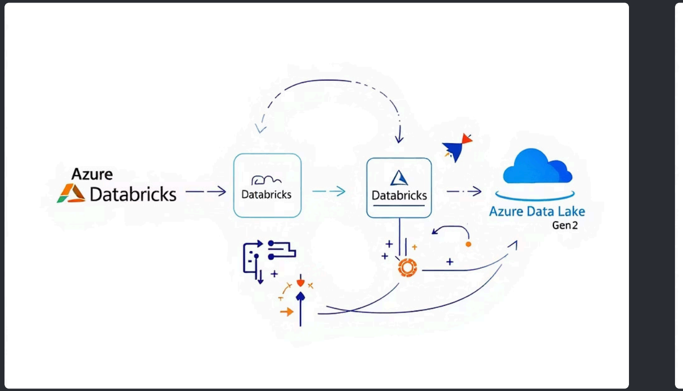
Olympics Data Analytics Pipeline using Azure

This project demonstrates a complete end-to-end **modern data pipeline** using Microsoft Azure.

This comprehensive solution showcases the power of Azure's cloud ecosystem for handling large-scale data analytics, from ingestion through transformation to visualization, creating a robust foundation for Olympic historical data analysis.

Azure Cloud Foundation

Data acquisition from GitHub repositories and initial staging in Azure Storage Accounts to establish a robust and scalable data lake foundation.



Streamlined Data Flow

Efficient data transformation using Azure Databricks for processing and refining raw data, ensuring its readiness for analytical workloads, and storing in Azure Data Lake Gen2.

Insightful Analytics & Visualization

Advanced querying capabilities with Azure Synapse Analytics for complex data analysis, and interactive visualization through Power BI dashboards to extract actionable insights.

Project Overview

Data Source

Olympics dataset from Kaggle, featuring 120 years of Olympic history with athlete and results data

