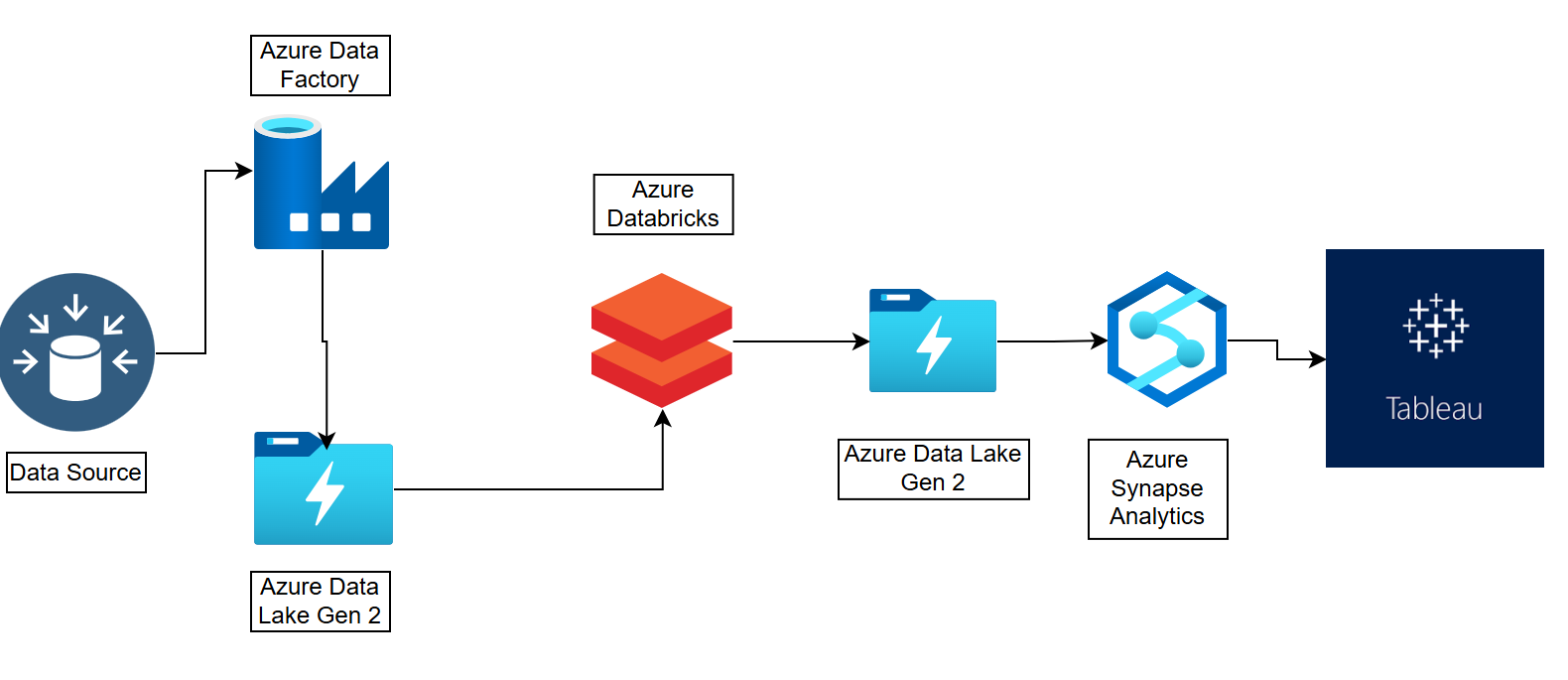
**ARCHITECTURE DIAGRAM**

****

**Azure Data Engineering: Streamlined Integration with Data Factory, Databricks, Synapse Analytics, and Tableau**

**Table of Contents**

1. Project Overview
2. Architecture
3. Setup
4. Execution

**Project Overview**

This project outlines a complete data engineering solution implemented on Microsoft Azure, showcasing how to transform raw data into actionable insights through a series of orchestrated steps. The solution employs a layered architecture to organize data and utilizes Azure Data Factory, Azure Databricks, Azure Synapse Analytics, and Tableau for data visualization.

*GitHub Dataset link*: [GitHub Datasets](https://github.com/VarshaSri21/tokyo-olympic/tree/main/Datasets)

**Architecture**

The solution comprises two key layers for data management:

1. **Raw Data Layer**: Ingests unprocessed data.
2. **Transformed Data Layer**: Cleans, processes, and refines data for analysis.

**Azure Services Used:**

* **Azure Data Factory**: Manages and automates data pipelines.
* **Azure Databricks**: Processes and transforms data using Apache Spark.
* **Azure Synapse Analytics**: Performs data warehousing and advanced analytics. It also supports the creation of external tables in a serverless SQL pool, referencing data in Azure Data Lake Gen 2. Updates in ADLS are automatically reflected in the SQL pool, enabling seamless querying without data duplication.
* **Azure Data Lake Storage Gen 2 (ADLS)**: Stores data across the Raw Data and Transformed Data layers.
* **Tableau**: Visualizes and reports data insights.

**Languages Used:**

* **PySpark**: For data transformations in Azure Databricks.
* **SQL**: For analytics and external table creation in Azure Synapse Analytics.

**Setup**

**Prerequisites**

* An active Azure subscription.
* Create a resource group to manage solution components collectively.
* Access to Azure Data Factory, Azure Databricks, Azure Synapse Analytics, Azure Data Lake Storage Gen 2, and Tableau.

**Configuration Steps**

**Data Lake Storage Gen2 Setup**

1. Create an Azure Storage Account for ADLS.
2. Establish containers for each data layer:
   * Raw Data Layer
   * Transformed Data Layer

**Azure Data Factory**

1. Create an instance of Azure Data Factory.
2. Set up Integration Runtime to connect local systems with Azure.
3. Define linked services to connect ADF with Databricks and ADLS.
   * Build pipelines to automate data movement and transform data in Azure Databricks.

**Azure Databricks**

1. Create an Azure Databricks workspace.
2. Use PySpark to:
   * Ingest raw data from the Raw Data layer.
   * Clean, process, and aggregate data for the Transformed Data layer.

**Azure Synapse Analytics**

1. Set up an Azure Synapse workspace.
2. Configure serverless SQL pools to access data in the Transformed Data layer on ADLS.
3. Create external tables to facilitate data querying.
4. Execute analytical SQL queries on the transformed data.

**Tableau**

1. Connect Tableau to Azure Synapse Analytics.
2. Develop interactive dashboards and reports.

**Execution**

1. **Ingest Data**: Use ADF to move raw data into the Raw Data layer on ADLS.
2. **Transform Data**: Run Databricks notebooks to process data through the Raw Data and Transformed Data layers.
3. **Analyze Data**: Query the Transformed Data layer using Synapse SQL.
4. **Visualize Data**: Develop and share reports using Tableau.