**BIG DATA COMPUTING 2018-19 – HOMEWORK 4 – GROUP 13**

1. **Required tests.** Do the tests with the parameters indicated in the following table and, for every test, report the following values: Ti = running time (in seconds) of Round i of MR\_kmedian (i=1,2,3), and Obj = value of the objective function (average distance of points from centers). If you notice anomalies in the values of a test try to repeat the test 2-3 times and take the average values*. Some anomalies are to be expected!*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **K** | **L** | **iter** | **num-executors** | **T1** | **T2** | **T3** | **Obj** |
| **HIGGS11M7D** | **10** | **16** | **0** | **16** | **1773** | **10** | **680** | **0,603550** |
| **HIGGS11M7D** | **25** | **16** | **0** | **16** | **3191** | **16** | **735** | **0,491572** |
| **HIGGS11M7D** | **50** | **16** | **0** | **16** | **5233** | **31** | **1048** | **0,425698** |
| **HIGGS11M7D** | **50** | **16** | **10** | **16** | **20959** | **82** | **1083** | **0,361068** |
| **HIGGS22M7D** | **50** | **32** | **10** | **8** | **71319** | **133** | **6287** | **0,432347** |
| **HIGGS22M7D** | **50** | **32** | **10** | **16** | **51228** | **134** | **6715** | **0,431728** |
| **HIGGS22M7D** | **50** | **32** | **10** | **32** | **27377** | **162** | **1365** | **0,431336** |

1. **Other observations** (at your discretion)

From the tests we can see that the biggest improvement of the objective function is provided by the application of Lloyd’s algorithm. This comes at the cost of much longer computation times, especially during Round 1 (even though Round 2’s computation time doubles, it stays negligible overall).

In the last three tests we can see that the algorithm scales really well with the number of executors in Round 1. Round 2 takes always the same time to complete, because it is a sequential operation and the parameters are the same. Round 3’s times are a bit anomalous, probably due to other tasks being performed on the cloud.

Because of this, we had to repeat the tests multiple times in order to get some meaningful values and some of them are still too high for the algorithm.