**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** | **2072** | **10** | **745** | **0,608575341267478** |
| **HIGGS11M7D** | **25** | **16** | **0** | **16** | **3191** | **16** | **735** | **0,491572137758809** |
| **HIGGS11M7D** | **50** | **16** | **0** | **16** | **5407** | **31** | **1125** | **0,422305425723374** |
| **HIGGS11M7D** | **50** | **16** | **10** | **16** | **20959** | **82** | **1083** | **0,361068175556586** |
| **HIGGS22M7D** | **50** | **32** | **10** | **8** | **71319** | **133** | **6287** | **0,432347917857593** |
| **HIGGS22M7D** | **50** | **32** | **10** | **16** | **54148** | **118** | **7972** | **0,432862920160947** |
| **HIGGS22M7D** | **50** | **32** | **10** | **32** | **40935** | **131** | **5705** | **0,431200437363617** |

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).

We can see that the algorithm scales really well with the number of executors.