Olympic Data Analytics

GROUP 10

Prerak Panwar (Roll number 30), Yi Hsiang Hung (Roll number 17)

In this project, our objective is to conduct a comprehensive analysis of the Olympic dataset available on Kaggle. To achieve this, we will implement an end-to-end data engineering project, incorporating various Azure services.

1. Data Extraction using Azure Data Factory:

We will initiate the project by extracting data from the Olympic API using Azure Data Factory.

Azure Data Factory, a powerful data pipeline tool in the Azure ecosystem, will be employed to create a seamless flow for retrieving and loading our data onto Azure Data Lake storage.

2. Data Transformation with Azure Databricks:

Following the data extraction phase, we will utilize Azure Databricks to employ Spark code for data transformation.

This step involves cleaning, structuring, and enhancing the raw data. The transformed data will then be stored back into the Azure Data Lake storage for further processing.

3. Data Analysis using Azure Synapse Analytics:

Azure Synapse Analytics will be employed to run SQL queries on the transformed data.

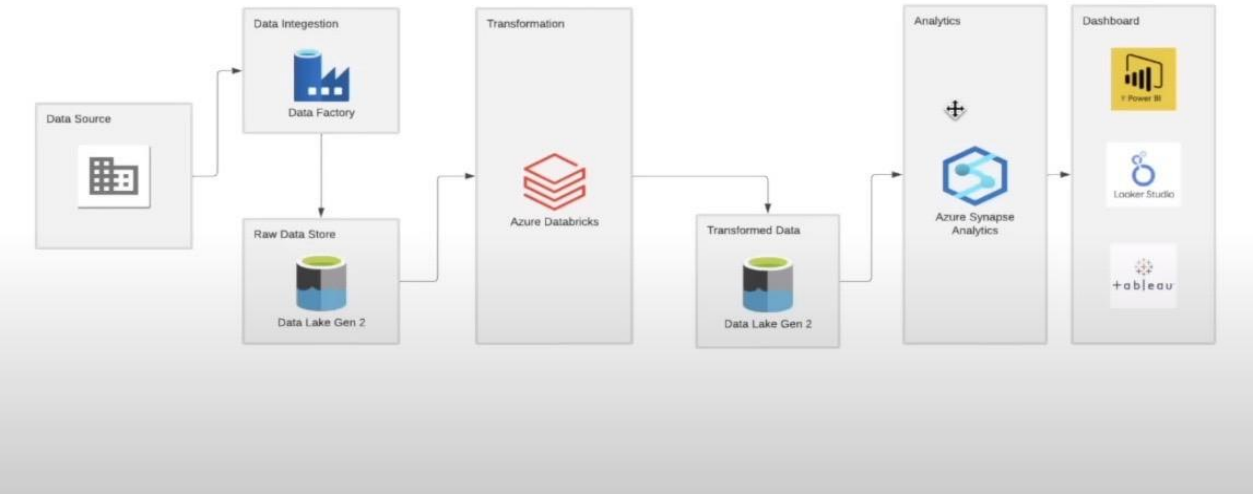
This step will provide valuable insights into the dataset, enabling us to derive meaningful information.

4. Data Visualization in PowerBI (Optional):

For enhanced visualization and presentation of our findings, we have the option to integrate with business intelligence tools like PowerBI. This step will allow us to create visually appealing and informative dashboards based on the analyzed Olympic dataset.

This structured approach ensures a systematic and efficient execution of the Olympic dataset analysis, leveraging the capabilities of Azure services at each stage of the data engineering process.

This picture represents the complete Architecture of the Project work:



More information about the Azure services:

1-Azure Data Factory (for Data Ingestion):
Data integration service that enables you to create, schedule, and manage data pipelines for efficient data movement and transformation between various sources and destinations in Azure and beyond. It simplifies ETL (Extract, Transform, Load) and data integration tasks.

2-Data Lake Gen 2 (for Data Storage):
Data lake solution that combines the capabilities of a data lake with the power of Azure Blob Storage, allowing you to store and analyze large volumes of structured and unstructured data with enhanced performance, security, and analytics capabilities.

3-Azure Databricks (for Data Transformation)
Databricks is a unified analytics platform built on top of Apache Spark, designed to help data engineers and data scientists collaborate on big data processing and machine learning tasks. It provides tools for data exploration, data processing, and building machine learning models in a collaborative and scalable environment.

4-Azure Synapse Analytics (for Data Analysis)
SQL Data Warehouse is a cloud-based analytics service provided by Microsoft Azure. It combines big data and data warehousing into a single integrated platform, allowing organizations to analyze and process large volumes of data for business intelligence and data analytics purposes.