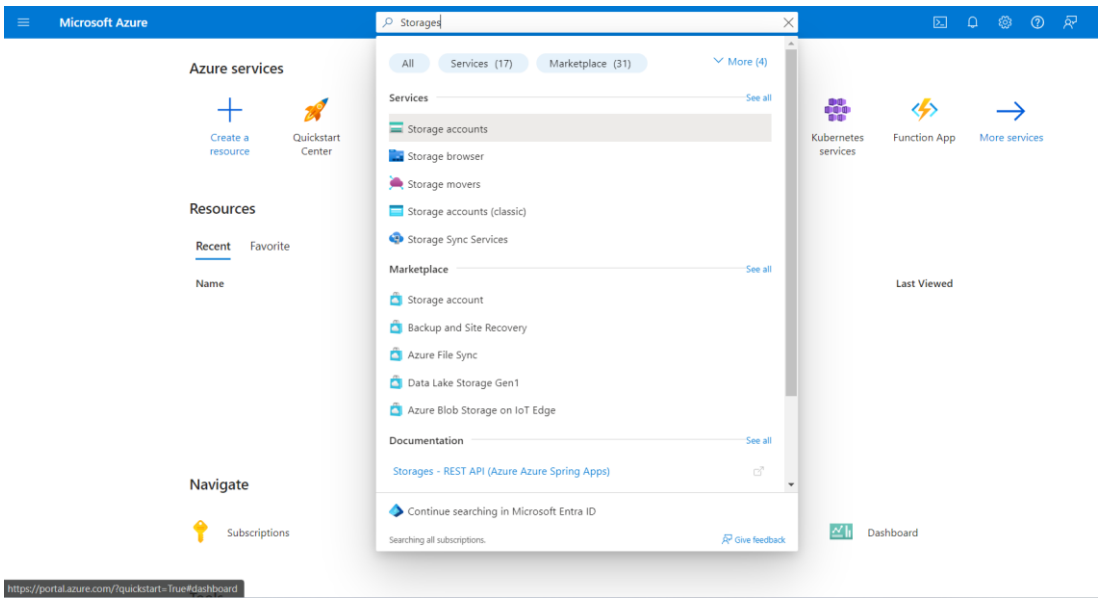
Raw data (Provided on GitHub)
Azure Data Factory: For streamlining the pipeline
Azure Data Lake Gen2:To store the data
Databricks-To perform data transformations using PySpark
Azure Synapse Analytics-For data warehousing and querying
Power BI- For data visualization and dashboarding

The GitHub Link to the Project-https://github.com/SumerPariani/azure_cloud

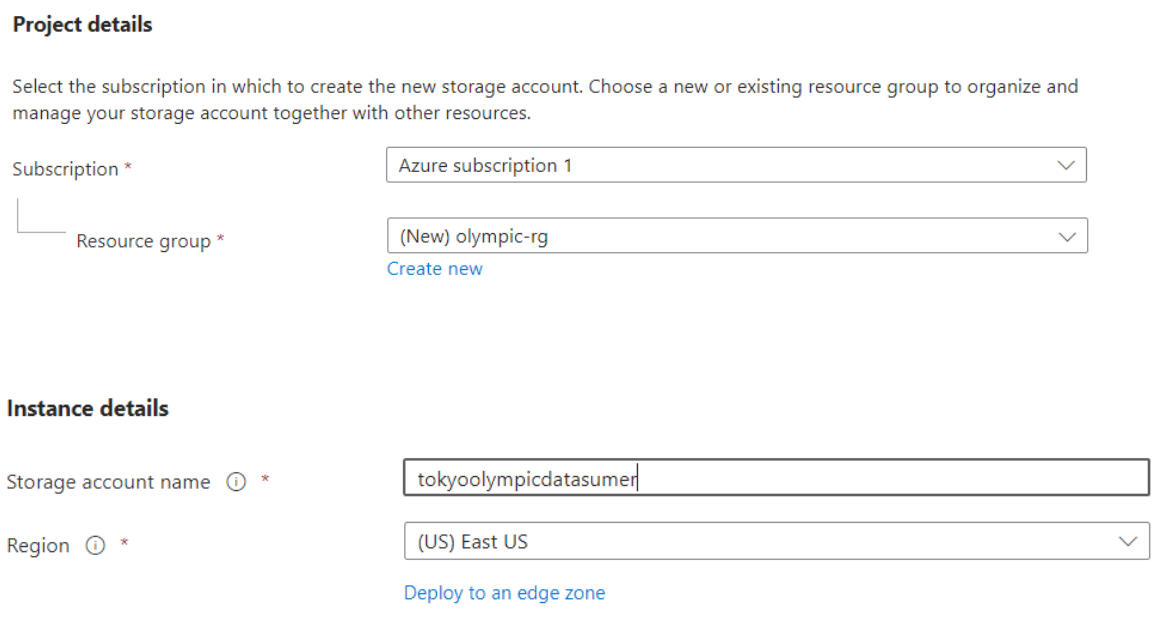
Dataset -
This dataset contains the details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics. This dataset contains the details of the Athletes, Coaches, Medals, Teams participating as well as the Entries by gender. It contains their names, countries represented, discipline, gender of competitors, name of the coaches, medals(gold, silver, bronze).

Let's dive into the project and get started

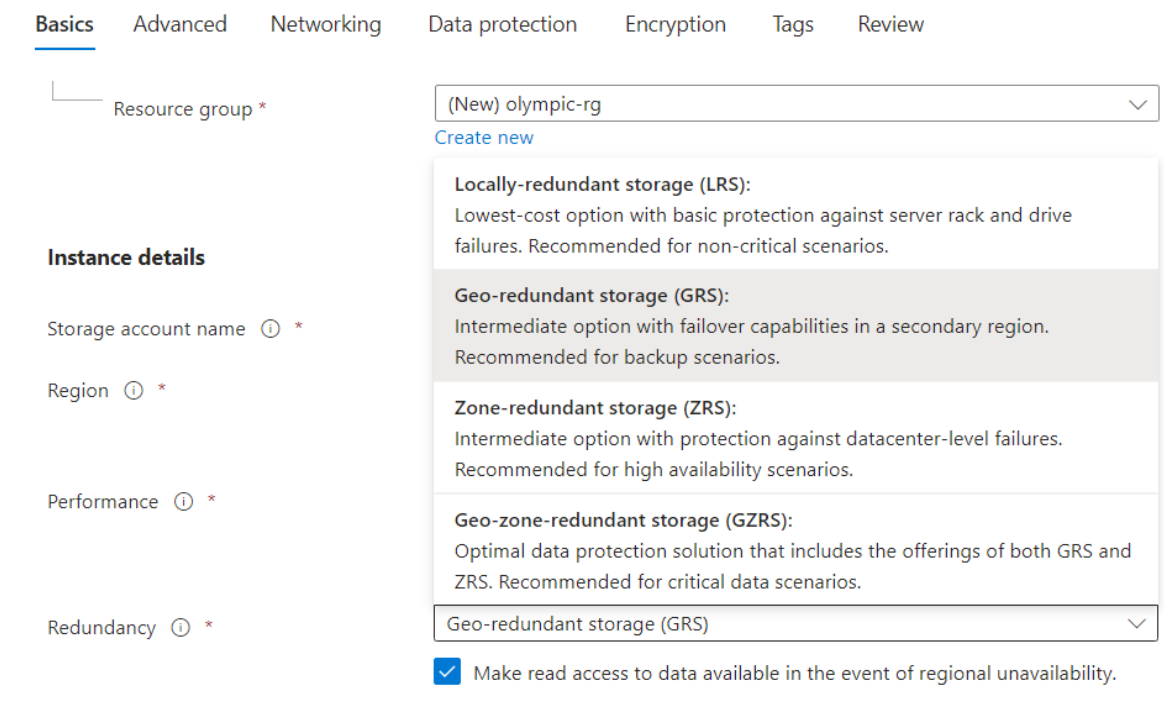
1.Creating Blob storage account
Once you are on your Azure Portal ,you can search for a Storage account on the search bar. Go to the storage account page and create a new one .



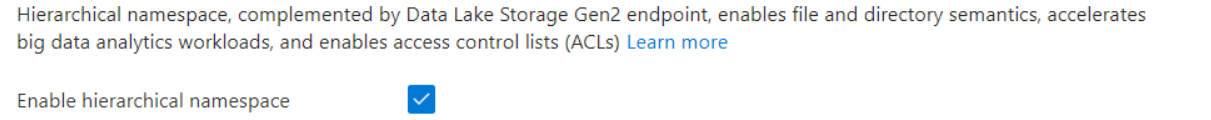
Give a name to your resource group to access all the resources deployed in this project via resource group. Give a name to your storage account and select a region (select the region that is closest to your location)



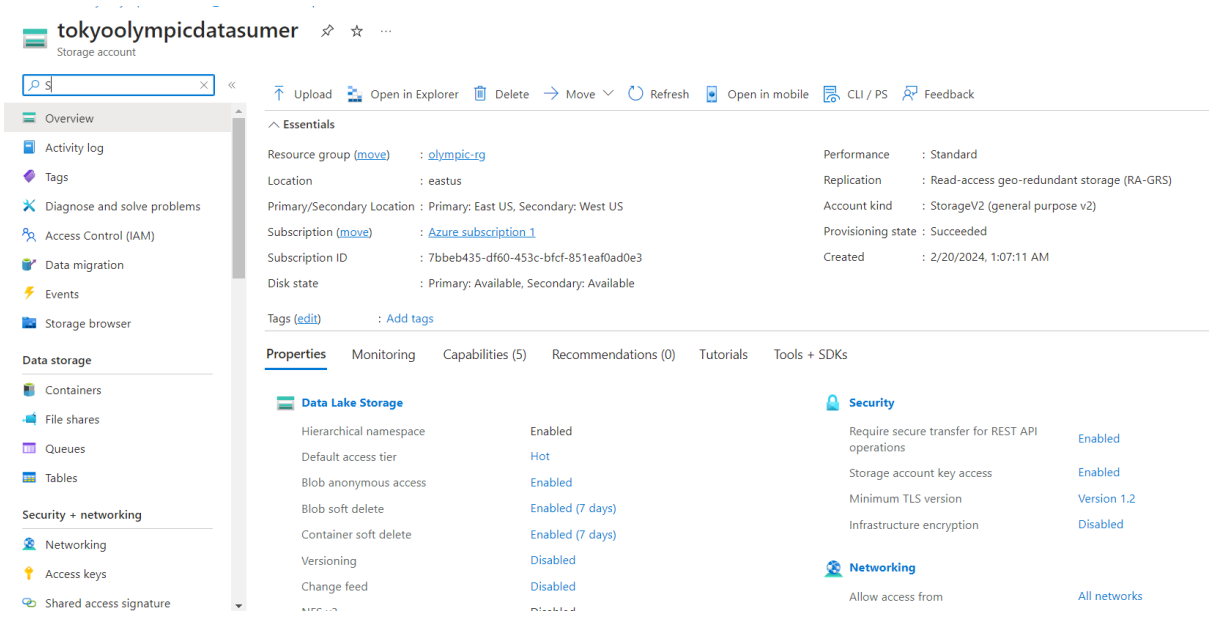
You will also get to choose options about performance and redundancy.I went for Locally-redundant storage because that is the cheapest one and the data is not of much importance to me .



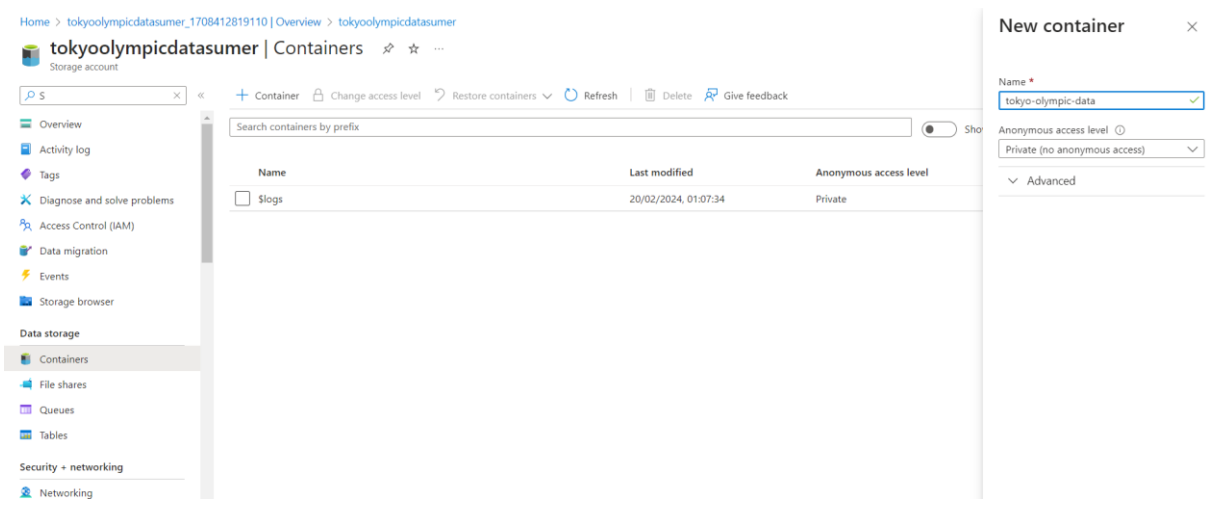
Under the Advanced option -> Hierarchical Namespace select the “Enable hierarchical namespace checkbox”. This will allow you to store your files in hierarchical format and not flat files. This hierarchical format will make storing and querying for files easier.



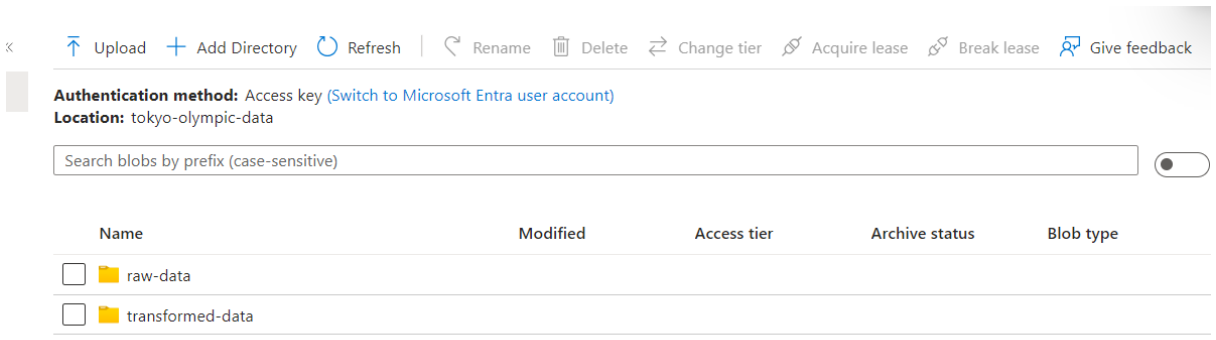
Here you can see that you have created your storage Blob



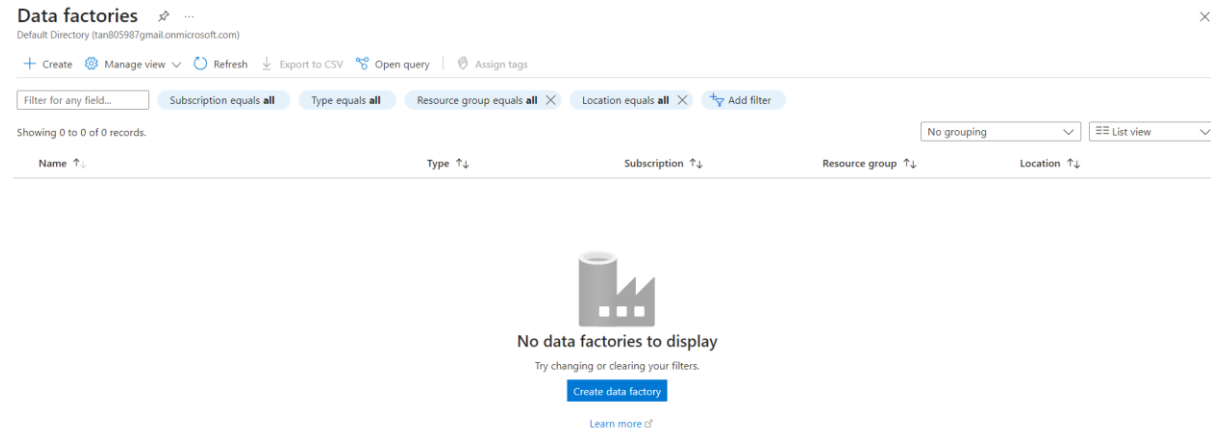
Then create a container by selecting the container option on the left side pane of the storage account. This container will allow you to add directories and add directories to keep the raw-data and transformed data separately.



