Project Report

End-to-End ETL Data Pipeline and Dashboarding using Azure Cloud Services

CSCI 516 – Engineering Cloud Computing

Sri Padmavathi Manoharan

Spring 2025

Abstract

This project aims to build a comprehensive ETL (Extract, Transform, Load) pipeline using Microsoft Azure cloud services to analyze the Olympic data from Tokyo 2021. The project begins by importing data from GitHub using Azure Data Factory, which automates the entire data pipeline. I will then use Azure Data Lake Gen2 to store the raw data. Next, Azure Databricks will be utilized with PySpark to process, clean, and transform the data. After processing, the structured data will be queried using SQL in Azure Synapse Analytics. To showcase the findings, I created a series of dashboards using Tableau Public. These dashboards highlight various trends, such as medal distribution by country, the number of athletes, and the participation of males and females in sports. This project demonstrates how well-integrated Azure cloud services, combined with the powerful visualization capabilities of Tableau, can provide scalable, automated, and intelligent solutions for real-life problems in sports analytics.

Table of Contents

1. Introduction

In today's era of digital transformation, making decisions based on data has become essential across all sectors. Businesses are increasingly leveraging data analytics to evaluate their operations and strategize for the future. Given the complexity, size, and unstructured nature of modern data, there is a growing need for automation to make data handling more scalable. The advent of platforms like Microsoft Azure has changed how data is collected, processed, stored, and visualized, enabling the end-to-end initiation of processes [1][4][7].

The Extract, Transform, Load (ETL) pipeline is a fundamental component of modern data analytics. This process involves moving and transforming data between the source and the analytical platform. Traditionally, ETL workflows were managed through on-premise tools, which limited scalability and flexibility. However, with advancements in cloud technology, it is now possible to build ETL pipelines that are more efficient and fault-tolerant, integrating various data environments [1][4][8]. Microsoft Azure offers a suite of integrated services that facilitate the creation of these data pipelines [7][9].

This project aims to construct an automated ETL pipeline using Azure cloud services to demonstrate the capabilities of these tools. The chosen dataset for this case study is from the Tokyo 2021 Olympic Games, which is recognized worldwide and is rich in data. The project works with multiple CSV files containing information about athletes, teams, medal counts, gender participation, and coaches. This dataset presents real-life challenges that one might encounter with data—challenges that cloud computing can effectively address [4][6].

The workflow begins by ingesting the raw datasets from a public GitHub repository into Azure Data Factory (ADF). The ADF pipelines then transfer the data to Azure Data Lake Storage Gen2, which is logically organized into two layers: a raw data zone and a transformed data zone. This structured approach effectively manages the data, ensuring proper footprint, staging, and governance [7][9]. Azure Databricks, an analytics platform built on Spark, manages the data transformation step after ingestion. To prepare raw data for analysis, PySpark scripts are used to clean, standardize, and reformat the data into structured representations [4][8].

Once the data is cleaned, structured queries are executed in Azure Synapse Analytics to identify significant trends. The SQL analysis focuses on key topics such as gender representation in sports, medal distribution by nation, and athlete participation [10]. These insights serve as a foundation for further investigation and visualization. Tableau Public is then utilized to visualize Synapse data, enhancing the workflow and facilitating the creation of interactive dashboards that display trends in Olympic performance. These dashboards convert raw data into visually appealing and informative insights, allowing users to explore key parameters and narrate engaging data stories [2][3][5].

2. Problem Statement

The Olympic Games generate massive amounts of data spanning various dimensions—athletes, teams, events, medals, nationalities, gender participation, and more. This data, often collected from multiple sources and stored in semi-structured formats like CSV or JSON, tends to be scattered,

inconsistent, and difficult to analyze directly. Without a proper data management process, deriving insights from such datasets becomes time-consuming, error-prone, and inefficient [1][4].

Several core challenges arise in working with Olympic data:

- Inconsistent formatting across files, such as mismatched column names or varied data types, hinders direct integration.
- Data silos exist when related information (e.g., medals and athlete details) is stored separately, making unified analysis difficult.
- Lack of visualization tools makes it hard for non-technical users to understand key findings from static tables or raw data.
- Manual spreadsheet processing is not scalable for large or evolving datasets, especially when repeatable workflows are needed.

Given these limitations, there is a clear need for an automated, centralized, and cloud-based ETL solution that streamlines data ingestion, transformation, and analysis—while also enabling meaningful visualizations for stakeholders [4][7][9].

This project addresses the above challenges by designing an end-to-end data pipeline using Microsoft Azure Cloud Services. The pipeline begins with automated data ingestion from GitHub using Azure Data Factory, followed by data transformation with Azure Databricks using PySpark [4][8]. Analytical queries are performed in Azure Synapse Analytics to derive insights, which are then visualized in Tableau Public for a broader audience [2][10]. By automating the entire pipeline—from extraction to dashboarding—the project not only saves time and reduces manual errors but also ensures that insights can be updated and scaled effortlessly in the future. This approach demonstrates how cloud technologies can transform complex datasets into reliable, actionable intelligence—especially in fields like sports analytics [4][5].

3. Objectives

The main objective of this project is to design and implement a fully automated ETL (Extract, Transform, Load) pipeline using Microsoft Azure Cloud Services to analyze the Tokyo 2021 Olympic Games dataset. The project addresses key data engineering challenges—such as manual handling, data inconsistency, and lack of insights—through a modern, cloud-based solution [4][7][8].

The specific objectives include:

- Automate data ingestion from a public GitHub repository using Azure Data Factory, enabling reliable and repeatable data flow into Azure Data Lake [7][9].
- Clean and transform datasets using Azure Databricks and PySpark, addressing inconsistencies, null values, and converting raw data into structured formats for analysis [4][8].
- Store structured outputs in a dedicated “transformed data” zone in Azure Data Lake Gen2 and enable analytical querying via Azure Synapse Analytics [7][10].

- Build interactive dashboards using Tableau Public to visualize key metrics such as medal counts, gender participation, and athlete distribution, supporting intuitive exploration of Olympic data [2][3][5].
- Showcase the value of cloud analytics by demonstrating how Azure's integrated services can efficiently solve real-world data problems with scalability and automation [1][4][10].

These objectives align with the broader goal of turning raw, scattered Olympic datasets into meaningful insights—delivered through a seamless and repeatable cloud-based analytics pipeline.

4. Literature Review

As data continues to expand in volume, variety, and velocity, cloud computing has become essential for building efficient and scalable data pipelines. The use of cloud-based big data technologies for Extract, Transform, Load (ETL) processes has been explored by Oladimeji, who compares AWS and Microsoft Azure in enterprise environments. The study shows that Azure, with its deeply integrated services, provides greater flexibility and scalability for data engineering workflows [1]. Microsoft Azure's ecosystem supports the complete data pipeline lifecycle—from ingestion to transformation and analysis—making it a strong choice for large-scale data analytics [1][4].

One of the most widely used services in Azure is Azure Data Factory (ADF). It is a fully managed cloud service that enables the creation of automated workflows for moving and transforming data across various sources. According to Microsoft documentation, ADF supports both batch and streaming pipelines, integrating easily with Azure Data Lake Storage Gen2 and other services like Azure SQL Database and Azure Synapse Analytics [7][9].

The transformation stage is enhanced by Azure Databricks, which combines Apache Spark's distributed computing power with the collaborative features of Azure. Thoutam emphasizes that Azure Databricks is particularly effective for transforming large datasets in real time, supporting Python, SQL, and Spark APIs within a unified workspace [4]. The integration with Delta Lake ensures ACID-compliant data storage, which is vital for repeatable and reliable analytics processes [8].Once transformed, data is analyzed using Azure Synapse Analytics, a cloud-native data warehouse and analytics service. Synapse allows SQL-based querying of structured data at scale and provides connectors to Power BI and Tableau for visualization. According to Microsoft, it supports on-demand query execution, parallel processing, and advanced visualization, which helps in identifying trends across large Olympic datasets [10].

In the final phase of the pipeline, insights are communicated through data visualization. Tableau has become a leading choice for building interactive dashboards due to its ease of use, rich features, and compatibility with cloud storage systems. Parthe compares Tableau and Power BI, concluding that Tableau is better suited for flexible, highly visual storytelling and analytical exploration, especially in academic and research contexts [2]. Martins reinforces this by focusing on dashboard design principles, noting that effective dashboards should reduce cognitive load and highlight key metrics without overwhelming users [3].

Recent research by Purich explores how dashboards on platforms like Tableau Public reflect real-world analytical needs and user behavior. This study supports the idea that public dashboards can be powerful tools for educational and collaborative analytics [5]. Additionally, demonstrates how Tableau can integrate with APIs and geographic data to create layered visualizations, reinforcing its capability to handle spatial and statistical data simultaneously [6].

Overall, the literature confirms that cloud services like Azure and visualization platforms like Tableau can be successfully combined to build powerful end-to-end analytics pipelines. This project draws upon these insights, applying them to analyze the Tokyo 2021 Olympic dataset through a cloud-native, scalable, and automated ETL pipeline. The implementation reflects current best practices in data engineering, cloud integration, and dashboard design, offering a real-world example of modern analytics in action.

5. Dataset Description

The dataset used for this project was sourced from a public GitHub repository**,** originally published on Kaggle, and contains comprehensive records related to the 2021 Tokyo Olympic Games. The dataset is publicly available and includes five CSV files that capture various dimensions of the Olympic event such as athletes, teams, events, coaches, and medal distributions. These files offer a rich and diverse structure for building a meaningful data pipeline and performing analytics.

Each file in the dataset serves a specific purpose:

- **athletes.csv**
  This file contains detailed information about each individual athlete, including their name, nationality (NOC)**,** gender, and discipline (the sport in which they competed). It forms the foundation for athlete-level analysis.
- **coaches.csv**
  This file includes data on Olympic coaches, such as the name of the coach, the sport they are associated with, and their country. It enables insights into coaching participation across sports and regions.
- **entriesgender.csv**
  This file outlines gender-wise participation statistics for each Olympic discipline. It lists the number of male and female athletes participating in each event category, helping assess gender representation across the games.
- **medals.csv**
  This dataset captures the total number of, gold**,** silver and bronze medals won by each country. It also includes the overall medal count and ranking based on both gold medals and total medals won. This file is critical for performance analysis by nation.
- **teams.csv**
  The teams file includes information about national teams**,** event types, and other group-level participation details. It offers a broader view of team entries beyond individual athletes.

Together, these five files offer a multi-dimensional and semi-structured dataset**.** Since the files originated from different sources and varied in formatting, they required extensive preprocessing

before analysis. Tasks such as column renaming, handling missing values, and converting data types were performed using PySpark in Azure Databricks to ensure consistency and structure across all datasets.

By combining these data sources, the project was able to perform deep and interactive analysis, leading to valuable insights about country-level performance, gender distribution in sports, and Olympic participation trends.

6. Tools and Technologies Used

To implement the end-to-end ETL pipeline and analytics workflow, this project leverages a combination of Microsoft Azure cloud services and Tableau Public for visualization. Each tool plays a specific role in the data pipeline, contributing to a fully automated, scalable, and efficient solution. The table below summarizes the tools and their purposes:

Tool	Purpose
Azure Data Factory	Automate ingestion from GitHub
Azure Data Lake Gen2	Store raw and transformed data
Azure Databricks	Clean and transform data using PySpark
Azure Synapse Analytics	Query data using SQL
Tableau Public	Visualize results as interactive dashboards

6.1 Azure Data Factory (ADF)

ADF is a cloud-based ETL and data integration service used to automate the data ingestion process. In this project, ADF connects directly to GitHub to retrieve the Tokyo Olympics dataset. It schedules and manages pipelines that move raw CSV files from the source into the raw zone of Azure Data Lake Gen2. ADF ensures a repeatable, error-free, and fully automated ingestion process. According to Microsoft documentation, ADF supports various data sources and integrates with other Azure services for scalable pipeline execution [7][9].

