# 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)

Alongside those, the dataset also contains country level data on the following:

1. A series of educational statistics on the population over the years, such as “Percentage of population age 25+ with at least a completed master's degree or equivalent”. These statistics are indicated by Age range, sometimes also by gender.
2. How much countries spend on their education (by the sector type (private/public)) over the years, as a percentage of their expenses.

## Database Schema

We decided to arrange the data in 3 tables:

1. **“University\_World\_Ranking”**:

Stores the CWUR, Shanghai and Times rankings of the universities over the years (along with Shanghai and Times total score, that are used to determine the rank)

* 1. Partition Column: Year. Since most of the insights will be based on comparing the rankings over the years.
  2. Clustering Columns: country (to have the data pre ordered by it) and ‘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.

1. **“Country\_Statistics”:**

Unites the Expenditure and Attainment supplementary data into a single generic table. The idea is that metric is determined by ‘series\_name’ in attainment, and by the composite of ‘institute\_tpye’ with ‘direct\_expenditure’ in expenditure to represent a global metric. And the measure to be the measure specified at the corresponding year.

* 1. Partition Column: ‘country’ + ‘year’ assuming that we will access this table given those indicators.
  2. Clustering Columns: metric. To make PK unique.