|  |
| --- |
| TECHNICAL REPORT |

|  |
| --- |
| **Distributed and Scalable Data Engineering**  **(DSCI-6007)** |

|  |  |
| --- | --- |
| SPRING24 |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 5](#_Toc96341556)

[Discussion 6](#_Toc96341557)

[Conclusion 6](#_Toc96341558)

[Contributions/References 7](#_Toc96341559)

|  |
| --- |
| Tokyo Olympic Data Analysis |

|  |  |  |
| --- | --- | --- |
| Executive Summary This project uses Olympic data as a real-world dataset to demonstrate Azure Cloud's ability to handle end-to-end data engineering activities. We do an extensive study of the Olympic Games' athlete performances, historical patterns, and cultural effects by utilising Azure's capabilities. The project highlights Azure's usefulness in practical data analytics settings by showcasing its seamless integration for data extraction, processing, and visualisation. With this example, we clarify how Azure Cloud can support scalable and effective data processes, improving comprehension of intricate datasets and encouraging the platform's use in various data engineering endeavours. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Harika**  **Sri Charana**  **Amarnath Reddy**  **Venkata Naga Sandeep** | **Questions?**  Contact : @unh.newhaven.edu |  |

|  |
| --- |
| Tokyo Olympic Data Analysis |

|  |  |
| --- | --- |
| **Title of Project:**  **Tokyo Olympic Data Analysis** | Submitted on: 04/23/2024 |
| Highlights of Project  * The project showed how to convert data using Azure Databricks and extract data using Azure Data Factory, resulting in a simplified process that makes managing and analysing Olympic data easier. * The project made use of Azure services to mine the vast Olympic dataset for useful insights. This allowed stakeholders to make well-informed decisions on performance optimisation and strategic planning.  Submitted on: |

## Abstract

This project uses Olympic data as a real-world dataset to demonstrate how Azure Cloud supports end-to-end data engineering activities. We demonstrate smooth data extraction, processing, and visualization by utilizing Azure's capabilities, revealing insights on Olympic athlete performances, historical patterns, and cultural effects. With this example, we showcase Azure's adaptability and effectiveness in managing intricate data processes, providing useful information about its applicability to actual data engineering projects. In the end, this project highlights the potential of Azure Cloud to support data-driven decision-making and develop analytical capabilities across several areas, therefore promoting creativity and well-informed decision-making.

**Introductory Section**

* The Olympics draw athletes and fans from all over the world as a symbol of global sports achievement and cultural relevance.
* Analyzing Olympic data provides priceless information on the accomplishments of individual athletes, historical patterns, and national sports teams.
* By utilizing Azure cloud computing, this research thoroughly examines Olympic data from a variety of sectors, revealing trends and dynamics that influence the event's importance in both culture and sport.

Data Source Link

Through the following link, we have collected the data.

[2021 Olympics in Tokyo (kaggle.com)](https://www.kaggle.com/datasets/arjunprasadsarkhel/2021-olympics-in-tokyo)

.

## Methodology

This technique offers an organized way to organize a project using data mining. It is a strong and tested approach.

Business Understanding:

* Being aware of the significance of using data analytics to support strategic decision-making in the sports sector.
* Realizing that Olympic data analysis may maximize the performance of athletes, the distribution of resources, and strategic initiatives, creating a competitive edge and encouraging creativity.

## Data Understanding:

## The dataset includes information on 11,000 competitors who took part in 47 different Olympic sports in Tokyo in 2021. It contains details for 743 Olympic-represented teams. To make data processing simpler, the data is supplied in CSV format. To facilitate analysis and interpretation, the dataset is divided into discrete groups, such as athletes, coaches, entries, medals, and teams.

## Data Preparation

During this phase, you should understand what data you already have, where to get more, how much data is available, and what tools to use to get it. Knowing your data from the start will make your data science endeavour more coherent.

In our project,

* To get Olympic data, Azure Data Factory is used to collect data from several sources, including APIs.
* Data transformation operations carried out in Azure Databricks using Spark code, including cleaning, aggregation, and enrichment.
* 3. Organizing the data in a way that makes it ready for analysis, visualization, and interpretation using Python libraries or tools like Azure Synapse Analytics.

Modeling and Evaluation

* Model selection: Considering variables like interpretability and forecasting accuracy, pick the best statistical or machine learning models to evaluate Olympic data.
* Model assessment and training: Utilizing past Olympic data, train the chosen models, then assess their effectiveness with measures such as F1 score, accuracy, precision, and recall.
* Iterative refinement: Based on assessment findings, make iterative adjustments to parameters or features in the models to enhance performance and guarantee robustness when forecasting athlete performance or other pertinent outcomes.

**Data Pipeline Architecture**



**Results Section**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

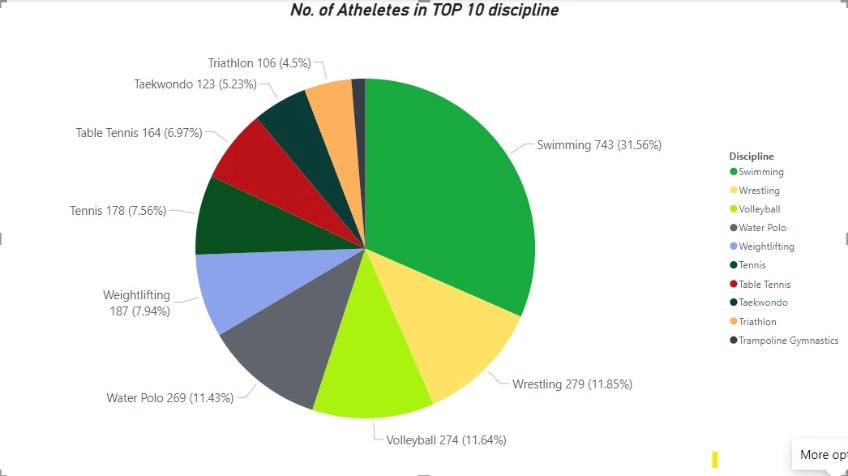
Description automatically generated

Visualization:

To generate visualisations and dashboards based on analytical findings, use Azure Synapse Analytics' visualisation capabilities or external tools such as PowerBI. Charts, graphs, and interactive dashboards are examples of visualisations that offer insights in a clear and understandable format.

A pie chart with numbers and text

Description automatically generated



A graph of blue and yellow bars

Description automatically generated with medium confidence

A graph of a number of people

Description automatically generated

## Conclusion

To sum up, our study has shown how valuable it is to use Azure cloud computing to analyze Olympic data in-depth. Our understanding of athlete performances, time patterns, and national athletic accomplishments has improved as a result of our use of Azure services for data extraction, transformation, and visualization. These revelations might help the sports sector make more strategic decisions and maximize athlete performance. In the future, further research and development into data analytics methods will be necessary to uncover new information and spur creativity in the field of sports administration and performance enhancement.

## Git hub Link

https://github.com/The-Team-12/Tokyo-Olympics-Data-Analysis