TOKOYO OLYMPIC DATA ANALYZATION USING AZURE SERVICES

PROJECT REPORT DONE BY :-

Ms.Mansi Jadhav

Mr.Jayesh Waingankar

Mr.Yogesh Pawar

TABLE OF CONTENTS

SR.NO	CONTENTS	PAGE NO.
1	Synopsis	3 - 4
2	Description of services	5
3	Usage of services	6-7
2	Data Flow	8
3	Diagram	9
4	Project Design	10-53
5	Advantages and Conclusion	54-55
6	Bibliography	56

SYNOPSIS

- **Introduction :-**

We have done analysis of data set using azure services .Here we have our data source and dataset of Tokyo Olympic. We will copy the data from data source to our Azure Data Storage using Data Factory service so we will build pipeline where we will upload raw data on data storage which is Data Lake Gen2.Once we have our data available onto raw storage then we will use azure data picks and write our code and do some transformation. Main Goal is to build a data file plan and show how you can build project so it just simple pipeline and try to remove duplicates and also drop columns rename it and then upload our data back to data lake storage which is transform data so this is initial pipeline where we will load our data do transformation and then again load that data onto data lake and then we will use synapse analytics to understand our data so we can easily write SQL query on top of these data that we have transform and extract the insight from it. So we will perform each and every steps.

- **Services used:-**

1. Rescource Group
2. Azure Data Factory
3. Azure Synapse Analyze
4. Azure Data Bricks
5. App registration

- **Requirements :-**

1. Laptop and stable internet connection
2. Basics of SQL query
3. Azure Services

- **Scope :-**

1. Data Source and Dataset: The project focuses on analyzing data related to the Tokyo Olympics. This includes data from various sources such as athletes, teams, medals, etc.

2. Azure Services Utilization: The project leverages various Azure services like Data Factory, Data Lake Storage Gen2, Databricks, Synapse Analytics, and Power BI.

3. Data Ingestion and Storage: The project involves ingesting raw data from a GitHub repository and storing it in Azure Data Lake Storage Gen2. This allows for efficient data management and retrieval.

4. Data Transformation: Azure Databricks is used for data transformation. This includes tasks like removing duplicates, dropping unnecessary columns, and renaming columns.

5. Data Analysis: Azure Synapse Analytics is employed for data analysis. This involves writing SQL queries to extract insights from the transformed data.

6. Data Visualization: Power BI is used for creating visually appealing and informative reports and dashboards based on the analyzed data.

❖ **Objective :-**

1. Data Pipeline Establishment: Create a robust data pipeline for ingesting, storing, transforming, and analyzing Tokyo Olympic data.

2. Data Quality Assurance: Ensure the quality of data through transformation steps like removing duplicates and dropping unnecessary columns.

3. Insight Extraction: Extract meaningful insights from the data through SQL queries in Azure Synapse Analytics.

4. Interactive Reporting: Generate visually engaging reports and dashboards in Power BI to facilitate easy understanding and decision-making.

❖ **Problems faced:-**

1. Access Control Configuration: Setting up proper access control in Azure Data Lake Storage Gen2 required careful attention to ensure secure data handling.

2. Data Integration Complexities: Managing different data formats and ensuring seamless integration from a variety of sources presented initial challenges.

3. Data Quality Issues: Dealing with inconsistent data quality in the raw dataset necessitated careful handling to ensure accurate insights

4. Power BI Customization: Designing a visually appealing and informative dashboard in Power BI required experimentation and learning to leverage its full potential.

DESCRIPTION OF SERVICES

1. **Resource Group :**

A Resource Group in Azure is a logical container that holds related Azure resources (such as virtual machines, databases, and websites) together. It helps you manage, organize, and monitor these resources as a single entity. Resource Groups allow you to apply policies, access control, and billing to multiple resources at once.

2. **Azure Data Factory:**

Azure Data Factory is a cloud-based data integration service that allows you to create, schedule, and manage data pipelines. It enables the extraction, transformation, and loading (ETL) of data from various sources into data stores and analytics platforms, both in the cloud and on-premises. Data Factory supports a wide range of data connectors and provides tools for data orchestration and monitoring.

3. **Azure Synapse Analytics:**

Azure Synapse Analytics is an analytics platform that integrates enterprise data warehousing, big data processing, and data integration. It allows you to analyze large volumes of data using both serverless and provisioned resources. It provides features for data integration, data warehousing, big data processing, and machine learning capabilities.

4. **Azure Databricks:**

Azure Databricks is an Apache Spark-based analytics platform optimized for Azure. It provides a collaborative environment for building, training, and deploying machine learning models and performing data analysis at scale. It includes features for data engineering, data science, and business intelligence.

5. **App Registration:**

App Registration in Azure Active Directory (Azure AD) is a way to define the characteristics and permissions of an application that wants to integrate with Azure AD. It involves creating an identity for the application, configuring authentication settings, and specifying which resources the application can access. This is a fundamental step in enabling secure access to various Azure services and APIs.

USAGE OF SERVICES

Azure Data Factory:

Usage: Data Factory was used for creating data pipelines to ingest, transform, and load data from different sources into Azure Data Lake Storage Gen2.

Reasoning: Data Factory is designed for building and managing data pipelines. It supports data movement and transformation activities, making it an ideal choice for handling data workflows in this project.

Azure Data Lake Storage Gen2:

Usage: Data Lake Storage Gen2 was used as the target storage location for the raw and transformed data.

Reasoning: Data Lake Storage Gen2 is optimized for big data analytics workloads and provides a hierarchical namespace, which makes it easier to organize and manage large amounts of data.

Azure Databricks:

Usage: Databricks was used for data transformation tasks, including removing duplicates, dropping unnecessary columns, and performing other data preparation steps.

Reasoning: Databricks is a powerful tool for big data processing and analytics. It provides a collaborative environment for data scientists and engineers to work on data transformation tasks at scale.

Azure Synapse Analytics:

Usage: Synapse Analytics was used for data analysis using SQL queries to extract insights from the transformed data.

Reasoning: Azure Synapse Analytics is an enterprise-grade analytics platform that integrates data warehousing, big data processing, and data integration. It's well-suited for performing complex analytical queries on large datasets.

Azure Power BI:

Usage: Power BI was used for data visualization, creating visually appealing reports and dashboards based on the analyzed data.

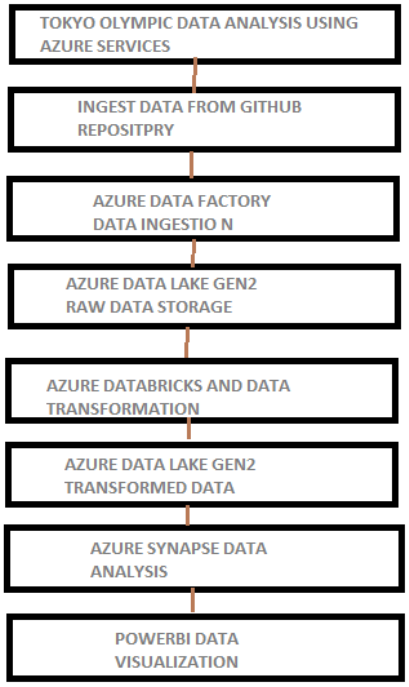
Reasoning: Power BI is a leading business analytics tool that enables users to visualize and share insights from their data. It's highly effective for creating interactive and informative reports.

App Registration (Azure Active Directory):

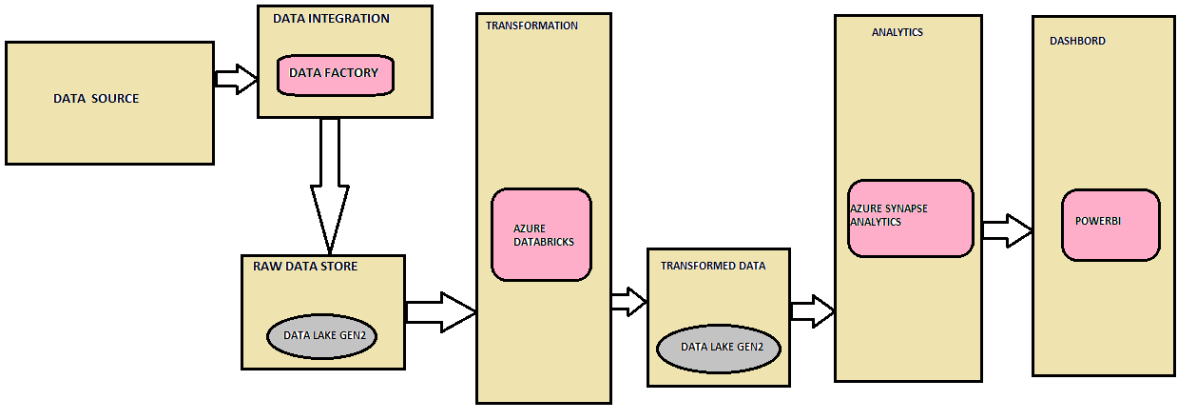
Usage: App Registration was used to define the characteristics and permissions of an application that integrates with Azure AD. It's a crucial step in enabling secure access to various Azure services and APIs.

Reasoning: App Registration ensures that only authorized applications have access to the resources in Azure AD, providing an additional layer of security for the project.

DATA FLOW

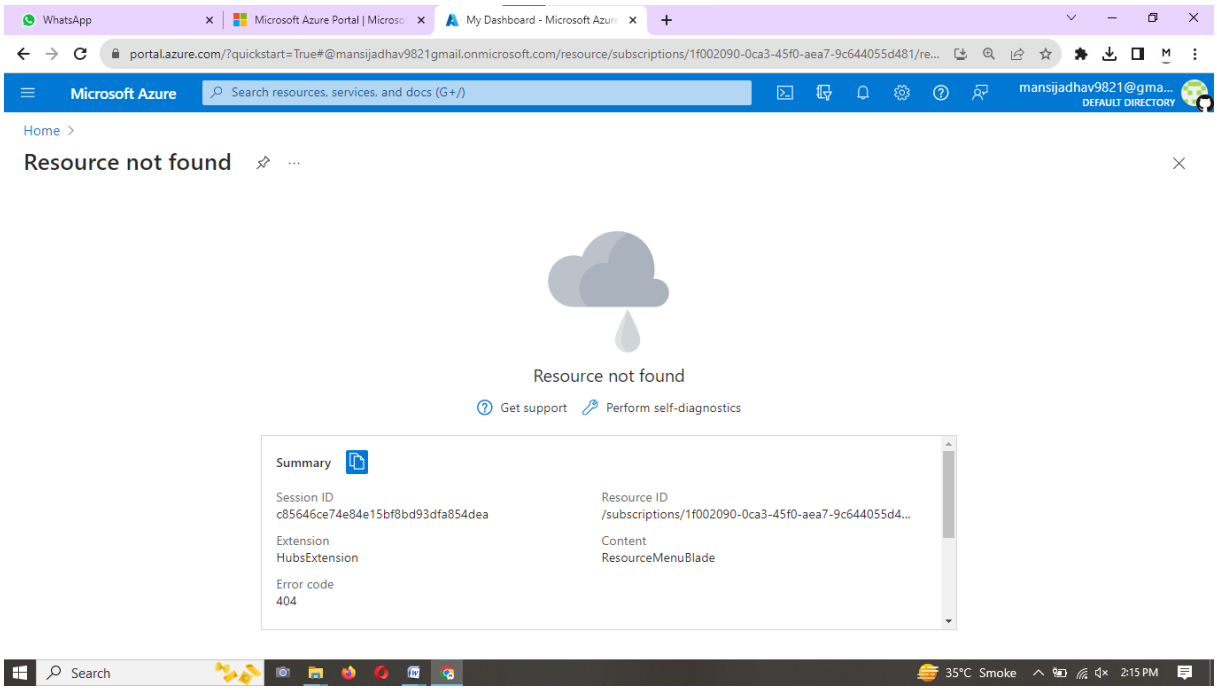

TOKYO OLYMPIC DATA ANALYSIS USING AZURE SERVICES
INGEST DATA FROM GITHUB REPOSITPRY
AZURE DATA FACTORY DATA INGESTIO N
AZURE DATA LAKE GEN2 RAW DATA STORAGE
AZURE DATABRICKS AND DATA TRANSFORMATION
AZURE DATA LAKE GEN2 TRANSFORMED DATA
AZURE SYNAPSE DATA ANALYSIS
POWERBI DATA VISUALIZATION

DIAGRAM

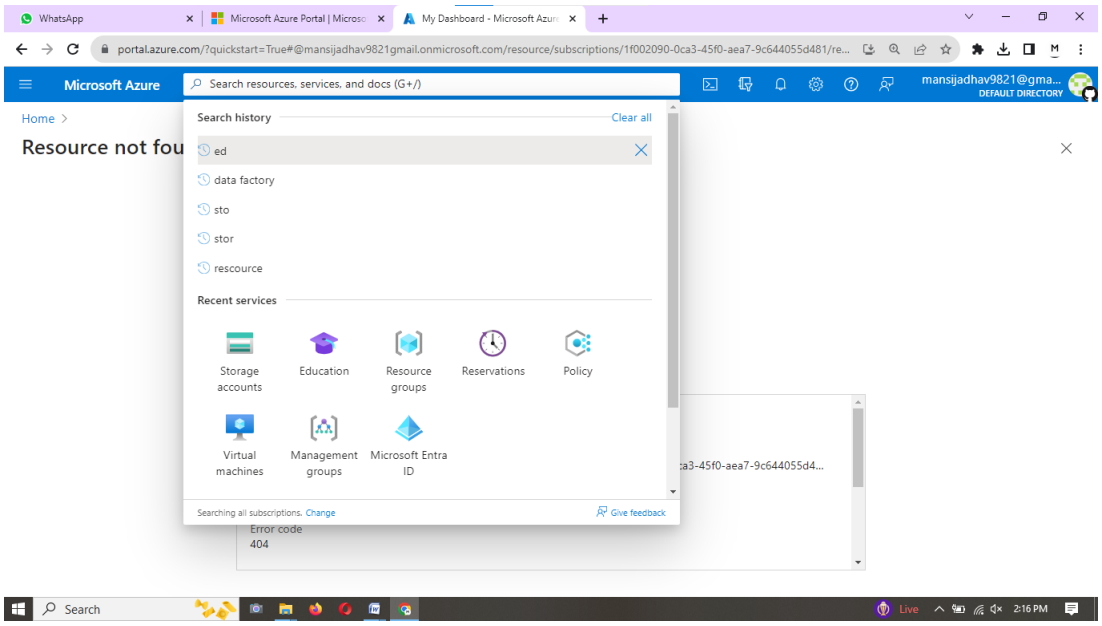

DATA SOURCE
DATA INTEGRATION
DATA FACTORY
RAW DATA STORE
DATA LAKE GEN2
TRANSFORMATION
AZURE DATABRICKS
TRANSFORMED DATA
DATA LAKE GEN2
ANALYTICS
AZURE SYNAPSE ANALYTICS
DASHBORD
POWERBI

PROJECT DESIGN

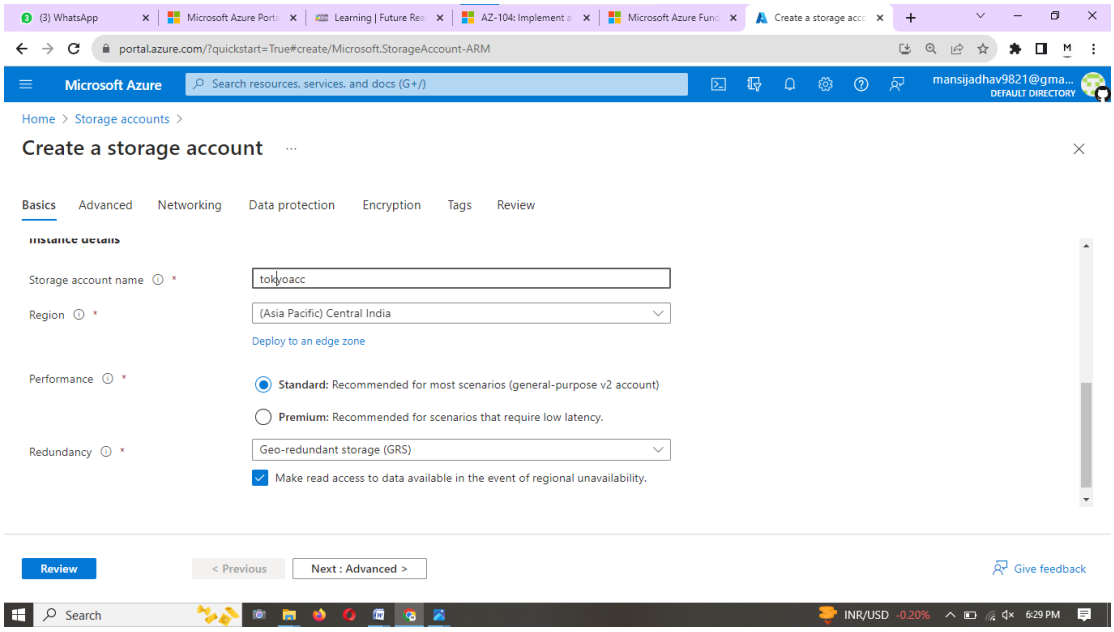
So first thing we will open azure portal and login.Open azure portal https://azure.microsoft.com/en-in/get-started/azure-portal and once you login in it will be as shown below.



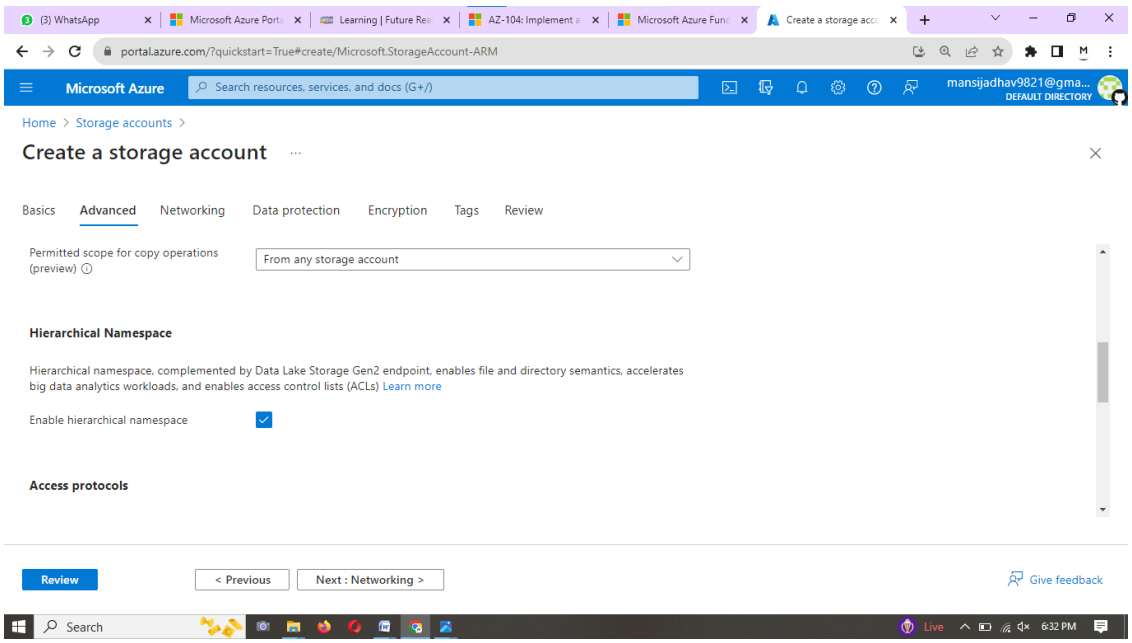
So First we have to do is ingest our data to data external. So we need two things one is Data Factory to create simple pipeline and second is Gen2 Which is where we will store raw data so lets start with it so we will these two things individually so first we will create Storage Account So in search bar search storage account so click on it and create account.



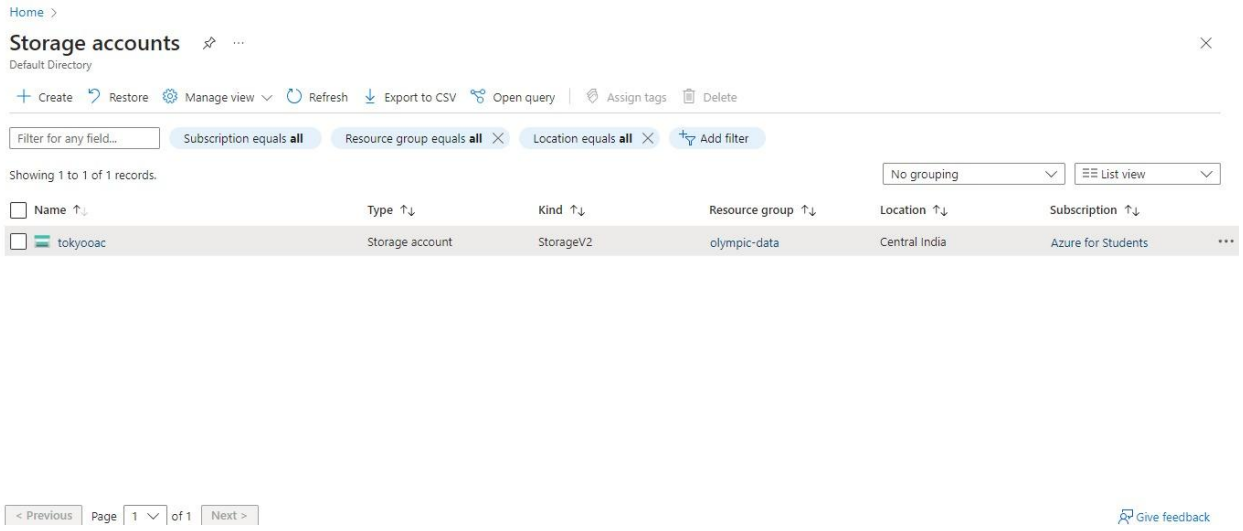
Storage account is created. Name of the storage account should be globally unique. Here performance is Standard as if is for General purpose v2 account and redundancy is Geo-Redundant Storage means if you want to replicate your data across different data centers or regions.



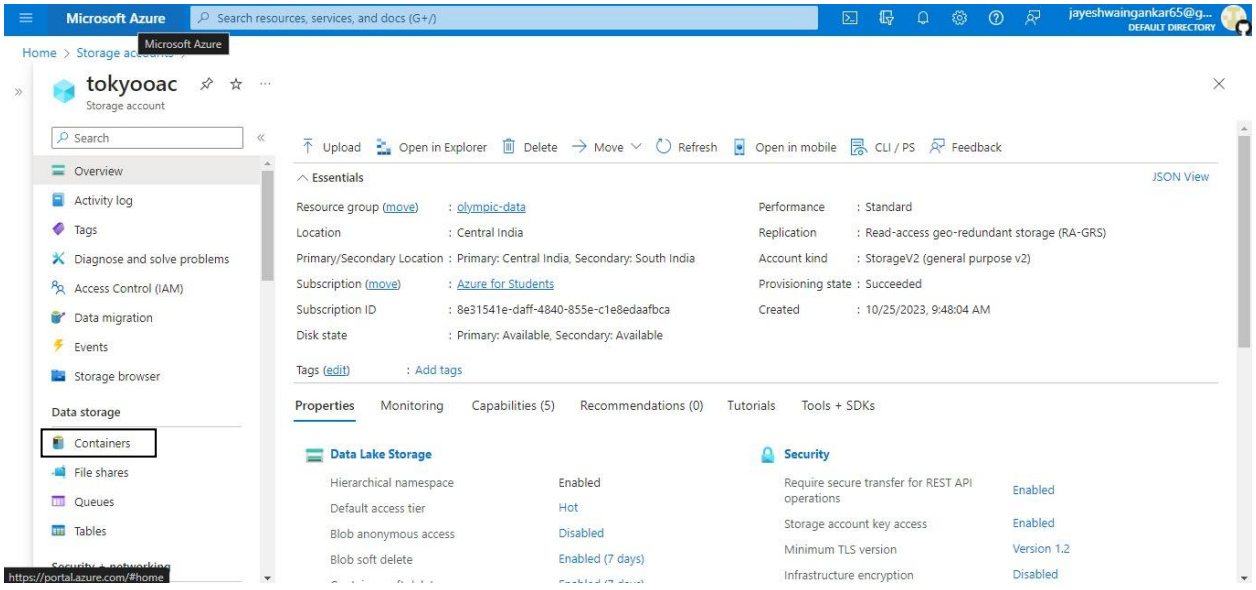
In advanced we will enable hierarchical Namespace means against all of the files that you store inside the container of your storage account will be available in hierarchical format the way it stores any data to any abject storage such as S3 or Google or just Azure Block Storage they store the entire data as an object but once you enable it you can access your data just like simple directory the way you do it on local computer so it is important to select the hierarchical namespace when we are creating storage account. And then review and create



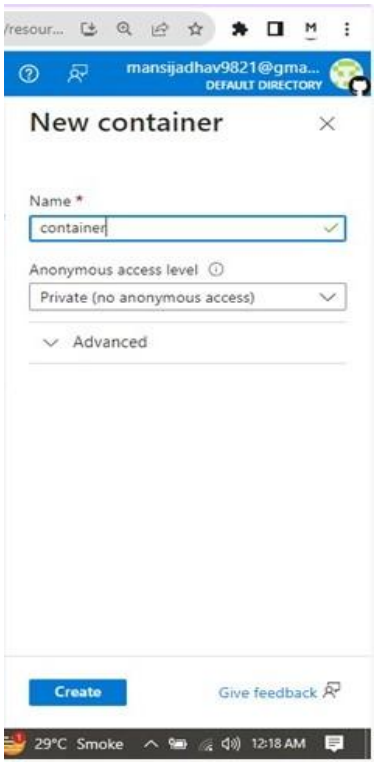
Storage account created



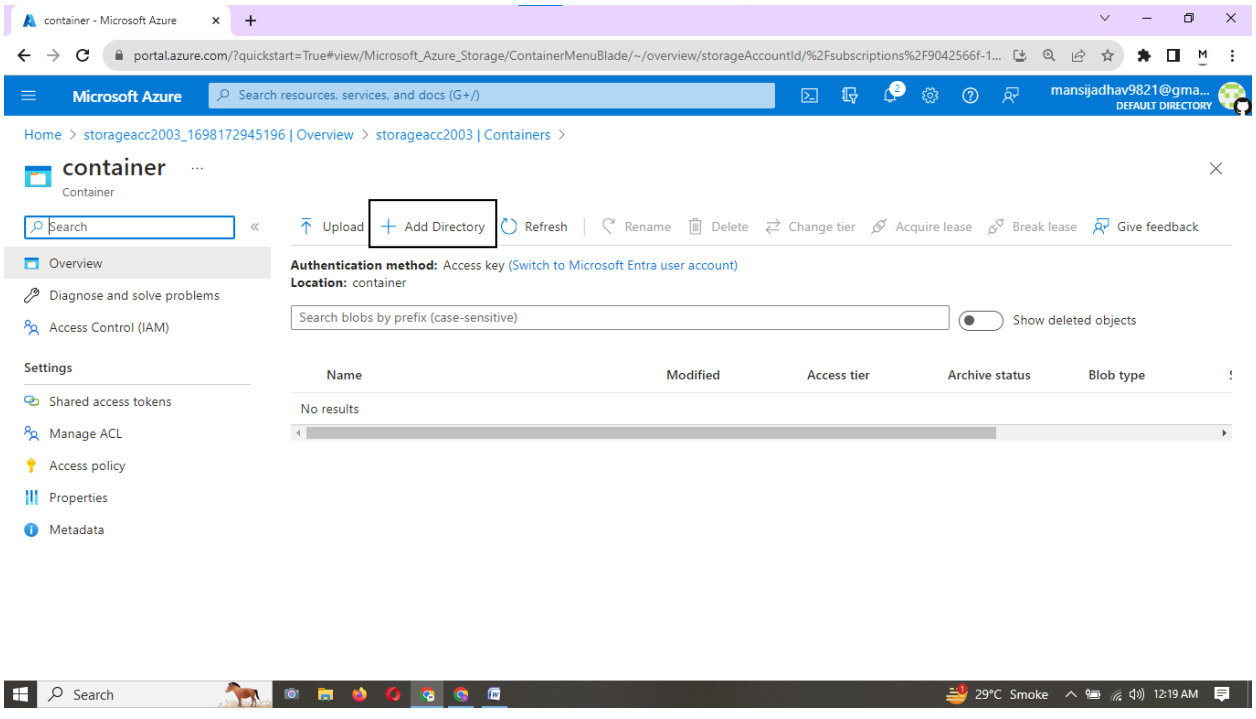
So these is the panel of storage account .here in overview we have all details about storage account its rescource group, subscription and many more. In azure storage we have container where you can store data as an object. Now we will create container in storage account.