6.2 Azure Data Lake Storage Gen2 (ADLS Gen2)

ADLS Gen2 is used for cloud-based storage of both raw and transformed data. It provides high-performance hierarchical file storage that integrates seamlessly with Azure analytics tools. Two logical folders were created: a raw-data folder to store the unprocessed files from GitHub, and a transformed-data folder to store the cleaned and structured outputs from Databricks. This approach ensures proper data zoning and governance [9].

6.3 Azure Databricks

Azure Databricks is a collaborative Apache Spark–based platform used in this project to perform data transformation tasks. Using PySpark, the data was cleaned, filtered, standardized, and formatted — including removing nulls, renaming columns, and converting data types. Databricks handles large-scale data efficiently and integrates with Delta Lake for ACID-compliant storage. This makes it ideal for real-time and batch processing pipelines, as supported in the literature by Thoutam [4] and Microsoft [8].

6.4 Azure Synapse Analytics

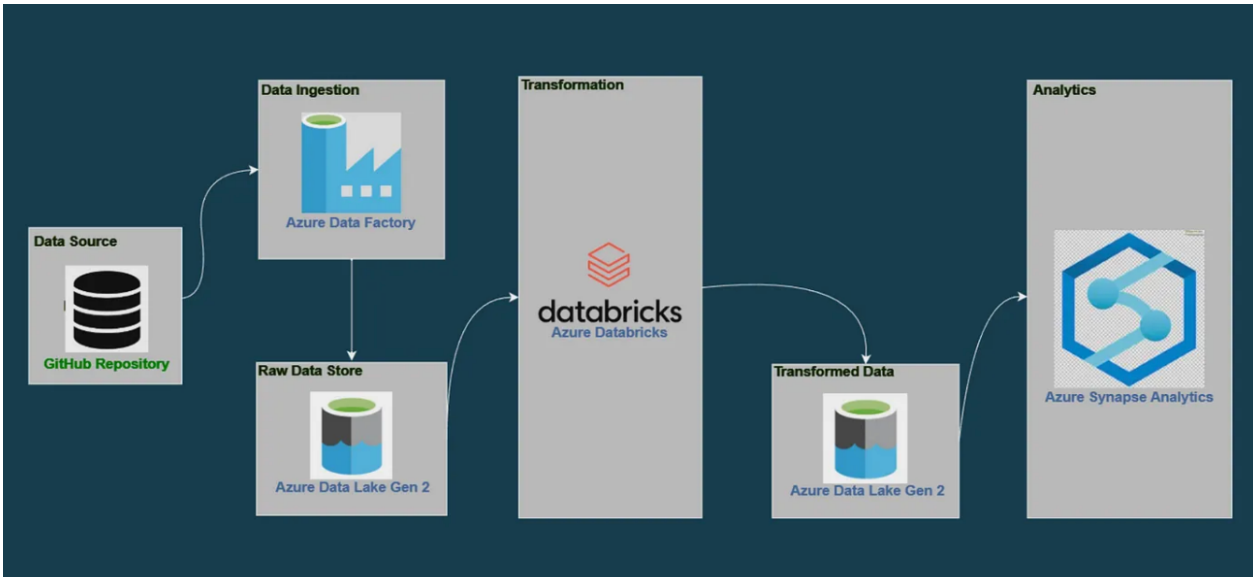
Azure Synapse Analytics enables querying and analyzing the transformed data using T-SQL. It connects directly to files stored in ADLS Gen2, allowing SQL-based analysis without the need for database import. The project used Synapse to generate insights such as total medals by country, gender participation, and discipline-wise athlete counts. As Microsoft notes, Synapse supports advanced querying, visualization integrations, and high-speed analytics across large datasets [10].

6.5 Tableau Public

Tableau Public is a widely-used visualization platform that allows users to create interactive dashboards. After data transformation and analysis, the cleaned CSV files were imported into Tableau. The dashboards developed included bar charts, treemaps, and filters to make Olympic data insights accessible to non-technical users. Comparative studies by Parthe [2] and design principles suggested by Martins [3] highlight Tableau's strengths in visual storytelling and user-centric reporting.

7. Methodology

The methodology follows a step-by-step ETL and visualization process:



Stage 1: Data Ingestion

The data ingestion phase of the project was carried out using Azure Data Factory (ADF), a cloud-based ETL service designed to move and integrate data from various sources. ADF pipelines were configured to extract the required CSV files directly from a GitHub repository hosting the Tokyo 2021 Olympic datasets. These pipelines utilized the HTTP connector in ADF to access the raw data links for each file, such as athletes.csv, medals.csv, and others. Once the source URLs were connected, ADF handled the automated transfer of data into Azure Data Lake Storage Gen2, where a dedicated "raw" folder was created to store the incoming files. This process eliminated the need for manual downloads and ensured that data ingestion was repeatable, scalable, and well-organized, setting the foundation for the subsequent transformation phase [7][9].

Stage 2: Data Transformation

Once the raw data was successfully ingested into Azure Data Lake Storage, the next stage focused on data transformation using Azure Databricks. A Databricks workspace was set up, and a Spark cluster was launched to enable parallel processing of the datasets. Using PySpark, each CSV file was read from the “raw” folder and subjected to various transformation steps. These included removing null or missing values, renaming inconsistent column headers, and casting data into appropriate types to ensure consistency across files. For example, numerical values such as medal counts were converted to integer types, while categorical data like country names and sports disciplines were standardized. After transformation, the cleaned and structured data was written back into a new “transform” folder within the same Data Lake. This separation of raw and processed data ensured traceability and clear data lineage throughout the pipeline [4][8].

Stage 3: Data Analysis

After the transformed data was stored in Azure Data Lake, the next step was to perform analysis using Azure Synapse Analytics. In this stage, external tables were created in Synapse that directly referenced the cleaned CSV files stored in the “transformed” folder of Azure Data Lake Gen2. This allowed the data to be queried without physically moving it into a database, making the process efficient and scalable. Using SQL queries, exploratory analysis and aggregations were performed to extract insights such as total medal counts by country, athlete participation by gender, and discipline-wise performance. These queries served as the foundation for generating summary statistics and trends, which were later used for data visualization in Tableau [10].

Stage 4: Dashboarding

In the final stage of the project, the analyzed data from Azure Synapse was exported as CSV files to be used for visualization. These CSVs contained summarized outputs such as medal counts, athlete participation, and gender-based statistics, which were then imported into Tableau Public. Tableau was used to design and build interactive dashboards that visually represent Olympic insights in a user-friendly format. The dashboards included visualizations such as bar charts, heatmaps, and filters to allow dynamic exploration of medal distribution by country, gender participation across sports, and athlete representation. This stage effectively translated raw

analytical results into engaging visual stories, making it easy for non-technical users to interpret and explore the Olympic data [2][3][5].

8. Implementation

The implementation of this project involved configuring and connecting various Microsoft Azure services to establish a smooth and automated data pipeline. The process began in Azure Data Factory, where a pipeline was designed to pull multiple CSV files directly from a public GitHub repository using the HTTP source option. These files were then saved into the "raw-data" folder within Azure Data Lake Storage Gen2, which was set up with a hierarchical namespace to support structured file access and folder-based organization [7][9].

Two main folders: raw and transform were created to separate the original files from the processed outputs. Next, in Azure Databricks, a compute cluster was initialized, and a PySpark notebook was developed to perform data transformation tasks. The raw CSVs were loaded from Data Lake, cleaned by removing null values, standardized through column renaming and formatting, and then written into the transformed-data folder. This transformation ensured the data was ready for structured querying [4][8].

Once the data was prepared, Azure Synapse Analytics was used to create external tables directly referencing the transformed files. SQL queries were executed to perform aggregations and extract meaningful insights such as country-wise medal tallies and gender-based participation [10]. Finally, the processed datasets were exported as CSV files and loaded into Tableau Public, where multiple interactive dashboards were built. These dashboards included visualizations for total medals by country, athlete participation by gender, and discipline-level comparisons, offering an intuitive way to explore and present Olympic insights [2][3][5].

9. Results and Analysis

The successful implementation of the ETL pipeline is clearly demonstrated through the visual outputs and execution logs from Azure Data Factory. The project was initiated by setting up a data ingestion pipeline that connected to publicly hosted CSV files on GitHub. Using the Copy Data activities in Data Factory, five different datasets—Athletes, Coaches, EntriesGender, Medals, and Teams—were extracted and loaded into Azure Data Lake Gen2.

In Fig: 1 the Azure Data Lake Storage Gen2 container named `tokyogopidemo`, specifically showcasing the `Raw` directory. It contains five CSV files (`Athletes.csv`, `Coaches.csv`, `EntriesGender.csv`, `Medals.csv`, and `Teams.csv`), all stored in block blob format under the *Hot (Inferred)* access tier. This raw zone serves as the initial landing area for ingested datasets before they undergo transformation in the ETL pipeline.

