IST 615

Project Proposal

Tokyo Olympics in Data

Shashank Guda

Rithika Gurram

Project Goals

The objective of this project is to build a data pipeline within Azure to analyze the 2021 Tokyo Olympics dataset. This project aims to use Azure cloud services to manage, analyze, and visualize data on the 2021 Tokyo Olympics. We will build a cloud-based data engineering pipeline that processes athlete, team, and event information, unlocking insights about athlete demographics, country performance, and event participation.

- **Data Ingestion**: Create a repeatable data ingestion pipeline that securely pulls data from a GitHub-hosted CSV file and stores it within Azure.
- **Data Storage and Organization**: Implement a scalable, secure storage solution to manage raw and processed datasets in Azure Data Lake Storage Gen2.
- **Data Transformation**: Use Databricks to process, cleanse, and transform the data, preparing it for advanced analytics.
- **Data Analysis and Visualization**: Generate insights and actionable information through SQL-based analytics in Synapse and, optionally, Power BI visualizations.
- **End-to-End Integration**: Achieve a seamless flow from data ingestion to visualization, allowing stakeholders to gain insights from the Tokyo Olympics dataset in a single, unified platform.

Cloud Services and Tools

Below are the Azure services and tools that will be used in this project:

1. **Azure Data Factory**:
    - **Purpose**: Data Factory acts as the primary tool for orchestrating and automating data ingestion pipelines. It enables data extraction from the CSV files on GitHub and facilitates movement into the Azure ecosystem.
    - **Features Used**: The **Copy Data** activity within a data pipeline will be configured to retrieve the dataset over HTTP and save it into Azure Data Lake Storage Gen2.
    - **Justification**: Data Factory offers flexibility and scheduling capabilities, which allow data engineers to control the timing and reliability of data ingestion for real-time or batch processing needs.

    Azure Data Factory (ADF) initiates the process by fetching the data from the GitHub source and storing it in **Azure Data Lake Storage Gen2**. Data Factory enables an automated, scheduled data load, ensuring data is ingested into the system with minimal manual intervention. ADF's **Linked Services** and **Copy Data** activities support moving data seamlessly to and from ADLS, leveraging the HTTP protocol to pull data from GitHub.

2. **Azure Data Lake Storage Gen2**:

    - **Purpose**: This service is the primary data repository for storing both raw and processed data files.

    - **Features Used**: Gen2's hierarchical namespace provides organized storage for different stages of the data. For instance, separate folders such as “raw_data” and “transformed_data” help organize data as it progresses through the pipeline.

    - **Justification**: With its scalability, security features, and integration with other Azure services, ADLS Gen2 ensures that the data is stored in a cost-effective and high-performance environment.

    Azure Data Lake Storage Gen2 acts as a centralized data storage point. The incoming data from ADF is stored under a dedicated folder (e.g., "raw_data") within ADLS Gen2, ensuring raw data is securely stored and readily available for transformations. Post-transformation data from Databricks is written back to ADLS Gen2 under a different folder (e.g., "transformed_data"), creating a clear separation of raw and processed data.

3. **Azure Databricks**:

    - **Purpose**: Databricks is used to perform data transformations, data cleaning, and any required data engineering tasks.

    - **Features Used**: This project utilizes Databricks’ **notebooks** to execute transformations on the Tokyo Olympics data, including aggregations, filtering, and mounting ADLS Gen2 for seamless access to the data.

    - **Justification**: Databricks provides a collaborative, scalable platform for large-scale data transformations using Spark, which improves processing speed and flexibility for complex data engineering tasks.

    Azure Databricks transforms the data to make it suitable for analysis. It connects directly to ADLS Gen2 to read raw data and write transformed outputs back. Using Databricks’ **Spark clusters** and **notebooks**, data cleansing, enrichment, and aggregation are performed, preparing the data for analytical tasks.

4. **Azure Synapse Analytics**:

    - **Purpose**: Synapse serves as the central hub for data analysis, enabling SQL-based querying and analytical workflows on the transformed Tokyo Olympics dataset.

    - **Features Used**: By creating tables and databases in Synapse from the transformed datasets, we perform data analysis and SQL-based explorations.

