Global Stadiums Data Analysis Project Report

Team Members:

Eng. Salah Mohamed
Eng. Omar Nouh
Eng. Kerolos Magdy
Eng. Ahmed Hassan

1. Introduction

1.1 Project Overview

The Global Stadiums Data Analysis Project aims to provide insights into the world's stadiums by leveraging data engineering techniques. Using data scraped from Wikipedia and YouTube, we built a pipeline that processes, cleans, stores, and visualizes this information using Microsoft Azure services and Tableau. The project provides insights into stadium distributions, capacities, and their hosted events.

1.2 Objectives

- Collect data on stadiums from Wikipedia and other sources.
- Clean and standardize the data for analysis.
- Build an automated pipeline using Azure Data Factory, Azure Data Lake, Azure Synapse, and Azure Databricks.
- Visualize the final insights in Tableau.

2. Data Collection

2.1 Data Sources

Wikipedia: We scraped structured information about stadiums from Wikipedia, including:
- Stadium name
- Location (city, country)
- Seating capacity
- Year opened
- Events hosted

2.2 Tools for Web Scraping

We used the following Python libraries:
- BeautifulSoup: For parsing the HTML structure of Wikipedia pages.
- Selenium: To automate browsing and scraping stadium information from YouTube descriptions.
- Requests: For sending HTTP requests to fetch Wikipedia pages.

3. Data Cleaning

3.1 Cleaning Techniques

- Duplicate Removal: Ensured no duplicate stadium entries were present.
- Handling Missing Values: Used the following strategies:
  - For missing seating capacity, we used median values based on location.
  - For missing years, used placeholder values (e.g., 'Unknown').
- Data Standardization:
  - Seating Capacity: Converted to integer values, removing commas or other formatting symbols.
  - Dates: Standardized date formats (e.g., 'MM-DD-YYYY').

3.2 Tools Used

- Pandas: For data cleaning and manipulation.
- NumPy: For handling missing data and numerical operations.

4. Data Pipeline Architecture

4.1 Overview of Data Pipeline

Our project involved building a robust data pipeline using several Azure services to ensure seamless data flow from collection to visualization. The following steps outline the pipeline:
1. Data Collection: Web scraping data from Wikipedia and YouTube using Python scripts.
2. Data Cleaning: Cleaned and standardized the scraped data using Pandas.
3. Data Storage: Raw and cleaned data stored in Azure Data Lake.
4. Data Transformation: Data was processed in Azure Data Factory for ETL (Extract, Transform, Load) operations.
5. Data Processing and Querying: Azure Synapse for large-scale data querying. Azure Databricks for further data analysis and processing.
6. Data Visualization: Insights visualized in Tableau.

4.2 Azure Services Used

- Azure Data Lake: For storing both raw and cleaned data.
- Azure Data Factory: For creating the ETL pipeline.
- Azure Synapse Analytics: For querying and analyzing large datasets.
- Azure Databricks: For advanced data processing and analysis.

5. Data Analysis

5.1 Querying and Analyzing Data

We used Azure Synapse and Azure Databricks to query and analyze the stadium data. Key analyses included:

- Top 10 Largest Stadiums: By seating capacity.
- Distribution by Continent: The number of stadiums in each continent.
- Stadium Age: Analysis of the oldest and newest stadiums worldwide.
- Events Hosted: Stadiums that have hosted the most major international events.

5.2 Results from Analysis

- Largest Stadium: The Rungrado 1st of May Stadium in North Korea, with a seating capacity of 114,000.
- Region with Most Stadiums: Europe, with the highest number of stadiums.
- Stadium Age Insights: Many of the largest stadiums were built in the 20th century, while newer stadiums are mostly in developing regions.

6. Data Visualization in Tableau

6.1 Visualizing Key Insights

We used Tableau to create interactive dashboards that allow users to explore the stadium data. The following visualizations were created:

- Global Map of Stadiums: Stadiums plotted on a world map, color-coded by seating capacity.
- Top 10 Stadiums by Capacity: A bar chart showing the largest stadiums worldwide.
- Distribution by Continent: A pie chart depicting the number of stadiums in each continent.

6.2 Interactive Features

- Filters: Users can filter by continent or seating capacity.
- Tooltips: Each stadium shows additional details when hovered over.

7. Conclusion and Future Work

7.1 Conclusion

The Global Stadiums Data Analysis Project successfully utilized Microsoft Azure's suite of tools to create an automated data pipeline for scraping, cleaning, storing, analyzing, and visualizing stadium data from around the world. The final visualizations in Tableau provide valuable insights into stadium distributions, capacities, and events.

7.2 Future Improvements

- Real-Time Data Updates: Automating the scraping process to include real-time updates from sources like Wikipedia.
- Additional Data Sources: Expanding the dataset to include other types of venues, such as

arenas and concert halls.
- Deeper Analysis: Performing advanced analyses on stadium usage for specific sports and events.

8. Team Contributions

Eng. Omar Nouh: Focused on data scraping and analysis and data cleaning.
Eng. Kerolos Magdy & Eng. Ahmed Hassan: Worked on Azure Data Factory setup , Azure Synapse and Databricks for large-scale querying and processing.

Eng. Salah Mohamed: Led the project, Designed the Tableau dashboards and contributed to data visualization.