11756 Storage and Data Recovery
Master's Degree in Intelligent Systems
Final Project – option B

By means of web scraping techniques, a database has been downloaded containing data on passengers and crew who sailed in the Americas, mainly during the 19th century. This database, **without making any changes to its content except for the format** (which has been converted to JSON format in the file *CrewJSON.json*), contains the following information fields for each passenger or crew member:

- **FirstName** and **LastName**: First name and last name.
- **Vessel**: Name of the ship on which the person sailed.
- **Rig**: Type of vessel.
- **ApproximateDeparture**: Approximate date of the ship's departure
- Personal characteristics: **eye** colour, **skin** colour, **hair** colour, **age**, **height**, and place of origin (**residence**).
- **Rank**: In the case of crew members, the position or role held by the person. For passengers, this information does not appear (which allows passengers to be distinguished from crew).

Example:

{"LastName":"Chace",

"FirstName":"George A.",

"Vessel":"Edward Everett",

"Rig":"Bark",

"ApproximateDeparture":"11/3/1869",

"Skin":"Light",

"Hair":"Brown",

"Residence":"Acushnet"}

Using MapReduce techniques, one of the following studies (only one) must be carried out:

- **Vessel study**: Analysis of voyages made, considering ships, vessel types, number of trips, total number of passengers, and total crew. An OLAP cube should be obtained linking vessel and rig with total passengers and crew.
- **Crew study** (for cases where Rank is specified): Analysis of crew characteristics, relating height, age, and occupation. An OLAP cube should be obtained linking rank, height, and age.
- **Anthropomorphic study of the people who sailed**: Analysis of the physical characteristics of the people who sailed, relating eye color, hair color, and skin color with place of origin. An OLAP cube should be obtained linking these parameters.

- **Study** of migratory movements and occupations: Relation of the occupations carried out on each ship with places of origin, in order to determine whether there is any relationship between vessel, rank, and place of origin. This analysis should also be carried out for passengers on the vessels.

In addition to the OLAP cube, charts must be produced to allow comparison of the results.

The Final Project consists of:

- Choosing one of the studies to be carried out
- Obtaining results from the analysis of the input data
- Writing a document of approximately 10–15 pages with the following structure: brief historical context, description of the work carried out, explanation of the methodology used (MapReduce), explanation of the practical application of MapReduce, results obtained (with the corresponding charts), and conclusions.

The documentation to be uploaded consists of:

- The programs used (source code)
- The aforementioned PDF document
- Database structure