A

SEMINAR REPORT ON

BIG DATA ANALYSIS: END TO END WORKFLOW

SUBMITTED TO SAVITRIBAI PHULE PUNE UNIVERSITY
FOR PARTIAL FULFILLMENT
OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

HONOURS
In
DATA SCIENCE

By

ANANT SURAJ NIKAM B400050014

GUIDE
Mr. R.G. YELALWAR

DEPARTMENT OF
ELECTRONICS AND TELECOMMUNICATION ENGINEERING
PUNE INSTITUTE OF COMPUTER TECHNOLOGY
PUNE – 43

A.Y. 2024-25

Department of Electronics and Telecommunication Engineering
Pune Institute of Computer Technology, Pune – 43

CERTIFICATE

This is to certify that the seminar Report entitled

Big Data Analysis: End to End Workflow

has been successfully completed by

Anant Suraj Nikam B400050014

Is a bona fide work carried out by them under the guidance of Prof. R.G. Yelalwar and it is approved for the partial fulfillment of the requirement of the Savitribai Phule Pune University, Pune for the award of the degree of the Honors in Data Science (Electronics and Telecommunication Engineering). This project work has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Mr. R. G. Yelalwar
Guide

Dr. M.V. Munot
HOD, E&TC Dept

Place: Pune
Date: 21/04/2025

ACKNOWLEDGEMENTS

I wish to express my heartfelt gratitude to my guide, Mr. R.G. Yelalwar for his invaluable guidance, constant encouragement, and thoughtful mentorship throughout the course of this project. His deep knowledge, patient support, and insightful feedback have greatly enriched my learning experience. Working under his supervision has not only enhanced my academic perspective but has also been a personally inspiring journey. I remain sincerely thankful for the time and attention he dedicated to this work. We extend our sincere thanks to our Head of Department Prof. Mousumi V. Munot ma'am for providing all kinds of cooperation during the course.

Anant Suraj Nikam

CONTENTS

ABSTRACT

In today's data-driven world, organizations are dealing with an overwhelming amount of information generated every second—from customer interactions to real-time sensor readings. Making sense of this massive and often unstructured data requires a well-planned, step-by-step approach. This project explores the full journey of big data analysis, beginning from the moment raw data is collected, all the way to the point where meaningful conclusions is drawn and decisions are made. The journey begins by pulling data from a data source. This data is first brought into Azure Data Factory, which manages and automates the movement of data into a secure raw storage layer using Data Lake Gen 2. At this stage, the information is unorganized and not yet ready for use. Using Azure Databricks, the data is cleaned, reshaped, and transformed into a structured format. This transformed version is again stored in Data Lake Gen 2, now ready for deeper analysis. The refined data is then passed to Azure Synapse Analytics, which allows for large-scale queries, reporting, and advanced insights. Finally, tools like Power BI, Looker Studio, and Tableau are used to build user-friendly dashboards that communicate key trends—such as medal performance, athlete statistics, or country-wise comparisons.

This end-to-end setup not only showcases the power of cloud-based data engineering but also demonstrates how raw, unstructured information can be turned into real-time, actionable insights for global events like the Olympics.

Abbreviations and Acronyms

OLAP	Online Analytical Processing
SQL	Structured Query Language
BI	Business Intelligence
ADF	Azure Data Factory
API	Application Programming Interface
CSV	Comma Separated Value
DAG	Directed Acyclic Graph
RBAC	Role Based Access Control
HTTP	Hyper Text Transfer Protocol
ML	Machine Learning
CAGR	Compound Annual Growth Rate

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background

Massive international events such as the Olympics create not just buzz but also an incredible amount of data in real time. From scores and results of athletes and matches to medal counts by country, viewer metrics, and live social media conversations, the volume of information is daunting. Historically, preparation and analysis of such data have been done manually or through fragmented systems that usually resulted in delayed insight generation. Traditional approaches involved manual work which is difficult and leads to inconsistencies, errors and even ambiguity in making decisions. The need is obvious: a new, automated, cloud-based system capable of processing high data velocity and variety without sacrificing speed or accuracy. This is where cloud-based data engineering steps in. Platforms like Microsoft Azure offer a powerful set of tools that can handle the entire data lifecycle—from ingestion and storage to processing, analysis, and visualization. By leveraging these tools, we can transform raw Olympic data into clear, actionable insights within minutes rather than days.

