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

**JAVA VERSION**

Run your algorithm on the cluster on CloudVeneto using the following datasets: **HIGGS10M7D.txt**(about 10M 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 tablesbelow, 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **HIGGS10M7D.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)** | 47502 | 30284 | 19157 | 12214 |
| **Time of ROUND 1 (in ms)** | 39794 | 21306 | 11387 | 5990 |
| **Time of ROUND 2 (in ms)** | 60 | 171 | 716 | 2066 |
| **Time to compute objective function (in ms)** | 2491 | 1474 | 879 | 652 |
| **Value of objective function** | 12.212318096918224 | 10.49952379948082 | 9.188191484951549 | 8.285848589407689 |

The second table collects results aimed at comparing the **accuracy** attained by the algorithm against the one attained by the sequential algorithm from Homework 2on 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 and z=200** |
| **Value of objective function** | 12.829787371581807 | 12.897592410988958 | 11.662186973291075 | 11.570155616931 | 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):

What we noticed with the increasing number of executors is that :

* Doubling the executors number the Time of round 1 halved, Time of round 2 ,instead, increases with an exponential behavior, also the time to read the file and the time to compute objective function will decrease drastically
* The value of the objective function will tend to the value obtained with the sequential algorithm

Notice that, the efficiency between 8 and 16 executors increases slightly, and this does not justify the increase of executors because the improvement is not that much.