**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 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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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)** | **47505** | **31081** | **18790** | **11466** |
| **Time of ROUND 1 (in ms)** | **40967** | **22250** | **12305** | **5876** |
| **Time of ROUND 2 (in ms)** | **49** | **99** | **253** | **694** |
| **Time to compute objective function (in ms)** | **1563** | **825** | **1383** | **546** |
| **Value of objective function** | **11.48042** | **10.44580** | **9.28175** | **8.61666** |

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 and z=200** |
| **Value of objective function** | **12.71808** | **13.26079** | **11.82715** | **11.45607** | **11.57693** |

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 the number of partitions increases, the execution time of Round 2 increases and the value of the objective function improves. Also, by increasing the number of partitions of Artificial9000.txt, the objective function gets closer to the values obtained in HW2. This is because, for a greater number of partitions, the number of points over which SeqWeightedOutliers is computed increases. With the corset-based approach, Round 1 complexity is dominant (while still being asymptotically lower than SeqWeightedOutliers with small k and z), while Round 2 has much less impact on performance.