This project is cantered on designing and implementing a full end-to-end data pipeline using the Azure ecosystem to analyse data from the Tokyo Olympics. By combining tools like Azure Data Factory, Data Lake Gen 2, Databricks, Synapse Analytics, and popular dashboarding tools like Power BI, Looker Studio, and Tableau, it becomes possible to create a seamless system that can process Olympic-level data at scale.

This project leverages that very approach—bringing the power of Azure to demonstrate how raw, unstructured data from the Tokyo Olympics can be turned into meaningful, interactive dashboards that inform, engage, and inspire.

1.2 Relevance

The project enables a user to understand the modern workflow for data analytics. Right from data collection to data analytics the approach mentioned in this approach is the most widely used in industries and companies. This project is relevant in the following ways:

- **Real-World Data Engineering**: At its core, this project is a hands-on application of data engineering principles. Data engineering is the discipline focused on building systems that collect, store, and move data in a way that it becomes usable for analytics, reporting, or machine learning. The raw .csv files—like Athletes.csv, Teams.csv, Medals.csv, and others—represent structured datasets, but in isolation, they're not immediately insightful. The real challenge lies in integrating them, cleaning them, understanding relationships between them, and making them analysis-ready. This mirrors exactly what data engineers do in industries like finance, e-commerce, sports, and healthcare. Further data analysts make use of the meaningful data and prepare dashboards and draw conclusions so that better decision can be taken!
- **Data Analytics:** The global data analytics market size was valued at USD 64.99 billion in 2024. The market is projected to grow from USD 82.23 billion in 2025 to USD 402.70 billion by 2032, exhibiting a CAGR of 25.5% during the forecast period [1]. From an analytics point of view, the datasets that we have used are goldmines. Once cleaned and structured, they can be queried to answer highly practical questions such as which country had the most balanced gender participation? Or did the number of medals correlate with coach-to-athlete ratio etc. Such analysis isn't academic—it reflects real business needs in the sports and broadcasting industries. Sports federations, coaches, trainers etc. can use this to allocate training resources. Further media companies can use it to build engaging data-driven content and even policymakers can use it to promote gender equity in sports.

1.3 Literature Survey

The literature survey analyzes the previous work done by authors in the field big data and analytics.

Table 1.3.1 Various Analytical Techniques

Technique	Description	Strengths	Limitations
Excel-Based Analysis	Manual data entry and analysis using spreadsheets like MS Excel	Easy to learn, widely available, suitable for small datasets	Not scalable, prone to human error, limited automation
SQL Querying	Using SQL to pull and filter data from relational databases	Precise data retrieval, powerful for structured data	Not ideal for unstructured or large-scale data; lacks advanced visualization
OLAP Cubes	Multi-dimensional data models for business intelligence	Fast slicing and dicing of pre-aggregated data	Static in nature, requires pre-configuration, poor support for real-time data
Desktop BI Tools	Standalone analytics platforms (e.g., SAP BO, SAS, Crystal Reports)	Good for enterprise reporting, user-friendly dashboards	High cost, limited integration with cloud or streaming data sources
Scripting in R/Python	Use of custom scripts to clean, analyze, and visualize data	Flexible, supports advanced analytics and visualization	Requires coding expertise, lacks drag-and-drop simplicity
On-Premise Warehousing	Local server-based data storage and analysis setups	Data control, suitable for confidential data	Expensive to maintain, difficult to scale, lacks agility

Table 1.3.2 Previous Work Done

Author	Analytical Approach	Advantages	Disadvantages
Alfredo Cuzzocrea et al. [2]	OLAP & Data Warehousing based approach	Better quality, easy integration, interactive exploration.	Poor performance on unstructured data.
S. Sagiroglu & D. Sinanc [3]	Statistical Analysis	Simple to implement, easy to visualize, multi-integrated model	Required extensive knowledge
Samuel Fosso Wamba et al. [4]	Proposed a research model for big data analytics	Dynamic approach, flexibility, factor analysis	Unstable, cannot be used in large industry application
Erik Brynjolfsson & Kristina McElheran [5]	Indicators for data driven decision making were calculated.	Used human capital, plant size etc. for consideration.	Average adoption rate is 0.3
Prakhar Maheshwari et al. [6]	Cloud computing approach for big data analysis.	Practical approach, Implementation of ML Algorithms.	Latency, high costs.