Storage Layer

Azure Blob Storage and Azure Data Lake Storage Gen2 for raw and processed data

Processing Engine

Azure Databricks for data transformation, cleaning, and advanced analytics

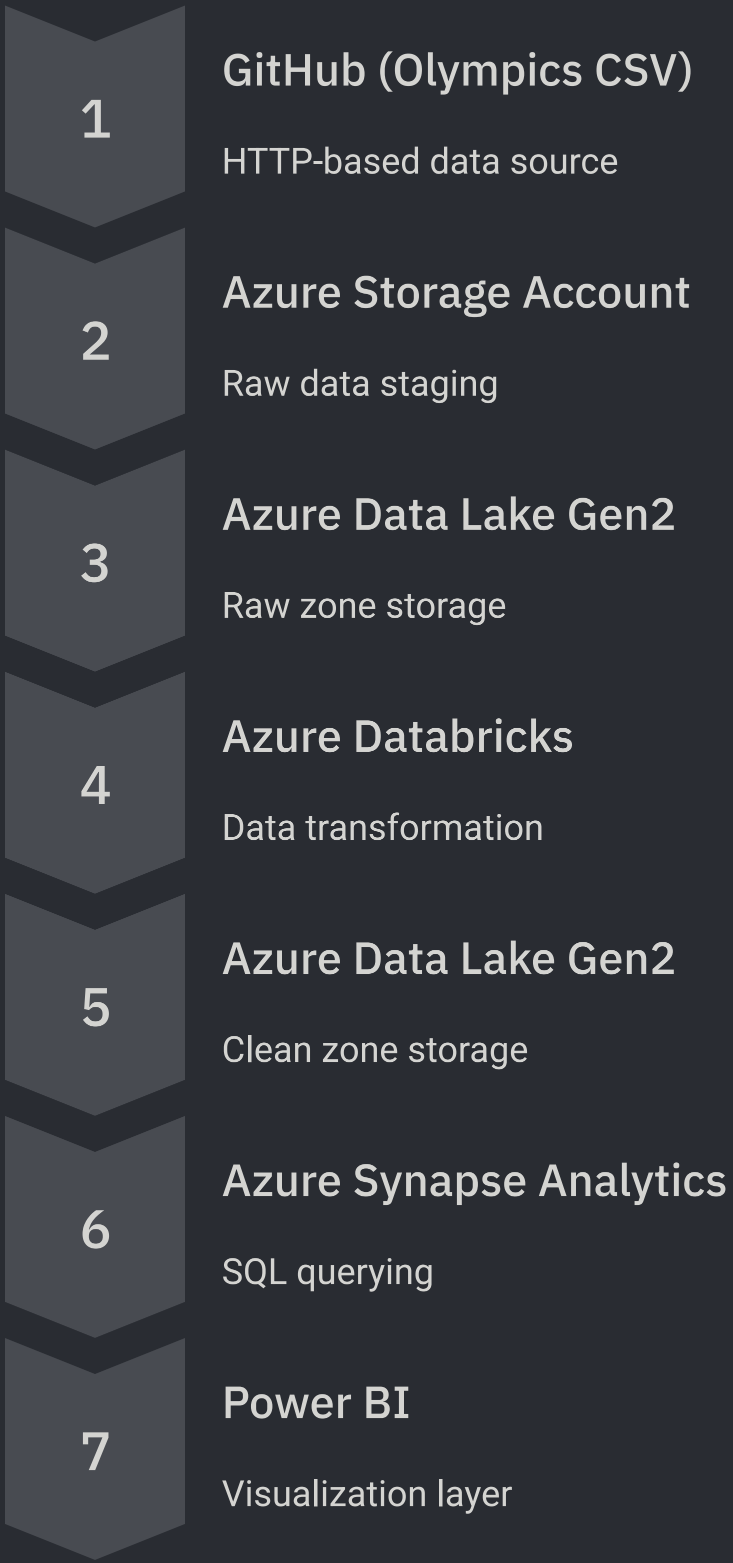
Analytics & Visualization

Azure Synapse Analytics for querying with Power BI dashboards for insights

Data Hosting: Dataset uploaded to GitHub [AkshathD2298] for HTTP-based access, enabling seamless integration with Azure services and automated data pipeline workflows.

Architecture & Tech Stack

📁 Architecture Flow



1 GitHub (Olympics CSV)

HTTP-based data source

2 Azure Storage Account

Raw data staging

3 Azure Data Lake Gen2

Raw zone storage

4 Azure Databricks

Data transformation

5 Azure Data Lake Gen2

Clean zone storage

6 Azure Synapse Analytics

SQL querying

7 Power BI

Visualization layer

Technology Components

GitHub (HTTP URL): Hosting Olympics CSV file for HTTP-based ingestion

Azure Blob Storage: Initial staging of raw data

Azure Data Lake Gen2: Raw + Processed data storage

Azure Databricks: Data transformation and cleansing

Azure Synapse Analytics: Analytics SQL querying on clean data

Azure Data Factory: Orchestrating data movement and pipeline automation

Power BI: Dashboard and data visualization

Python / PySpark: Data wrangling and transformation within Databricks

Data Pipeline Flow

01

Data Upload

The original Olympics CSV file is hosted on GitHub for HTTP access

02

Ingestion

Azure Data Factory fetches the CSV and loads it into an Azure Storage account

03

Raw Zone

Data is moved into the raw zone of Azure Data Lake Gen2

04

Transformation

Azure Databricks reads the raw data, performs cleaning (nulls, duplicates, column formatting), and writes to the clean zone

05

Querying

Azure Synapse Analytics connects to the clean data zone for querying

06

Visualization

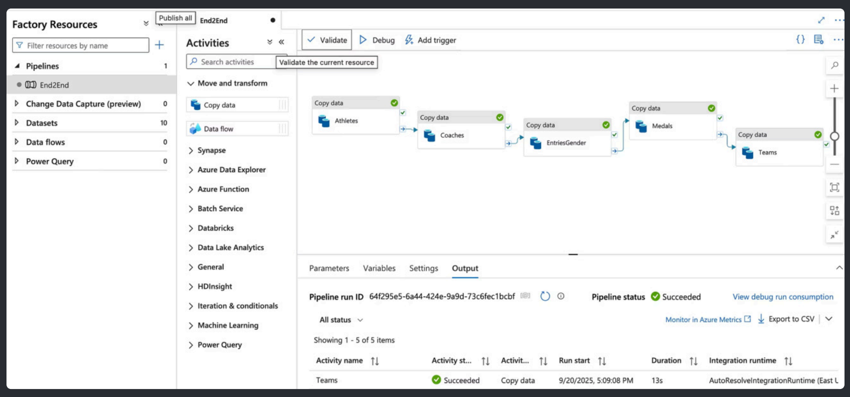
Power BI imports data from Synapse to create visual analytics and dashboards

