**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.

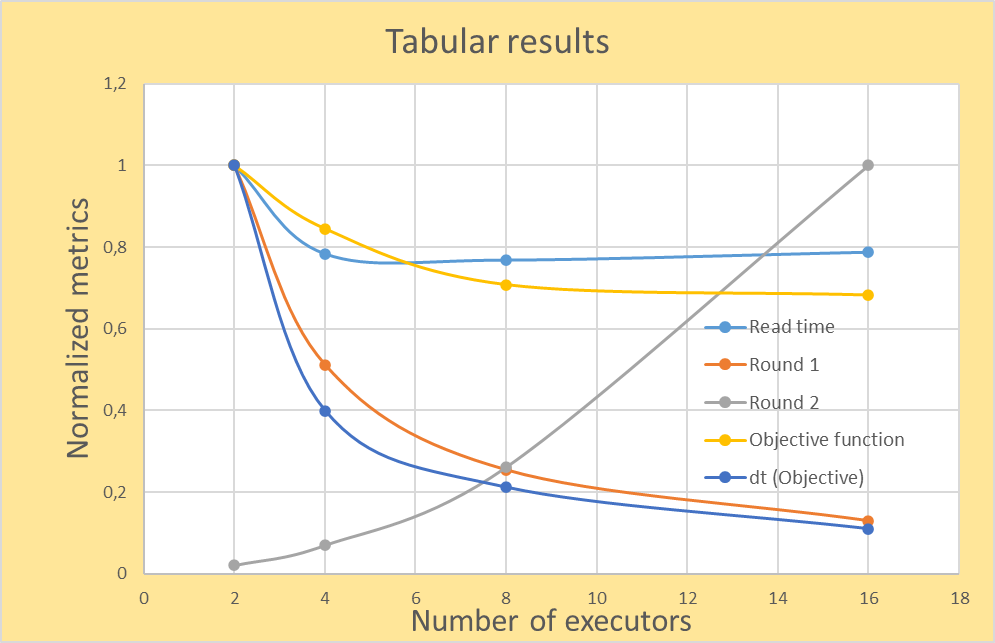
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **HIGGS-REDUCED-7D.txt** | **2 executors**  **k=10, z=150, L=2** | **4 executors**  **k=10, z=150, L=4** | **8 executors**  **k=10, z=150, L=8** | **16 executors**  **k=10, z=150, L=16** |
| **Time to read input from file (in ms)** | 842.5472 | 659.8680 | 646.9817 | 663.3165 |
| **Time of ROUND 1 (in ms)** | 378297.3602 | 193474.4728 | 96074.9838 | 48899.4610 |
| **Time of ROUND 2 (in ms)** | 453.9123 | 1533.2470 | 5703.5975 | 21799.4530 |
| **Time to compute objective function (in ms)** | 38368.1502 | 15310.7643 | 8111.31453 | 4200.3787 |
| **Value of objective function** | 9.0551 | 7.6568 | 6.4120 | 6.1867 |

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** | 12.9865 | 12.5855 | 12.2959 | 11.8318 | 15.36 |

**(Please, refer to summary on the second page)**

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



* + **Reading time** and **objective function –** reach plateau for >8 executors.
  + Exponential decrease of **calculating objective** and **round 1** (for8 executors is **70%** less than for 2 executors).
  + **Rapid** **raise in Round 2 time**, because more points are passed to round 2.
  + Overall conclusion – **problem is scalable for Round 1 and computation of objective. Others metrics reach plateau or grow**, 🡪 need optimization.
  + **8 executors** are be **optimal** for this task.