1.4 Motivation

The rapid growth of digital systems and connected devices has led to significant increase in the volume, velocity, and variety of data generated daily. Traditional data processing methods are no longer sufficient to handle the scale and complexity of this information. This gap between data availability and actionable insight has created the need for robust big data analysis frameworks.

The motivation for undertaking big data analysis stems from the demand to extract structured meaning from unstructured or semi-structured data sources. Organizations and institutions across sectors are now required to make faster, data-driven decisions—often in real time. These decisions may relate to operational efficiency, customer behaviour, risk mitigation, or long-term strategic planning. From a technical standpoint, big data analysis involves designing and implementing workflows that can efficiently process large datasets across distributed environments. These workflows typically include data acquisition, cleansing, transformation, modelling, and visualization. Each stage presents unique challenges and opportunities for innovation. The integration of automated pipelines, parallel processing, and scalable storage solutions has become essential to manage such data-intensive tasks effectively.

Furthermore, developing an end-to-end big data pipeline helps bridge the gap between raw data collection and final decision-making. The ability to detect patterns, generate forecasts, and build adaptive systems relies heavily on the precision and performance of the analytical models deployed.

1.5 Aim

The aim of the project can be broken down into 5 concise technical points:

- Ingest and integrate raw Tokyo Olympics datasets from multiple CSV HTTP APIs into Azure Data Lake Storage for centralized access.
- Build and orchestrate a scalable data pipeline using Azure Data Factory to automate data movement and transformation.
- Perform data cleaning and transformation using Azure Databricks to prepare the datasets for structured analysis.
- Store processed data in Azure Synapse Analytics to support querying and analytical workloads efficiently.
- Visualize key insights (e.g., medal tallies, athlete stats, gender distribution) using Databricks or a suitable dashboarding tool such as Power BI.

1.6 Scope and Objectives

This project focuses on building a fully functional data engineering pipeline that demonstrates the capabilities of cloud-based tools in processing and analyzing real-world datasets. Using the Tokyo Olympics dataset as the central input, the scope includes data ingestion, transformation, storage, querying, and visualization—delivered through a cohesive Azure-based infrastructure. The pipeline aims to handle structured CSV files, ensuring data reliability through validation and preprocessing steps. The scope extends from initial raw data collection all the way to generating business-relevant dashboards that provide clear, visual insights. The pipeline will be modular, scalable, and reusable—allowing similar datasets to be processed with minimal configuration changes.

Additionally, the project illustrates how multiple Azure services (Data Lake, Data Factory, Databricks, Synapse Analytics, and Power BI) can be integrated seamlessly to create a robust data solution suitable for enterprise-level analytics.

Key constraints include:

- Processing static CSV files (no streaming data involved)
- Focus on analysis and reporting; no/little predictive modelling
- Azure tools only; no hybrid or on-premise components

To accomplish the stated aim, the project will pursue the following objectives:

1. **Design a modular Azure-based architecture** that supports end-to-end data processing workflows suitable for large-scale datasets.
2. **Ingest multiple CSV datasets** related to the Tokyo Olympics into Azure Data Lake Storage using automated data pipelines. This is done using an online HTTP API.
3. **Clean and transform raw data** using Azure Databricks, including handling null values, standardizing formats, and merging related datasets.
4. **Load the processed data** into Azure Synapse Analytics to enable efficient querying, aggregation, and analysis. Here we can perform SQL Querying.
5. **Develop visualizations** that offer visual insights into key metrics such as medal counts, country-wise performance and sport-wise trends.
6. **Ensure maintainability and reusability** of the pipeline components for future datasets or similar sports data use cases.
7. **Document each stage** of the workflow clearly to support scalability, transparency, and possible deployment in real-world environments.

