**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)** | 10946.742057800293 | 7929.847478866577 | 6677.440643310547 | 5660.010814666748 |
| **Time of ROUND 1 (in ms)** | 380570.1034069061 | 188866.71376228333 | 115147.89366722107 | 48950.02245903015 |
| **Time of ROUND 2 (in ms)** | 299.7448444366455 | 1065.901756286621 | 4427.877187728882 | 16638.9377117157 |
| **Time to compute objective function (in ms)** | 43514.61958885193 | 22196.165084838867 | 13318.256139755249 | 6446.886777877808 |
| **Value of objective function** | 8.645875557885208 | 7.715136959555597 | 6.4367787997870005 | 5.985465377224249 |

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** | 12.909866304497502 | 12.693431253999055 | 11.918118307853804 | 11.450581862944786 | 11.11.576939707884812 |

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

In the first table, the FFT algorithm has an advantage when the number of executors and partitions increases from 2 to 16, while the sequential Kcenteroutliers algorithm has no advantage of parallel workers. As a result, round 1 takes less time and round 2 takes more time as the size of the core set increases with the number of partitions(L). The computation of the objective function is also parallelized and therefore takes less time.

In the second table, as the number of partitions increases, the final coreset size (k+z+1)\*L becomes larger and the value of the objective function is close to the value in the last column calculated for the entire point set (P).