Project Implementation: Visual Walkthrough

This section provides a visual walkthrough of the key components and configurations within the Azure portal, Databricks notebooks, Synapse Analytics, and Power BI dashboards, demonstrating the end-to-end data pipeline in action.

Azure Data Factory: Ingestion Pipeline

Observe the Azure Data Factory pipeline, meticulously configured to ingest the Olympics CSV data directly from GitHub. This pipeline skillfully orchestrates data movement, ensuring efficient loading of raw data into Azure Storage for subsequent processing.



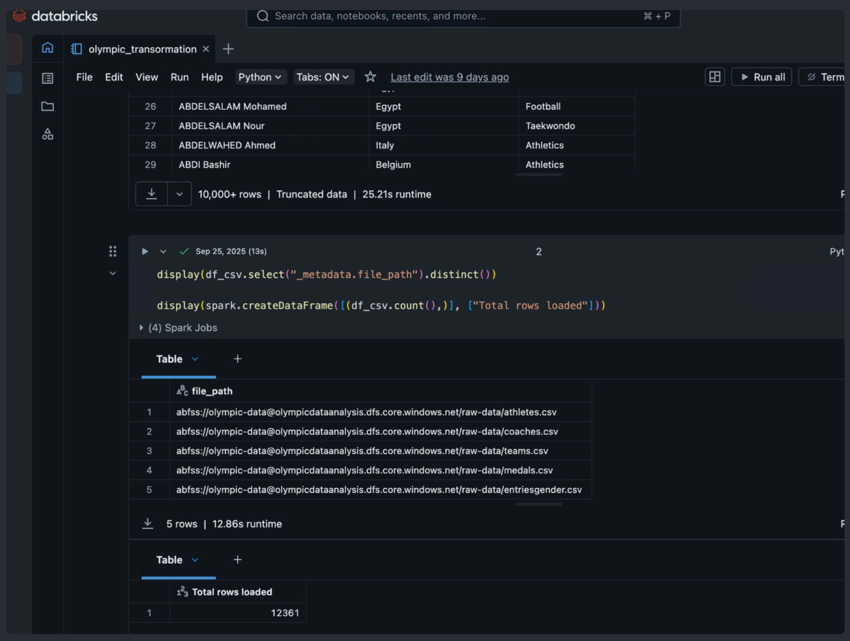
Pipeline run ID 64f295e5-6a44-424e-9a9d-73c6fec1bcbf   Pipeline status ✓ Succeeded

Showing 1 - 5 of 5 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime
Teams	✓ Succeeded	Copy data	9/20/2025, 5:09:08 PM	13s	AutoResolveIntegrationRuntime (East L

Azure Databricks: Data Transformation Notebook

Explore this Azure Databricks notebook, where PySpark code is applied for robust data cleaning, transformation, and enrichment. Steps encompass handling null values, removing duplicates, and standardizing column formats, all prior to writing the refined data to the clean zone in Azure Data Lake Gen2.



26	ABDELSALAM Mohamed	Egypt	Football
27	ABDELSALAM Nour	Egypt	Taekwondo
28	ABDELWAHED Ahmed	Italy	Athletics
29	ABDI Bashir	Belgium	Athletics

10,000+ rows | Truncated data | 25.21s runtime

✓ Sep 25, 2025 (13s)   2

```python
display(df_csv.select("_metadata.file_path").distinct())

display(spark.createDataFrame([(df_csv.count(),)], ["Total rows loaded"]))
```

