Title: End-to-End Azure Data Engineering for Tokyo Olympic Analytics.


| Name | Registration Number |
| --- | --- |
| Aniruddha S | 240913012 |
| Kedaresh Inamdar | 240913017 |


**Aim :** To create a pipeline to analysis the data and visualize it using Azure Cloud.

## Introduction

The Tokyo Olympic Data Analytics project is an end-to-end data engineering solution built on Microsoft Azure to process and analyze Olympic data efficiently. It leverages Azure Data Factory, Data Lake Storage, Azure Databricks, and Azure Synapse Analytics for seamless data ingestion, transformation, and analytics. The project enables scalable, automated ETL/ELT pipelines to process structured and semi-structured data for real-time insights. Interactive dashboards in Power BI, Looker Studio, and Tableau help visualize key performance trends. This paper presents the architecture, implementation, challenges, and insights gained from the project.

## System Architecture

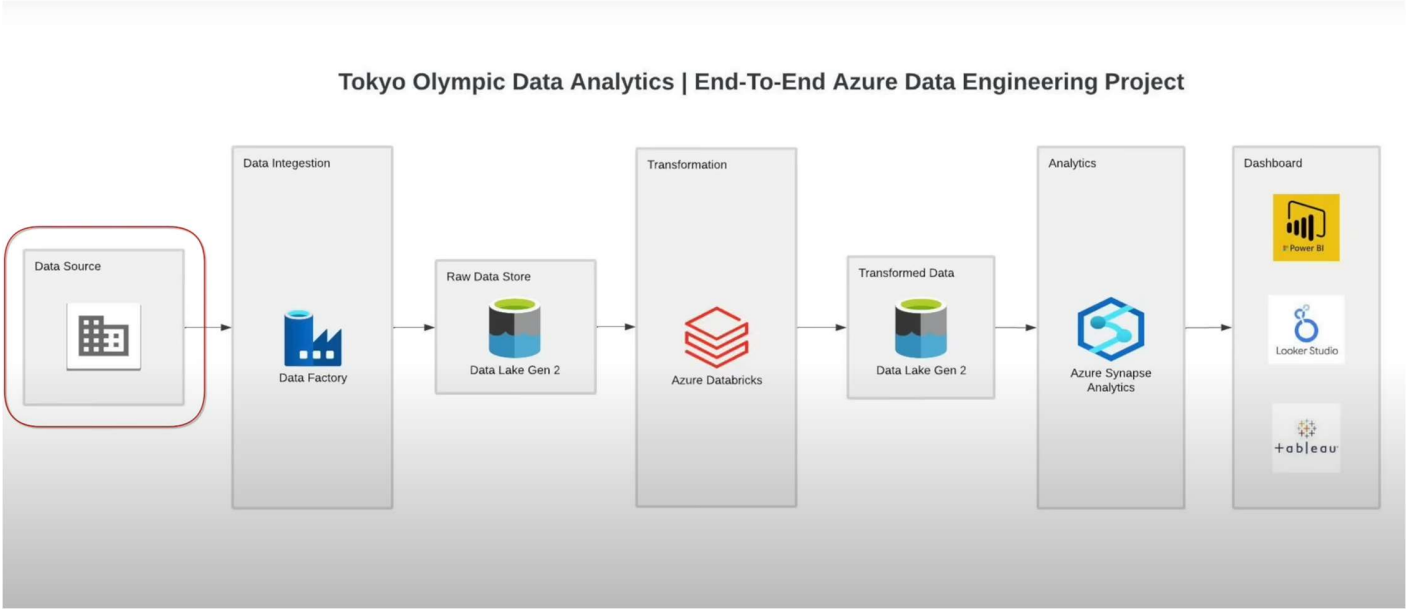


Figure 1 : System Architecture

## Functional Requirements

1. Data Integration & Storage
    - Azure Data Factory for automated data ingestion from multiple sources.
    - Azure Data Lake Storage Gen2 for scalable and cost-effective raw data storage.
    - Support for structured and semi-structured data formats (CSV, JSON, Parquet).
2. Data Transformation & Processing
    - Azure Databricks for distributed data processing using Spark.
    - ETL/ELT Pipelines for data cleansing, aggregation, and transformation.
    - Schema validation and error handling to ensure data consistency.
3. Analytics & Data Modeling
    - Azure Synapse Analytics for querying and analyzing large datasets efficiently.
    - Data partitioning and indexing for performance optimization.
    - Dimensional modeling techniques (Star Schema, Snowflake Schema) for effective analytics.
4. Visualization & Reporting
    - Power BI, Looker Studio, Tableau for interactive dashboards and reports.
    - Role-based access control (RBAC) and IAM policies for data security.
    - Real-time and historical trend analysis of Olympic data.
5. Scalability & Performance Optimization
    - Parallel processing and distributed computing for handling large datasets.
    - Cost-optimized storage and compute solutions using Azure pricing models.
    - Auto-scaling capabilities to handle varying workloads dynamically.

## Output:

- Transformed and Analyzed Olympic Data – Cleaned, structured, and optimized datasets stored in Azure Data Lake Gen2 and processed using Azure Databricks for analytics.
- Interactive Dashboards & Reports – Visual insights on country-wise medal counts, athlete performance trends, and event-wise statistics using Power BI, Looker Studio, and Tableau.
- Scalable & Automated Data Pipeline – A robust Azure-based ETL/ELT pipeline that automates data ingestion, transformation, and analysis, ensuring efficiency and scalability for large datasets.