哈尔滨工业大学

<<大数据分析>>

实验报告

(2021 年度春季学期)

姓名:	卢兑玑
学号:	L170300901
学院:	计算机学院
教师:	

实验三 基于预计计算技术的大数据分析

一、实验目的

1. 了解预计算的优化效果。
2. 理解 Data Canopy 中线段树的设计原理及实现。
3. 体会并掌握系统优化技术。

二、实验环境

Windows 10 操作系统，python

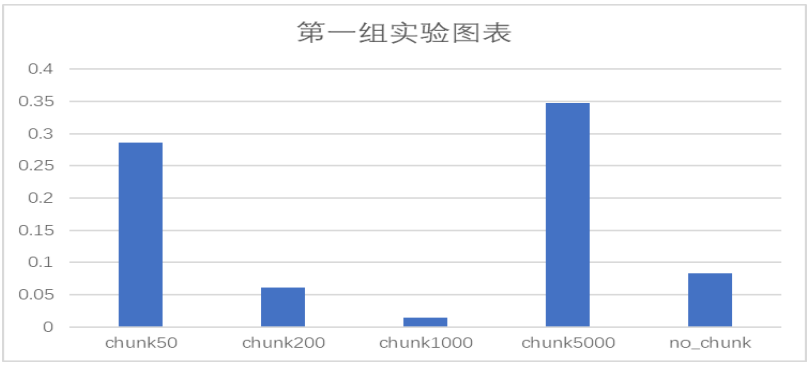
三、实验过程及结果

第一组实验：

设置参数 chunk size 为线段树中块的大小。在已经建立好的大数据分析系统中，执行查询 3.3 节中的工作负载集合[1]，通过改变 chunk size 的大小，来探究与分析 chunk size 参数对性能的影响。。

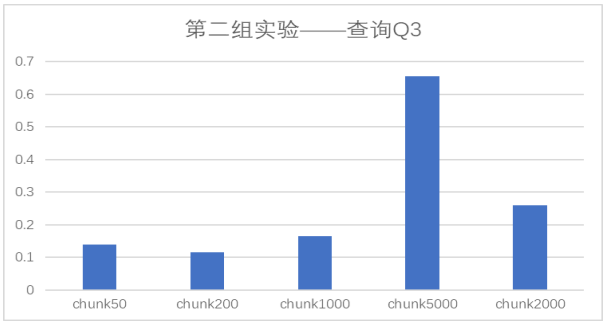
Chunk size 参数设置：

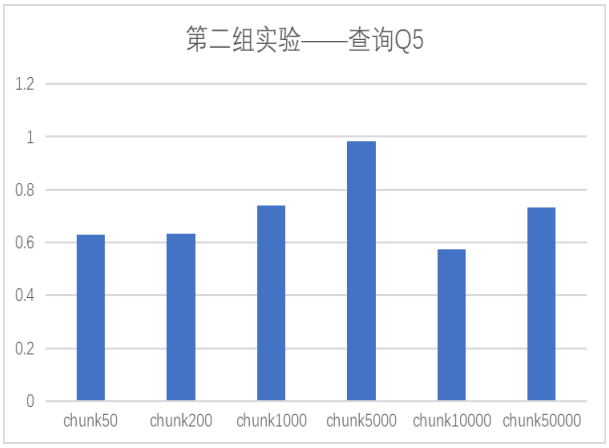
分别设置 chunk size 参数为 50、200、1k、0.5M 进行查询实验。

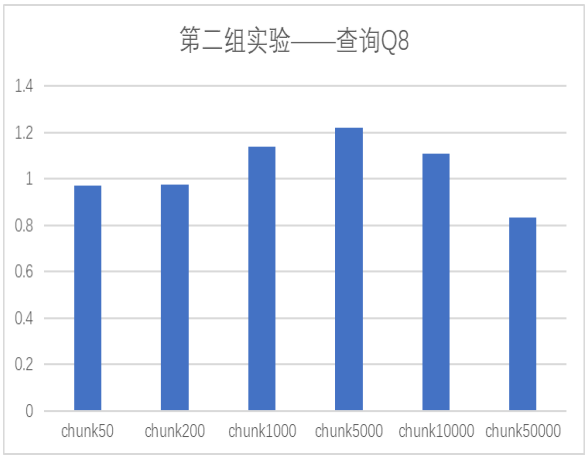


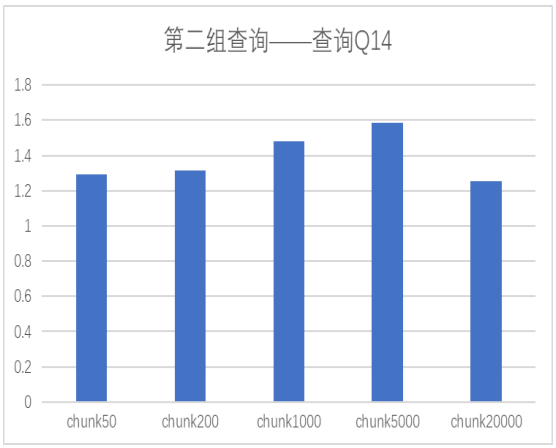
Q3 Q7 Q11 Q12 Q13 更适合硬扫描的方法，其它的更适合块查询的方法。因为块查询的方法更适合数据较多而且数据集中在几个块内部的情况，如果要查询的数据比较少，直接硬扫描会更快，如果数据较多，但是需要查询的数据中，起始部分或者结束部分的某些数据不是一个块中的所有数据，这个时候如果使用块查询，就需要对部分数据继续使用硬扫描或者对部分数据重新建立线段树再进行查询，时间成本会高于硬扫描查询。

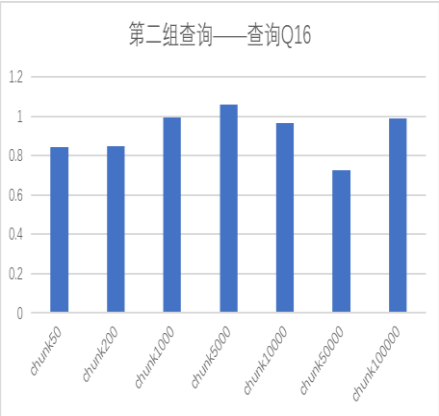
第二组实验：各个查询结果的对比图如下。


第二组实验——查询Q3
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
chunk50
chunk200
chunk1000
chunk5000
chunk2000


第二组实验——查询Q5
1.2
1
0.8
0.6
0.4
0.2
0
chunk50
chunk200
chunk1000
chunk5000
chunk10000
chunk50000


第二组实验——查询Q8
1.4
1.2
1
0.8
0.6
0.4
0.2
0
chunk50
chunk200
chunk1000
chunk5000
chunk10000
chunk50000


第二组查询——查询Q14
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
chunk50
chunk200
chunk1000
chunk5000
chunk20000



四、实验心得

通过这个实验，学会了一种全新的数据结构 segment tree。知道了这种数据结构在查询方面的巨大优势。并学会了如何熟练地使用这种数据结构。

体会到了基于预计计算技术在查询方面的优势，激发了我更深一步的研究这方面技术的热情。