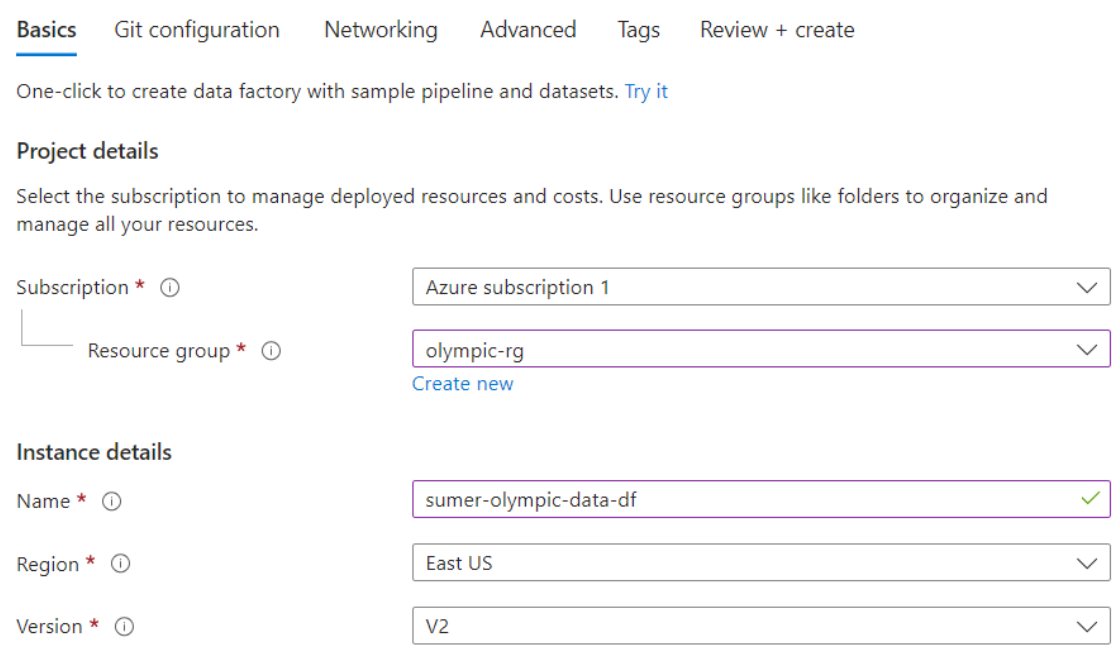
Click on add directories to create two directories
1. raw-data-To store the raw files
2.transformed-data-To store the data files after we perform the Pyspark transformations



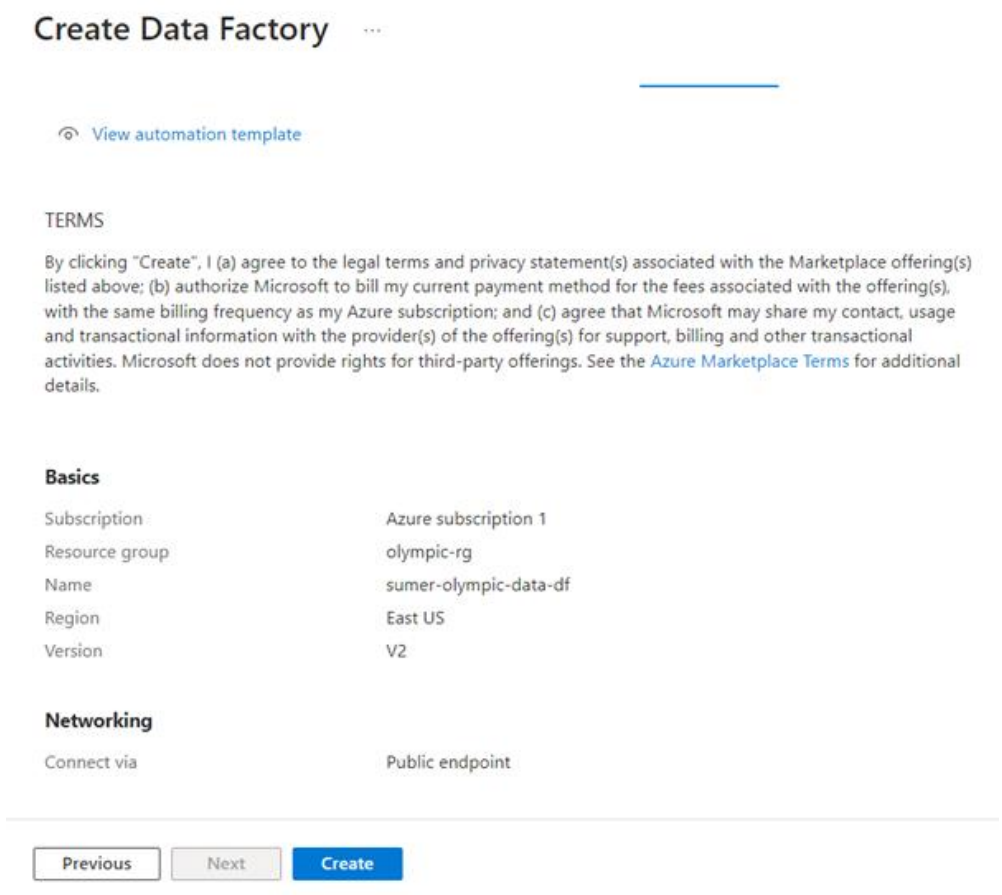
Then let's go on to create our data factory. You can search for Azure Data Factory in the search box and then click on "Create data factory" option to create a new data factory.



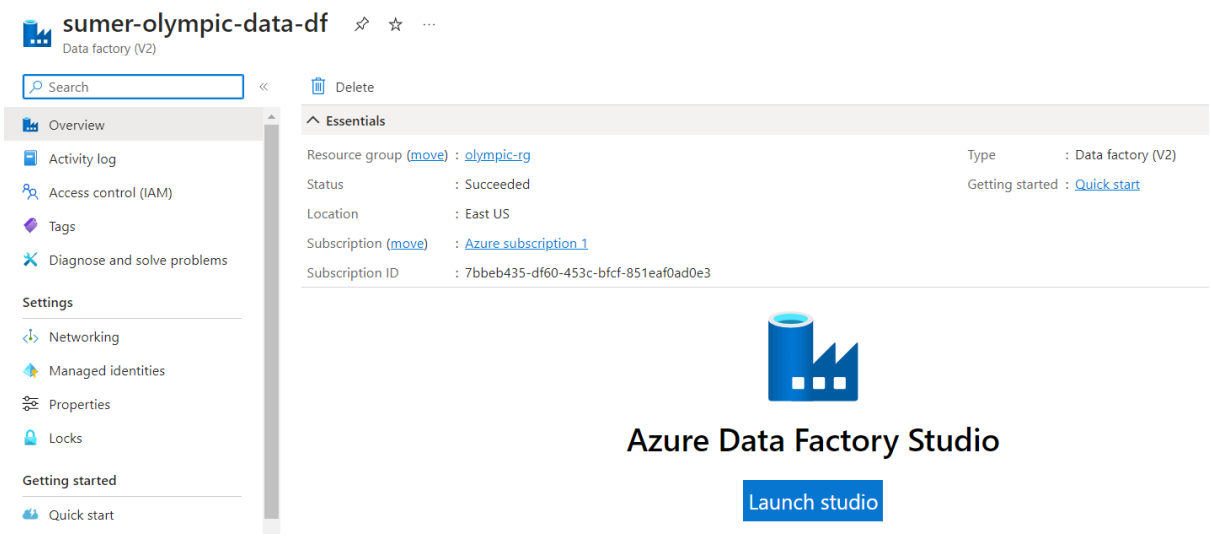
We can keep the same resource group as the storage account as these services are deployed in the same project. Give a name to your azure data factory and select a region.



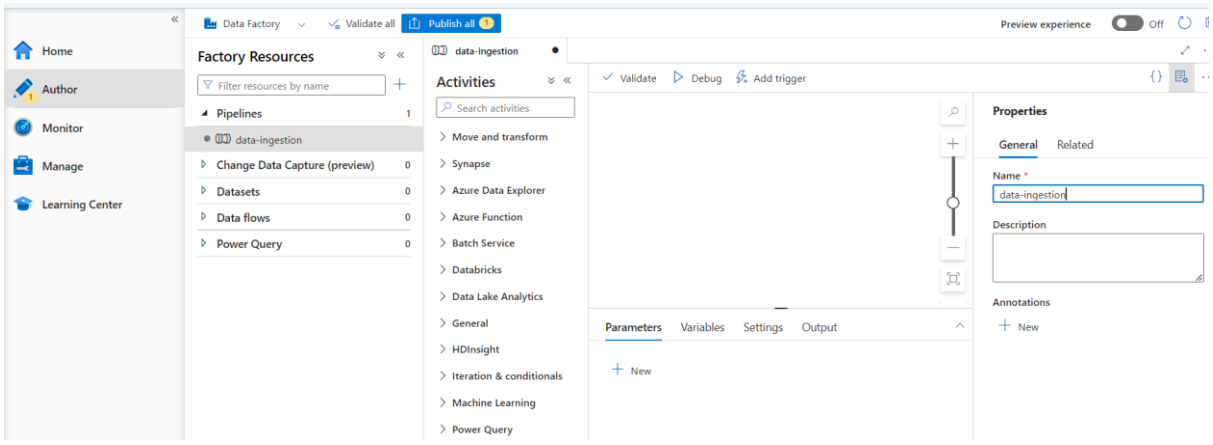
Go-on clicking next-next in the creation dialog box and you will get an option to validate the resource and create it .



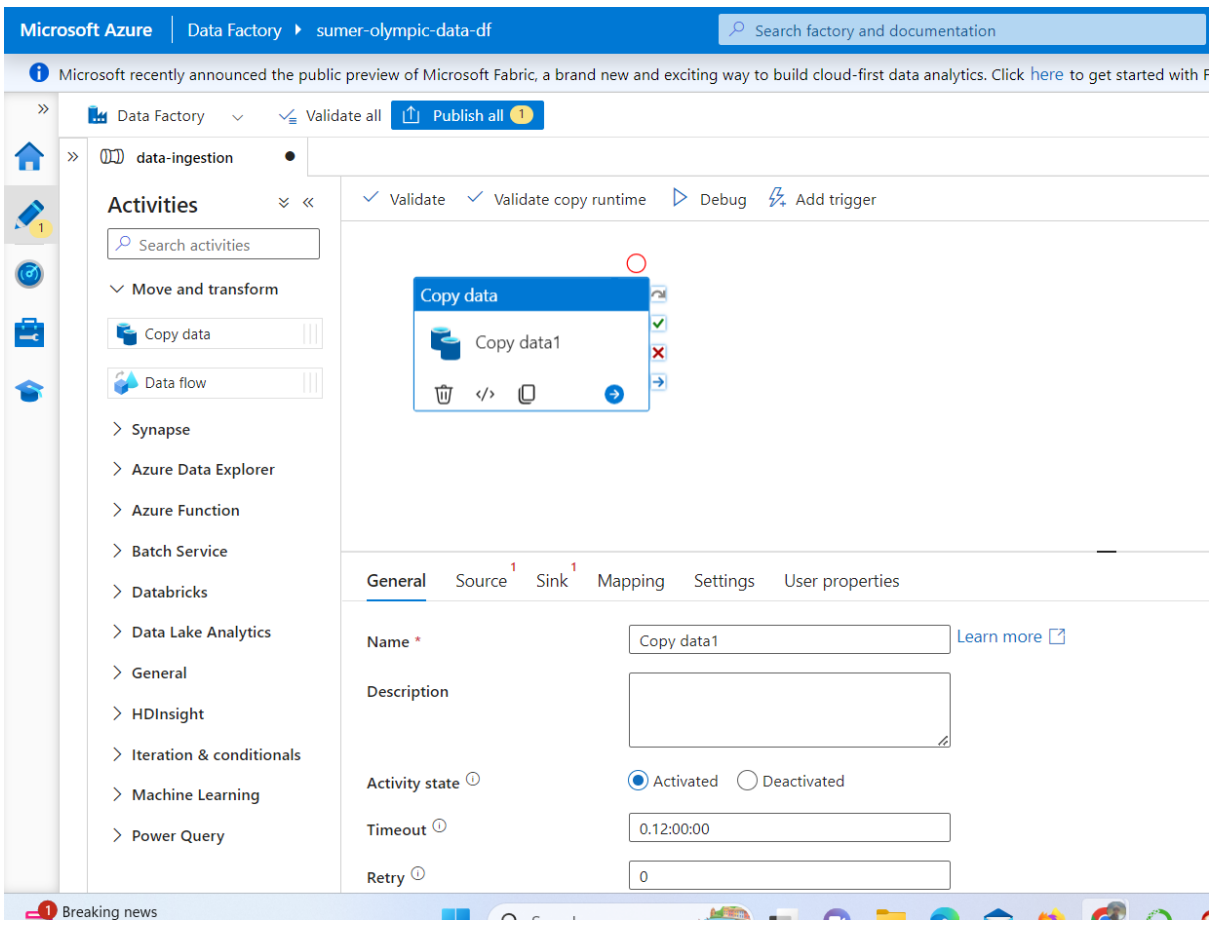
Here you can see that Azure Data Factory is Deployed .



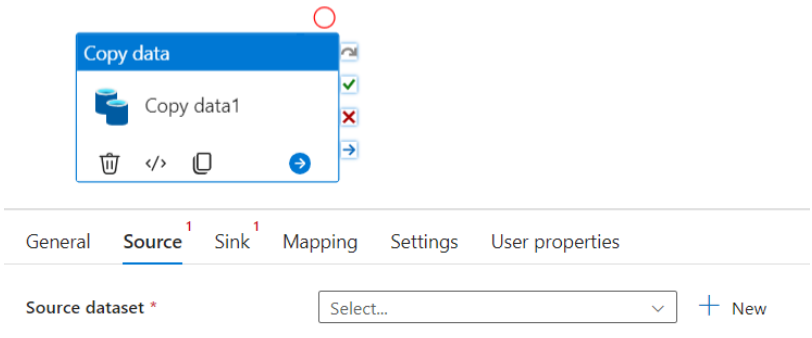
Now we have to create a pipeline to streamline our flow of data .For this create a new pipeline by selecting options Author->Pipelines->Move and transform. Give name to your data pipeline to avoid any kind of confusion later.