1.7 Technical Approach

The project follows a five-stage approach:

1. Data Ingestion: The process begins with uploading the raw Tokyo Olympics datasets (in CSV format) to Azure Data Lake Storage Gen2. As we are using Azure services or this, we will not be uploaded our data sources locally but rather through an HTTP API to Azure Data Factory in the form of databases These datasets include information on athletes, events, countries, medals, and other key metrics. Data is uploaded either manually or through Azure Data Factory pipelines to ensure consistent access and version control. Folder structures and naming conventions are used to logically organize raw and processed files.

2. Data Orchestration: Once the data is available in the Data Lake, Azure Data Factory (ADF) is used to automate and schedule the movement of data between services. ADF pipelines are configured to:

- Trigger jobs based on time or events
- Move raw data to processing zones
- Log each activity to ensure traceability The orchestration layer ensures repeatability and scalability of the entire workflow.

3. Data Transformation: Data transformation is carried out using Azure Databricks, a distributed computing platform based on Apache Spark. Here, the raw CSV files undergo several pre-processing steps, including:

- Removing null or duplicate entries
- Standardizing date, name, and numeric formats
- Joining datasets (e.g., linking athlete IDs with their medal records)
- Generating additional columns (e.g., total medals per country) Python and PySpark scripts are developed within notebooks to execute these operations efficiently in a distributed environment.

4. Data Storage and Querying: Transformed data is then loaded into Azure Synapse Analytics, a scalable SQL-based data warehouse. Tables are designed to support both detailed and aggregated views, enabling fast querying. This acts as the analytical engine of the project. We can perform retrieval of data from tables and even see the visualizations if needed.

5. Data Visualization

The final step involves connecting Power BI to Azure Synapse to build interactive, real-time dashboards. The dashboards include:

- Bar charts for medal tallies by country
- Pie charts for gender and sport distributions
- Filters for dynamic exploration (e.g., by year, sport, or nation) Power BI dashboards are published to the cloud and shared via secure links.

6. Security and Monitoring

Throughout the pipeline, appropriate access controls are applied via Azure Role-Based Access Control (RBAC). Logs are maintained within ADF and Databricks to track failures, performance bottlenecks, and job success rates.

