رواد مصر الرقمية

وزارة الاتصـــــــالات
وتكنولوجيا المعلومات

Graduation Project Document

Global Stadiums Data Analysis

Team Members:

Eng. Salah Mohamed

Eng. Omar Nouh

Eng. Kerolos Magdy

Eng. Ahmed Hassan

Supervised By:

Eng. Ahmed Essam Azab

Global Stadiums Data Analysis

1. Introduction:

1.1 Project Overview:

The Global Stadiums Data Analysis Project aims to provide a comprehensive understanding-of the world's stadiums by employing data engineering techniques. The project involves building a data pipeline that collects, processes, and visualizes data about stadiums from various sources such as Wikipedia. The resulting insights focus on stadium distributions, capacities, and major events hosted.

1.2 Objectives:

- To collect stadium data from Wikipedia and other online sources.
- Clean and standardize the data for effective analysis.
- Build an automated data pipeline using Microsoft Azure services like Azure Data Factory, Azure Data Lake, Azure Synapse Analytics, and Azure Databricks.
- Visualize the data using interactive Tableau dashboards.

2. Data Collection:

2.1 Data Sources :

Data was collected from:

- Wikipedia: Information such as stadium name, location, seating capacity, year opened, and events hosted was scraped.

2.2 Tools for Web Scraping :

- Beautiful Soup: For parsing HTML content from Wikipedia.
- Selenium: Used for automating web scraping on YouTube.
- Requests: For sending HTTP requests to access webpage content.

3. Data Cleaning:

3.1 Cleaning Techniques :

- Duplicate Removal: Ensured no repeated entries.
- Handling Missing Values: For missing seating capacities, we used median values based on location; placeholder values were used for unknown opening years.
- Data Standardization: Formatting seating capacities as integers and standardizing date formats.

3.2 Tools Used :

- Pandas: For cleaning and manipulating data.
- NumPy: To handle missing values and perform numerical operations.

4. Data Pipeline Architecture:

4.1 Overview :

The project utilized several Azure services to automate the flow of data:
1. Data Collection: Web scraping using Python.
2. Data Cleaning: Data was cleaned and formatted using Pandas.
3. Data Storage: Stored in Azure Data Lake.
4. Data Transformation: ETL processes were implemented with Azure Data Factory.
5. Data Processing and Querying: Used Azure Synapse for querying and Azure Databricks for deeper analysis.
6. Data Visualization: Insights visualized using Tableau.

4.2 Azure Services Used :

- Azure Data Lake: For storing raw and cleaned datasets.
- Azure Data Factory: To implement the ETL pipeline.
- Azure Synapse Analytics: For large-scale data querying.
- Azure Databricks: For advanced data analysis.

4. Data Analysis:

5.1 Querying and Analyzing Data :

We performed various analyses, including:
- Top 10 Largest Stadiums by Capacity.
- Distribution by Continent: Analysis of the number of stadiums in different regions.
- Stadium Age Analysis: Insights into the oldest and newest stadiums.
- Events Hosted: Identified stadiums hosting the most international events.

5.2 Results from Analysis:

- **Largest Stadium:** Rungrado 1st of May Stadium in North Korea (114,000 capacity).
- **Region with Most Stadiums:** Europe.
- **Stadium Age Insights:** Most large stadiums built in the 20th century; newer stadiums are more common in developing areas.

6. Data Visualization in Tableau:

6.1 Visualizing Key Insights:

We created interactive dashboards showcasing:

- **Global Map of Stadiums:** Plotted locations, color-coded by seating capacity.
- **Top 10 Stadiums by Capacity:** Bar chart visualization.
- **Distribution by Continent:** Pie chart representation.

6.2 Interactive Features:

- **Filters:** Users can filter visualizations by continent or capacity.
- **Tooltips:** Display extra information when hovering over specific data points.

7. Conclusion and Future Work:

7.1 Conclusion:

The project successfully demonstrated the use of Azure services for building an end-to-end data pipeline, offering valuable insights into global stadiums.

7.2 Future Improvements:

- **Real-Time Data Updates:** Automate data scraping for real-time updates.
- **Additional Data Sources:** Include arenas and concert halls.
- **Deeper Analysis:** Investigate stadium usage for different sports.

8. Team Contributions

- **Eng. Omar Nouh:** Focused on data scraping, data cleaning, and initial data analysis.
- **Eng. Kerolos Magdy & Eng. Ahmed Hassan:** Worked on setting up Azure Data Factory, configuring Azure Synapse, and utilizing Databricks for large-scale data querying and processing.
- **Eng. Salah Mohamed:** Led the project, designed the Tableau dashboards, and contributed to the data visualization aspects.