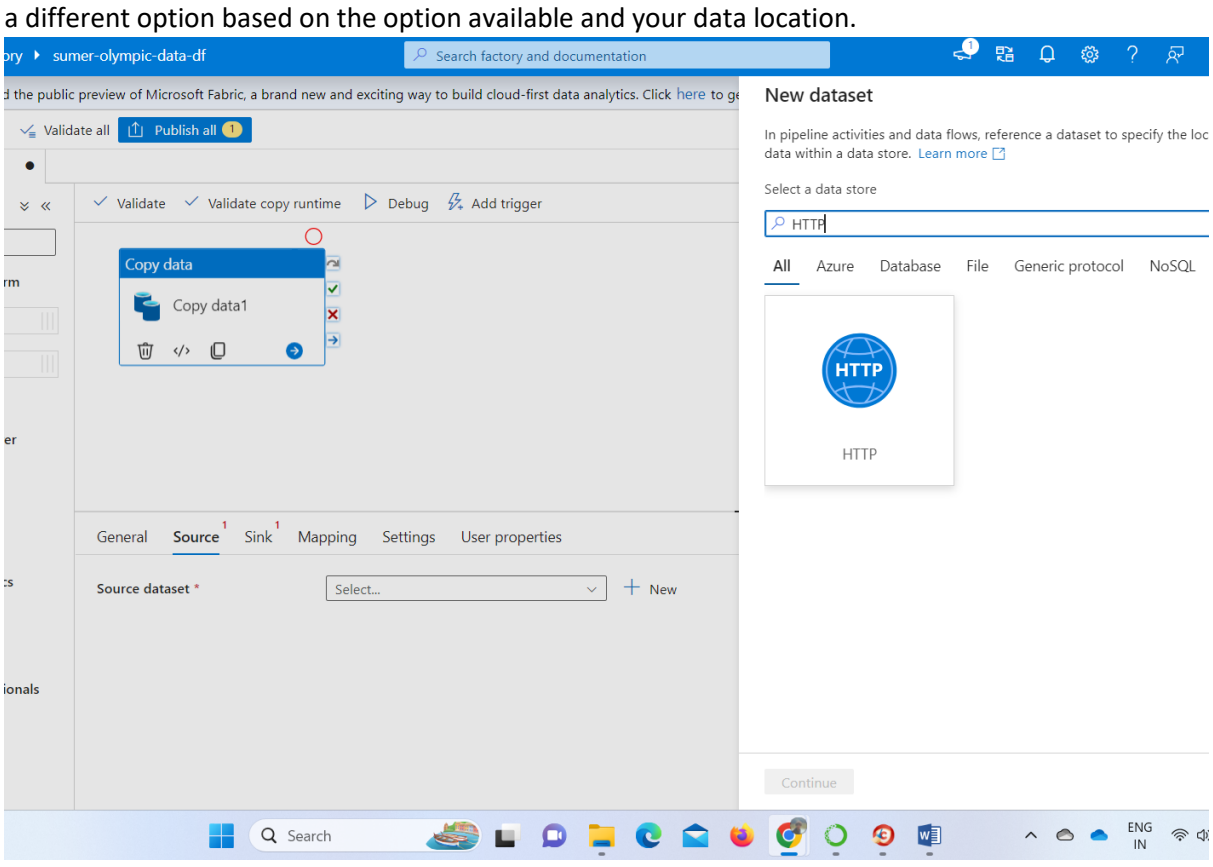
Drag the Copy data option to your window to start your data ingestion process. You can give a name to copying your data resource under the “General” tab.



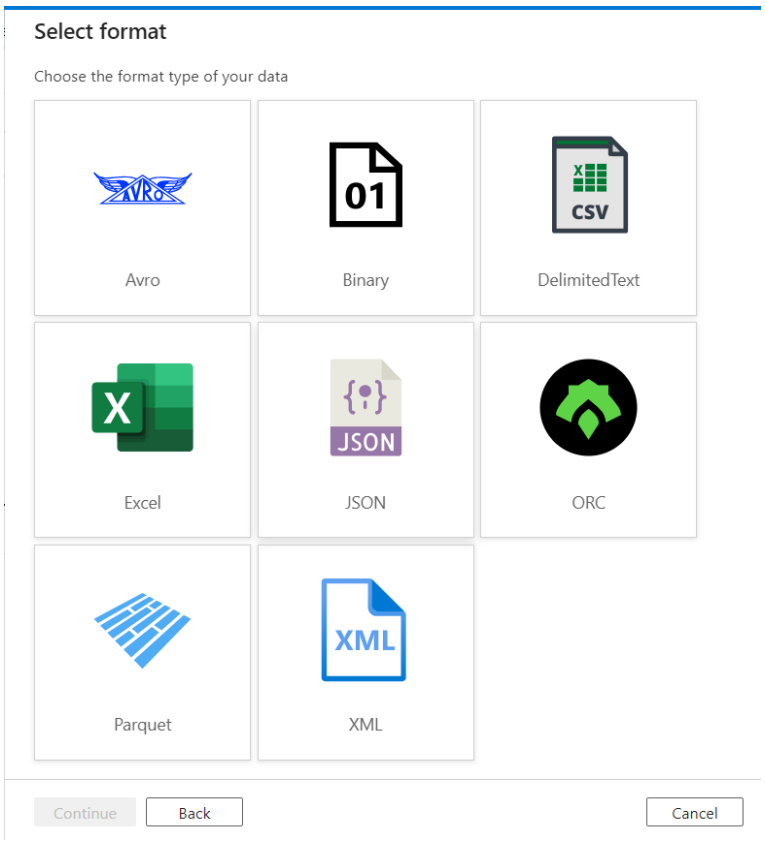
Select the Source tab and you define a location from where you have to ingest your data .Once you create a new source ,the portal will ask you to name the source and mention the path.



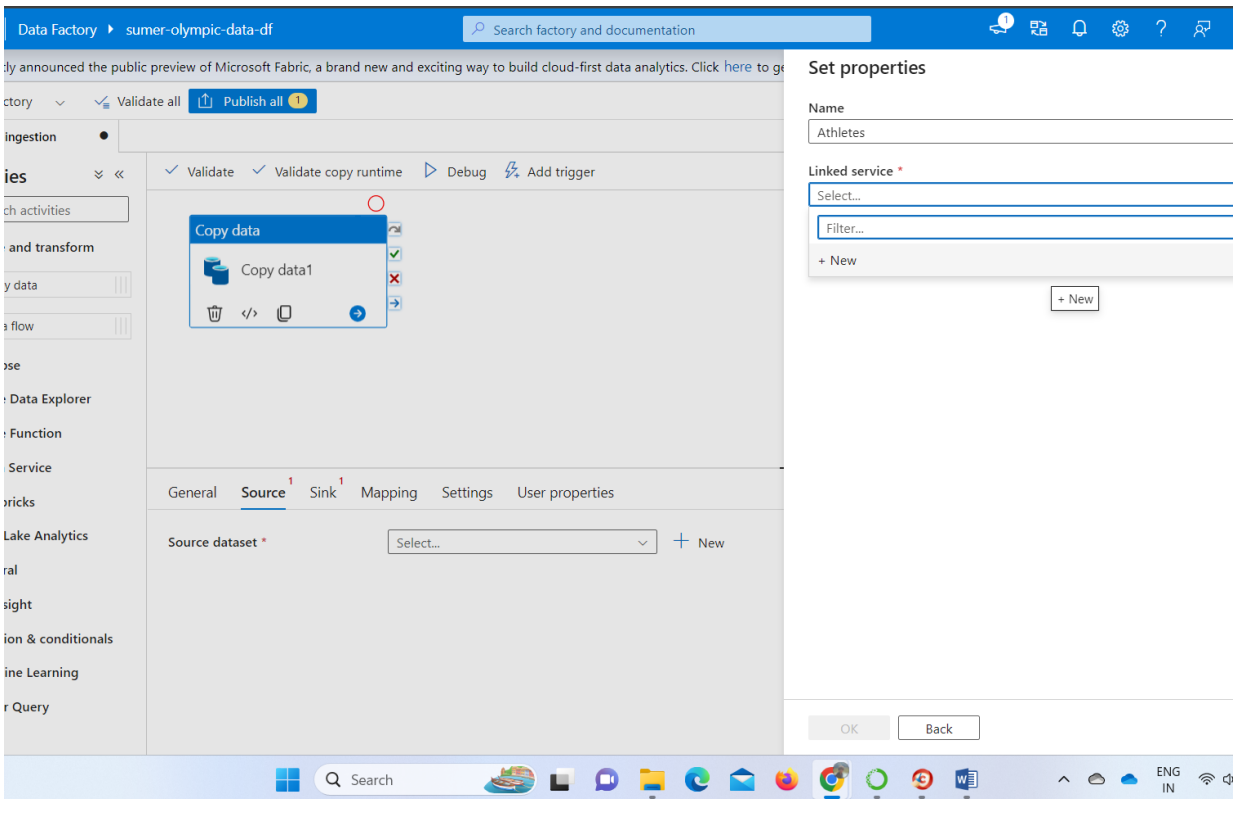
Here we are selecting the source from where we will ingest our data .I have selected through HTTP request because the data is stored on my GitHub repository. You can choose a different option based on the option available and your data location.



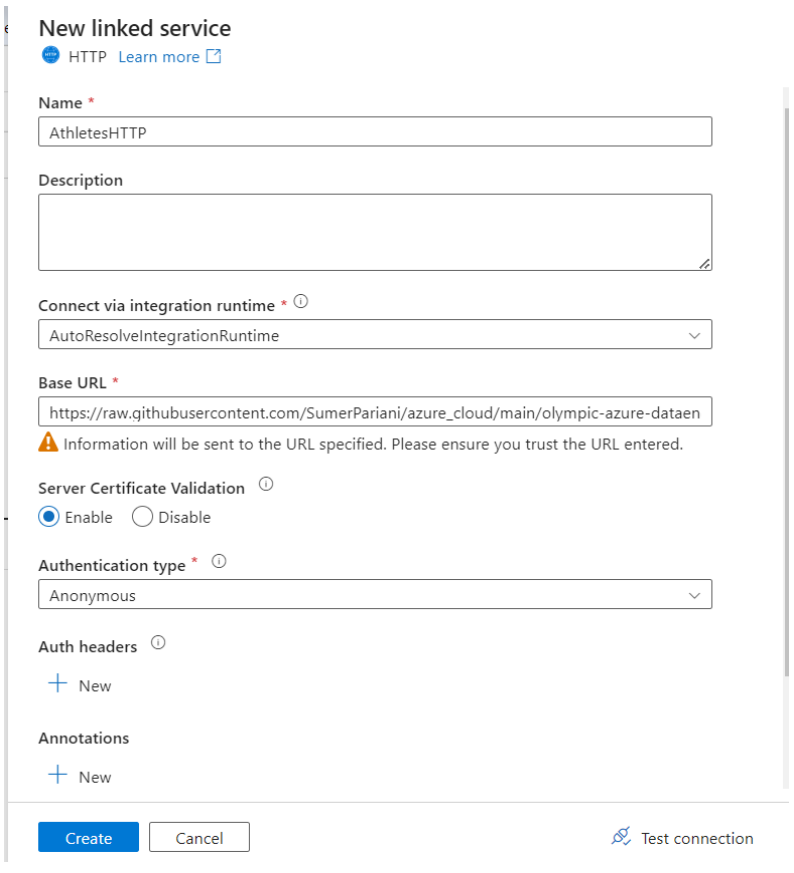
Select the format of the data you are ingesting .In my case it is a CSV file .



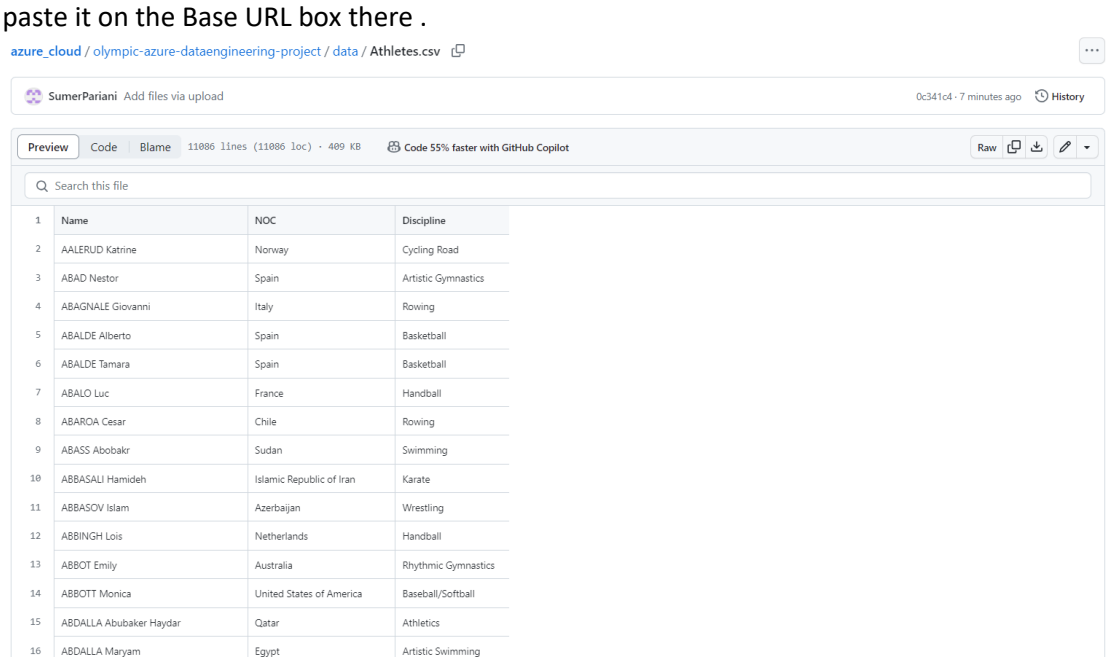
Give a name to your resource. In my case I am ingesting the athletes table from my repository so I gave the name as “Athletes”. Create a new linked service.



Then let’s give a name to our new linked service. I have named it Athletes HTTP to make things easier. I have kept the authentication as “anonymous” . It's not the best practice but you will have options to choose where you can have a secret key in your key vault through which you can access your linked service.

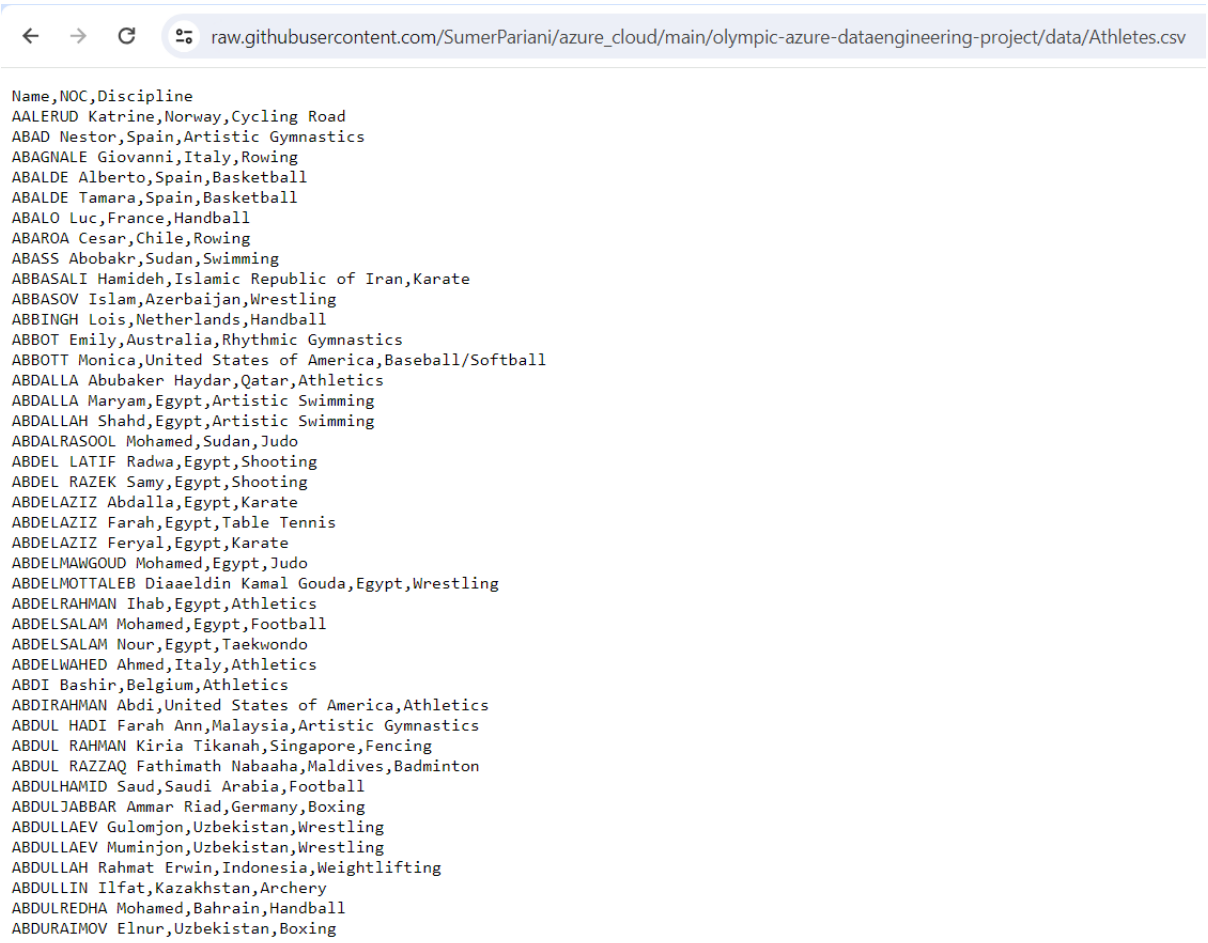


In this project I am ingesting the data from my GitHub repository so for this .I will go to my GitHub repository ,dataset .Select the Raw form and copy the URL of the table and paste it on the Base URL box there .