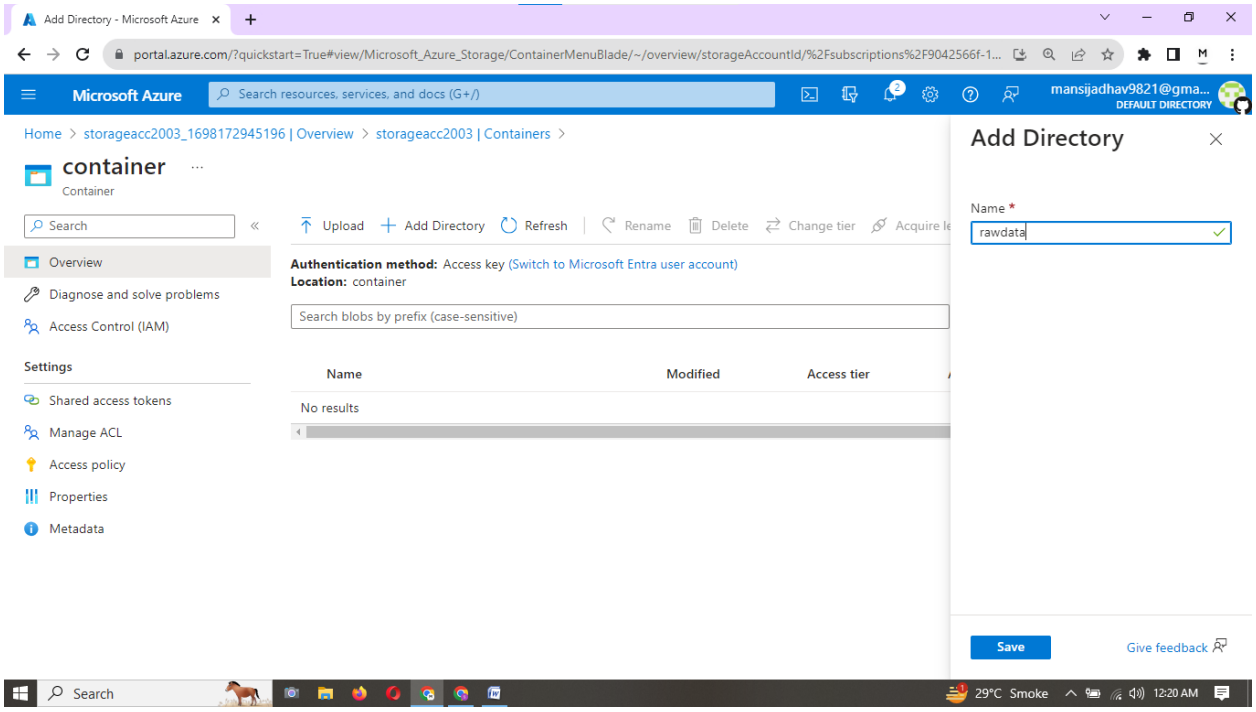
Click on container and write name of the container and then create it



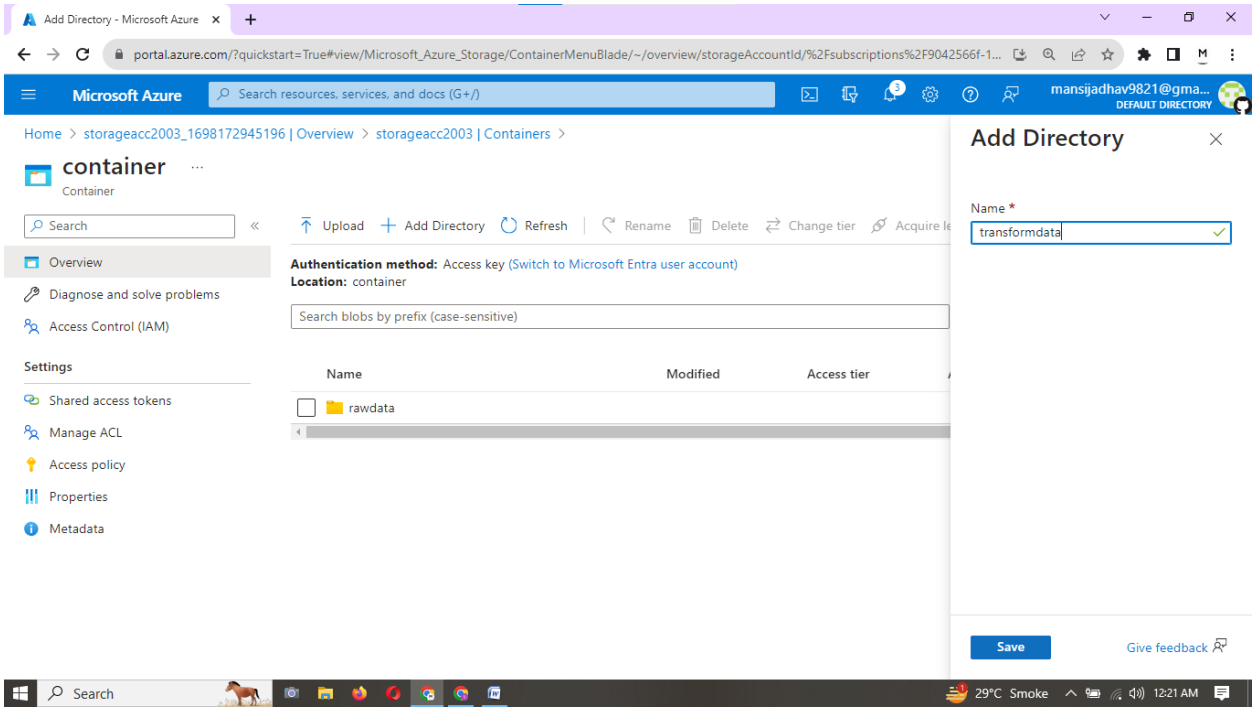
Click on Container and there we will create 2 folders so we will click on add directory .Now we will add 2 directory for raw where there will be raw data as it is from data source that we extract and transformed data where we will have transform data using Database transformation



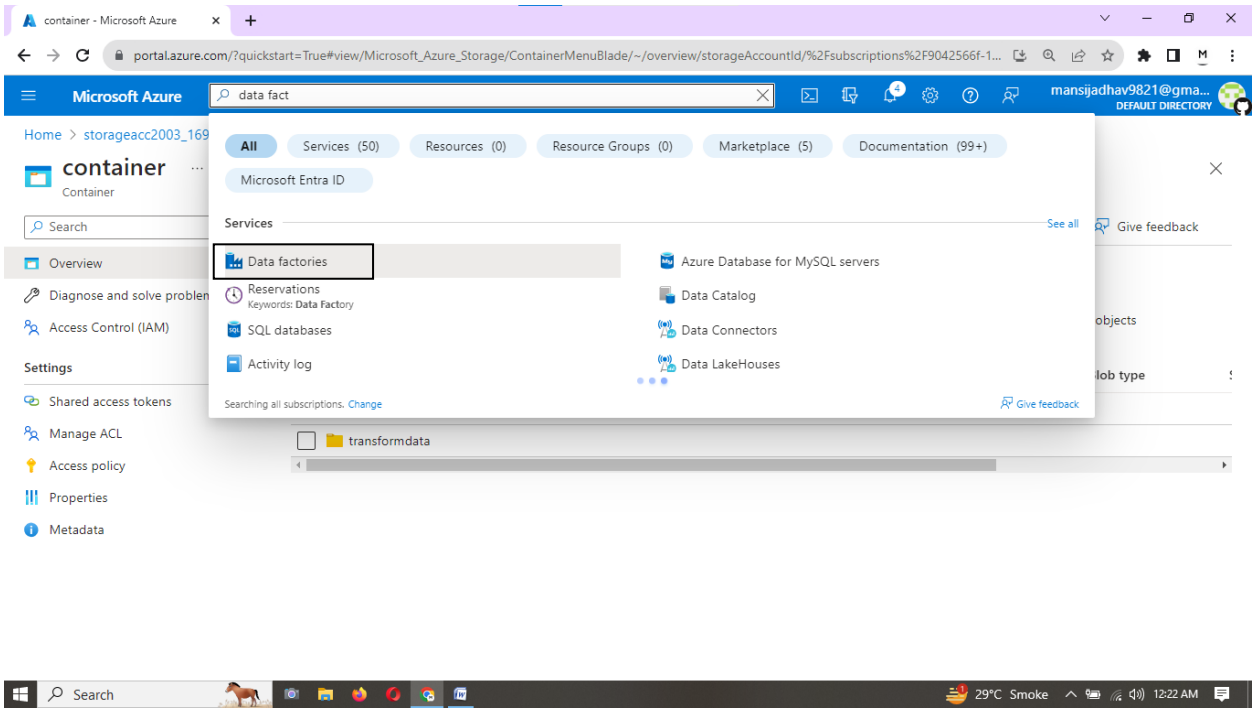
Adding of raw data



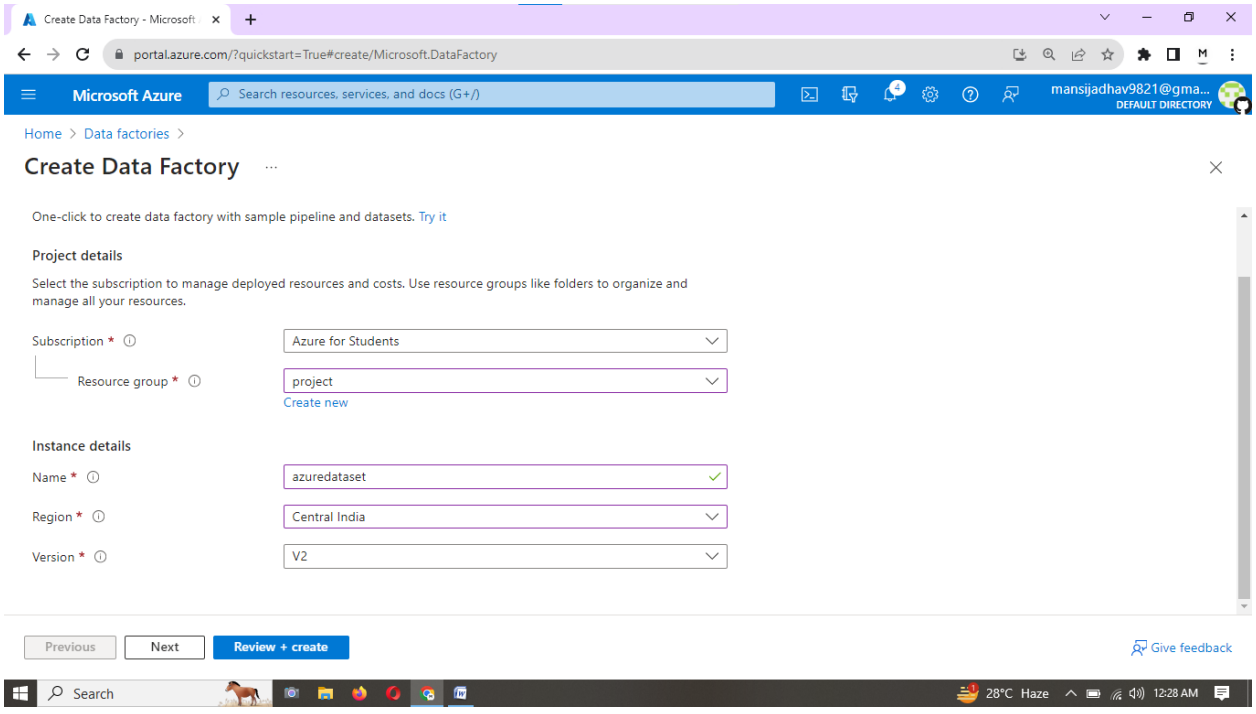
For transformed data.



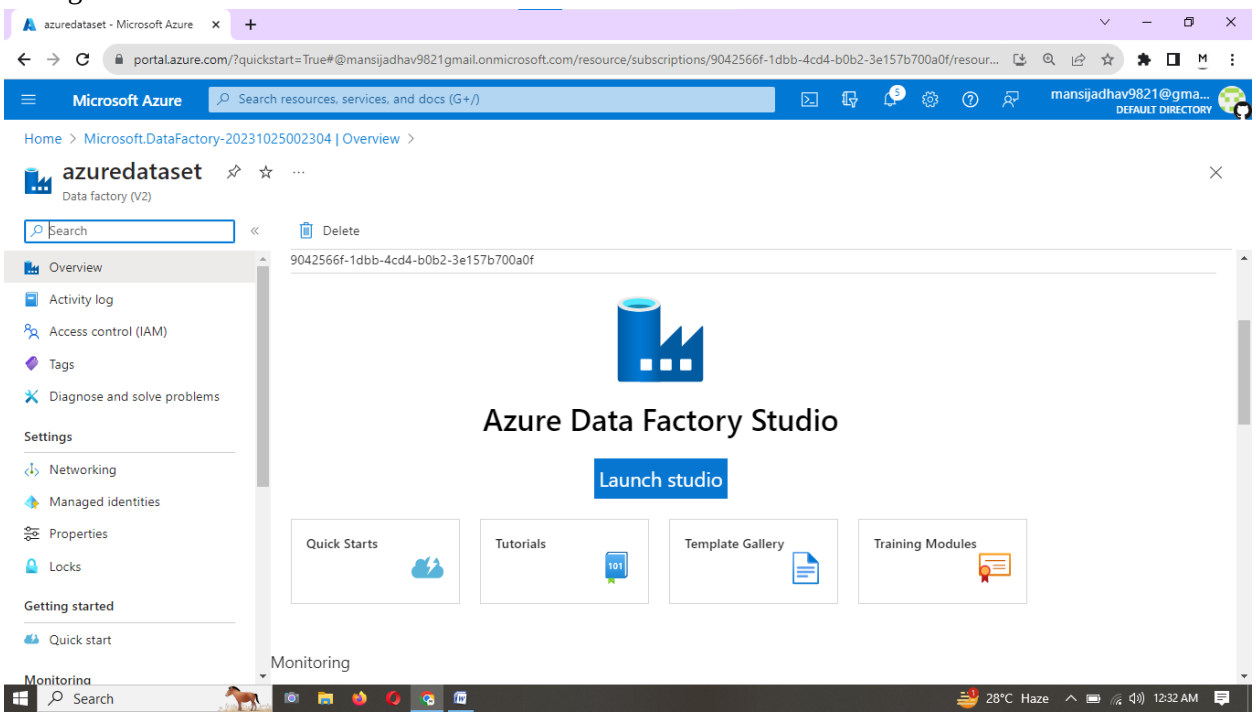
Now we have set our Azure data Lake now we have to copy the data from data source and put data on Target location and for that we need to use Azure Data Factory so search for Data Factory



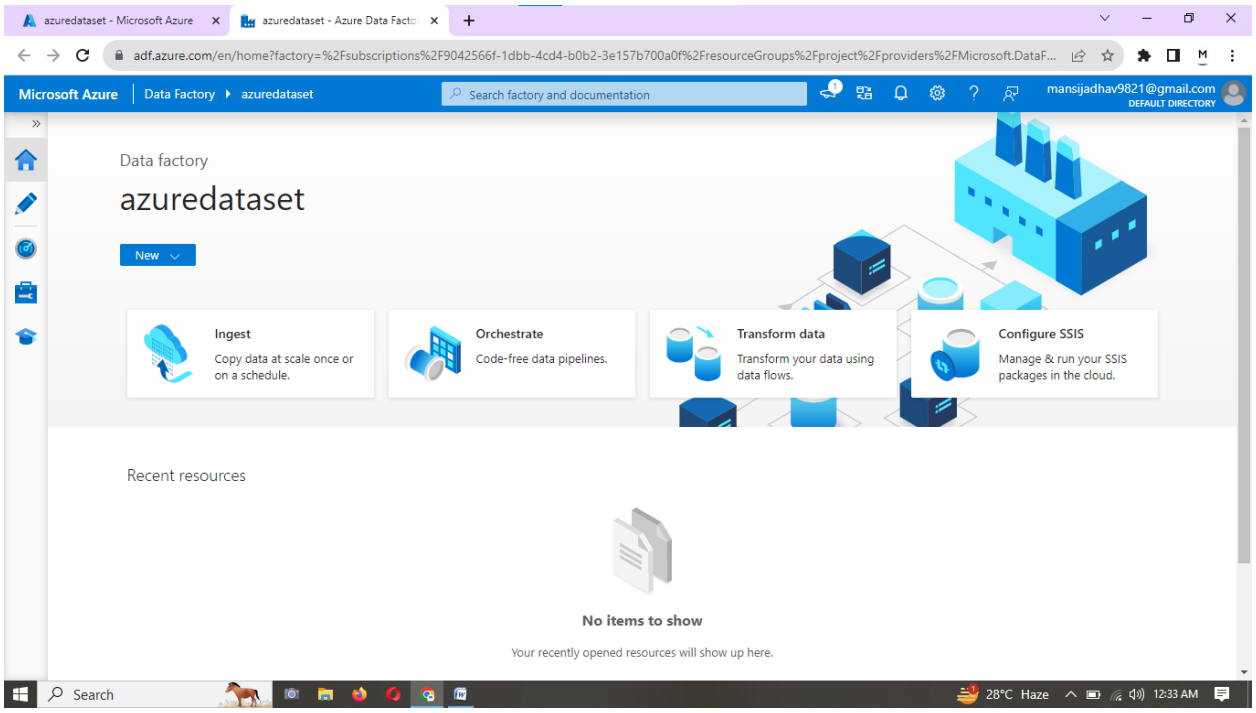
This is the console and click on Create and it will redirect to page. Now rescource group will be same and then review and create



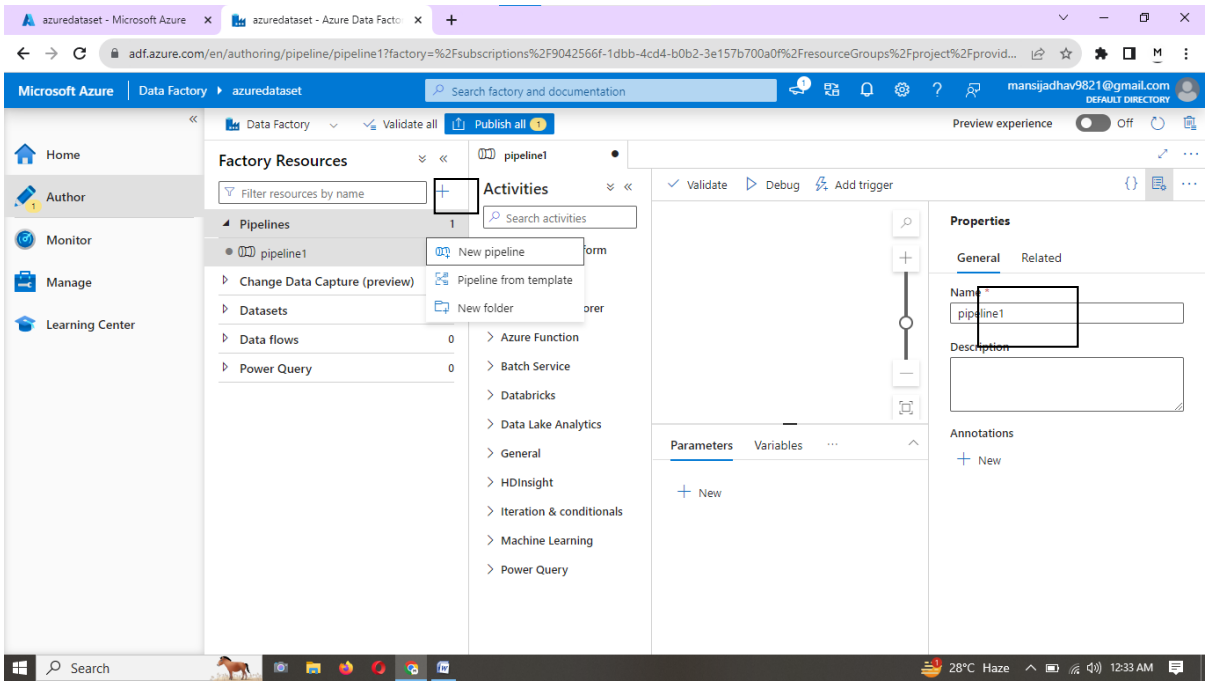
Then go to rescource and Launch Studio



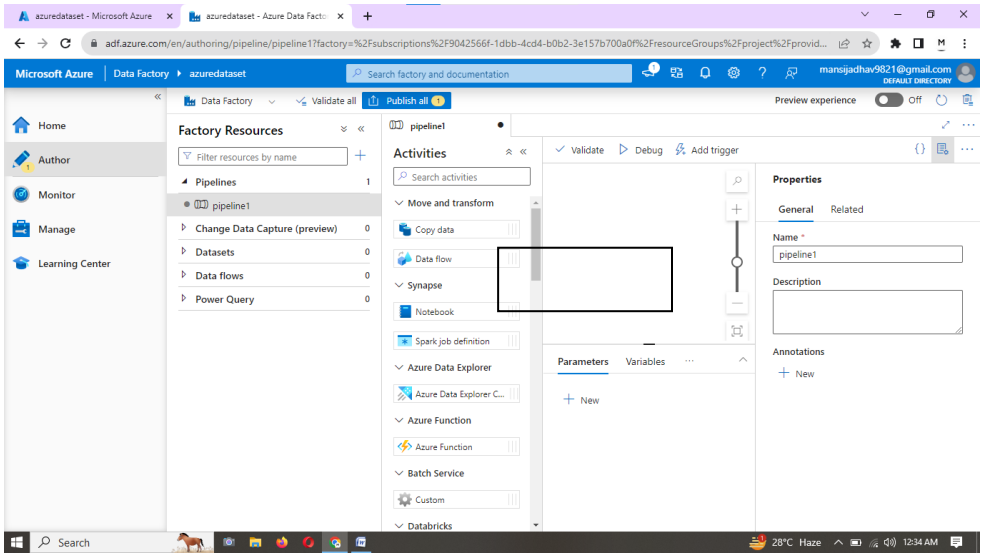
This is the Azure Data set ie Data factory panel. Here we can create Pipeline to extract data from different sources and upload that data on their Target Location so we have ingest option and other options also. In home all data factory are shown .



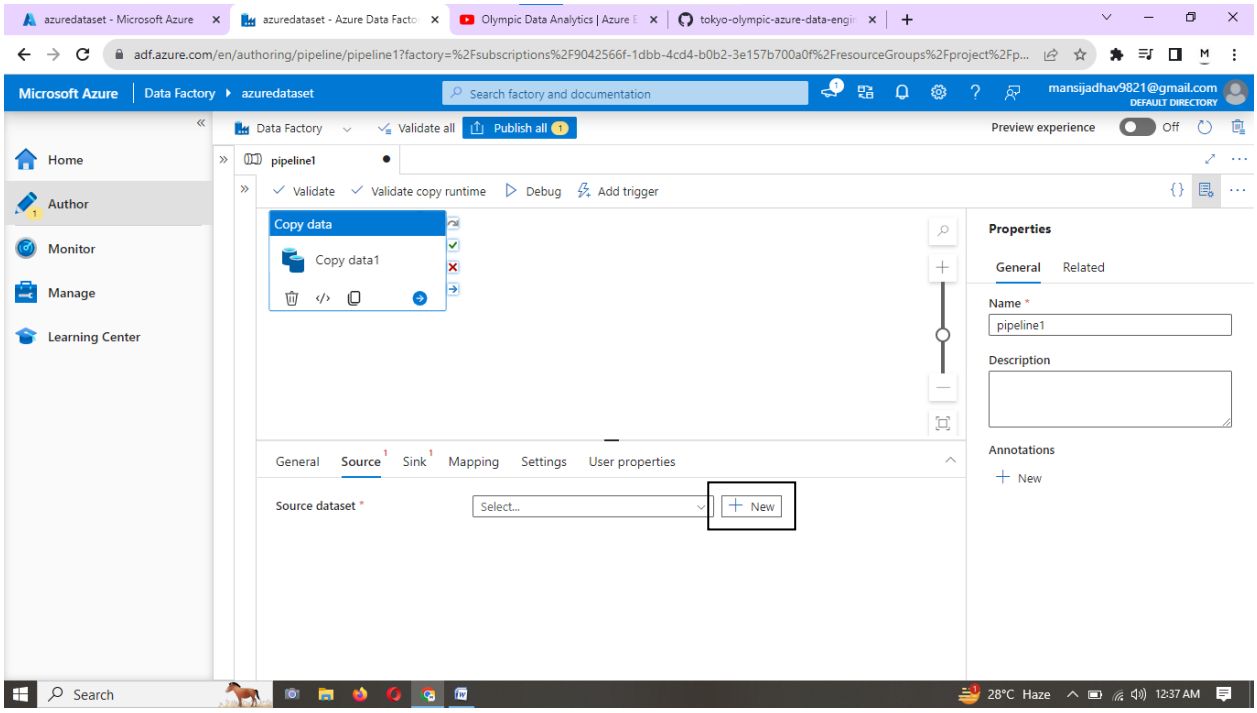
So we will click on Author where we actually create pipelines. Then click on plus sign and then select pipeline and we will create simple pipeline. Give the name for pipeline



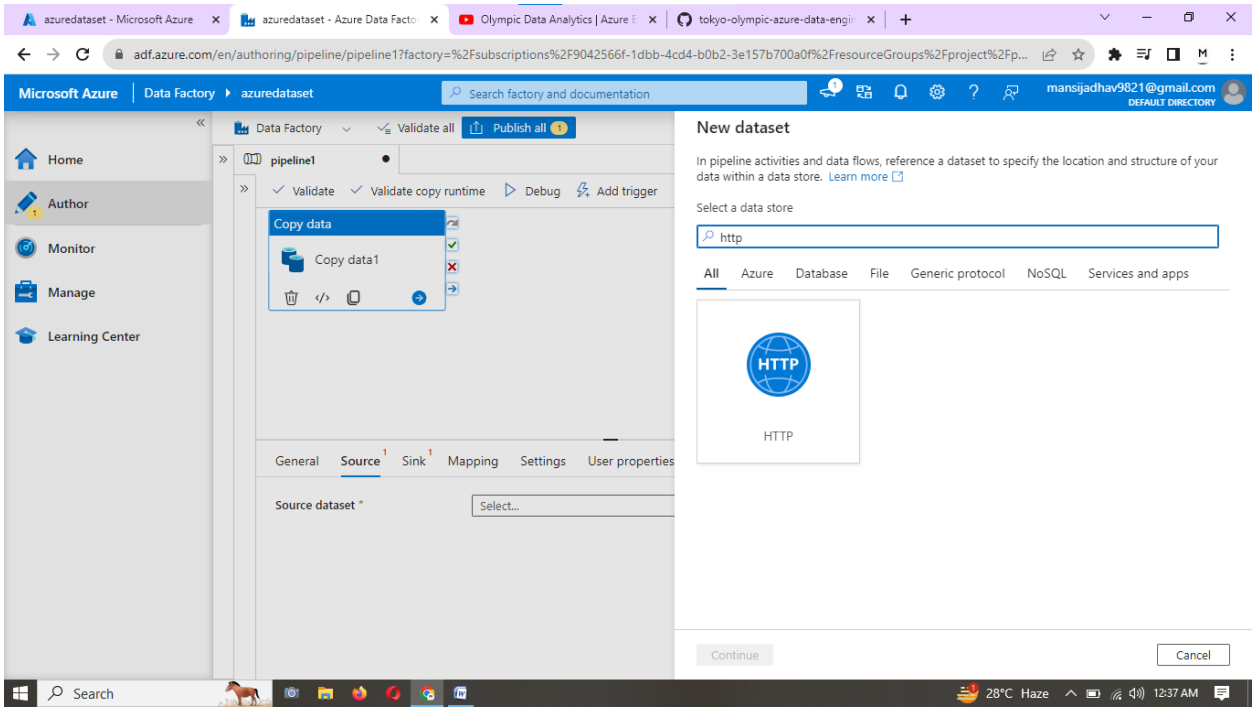
So this is the pipeline window. here we have activities and there we will select move and transform as we have to copy the data and drag copy data in window. So first we will copy the data from our API or this data source to our location which is data storage this is our copy activity inside the copy activity we want to give the source where our data is actually stored and then sink where we want to load our data so lets do it. All the data is uploaded on github so you can refer that. Using Github repository as our data source we will extract data from repository and load our data on our azure location so we have five files available.