CHAPTER 2

System Architecture

2.1 System Architecture

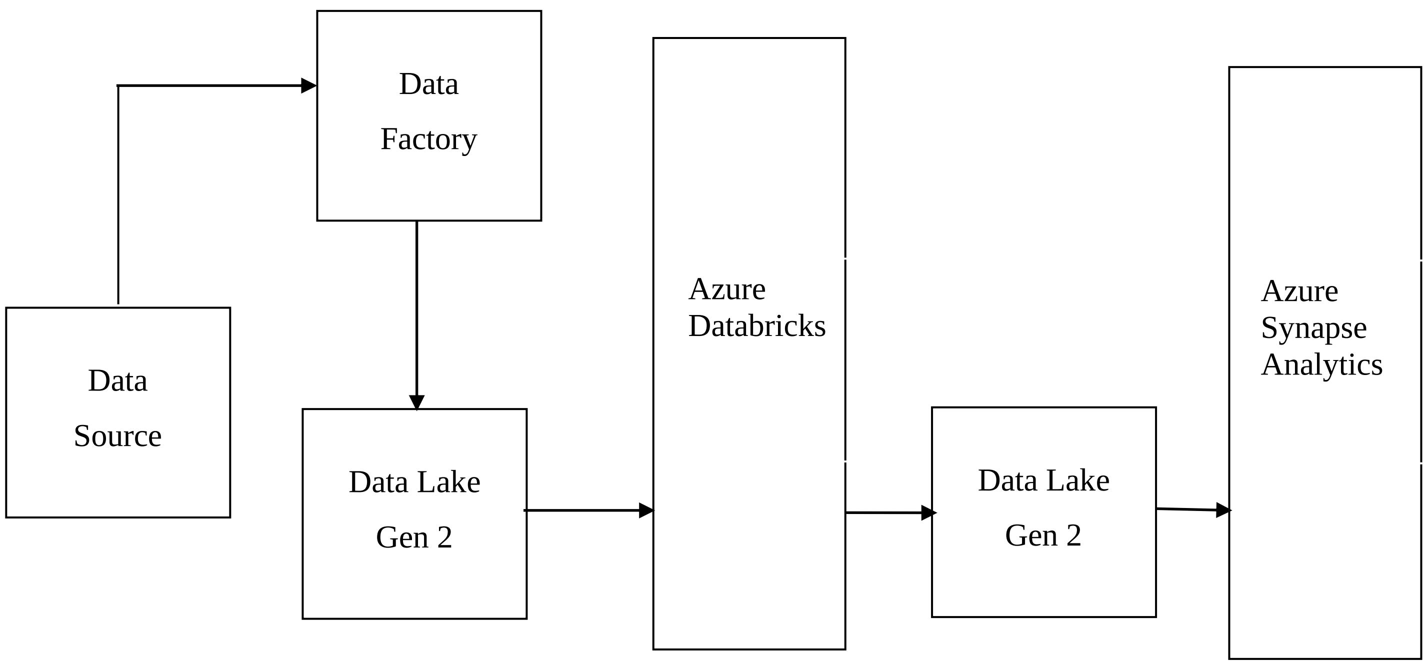


Fig 1. Block Diagram

The system block diagram provides a structured view of the overall data engineering pipeline used for processing the Tokyo Olympics datasets on Microsoft Azure.

1. Data Source: The process initiates with structured datasets containing information about Olympic events, participants, results, and related attributes. These files are assumed to be available either from open repositories or APIs and serve as the base input for the pipeline. This data is not hosted locally but streamed or uploaded via HTTP routes into the Azure ecosystem.

2. Data Integration – Azure Data Factory: Azure Data Factory (ADF) operates as the intake mechanism, automating the import of external files into the platform. It functions as the controller for triggering data pipelines.

3. Raw Data Storage – Azure Data Lake Gen2: Once data enters the system, it is deposited in Azure Data Lake Gen2. This component offers reliable cloud-based storage with hierarchical folder management. Here, raw files are held in their original format and organized by type or category (e.g., medals, athletes, sports). This layer serves as a temporary resting point before processing begins.

4. Data Transformation – Azure Databricks: Azure Databricks is responsible for processing the raw datasets. Built on Apache Spark, it allows scalable, distributed computation. Within this block, scripts written in PySpark cleanse the data, fix formatting inconsistencies, remove null values, join related tables, and derive new features such as medal totals or participant counts. The output of this transformation is stored separately to preserve the integrity of original inputs.

5. Transformed Data – Azure Data Lake Gen2: After processing, the cleaned and enriched datasets are written back to the Data Lake but under a new path labelled as "Transformed Data." This step ensures separation of concerns between unprocessed and ready-to-analyse files. Data in this zone is optimized for analytical workloads.

6. Analytics Layer – Azure Synapse Analytics: Transformed datasets are then ingested into Azure Synapse Analytics. This service enables structured querying, data modelling, and performance optimization for high-volume analytics. It provides support for SQL-like syntax to interact with datasets and acts as the intermediary between storage and dashboard layers. The system can aggregate and filter information based on conditions such as sport type, year, or country.

2.2 Performance Design

Apache Spark serves as the computational engine within Azure Databricks, enabling large-scale data processing across multiple nodes in a distributed fashion. Rather than processing data sequentially on a single machine, Spark breaks the data into smaller partitions and processes them concurrently across a cluster. The data preprocessing tasks like removing duplicates and nulls are distributed automatically and combining data sources like medals, athletes, teams etc are done effectively and optimized using shuffle and broadcasting techniques depending on the dataset size.

Moreover, Spark maintains a directed acyclic graph (DAG) to manage dependencies between operations. This allows Spark to optimize execution plans and recover from failures without reprocessing the entire data, enhancing reliability.