(4) Spark Jobs

Table

	file_path
1	abfss://olympic-data@olympicdataanalysis.dfs.core.windows.net/raw-data/athletes.csv
2	abfss://olympic-data@olympicdataanalysis.dfs.core.windows.net/raw-data/coaches.csv
3	abfss://olympic-data@olympicdataanalysis.dfs.core.windows.net/raw-data/teams.csv
4	abfss://olympic-data@olympicdataanalysis.dfs.core.windows.net/raw-data/medals.csv
5	abfss://olympic-data@olympicdataanalysis.dfs.core.windows.net/raw-data/entriesgender.csv

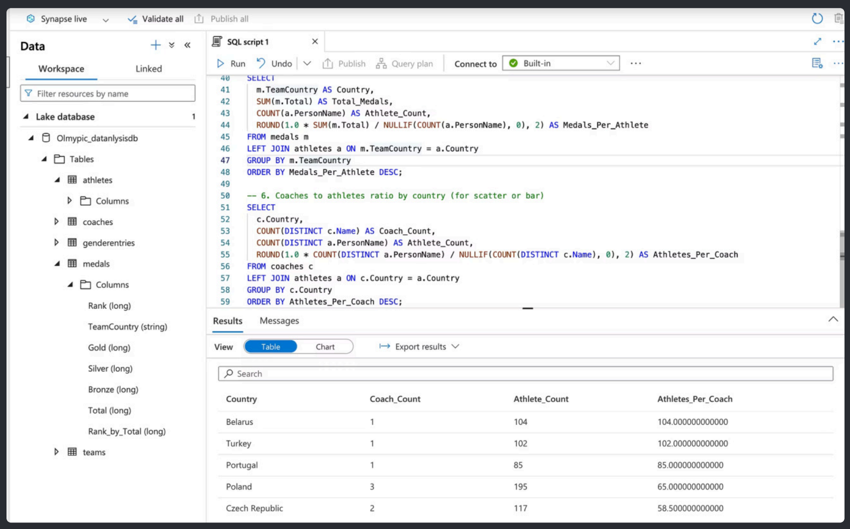
5 rows | 12.86s runtime

Table

	Total rows loaded
1	12361

Azure Synapse Analytics: SQL Querying

Witness Azure Synapse Analytics in action, querying the meticulously cleaned data. This environment facilitates interactive analysis of the processed Olympics data using SQL, providing a powerful interface for data exploration and preparing it for dynamic visualizations.



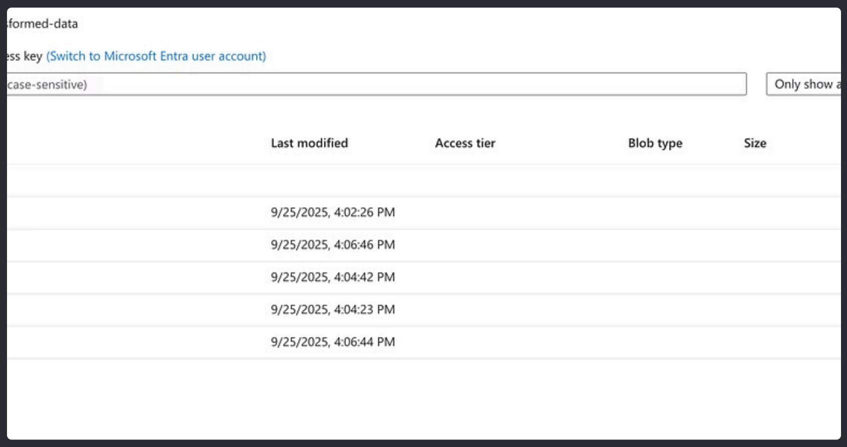
```sql
SELECT
  m.TeamCountry AS Country,
  SUM(m.Total) AS Total_Medals,
  COUNT(a.PersonName) AS Athlete_Count,
  ROUND(1.0 * SUM(m.Total) / NULLIF(COUNT(a.PersonName), 0), 2) AS Medals_Per_Athlete
FROM medals m
LEFT JOIN athletes a ON m.TeamCountry = a.Country
GROUP BY m.TeamCountry
ORDER BY Medals_Per_Athlete DESC;

-- 6. Coaches to athletes ratio by country (for scatter or bar)
SELECT
  c.Country,
  COUNT(DISTINCT c.Name) AS Coach_Count,
  COUNT(DISTINCT a.PersonName) AS Athlete_Count,
  ROUND(1.0 * COUNT(DISTINCT a.PersonName) / NULLIF(COUNT(DISTINCT c.Name), 0), 2) AS Athletes_Per_Coach
FROM coaches c
LEFT JOIN athletes a ON c.Country = a.Country
GROUP BY c.Country
ORDER BY Athletes_Per_Coach DESC;
```