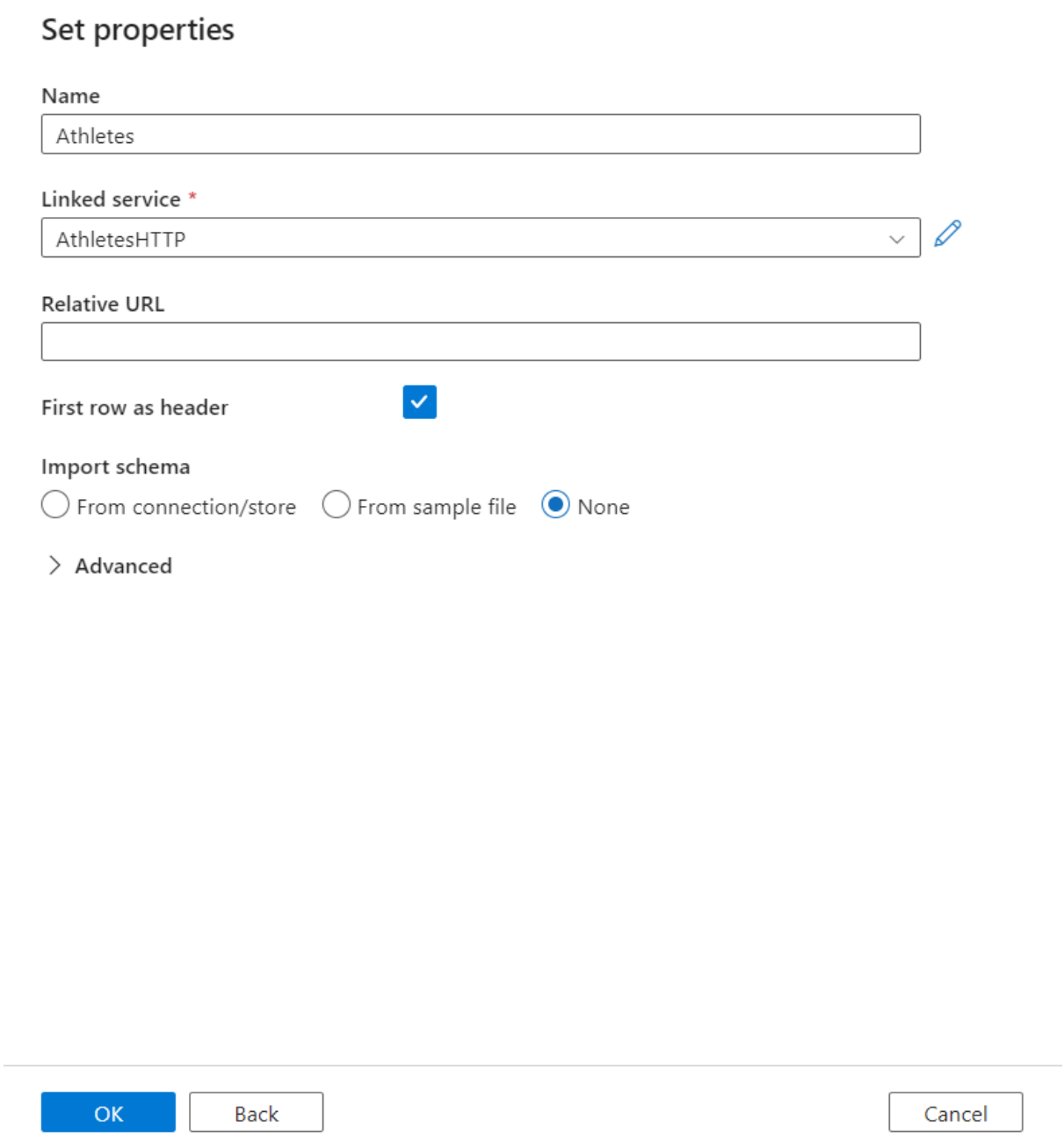
	Name	NOC	Discipline
2	AALERUD Katrine	Norway	Cycling Road
3	ABAD Nestor	Spain	Artistic Gymnastics
4	ABAGNALE Giovanni	Italy	Rowing
5	ABALDE Alberto	Spain	Basketball
6	ABALDE Tamara	Spain	Basketball
7	ABALO Luc	France	Handball
8	ABAROA Cesar	Chile	Rowing
9	ABASS Abobakr	Sudan	Swimming
10	ABBASALI Hamideh	Islamic Republic of Iran	Karate
11	ABBASOV Islam	Azerbaijan	Wrestling
12	ABBINGH Lois	Netherlands	Handball
13	ABBOT Emily	Australia	Rhythmic Gymnastics
14	ABBOTT Monica	United States of America	Baseball/Softball
15	ABDALLA Abubaker Haydar	Qatar	Athletics
16	ABDALLA Maryam	Egypt	Artistic Swimming

This is how your data looks in Raw form .

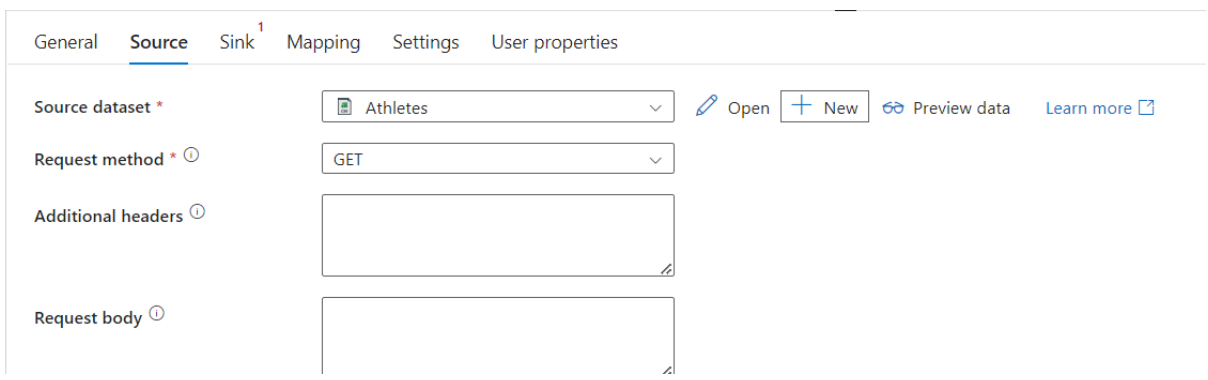



```
Name,NOC,Discipline
AALERUD Katrine,Norway,Cycling Road
ABAD Nestor,Spain,Artistic Gymnastics
ABAGNALE Giovanni,Italy,Rowing
ABALDE Alberto,Spain,Basketball
ABALDE Tamara,Spain,Basketball
ABALO Luc,France,Handball
ABAROA Cesar,Chile,Rowing
ABASS Abobakr,Sudan,Swimming
ABBASALI Hamideh,Islamic Republic of Iran,Karate
ABBASOV Islam,Azerbaijan,Wrestling
ABBINGH Lois,Netherlands,Handball
ABBOT Emily,Australia,Rhythmic Gymnastics
ABBOTT Monica,United States of America,Baseball/Softball
ABDALLA Abubaker Haydar,Qatar,Athletics
ABDALLA Maryam,Egypt,Artistic Swimming
ABDALLAH Shahd,Egypt,Artistic Swimming
ABDALRASOOL Mohamed,Sudan,Judo
ABDEL LATIF Radwa,Egypt,Shooting
ABDEL RAZEK Samy,Egypt,Shooting
ABDELAZIZ Abdalla,Egypt,Karate
ABDELAZIZ Farah,Egypt,Table Tennis
ABDELAZIZ Feryal,Egypt,Karate
ABDELMAWGOUD Mohamed,Egypt,Judo
ABDELMOTTALEB Diaaeldin Kamal Gouda,Egypt,Wrestling
ABDELRAHMAN Ihab,Egypt,Athletics
ABDELSALAM Mohamed,Egypt,Football
ABDELSALAM Nour,Egypt,Taekwondo
ABDELWAHED Ahmed,Italy,Athletics
ABDI Bashir,Belgium,Athletics
ABDIRAHMAN Abdi,United States of America,Athletics
ABDUL HADI Farah Ann,Malaysia,Artistic Gymnastics
ABDUL RAHMAN Kiria Tikanah,Singapore,Fencing
ABDUL RAZZAQ Fathimath Nabaaha,Maldives,Badminton
ABDULHAMID Saud,Saudi Arabia,Football
ABDULJABBAR Ammar Riad,Germany,Boxing
ABDULLAEV Gulomjon,Uzbekistan,Wrestling
ABDULLAEV Muminjon,Uzbekistan,Wrestling
ABDULLAH Rahmat Erwin,Indonesia,Weightlifting
ABDULLIN Ilfat,Kazakhstan,Archery
ABDULREDHA Mohamed,Bahrain,Handball
ABDURAIMOV Elnur,Uzbekistan,Boxing
```

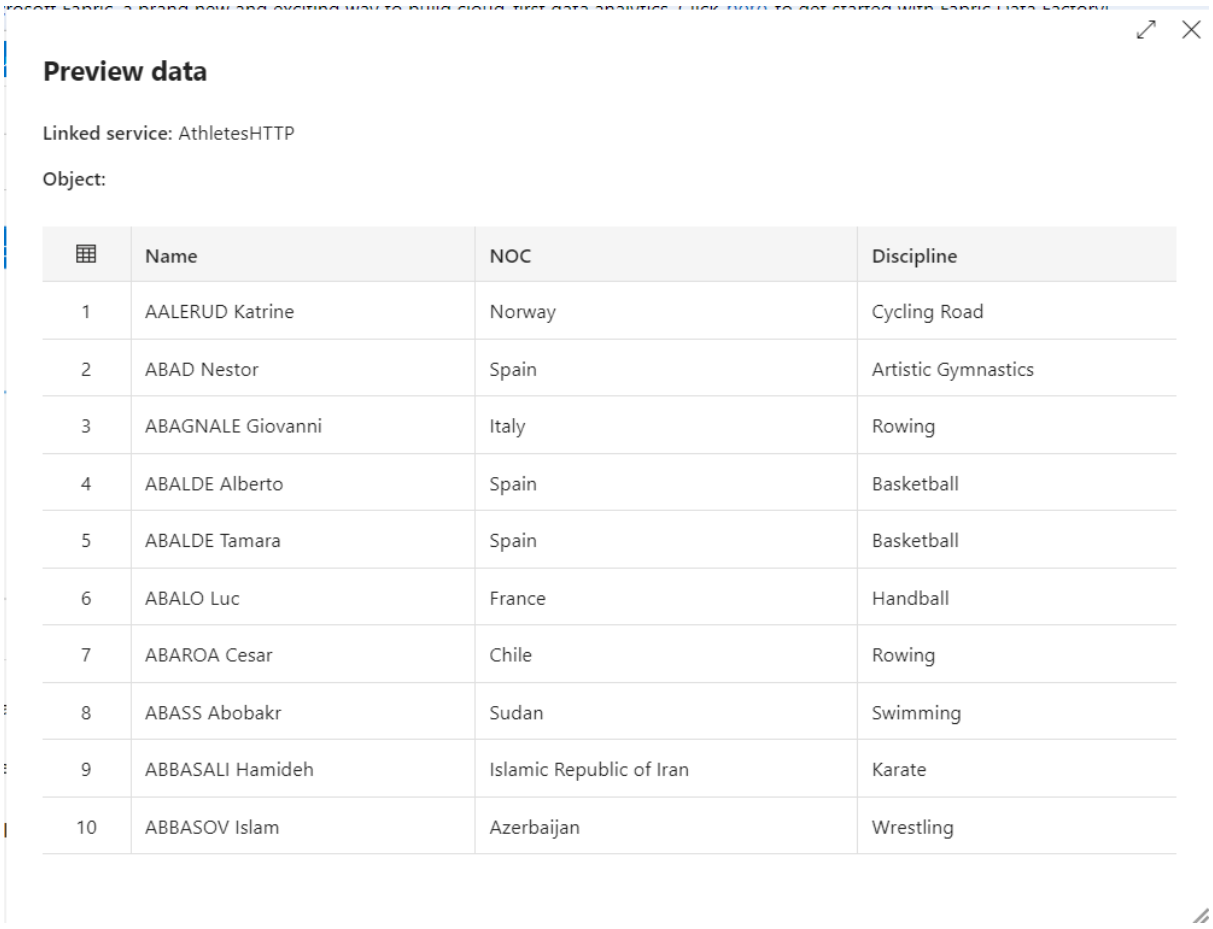
let’s select the option to keep the first row as header and import schema as none .By clicking Okay we have successfully created the source of the data resource to be ingested



You can preview the data by going on the source tab and clicking the “preview data” option