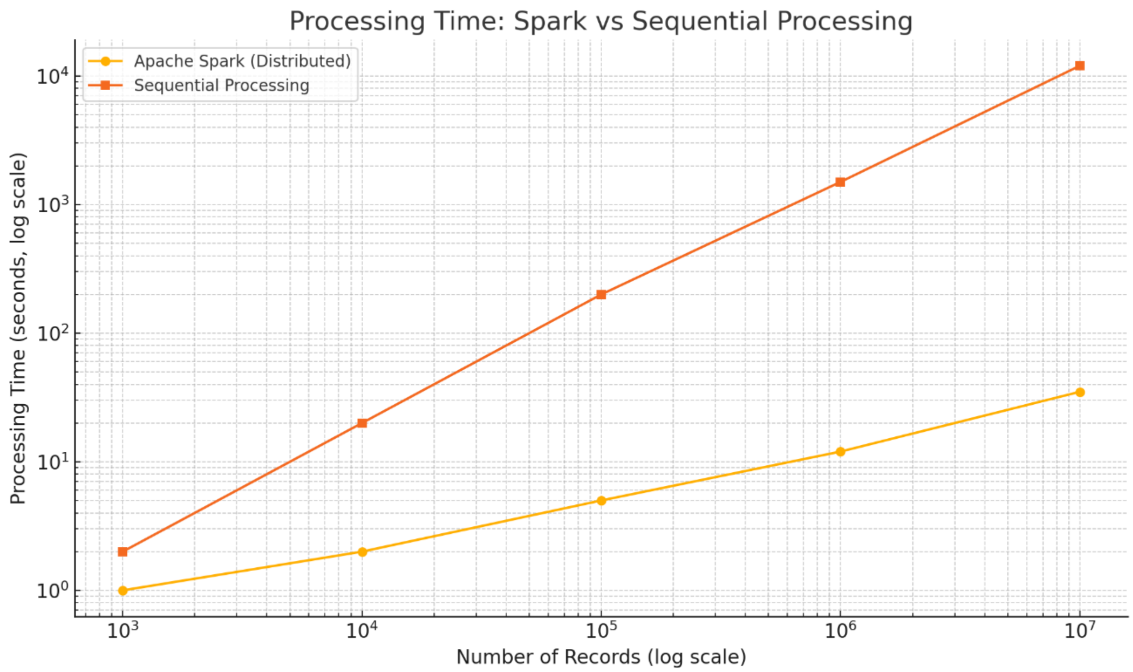


Fig 2. Processing Comparison

CHAPTER 3

Results and Discussion

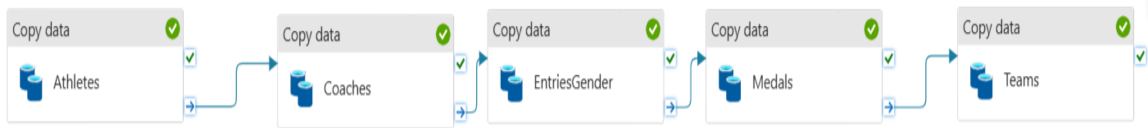
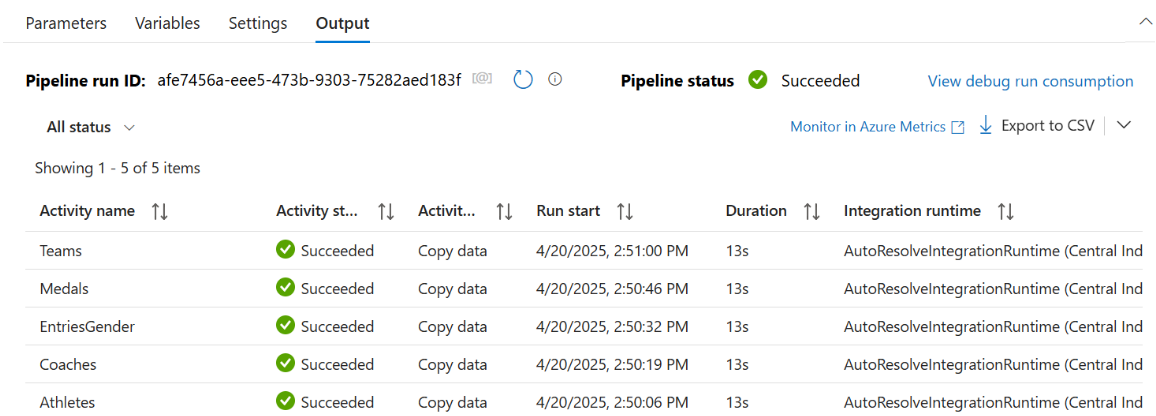


Fig 3. Data Ingestion

The above figure shows that data has been loaded and ingested successfully. Now our data pipeline for further processing is ready.



Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime
Teams	Succeeded	Copy data	4/20/2025, 2:51:00 PM	13s	AutoResolveIntegrationRuntime (Central Ind
Medals	Succeeded	Copy data	4/20/2025, 2:50:46 PM	13s	AutoResolveIntegrationRuntime (Central Ind
EntriesGender	Succeeded	Copy data	4/20/2025, 2:50:32 PM	13s	AutoResolveIntegrationRuntime (Central Ind
Coaches	Succeeded	Copy data	4/20/2025, 2:50:19 PM	13s	AutoResolveIntegrationRuntime (Central Ind
Athletes	Succeeded	Copy data	4/20/2025, 2:50:06 PM	13s	AutoResolveIntegrationRuntime (Central Ind

Fig 4. Completed Data Pipeline

The status of the data pipeline is shown above. The activity log and integration runtime are noted and ensure that any errors (if occurred) are clearly reported for debugging.

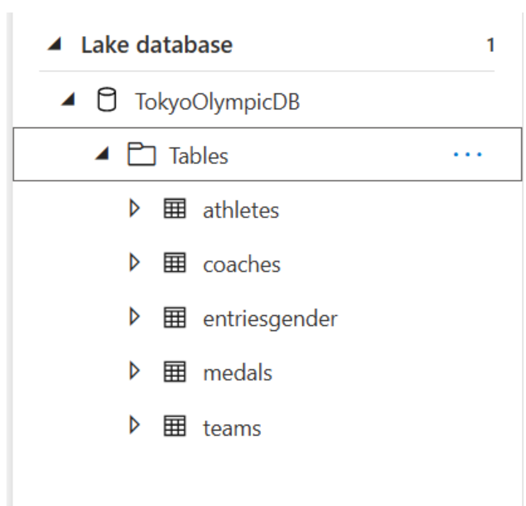


Fig 5. Transformed Data Lake Gen 2

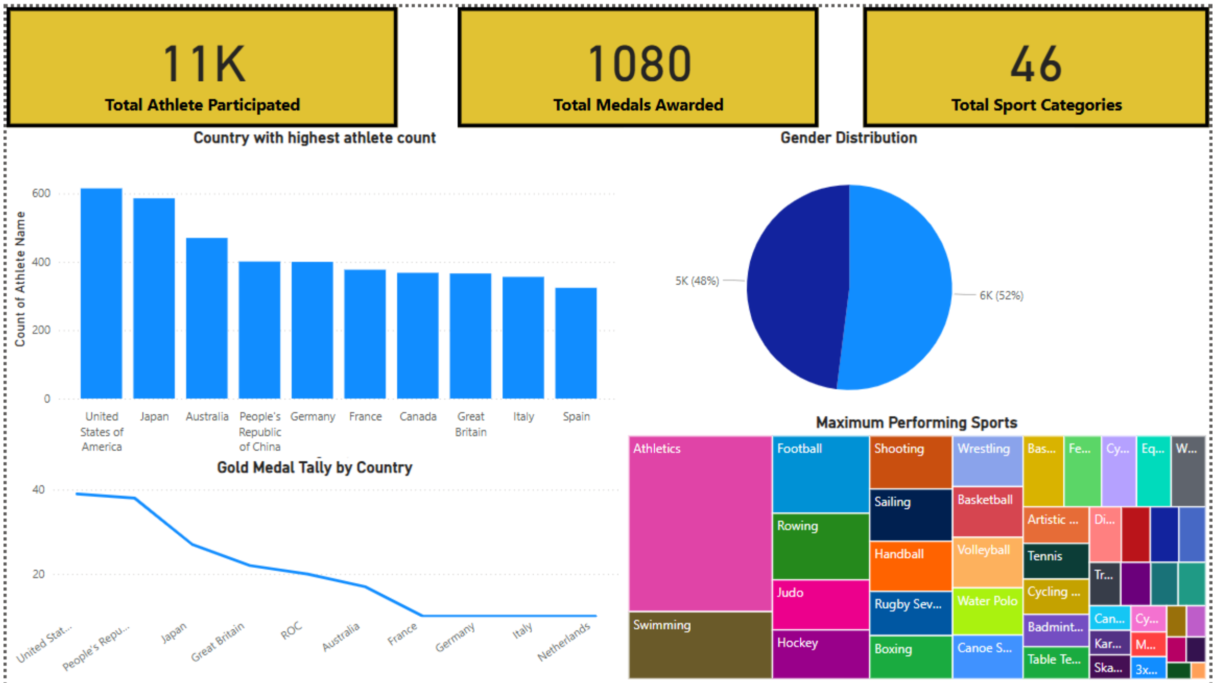


Fig 6. Power BI Dashboard

The above dashboard has been created in Power BI. Power BI provides an intuitive, drag-and-drop interface that allows users to create visually engaging dashboards without the need for extensive coding knowledge. It supports real-time data connections with Azure Synapse Analytics, ensuring that insights are always based on the most updated data. The wide variety of visualization options such as bar charts, pie charts, slicers, and filters enables interactive exploration of data, making it easier to identify trends, patterns, and outliers. Additionally, Power BI integrates seamlessly with other Microsoft Azure tools, streamlining the workflow from data ingestion to final presentation

CHAPTER 4

Conclusions

This project successfully demonstrates the construction of an end-to-end data engineering pipeline using Azure cloud services, specifically designed to analyse and visualize insights from the Tokyo Olympics dataset. Through the integration of various Azure components—such as Data Factory, Databricks, Synapse Analytics, and Power BI—the pipeline automates the flow from raw ingestion to interactive dashboards.

The data architecture efficiently handles diverse datasets including athletes, coaches, teams, and medal distributions. Using distributed processing within Azure Databricks, large volumes of records were transformed and aggregated without performance bottlenecks. The results were stored in a structured format within Synapse Analytics, enabling fast, SQL-like querying and seamless dashboard connectivity. We observed a direct correlation between number of athletes and the medals won.

Some key insights derived from the processed data include:

- The United States led the overall medal tally, contributing approximately 28.6% of total medals among the top 5 countries.
- People's Republic of China followed with 22.3%, and ROC with 18%, showcasing a competitive medal distribution at the global level.
