**BIG DATA COMPUTING 2018-19 – HOMEWORK 4 – GROUP 06**

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** | **7.048** | **0.345** | **1.308** | **0.61067187** |
| **HIGGS11M7D** | **25** | **16** | **0** | **16** | **12.825** | **0.187** | **1.762** | **0.49854682** |
| **HIGGS11M7D** | **50** | **16** | **0** | **16** | **22.573** | **0.350** | **1.872** | **0.41905077** |
| **HIGGS11M7D** | **50** | **16** | **10** | **16** | **42.437** | **0.379** | **1.826** | **0.36116315** |
| **HIGGS22M7D** | **50** | **32** | **10** | **8** | **85.613** | **0.523** | **6.593** | **0.43252664** |
| **HIGGS22M7D** | **50** | **32** | **10** | **16** | **63.689** | **0.372** | **6.851** | **0.43280421** |
| **HIGGS22M7D** | **50** | **32** | **10** | **32** | **52.302** | **0.375** | **3.405** | **0.43302554** |

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

**HIGGS11M7D**

Smaller set of points. The number of executors is fixed for every kind of run. It’s interesting to observe, in our opinion, the comparisons between the first three runs and the third with the fourth run.

1. The variation is the number of clusters. We observe that the bigger K is, the more the time to compute the clusters. Moreover, the average distance of points from centers decreases with the increasing of K. This is absolutely expected since with more clusters there are more centers and every point is associated to the nearest center.
2. By making some iterations with Lloyd's algorithm, the centers are refined. We see that the time to compute the first round is bigger (because of the refinement of the centers) but in this way the average distance from the centers is smaller (better clustering). So, in conclusion, with just 10 iterations of Lloyd's algorithm there are good results, even though the time to compute round 1 doubles.

**HIGGS22M7D**

Bigger set of points: K, L and iter are the same. The objective function value converges to the same value for every type of run. The change is about the time to complete the runs and decreases as the number of executors increases. This is expected, in particular for the first round which is where most of the operations are done.