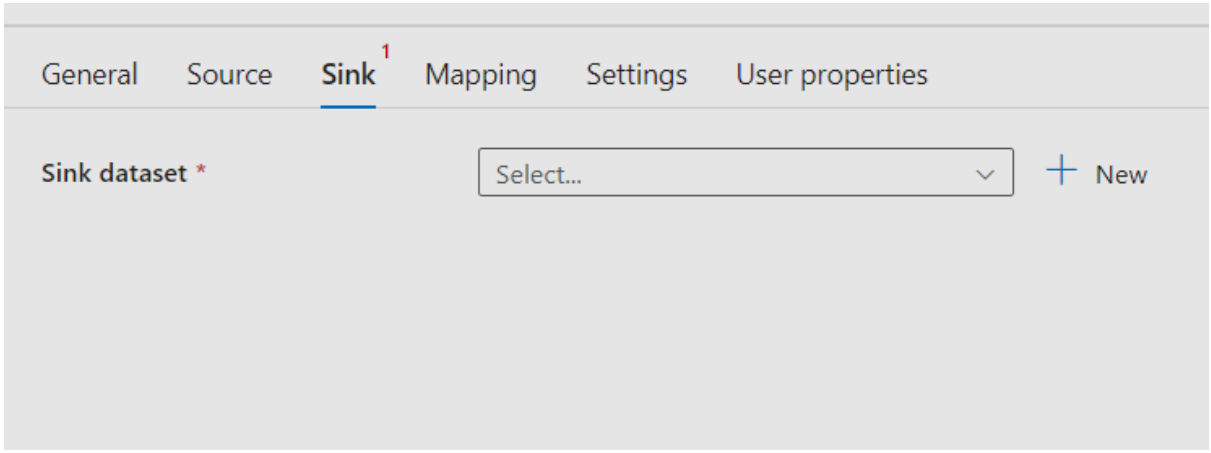


Here I have previewed the data

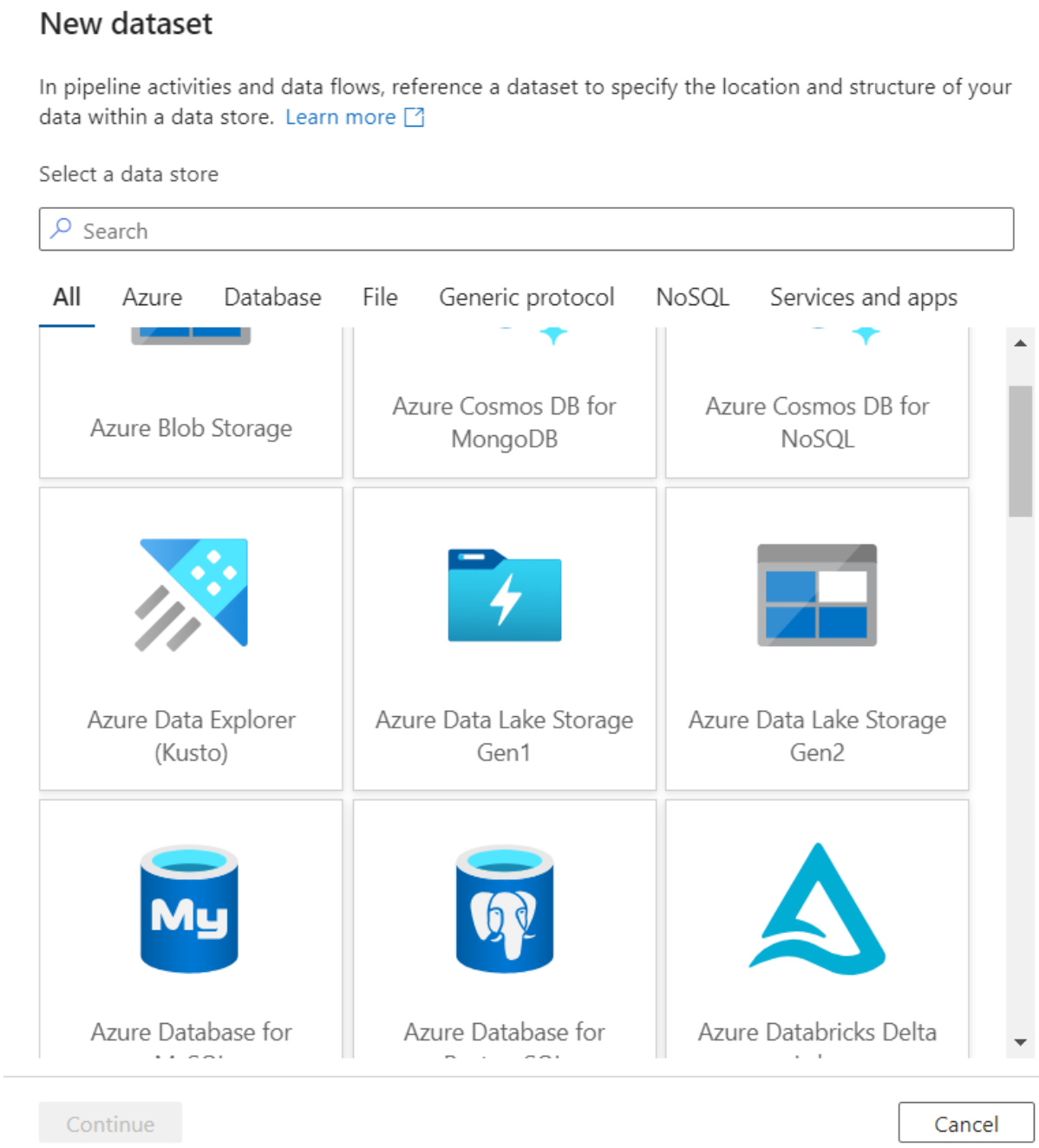


	Name	NOC	Discipline
1	AALERUD Katrine	Norway	Cycling Road
2	ABAD Nestor	Spain	Artistic Gymnastics
3	ABAGNALE Giovanni	Italy	Rowing
4	ABALDE Alberto	Spain	Basketball
5	ABALDE Tamara	Spain	Basketball
6	ABALO Luc	France	Handball
7	ABAROA Cesar	Chile	Rowing
8	ABASS Abobakr	Sudan	Swimming
9	ABBASALI Hamideh	Islamic Republic of Iran	Karate
10	ABBASOV Islam	Azerbaijan	Wrestling

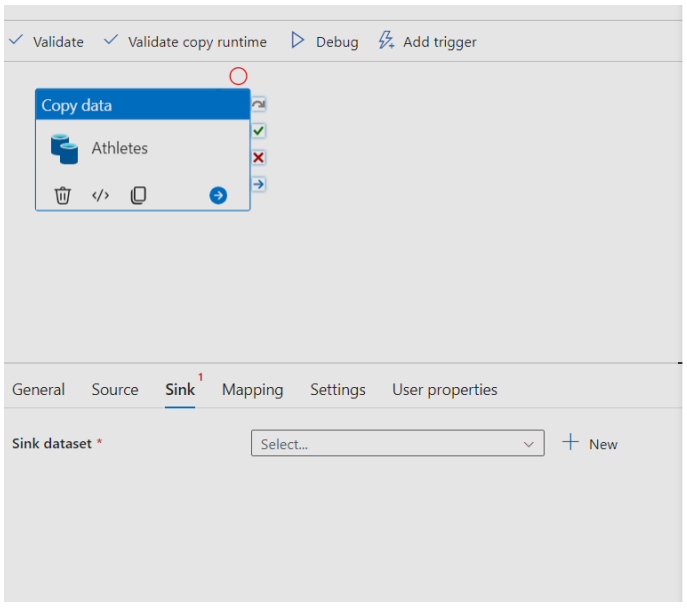
Now the next step is to create Sink for the data .Let’s create a sink for the source data by selecting the source tab and clicking the create new option.

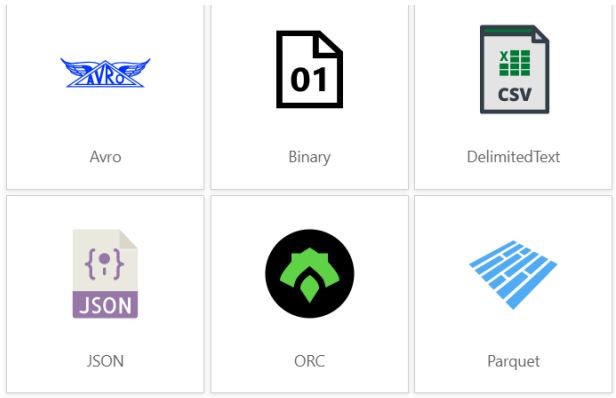


Let’s select the Sink location for the data source we have created .I have selected Data Lake Gen2 (ADLS).You can select any other option your prefer.

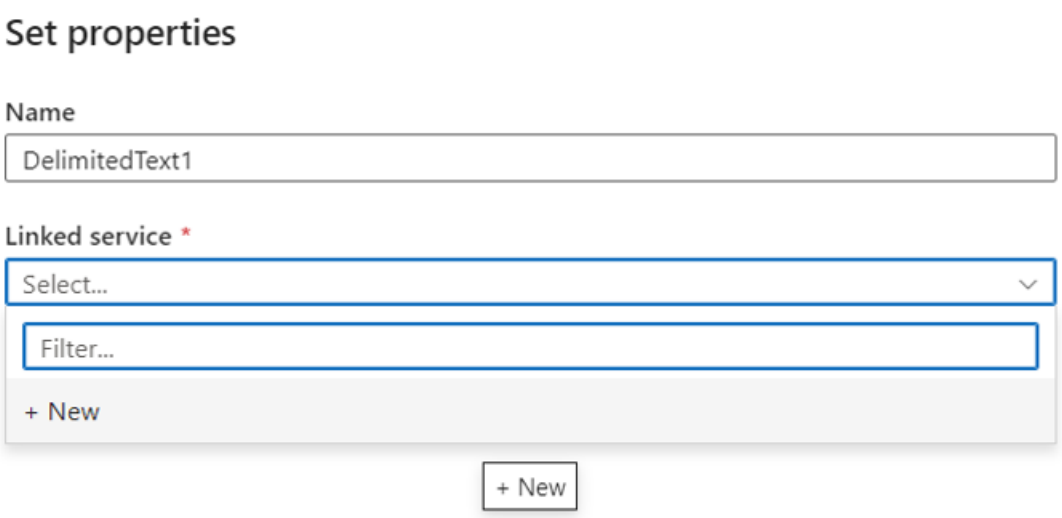


Now we have to select the format in which we have to keep our data in our ADLS .I have selected the CSV format .

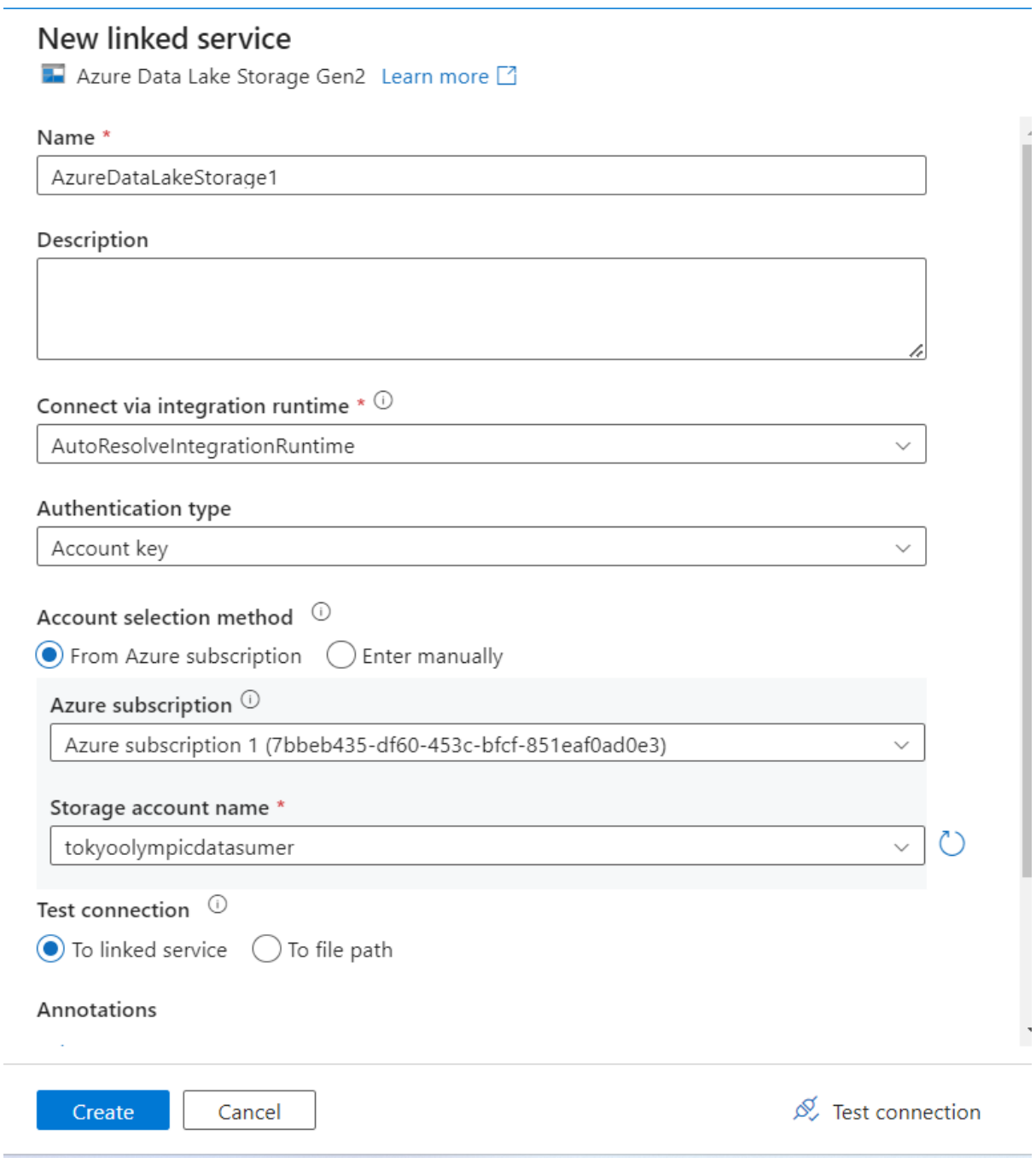




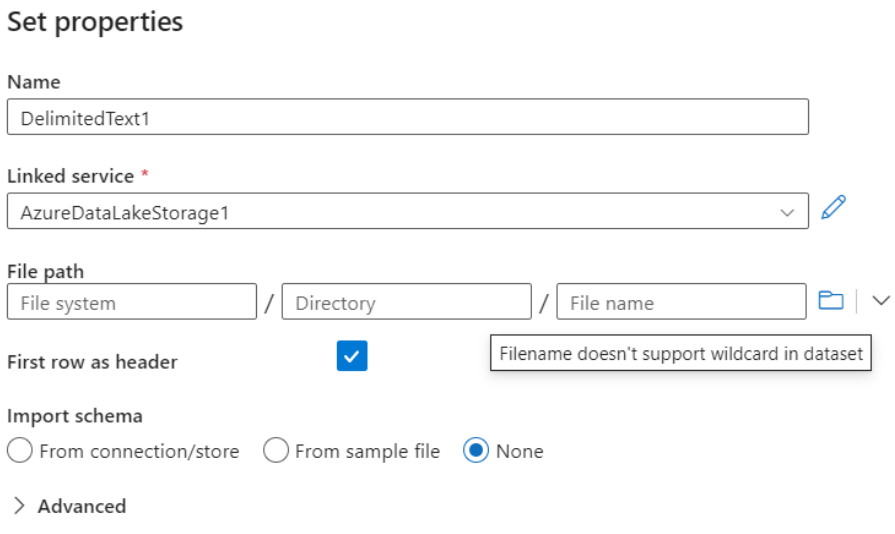
Set the properties of your sink service ,Give name and create a new linked service



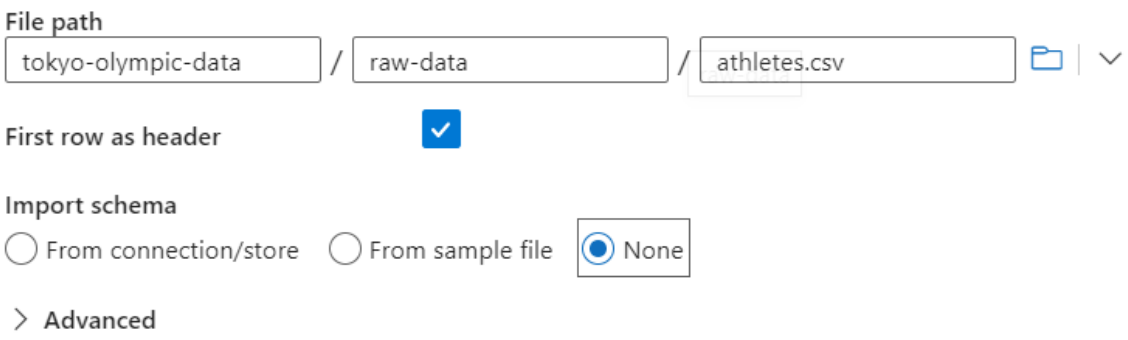
I have given the name AzureDataLakeStorage1 and then selected your subscription and the storage account (tokyoolypicdatasumer in my case).



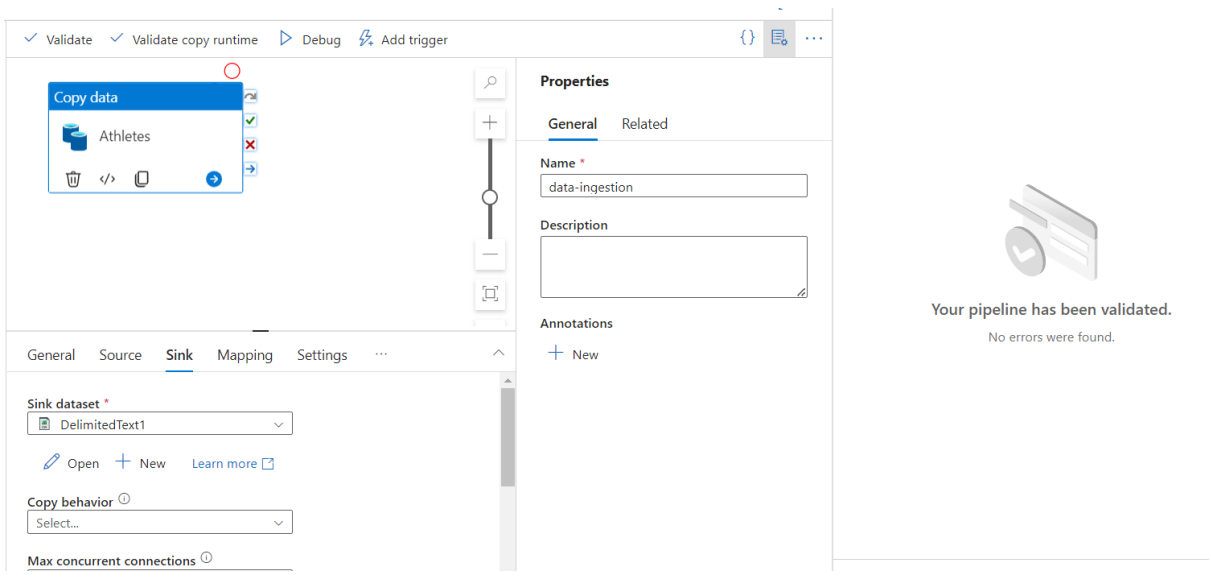
select the file path by browsing the “raw-data” directory we have created in our storage Blob->container->directory.



