数据库课程设计 1：可扩展哈希

1. 实验平台

Linux 或者 Windows

2. 编程语言

仅能使用 C/C++，不允许使用 STL。

3. 实现算法

可扩展哈希，具体算法请见课件*<lec7 Hash-Based Indexes>*中的第 5 至 12 张。

算法实现分为两大部分，第一部分是建立索引，第二部分是查询。建立索引是将输入的每一条记录根据指定的键值放入合适的哈希桶内，当哈希桶已满时，需要进行分裂。查询是根据输入的键值返回具有相同键值的记录，返回的记录可能有不止一条。

4. 实现过程

(1) 读入由 tpc-h 生成的 lineitem.tbl，以 L_ORDERKEY 属性作为键值将记录放入合适的哈希桶内。

(2) 读入测试文件 testinput.in 内的数据，数据中包含多个需要查询的键值，具体格式请见[6. 数据输入输出说明]。

(3) 将通过键值查询得到的所有记录都输出到 testoutput.out 文件中，具体格式请见[6. 数据输入输出说明]。

5. 实现细节

(1) 只能使用 P 个页，每个页的大小为 8K bytes，一个哈希桶的大小和一个页的大小相同。考虑以下两种情况：

a) P=8（也就是整个内存中仅能使用 8 个页，这 8 个页用于存放索引和哈希桶的数据）

b) P=128（也就是整个内存中仅能使用 128 个页，这 128 个页用于存放索引和哈希桶的数据）

由于页的数量有所限制，lineitem.tbl 的数据和对其建立的索引不可能都放在内存里，所以频繁的文件读写(I/O)是不可避免的。请比较 P=8 和 P=128 两种情况下，I/O 的次数、目录的大小、哈希桶的数量、查询的速度（每秒执行查询的数量）、I/O 用时占查询总用时的比例等。

(2) 对于哈希桶内的数据，采用<键，数据记录>的方式进行存储。详细内容请参考课件*<lec3 File and Indexing>*中的第 16、17 张。

(3) 对于存储哈希桶数据的页面，使用变长记录的方式进行存储。关于变长记录，请参考课件*<lec 3 Storing Data>*中的第 31 张。

(4) 使用时钟页面算法进行页面置换，请参考*<lec 3 Storing Data>*中的第 19 张。

(5) 分别实现从低位和从高位进行扩展的哈希。详细内容请参考课件*<lec 7 Hash-Based Indexes>*中的第 10 张。比较这两种哈希方法，包括桶的分裂方式、桶分裂时桶内数据的分配方式、I/O 的次数、目录的大小、桶的数量、查询的速度（每秒执行查询的数量）、I/O 用时占查询总用时的比例等。

(6) 将建立的哈希索引输出到新的文件中，命名为 hashindex.out，格式自定。

(7) 对于每个查询结果，请按照L_PARTKEY属性的值进行排序。

6. 数据输入输出说明

将文件路径作为启动参数传入程序，例如执行 ExtendibleHash.exe 程序， 执行命令为“ExtendibleHash.exe D:\database”。该路径表示 lineitem.tbl 和 testinput.in 所在的文件路径，同时也表示 testoutput.out 和 hashindex.out 输出的文件路径。以下是对文件的格式进行说明，请严格按照文件格式进行输入输出。hashindex.out 格式自定，因此不作说明。

(1) lineitem.tbl

通过 tpc-h 的 dbgen 程序生成，Scale Factor 设置为 1，所以记录的数量一定是 6001215。选择 L_ORDERKEY 属性作为键值建立哈希索引。

(2) testinput.in

数据第一行是 n，表示查询数量。接下来的 n 行，每行分别是一个整数，表示要查询的键值。

(3) testoutput.out

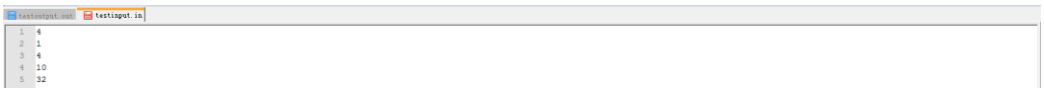
对于每一个查询，输出所有满足要求的记录，并且按照 L_PARTKEY 属性的值进行排序。每行一条记录。对于每个查询的结果，都以-1 作为结束。

