# HW1 Big Data Platform

## Dataset

Our chosen dataset is [World University Rankings](https://www.kaggle.com/datasets/mylesoneill/world-university-rankings/data).  
It gathers University Rankings and their affecting criteria’s from 3 popular rankings over the years. These rankings are:

* [CWUR](https://cwur.org/) (Center of World University Rankings)
* [Times Higher Education World University Rankings](https://www.timeshighereducation.com/world-university-rankings)
* [Shanghai Ranking](https://www.shanghairanking.com/) (The Center of World University Rankings)

## Database Schema

We decided to arrange the data in 3 tables:

1. **“schools\_combined\_worlds\_ranks”**:

Stores the CWUR, Shanghai and Times rankings and total scores of the universities over the years, partitioned by country.

* 1. Partition Column: country’. To filter and group by it. To collect per country stats. And to be able to effectively compare country ‘s schools.
  2. Clustering Columns: ‘year’ – to be able aggregate by it, and limit result to given range of years.

‘university\_name’ – to make the PK unique.

1. **“University\_Criteria”:**

Gathers all the supporting criteria’s (scores, ranks, statistics) the different rankings use to rank a university in a single table. To be able to compare them easily.

* 1. Partition Column: ‘university\_name’. Assuming that this table will be accessed given a university, to support rankings.
  2. Clustering Column: ‘year’. Order by it and make PK unique.