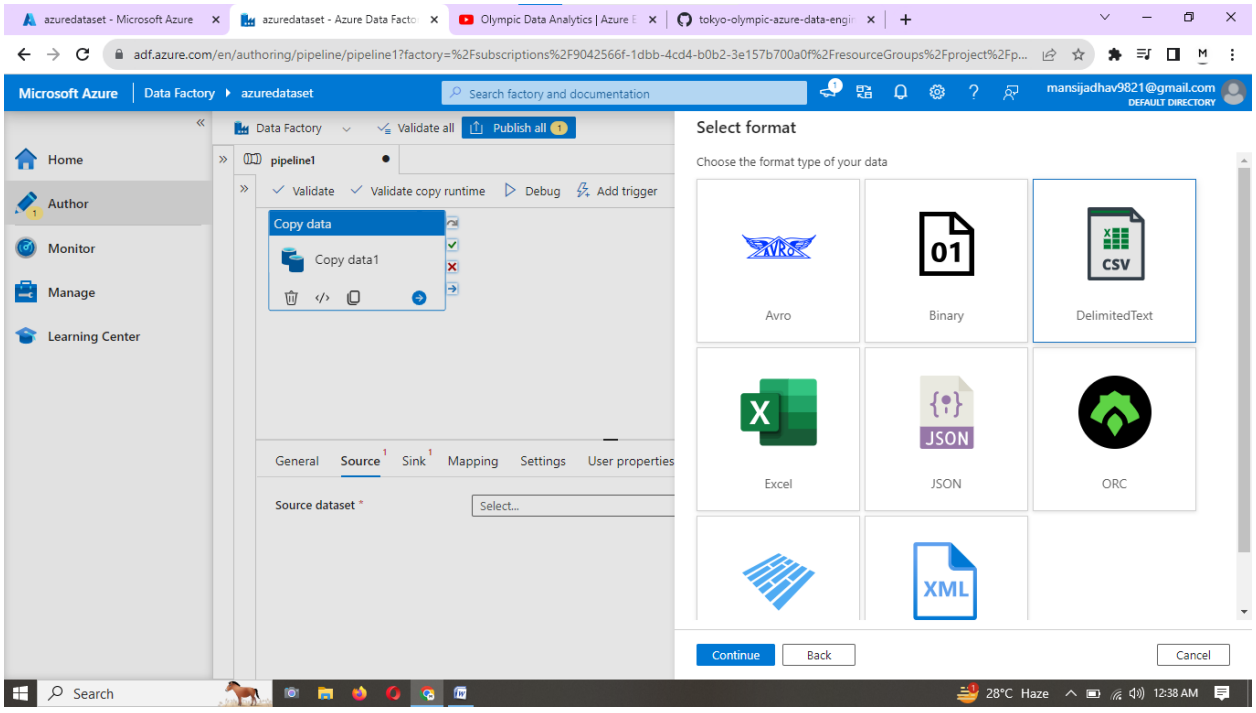
Now here select Source and click on New



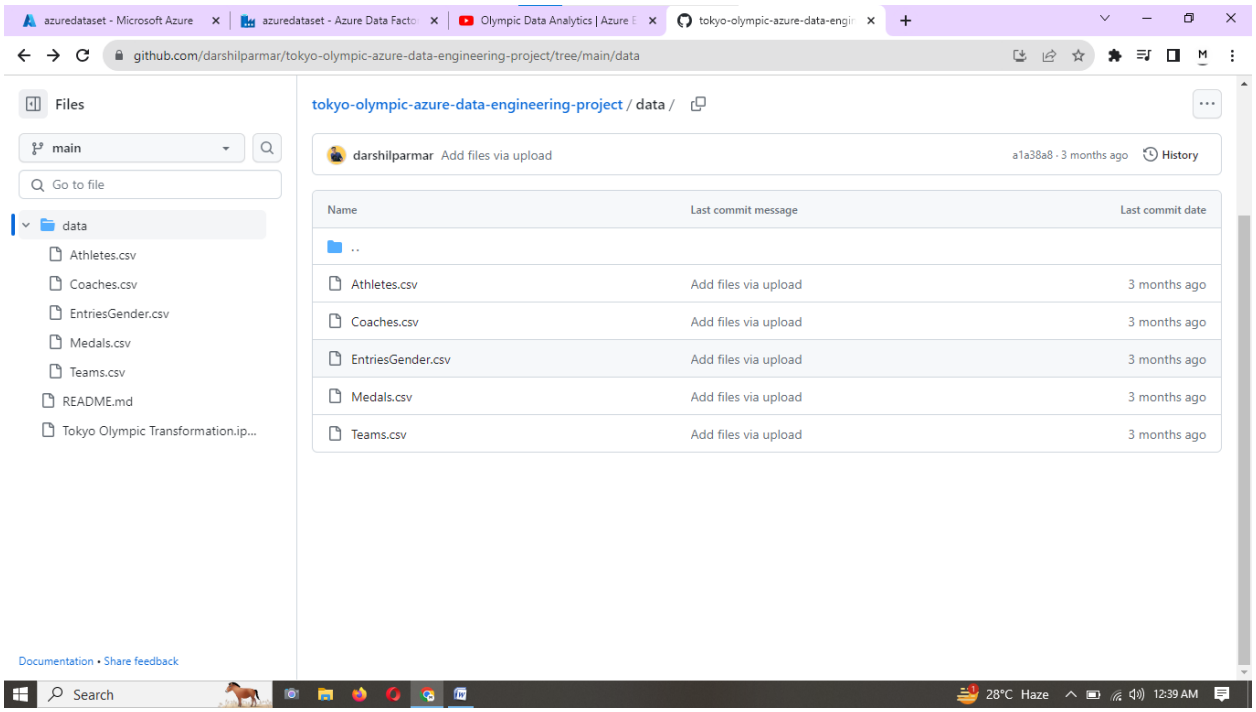
Here we select dataset as HTTP. We have multiple datasets to access data we select HTTP because this raw data is accessed through HTTP server.



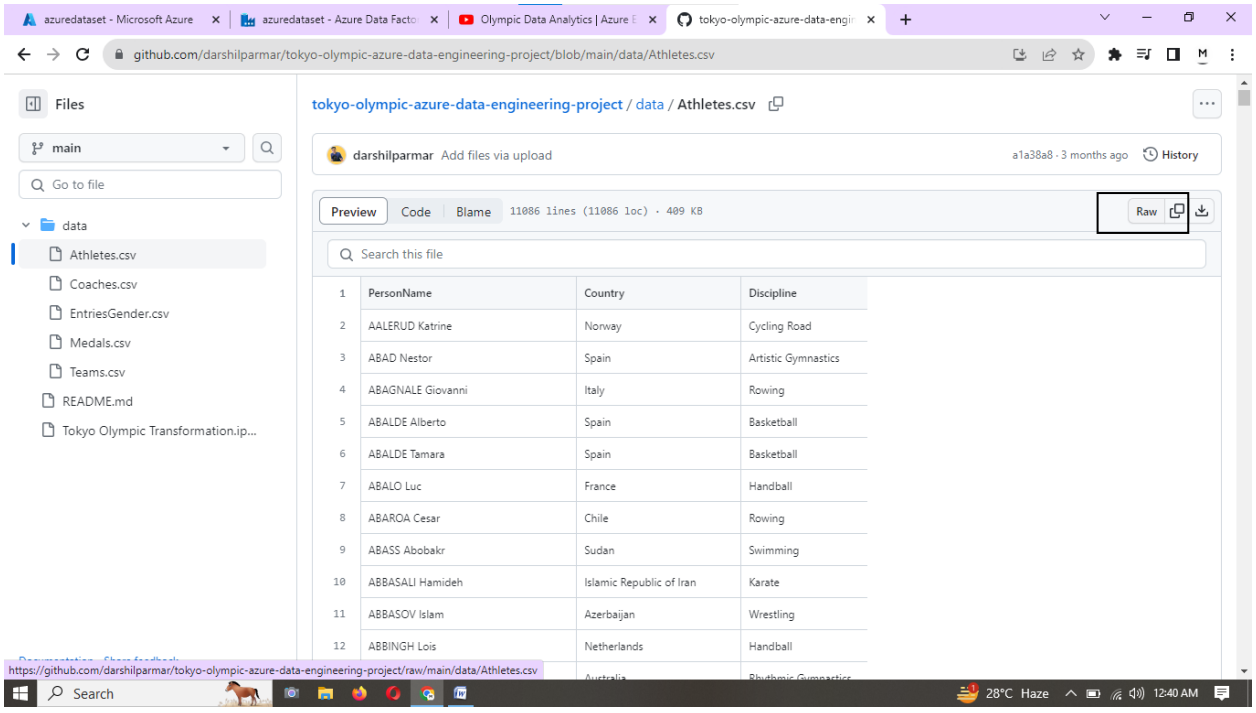
The file format is csv file so select DelemittedText and continue



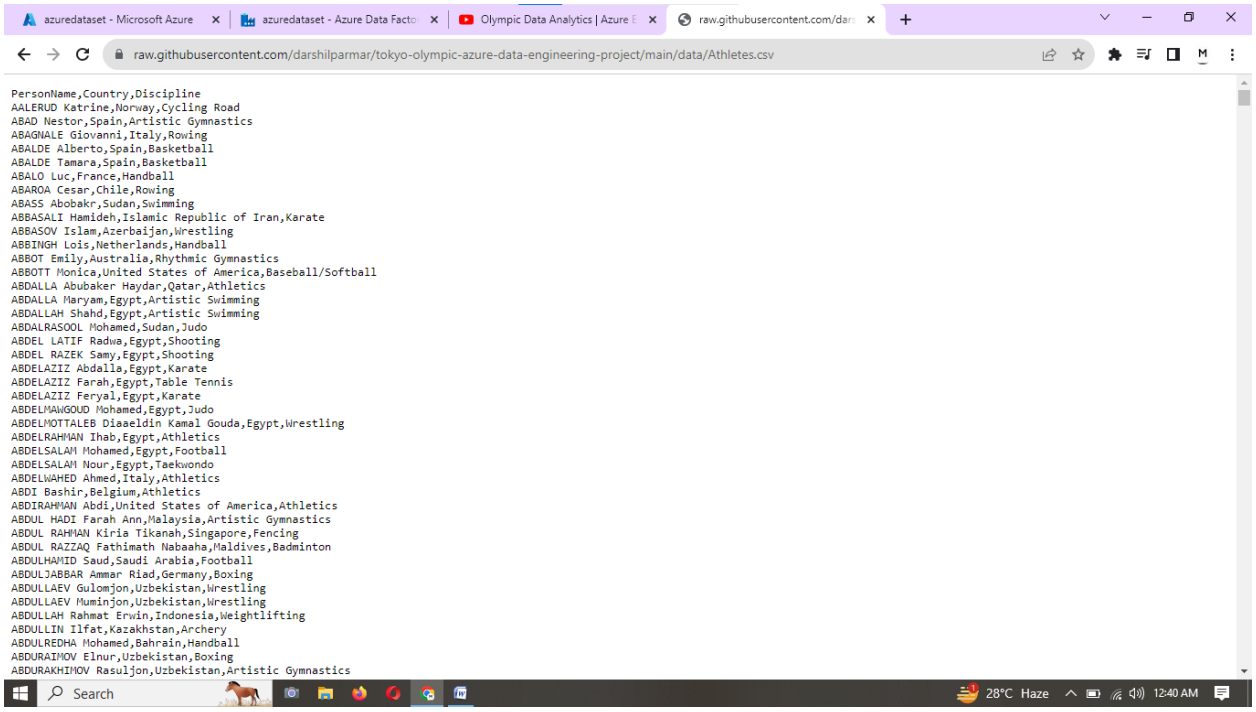
Now this is the GitHub account where we have our dataset and we have 5 files so we have to copy all the data of 5 files so first we will do for Athletes.



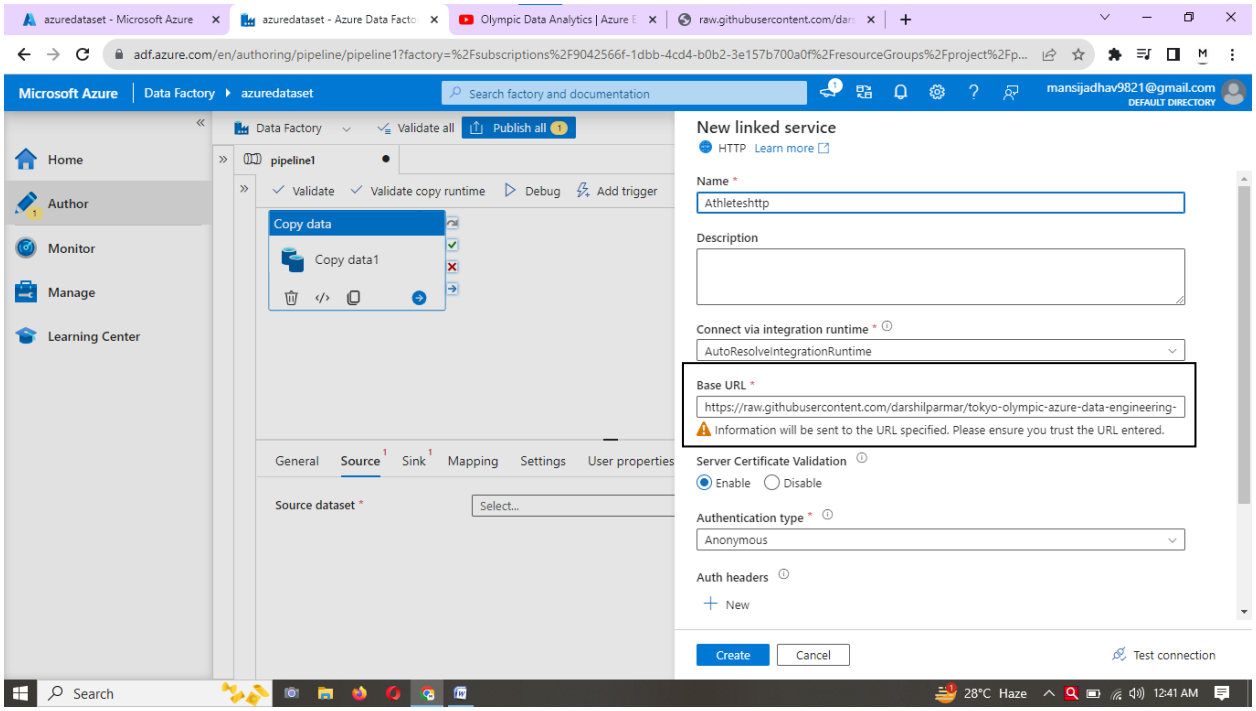
So click on Athletes and then at the right bottom up we have raw so click on it



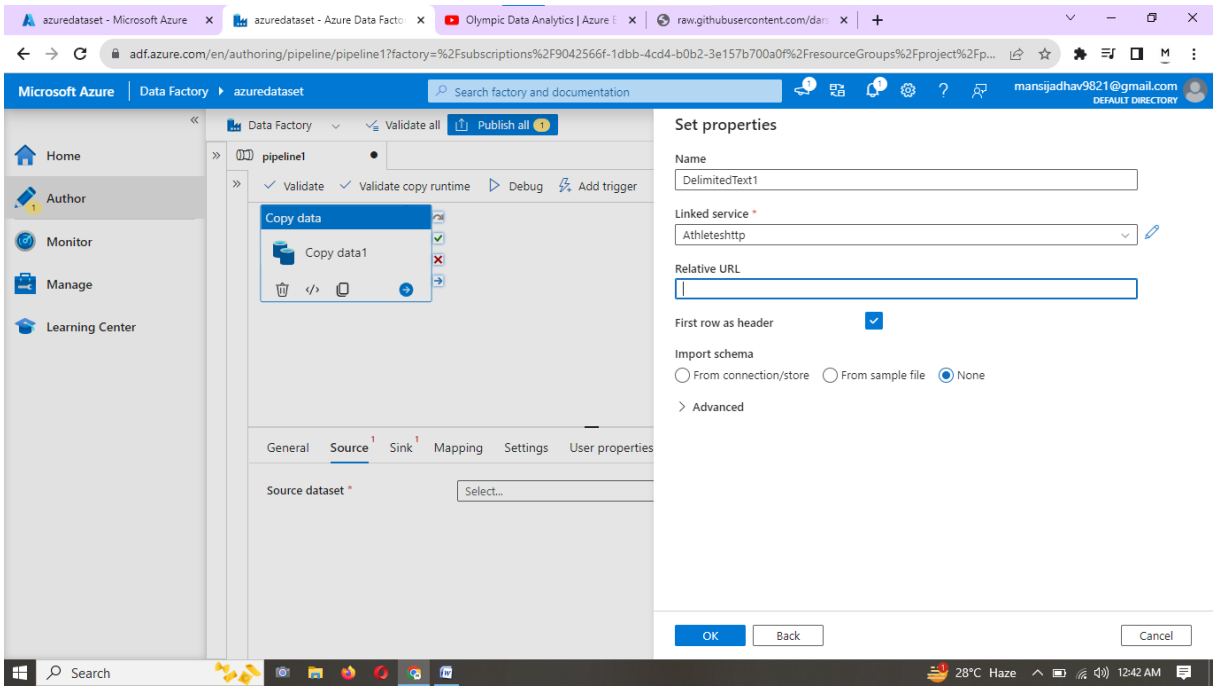
After clicking on it we get raw data so the URL is the base link. So copy that link and paste it in Base link



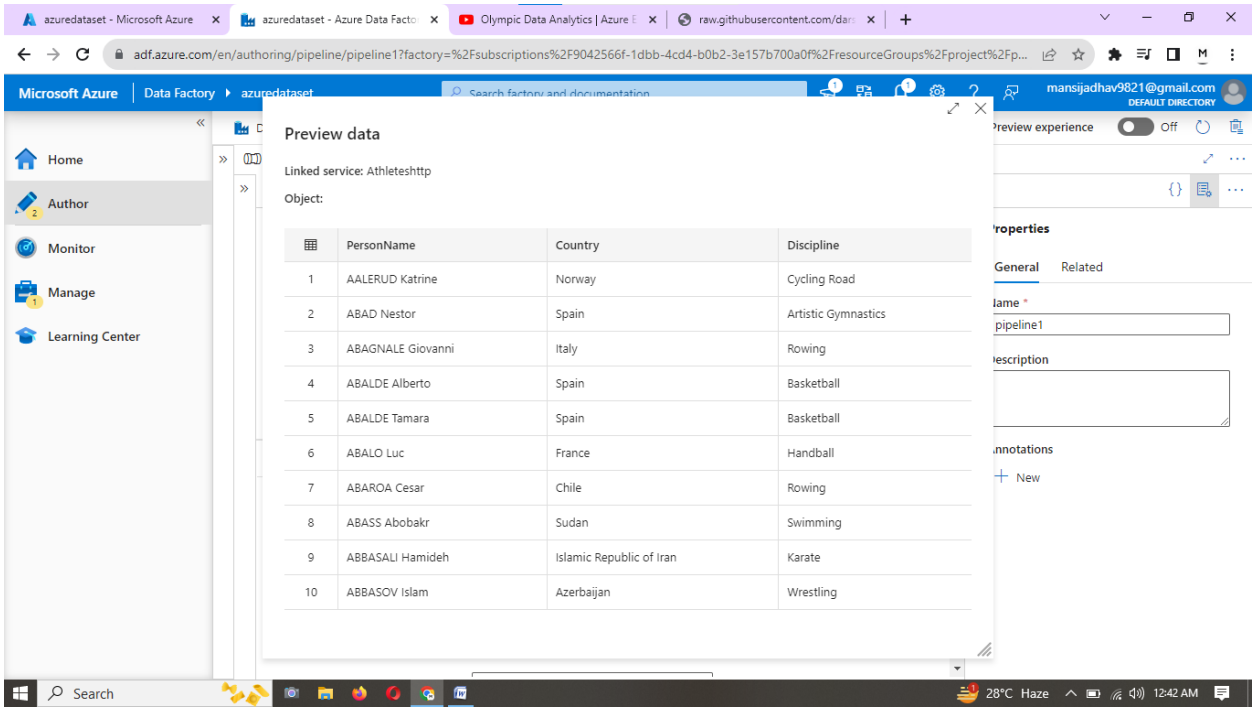
After copy create the linked service



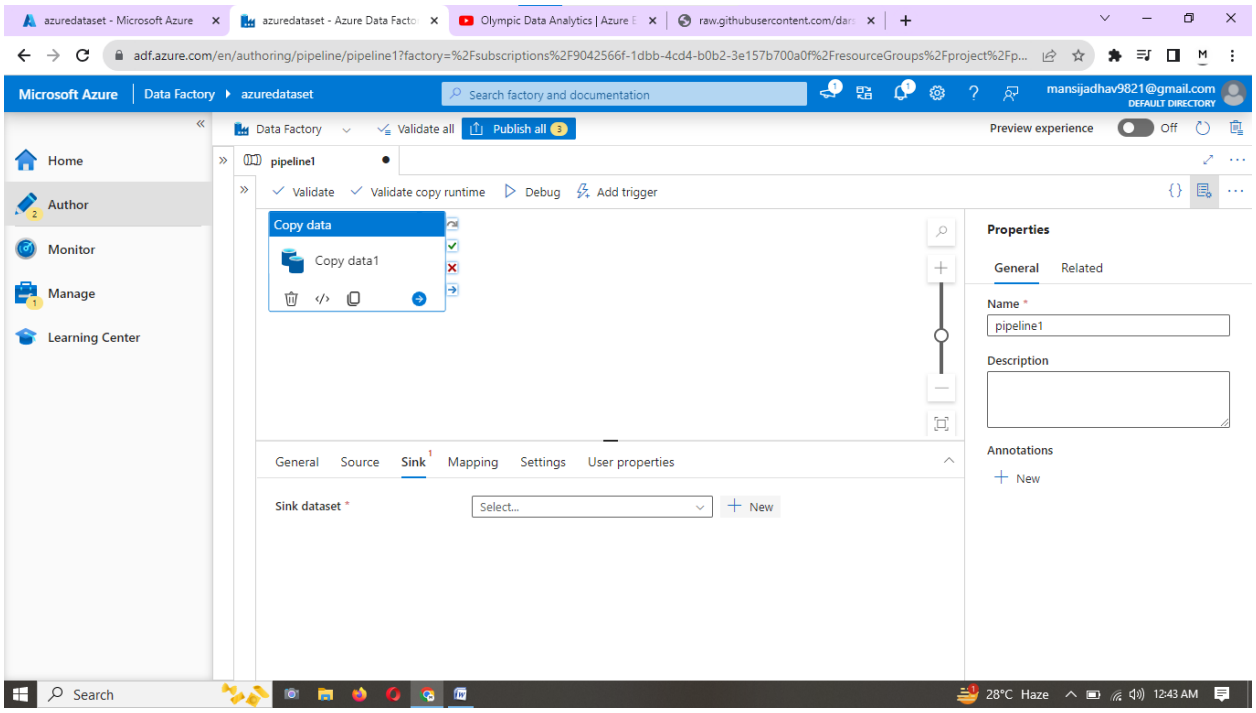
Now we will give the path for data so click on ok.



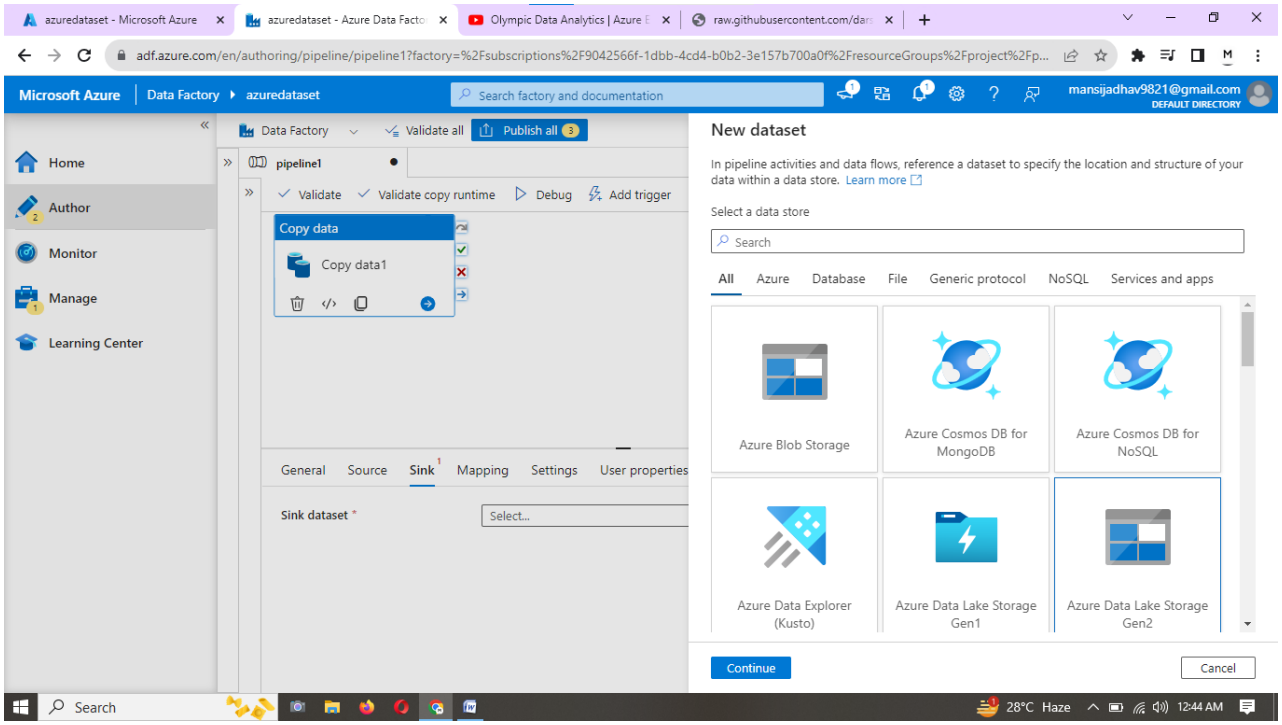
Now click on preview data if the data occurs means you are on right track



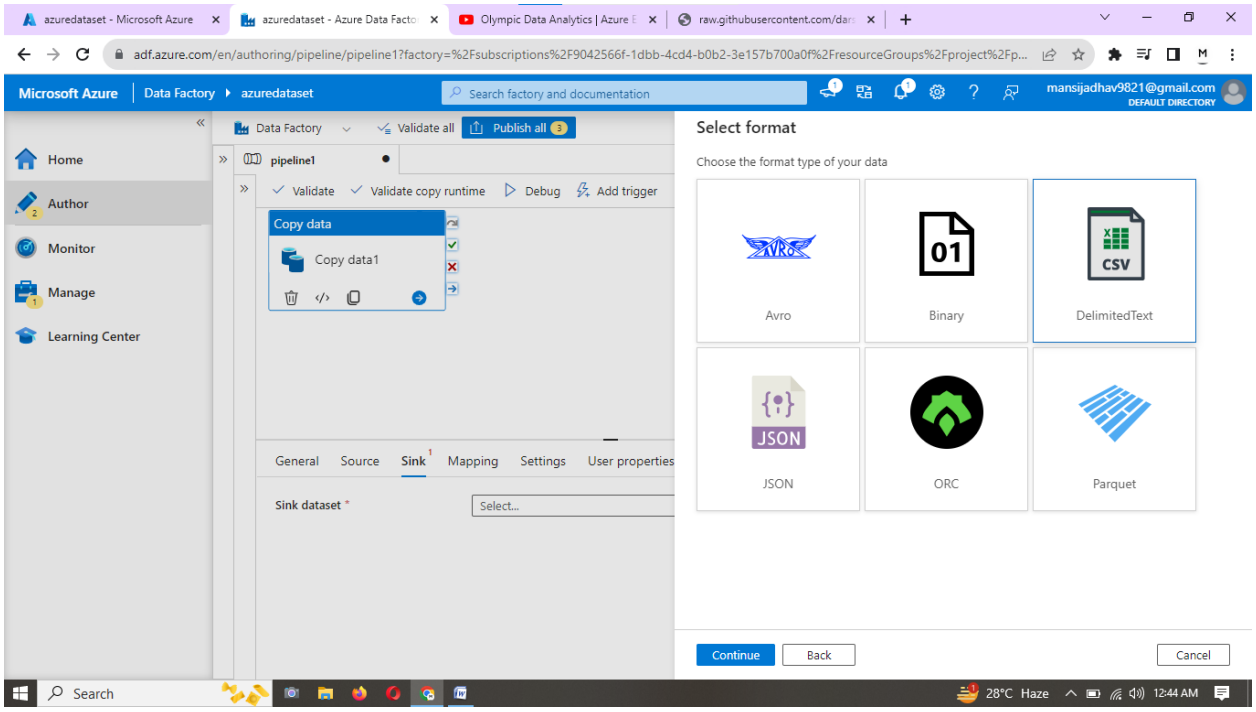
Now we will sink so click on new



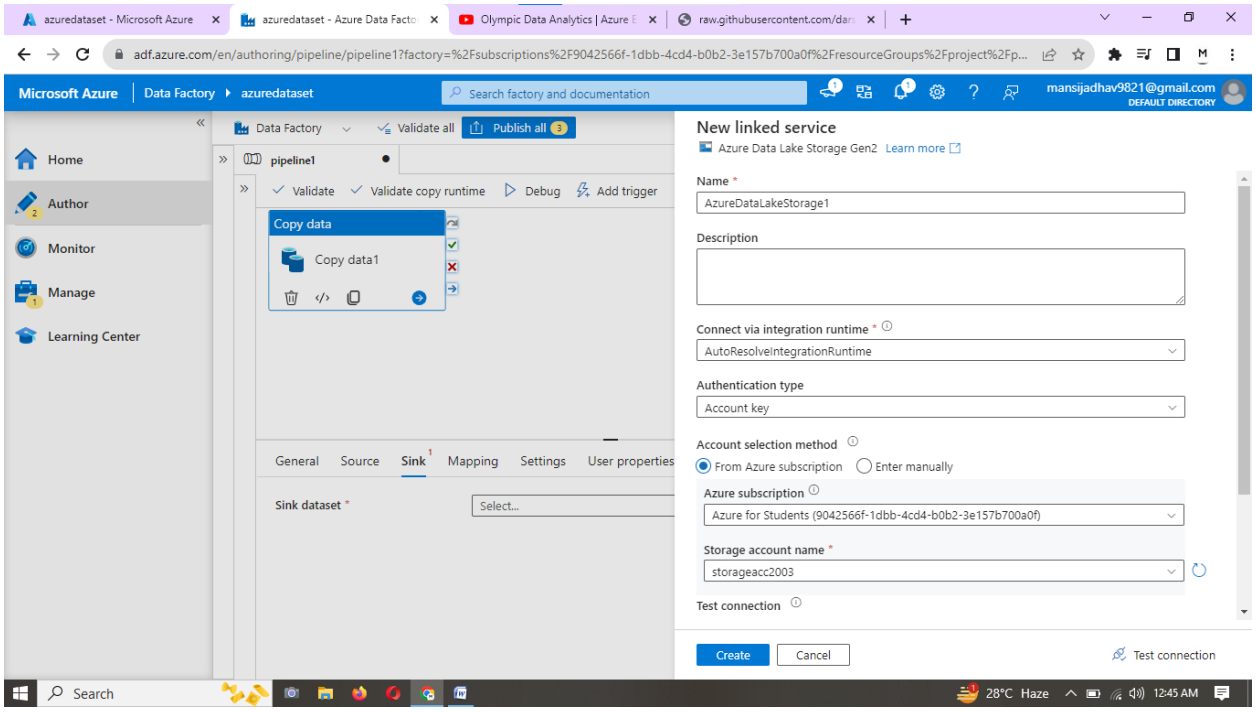
And here we will select Azure Data Lake Storage Gen2