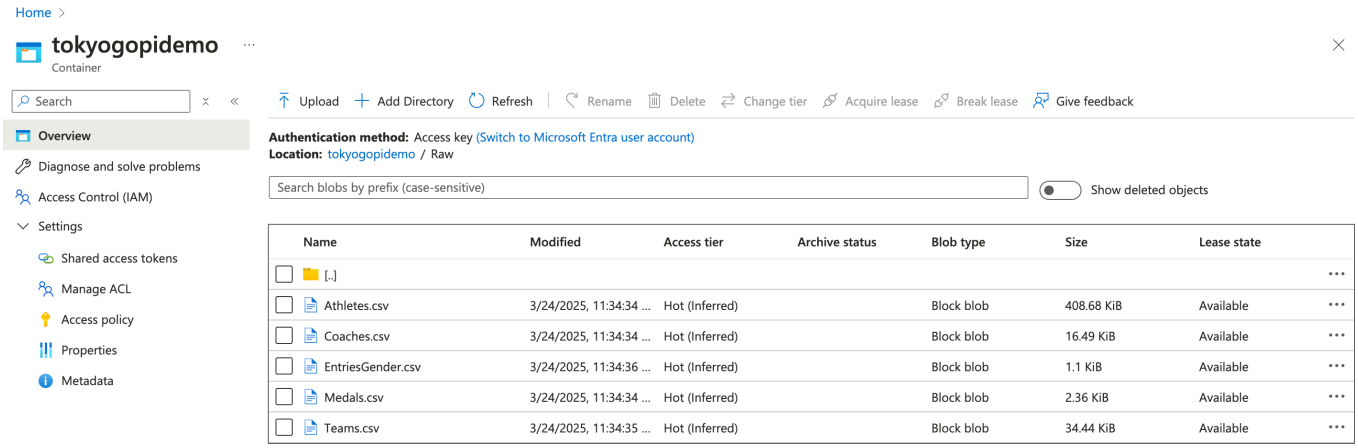


Fig 1: Azure Blob Storage – Raw Zone in Container

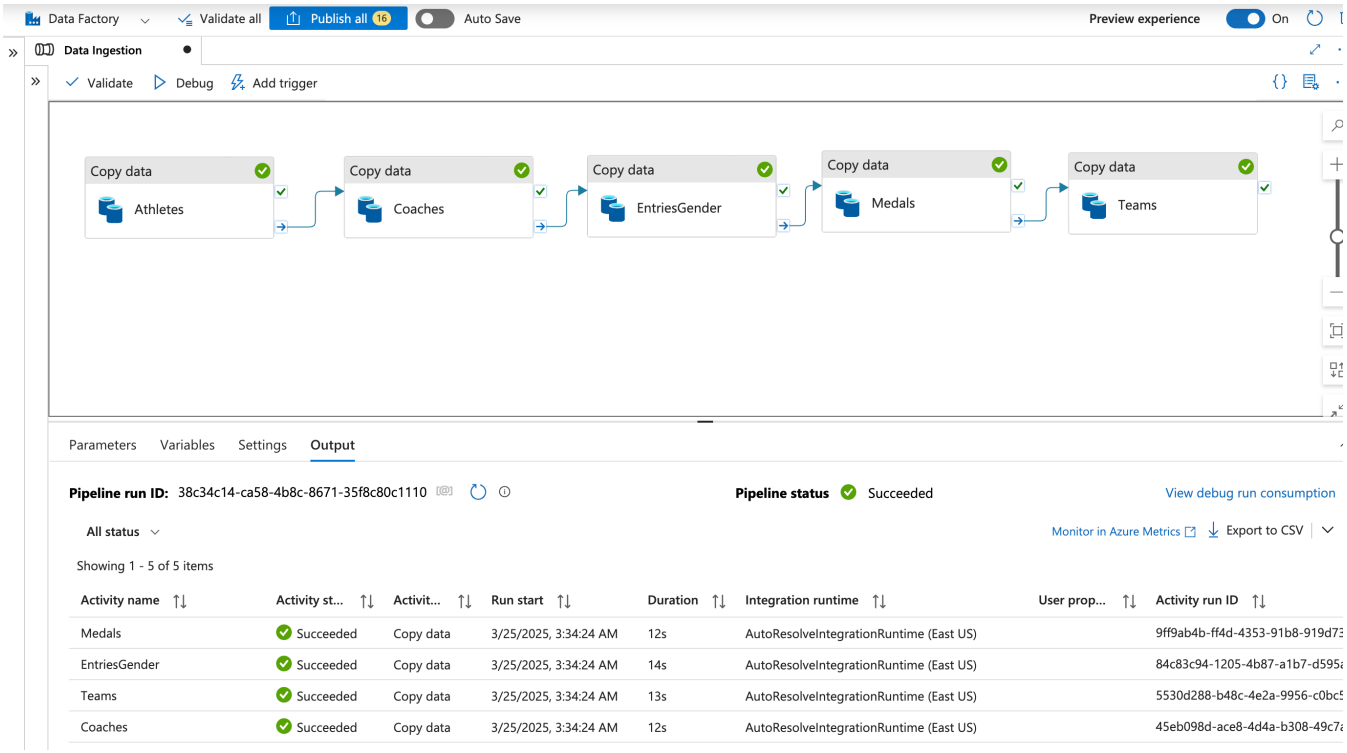


Fig 2: Azure Data Factory Pipeline for Data Ingestion

In Fig: 2 Azure Data Factory Pipeline for Data Ingestion, completed pipeline in Azure Data Factory designed for automated data ingestion. Each activity—such as copying data from source CSVs (`Athletes`, `Coaches`, `EntriesGender`, `Medals`, `Teams`)—has successfully executed as indicated by the green check marks. The output section confirms that all data movement tasks succeeded with low latency (12–14 seconds each), using the AutoResolveIntegrationRuntime in the East US region. This visual validates the smooth execution of the ETL pipeline's ingestion phase.

Once the raw data was available in the "raw-data" folder in Azure Data Lake Storage Gen2 in the next stage involved building a Data Flow within Azure Data Factory to manage transformations.

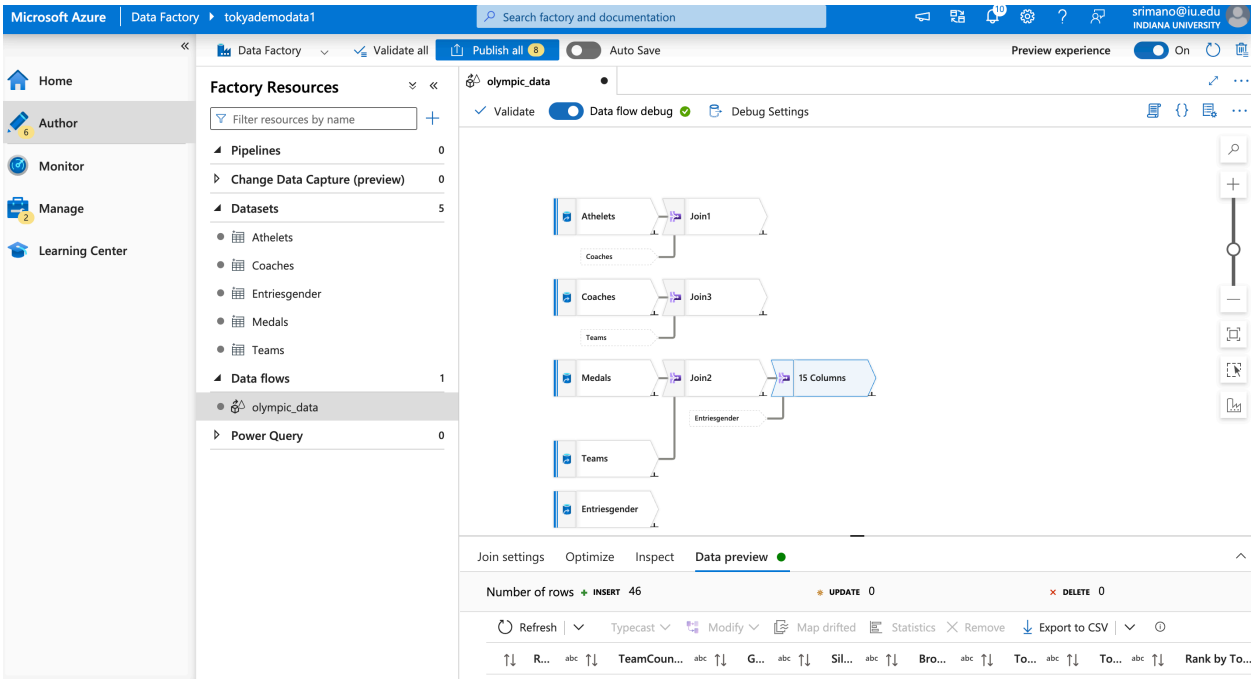


Fig: 3 Data Flow for transformation

The visual flowchart highlights in Fig: 3 Data Flow for transformation, how datasets were joined using keys such as Team, Country and Discipline, merged appropriately to prepare a unified dataset. This setup ensured consistent formatting, column standardization, and clean joins.

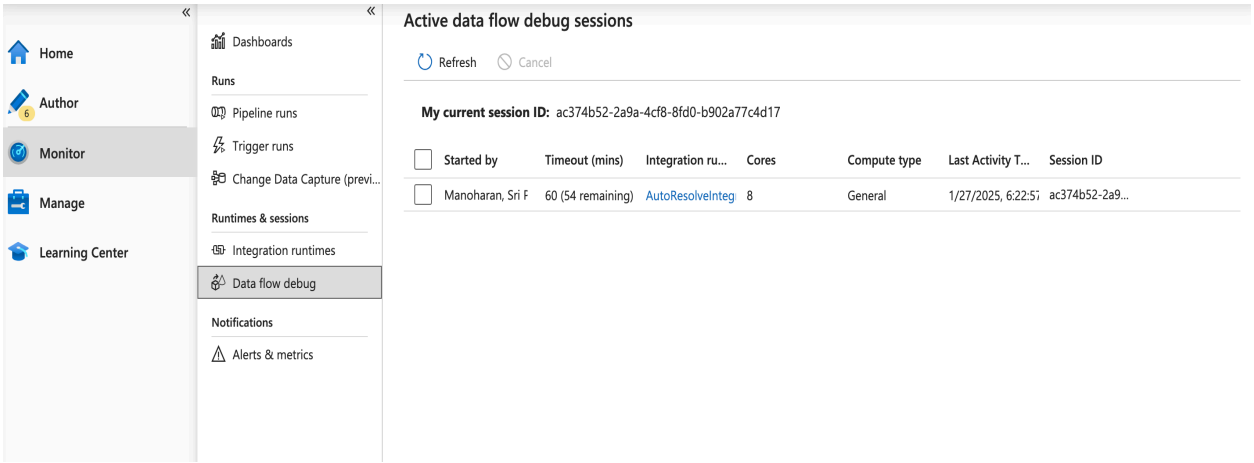


Fig:4 Active Data Flow Debug Session

The data flow debug session was successfully activated in Azure Data Factory to test and validate the pipeline transformations in real-time Shown in Fig: 4.

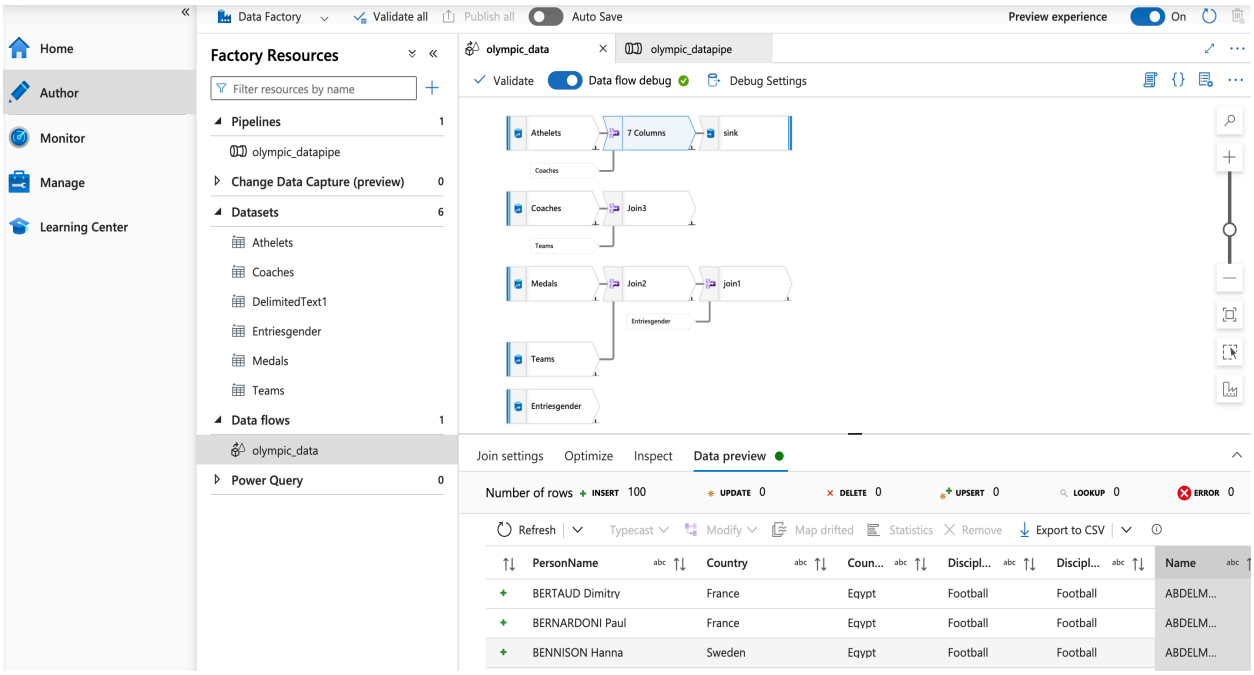


Fig: 5 Data Flow Design for Olympic Dataset

In Fig:5 the live data preview within the Data Flow pane confirms the integration of multiple datasets, displaying a cleaned and consolidated table ready for analysis.

