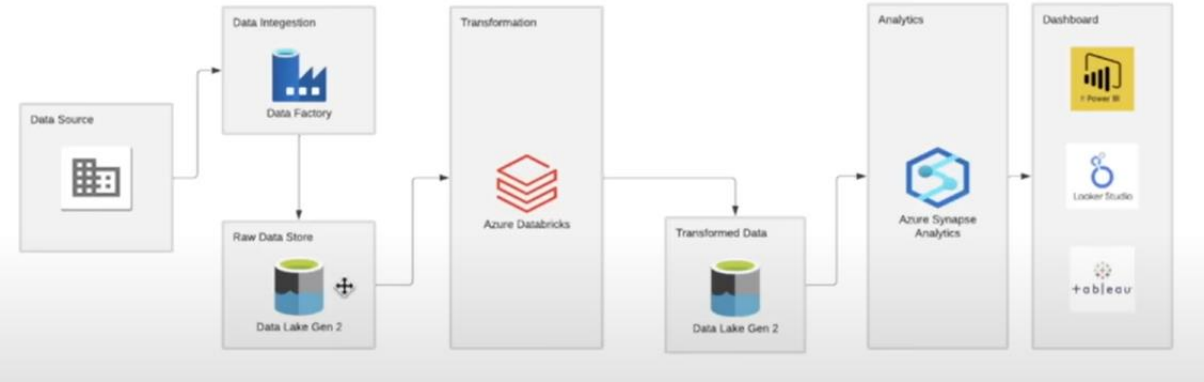
Azure – Tokyo Olympics Analysis

ARCHITECTURE



This project leverages the capabilities of Azure services to analyze Olympic data sourced from Kaggle. The workflow involves multiple stages, each utilizing different Azure tools to ensure efficient data processing and insightful analysis.

1. **GitHub Repository:**
    - **Data Source:** The Olympic dataset is stored in a GitHub repository.
    - **Initial Storage:** Acts as the initial staging area for raw data.
2. **Azure Data Factory:**
    - **Data Extraction:** Azure Data Factory is used to extract data from the GitHub repository.
    - **Data Loading:** The extracted data is loaded into Azure Data Lake Storage Gen2.
3. **Azure Data Lake Storage Gen2:**
    - **Raw Data Storage:** Serves as the storage for raw data in its staging layer.
    - **Cleaned Data Storage:** Hosts the transformed and cleaned data in a dedicated Datamart folder.
4. **Azure Databricks:**
    - **Data Transformation:** Reads raw data from the staging layer in Data Lake Storage Gen2.
    - **Data Processing:** Performs transformations such as changing data types.
    - **Data Cleaning:** Ensures data quality and prepares it for analysis.
    - **Data Loading:** Loads the cleaned data back into the Datamart folder of Data Lake Storage Gen2.

5. **Azure Synapse Analytics:**
    - **Data Connection:** Establishes a connection with the cleaned data in Data Lake Storage Gen2.
    - **Data Analysis:** Executes SQL queries for comprehensive data analysis.
6. **Power BI:**
    - **Data Visualization:** Connects to data from Azure Synapse Analytics.
    - **Dashboard Creation:** Builds interactive dashboards to visualize the analyzed data.
    - **Insights Generation:** Enables intuitive and insightful analysis through data visualizations.

This end-to-end pipeline demonstrates the integration of various Azure services to process and analyze the Olympic dataset, providing valuable insights through a robust and scalable data pipeline.