**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)** | 10856.874 | 7437.933 | 5633.421 | 5217.565 |
| **Time of ROUND 1 (in ms)** | 398706.540 | 186833.512 | 96334.149 | 49221.668 |
| **Time of ROUND 2 (in ms)** | 260.933 | 865.168 | 3405.266 | 13959.338 |
| **Time to compute objective function (in ms)** | 48126.496 | 23450.886 | 12822.713 | 6939.570 |
| **Value of objective function** | 9.209 | 7.837 | 6.412 | 6.190 |

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** | 13.033 | 12.910 | 12.340 | 11.592 | 11.576 |

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

Increasing the number of partitions L and the number of executors causes a decrease in the time to read the input, the time of round 1 and to compute the objective function. This is because the larger L, the smaller portion of data each executor works on in parallel. The time of round 2 increases since it is a sequential operation performed on a number of points that increases with L (the size of the merged coresets is in fact (k+z+1)\*L). The accuracy of our objective function improves, since with larger L more input points are extracted by the FFT algorithm, thus leading to more precises distances computations.

Additionally, we tested our algorithm on HIGGS-10M-7D.txtdataset, and we report our results in the table below.

|  |  |
| --- | --- |
| **HIGGS-10M-7D.txt** | **16 executors**  **k=10, z=150, L=16** |
| **Time to read input from file (in ms)** | 13247.522 |
| **Time of ROUND 1 (in ms)** | 429069.978 |
| **Time of ROUND 2 (in ms)** | 14514.979 |
| **Time to compute objective function (in ms)** | 55118.165 |
| **Value of objective function** | 8.486 |