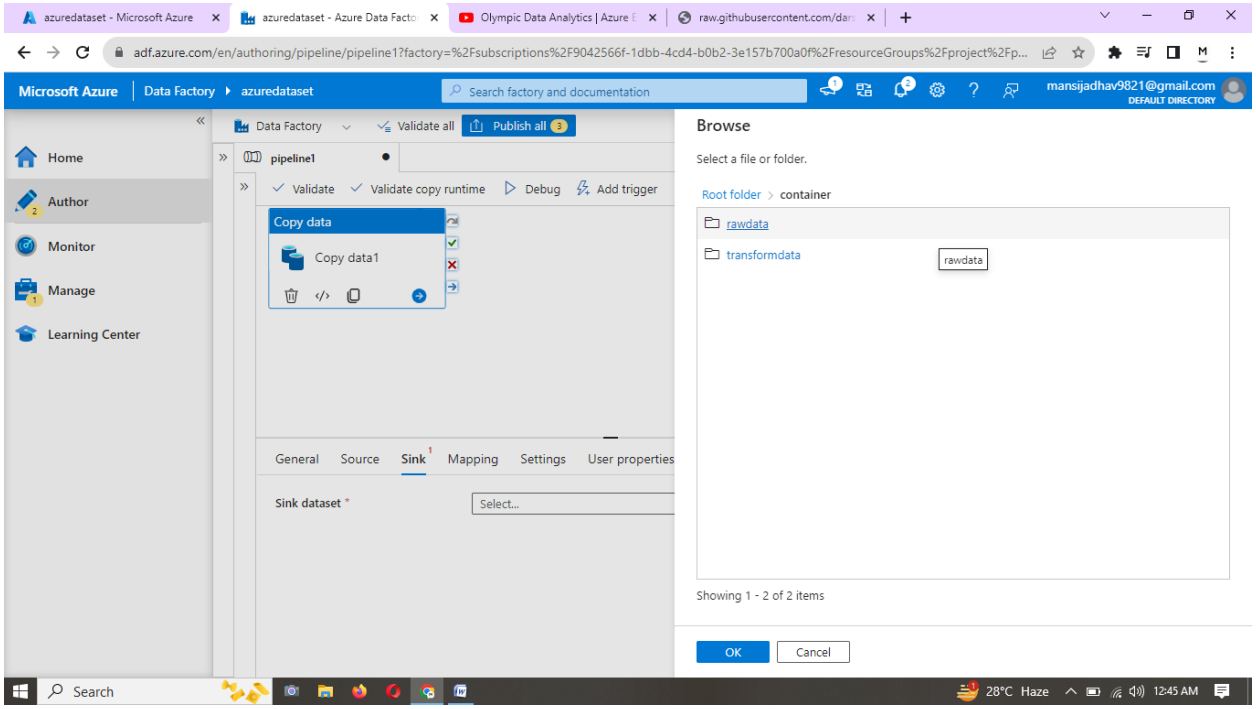


Same here also we select csv file amd then continue

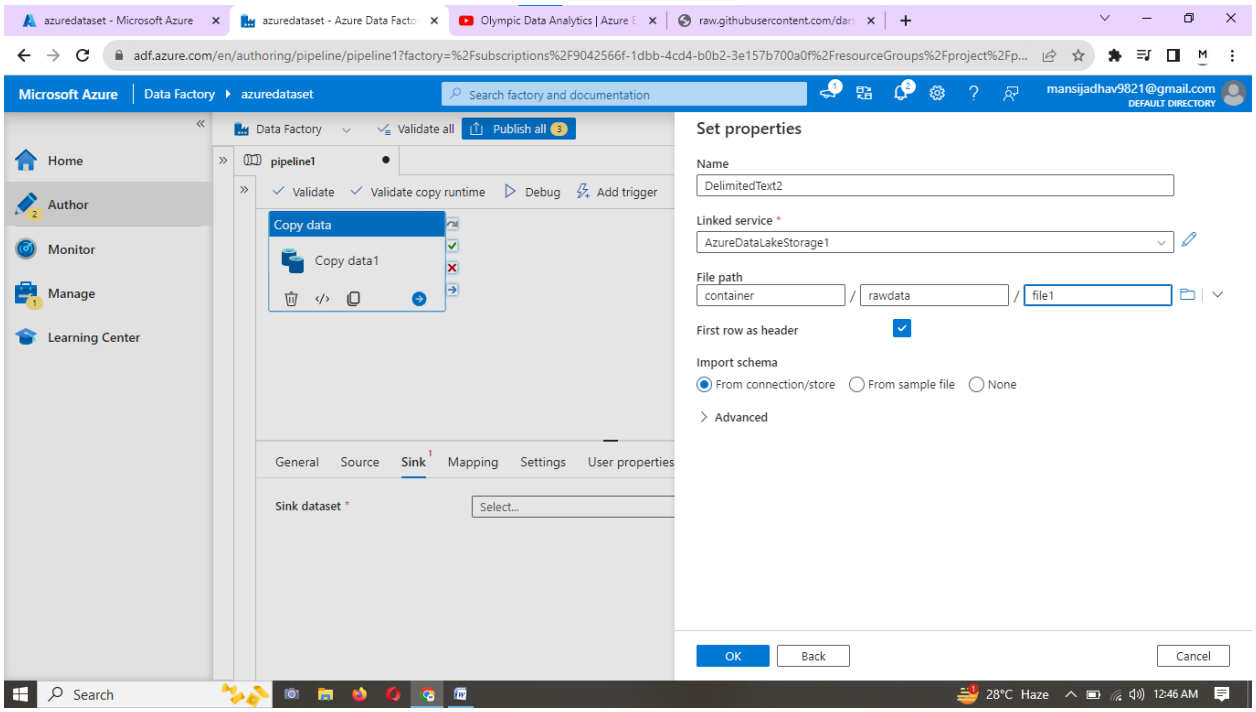


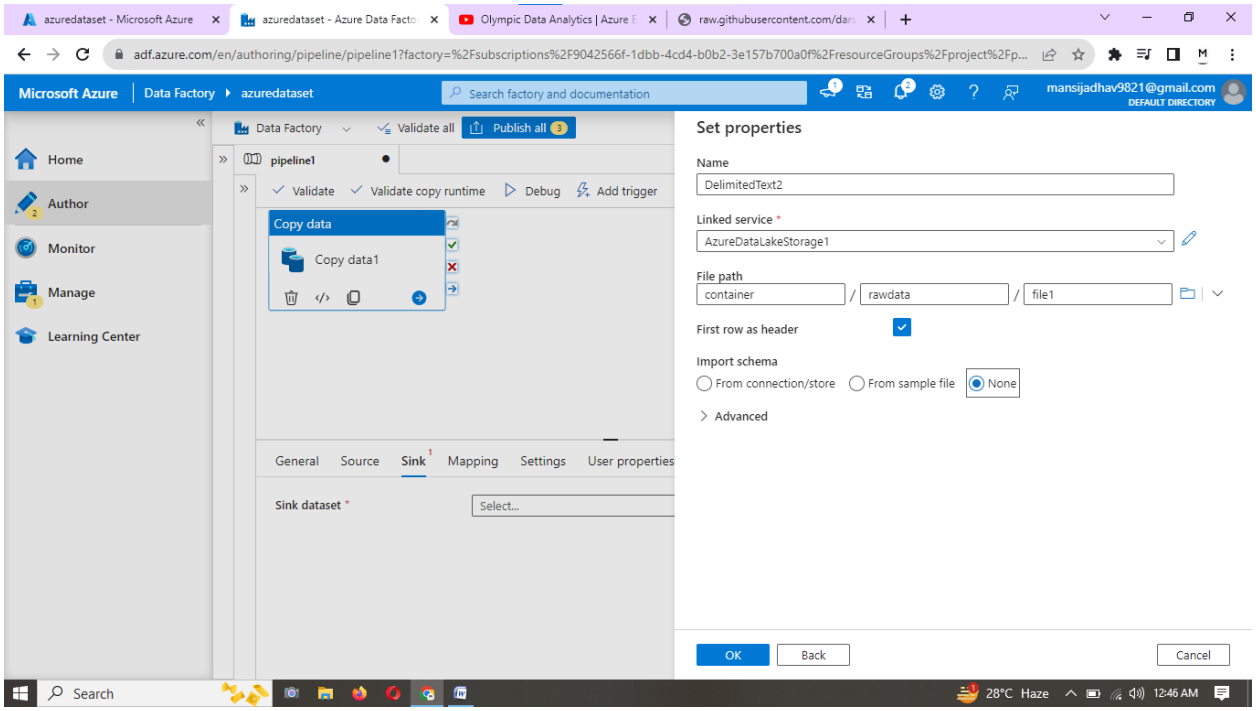
Here select the subscription and storage account we have created and then create it



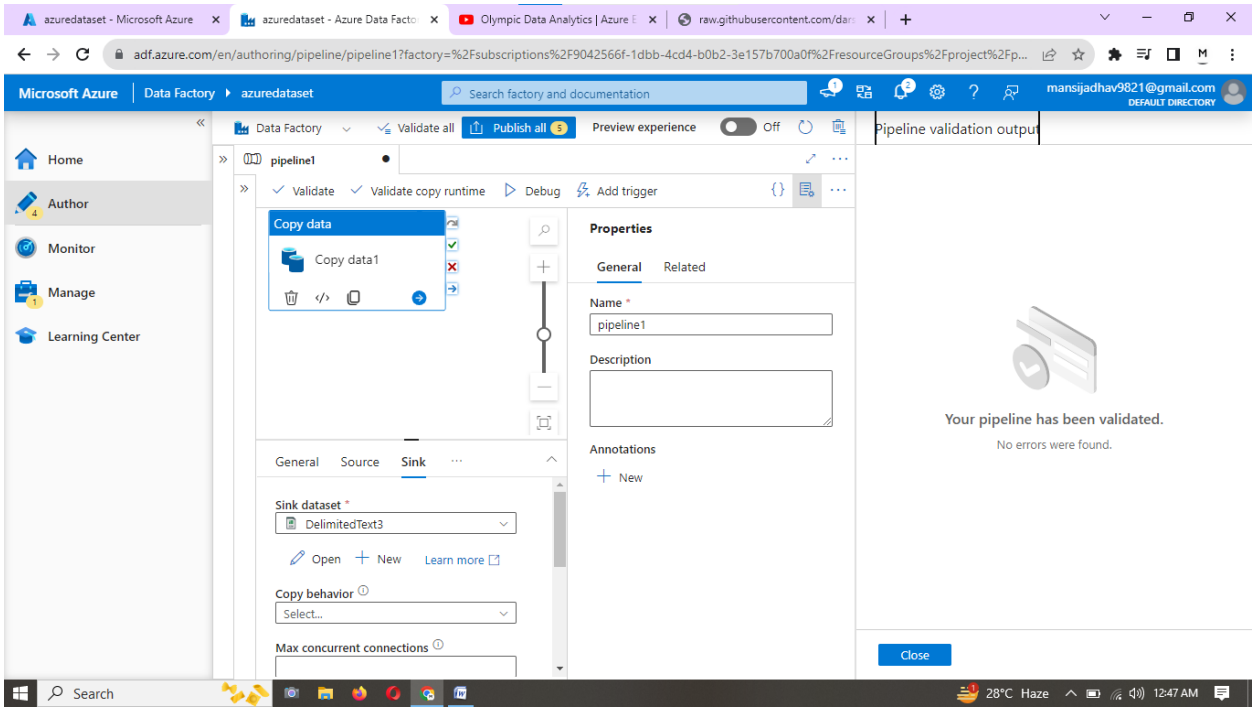


Now here we will browse the location ie container and there raw data and click on ok and give the file name.

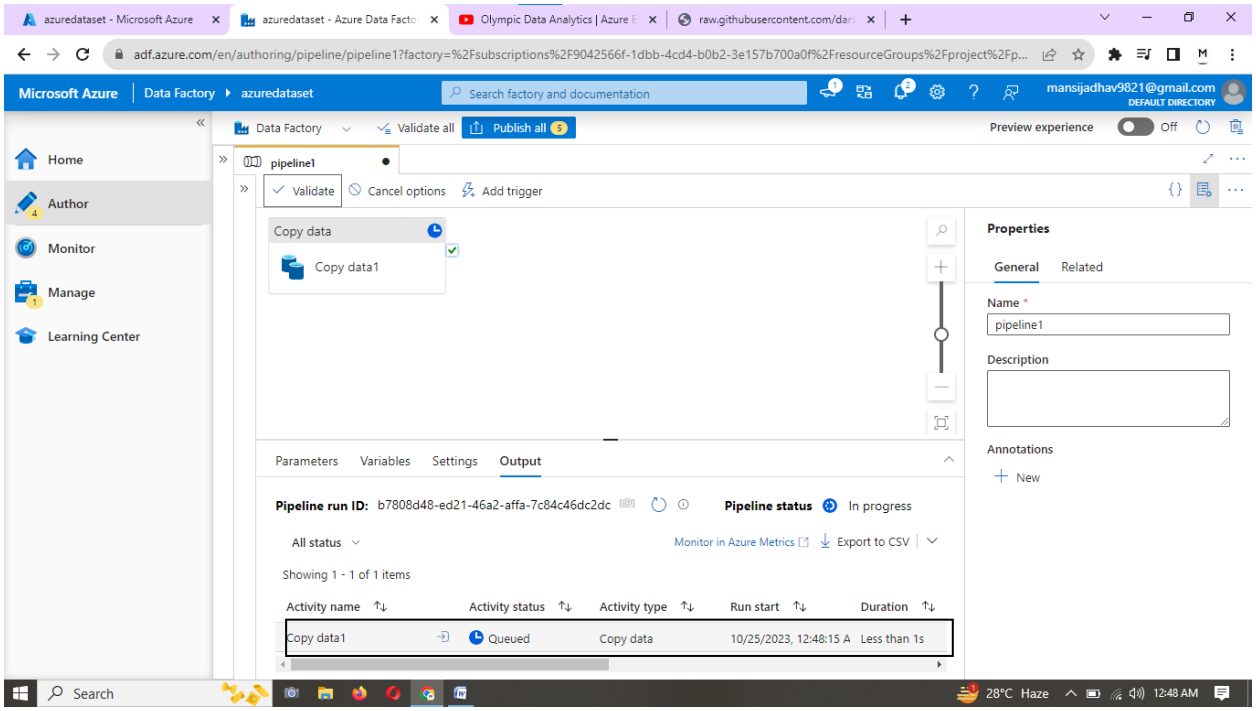




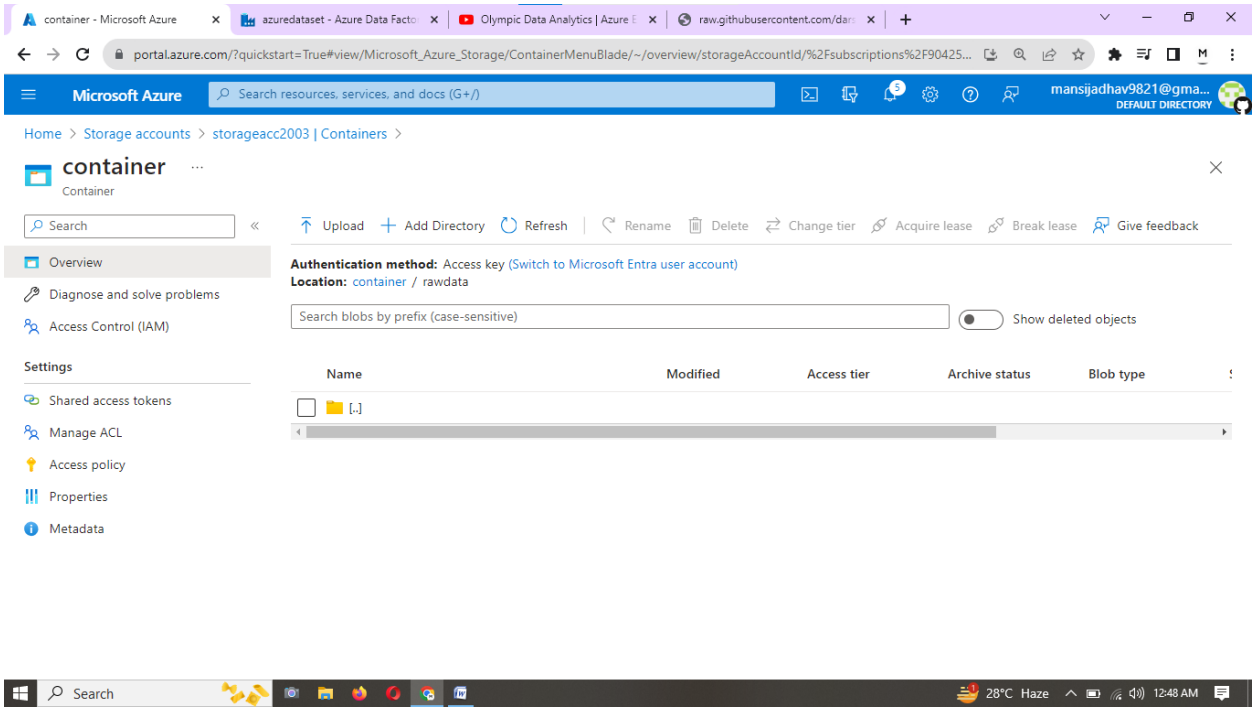
After that click on Validate and if no error occurs means we have done all right.



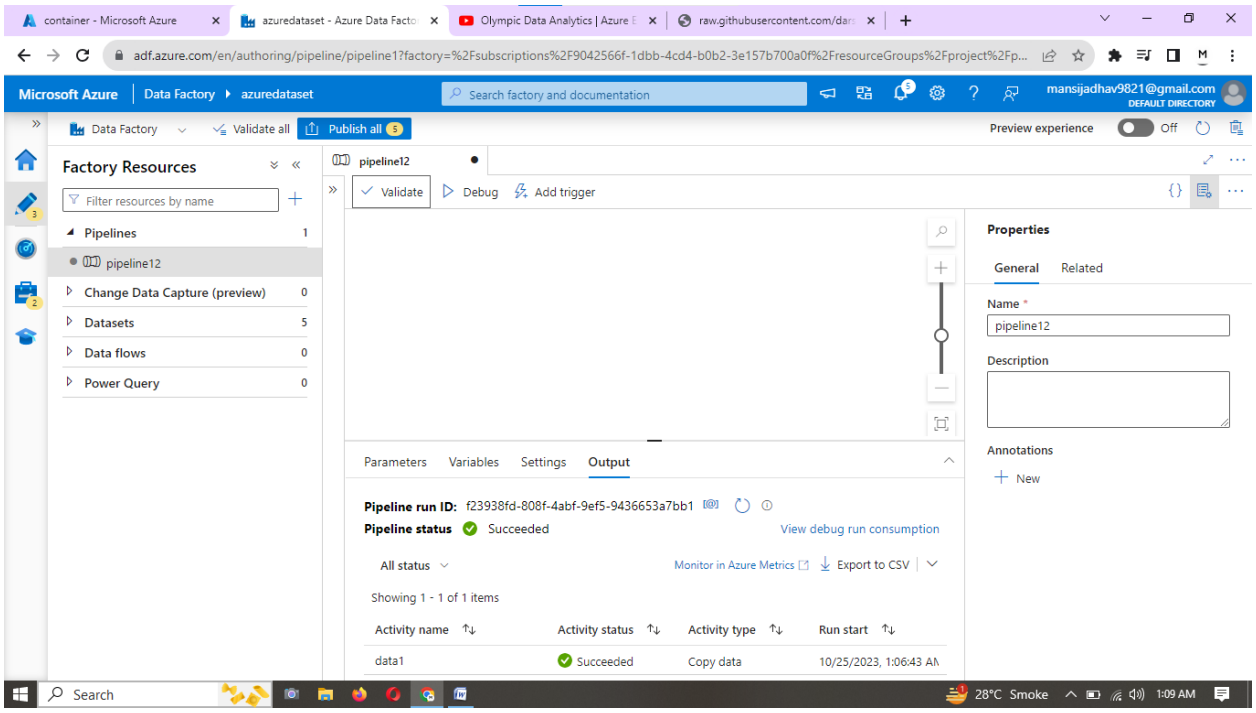
So now we will Debug



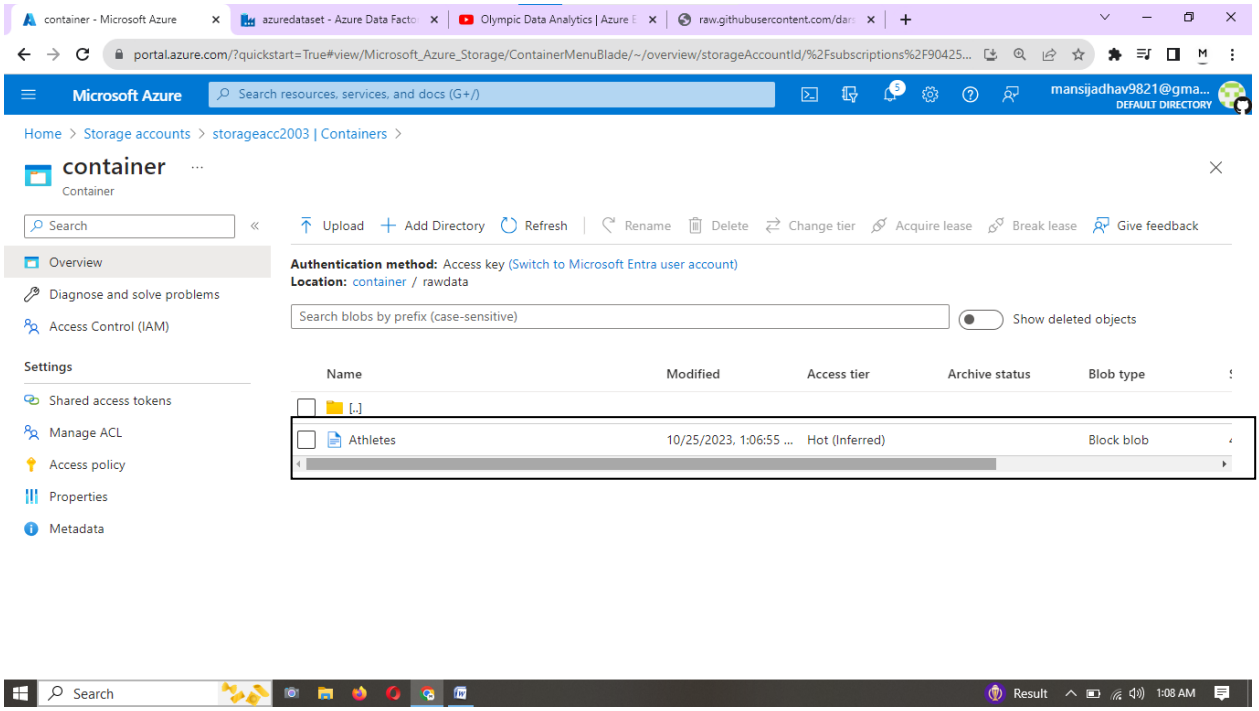
Before the raw data was empty no file was there



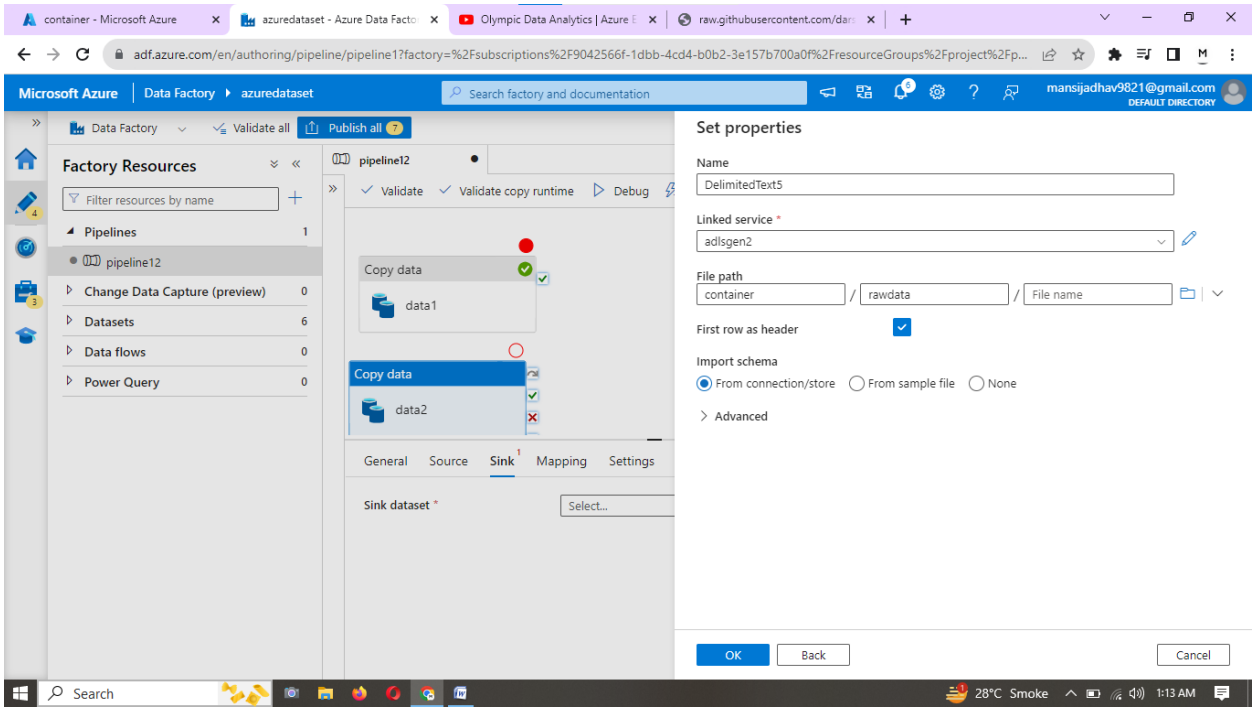
Now it is successful and go and check in azure portal Athletes file will be shown there

