**BIG DATA COMPUTING 2021/22 - HOMEWORK 3**

**PYTHON VERSION**

Run your algorithm on the cluster on CloudVeneto using the following datasets: **HIGGS-REDUCED-7D.txt** (about 1.2M points in 7 dimensions),and **artificial9000.txt** (9200 points in 2 dimensions).The datasets are in the **directory /data/BDC2122** of the HDFS. You must fill the two tables below, one for each dataset, where the headers of the rows indicate the values to report, and the headers of the columns indicate the configurations of parameters to be used.

The first table collects results aimed at assessing the **scalability** of the algorithm.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **HIGGS-REDUCED-7D.txt** | **2 executors**  **k=10, z=150, L=2** | **4 executors**  **k=10, z=150, L=4** | **8 executors**  **k=10, z=150, L=8** | **16 executors**  **k=10, z=150, L=16** |
| **Time to read input from file (in ms)** | 12567 | 7454 | 6112 | 5978 |
| **Time of ROUND 1 (in ms)** | 394280 | 188890 | 95240 | 50162 |
| **Time of ROUND 2 (in ms)** | 601 | 2458 | 9075 | 37983 |
| **Time to compute objective function (in ms)** | 8705 | 4223 | 2426 | 1233 |
| **Value of objective function** | 8.91507526715 | 7.63266544061 | 6.45591354044 | 5.93477711865 |

The second table collects results aimed at comparing the **accuracy** attained by the algorithm against the one attained by the sequential algorithm from Homework 2 on the entire dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Artificial9000.txt** | **2 executors**  **k=9, z=200, L=2** | **4 executors**  **k=9, z=200, L=4** | **8 executors**  **k=9, z=200, L=8** | **16 executors**  **k=9, z=200, L=16** | **Sequential algorithm from Homework 2 with k=9, z=200** |
| **Value of objective function** | 13.0744587651 | 12.9032108407 | 11.7110452138 | 11.7121112102 | 11.576939707884812 |

Provide below a brief comment to justify the scalability and accuracy observed (your answer should be of at most 6 lines, font 12 points):

By increasing the number of partitions (we consider more point from our original dataset) and executors my objective function starts to reduce (gets closer and closer to the objective function of the second homework if we consider the Artifical9000 datasets): that because I’m considering more points to be processed : this fact increments the time of ROUND 2 but at the end we will conclude with a more accurate objective function. Last thing we can notice: the closer we are to the best approximation, the slower our function improves.