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