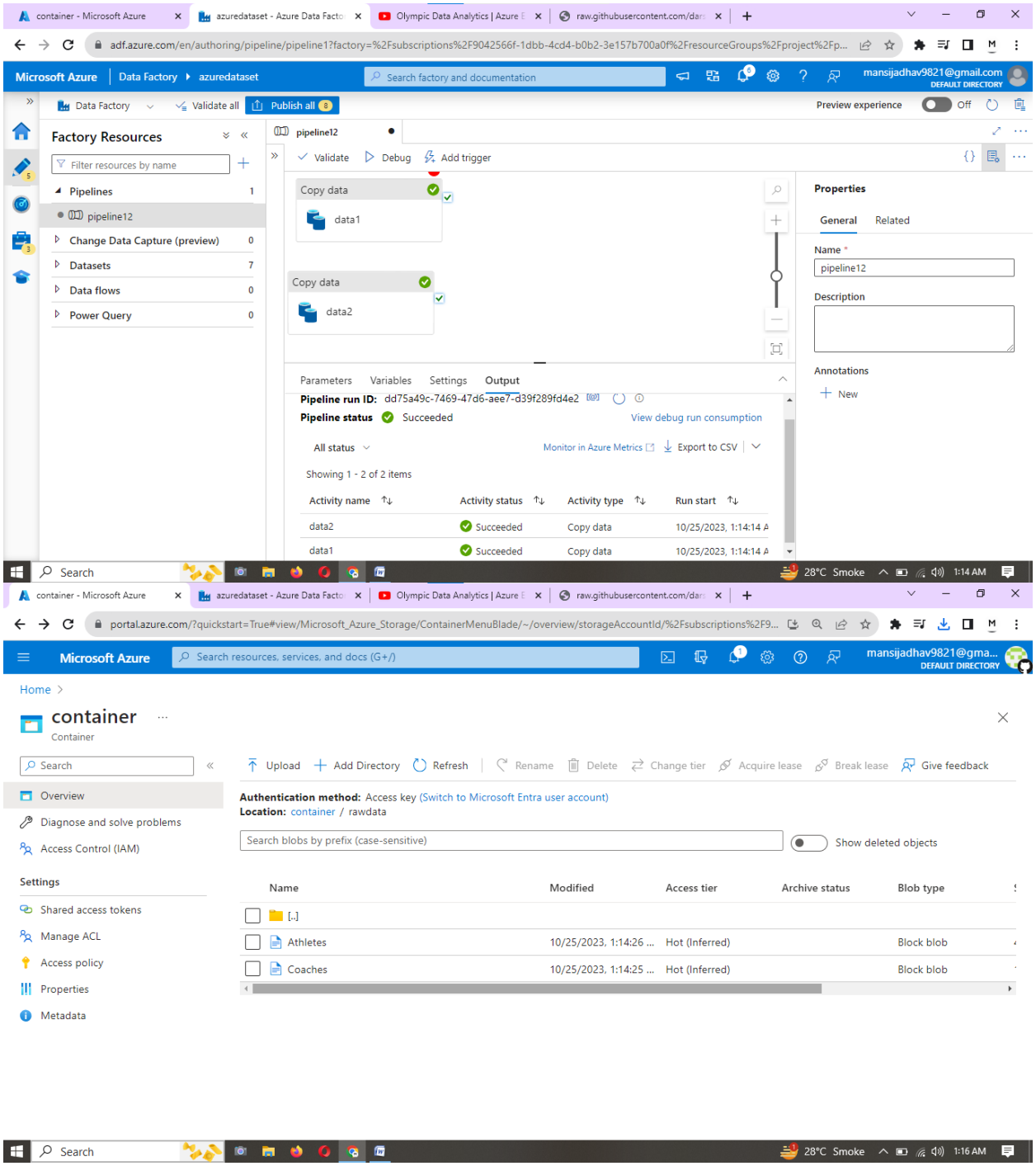


See we have atheletes file in container

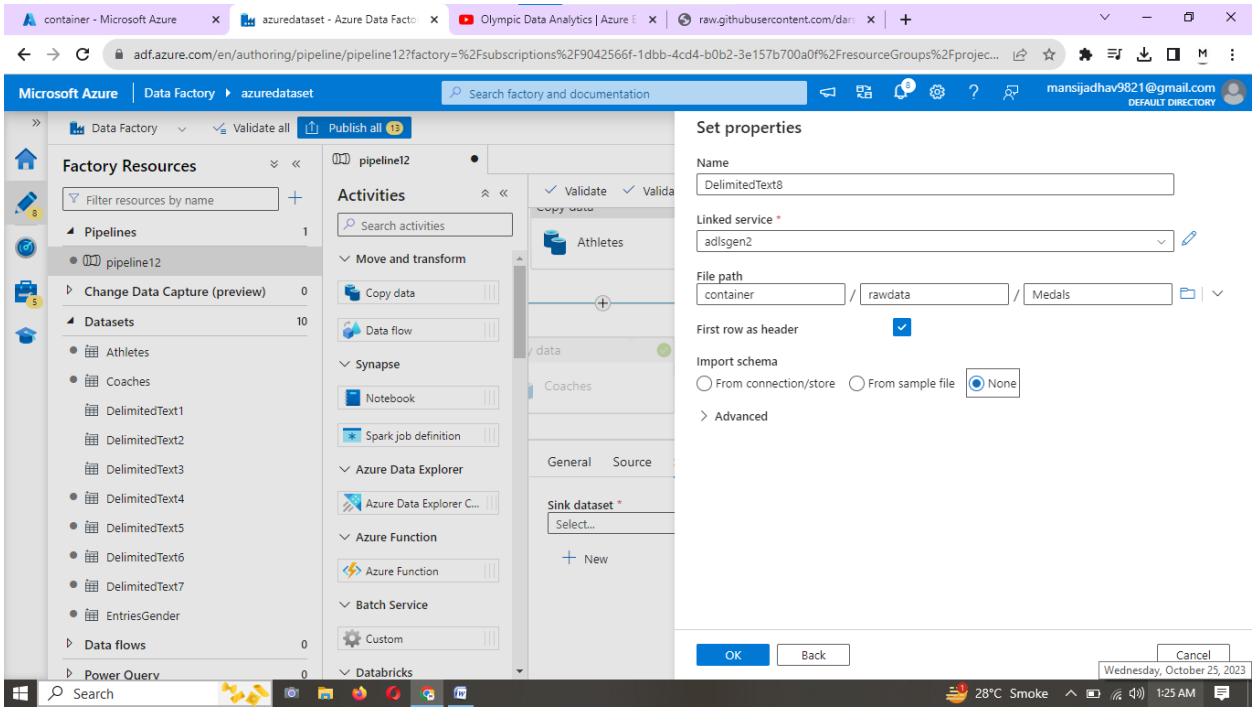


Now we will do same steps for other files perform same step

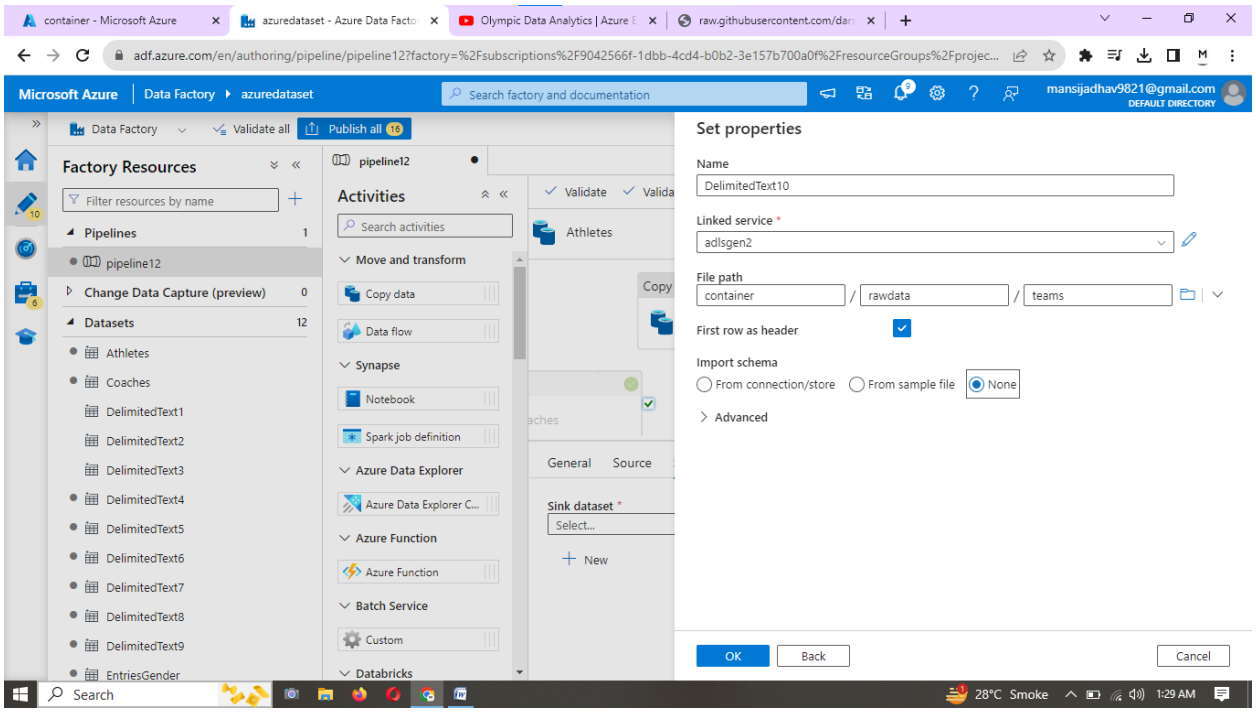



container - Microsoft Azure
azuredataset - Azure Data Facto
Olympic Data Analytics | Azure E
raw.githubusercontent.com/dar
adf.azure.com/en/authoring/pipeline/pipeline1?factory=%2Fsubscriptions%2F9042566f-1dbb-4cd4-b0b2-3e157b700a0f%2FresourceGroups%2Fproject%2Fp...
Microsoft Azure
Data Factory
azuredataset
Search factory and documentation
mansijadhav9821@gmail.com
DEFAULT DIRECTORY
Data Factory
Validate all
Publish all
Preview experience
Off
Factory Resources
Filter resources by name
Pipelines
pipeline12
Change Data Capture (preview)
Datasets
Data flows
Power Query
pipeline12
Validate
Debug
Add trigger
Copy data
data1
Copy data
data2
Properties
General
Related
Name *
pipeline12
Description
Annotations
New
Parameters
Variables
Settings
Output
Pipeline run ID: dd75a49c-7469-47d6-aee7-d39f289fd4e2
Pipeline status Succeeded
View debug run consumption
All status
Monitor in Azure Metrics
Export to CSV
Showing 1 - 2 of 2 items
Activity name
Activity status
Activity type
Run start
data2
Succeeded
Copy data
10/25/2023, 1:14:14 A
data1
Succeeded
Copy data
10/25/2023, 1:14:14 A
Search
28°C Smoke
1:14 AM
container - Microsoft Azure
azuredataset - Azure Data Facto
Olympic Data Analytics | Azure E
raw.githubusercontent.com/dar
portal.azure.com/?quickstart=True#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2F9...
Microsoft Azure
Search resources, services, and docs (G+/)
mansijadhav9821@gma...
DEFAULT DIRECTORY
Home
container
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Shared access tokens
Manage ACL
Access policy
Properties
Metadata
Upload
Add Directory
Refresh
Rename
Delete
Change tier
Acquire lease
Break lease
Give feedback
Authentication method: Access key (Switch to Microsoft Entra user account)
Location: container / rawdata
Search blobs by prefix (case-sensitive)
Show deleted objects
Name
Modified
Access tier
Archive status
Blob type
[..]
Athletes
10/25/2023, 1:14:26 ...
Hot (Inferred)
Block blob
Coaches
10/25/2023, 1:14:25 ...
Hot (Inferred)
Block blob
Search
28°C Smoke
1:16 AM

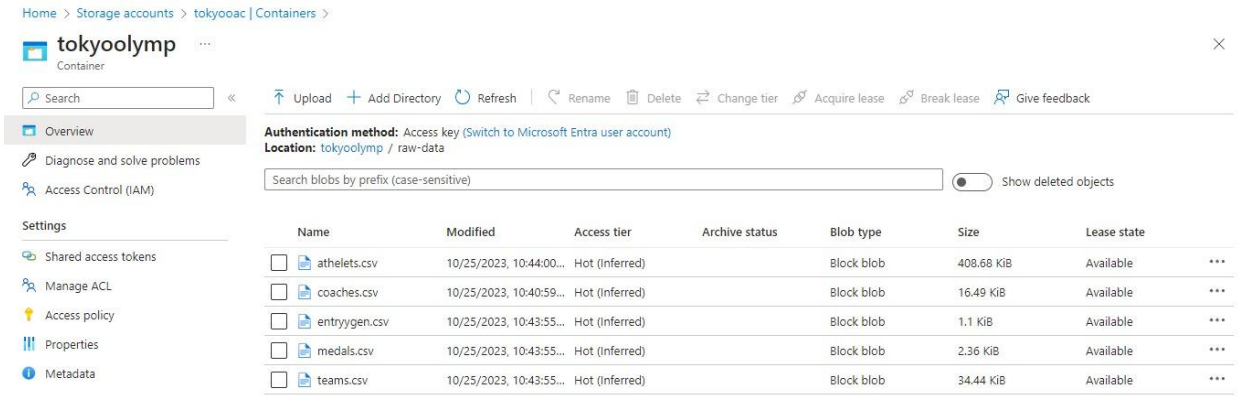
For Medals



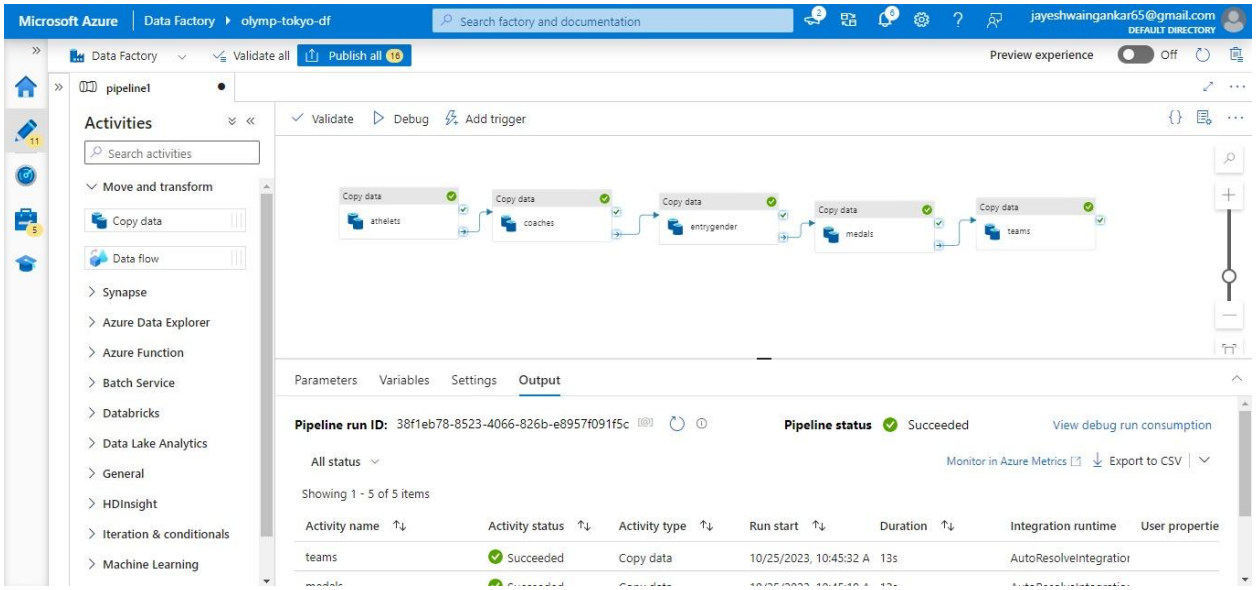
For Teams

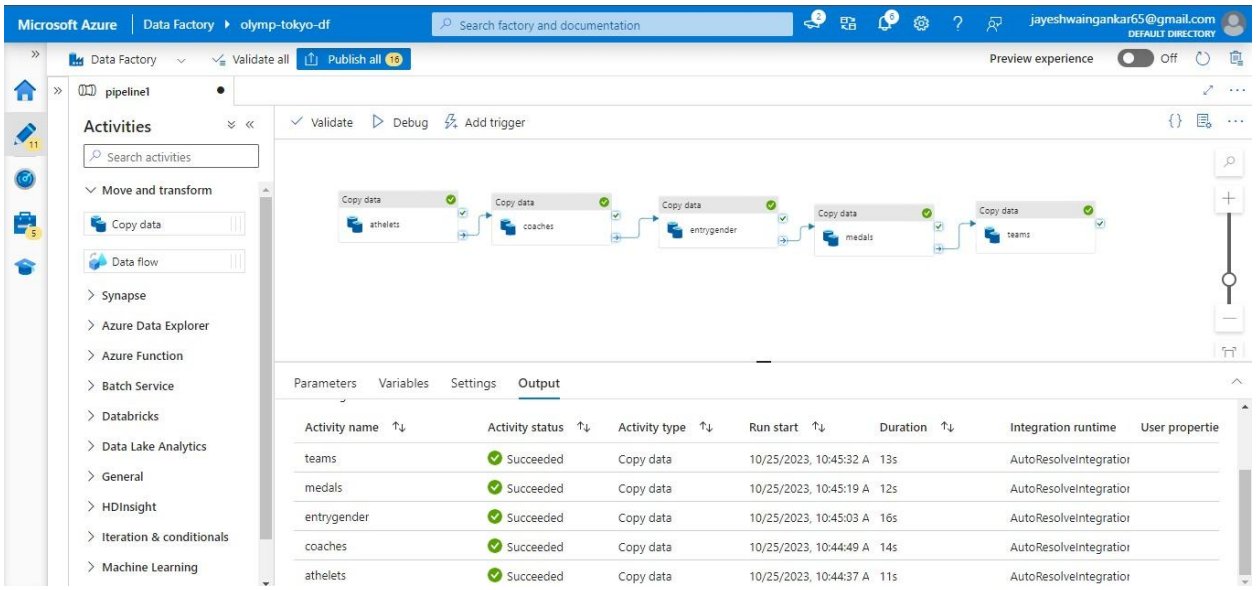


Here we have all our 5 files in azure container

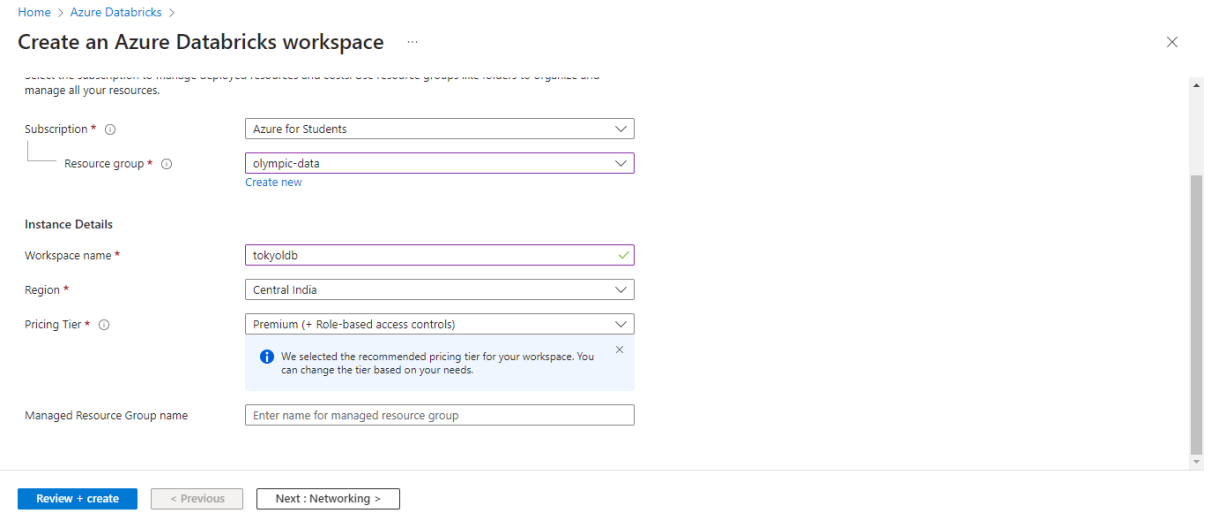


This is the linked diagram of the files in Data Factory and we have created pipeline.

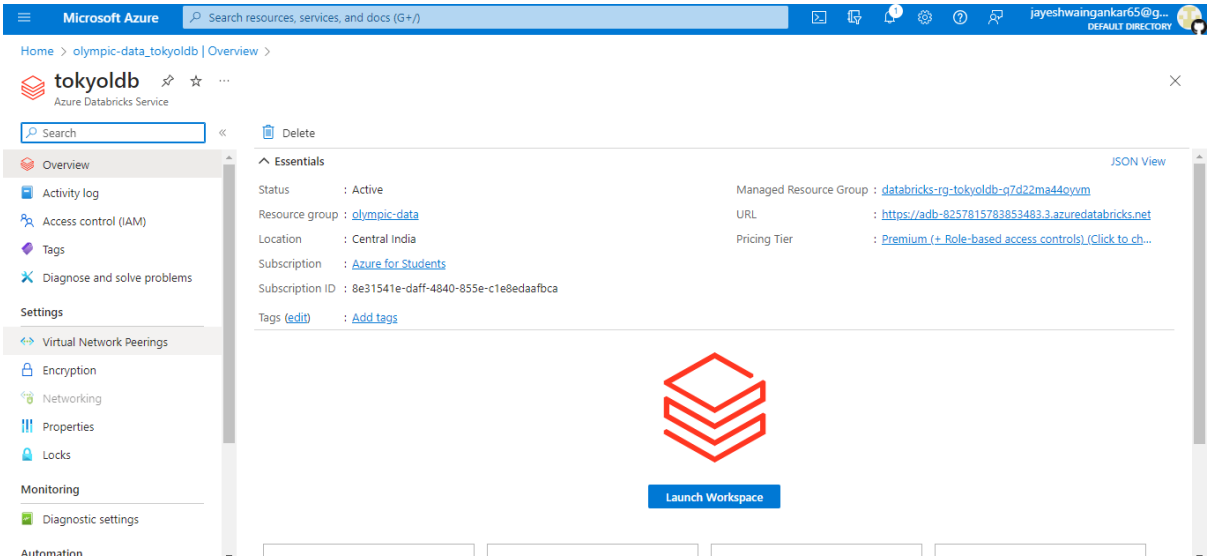




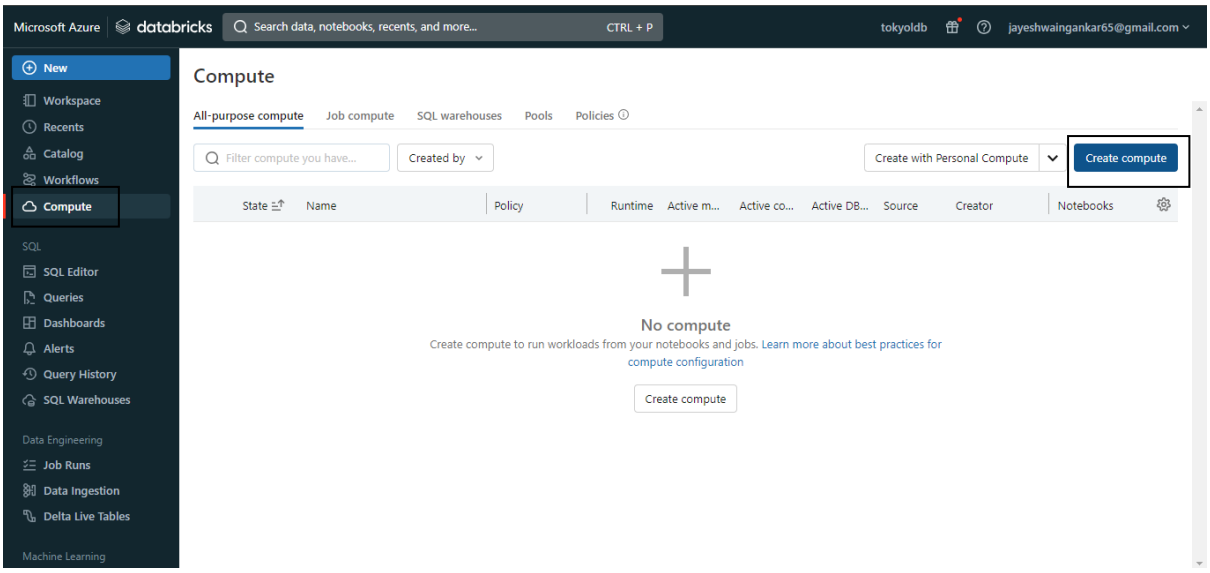
Now we will do Azure Databricks part where we will write transformation code. So search databricks in azure portal. Give the and and then review and create.



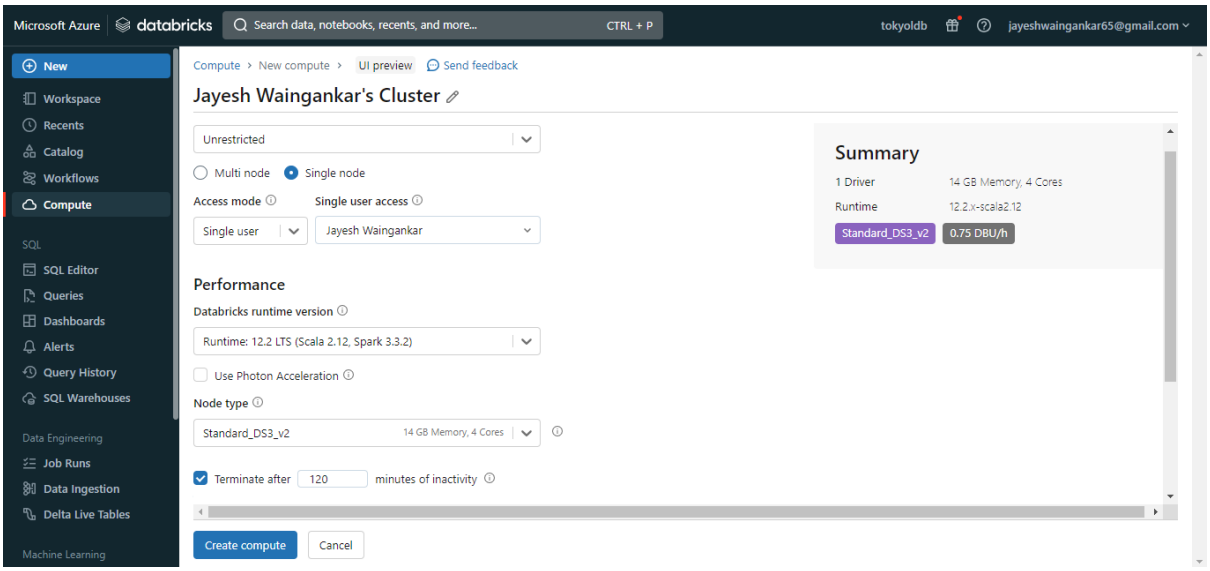
After deployment click on go to rescource and launch workspace



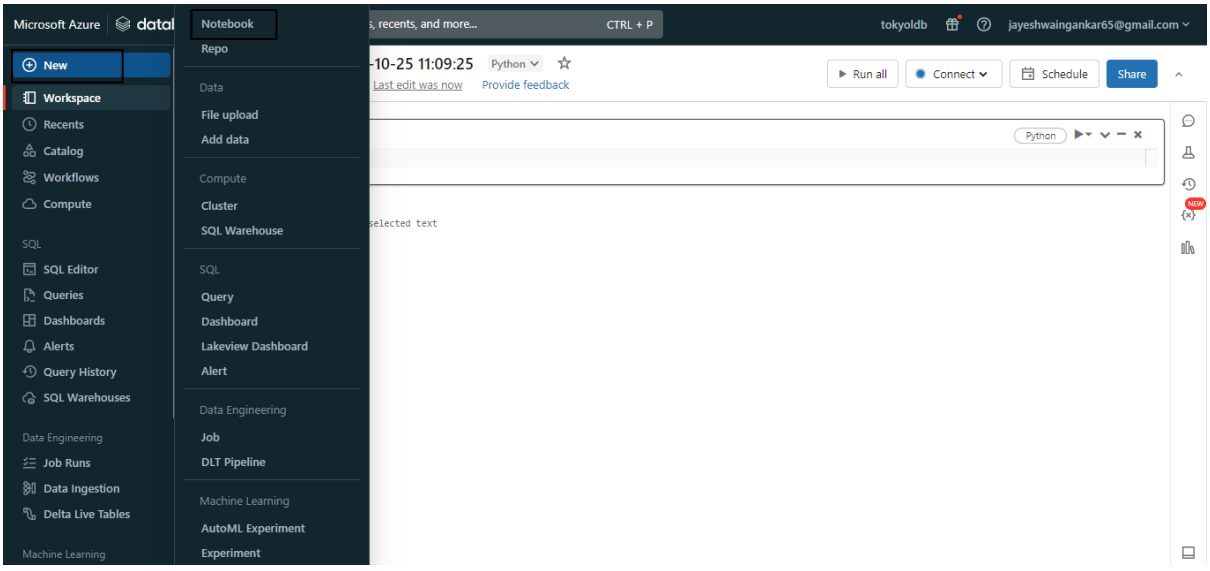
Now go to Compute and then create compute



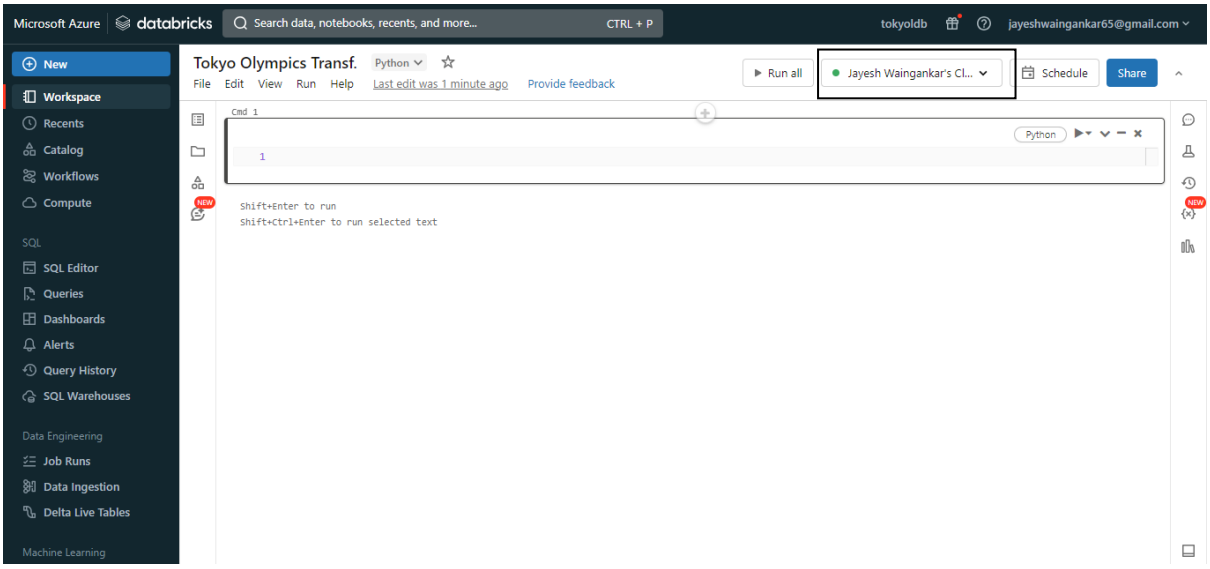
Now if we have multiple machines we can select MultiNode amd if single we can select single node and then create compute.



