# Olympics data analysis using Microsoft Azure

Seventh Semester Project Evaluation-I Report (PCC-CSBS781)

Prepared by
Debajit Samanta(18731120017)
Subhodeep Roy(18731120002)

Under the Guidance of
Prof. Bitan Misra

Batch:- 2020-2024 Semester :7 th (2023 –ODD)
Stream:- Computer Science and Business System
Year of Study: 4th

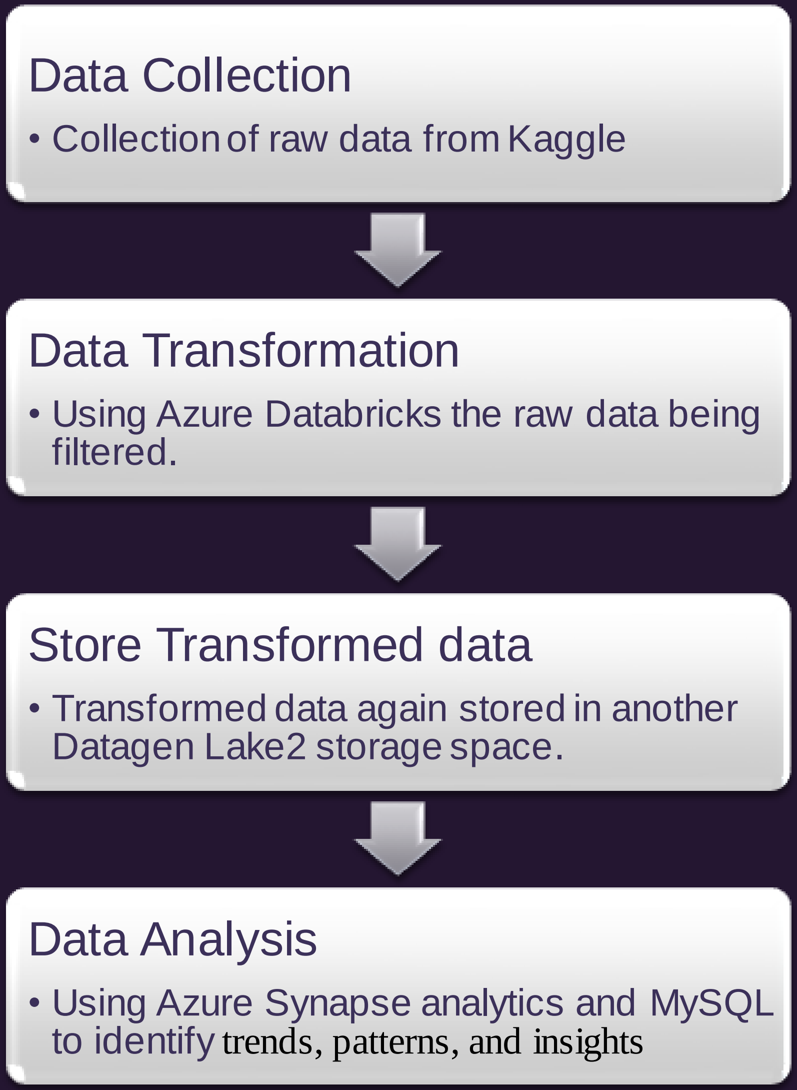
# Introduction

The "Olympics Data Analysis on Microsoft Azure" project is a comprehensive exploration of Olympic datasets leveraging the capabilities of Microsoft Azure's cloud services. The initiative involves raw data ingestion into Azure Data Lake Storage Gen2, subsequent transformation using Azure Databricks for cleansing and filtering, and secure storage back into the Data Lake. Azure Synapse Analytics facilitates advanced analysis with MySQL code, unveiling patterns and insights. Emphasis on documentation and adherence to security and compliance standards ensures transparency and data integrity. The project culminates in visually compelling representations of key findings using Power BI, showcasing proficiency in Azure services for holistic data engineering and analysis.

# Problem Definition

This project addresses the challenge of efficiently managing and analyzing large-scale Olympic Games data, emphasizing the need for a streamlined and scalable solution. The complexity of raw Olympic data, coupled with the increasing demand for real-time insights, poses a significant hurdle that necessitates advanced cloud-based technologies. The problem revolves around designing an end-to-end data processing pipeline using Microsoft Azure services to ingest, store, and analyze Olympic data, ultimately providing a platform for extracting meaningful patterns and trends. The project aims to overcome the intricacies associated with data volume, diversity, and processing speed, offering a comprehensive solution for researchers, analysts, and enthusiasts to glean valuable insights from the extensive historical and contemporary Olympic datasets.