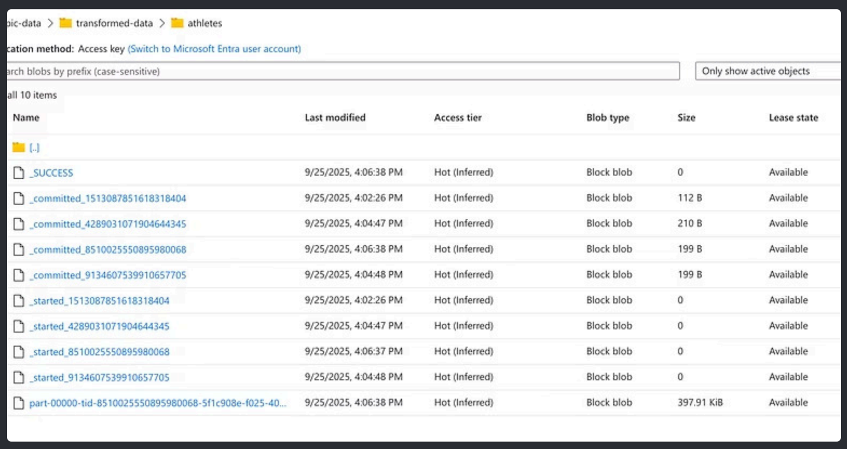
Country	Coach_Count	Athlete_Count	Athletes_Per_Coach
Belarus	1	104	104.000000000000
Turkey	1	102	102.000000000000
Portugal	1	85	85.000000000000
Poland	3	195	65.000000000000
Czech Republic	2	117	58.500000000000

Raw vs. Transformed Data

These visuals illustrate the critical data transformation process, showcasing the initial raw Olympics dataset alongside the cleaned and transformed results. Observe the improvements in data quality, consistency, and structure after Databricks processing, ready for advanced analysis and reporting.



Last modified	Access tier	Blob type	Size
9/25/2025, 4:02:26 PM			
9/25/2025, 4:06:46 PM			
9/25/2025, 4:04:42 PM			
9/25/2025, 4:04:23 PM			
9/25/2025, 4:06:44 PM			


ic-data > transformed-data > athletes
cation method: Access key (Switch to Microsoft Entra user account)
arch blobs by prefix (case-sensitive)
Only show active objects
all 10 items

Name	Last modified	Access tier	Blob type	Size	Lease state
[..]					
_SUCCESS	9/25/2025, 4:06:38 PM	Hot (Inferred)	Block blob	0	Available
_committed_1513087851618318404	9/25/2025, 4:02:26 PM	Hot (Inferred)	Block blob	112 B	Available
_committed_4289031071904644345	9/25/2025, 4:04:47 PM	Hot (Inferred)	Block blob	210 B	Available
_committed_8510025550895980068	9/25/2025, 4:06:38 PM	Hot (Inferred)	Block blob	199 B	Available
_committed_9134607539910657705	9/25/2025, 4:04:48 PM	Hot (Inferred)	Block blob	199 B	Available
_started_1513087851618318404	9/25/2025, 4:02:26 PM	Hot (Inferred)	Block blob	0	Available
_started_4289031071904644345	9/25/2025, 4:04:47 PM	Hot (Inferred)	Block blob	0	Available
_started_8510025550895980068	9/25/2025, 4:06:37 PM	Hot (Inferred)	Block blob	0	Available
_started_9134607539910657705	9/25/2025, 4:04:48 PM	Hot (Inferred)	Block blob	0	Available
part-00000-tid-8510025550895980068-5f1c908e-f025-40...	9/25/2025, 4:06:38 PM	Hot (Inferred)	Block blob	397.91 KiB	Available