Here then go to new and then notebook and give the name for notebook.

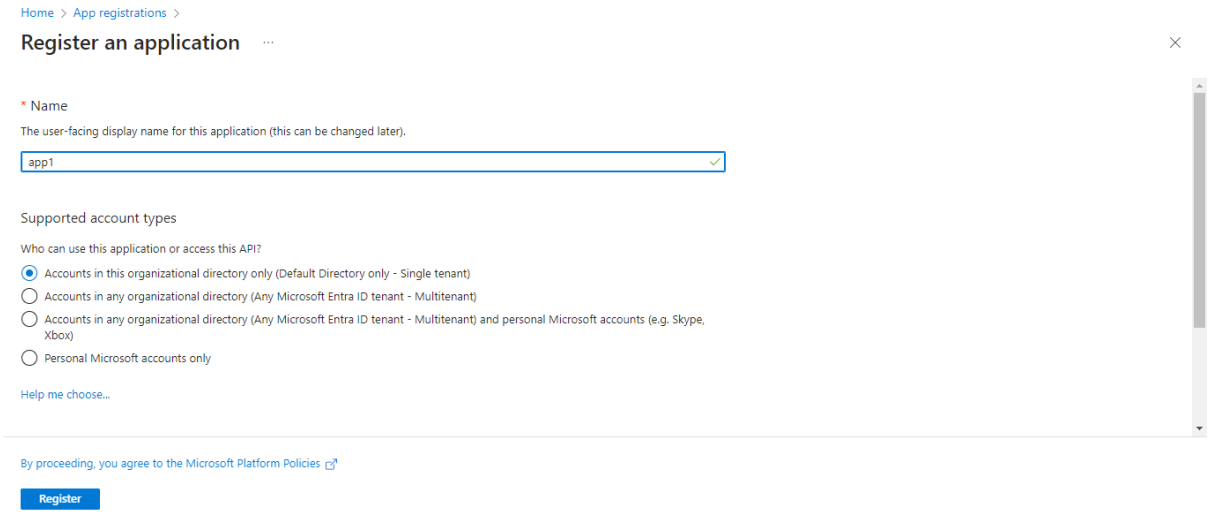


In Connect attach your cluster.

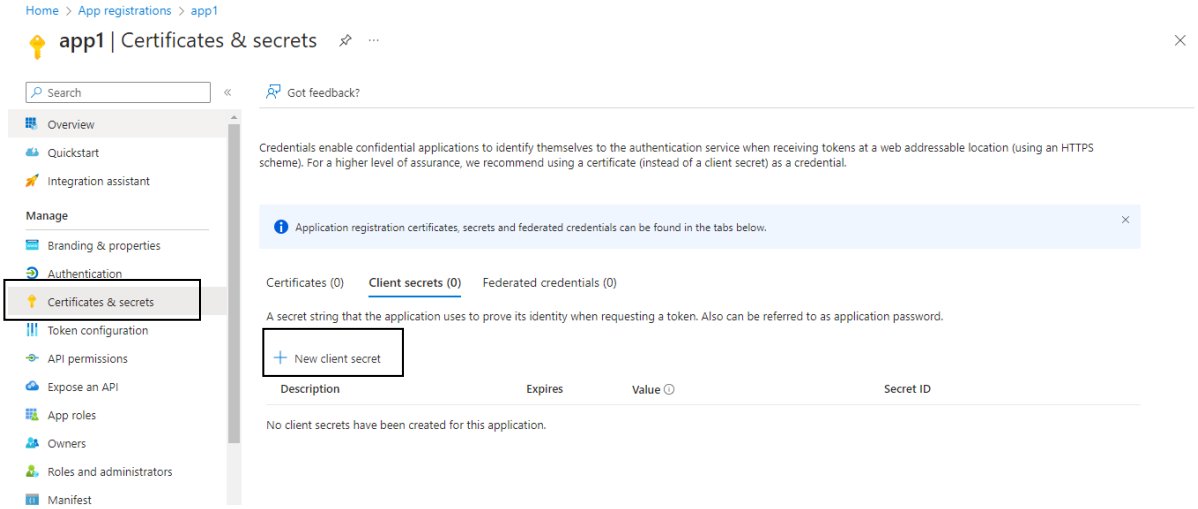


Now we have to create connection from this azure databricks to our azure data storage so that we can easily access the data so the steps to create connection between this is you need to mount this azure data lake storage to azure data factory so that you can access the file from it so mounting is basically attaching so just like we use USB cable to attach the hard disk physically we have to mount service to azure data factory. go to azure portal.

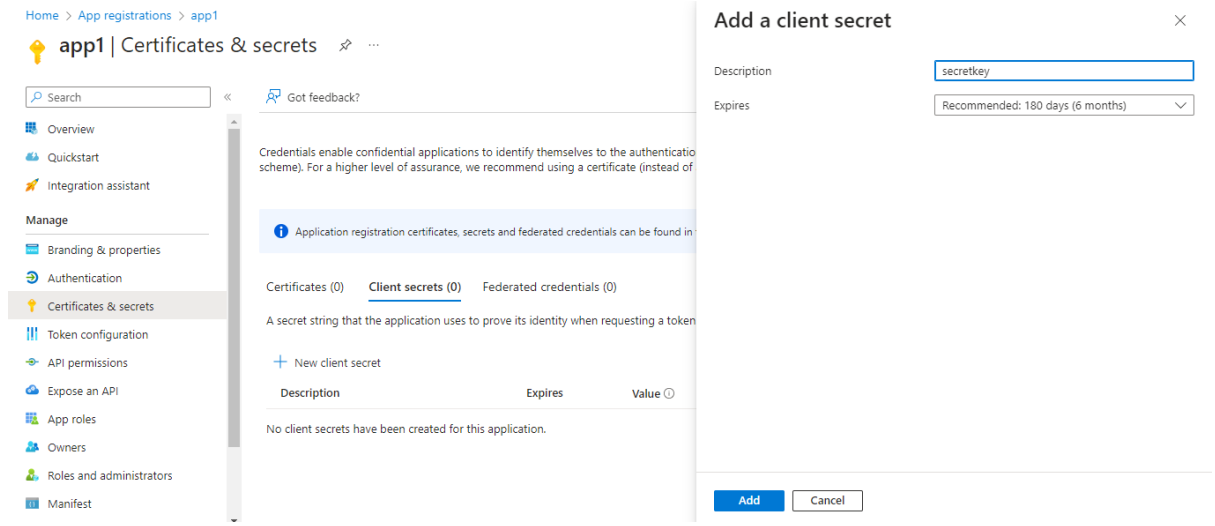
So in azure portal search app registration.



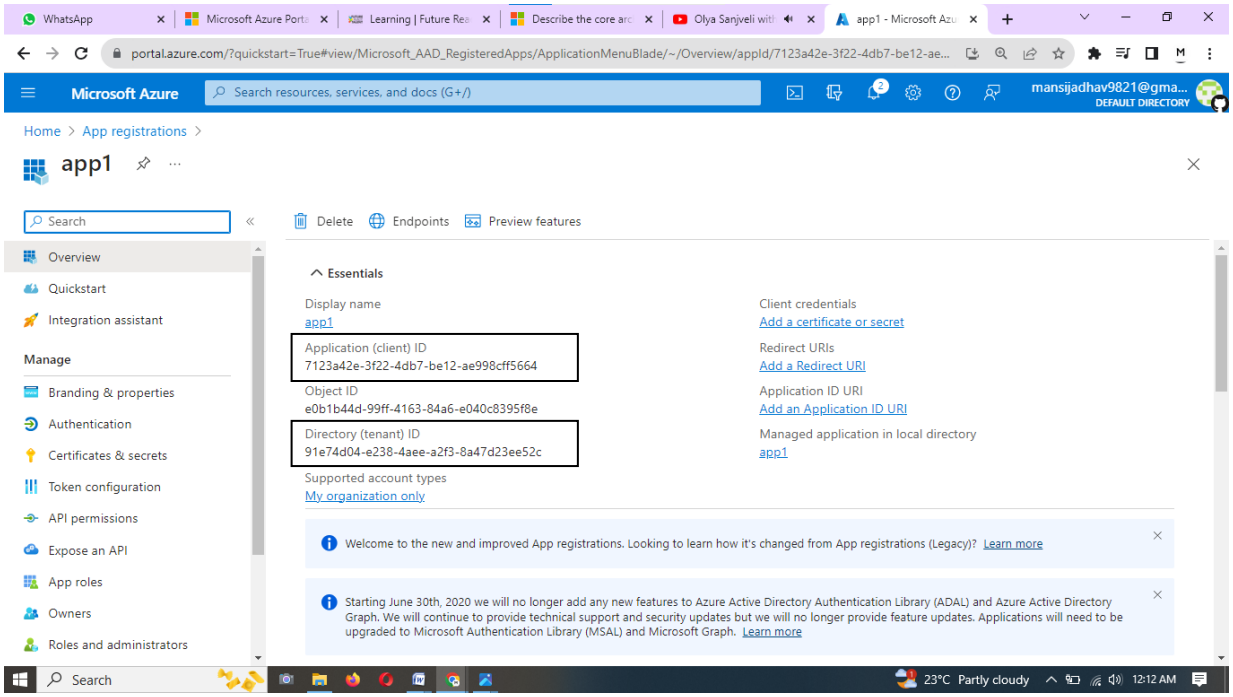
After creating it Go to resource and then in certificates and secrets click on new certificate



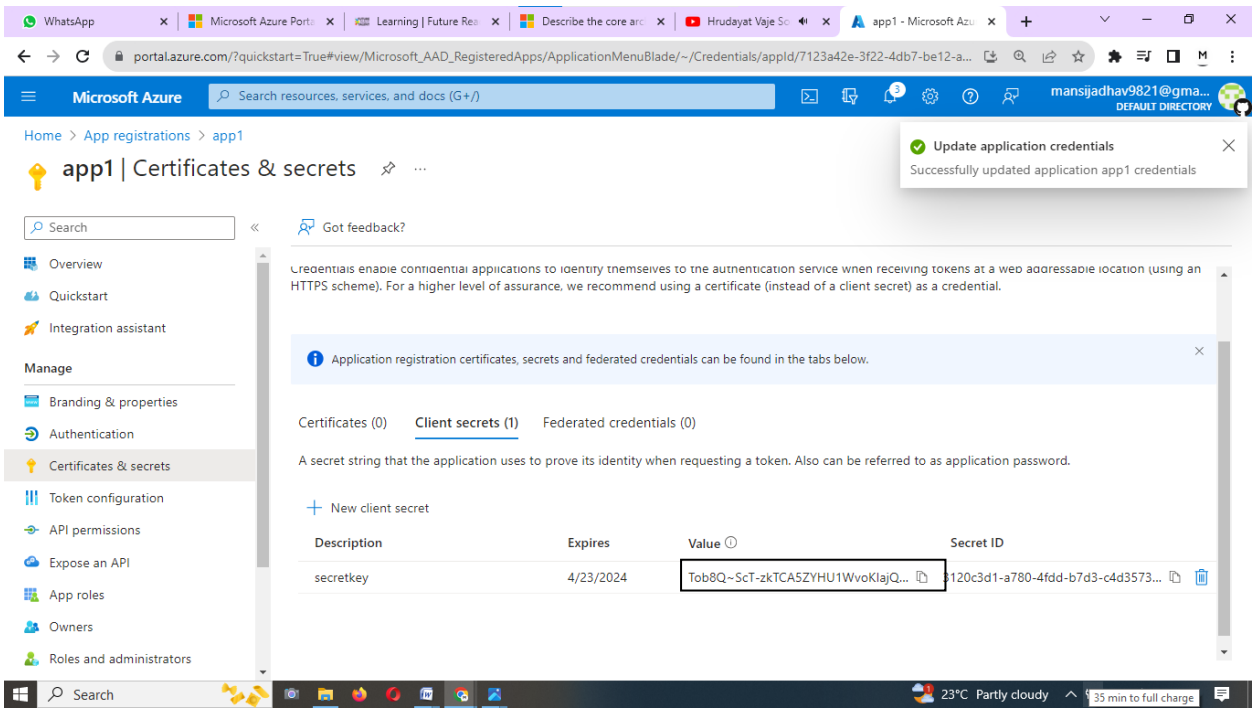
Give the name for client server and add it.



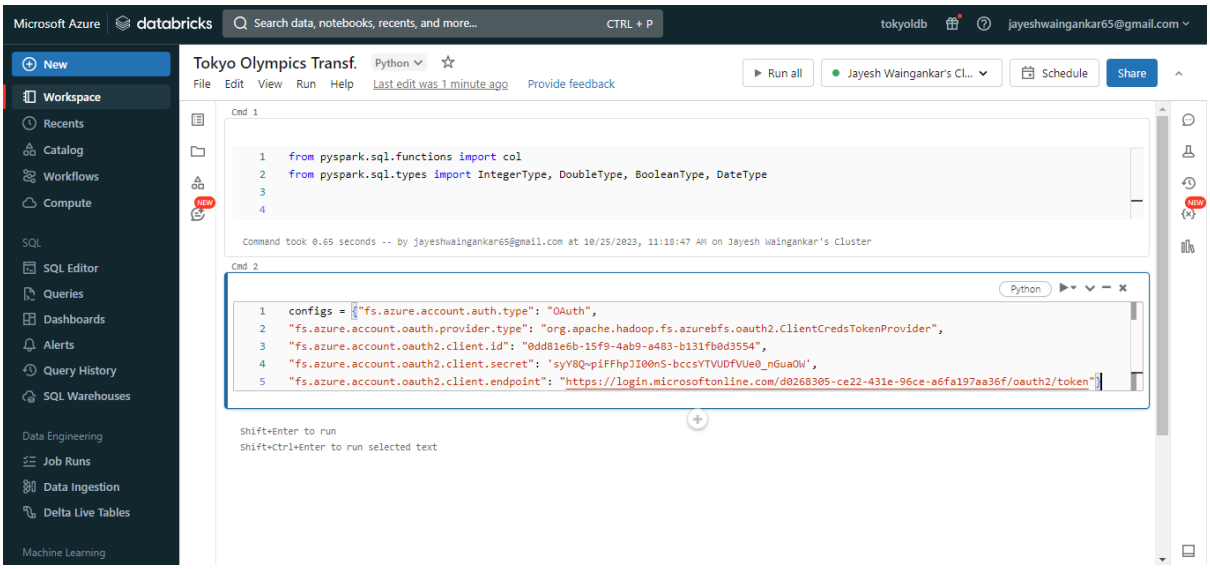
Here copy the client id , tenant id in notepad .



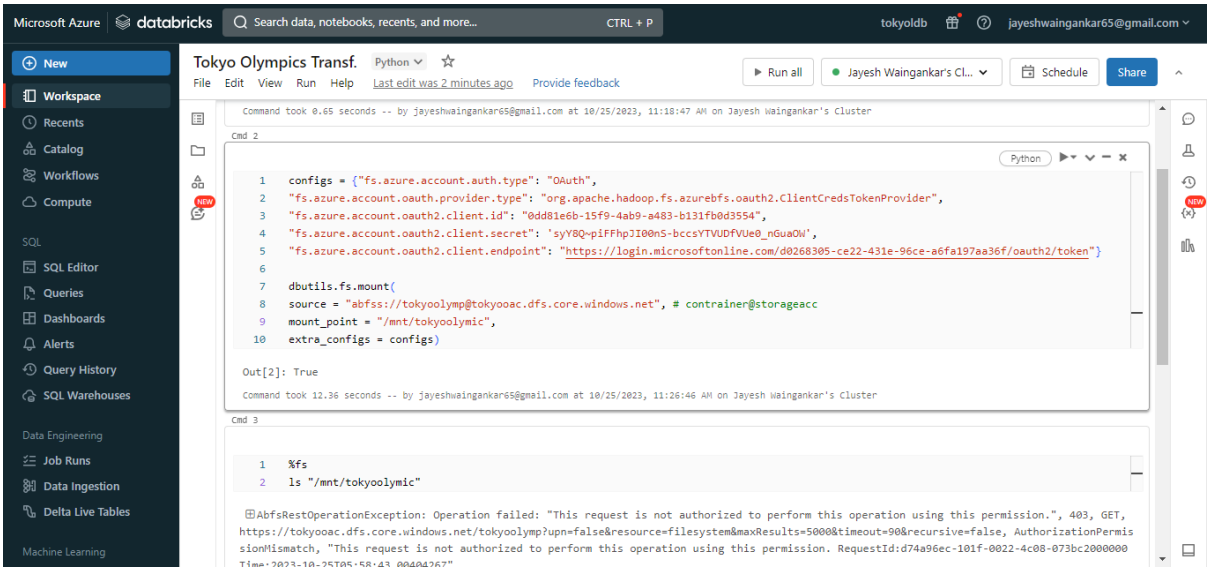
After creation of secret key copy the value in notepad



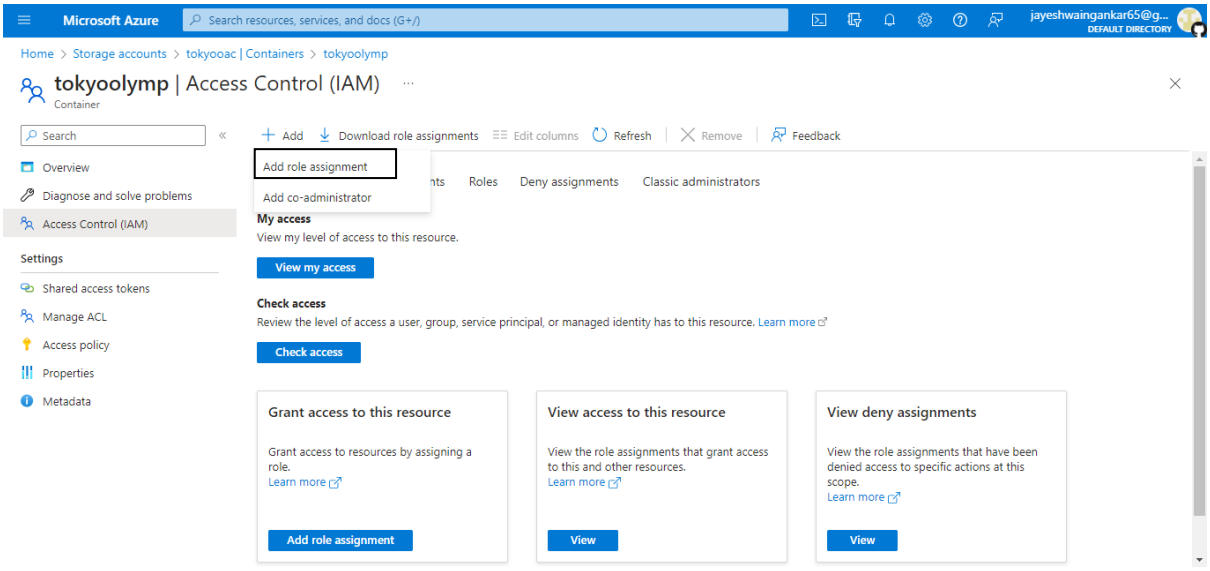
This is the code to run in databricks in the where we have tenant id copy that id here and where we have secret key copy the key here and then run. All the code are uploaded on Github link



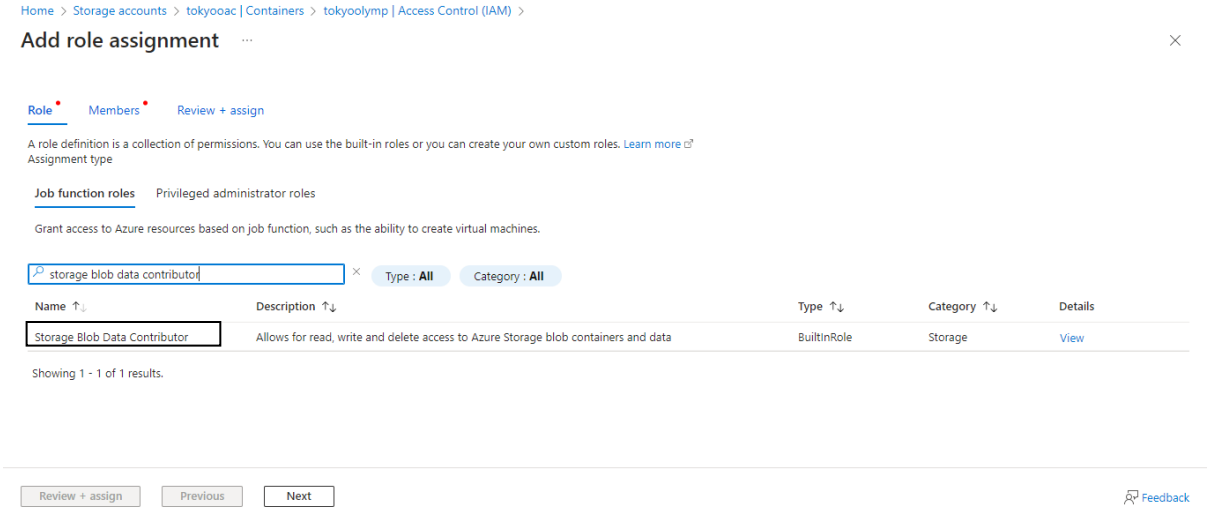
before @ write the name of container ie container@storageaccname.Now operation fail as access is not give so that we will five access.



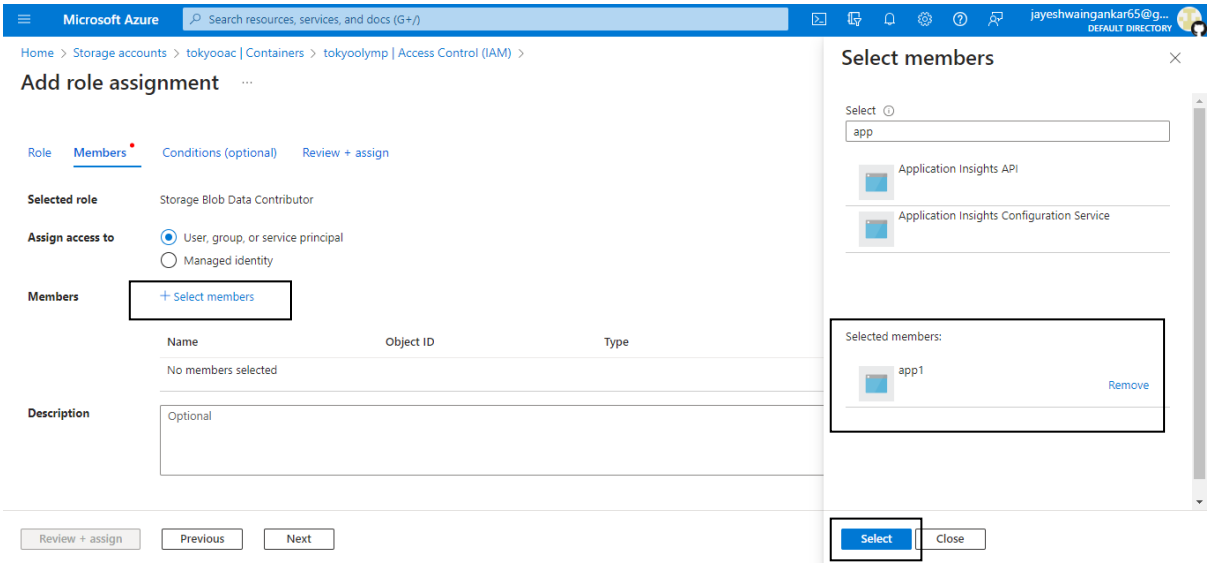
Go to azure portal container >> Access Control



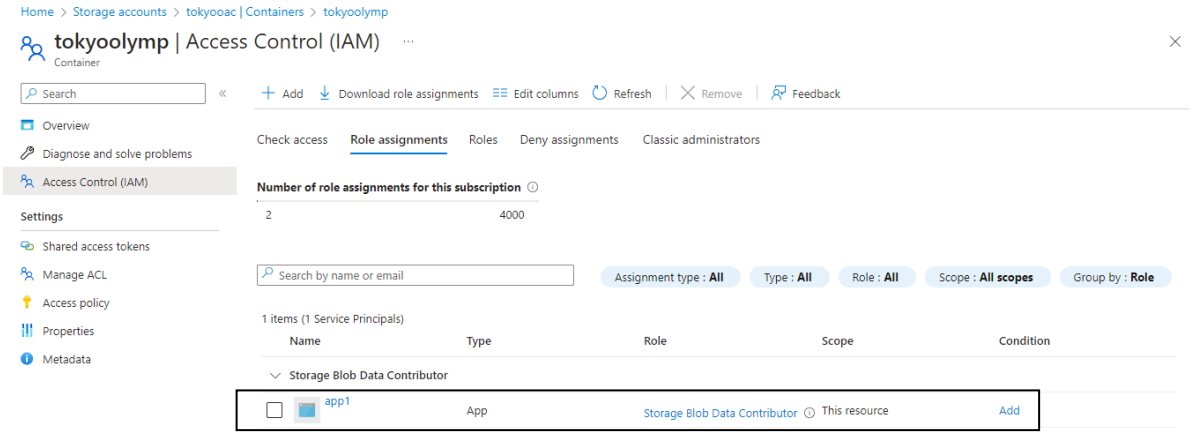
Here select storage blob data contributor



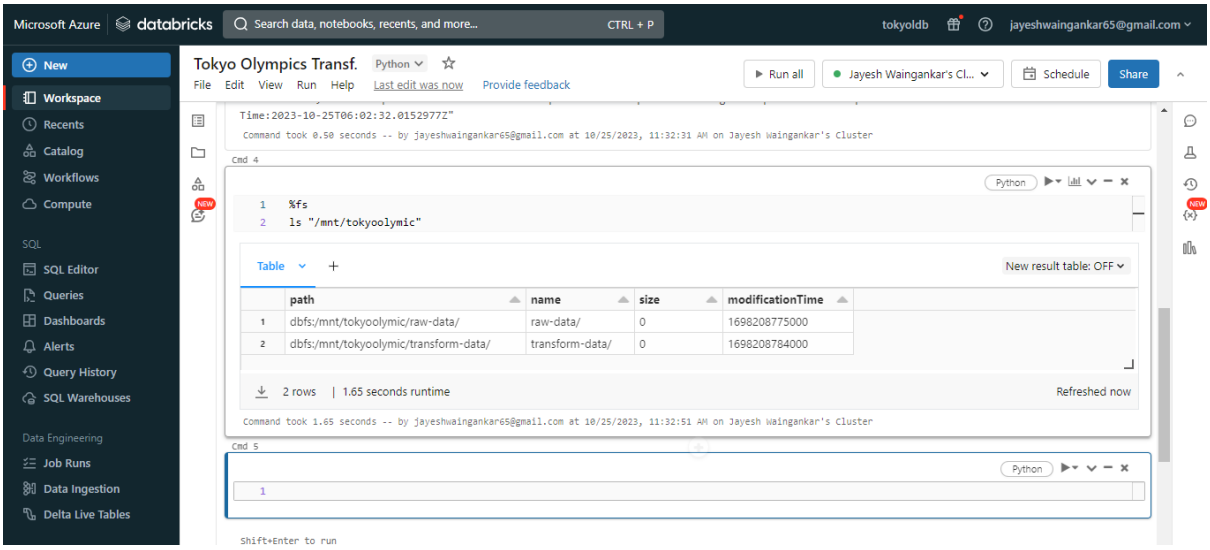
Here select members and select the app we have created

