FINAL PROJECT REPORT

Project Scenario

The project started out as a broad data engineering concept that made use of Microsoft Azure's powerful cloud computing capabilities. Our objective was to build an end-to-end data pipeline that can potentially manage and analyze large datasets, with the 2021 Tokyo Olympics data from Kaggle as our core dataset.

Problem Statement

The project addressed the challenge of developing a robust, scalable data engineering system that could handle, transform, and analyze large datasets in a cloud environment. This tech was designed to enable real-world applications, such as large-scale international athletic events like the Olympics. The goal was not limited to handling massive volumes of data, but also to extract meaningful insights that might deliver deep analytical insights.

Intended Audience

This project targets a wide range of professionals, including data engineers, cloud architects, analytics teams, and authorities in charge of large-scale athletic events like the Olympics. It serves as a foundation for developing cloud-based data solutions and is necessary for those who use dynamic dashboards for real-time monitoring or display player analytics that can improve fan experiences. Individuals learning azure data services may also find this useful by seeing applied data engineering principles in the real-world.

Data Source

The data for this project was taken from the 2021 Tokyo Olympics dataset on Kaggle, which consisted of five csv files totaling around 13,000 rows. This dataset was suitable because it featured different data types and formats which required data transformations. Also, the different types of data were used to derive insights about teams, disciplines of sports, coaches, gender parity in participation for different kinds of sports and much more.

Client Perspective

Our client, a friend, had also taken DWBI last semester. He was helpful in the evaluation process for our data dashboard. His unique and third-party perspective was critical in identifying significant areas for development that may have gone unnoticed by me. By thoroughly reviewing our work, he gave important suggestions that improved the dashboard's design and functionality, ensuring that the final product was both viable and user-friendly.

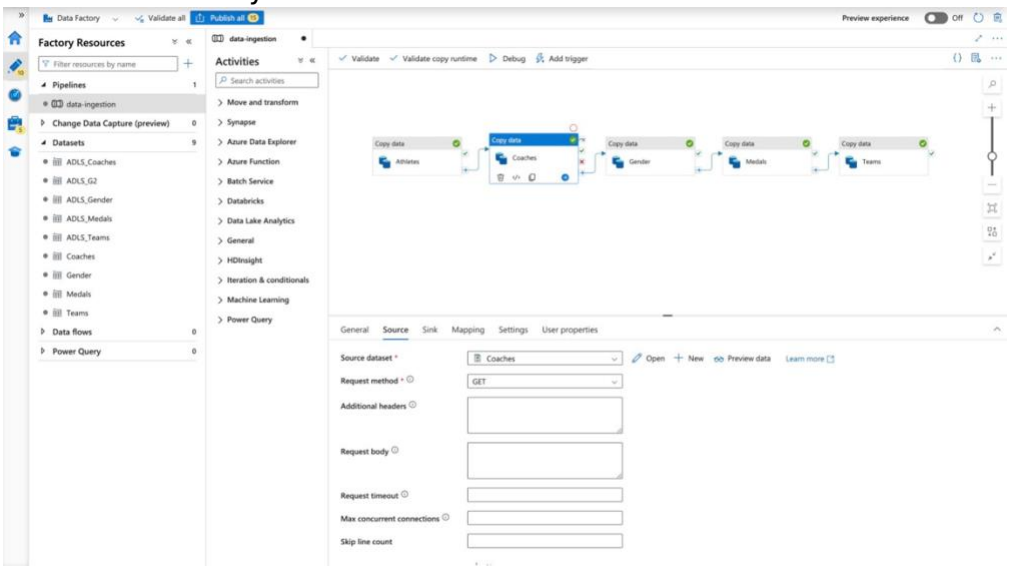
Client Contact Information

Name: Sundar Raghavan
Email: sundarr2@illinois.edu

Project Steps

1. Data Extraction with Azure Data Factory: Figure 1 depicts how I used Azure Data Factory, a cloud-based data integration solution, to automate the extraction of Olympic data from GitHub. This service simplifies the construction and scheduling of data-driven processes, allowing raw data to be fed straight into Azure Data Lake Storage with no user involvement.

