Lab 2: Spark RDD

Note: Save your answers and all the steps performed in a document for review (consultation).

- Launch the three containers from Lab 1.

Configuring Spark:

- Create the **slaves** configuration file in the directory **/usr/local/spark/conf**.
- Add the names of the worker containers to the **slaves** file.
- Start the Spark services on all nodes.

Spark RDD:

Write four Python programs to:

(1) Create an RDD from the file **arbres.csv** and display the number of lines in the created RDD.
(2) Calculate and display the average height of the trees.
(3) Display the genus of the tallest tree: The principle is to construct key-value pairs, where the tree heights serve as the key and their genus as the value. Then, sort the pairs in descending order by key using **sortByKey** and keep only the first pair.
(4) Display the number of trees for each genus: The principle is to construct a pair **(genus, 1)** for each tree in the file, then aggregate the values by genus using **reduceByKey**.

Execute these programs in the Spark cluster in two different ways: (1) Without using HDFS (i.e., by using the local file system), (2) Using HDFS.