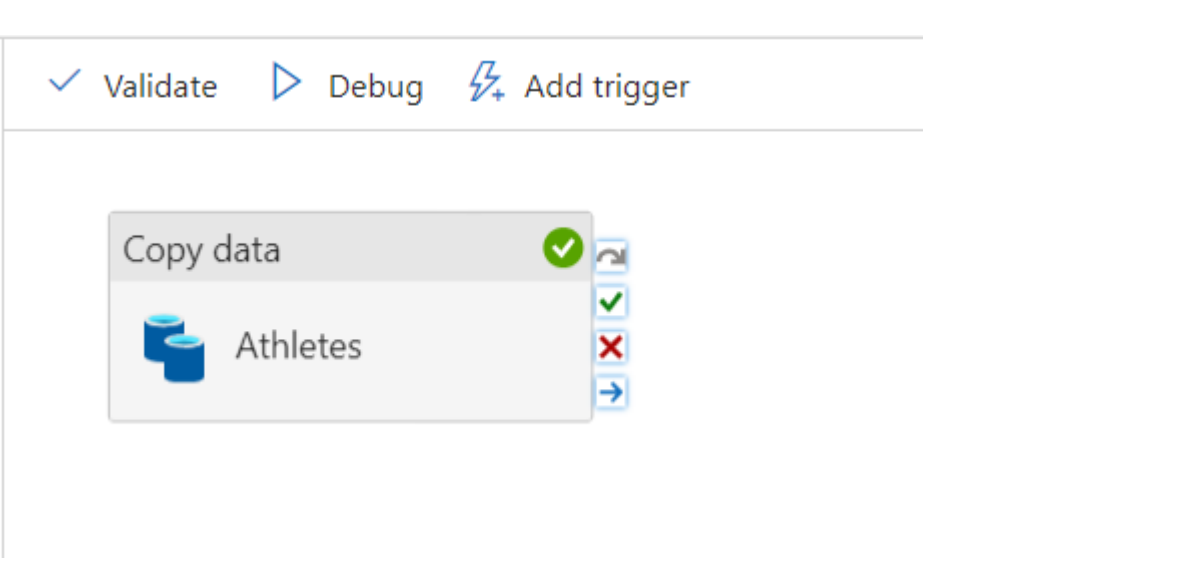
I have not imported the schema and then clicked okay.



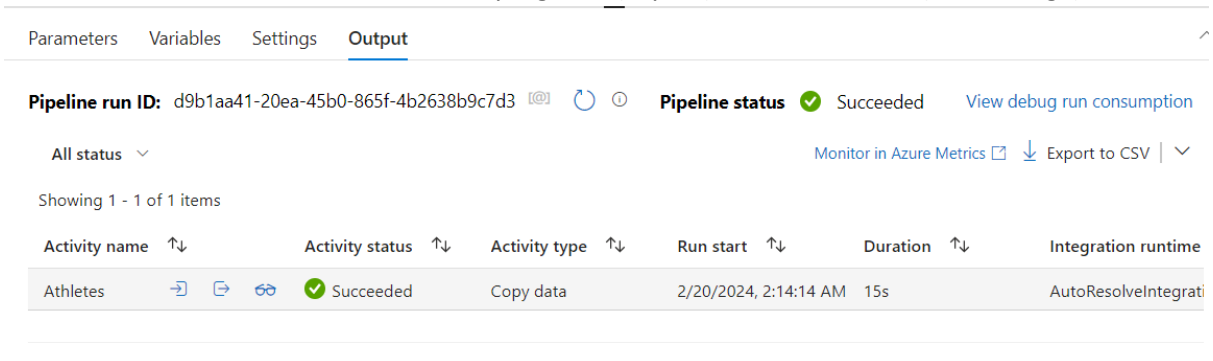
And ring ding ding ding we have successfully ingested our athletes data.



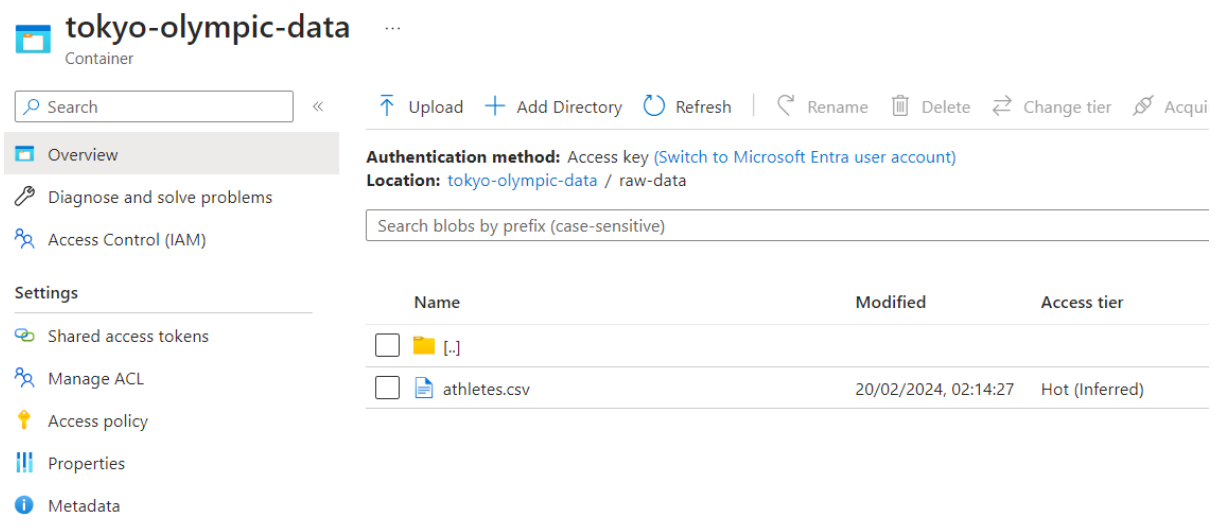
Validate the change you have made and then if the changes are validated click on publish option to save the changes



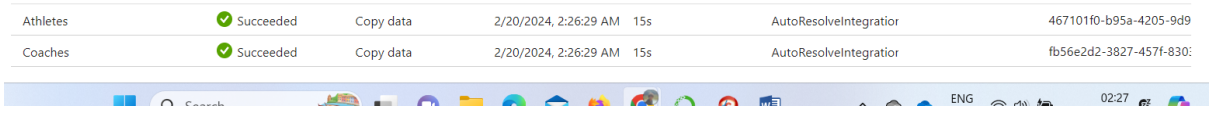
Now we can see that we have successfully ingested(copied) our data to ADLS(raw-storage).



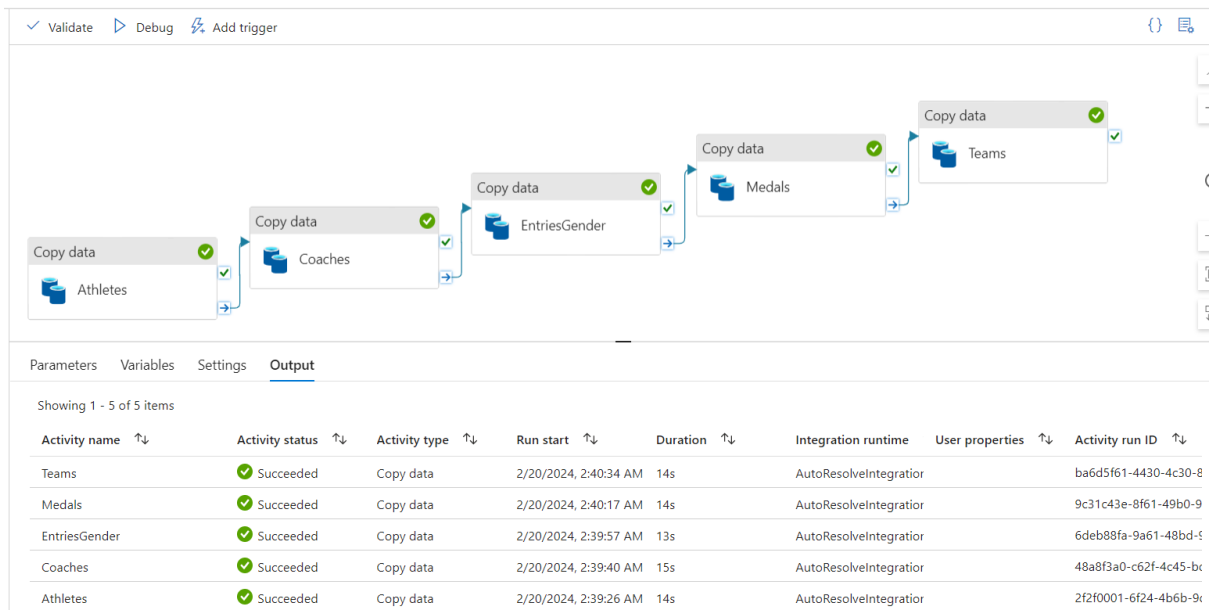
you can see the athletes CSV file by going to your raw-data directory in Azure BLOB



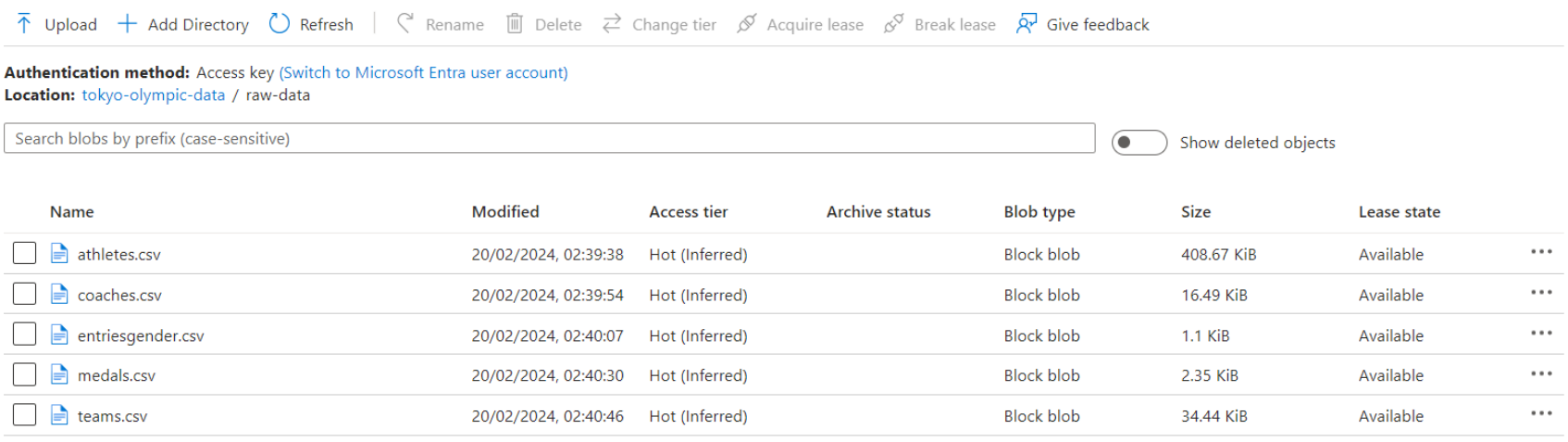
let’s repeat the same process for all the five data files we have



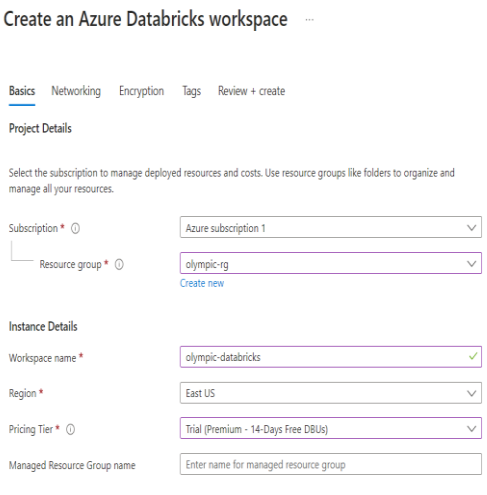
Once we are done with this we can see that we have ingested your data and created our ingestion pipeline .



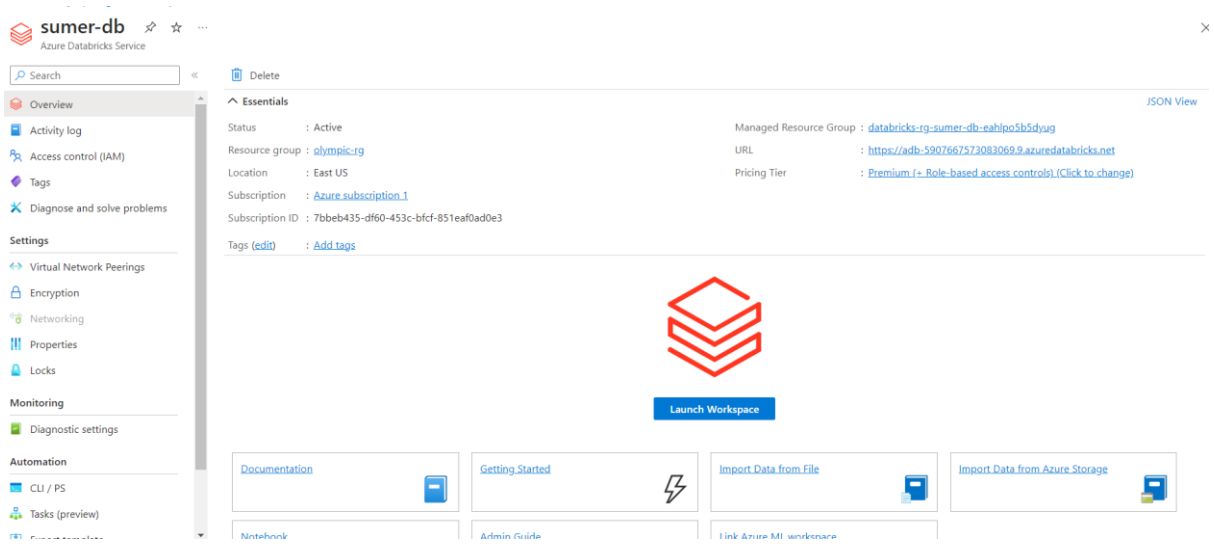
All the five CSV files are present in our raw-data storage container .



Now let us set up the Azure Databricks workspace to perform the necessary transformation in our data using Pyspark and then store the data in your transformed folder in a storage blob.create your Azure Databricks workspace by selecting the same resource group and region, give a name and click create .