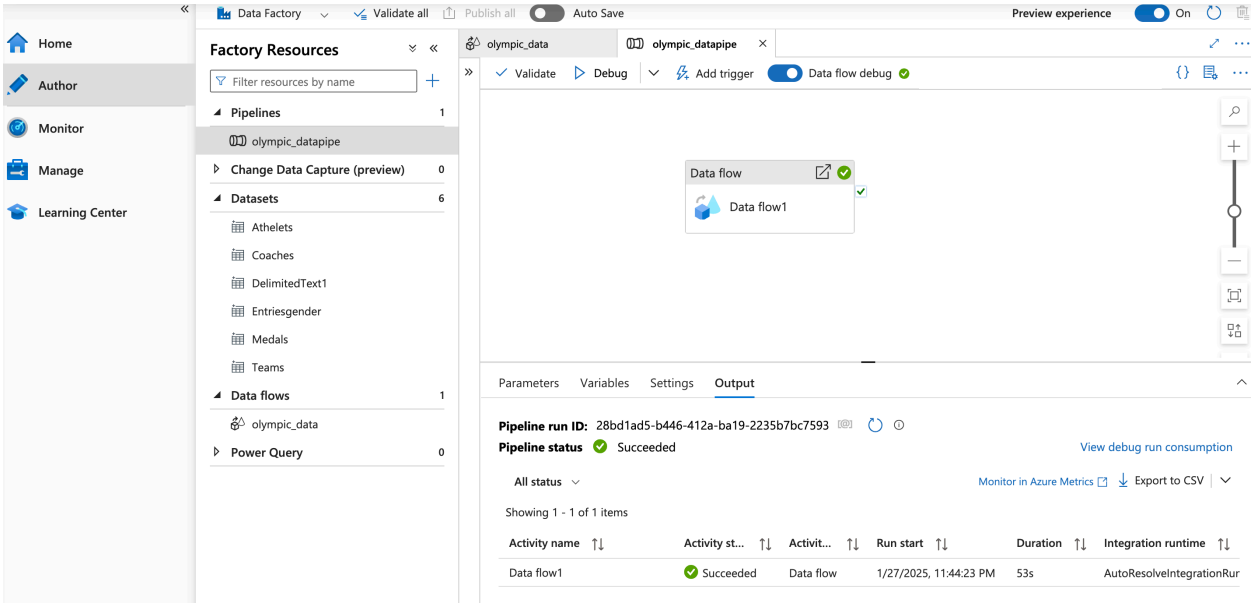


Fig:6 Data Flow

To validate the transformations, the final Data Flow debug sessions show that the pipeline was executed under an 8-core runtime environment and remained active during live testing. After

debugging and validation, the data was pushed into the "transform" zone, ready for analytical operations in Azure Synapse Analytics are shown in Fig: 6.

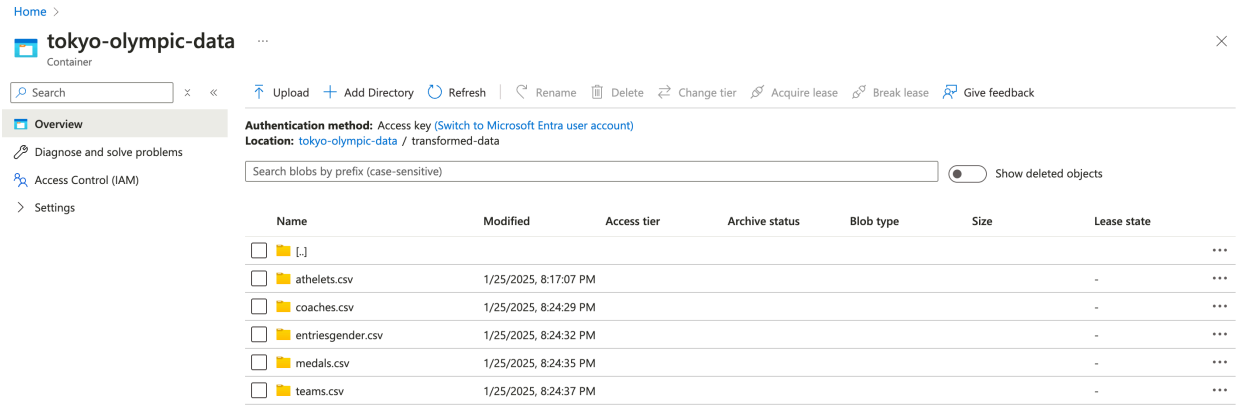


Fig: 7 Transform data in Azure Data Lake Gen2

The end-to-end pipeline was further encapsulated into a reusable pipeline named `olympic_datapipe`, which ran successfully as seen in the execution pane. This not only confirms operational success but also ensures that the pipeline can be triggered repeatedly for future data ingestion and transformation tasks.

Following ingestion, the data was cleaned and transformed using Azure Databricks**,** where PySpark code was executed on a Spark cluster. After successful data ingestion into Azure Data Lake Gen2, the next phase focused on transformation and analysis using Azure Databricks.

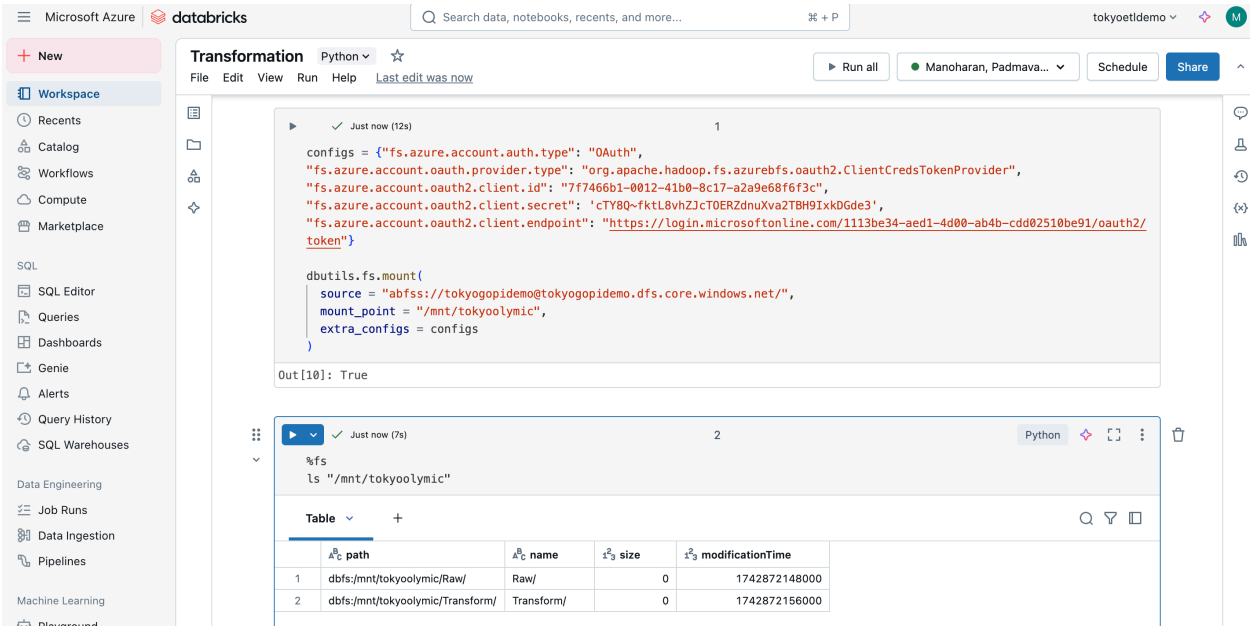


Fig: 8 Mounting Azure Data Lake in Azure Databricks

In Fig 8, The Databricks workspace was connected to the Data Lake using OAuth authentication, and both `Raw` and `Transform` folders were mounted to the notebook environment. This setup allowed seamless access to files stored in the lake.

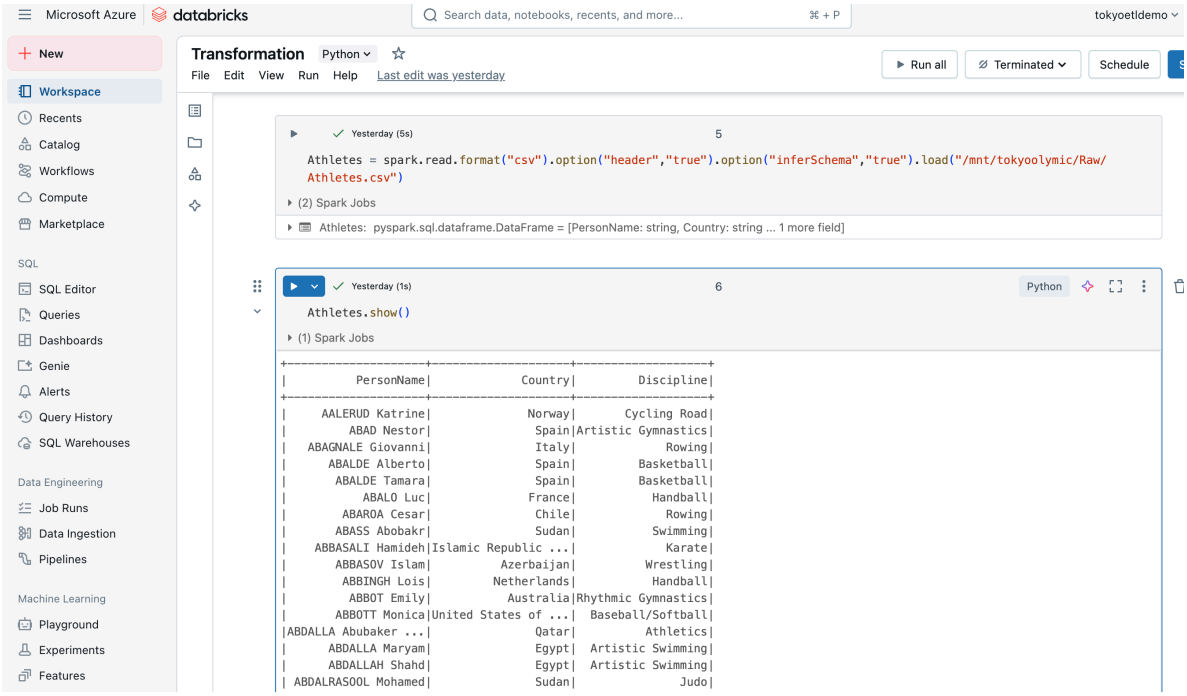


Fig: 9 Reading and Displaying Raw Athlete Data in Databricks

The `Athletes.csv` file was read using Spark DataFrame operations, with the header and schema inferred automatically. The dataset included athlete names, countries, and their respective disciplines. A `show()` function confirmed the data was correctly loaded and ready for transformation in Fig: 9.

In Fig: 10 Schema Inspection and Data Type Casting for Gender Entries, The `EntriesGender.csv` file was analyzed next. Using PySpark, a schema was printed to understand the structure, which included `Discipline`, `Female`, `Male`, and `Total` columns. These columns were then cast into appropriate data types (`IntegerType`). A new DataFrame was created to compute average male and female participation per discipline. The results showed interesting trends—for example, disciplines like Artistic Swimming had 100% female participation, while Boxing and Wrestling were more male-dominated. Balanced participation was observed in sports like Athletics and Swimming, reflecting gender equity.

