Systems for Big Data Processing 2019/20 – 1st semester
Project nº 1

You and your colleague were contracted by a non-profit-organization to extract some information on natural disasters from their data set and report on it. The data set is available in two versions:

1. https://www.dropbox.com/s/lgmoludarl4k4hm/natural_disasters_SMALL.csv?dl=0
   `(md5sum = fcab54defa153b6599958be4bb23eaa3)`
   a (smaller) sample of the data set, ideal for code development'; and
2. https://www.dropbox.com/s/b74bpkbr8918k43/natural_disasters_BIG.csv?dl=0
   `(md5sum = 6f4e1fd4371cf6160c59b9bc876ab73c)`
   a bigger one, which should be used in the report.

The data sets have the following columns:

Column #	Name	Type
0	year	integer
1	disaster_group	string
2	disaster_subgroup	string
3	disaster_type	string
4	disaster_subtype	string
5	continent	string
6	region	string
7	iso	string
8	country_name	string
9	occurrence	integer
10	Total deaths	integer
11	Injured	integer
12	Affected	integer
13	Homeless	integer
14	Total affected	integer
15	Total damage (K$)	integer

You are being asked to create three indexes for answering the following three queries:

1. *How many disasters occurred in continent C? (columns 5 and 9)*
2. *In which regions there were disasters of type X? (columns 3 and 6)*
3. *What are the probabilities of getting injured or dying in a natural disaster of type T in the continent C during decade D (190x, 191x, 192x, …, 199x, 200x, 201x)? (columns 0, 3, 5, 10, 11, and 14)*
4. An **optional 4th index** that will answer a non-trivial and interesting question over the given data set. If you add this to your project you will get some extra points.

In this setting, an index is a file with of pairs "key, value" that one could scan to quickly find the answer to the given question.

To succeed in this project, you must create the requested indexes using three different technologies (over Apache Hadoop):

1. Create a Map-Reduce program to create the three indexes;
2. Create a Spark program to create the three indexes;
3. Create a Spark DataFrame or SparkSQL program to create the three indexes.

To measure the time it takes to execute a python/hadoop program, do the following:

- In the command line, just prefix the command line with `time`, e.g.,
  ```
  $  time “hadoop jar …”
  $  time “python filename.py”
  ```
- In a Jupyter notebook, put
  ```
  %%time
  ```
  in the first line of your code box, and Jupyter will report the time that box took to execute.

Time will output 3 different values. Please report two values: *real* and *user+sys*. For more information see:
https://stackoverflow.com/questions/556405/what-do-real-user-and-sys-mean-in-the-output-of-time1

The project team has two members and must deliver:

1. A set of three Jupyter notebooks (one notebook per technology), which must include the code to be executed in the docker container used in the labs.
2. A report with a maximum (strict) of two-pages maximum[1] that describes briefly your rational for creating the required indexes, and the time it takes to process the complete (bigger) dataset in your computer, for each of the requested indexes in each of the required technologies.

The 4 project files (3 Jupyter notebooks and the report in PDF) shall be uploaded as a single ZIP file with the name **`Gnn_AAAAA_BBBBB.zip`** where:

Gnn -> group number, e.g., G04

AAAAA -> student 1 number, e.g., 45454

BBBBB -> student 2 number, e.g., 54321

(the numbers AAAAA and BBBBB must be in increasing order),

using the form at the address:

https://forms.gle/SkGeDJcQokcx3nBX6

no later than Sunday, November 17, 2019 @ 23:59.

[1] Please use the IEEE template available at: https://www.ieee.org/conferences/publishing/templates.html