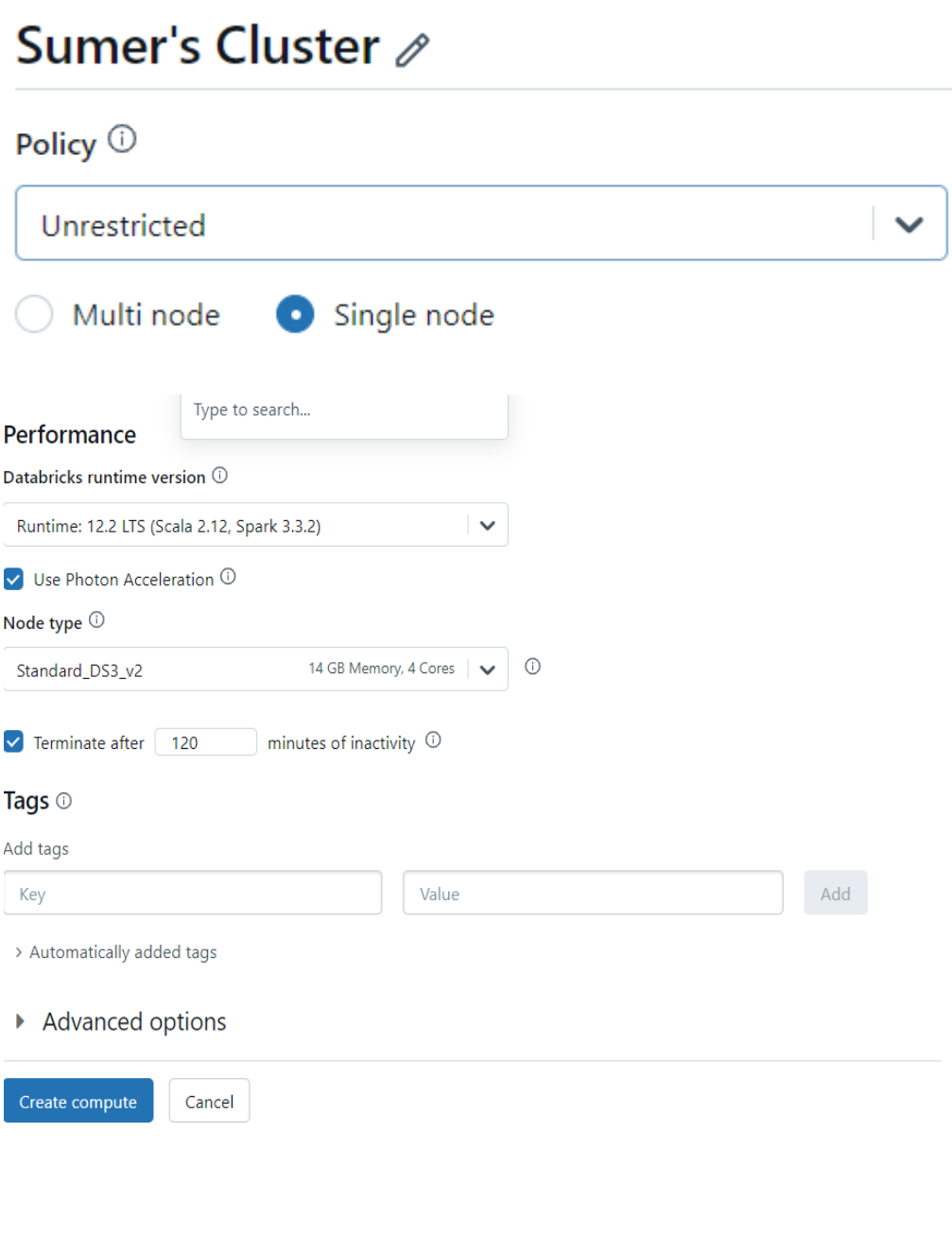


Launch your workspace

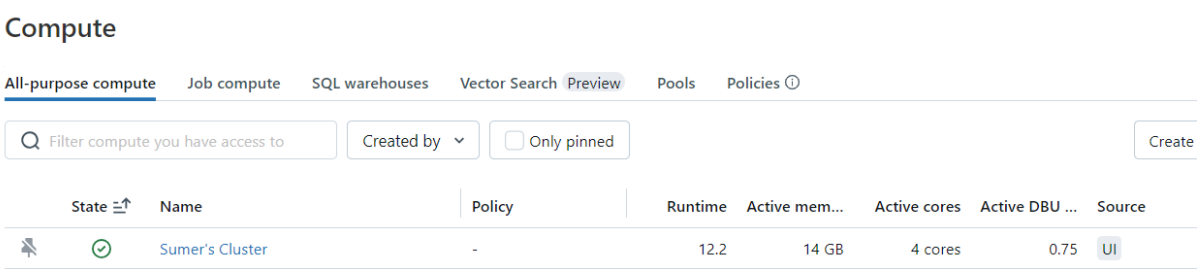


Select the Compute option to create your work cluster

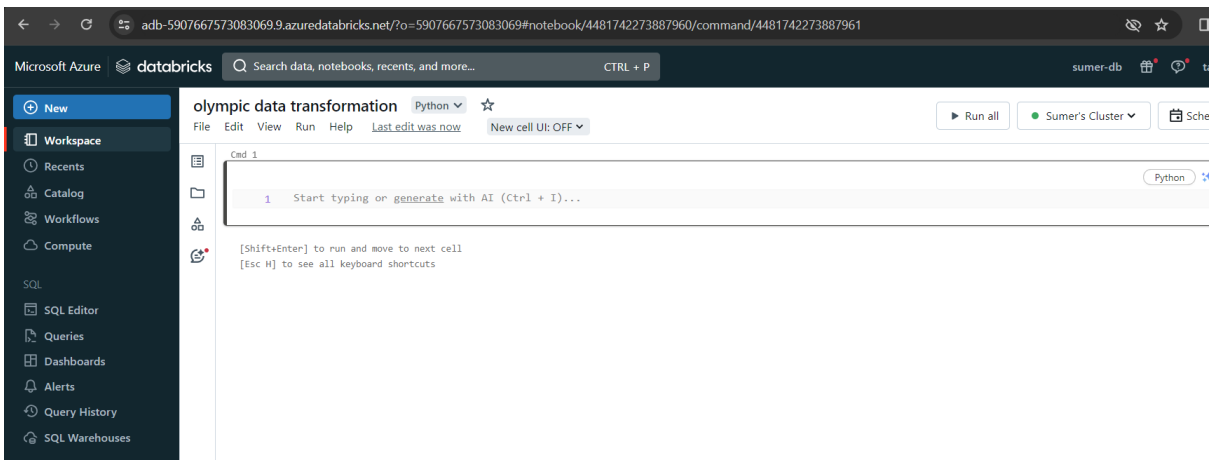
Once you are in your workspace (Sumer's Cluster),select the machine you want to work with by selecting the policy, number of nodes, node type, runtime version and click create compute .



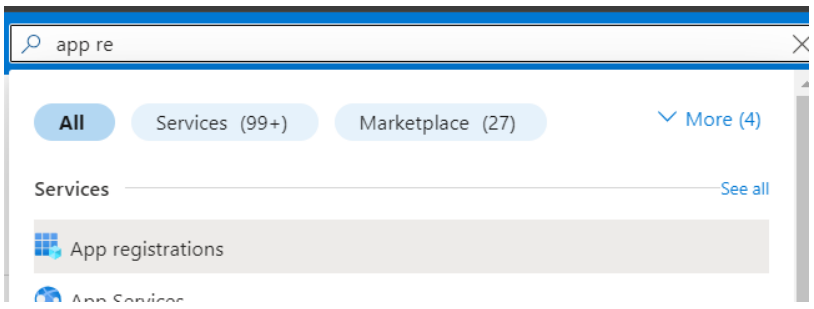
we can see that we have successfully created our cluster now on which we can perform our transformation in Pyspark



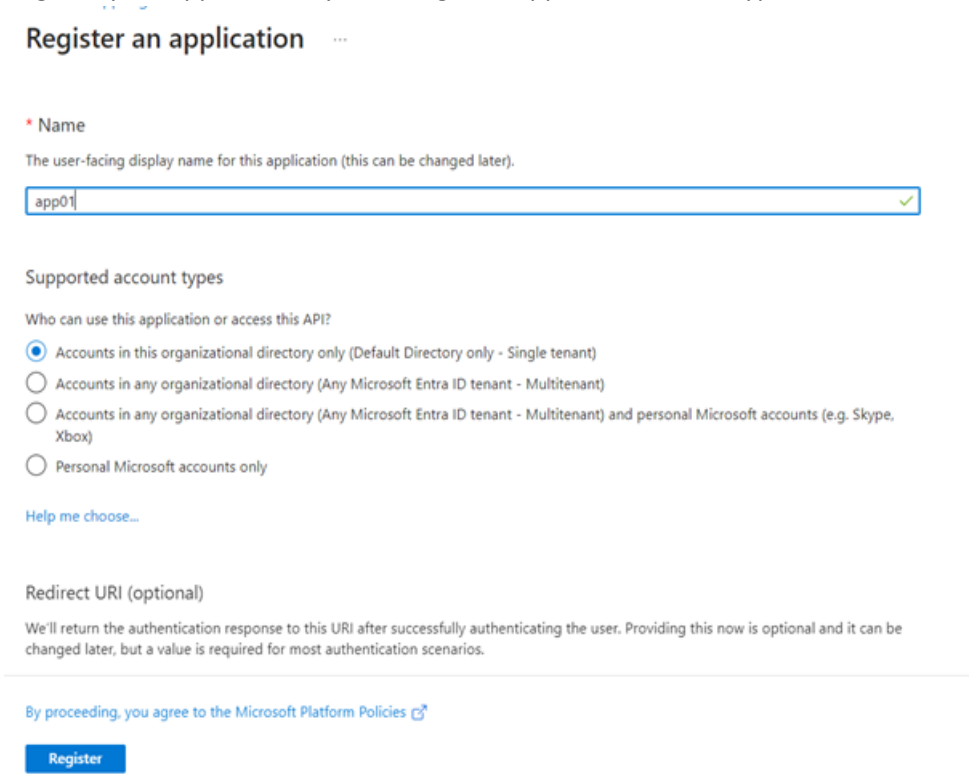
Let’s give a name to our data transformation code and start our data transformation process. Before we write our data transformation code we need to give permission to data brick to access our data.



For that we need to go to the App registration service by type App registration in search box

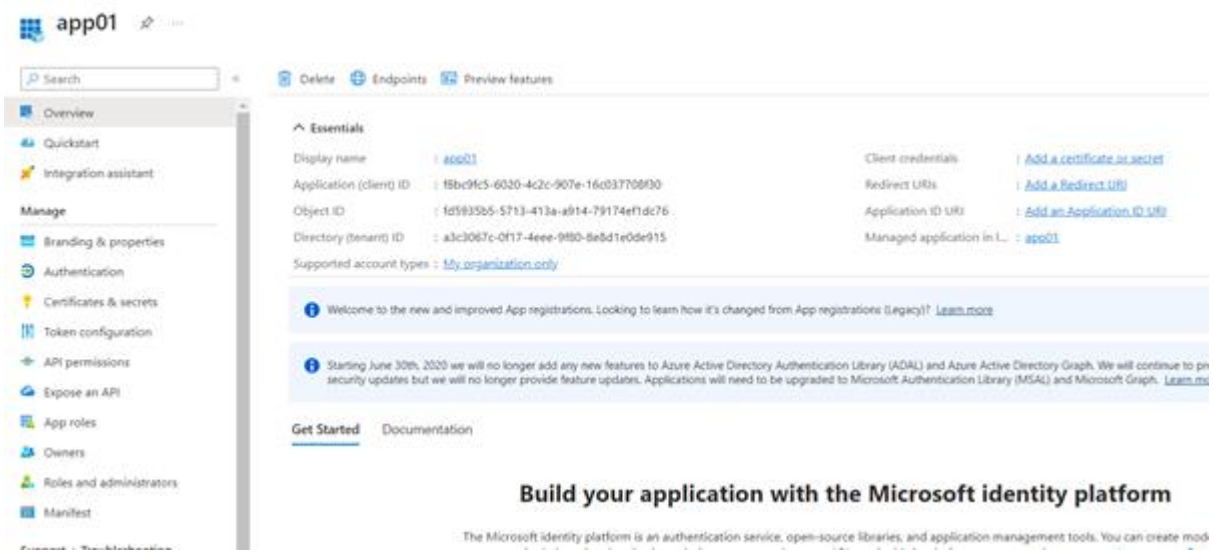


Click on new registration
Register your application by selecting the supported account types.

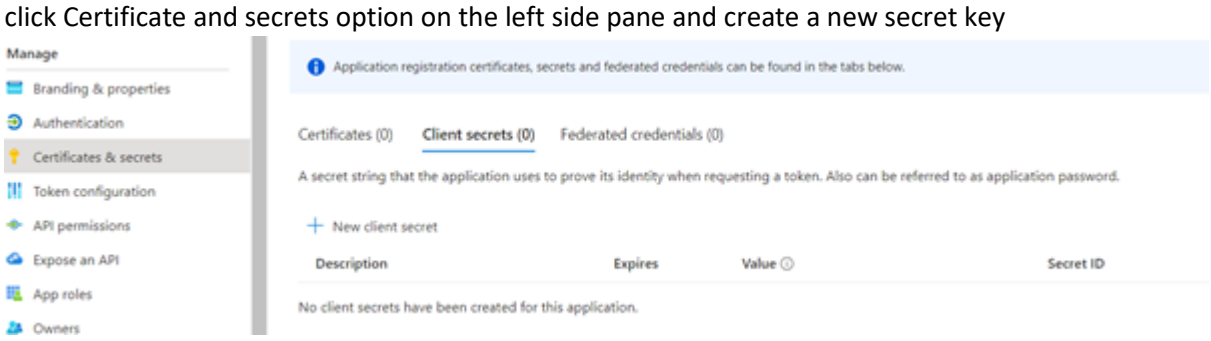


Once the app is created you will have your
Application (client ID)
Directory(Tenant ID)

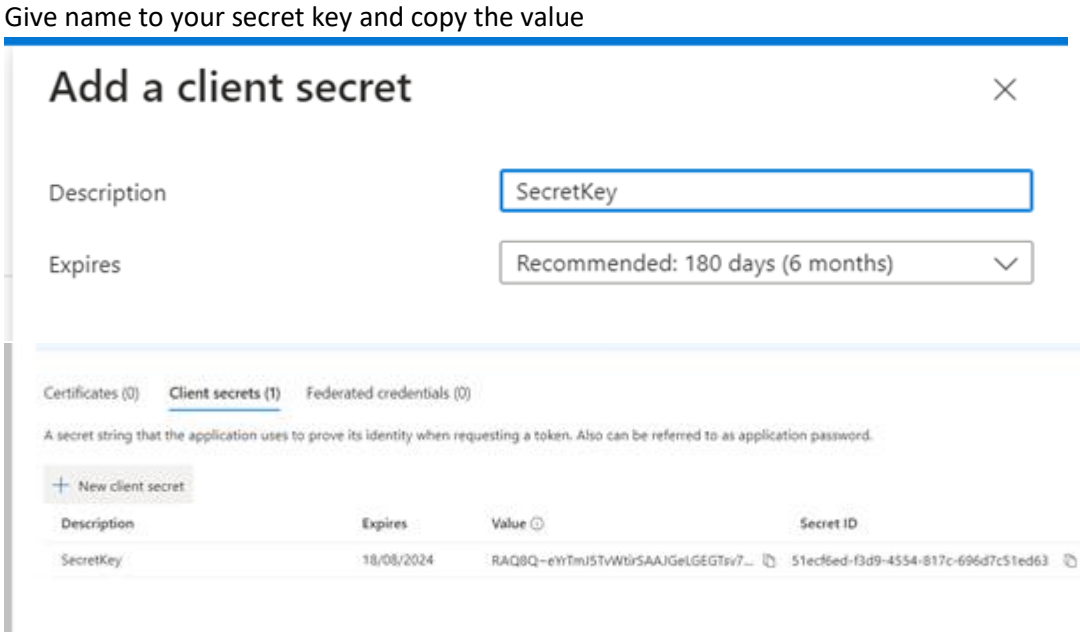
Copy paste these ID’s on a notepad



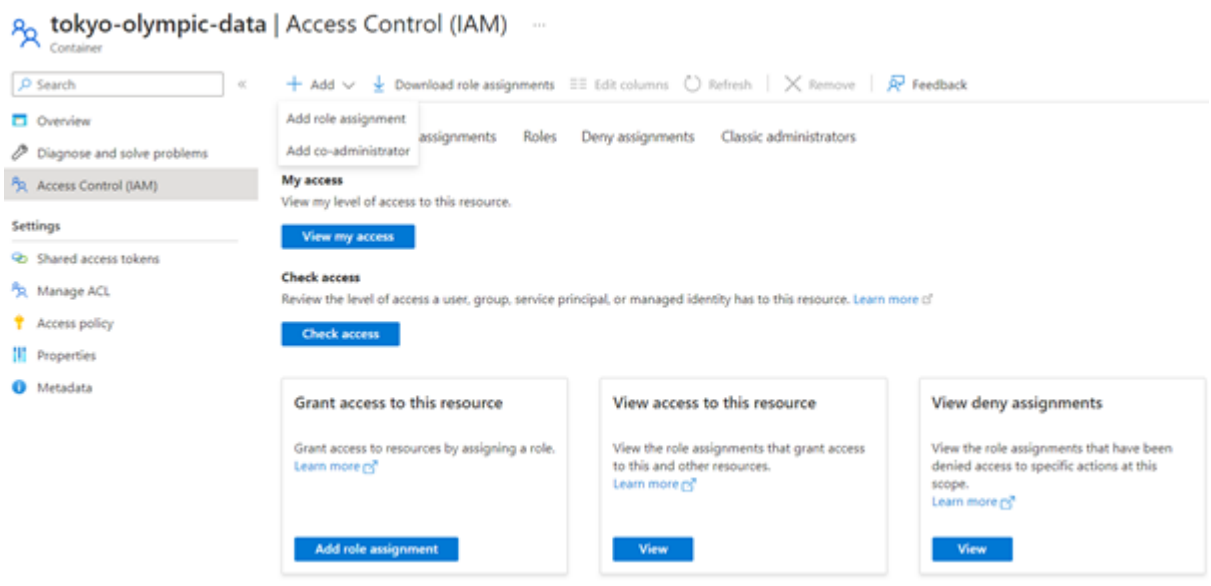
click Certificate and secrets option on the left side pane and create a new secret key



