**BIG DATA COMPUTING 2020/21 – HOMEWORK 3**

**Finding the optimal silhouette value**

Test your algorithm on the **synt2M.txt.gz** dataset for finding the number k of clusters that provide the best silhouette value. Search k in the range [8,12]. Report the silhouette value and the running times when running your algorithm with kstart equals to the optimal k and with M=2000, iter=10, L=16 and 16 executors.

**K = 10**

**Silhouette = 0.9989081**

**Time to read the input (in ms) = 630**

**Time to compute clustering (in ms) = 5780**

**Time to compute the silhouette (in ms) = 28107**

**Note for Python users:** if the total running time of the required run exceeds 15 minutes, reduce the value of M until the running time is below 15 minutes.

**Analyzing algorithm scalability**

Analyze the scalability of your algorithm on the **HIGGS11M7D.txt.gz** dataset with 2, 4, 8 or 16 executors. Run your algorithm with k=5, h=1, iter=10, L=16, and M=500 (Python users: use M=50), and fill the table below with the required values.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **2 executors** | **4 executors** | **8 executors** | **16 executors** |
| **Time to read input (in ms)** | 611 | 660 | 622 | 643 |
| **Time to compute clustering (in ms)** | 96838 | 89630 | 64784 | 51390 |
| **Time to compute the silhouette (in ms)** | 307581 | 130671 | 72051 | 33788 |

The read input times are not dependent on the number of executors, clustering times are sub-linearly dependent on the number of executors and silhouette computation times are linearly dependent on the number of executors.