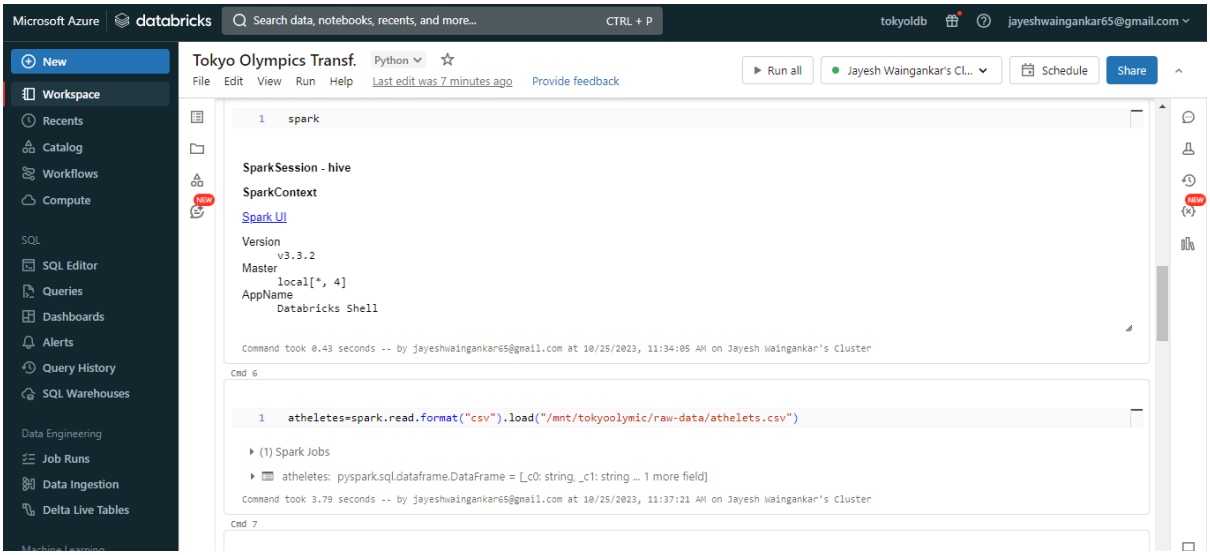


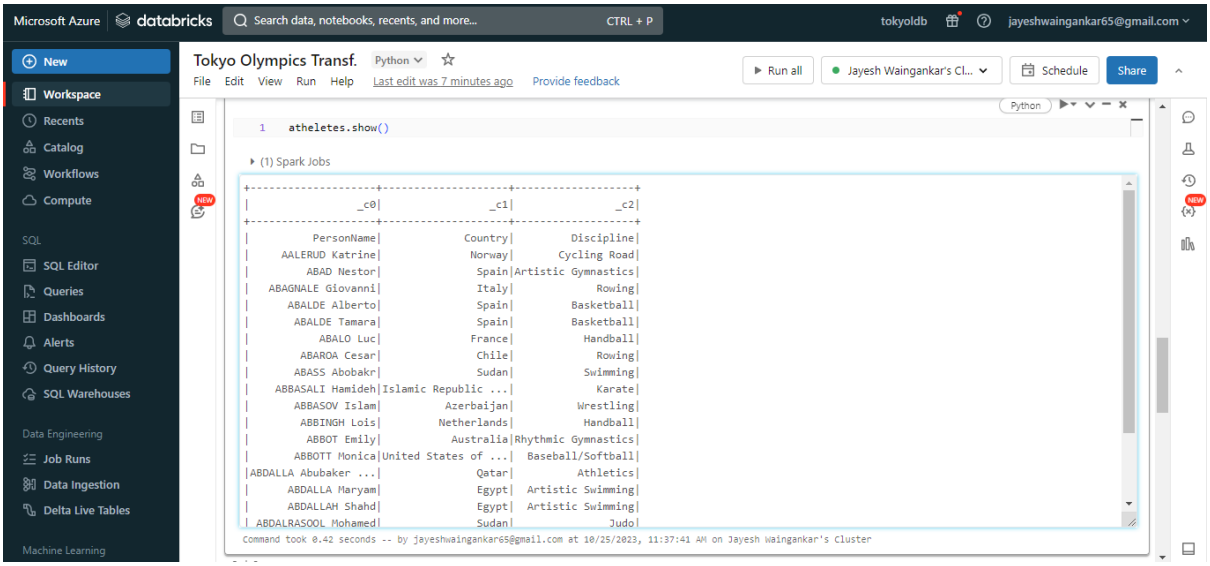
Created access control



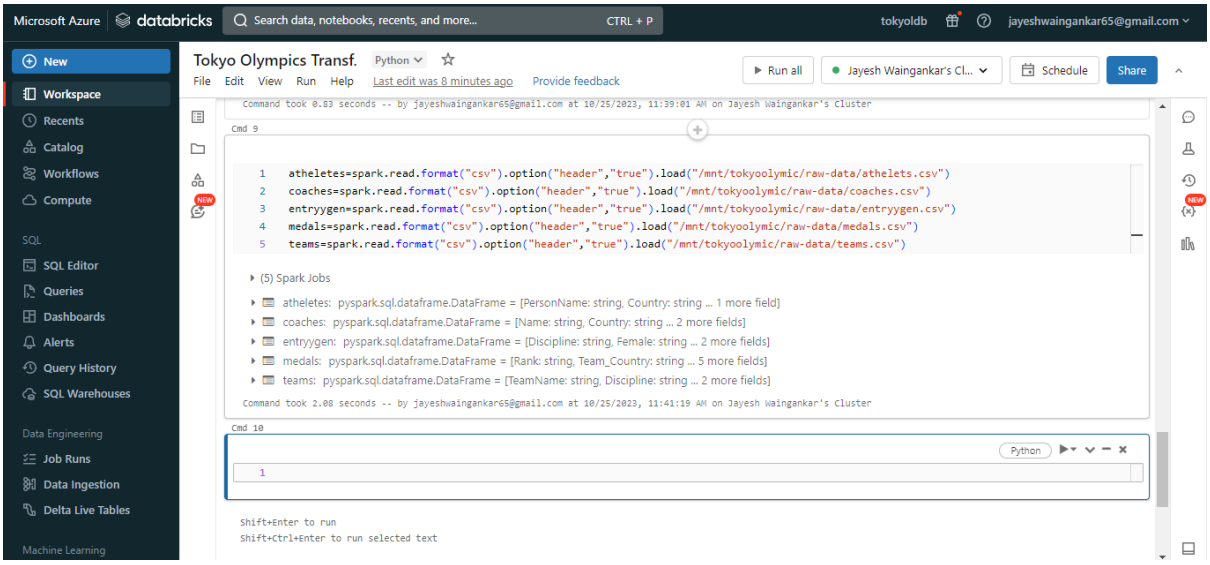
Now run the same code no error will come as now we have assigned role



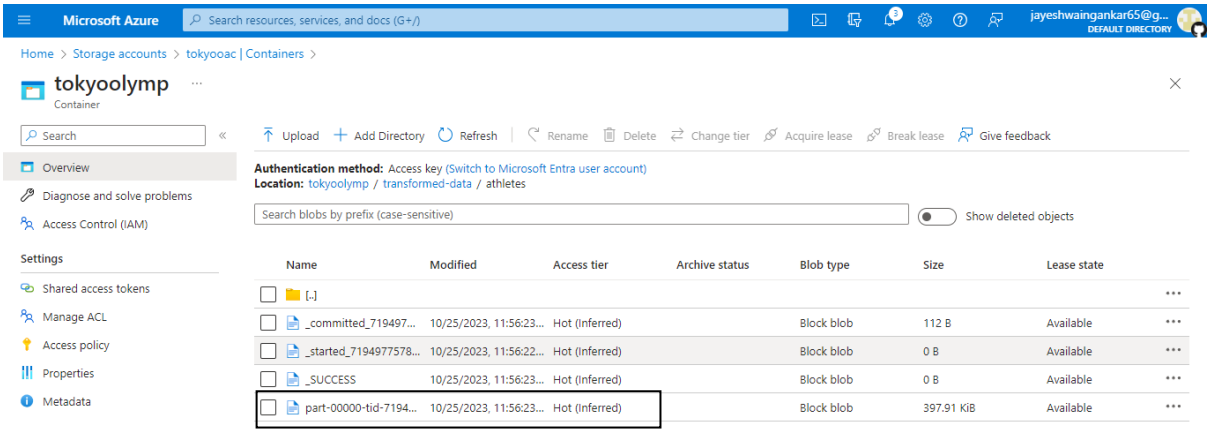


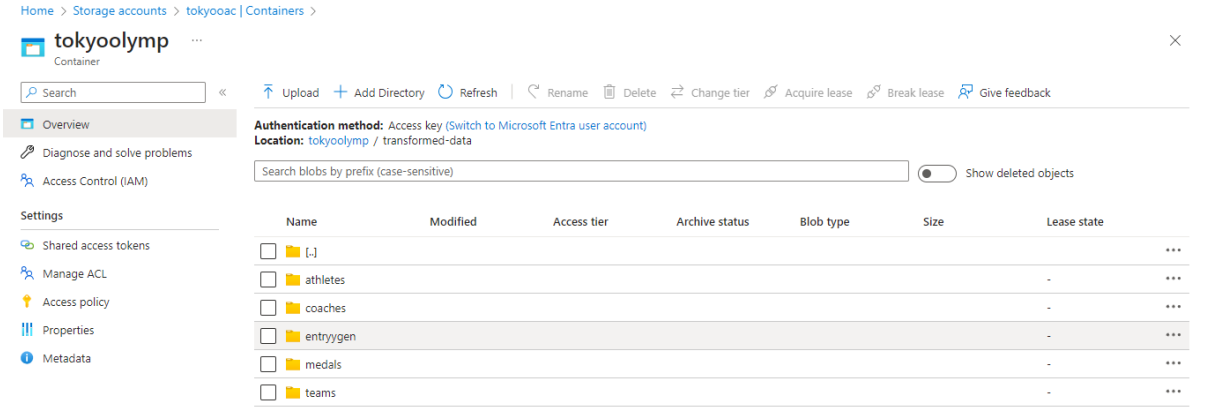


Here we have get the data for all 5 files

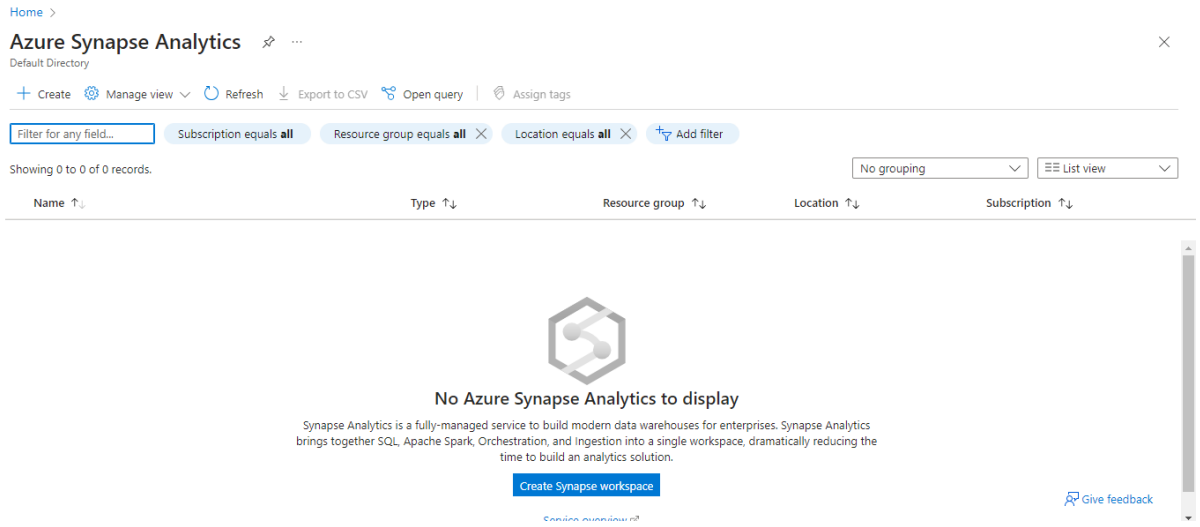


Created so now in azure portal go and check in each file all the files are created

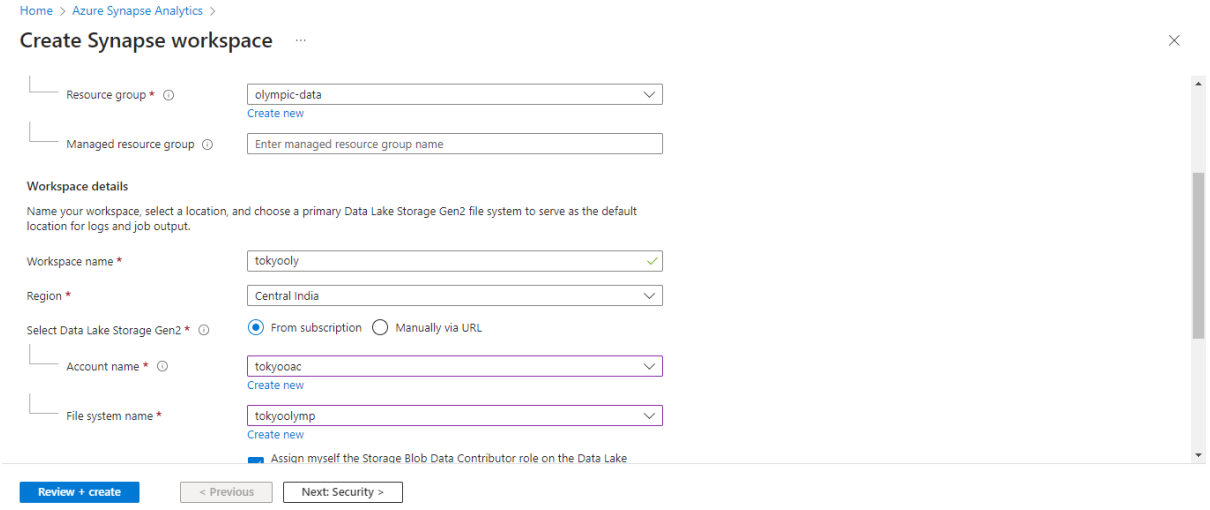




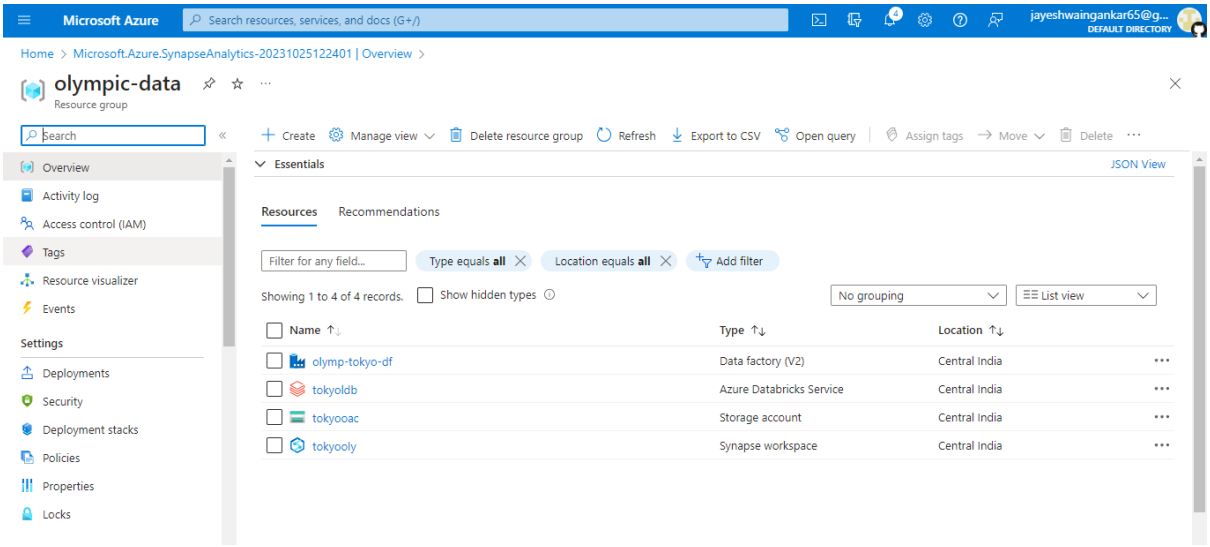
Now we will do Analysis of the data so search Synapse Analytics in azure portal and create it



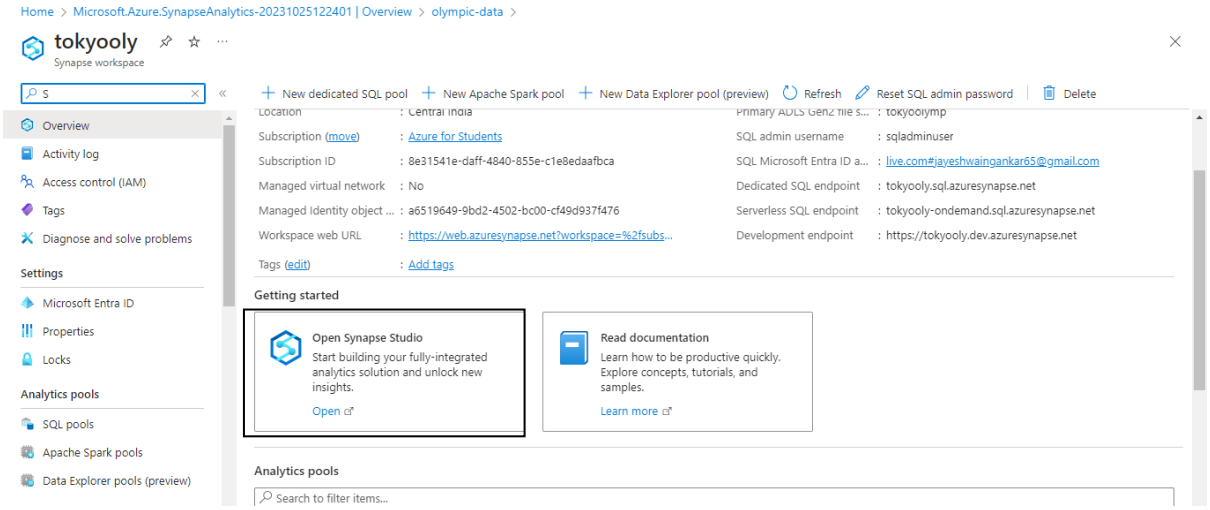
Then review and create



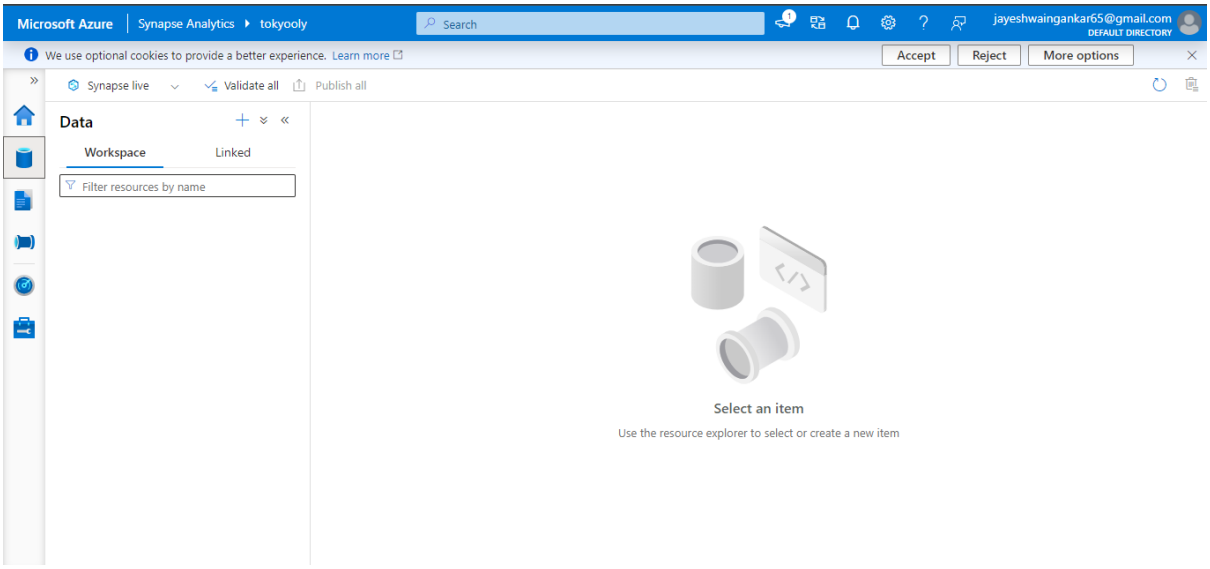
Here we have all our details about our project



Go to overview and then open synapse studio

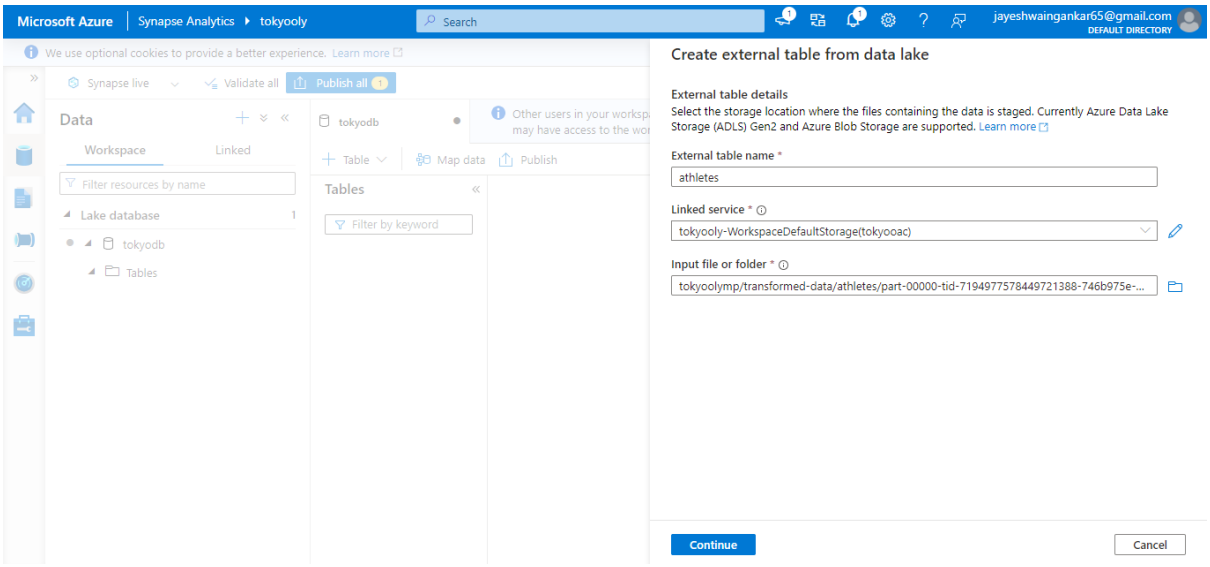


Here go to Data and then click on plus sign and there we will create data

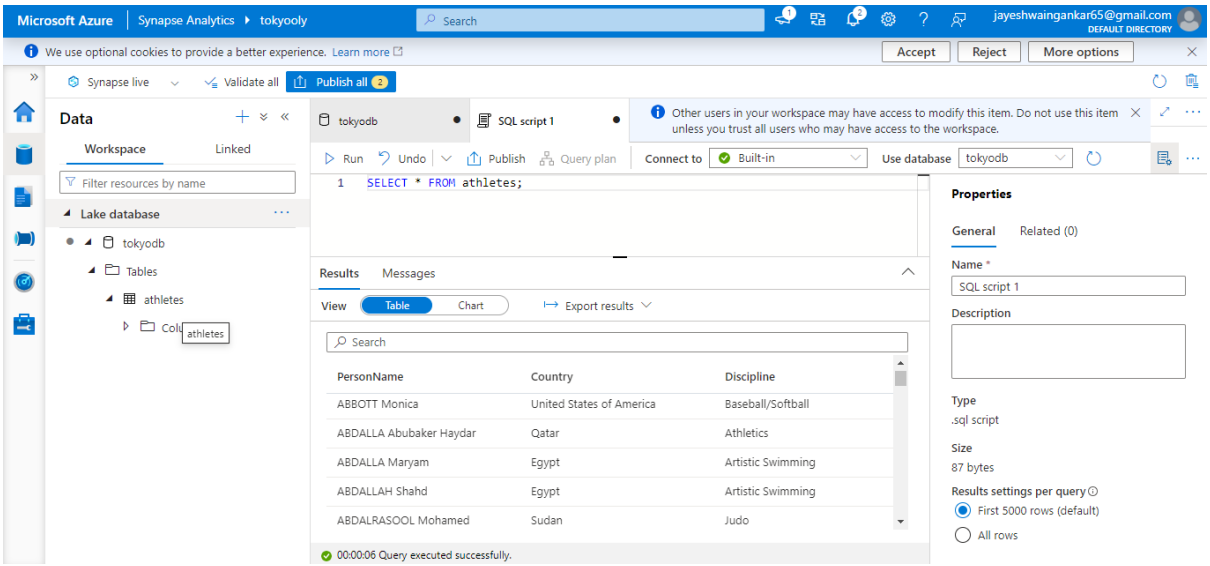


Here we will create table for all 5 files so in folder we will browse the part file here

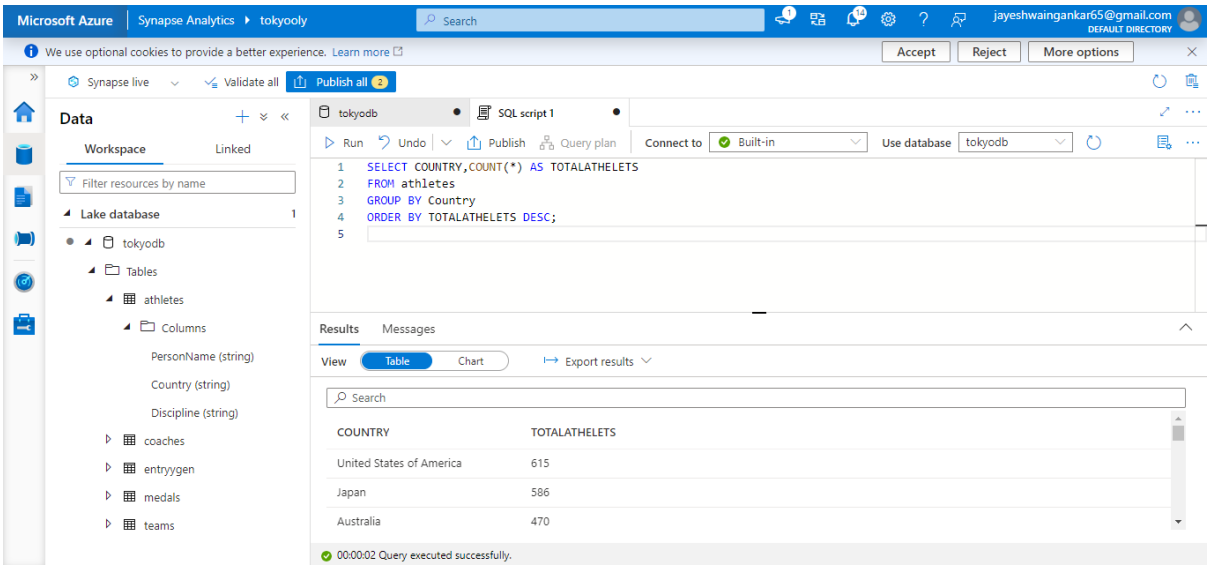
Container >> transformed data >> part file . Do these for all the 5 files



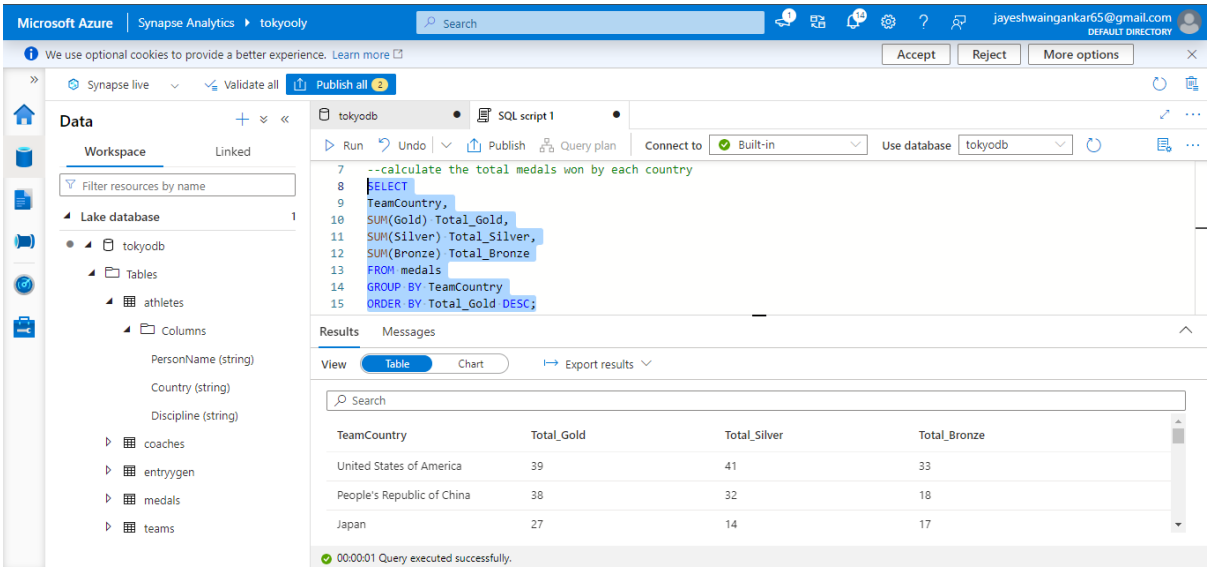
After creating tables for all we will write the SQL query so we will open SQL script and write code. Here we have all the athletes