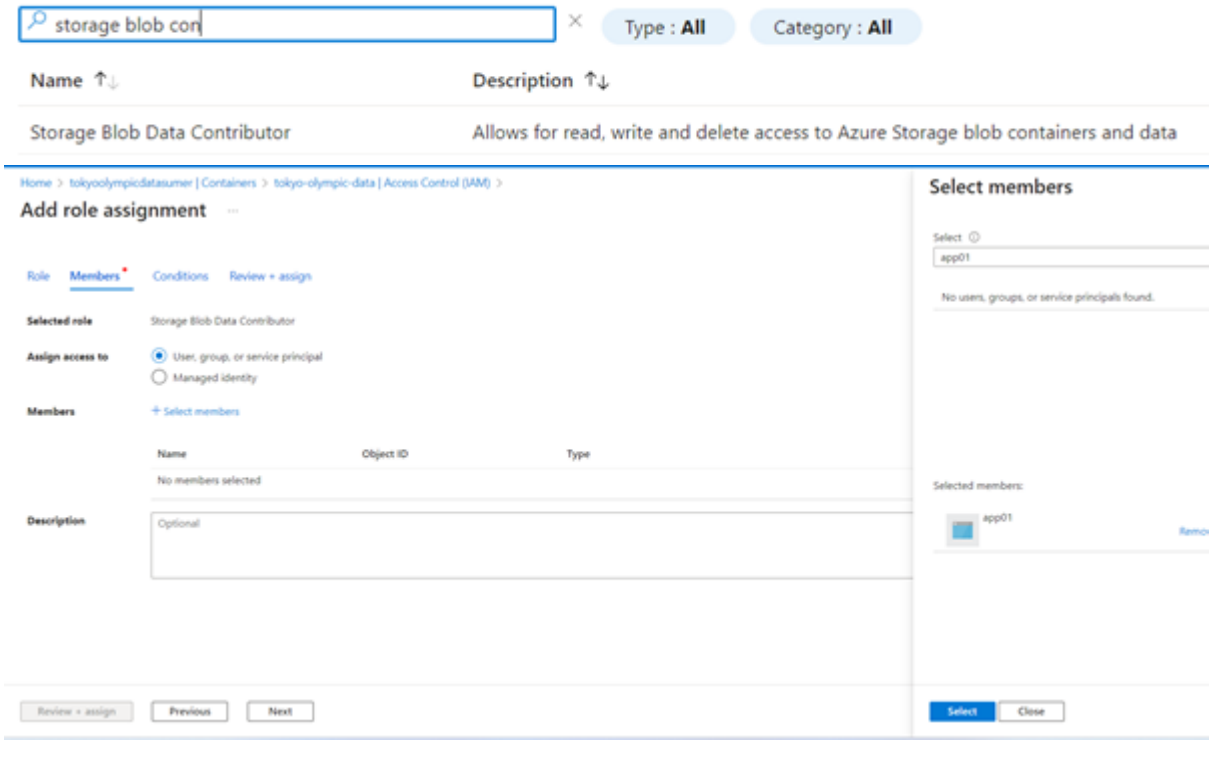
Give name to your secret key and copy the value

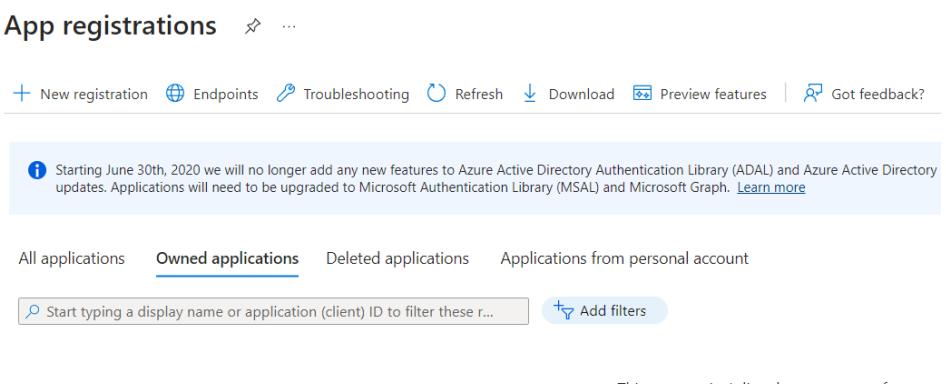


Now we have to give permission to our to be able to access the data kept in our raw storage .For that go to your storage blob->containers->Access control->Add->Add role assignment

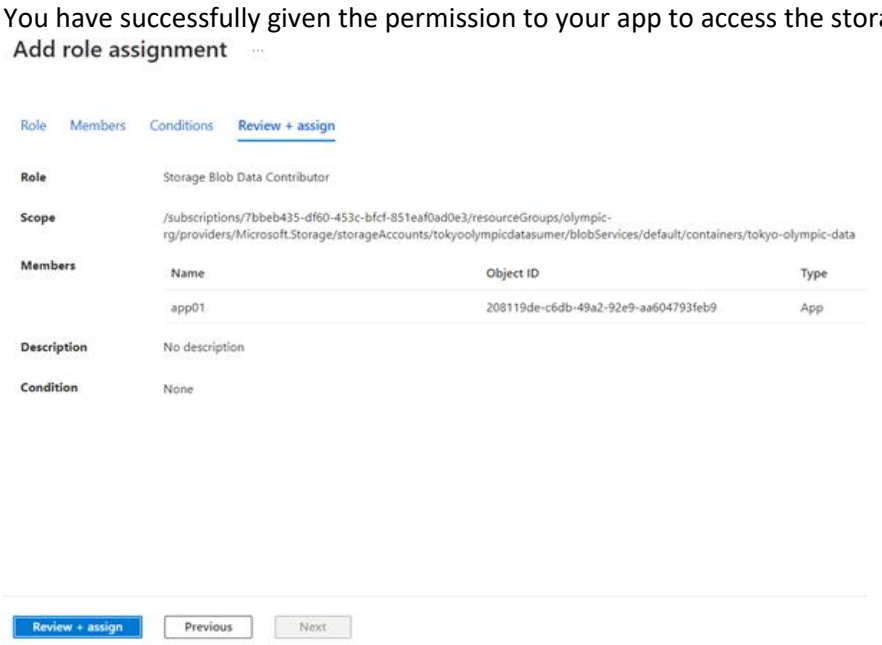


Select the role of Storage Blob data contributor and select the members as your app .





You have successfully given the permission to your app to access the storage account



Once we have successfully given the permission to our app we now have to connect with our app using the databrick workspace and mount our data on data bricks .

Below is the code to connect with your app and mount the data on databricks .

```
configs = {"fs.azure.account.auth.type": "OAuth",

"fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",

"fs.azure.account.oauth2.client.id": "f96d1bdb-fb31-4af5-96bf-2c7c68afa0d1",

"fs.azure.account.oauth2.client.secret": 'nh08Q~2_QV5P1lcuI8OoNKshcPpJAiAMhGsFxccF',

"fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/a3c3067c-0f17-4eee-9f80-8e8d1e0de915/oauth2/token"}

dbutils.fs.mount(

source = "abfss://tokyo-olympic-data@tokyoolympicdatasumer.dfs.core.windows.net", # contrainer@storageacc

mount_point = "/mnt/tokyoolymicsumer",

extra_configs = configs)
```

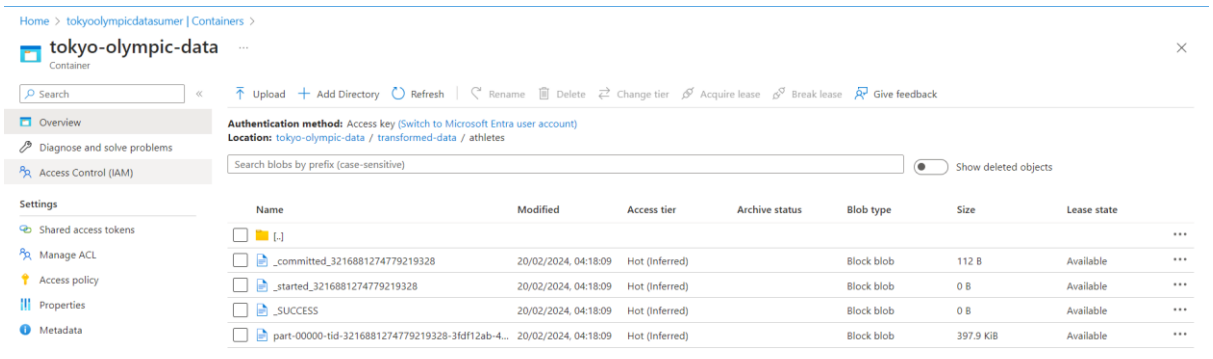
For the data transformation you need to start your spark session and explore your dataset for any anomalies .

I have inspected the dataset and changed the data types of a few columns by exploring their schema's. You can also set the first row as header here and deal with duplicates,missing values ,data type transformation ,new column creation and the transformation you need in your dataset here .

Link to Transformation notebook-Transformation_notebook

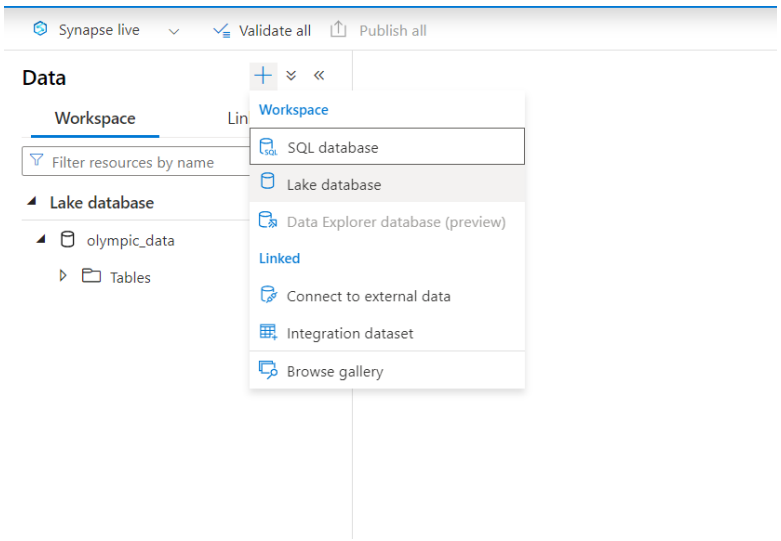
Now we have to load our transformed dataset to the transformed-data container in our storage block .Below is the code for that .

```
athletes.repartition(1).write.mode("overwrite").option("header",'true').csv("/mnt/tokyoolymicsumer/transformed-data/athletes")
coaches.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymicsumer/transformed-data/coaches")
entriesgender.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymicsumer/transformed-data/entriesgender")

medals.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymicsumer/transformed-data/medals")

teams.repartition(1).write.mode("overwrite").option("header","true").csv("/mnt/tokyoolymicsumer/transformed-data/teams")
```

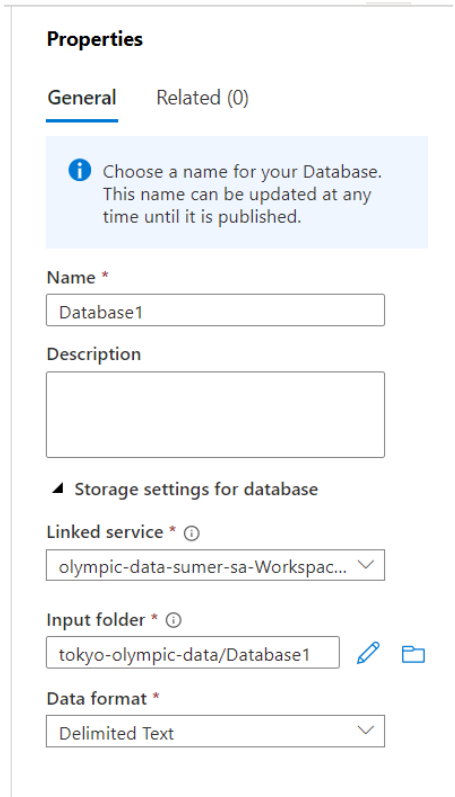


Now that we have our transformed data in the storage blob we need to create an environment to query this data using Azure Synapse analytics .For this follow the same process of searching for Azure Synapse analytics in the search box and open the workspace .

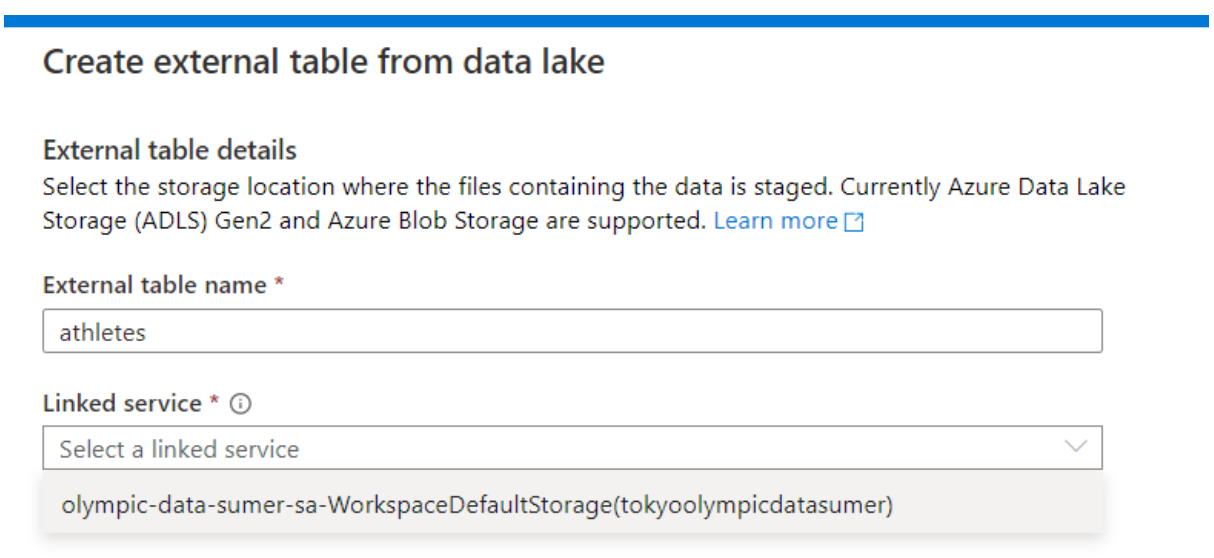
To ingest the data we need to click on the Add option under the Data pane and select the option of lake data .



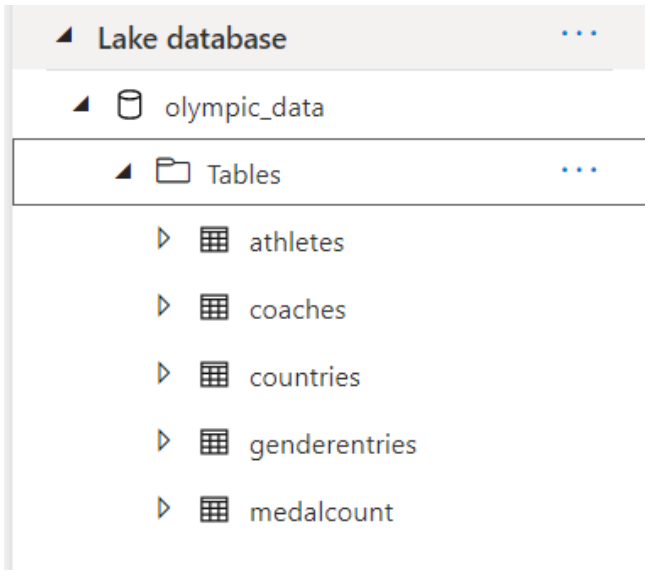
Give a name to our new Database, select the linked services and browse to the transformed data file



Once the database is created we need to import the table by adding table option ,give name to our table and attach the related linked services .Validate the data you are importing and publish it .



You will have all the data in your Azure Synapse analytics ,you can run queries on the data to get insights of the dataset .



The last and final step in the pipeline is to create visualization based on the insights which you got by connecting Azure Synapse analytics to Power BI or any other dashboarding tool.

Dashboard Link-https://github.com/SumerPariani/azure_cloud/blob/main/Olympic_data.pbix

Below is the Dashboard which I have created .