**Olympic Data Processing using Azure and Databricks**

**Project Overview**

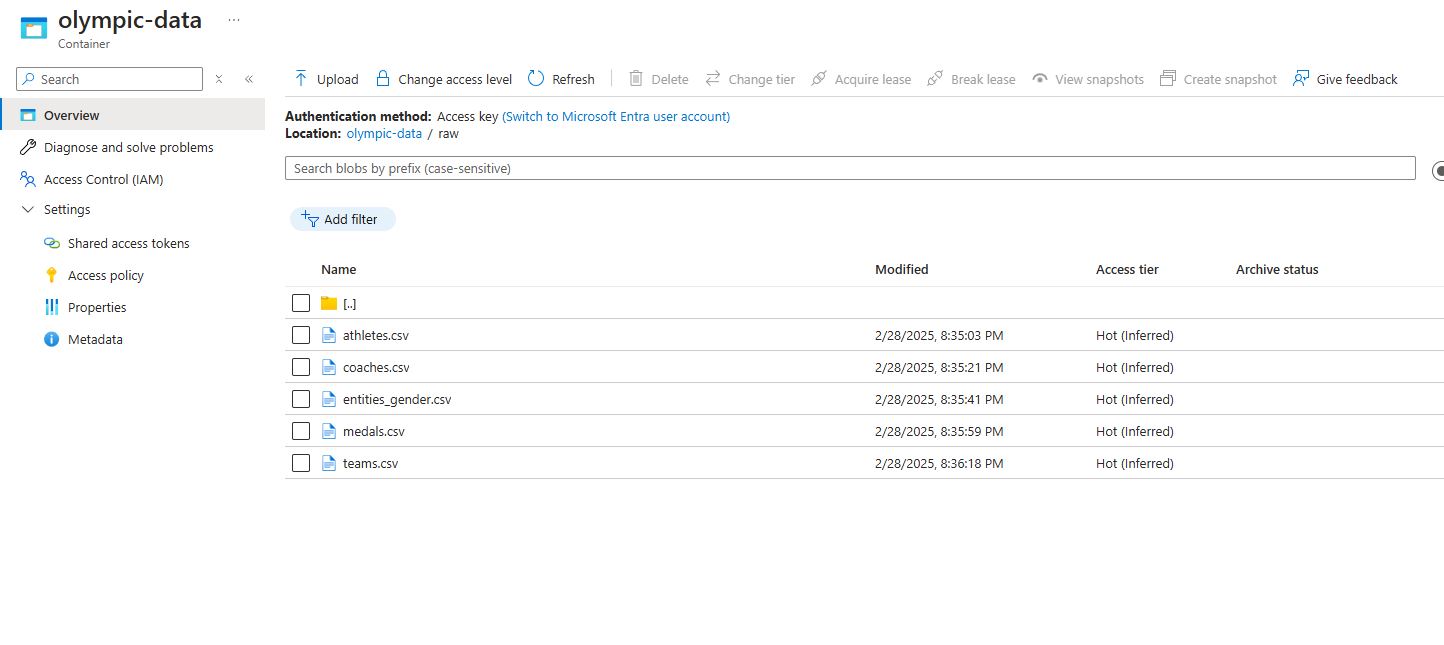
This project involves extracting, transforming, and analysing Olympic data using Azure Data Factory (ADF), Azure Data Lake Storage (ADLS) Gen2, and Databricks. The goal was to process multiple CSV files containing information about Olympic teams, athletes, and medal counts, perform data transformations, and store the processed data back in ADLS Gen2.

**1. Data Storage Overview**

The data used in this project consists of multiple CSV files stored in an **Azure Data Lake Storage (ADLS) Gen2** account. The files contain information on Olympic medals, teams, and rankings.

**Storage Account Details:**

* **Storage Account Name:** tokyolymic
* **Container Name:** olympic-data
* **File Format:** CSV