- In the comparison across countries, the USA not only topped in medals but also had the highest number of participating athletes, clearly reflecting its large delegation and wide participation across events.
- Countries like Germany and Australia had strong athlete counts and team entries, although their medal counts were lower in relative terms.

By employing this pipeline, the project proves the efficiency of modern cloud-based data analytics in extracting meaningful outcomes from complex datasets. The flexibility of the system allows for further enhancement, including the addition of streaming sources, predictive models, or real-time dashboards.

CHAPTER 5

Future Scope

One promising direction is to incorporate streaming data sources such as live event results or social media sentiment. By integrating services like Azure Event Hubs or Azure Stream Analytics, the pipeline can be extended to support real-time dashboards that update dynamically during live sports events.

The current setup is focused on descriptive analytics. A natural progression is to incorporate predictive models using Azure Machine Learning or Databricks MLlib. This can enable forecasting medal outcomes, athlete performance trends, or participation predictions based on historical data and demographic variables.

Additional data sources, such as athlete training logs, weather conditions, or country-wise funding in sports, can be added to provide more context-rich analytics. This will allow for deeper correlations and cross-variable insights (e.g., how funding impacts performance).

The same architecture can be reused and scaled to support data from other Olympic years, Paralympic games, or international sports events like the FIFA World Cup or Commonwealth Games. This would allow trend comparison across time and competitions.

References

1. https://www.fortunebusinessinsights.com/data-analytics-market-108882
2. Cuzzocrea, Alfredo, Ladjel Bellatreche, and Il-Yeol Song. "Data warehousing and OLAP over big data: current challenges and future research directions." *Proceedings of the sixteenth international workshop on Data warehousing and OLAP*. 2013.
3. Sagiroglu, Seref, and Duygu Sinanc. "Big data: A review." *2013 international conference on collaboration technologies and systems (CTS)*. IEEE, 2013.
4. Wamba, Samuel Fosso, et al. "Big data analytics and firm performance: Effects of dynamic capabilities." *Journal of business research* 70 (2017): 356-365.
5. Brynjolfsson, Erik, and Kristina McElheran. "The rapid adoption of data-driven decision-making." *American Economic Review* 106.5 (2016): 133-139.
6. Maheshwari, Prakhar, Alankar Singhal, and Mohammed A. Qadeer. "Data analytics using cloud computing." *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2017.