* 1. 小dataset k=3 不分離群值:

Bulkload time: 2.83701e+06

Total lookup time: 1.96949e+06 nanoseconds

Bulkload time: 3.59767e+06

Total lookup time: 1.74815e+06 nanoseconds

Bulkload time: 2.26868e+06

Total lookup time: 1.7855e+06 nanoseconds

* 1. 小dataset k=3 移除5%離群值:

Bulkload time: 8.88093e+06

Total lookup time: 1.60939e+06 nanoseconds

Bulkload time: 4.30147e+06

Total lookup time: 1.8539e+06 nanoseconds

Bulkload time: 6.28009e+06

Total lookup time: 1.6392e+06 nanoseconds

Center: 固定在16xxxx、50xxxx、83xxxx 1/6、1/2、5/6

* 1. 小dataset sort後硬切3等分

Bulkload time: 2.04083e+06

Total lookup time: 2.25626e+06 nanoseconds

Bulkload time: 1.94355e+06

Total lookup time: 2.15779e+06 nanoseconds

Bulkload time: 2.06621e+06

Total lookup time: 2.04582e+06 nanoseconds

* 1. 小dataset直接塞

Bulkload time: 1.70331e+07

Total lookup time: 3.04203e+06 nanoseconds

Bulkload time: 1.66395e+07

Total lookup time: 3.55097e+06 nanoseconds

Bulkload time: 2.23661e+07

Total lookup time: 2.97749e+06 nanoseconds

現有的前中後

1. 80% 暴力解應該做的不好　也做kmean
2. 30/40%的暴力解 也做kmean

2-1 k=3 有移除離群值

80%

Bulkload time: 1.30256e+10

Total lookup time: 4.03619e+09 nanoseconds

Bulkload time: 1.40537e+10

Total lookup time: 4.48975e+09 nanoseconds

Bulkload time: 1.41226e+10

Total lookup time: 4.34047e+09 nanoseconds

30%

Bulkload time: 1.44759e+10

Total lookup time: 5.88854e+09 nanoseconds

Bulkload time: 1.3367e+10

Total lookup time: 5.52157e+09 nanoseconds

Bulkload time: 1.48113e+10

Total lookup time: 5.804e+09 nanoseconds

40%

Bulkload time: 1.56288e+10

Total lookup time: 5.72444e+09 nanoseconds

Bulkload time: 1.95953e+10

Total lookup time: 7.17878e+09 nanoseconds

Bulkload time: 1.90824e+10

Total lookup time: 6.17742e+09 nanoseconds

2-2 k=3 沒有移除離群值

80%

Bulkload time: 9.71623e+09

Total lookup time: 4.66782e+09 nanoseconds

Bulkload time: 1.10457e+10

Total lookup time: 5.13747e+09 nanoseconds

Bulkload time: 8.13074e+09

Total lookup time: 4.25134e+09 nanoseconds

40%

Bulkload time: 1.33645e+10

Total lookup time: 5.37938e+09 nanoseconds  
Bulkload time: 2.12372e+10

Total lookup time: 7.04227e+09 nanoseconds

Bulkload time: 2.09467e+10

Total lookup time: 5.57172e+09 nanoseconds

30%

Bulkload time: 1.28537e+10

Total lookup time: 5.00184e+09 nanoseconds

Bulkload time: 1.96721e+10

Total lookup time: 5.26169e+09 nanoseconds

Bulkload time: 1.73979e+10

Total lookup time: 6.26713e+09 nanoseconds

2-3 分三等份

90

Bulkload time: 8.09337e+09

Total lookup time: 3.73037e+09 nanoseconds

Bulkload time: 1.27842e+10

Total lookup time: 6.43557e+09 nanoseconds

Bulkload time: 1.18299e+10

Total lookup time: 4.59697e+09 nanoseconds

80

Bulkload time: 1.18927e+10

Total lookup time: 4.78869e+09 nanoseconds

Bulkload time: 1.56188e+10

Total lookup time: 5.31674e+09 nanoseconds

Bulkload time: 1.48487e+10

Total lookup time: 5.59115e+09 nanoseconds

front one lookup time: 1.48475e+09 nanoseconds

middle one lookup time: 1.14432e+09 nanoseconds

back one lookup time: 1.60214e+09 nanoseconds

70

Bulkload time: 1.01022e+10

Total lookup time: 4.2137e+09 nanoseconds

Bulkload time: 1.53003e+10

Total lookup time: 5.48077e+09 nanoseconds

Bulkload time: 1.69117e+10

