**Olympic Data Analytics - Azure End-to-End Data Engineering Project**

This project leverages the Microsoft Azure data stack to ingest, process, analyze, and visualize Olympic Games data using a modern data engineering pipeline. The architecture supports scalability, performance, and real-time analytics, transforming raw data into actionable insights.

**Architecture Overview**

The data pipeline consists of the following stages:

1. **Data Integration** using **Azure Data Factory (ADF)**.
2. **Raw Data Storage** in **Azure Data Lake Gen 2**.
3. **Data Transformation** with **Azure Databricks**.
4. **Processed Data Storage** in **Azure Data Lake Gen 2**.
5. **Analytics & Querying** through **Azure Synapse Analytics**.
6. **Visualization and Dashboarding** using **Power BI**, **Looker Studio**, and **Tableau**.

**Project Workflow**

**1. Data Ingestion**

* **Tool**: Azure Data Factory (ADF)
* **Description**: ADF extracts raw data from various sources (e.g., databases, APIs, files). The data is then loaded into **Azure Data Lake Gen 2** for storage.

**2. Raw Data Storage**

* **Tool**: Azure Data Lake Storage Gen 2
* **Description**: Acts as a scalable storage system for unstructured and structured data. This is where raw Olympic data is initially stored.

**3. Data Transformation**

* **Tool**: Azure Databricks
* **Description**:
  + Perform data cleansing, transformations, and aggregations using Apache Spark within **Azure Databricks**.
  + The transformed data is stored back in **Azure Data Lake Gen 2** for further analysis.

**4. Data Analytics**

* **Tool**: Azure Synapse Analytics
* **Description**:
  + **Azure Synapse** is used for data warehousing and complex queries.
  + Processed data from **Azure Data Lake Gen 2** is loaded into Synapse, where advanced analytics and big data processing are performed.

**5. Visualization & Reporting**

* **Tools**: Power BI, Looker Studio, Tableau
* **Description**:
  + The insights derived from the processed data are visualized using popular business intelligence tools like **Power BI**, **Looker Studio**, and **Tableau**.
  + Dashboards present key Olympic metrics such as medal counts, athlete performance, and event statistics.

**Getting Started**

**Prerequisites**

Ensure you have the following Azure services and tools set up:

* **Azure Data Factory**
* **Azure Data Lake Gen 2**
* **Azure Databricks**
* **Azure Synapse Analytics**
* **Power BI** / **Tableau** / **Looker Studio**

**Setup Instructions**

**Azure Data Factory (ADF):**

* Configure pipelines in ADF to extract raw Olympic data from source systems and store it in **Azure Data Lake Gen 2**.

**Azure Data Lake Gen 2:**

* Create containers for both raw and transformed data storage.

**Azure Databricks:**

* Set up a Databricks cluster.
* Load raw data from **Azure Data Lake** into Databricks, apply necessary transformations, and save the results back to the Data Lake.

**Azure Synapse Analytics:**

* Load transformed data from **Azure Data Lake Gen 2** into **Azure Synapse Analytics** for querying and analysis.

**Power BI / Looker Studio / Tableau:**

* Connect the BI tools to **Synapse** or **Data Lake** to visualize the final outputs and create interactive dashboards.

**Key Features**

* **End-to-End Data Pipeline**: Covers the entire process from data ingestion to visualization.
* **Scalable and Efficient**: Uses Azure cloud services to handle large datasets and provide scalability.
* **Data Transformation**: Utilizes **Azure Databricks** for advanced data transformation and preparation.
* **Real-time Analytics**: **Azure Synapse** enables fast querying and analysis of processed data.
* **Interactive Dashboards**: Visualize key Olympic statistics with popular BI tools.

**Technologies Used**

* **Azure Data Factory**: For orchestrating data ingestion.
* **Azure Data Lake Storage Gen 2**: For storing raw and processed data.
* **Azure Databricks**: For transforming data using Apache Spark.
* **Azure Synapse Analytics**: For large-scale data querying and analytics.
* **Power BI** / **Looker Studio** / **Tableau**: For data visualization and reporting.