Homework 4 **Group 18**

Department of Information Engineering
Big Data Computing
Prof. Andrea Pietracaprina
Due Date: 31/05/2018

Aladdine Ayadi
Giovanni Barbieri
Alessandro Pelizzo
Davide Talon

Test results

In the present report the times needed for the construction of the coreset, the running time of the sequential algorithm and the load time of different datasets are analyzed. Let N_{exec} be the number of cores per executor for a total number of N_{tot} cores used. The measurements evaluate the time t_c which denotes the time for the construction of the coreset, t_s represents the time for the run of sequential algorithm and finally t_l the time needed for the loading and count of points in the dataset. Figure 1 shows that the heaviest operations of the algorithm are the loading of the dataset and the construction of the coreset, while the time needed by the sequential algorithm for the construction of the final centers is proportional to the number of partitions, this is probably due to the cost of communication among different partitions. Incresing the number of centers greater times are noticed in terms of t_c and t_s while the time for the loading of the dataset is steady. As expected the average distance among points deacreases with the number of centers, as shown in Figure 2. Different times for the same simulation settings lead us to conjecture that the load of the cluster influences the performance obtained.

N_{tot}	N_{exec}	$t_c[ms]$	$t_s[ms]$	$t_l[ms]$
4	2	14785	165	169256
8	2	7262	194	92361
8	4	12393	174	114576
8	8	23183	324	229915
16	4	22651	229	147105
16	8	22902	212	215455
32	4	14832	256	257419
32	8	24426	516	227701
64	4	19192	299	148242
64	8	22688	188	206375

Table 1: Results with dataset *all*, $P = 32$, $k = 20$ and increasing cores.

P	$t_c[ms]$	$t_s[ms]$	$t_l[ms]$
4	12402	21	346732
8	7808	36	42784
16	4885	80	40415
32	3037	195	33901
64	9715	607	44835
128	4139	2050	43149
256	5222	8189	38288

Table 2: Results with dataset *all*, $N_{tot} = 32$, $N_{exec} = 4$, $k = 20$ and increasing P.

Dataset	$t_c[ms]$	$t_s[ms]$	$t_l[ms]$
vectors-50-500000	1848	170	26885
vectors-50-1000000	2190	193	31254
vectors-50-2000000	5325	179	73585
vectors-50-3000000	2486	232	40012
vectors-50-all	3037	195	33901

Table 3: Results with $N_{tot} = 32$, $N_{exec} = 4$, $k = 20$, constant number of partitions P and increasing size of the dataset.

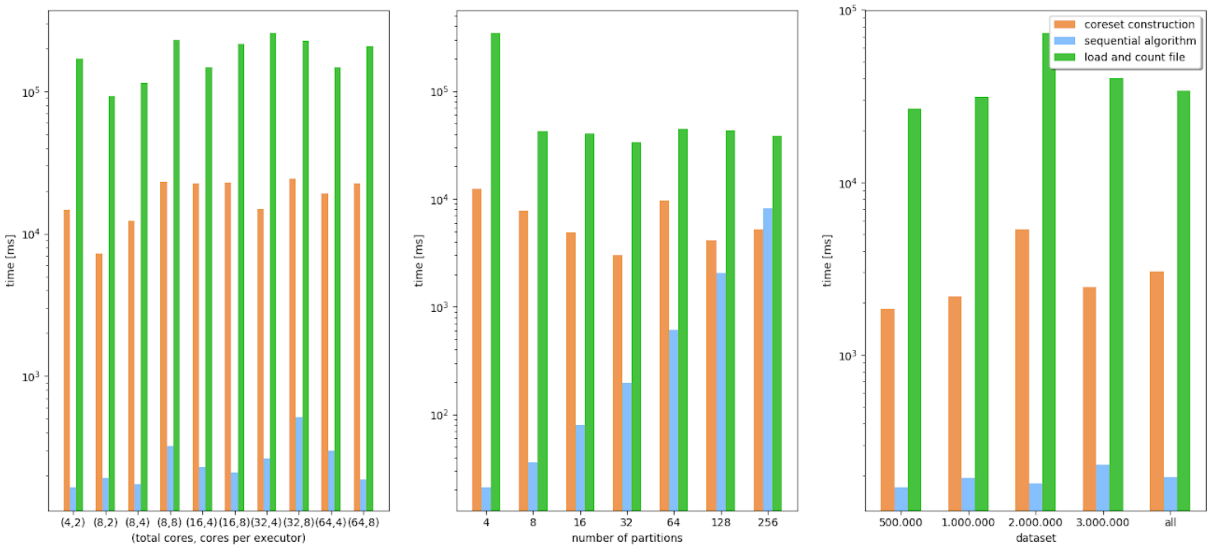

Figure 1: Results of the different simulations respectively with increasing cores, partitions and dataset size, times are represented in a logaritmic scale.

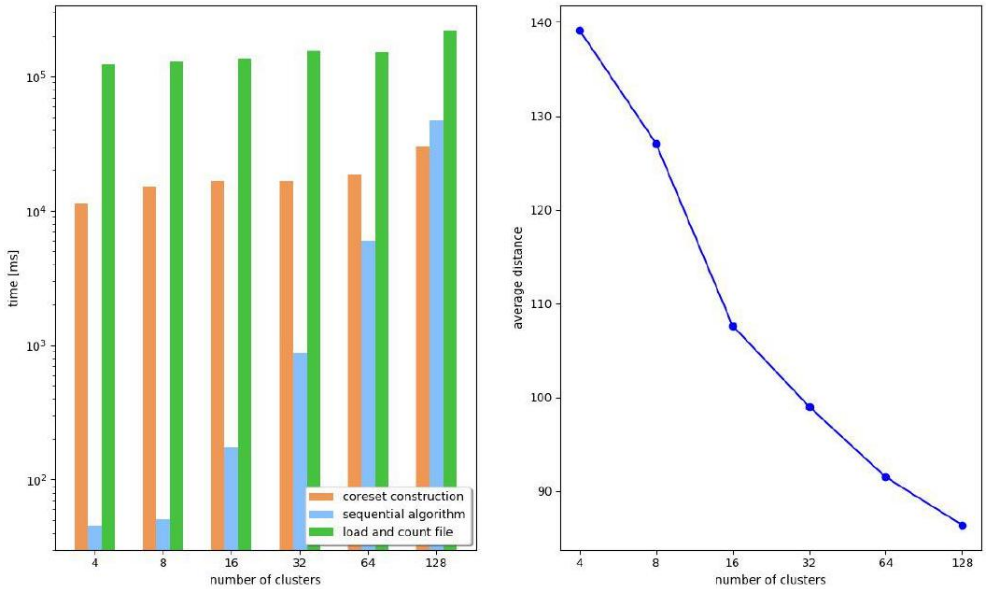

Figure 2: Results for the simulations with increasing number of centers with $N_{tot} = 32$, $N_{exec} = 4$ and $P = 32$. The image on the left represents the performances in terms of times expressed in a logaritmic scale, while the one on the right shows the average distance, i.e.the sum of centers pairwise distances divided their number.