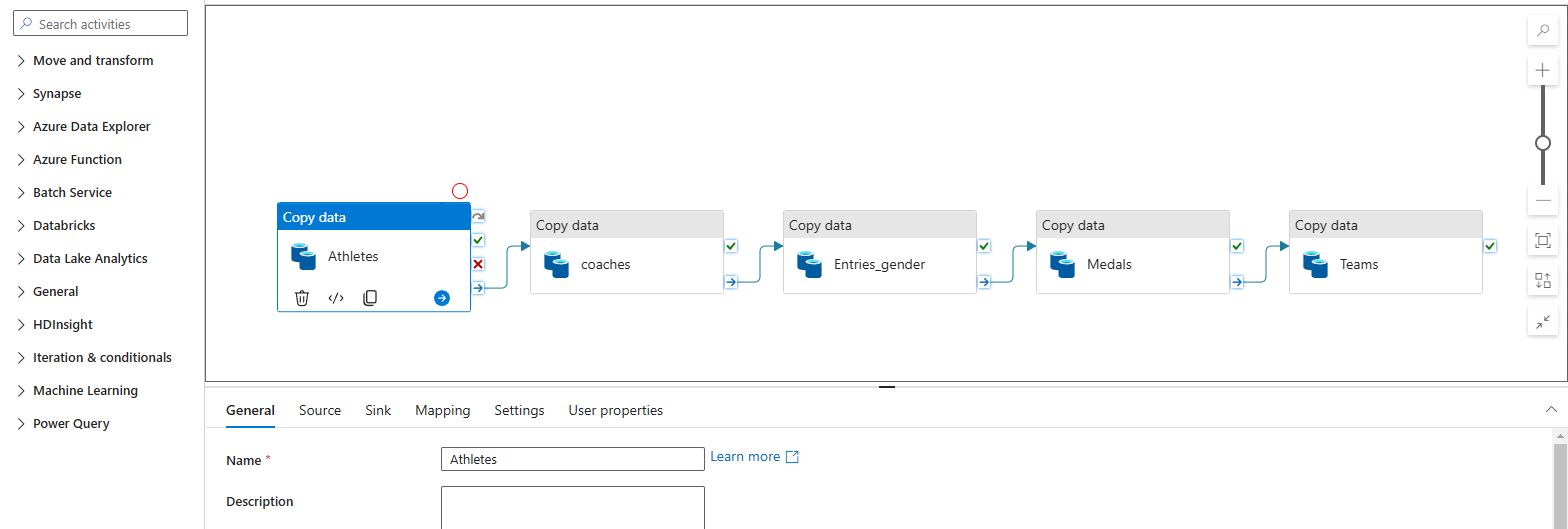


**2. Extracting Data using Azure Data Factory (ADF) Pipeline**

An **Azure Data Factory (ADF) pipeline** was created to extract data from CSV files and load it into **Azure Data Lake Storage (ADLS) Gen2**.

**ADF Pipeline Steps:**

1. **Source:** CSV files stored in the Storage Account.
2. **Copy Activity:** Extracts data from the source and loads it into the destination.
3. **Destination:** ADLS Gen2 (hierarchical namespace enabled).



**3. Mounting ADLS in Databricks**

To access the data from Databricks, we mounted the ADLS Gen2 container using the following command:



**4. Reading CSV Files in Databricks**

Once the data was mounted, we loaded all five CSV files into Spark DataFrames for processing.

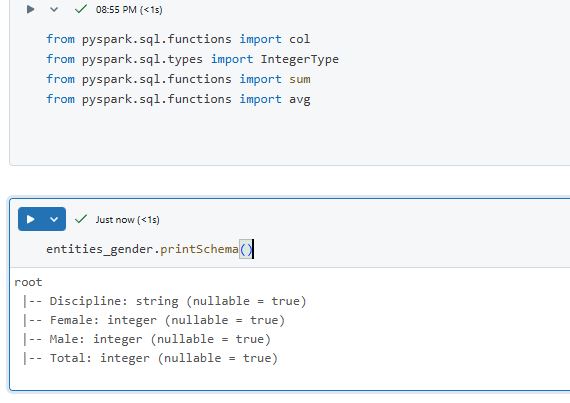
**Reading CSV Files from ADLS Gen2:**



**5. Importing Functions for Data Processing**

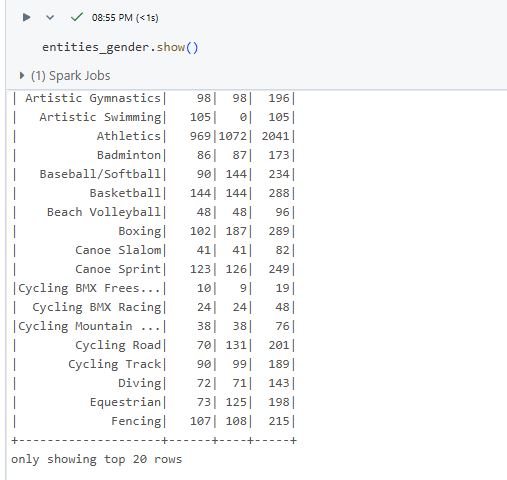
To perform transformations, we first import the necessary functions from pyspark.sql.functions.

**Importing Required Functions:**



**6. Displaying DataFrames**

After reading the data, we display the contents of a DataFrame to verify the data.



**7. Casting Data Types using withColumn**

To ensure the correct data types, we used withColumn to cast specific columns to appropriate types.

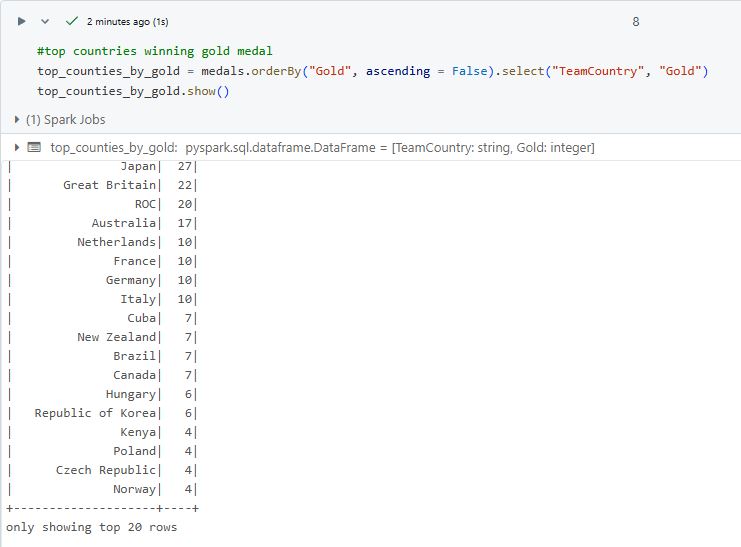
**Casting Columns to Correct Data Types:**



**8. Loading and Transforming Data in Databricks**

After reading the data, we performed some basic transformations.

**Basic Data Transformations:**



**9. Loading and Transforming Data in Databricks**

**Aggregation - Summing Up Medals**



**10. Exporting Transformed Data to ADLS Gen2**

After performing transformations, we exported the processed DataFrame back to ADLS Gen2 in CSV format

