8/4/2021

Olympics 2021

ETL Project

By Heejun Seo, Nithiya, Litesh Samji

# Introduction

Our project is based on the 2021 Olympics, compared data is list of all athletes participating in the games, their coaches, and the teams participating across the globe.

This contains the details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics.

## Extraction

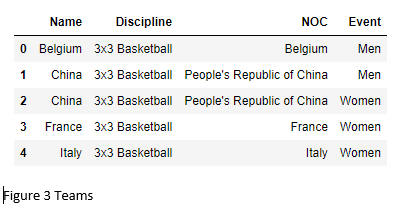
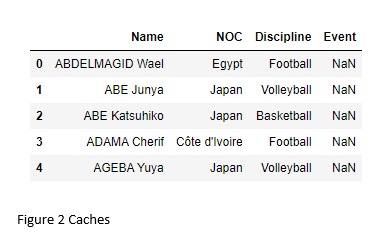
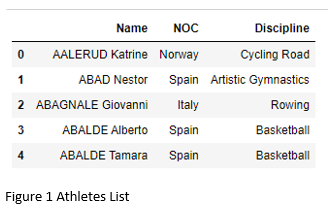
We used 3 datasets from the public Kaggle website.

* Countries
* Discipline
* Athletes
* Coaches

Link: <https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo?select=Coaches.xlsx>

In order to transform the public data and use it in our study we performed the following:

Installed necessary dependencies to read files in xlsx format, and then confirmed the columns from each dataset.



## Transformation

We needed to create 4 tables with the dataset above. For the transformation, data validation was required before hand in order to get the expected results.

The step we took was to create each table is

1. Extracted all countries from Athletes dataset for countries table, dropping duplicate value, set an ID for each country which will be used as a primary key.
2. Extracted all disciplines from Athletes dataset for disciplines table, dropping duplicate value, set an ID for each discipline which will be used as a primary key.
3. Extracted all athletes from Athletes dataset for athletes table, set an ID for each athlete which will be used as a primary key, then merged with countries, disciplines to set id as FK.
4. Extracted all athletes from Athletes dataset for athletes table, set an ID for each athlete which will be used as a primary key, then merged with countries, disciplines to set id as FK.

## Load

We have used a relational database, Postgres SQL to load data and created a table schema for the 4 tables (countries, discipline, athlete, coaches) transformed from the dataset. All data have been inserted to the database, and confirmed by making selection query to each table.