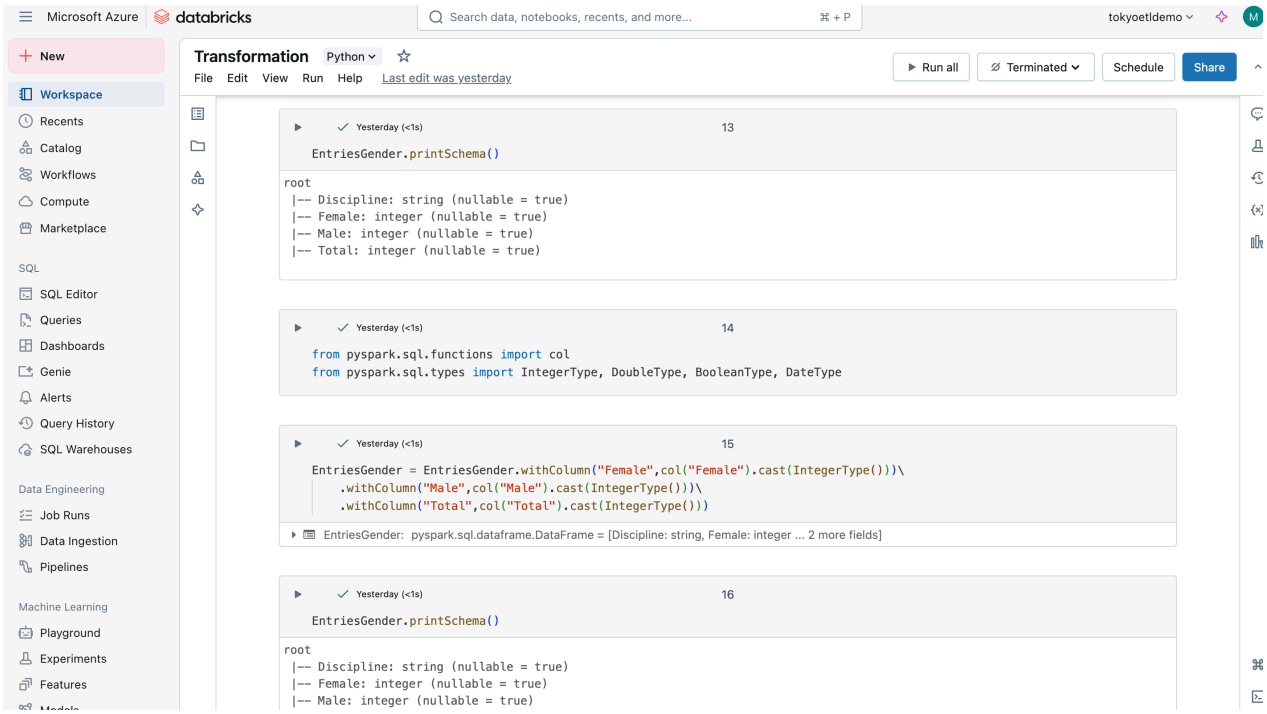


Fig: 10 Schema Inspection and Data Type Casting for Gender Entries

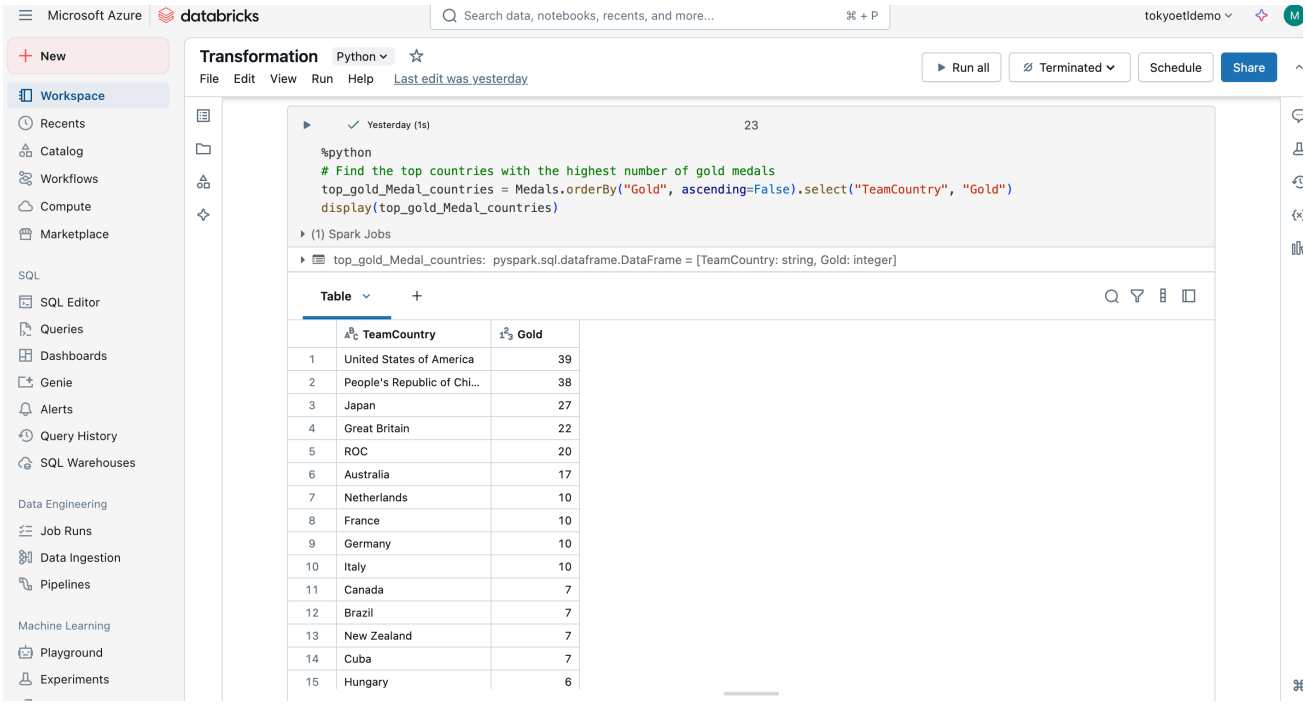


Fig: 11 Extracting Top Countries by Gold Medal Count

Additionally, a PySpark query was used to rank countries based on the number of Gold medals. The resulting table in Fig: 11 displays that the United States led with 39 gold medals, followed by China with 38, and Japan with 27.

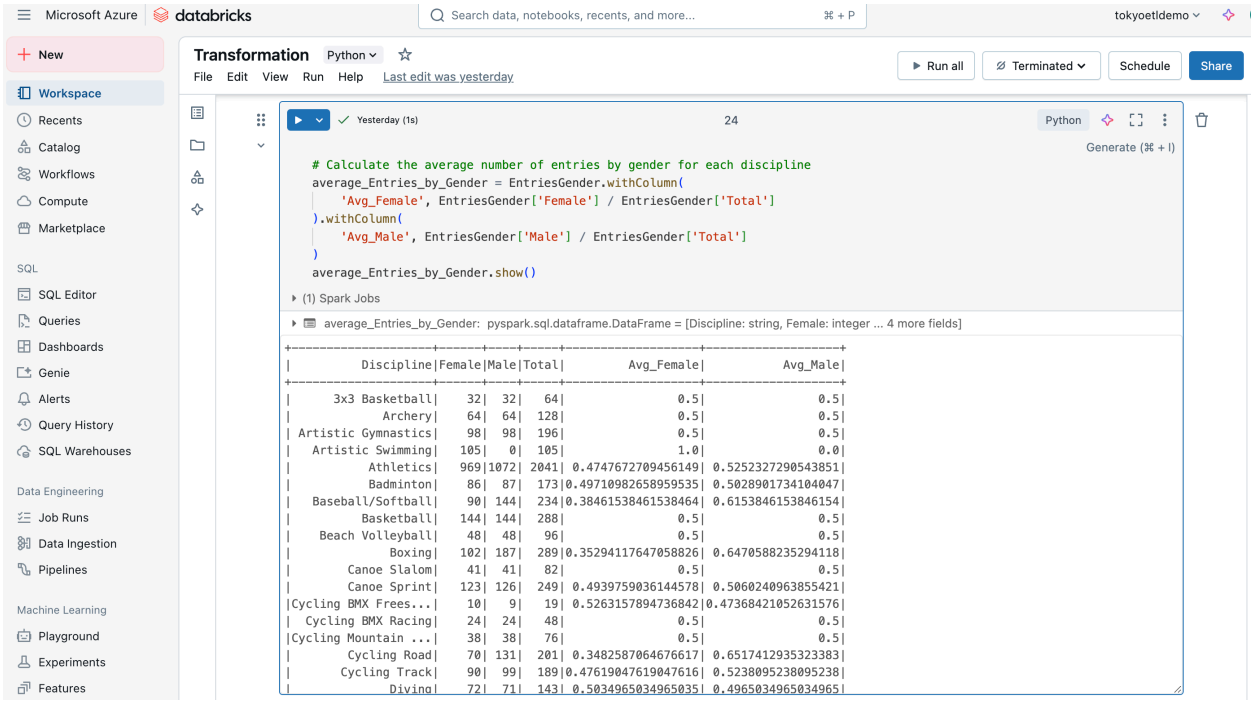


Fig:12 Gender-Based Participation Analysis Using PySpark

This Fig: 12 illustrates the use of PySpark in Azure Databricks to calculate the average male and female entries per discipline in the Tokyo 2021 Olympics dataset. By dividing the respective gender counts by the total entries, the code computes the average participation rate for each gender across various sports. The output helps highlight gender distribution patterns, showing balanced participation in many disciplines, while others exhibit a notable gender gap.

The transformed data was then analyzed using Azure Synapse Analytics, where SQL queries were executed on external tables created from the cleaned data.

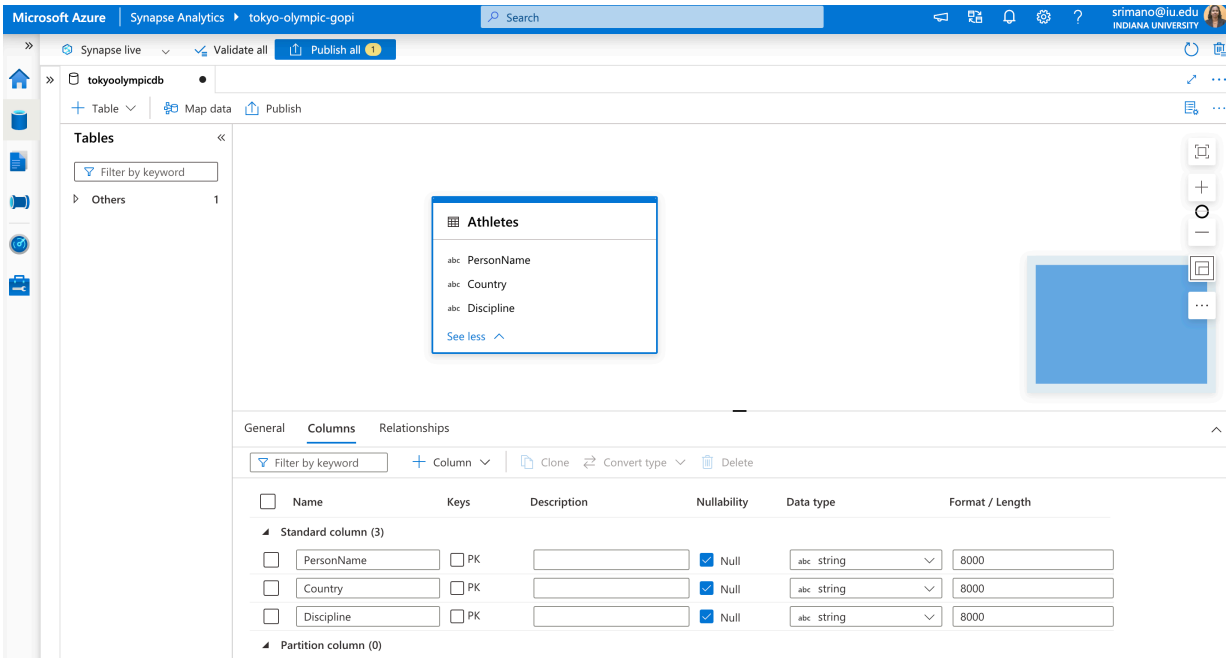


Fig:13 Athlete Table in Azure Synapse Analytics

This Fig: 13 presents the "Athletes" table created in Azure Synapse Analytics as part of the data analysis phase. The table includes structured columns such as `PersonName`, `Country`, and `Discipline`, all defined as string data types. This table enables efficient querying of athlete-related data for further SQL-based analytics, supporting insights into participation trends by country and discipline.