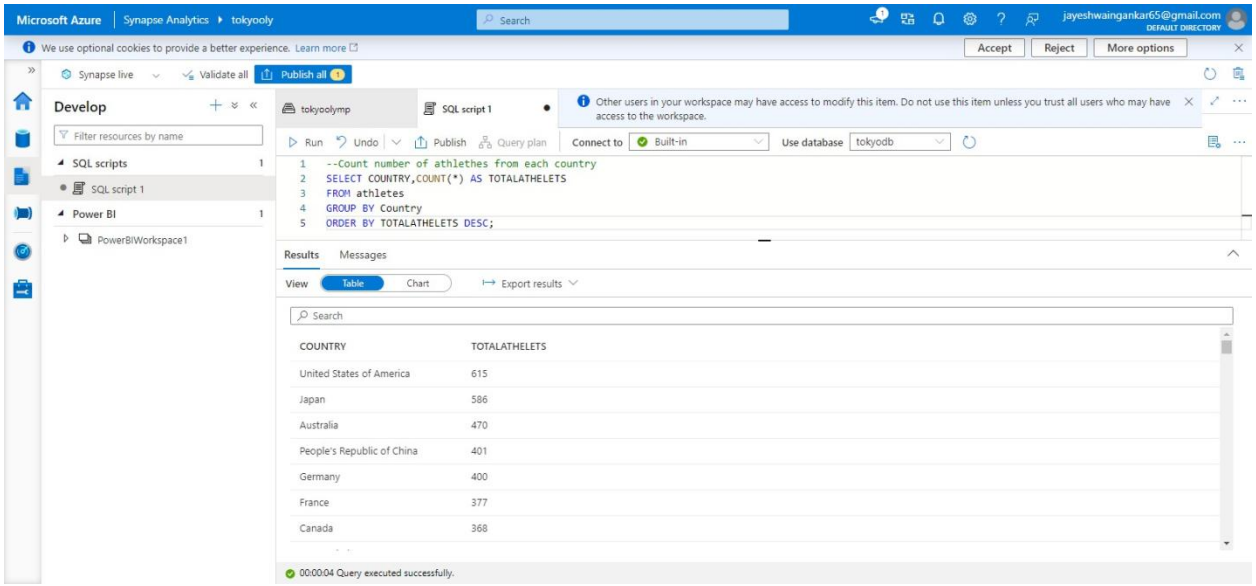
These is the athletes in descending order



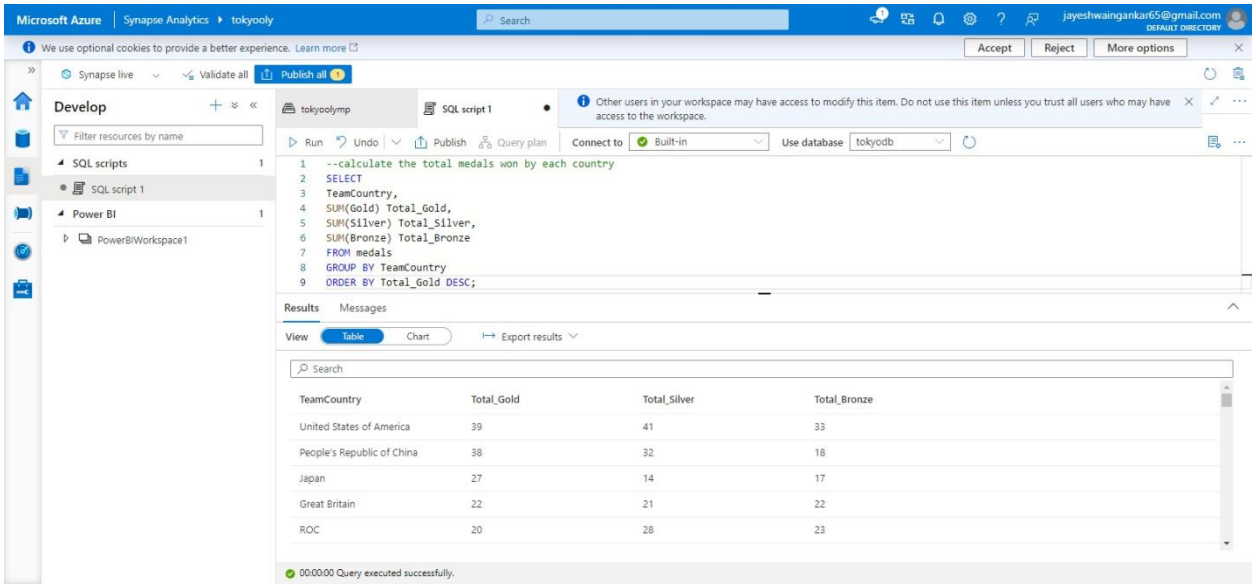
This is the code for medals where we have how many medals are there in total in descending order.



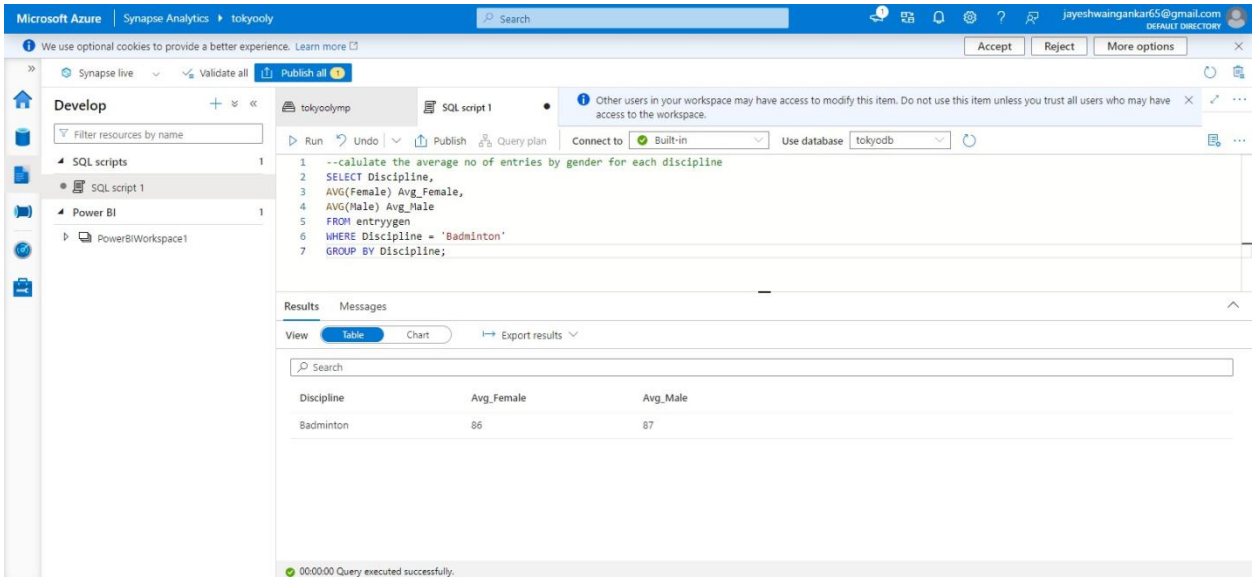
This is count number of athletes from each country



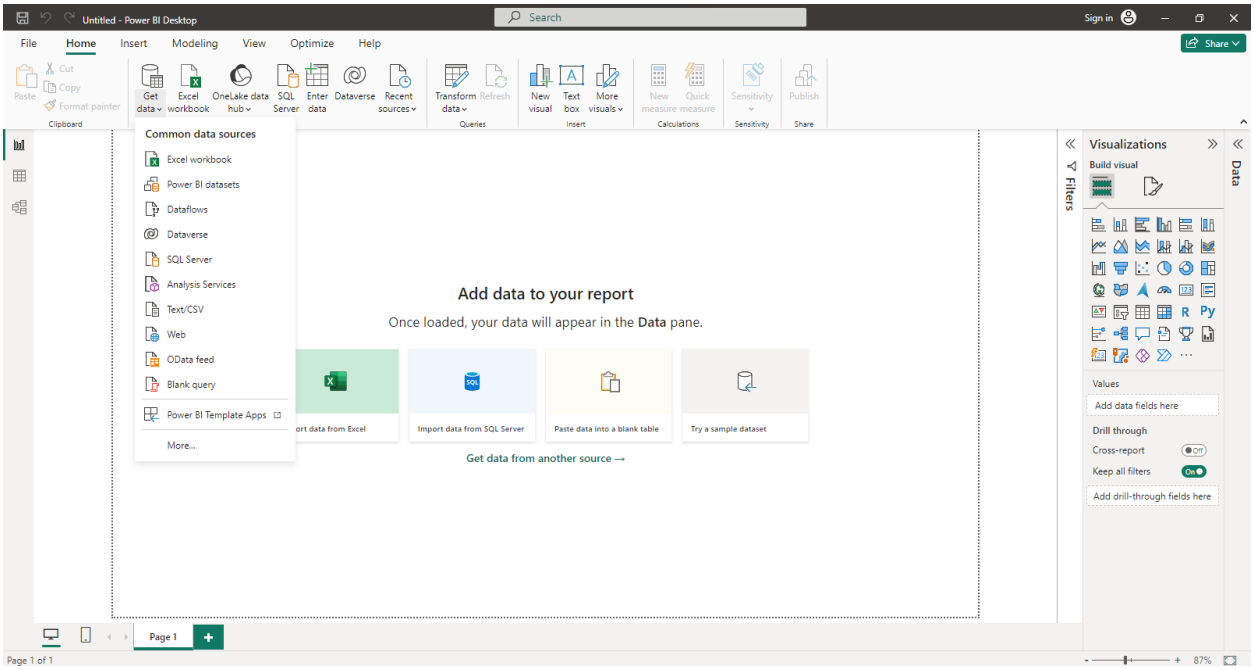
Total number of medals won by each country



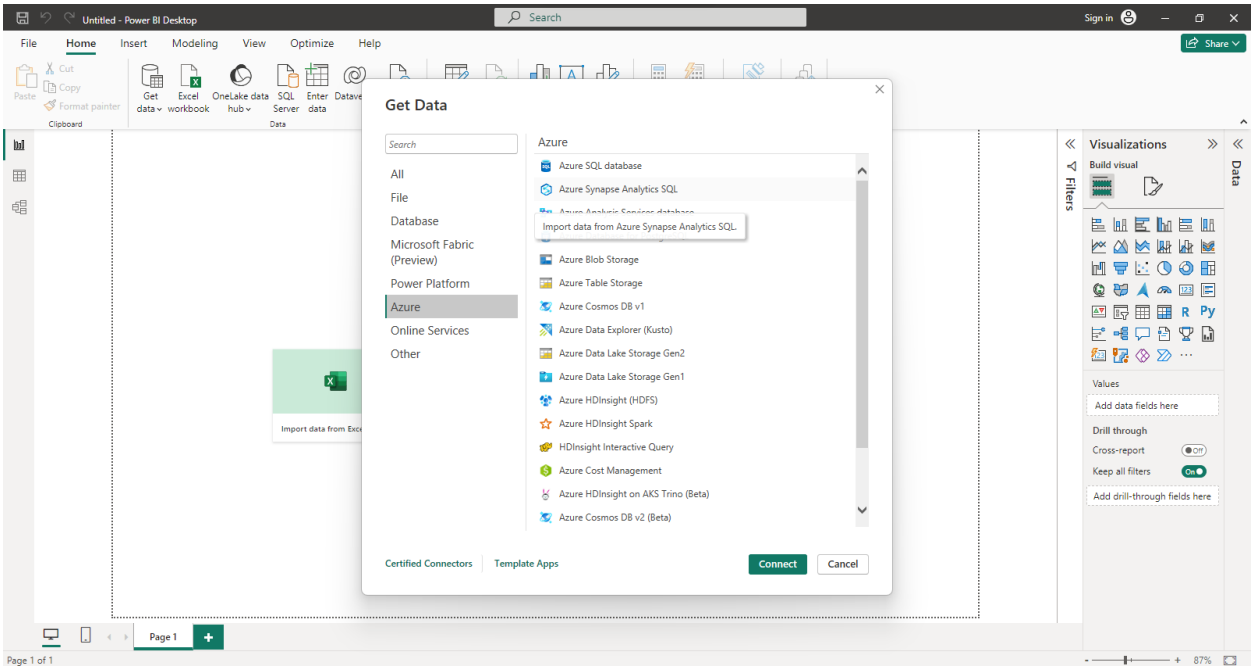
This is the average of entries by gender in badminton. So we can analyze the data using SQL query



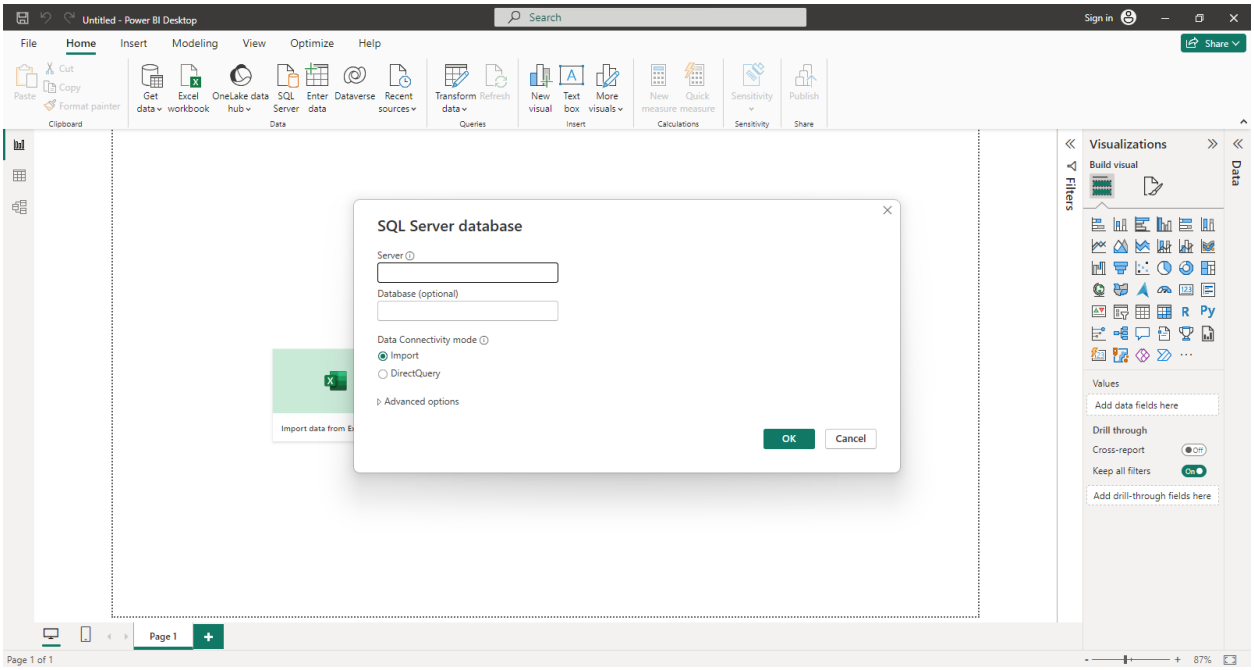
Now we will Do the visualization using Powerbi tool. Select the option get data and click on more



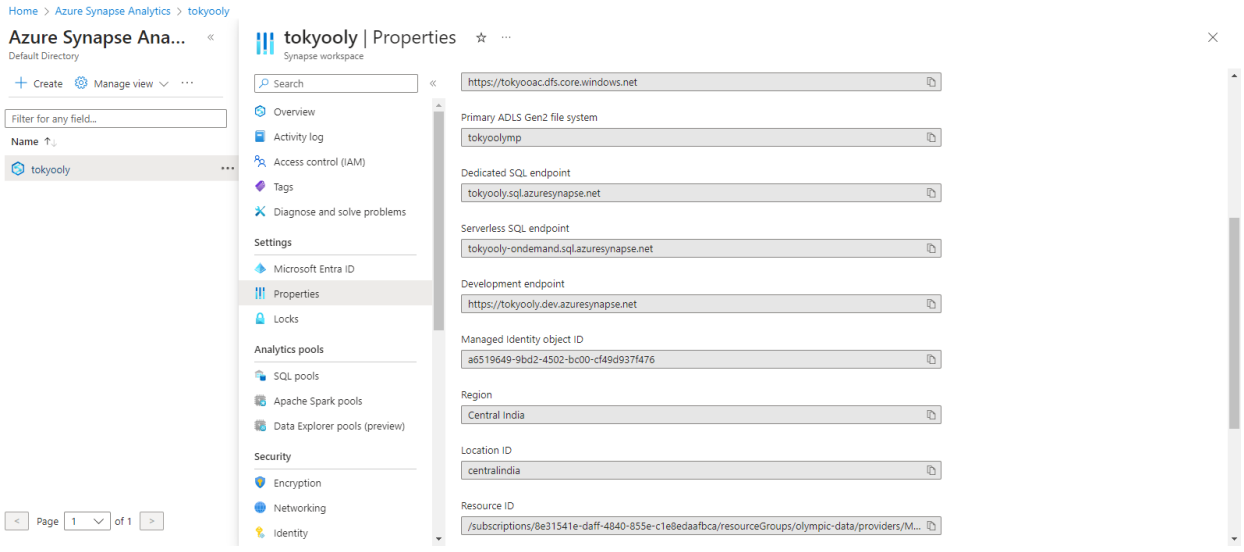
And then select azure synapse analytics SQL and then connect



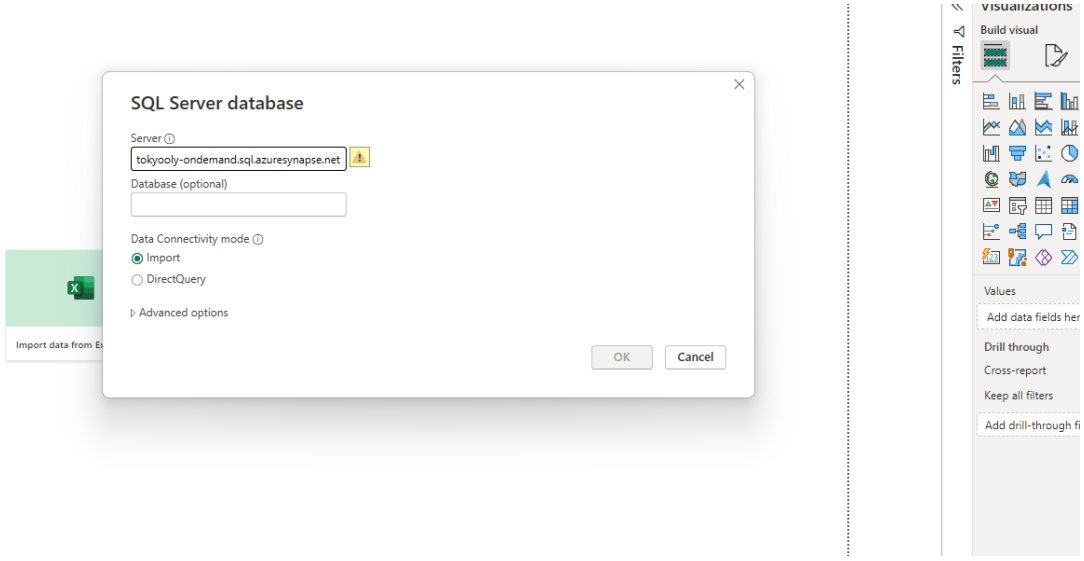
And then enter the credentials



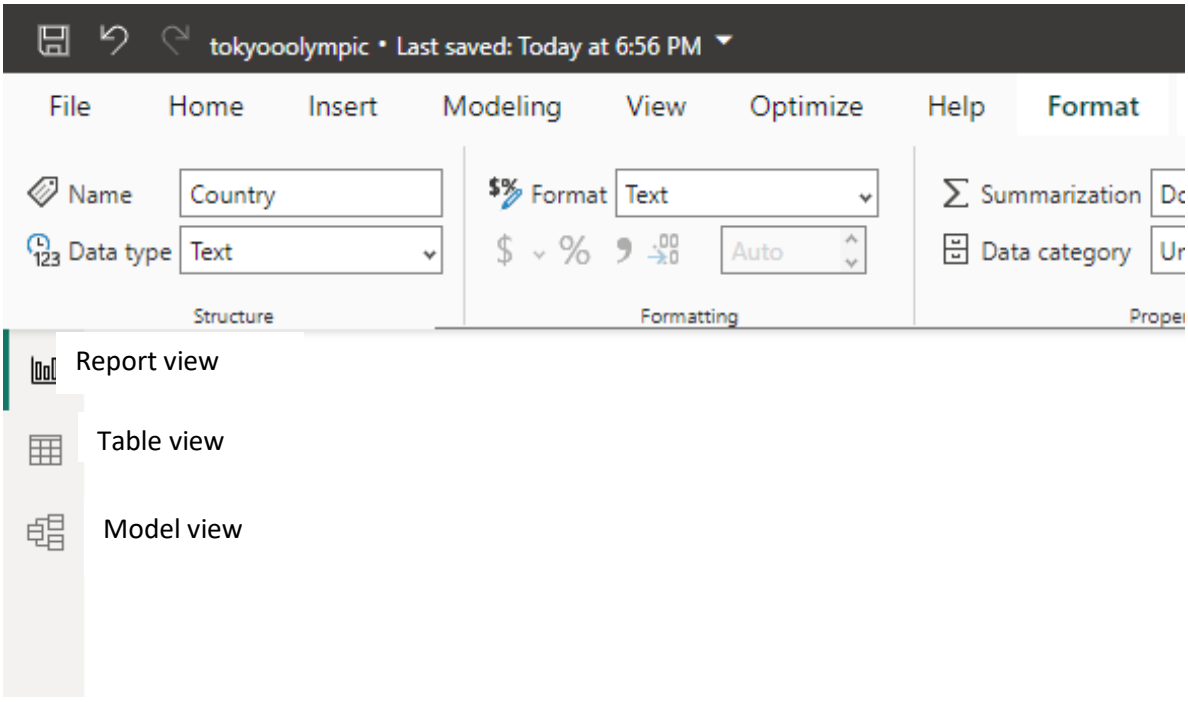
For creadentials go to the portal azure synapse analytics and go to properties and copy the serverless SQL endpoint



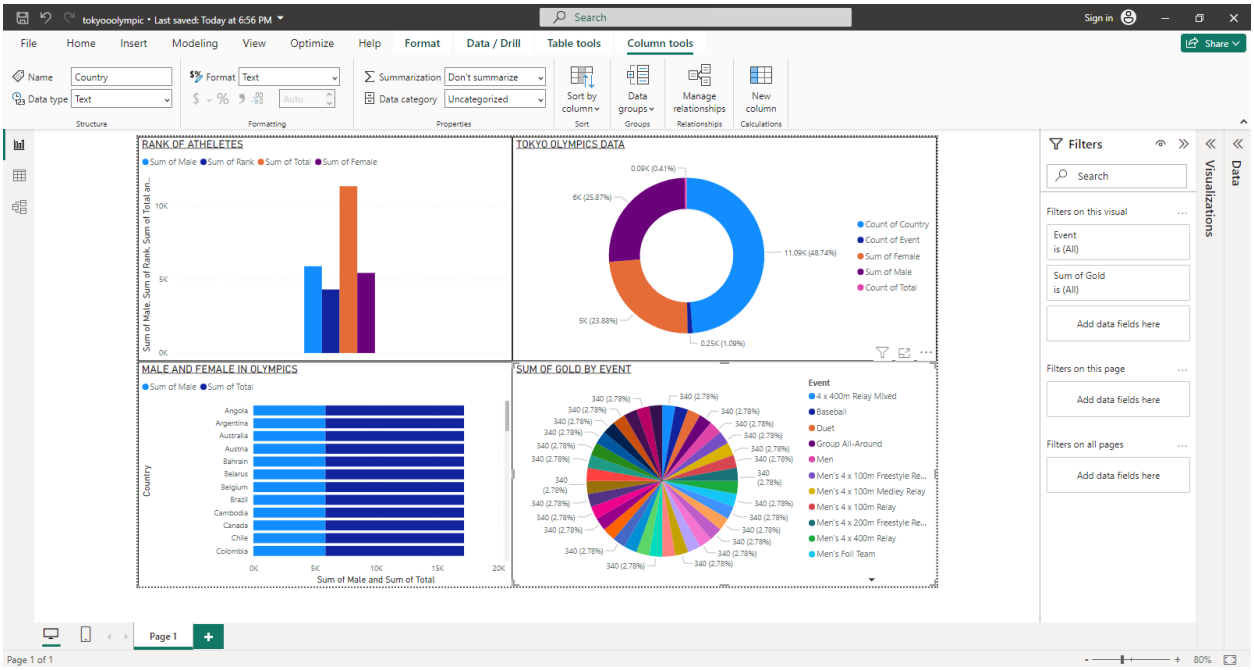
And go to power bi and then enter the copied text and then select import and ok



And then the data will get imported successfully and then create relationship through relationship manager and then make a report through report view



And visualize the data as you want



ADVANTAGES AND CONCLUSION

Advantages:-

1. Comprehensive Data Analysis: The project provides a thorough analysis of Tokyo Olympic data, offering valuable insights into athlete performance, team achievements, and medal distributions.

2. Efficient Data Handling: By leveraging Azure services like Data Factory, Databricks, and Synapse Analytics, the project demonstrates an efficient pipeline for data ingestion, transformation, analysis, and visualization.

3. Scalability: Azure services are designed to handle large volumes of data, making this project easily adaptable to future Olympic events or similar large-scale datasets.

4. Learning Opportunities: The project offers valuable lessons in effective collaboration, data quality assurance, and the integration of Azure services, providing a learning platform for future endeavors.

5. Security and Access Control: Properly configuring access control in Azure Data Lake Storage Gen2 ensures secure data handling, protecting sensitive information.

6. Demonstration of Azure Capabilities: The project showcases the capabilities of Azure services, illustrating their role in handling and analyzing large datasets efficiently.

PROBLEMS FACED:-

1. Access Control Configuration: Setting up proper access control in Azure Data Lake Storage Gen2 required careful attention to ensure secure data handling.

2. Data Integration Complexities: Managing different data formats and ensuring seamless integration from a variety of sources presented initial challenges.

3. Data Quality Issues: Dealing with inconsistent data quality in the raw dataset necessitated careful handling to ensure accurate insights.

5. Power BI Customization: Designing a visually appealing and informative dashboard in Power BI required experimentation and learning to leverage its full potential.

Lesson Learned

1. Effective Collaboration: The project highlighted the importance of clear communication and collaborative efforts between team members, ensuring tasks were completed efficiently.

2. Azure Services Integration: Understanding the capabilities of Azure services and their seamless integration greatly enhanced the efficiency of data processing and analysis.

3. Continuous Learning: Regularly exploring new features and best practices in Azure services and Power BI can lead to more sophisticated and efficient solutions.

Future Enhancement :-

1. Integration with External Data Sources:Include data from external sources like social media analytics, weather data, or economic indicators to provide additional context for the analysis.

2. Cost Optimization Strategies:Explore cost-saving measures for Azure services, such as utilizing serverless computing options or optimizing resource allocation based on usage patterns.

3. Feedback Mechanism:Implement a feedback system to gather user feedback on the dashboard's usability, features, and any additional functionalities they would find valuable.

Conclusion

The "Tokyo Olympic Data Analysis using Azure Services" project successfully demonstrated the power of leveraging Azure services for comprehensive data analysis. By employing a well-structured pipeline, the team efficiently ingested, transformed, and analyzed data, providing valuable insights into the Tokyo Olympic dataset. The resulting Power BI dashboard offers a visually intuitive representation of the data, allowing for easy exploration and interpretation. This project serves as a solid foundation for future data analyses and showcases the capabilities of Azure services in handling large-scale datasets.

BIBLIOGRAPHY

We express our sincere gratitude to all those who helped us in gathering the information while preparing this project. To prepare this project we required information regarding how to use services and implement it in proper way.

References:-

1. https://learn.microsoft.com/en-us/azure/?product=popular
2. https://learn.microsoft.com/en-us/power-bi/
3. https://learn.microsoft.com/en-us/azure/data-factory/
4. https://learn.microsoft.com/en-us/azure/databricks/
5. https://learn.microsoft.com/en-us/azure/synapse-analytics/
6. You tube videos