Figure 1
Azure Data Factory



Note. The Figure demonstrates the Azure data ingestion to Data Lake Factory Gen 2's container. Own work

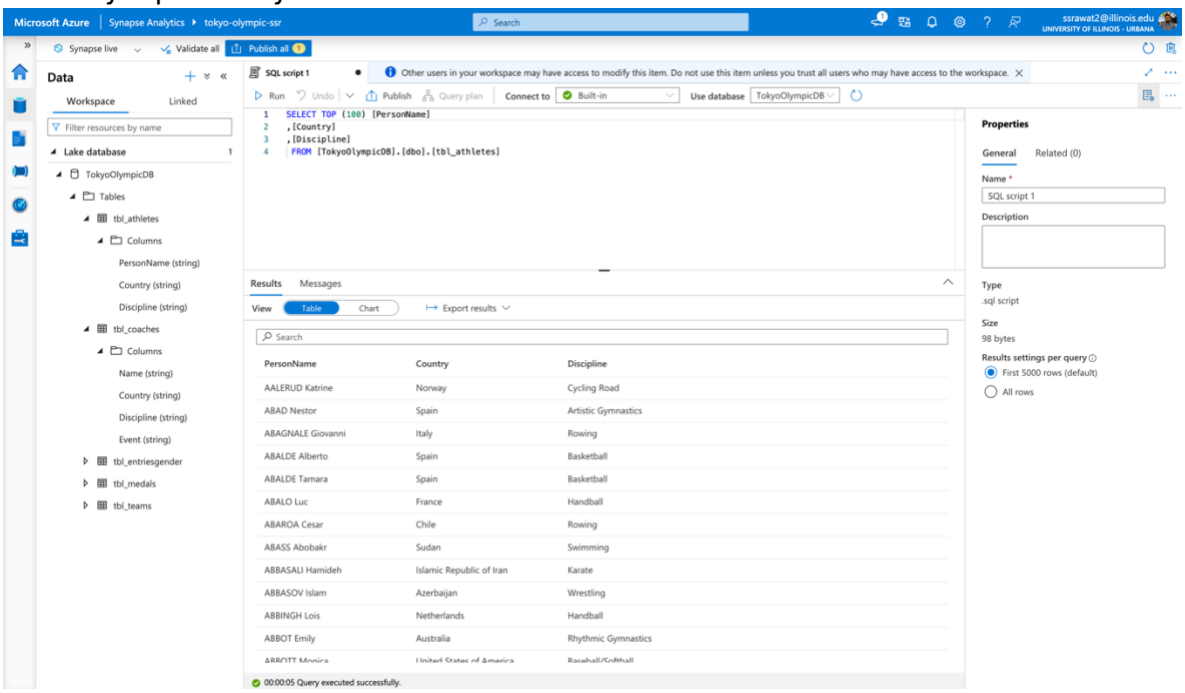
2. Data Transformation with Azure Databricks: Used Azure Databricks, a big data analytics platform powered by Apache Spark. This service provided a collaborative workspace for processing huge amounts of data. I conducted data cleansing, transformation, and normalization to prepare it for analysis. Databricks excels at managing complicated data operations at scale, including batch and real-time processing options.

3. Data Loading into Azure Data Lake Storage: After transformation, the data was placed into Azure Data Lake Storage, a scalable and secure repository for big data analytic applications. This is a large-scale storage system and provides security specifically with access control management.

4. Data Analysis with Azure Synapse Analytics: Used Azure Synapse Analytics, which is a service that combines big data and data warehousing. Synapse can execute complicated queries over processed datasets and the direct connection to data viz tools such as Tableau. I also executed some SQL queries to get some data insights and validation. The multidimensional dashboard can be seen in Figure 2.

