**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)** | 10958.423 | 7344.643 | 6300.338 | 5174.194 |
| **Time of ROUND 1 (in ms)** | 6683.121 | 3116.680 | 1503.068 | 866.346 |
| **Time of ROUND 2 (in ms)** | 35.400 | 97.563 | 271.466 | 897.998 |
| **Time to compute objective function (in ms)** | 1489.037 | 836.613 | 501.024 | 336.021 |
| **Value of objective function** | 8.85256289329 | 7.64630185804 | 6.43736832626 | 5.90730148351 |

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** | 12.7859274204103 | 12.712556037241294 | 12.258389168239031 | 11.649088419271267 | 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):

Round 2 times increase with the number of executors as each FFT in round 1 returns K points, so more executors means more points to process. At the same time, a greater number of points will also give better estimates, which greatly improves accuracy. On the other hand, input file processing, round 1 and objective function computation times all decrease (round 1 essentially inversely proportional to the number of executors, the others less so) as more executors are added, as they simply benefit from the same number of elements being split into more partitions.