NEM Data Challenge

1st phase guidelines

The participant will receive two files. The first one is a "csv" with 5 columns ("asset", "date", "time", "variable" and "value"). The file contains the speed, velocity, temperature and pressure sensor measurements associated to the different wind turbines with a high measurement frequency (a measurement per second).

- The column "asset" contains the asset's identification to which the measurement is associated.
- The date and time of each measurement is registered in two columns. The "date" one contains the date in the format "yyyy-mm-dd", whereas the "time" one contains the time in the format "hh:mm:ss"
- The "variable" column registers the sensor's identification of each asset and the value of the measurement obtained by the mentioned sensor is contained in the column "value"

Each sensor's identifier is registered in a different way in each asset. The second file is an "xml" which contains the unification of each sensor's identifiers among the different assets.

The participant will be asked to process that file and as a result, generate a new "csv" file. The file will have to contain a "timestamp" column with the date and time of each register in unix and all the values of each of the registered sensors in different columns.

The heading of each column will have to be the sensor's unified identifier. The file will contain a row every ten minutes and the average of the registered measurements in those 10 minutes in each one of the cells. The timestamp associated to each interval will have to be the superior limit of each interval.

Once the data is processed, the participant will have to make a descriptive analysis of the quality of the data, together with distributions or relations among the different variables. Once the analysis is carried out, the participant will have to generate a report of no more than 5 pages where he/she will have to register in a visual way that information the participant considers most relevant.

The outcome of this test will be the source code used for the transformation of the files, the resultant "csv" file and a PDF with the report including the visual description of the information contained in the latter mentioned files. The objective of this test is to assess the participant's capacity in the areas of data processing (ETL processes), data analysis (statistics and data mining) and information visualisation (visual analytics).