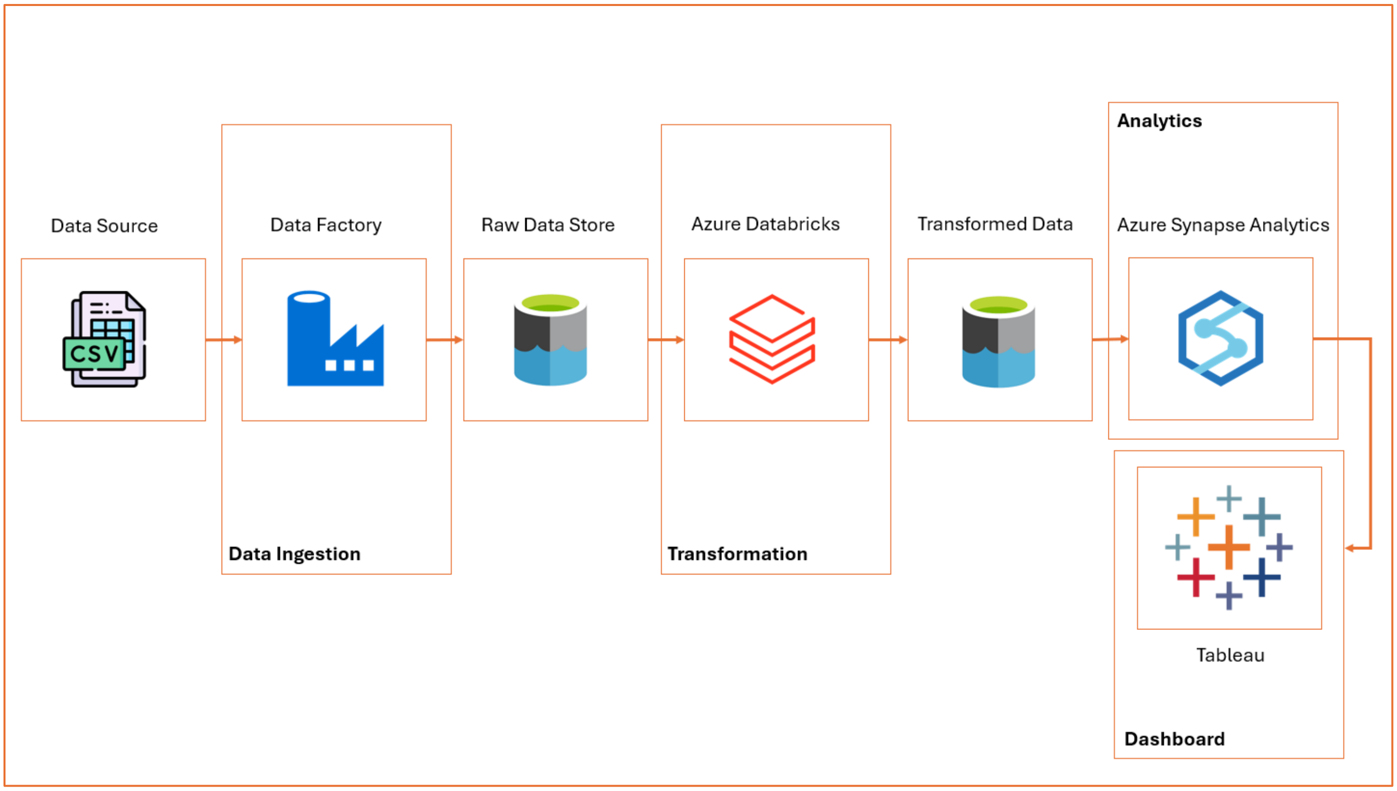
    - **Justification**: Synapse provides a powerful environment for handling data warehousing and analytics at scale. With its tight integration into the Azure ecosystem, it ensures efficient data processing and analytics workflows.

Azure Synapse Analytics creates tables from the transformed data stored in ADLS Gen2, allowing for powerful, SQL-based analysis. Within Synapse, data analysts can explore data relationships, run analytical queries, and generate insights. This component completes the pipeline by facilitating advanced analysis of the Tokyo Olympics data.

5. **Tableau:**
   - **Purpose:** Tableau will be used to create interactive dashboards and visualizations, displaying insights from the Tokyo Olympics data analysis.
   - **Features Used:** Tableau's direct connectivity with Azure Synapse allows it to pull analysis results in real time, displaying trends, comparisons, and performance metrics across different dimensions, such as country, sport, and athlete demographics.
   - **Justification**: Tableau's strong visualization capabilities and intuitive user interface make it an excellent choice for creating accessible, interactive dashboards to communicate project insights effectively.

Tableau connects to Azure Synapse, visualizing insights in an interactive and accessible format. With Tableau, dashboards will display comparisons across athletes, countries, and events, offering stakeholders a comprehensive view of trends and outcomes in the Tokyo Olympics dataset.

Architecture Diagram



1.Data Source

- **Output:** Raw data in its initial format, ready to be ingested by the Azure ecosystem.

2. Data Integration – Azure Data Factory

- **Input:** Raw data from the data source.
- **Process:** ADF extracts the data, applies basic validation (if needed), and transfers it to Azure Data Lake Storage Gen 2 for raw storage.
- **Output:** Raw data stored in Azure Data Lake Gen 2, the raw data storage layer.

3. Raw Data Storage – Azure Data Lake Storage Gen 2

- **Input:** Raw data ingested by Azure Data Factory.
- **Output:** Stored raw data, ready for transformation in subsequent steps.

4. Data Transformation – Azure Databricks

- **Input:** Raw data from Azure Data Lake Gen 2.
- **Process:** Azure Databricks loads the raw data, performs necessary transformations, and applies data processing steps like filtering, aggregation, and standardization.
- **Output:** Transformed, clean data stored back in Azure Data Lake Gen 2 for further analysis.

5. Transformed Data Storage – Azure Data Lake Storage Gen 2

- **Input:** Transformed data from Azure Databricks.
- **Output:** Ready-to-analyze data, stored and accessible for analytics services.

6. Analytics – Azure Synapse Analytics

- **Process:** Synapse loads and queries the transformed data, allowing for in-depth analysis to derive insights, trends, and metrics that are valuable for stakeholders.
- **Output:** Analytical results and datasets, which are ready to be visualized in the next stage.

7. Dashboard & Visualization - Tableau

- **Input:** Analytical outputs and query results from Azure Synapse Analytics.
- **Process:** Visualization tool connects to Synapse, pulling data to create dashboards that present key metrics, insights, and trends. These dashboards enable stakeholders to explore the data interactively.
- **Output:** User-friendly dashboard displaying analytics on Tokyo Olympic data, providing insights in real-time or near real-time.

References / Sources of Information

Below are links and resources you'll use in the project:

- **Dataset Source**: Kaggle - 2021 Tokyo Olympics Dataset
- **Azure Documentation**:
    - Azure Data Factory
    - Azure Data Lake Storage Gen2
    - Azure Databricks
    - Azure Synapse Analytics
    - Tableau