**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)** | 10915 | 7419 | 5387 | 5116 |
| **Time of ROUND 1 (in ms)** | 23152 | 11507 | 5959 | 3139 |
| **Time of ROUND 2 (in ms)** | 24 | 51 | 155 | 465 |
| **Time to compute objective function (in ms)** | 9063 | 4527 | 2510 | 1286 |
| **Value of objective function** | 9.32264109344 | 7.56472205987 | 6.33673072798 | 6.01539567448 |

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** | 12.7722883228 | 12.0586954933 | 11.9514118413 | 11.3488087921 | 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):

As shown in the table, the time to read the input from file is reduced as the number of executors increases, since the partitions of the inputs are read in parallel. The time to execute the first Round also is reduced because the partitions have to extract and process less points each. On the other hand, the time of the second Round slightly increases because the coreset is made by more partitions so the SeqWeightedOutliers takes more time to complete. The time to compute the objective function is instead reduced because we parallelized the execution by partitioning it with a mapPartition function. The value of the objective function is then reduced as well because SeqWeightedOutliers can work with a bigger coreset, which allows to compute better centers for calculation of the objective function. However, on the artificial9000 dataset, the objective function computed without partition has similar result to when is computed with 8 or 16 partition. This is probably because of the alpha=0, as the alpha=2 used when the algorithm is run on a coreset is needed to address the approximation introduced by the coreset.