**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)** | 46533 | 29863 | 18722 | 11960 |
| **Time of ROUND 1 (in ms)** | 41761 | 21224 | 11605 | 6509 |
| **Time of ROUND 2 (in ms)** | 78 | 111 | 184 | 476 |
| **Time to compute objective function (in ms)** | 3075 | 1459 | 1469 | 833 |
| **Value of objective function** | 11.4804217255222 | 10.14475642512223 | 9.196948767225818 | 8.547263907602714 |

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** | 12.829787371581807 | 12.678071777679758 | 11.77398564633064 | 11.398787698698488 | 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 expected, we see that an overall improvement in performance is obtained as the number of executors increases. We can see how the time for reading the file, running ROUND 1 and compute the objective function decreases although the time for ROUND 2 increases. We can notice that as the number of partitions and consequently the number of points taken into consideration in the calculation of the centers increases, the value of the objective function goes close to the result obtained by executing the algorithm sequentially.