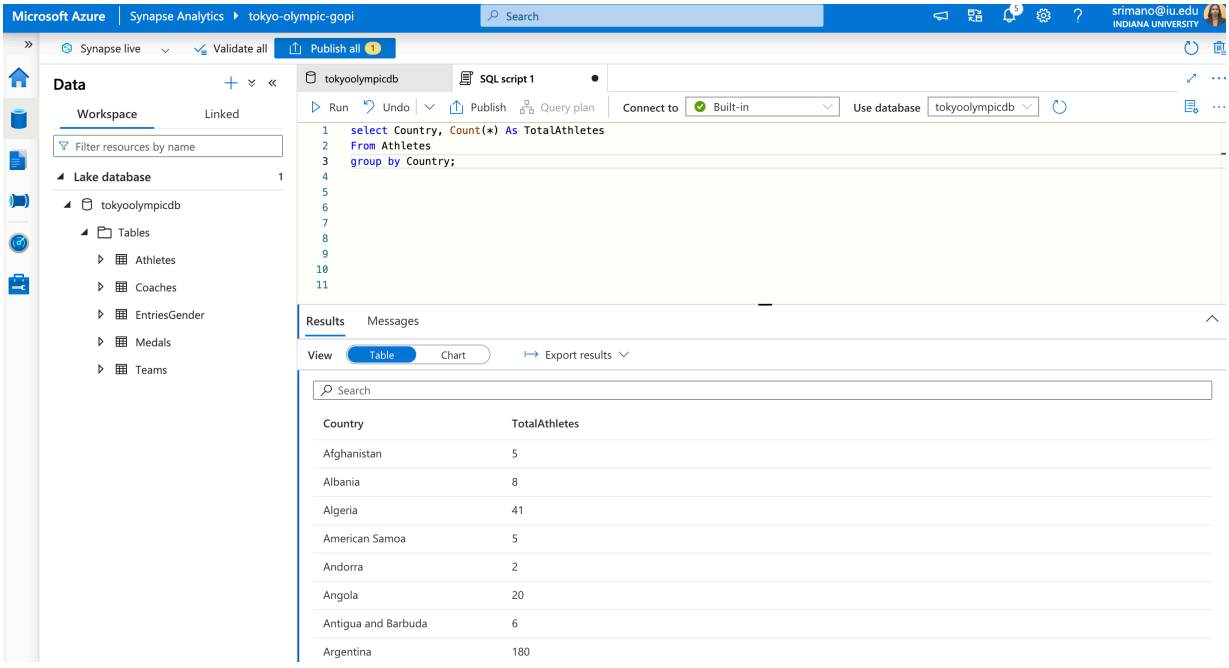


Fig:14 Athlete Count by Country in Synapse Analytics

This Fig: 14 displays the result of an SQL query executed in Azure Synapse Analytics that calculates the total number of athletes from each country. The query groups entries from the `Athletes` table by the `Country` column and counts the occurrences, providing a clear view of country-wise participation in the Tokyo Olympics dataset. This insight supports comparative analysis of athlete distribution across nations.

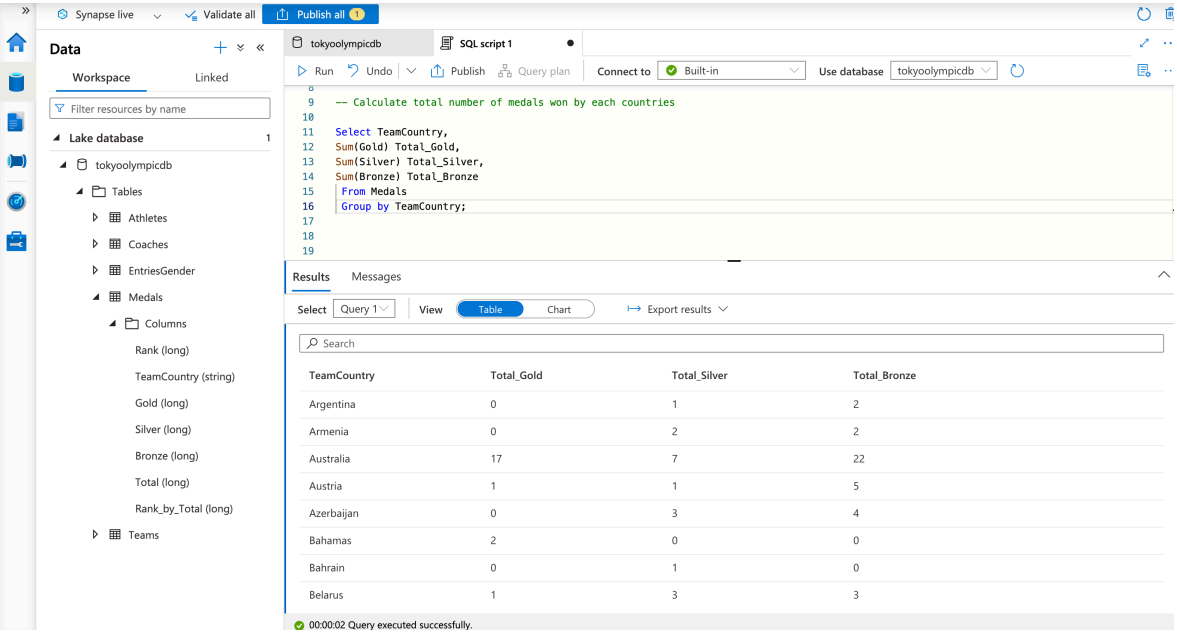


Fig:15 Total Medals Won by Country in Synapse Analytics

This Fig:15 presents the output of a SQL aggregation query in Azure Synapse Analytics, showing the total number of Gold, Silver, and Bronze medals won by each country. The data was grouped by the `TeamCountry` column in the `Medals` table and provides a detailed breakdown of medal counts per nation. This analysis supports ranking countries based on Olympic performance and reveals distribution trends among top-performing nations.

This chart in Fig 16 illustrates the average number of male and female athlete entries for each Olympic discipline, as calculated through SQL queries in Azure Synapse Analytics. The visual comparison highlights gender participation trends, showing variations in male and female representation across sports. This analysis supports evaluating diversity and gender balance in Olympic events.

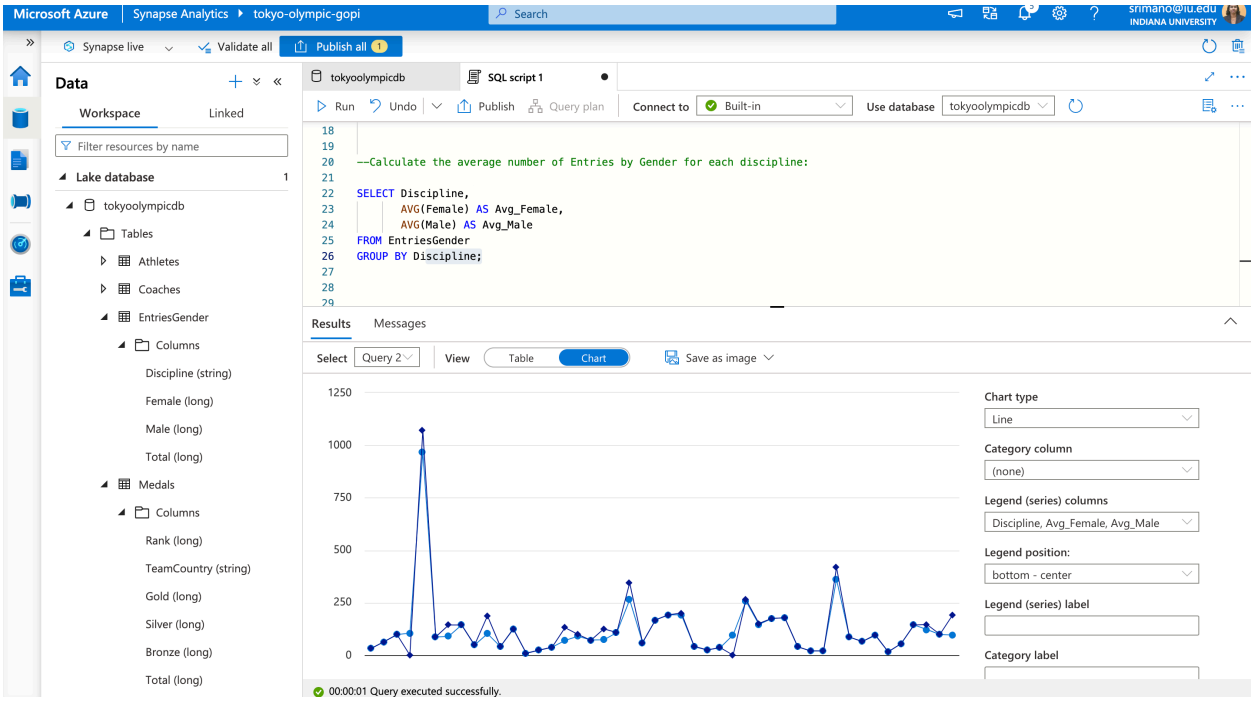


Fig: 16 Average Athlete Entries by Gender Across Disciplines in Synapse Analytics

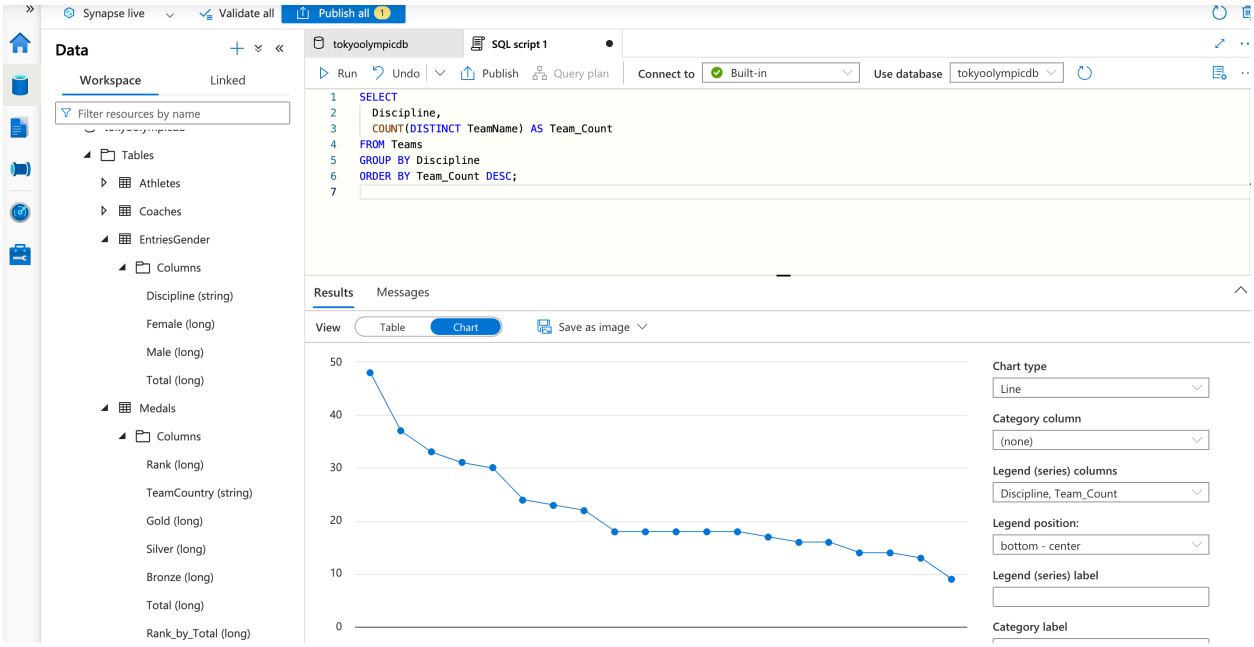


Fig: 17 Discipline-Wise Team Participation Count

This line chart in Fig: 17 displays the number of unique teams participating in each Olympic discipline, using data queried from the Teams table in Synapse Analytics. The disciplines are arranged in descending order based on team count, helping identify which sports had the highest team engagement during the Tokyo Olympics.