Figure 2

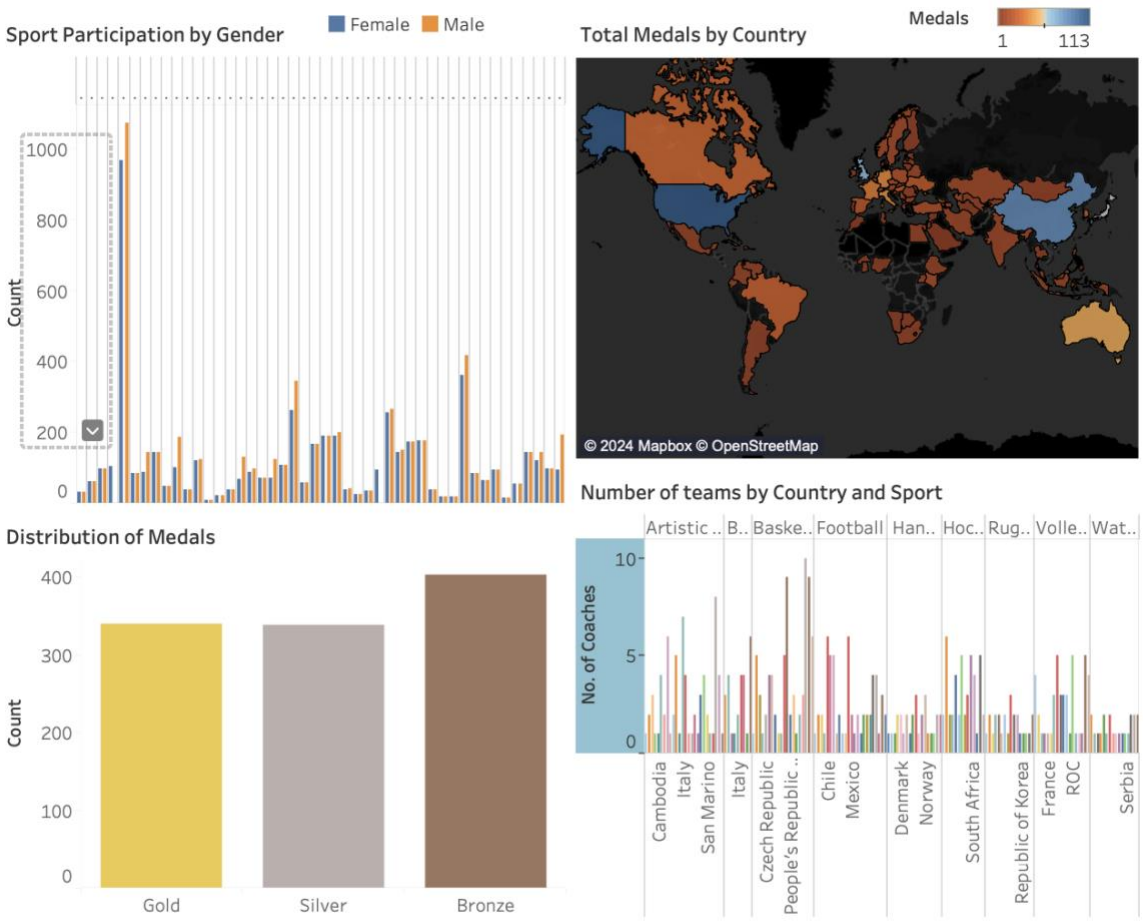
Azure Synapse Analytics Studio



Note. The Figure demonstrates the Azure Synapse Analytics studio dashboard screenshot, where I analyzed a few insights using SQL. Own work

5. Dashboard Creation in Tableau: Created an interactive dashboard using Tableau. The dashboard was connected to Azure Synapse via OAuth to retrieve real-time analytics. Tableau's features enabled me to design a user-friendly interface that displayed data through four visualizations.

Figure 3
Tableau Dashboard



Note. The Figure demonstrates Tableau Dashboard screenshot. Own work

Analysis and Discoveries

The project's analytical phase brought forward several key insights:

- Gender Participations: Visualizations provided detailed gender participation in several sports, emphasizing discrepancies and equality in athlete activity.

- Medal Distribution: Analysis of the medal count revealed significant insights into which nations lead, indicating the success of their sports programs.

- Team Participation: The stats emphasized the number of teams each country had, which gave an insight into each country's funding and participation levels in various sports.

- Interactive Insights: The introduction of interactive filters enabled dynamic querying, allowing the users to focus on certain parts of the data, such as a specific nation, sport, or discipline, so improving the dashboard's usefulness.

Challenges and Resolutions

I encountered several challenges and were subsequently resolved:

Data Connection Complexity: To maintain dashboard accuracy while dealing with external connections for data extractions, I faced issues in connecting to Azure Synapse Analytics. Learning the access management and connections protocols were difficult, but I managed by making use of the community support from Azure and Tableau

Client Feedback: During the review, the client identified specific flaws that impacted the usability of the dashboard. The map visualization had a single-color gradient at first, which lacked enough contrast to indicate major variations between data points. This makes it difficult for users to detect differences in data distribution across the map. Furthermore, the client saw that the dashboard's interactive filters were limited to only two visualizations, limiting the user's ability to play with the data across all elements of the dashboard.

Adjustments to original plan: Initial feedback on the dashboard's usability led to significant enhancements. Based on the client's feedback, I made the following adjustments to the dashboard design.

Enhanced Color Scheme: The color pattern of the map visualization was changed from a single-color to a two-color gradient. This increased the visual difference, allowing users to more easily detect trends and patterns across different countries.

Expanded Filter Functionality: The filtering feature has been expanded to all visualizations on the dashboard. This change made the dashboard more dynamic and user-friendly, allowing for a smooth data exploration across all the visualizations.

These modifications not only addressed the client's feedback but also improved the dashboard's effectiveness and aesthetics. The final product can now be used in a better way to both sports fans and professional analysts by providing more dynamic and insightful data visualizations.