# Proposed Methodology



## Data Collection

- Collection of raw data from Kaggle

## Data Transformation

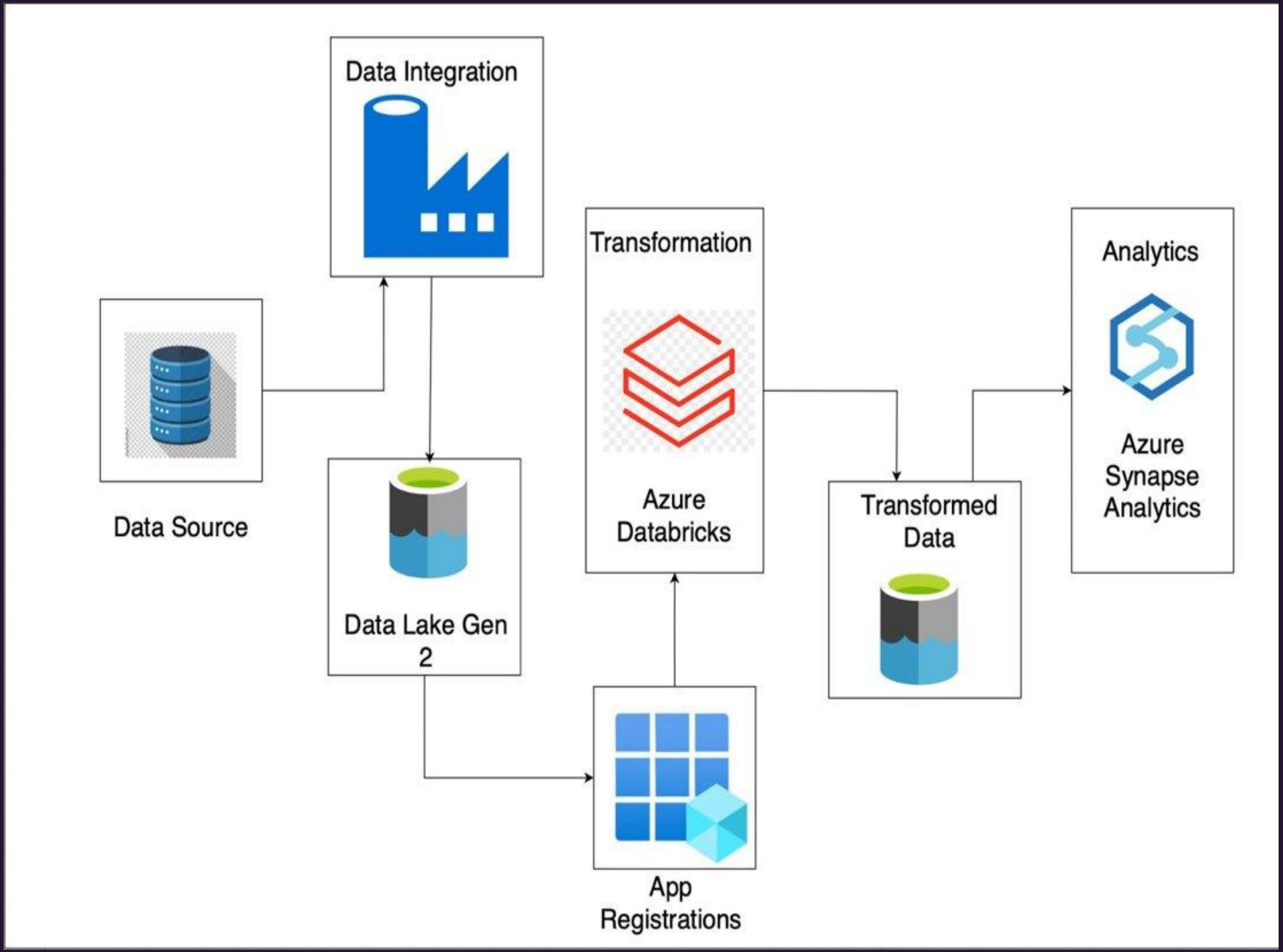
- Using Azure Databricks the raw data being filtered.

## Store Transformed data

- Transformed data again stored in another Datagen Lake2 storage space.

## Data Analysis

- Using Azure Synapse analytics and MySQL to identify trends, patterns, and insights



# Future Scope

This project addresses the challenge of efficiently managing and analyzing large-scale Olympic Games data, emphasizing the need for a streamlined and scalable solution. The complexity of raw Olympic data, coupled with the increasing demand for real-time insights, poses a significant hurdle that necessitates advanced cloud-based technologies. The problem revolves around designing an end-to-end data processing pipeline using Microsoft Azure services to ingest, store, and analyze Olympic data, ultimately providing a platform for extracting meaningful patterns and trends. The project aims to overcome the intricacies associated with data volume, diversity, and processing speed, offering a comprehensive solution for researchers, analysts, and enthusiasts to glean valuable insights from the extensive historical and contemporary Olympic datasets.

1. Real-Time Data Integration: The incorporation of real-time data streaming to provide up-to-the-minute analytics during ongoing Olympic events. Azure Stream Analytics can be leveraged for this purpose.
2. Machine Learning and Predictive Analytics: Integrate machine learning models to predict future trends, athlete performances, or medal distributions based on historical data. This could involve implementing Azure Machine Learning services.
3. Advanced Visualizations: Enhance the visualization layer with more advanced and interactive dashboards. Using tools like Tableau or advanced Power BI features for richer visual representations.

By exploring these avenues, your project can evolve into a dynamic and impactful platform that contributes not only to the understanding of Olympic data but also to the broader field of sports analytics and data science.

# Conclusion

In conclusion, this project has successfully demonstrated the effective utilization of Microsoft Azure services to address the multifaceted challenges associated with Olympic Games data analysis. From the initial data collection and storage in Azure Databricks and Data Lake Storage Gen2 to the refined analysis using Azure Synapse Analytics, the project has showcased a robust and scalable pipeline for processing large-scale datasets. The application of MySQL code for data analysis and the integration of visualization tools have enabled meaningful insights into various aspects of the Olympic Games. As technology continues to advance, the project's future scope encompasses real-time data integration, machine learning, advanced visualizations, and broader collaborations with sports organizations. The continuous optimization of the data processing pipeline and adherence to data security and privacy measures further contribute to the project's relevance and sustainability. In essence, this endeavor not only signifies a successful implementation of cloud-based data analytics but also opens avenues for ongoing exploration and contribution to the evolving landscape of sports analytics and data science.

# Thank You