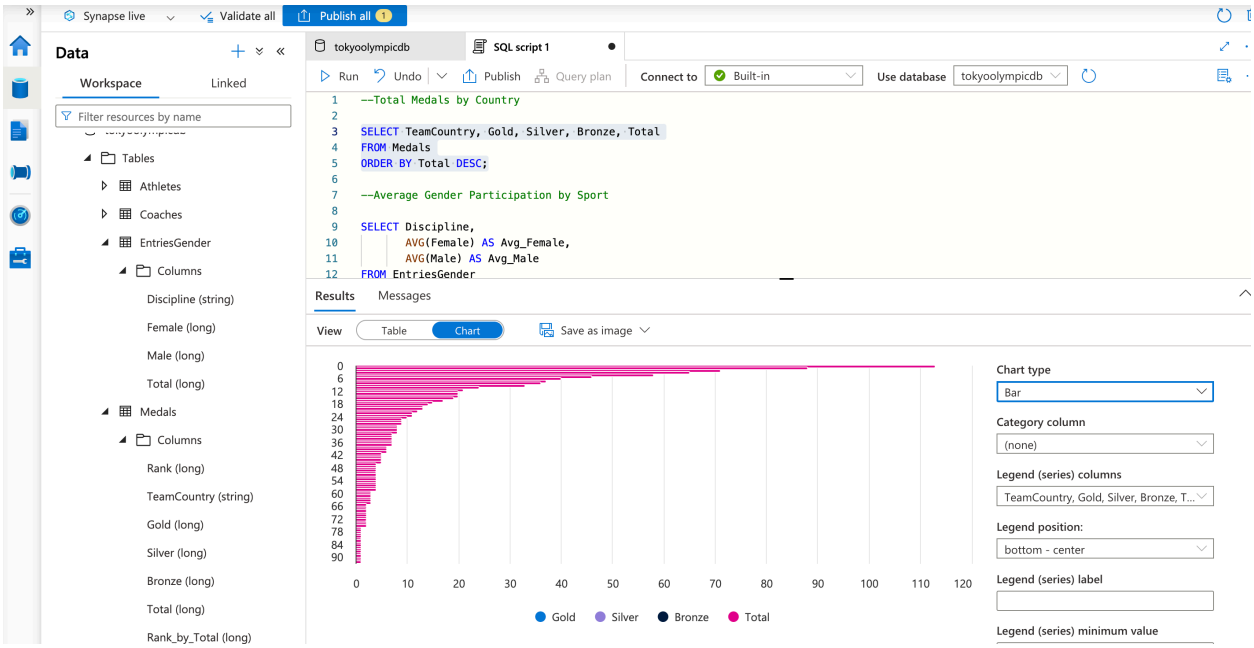


Fig: 18 Total Medals by Country

This bar chart displays in Fig:18 the distribution of gold, silver, and bronze medals along with the total medal count for each country, sorted in descending order. The data was extracted from the Medals table in Azure Synapse Analytics, helping identify top-performing nations in the Tokyo Olympics.

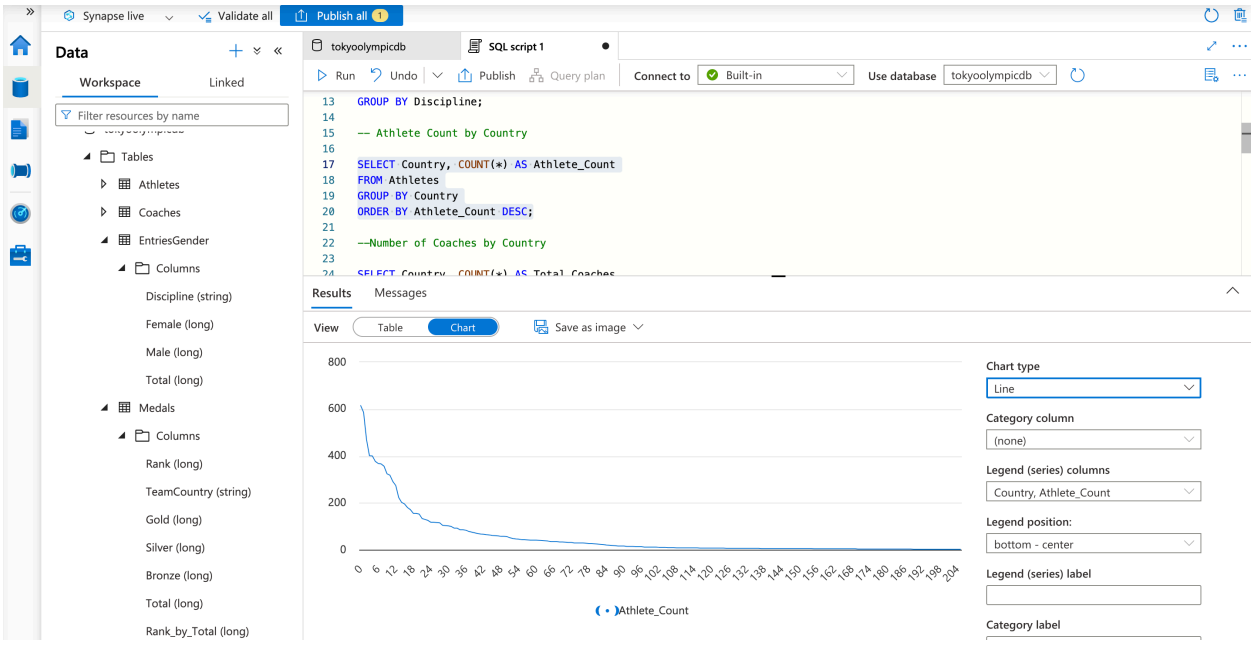


Fig: 19 Athlete Count by Country

This line chart visualizes in Fig:19 the number of athletes representing each country in the Tokyo Olympics. The data was aggregated from the Athletes table using Azure Synapse Analytics, showcasing the scale of participation and highlighting countries with the highest athlete representation.

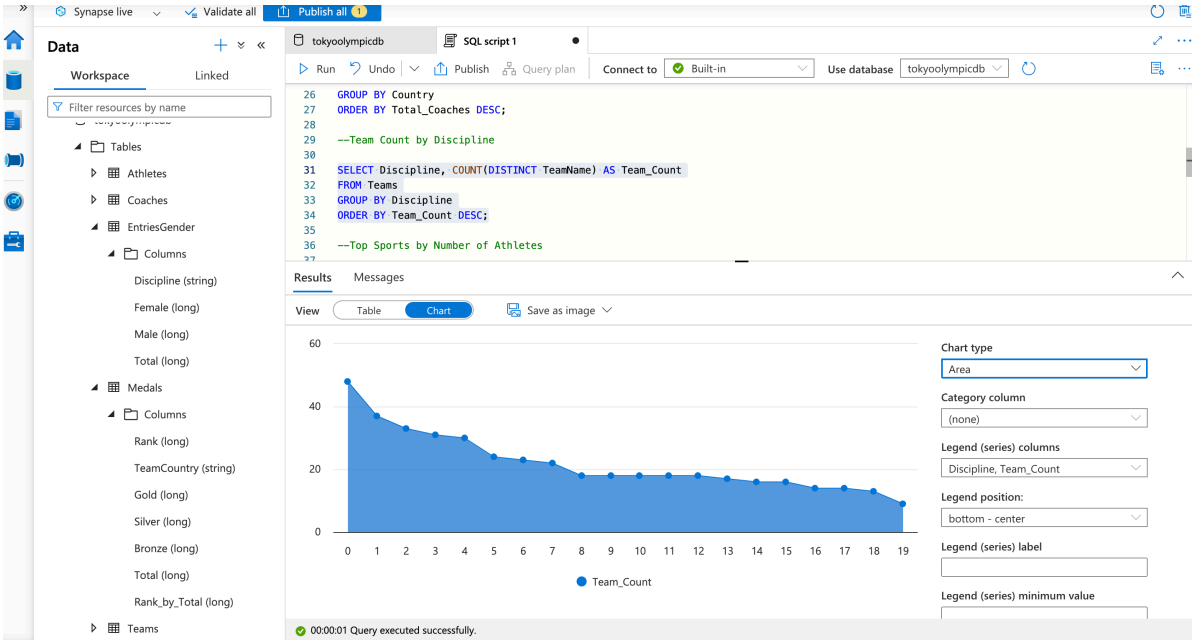


Fig: 20 Team Count by Discipline

This area chart in Fig: 20 illustrates the number of distinct national teams participating in each Olympic discipline. The data was derived using a COUNT(DISTINCT TeamName) query in Synapse Analytics, helping to highlight which sports saw the broadest international engagement.

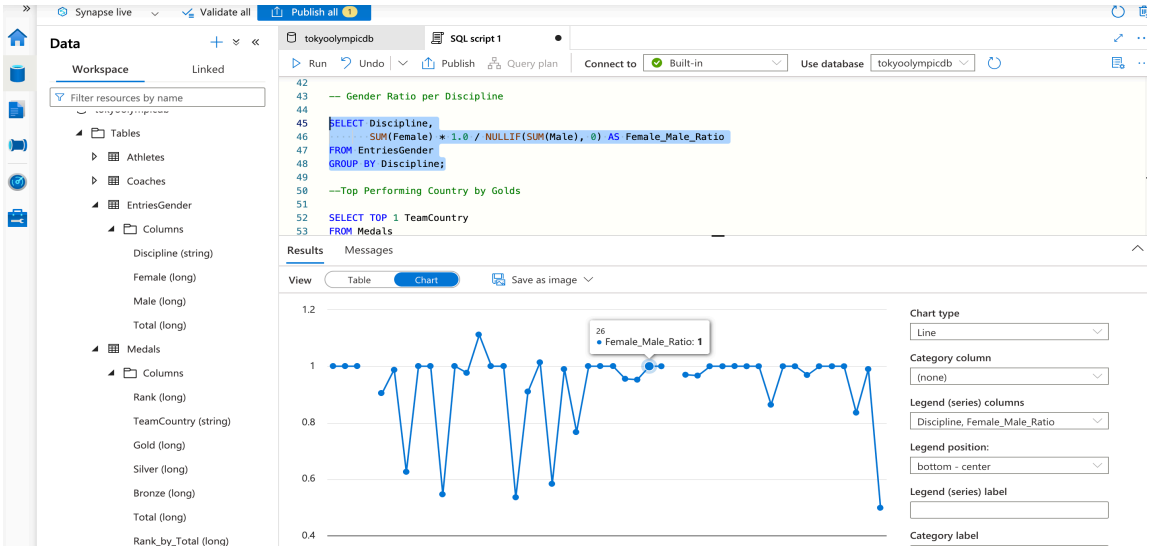


Fig: 21 Gender Ratio by Discipline

This line chart in Fig: 21 displays the Female-to-Male participation ratio for each Olympic discipline. Using aggregated data from the `EntriesGender` table in Synapse Analytics, the ratio was computed as `SUM(Female) / SUM(Male)`, highlighting how balanced or skewed participation was across different sports.