Total lookup time: 6.75981e+09 nanoseconds

60

Bulkload time: 1.05217e+10

Total lookup time: 4.86968e+09 nanoseconds

Bulkload time: 1.63362e+10

Total lookup time: 6.63432e+09 nanoseconds

Bulkload time: 1.65082e+10

Total lookup time: 6.92775e+09 nanoseconds

50

Bulkload time: 1.28339e+10

Total lookup time: 4.78316e+09 nanoseconds

Bulkload time: 2.05095e+10

Total lookup time: 6.61588e+09 nanoseconds

Bulkload time: 2.10169e+10

Total lookup time: 6.78443e+09 nanoseconds

40

Bulkload time: 2.02363e+10

Total lookup time: 7.9997e+09 nanoseconds

Bulkload time: 2.05085e+10

Total lookup time: 8.54791e+09 nanoseconds

Bulkload time: 1.95825e+10

Total lookup time: 6.91108e+09 nanoseconds

30

Bulkload time: 1.2725e+10

Total lookup time: 8.78933e+09 nanoseconds

Bulkload time: 1.22821e+10

Total lookup time: 9.46291e+09 nanoseconds

Bulkload time: 1.34034e+10

Total lookup time: 8.68267e+09 nanoseconds

33

Bulkload time: 1.25535e+10

Total lookup time: 1.06285e+10 nanoseconds

Bulkload time: 1.24714e+10

Total lookup time: 9.96699e+09 nanoseconds

Bulkload time: 1.2995e+10

Total lookup time: 9.93856e+09 nanoseconds

Total lookup time: 7.34868e+09 nanoseconds

front one lookup time: 1.54864e+09 nanoseconds

middle one lookup time: 2.92754e+09 nanoseconds

back one lookup time: 2.8725e+09 nanoseconds

Total lookup time: 1.23384e+10 nanoseconds

front one lookup time: 1.72822e+09 nanoseconds

middle one lookup time: 6.34756e+09 nanoseconds

back one lookup time: 4.26261e+09 nanoseconds

20

Bulkload time: 7.37413e+09

Total lookup time: 6.39488e+09 nanoseconds

Bulkload time: 1.22886e+10

Total lookup time: 9.43936e+09 nanoseconds

Bulkload time: 1.37593e+10

Total lookup time: 9.77176e+09 nanoseconds

10

Bulkload time: 6.80845e+09

Total lookup time: 6.13179e+09 nanoseconds

Bulkload time: 1.16819e+10

Total lookup time: 7.0323e+09 nanoseconds

Bulkload time: 1.03729e+10

Total lookup time: 8.56884e+09 nanoseconds

Total lookup time: 9.34243e+09 nanoseconds

front one lookup time: 1.76756e+09 nanoseconds

middle one lookup time: 2.9445e+09 nanoseconds

back one lookup time: 4.63037e+09 nanoseconds

2-4 一次全部

80 4.36978e+10

40 3.8462e+10

30 4.5852e+10

Improve kmean?

離群值真的不好嗎

先移除離群值再做kmean

* Range之間的距離調整k=3再做一次
* K增加到5(或7)

同現在方法增加到五群 (Random 100M)

* Suboptimal (三群分開bulkload)
* Knn找離群值之中的離群值 (各2.5%) 中心點為靠中間的底

0(((環境會不會有新資料 影響是否只有look up time是參考指標

(頻繁加入的過程中結構在調整，不能做lookup要等bulkload結束)))))))

33%做kmean:

狀況1

center: 545716 6.85576e+06 1.00001e+09

size: 43322461 23344206 33333333

center有明顯的小中大，預期應該是每區各33333333，前中稍微混在一起了

狀況2

center: 2.75485e+06 1e+09 1.00001e+09

size: 66666667 16665277 16668056

前中合併成一個，後面兩個center位置幾乎一樣，平分後

跟range有關嗎?

range1(0.0, 10000.0); e+04

range2(1000000.0, 10010000.0); e+06

range3(1000000000.0, 1000010000.0); e+09

把bulkload的順序改成後中前

Total lookup time: 4.80768e+09 nanoseconds

front one lookup time: 5.31813e+08 nanoseconds

middle one lookup time: 1.0968e+09 nanoseconds

back one lookup time: 3.17906e+09 nanoseconds

Total lookup time: 4.5467e+09 nanoseconds

front one lookup time: 5.37052e+08 nanoseconds

middle one lookup time: 9.36236e+08 nanoseconds

back one lookup time: 3.07342e+09 nanoseconds

?????????

真的是距離的關係嗎