To present these insights effectively, the final output was visualized using Tableau Public. Three interactive dashboards were created, each focused on a different aspect of the Olympics data:

- **Athlete count by Country**: Displayed country-wise athlete distribution, showing the United States, Japan, and Australia as top contributors.
- **Most Popular Sports (Tree Map):** Visualized which sports had the highest representation, with Athletics, Swimming, and Judo being most prominent.
- **Discipline-wise Athlete Analysis**: Visualized athlete counts across disciplines and countries.

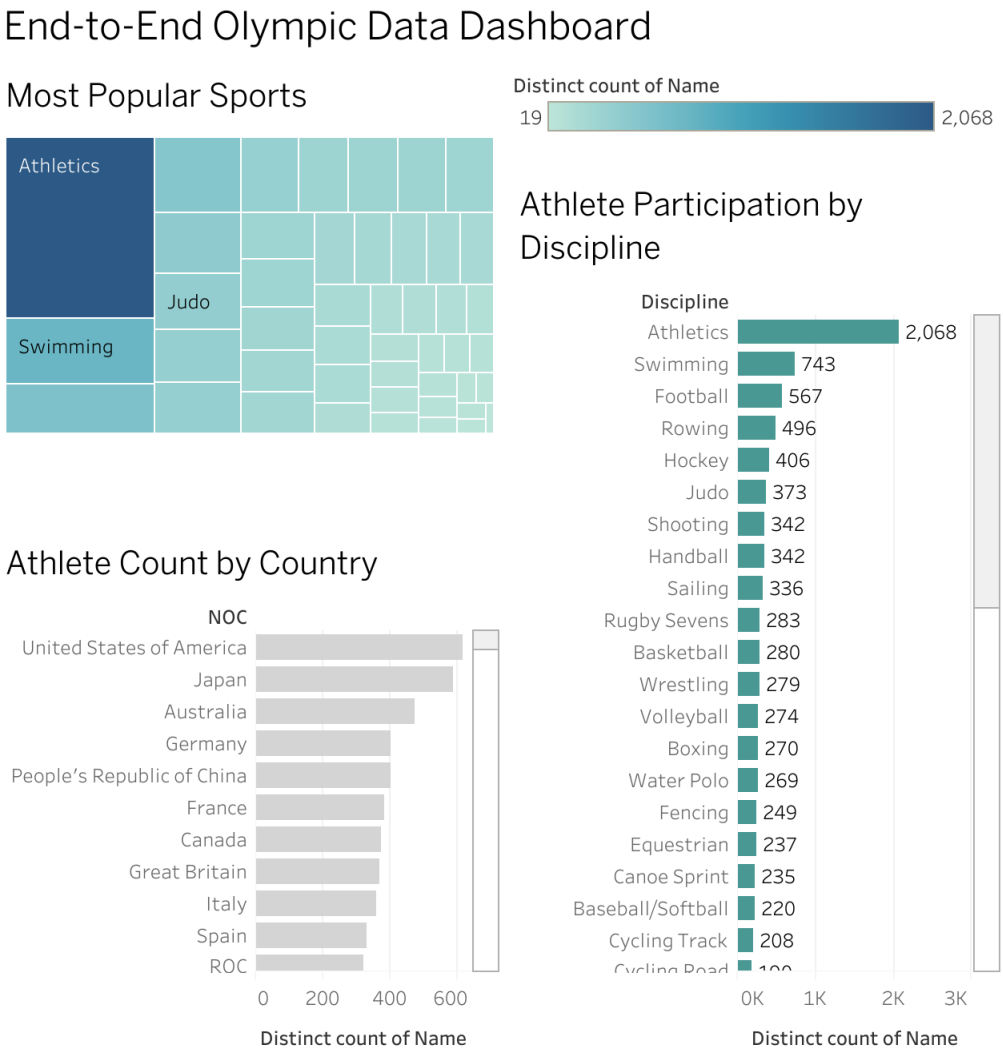


Fig:22 End-to-End Olympic Data Dashboard (Tableau Public Output)

This dashboard in Fig: 22, visually summarizes key insights from the Olympic dataset. The treemap shows the most popular sports by athlete count, with Athletics, Swimming, and Judo leading in participation. The bar chart titled "Athlete Participation by Discipline" highlights how different sports vary in athlete representation, with Athletics reaching over 2,000 participants. Another bar chart breaks down "Athlete Count by Country", revealing that the United States had the highest number of athletes, followed closely by Japan and Australia. These visualizations together help analyze patterns in global participation and sport popularity during the Tokyo 2021 Olympics.

10. Discussion

This project showcases the effectiveness of integrating cloud-based data engineering tools with modern business intelligence (BI) platforms to analyze real-world datasets. The seamless workflow from data ingestion to visualization reflects how Microsoft Azure and Tableau can be used together to build robust, end-to-end analytical solutions. The Tokyo 2021 Olympics dataset served as an excellent case study for demonstrating this potential.

One of the key advantages observed was the modularity and scalability of Azure services. Azure Data Factory enabled scheduled and automated data ingestion directly from GitHub, eliminating manual effort and reducing the risk of human error. Azure Data Lake Gen2 provided a scalable storage solution with logical separation between raw and transformed data. This allowed for organized data flow and easy retrieval at each stage of the pipeline [7].

The transformation of data using Azure Databricks highlighted the power of distributed computing with Apache Spark. Even though the dataset was semi-structured and included inconsistencies, the use of PySpark ensured that data could be cleaned, standardized, and made analysis-ready with minimal execution time. The reusability of PySpark notebooks also promotes maintainability and supports future enhancements [4].

Azure Synapse Analytics provided a SQL-based environment to perform exploratory data analysis and aggregation without having to duplicate data into a traditional database. By using external tables connected to files in Data Lake, the solution remained lightweight and efficient. This also demonstrated Synapse's ability to handle large files and provide rapid query execution, which is essential in time-sensitive analytics like sports reporting or live performance dashboards.

Finally, the use of Tableau Public for visualization ensured that complex data outputs could be translated into interactive dashboards that are accessible to both technical and non-technical stakeholders. The dashboards offered intuitive filters, visual summaries, and comparative insights that would be difficult to interpret using raw tables alone. This makes the solution ideal for decision-makers, coaches, or journalists looking to explore Olympic data without writing code.

Overall, the project demonstrates how cloud-based, scalable, and modular tools can be orchestrated to deliver fast, reliable, and visually appealing analytics. It validates the relevance of cloud data engineering in modern analytics and sets a foundation for applying similar techniques to other real-time or large-scale datasets beyond sports.

11. Conclusion and Future Scope

This project successfully demonstrated the creation and execution of a complete cloud-based ETL pipeline using Microsoft Azure services. Each stage—from data ingestion with Azure Data Factory to transformation in Azure Databricks and querying in Azure Synapse Analytics—was implemented with a focus on automation, scalability, and efficiency. The raw Tokyo Olympics 2021 dataset, which initially existed in a semi-structured format, was processed and structured in a way that enabled meaningful insights. Tableau Public was used effectively to visualize the outcomes, allowing end-users to interact with dashboards that provided clarity on key metrics such as medal distribution, gender participation, and discipline-level trends.

Looking ahead, there are several areas for future enhancement. The dashboards could be extended to Power BI for more advanced enterprise-level reporting and tighter integration with other Microsoft services. Additionally, the Azure Data Factory pipelines could incorporate scheduling, monitoring, and alerting features to support continuous data ingestion from dynamic sources. Another potential upgrade would be the use of machine learning models within Azure Databricks to predict medal outcomes or athlete performance based on historical trends. Lastly, implementing CI/CD pipelines for automated deployment and version control would align the solution with modern DevOps practices**,** enhancing reproducibility and collaboration.

Overall, this project not only achieved its initial objectives but also lays the groundwork for more complex, real-time analytics solutions that integrate data engineering, data science, and business intelligence in the cloud.

12. References

[1] O. Oladimeji, "Enhancing Data Pipeline Efficiency Using Cloud-Based Big Data Technologies: A Comparative Analysis of AWS and Microsoft Azure," *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/384958218_Enhancing_Data_Pipeline_Efficiency_Using_Cloud-Based_Big_Data_Technologies_A_Comparative_Analysis_of_AWS_and_Microsoft_Azure

[2] R. M. Parthe, "Comparative Analysis of Data Visualization Tools: Power BI and Tableau," *ResearchGate*, 2023. [Online]. Available: https://www.researchgate.net/publication/374957892_Comparative_Analysis_of_Data_Visualization_Tools_Power_BI_and_Tableau

[3] N. Martins, "Design Principles in the Development of Dashboards for Business Management," *ResearchGate*, 2022. [Online]. Available: https://www.researchgate.net/publication/355031055_Design_Principles_in_the_Development_of_Dashboards_for_Business_Management

[4] M. Thoutam, "Cloud-Native ETL: Integrating Databricks and Azure Data Factory for Scalable Data Processing in Enterprise Environments," *International Journal For Multidisciplinary Research (IJFMR)*, 2024. [Online]. Available: https://www.ijfmr.com/papers/2024/6/29886.pdf

[5] J. Purich, "Toward a Scalable Census of Dashboard Designs in the Wild: A Case Study with Tableau Public," *arXiv preprint arXiv:2306.16513*, 2023. [Online]. Available: https://arxiv.org/abs/2306.16513

[6] M. Kim, "Using Tableau and Google Map API for Understanding the Impact of Walkability on Dublin City," *arXiv preprint arXiv:2310.07563*, 2023. [Online]. Available: https://arxiv.org/abs/2310.07563

[7] Microsoft, "Azure Data Factory Documentation," *Microsoft Docs*, 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/data-factory/

[8] Microsoft, "Build ETL Pipelines with Azure Databricks and Delta Lake," *Microsoft Docs*, 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/architecture/solution-ideas/articles/ingest-etl-stream-with-adb

[9] Microsoft, "Introduction to Azure Data Factory," *Microsoft Docs*, 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/data-factory/introduction

[10] Microsoft, "Analytics End-to-End with Azure Synapse," *Microsoft Docs*, 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/architecture/example-scenario/dataplate2e/data-platform-end-to-end

13. Acronyms

ETL	Extract, Transform, Load
ADF	Azure Data Factory
ADLS	Azure Data Lake Storage
SQL	Structured Query Language
BI	Business Intelligence
ML	Machine Learning
CI/CD	Continuous Integration/Delivery