以下是样例。

lineitem.tbl

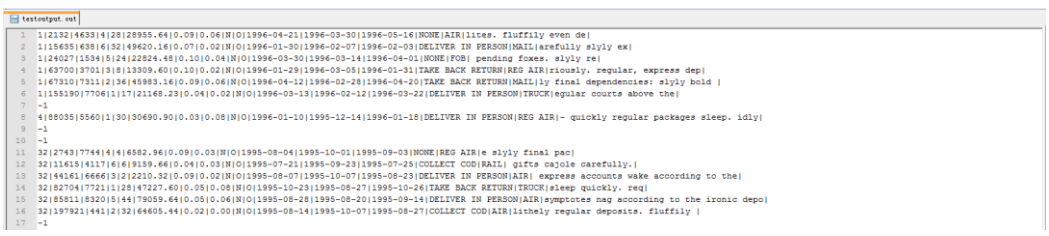
```
1|155190|7706|1|17|21168.23|0.04|0.02|N|O|1996-03-13|1996-02-12|1996-03-22|DELIVER IN PERSON|TRUCK|egular courts above the|
1|67310|7311|2|36|45983.16|0.09|0.06|N|O|1996-04-12|1996-02-28|1996-04-20|TAKE BACK RETURN|MAIL|ly final dependencies: slyly bold |
1|63700|3701|3|8|13309.60|0.10|0.02|N|O|1996-01-29|1996-03-05|1996-01-31|TAKE BACK RETURN|REG AIR|riously. regular, express dep|
1|2132|4633|4|28|28955.64|0.09|0.06|N|O|1996-04-21|1996-03-30|1996-05-16|NONE|AIR|lites. fluffily even de|
1|24027|1534|5|24|22824.48|0.10|0.04|N|O|1996-03-30|1996-03-14|1996-04-01|NONE|FOB| pending foxes. slyly re|
1|15635|638|6|32|49620.16|0.07|0.02|N|O|1996-01-30|1996-02-07|1996-02-03|DELIVER IN PERSON|MAIL|arefully slyly ex|
2|106170|1191|1|38|44694.46|0.00|0.05|N|O|1997-01-28|1997-01-14|1997-02-02|TAKE BACK RETURN|RAIL|ven requests. deposits breach a|
3|4297|1798|1|45|54058.05|0.06|0.00|R|F|1994-02-02|1994-01-04|1994-02-23|NONE|AIR|ongside of the furiously brave acco|
3|19036|6540|2|49|46796.47|0.10|0.00|R|F|1993-11-09|1993-12-20|1993-11-24|TAKE BACK RETURN|RAIL| unusual accounts. eve|
3|128449|3474|3|27|39890.88|0.06|0.07|A|F|1994-01-16|1993-11-22|1994-01-23|DELIVER IN PERSON|SHIP|nal foxes wake. |
3|29380|1883|4|2|2618.76|0.01|0.06|A|F|1993-12-04|1994-01-07|1994-01-01|NONE|TRUCK|y. fluffily pending d|
3|183095|650|5|28|32986.52|0.04|0.00|R|F|1993-12-14|1994-01-10|1994-01-01|TAKE BACK RETURN|FOB|ages nag slyly pending|
3|62143|9662|6|26|28733.64|0.10|0.02|A|F|1993-10-29|1993-12-18|1993-11-04|TAKE BACK RETURN|RAIL|ges sleep after the caref|
4|88035|5560|1|30|30690.90|0.03|0.08|N|O|1996-01-10|1995-12-14|1996-01-18|DELIVER IN PERSON|REG AIR|- quickly regular packages sleep. idly|
5|108570|8571|1|15|23678.55|0.02|0.04|R|F|1994-10-31|1994-08-31|1994-11-20|NONE|AIR|ts wake furiously |
5|123927|3928|2|26|50723.92|0.07|0.08|R|F|1994-10-16|1994-09-25|1994-10-19|NONE|FOB|sts use slyly quickly special instruc|
5|37531|35|3|50|73426.50|0.08|0.03|A|F|1994-08-08|1994-10-13|1994-08-26|DELIVER IN PERSON|AIR|eodolites. fluffily unusual|
6|139636|2150|1|37|61998.31|0.08|0.03|A|F|1992-04-27|1992-05-15|1992-05-02|TAKE BACK RETURN|TRUCK|p furiously special foxes|
7|182052|9607|1|12|13608.60|0.07|0.03|N|O|1996-05-07|1996-03-13|1996-06-03|TAKE BACK RETURN|FOB|ss pinto beans wake against th|
7|145243|7758|2|9|11594.16|0.08|0.08|N|O|1996-02-01|1996-03-02|1996-02-19|TAKE BACK RETURN|SHIP|es. instructions|
7|94780|9799|3|46|81639.88|0.10|0.07|N|O|1996-01-15|1996-03-27|1996-02-03|COLLECT COD|MAIL| unusual reques|
7|163073|3074|4|28|31809.96|0.03|0.04|N|O|1996-03-21|1996-04-08|1996-04-20|NONE|FOB|. slyly special requests haggl|
7|151894|9440|5|38|73943.82|0.08|0.01|N|O|1996-02-11|1996-02-24|1996-02-18|DELIVER IN PERSON|TRUCK|ns haggle carefully ironic deposits. bl|
7|79251|1759|6|35|43058.75|0.06|0.03|N|O|1996-01-16|1996-02-23|1996-01-22|TAKE BACK RETURN|FOB|jole. excuses wake carefully alongside of |
7|157238|2269|7|5|6476.15|0.04|0.02|N|O|1996-02-10|1996-03-26|1996-02-13|NONE|FOB|ithely regula|
32|82704|7721|1|28|47227.60|0.05|0.08|N|O|1995-10-23|1995-08-27|1995-10-26|TAKE BACK RETURN|TRUCK|sleep quickly. req|
32|197921|441|2|32|64605.44|0.02|0.00|N|O|1995-08-14|1995-10-07|1995-08-27|COLLECT COD|AIR|lithely regular deposits. fluffily |
32|44161|6666|3|2|2210.32|0.09|0.02|N|O|1995-08-07|1995-10-07|1995-08-23|DELIVER IN PERSON|AIR| express accounts wake according to the|
32|2743|7744|4|4|6582.96|0.09|0.03|N|O|1995-08-04|1995-10-01|1995-09-03|NONE|REG AIR|e slyly final pac|
32|85811|8320|5|44|79059.64|0.05|0.06|N|O|1995-08-28|1995-08-20|1995-09-14|DELIVER IN PERSON|AIR|symptotes nag according to the ironic depo|
32|11615|4117|6|6|9159.66|0.04|0.03|N|O|1995-07-21|1995-09-23|1995-07-25|COLLECT COD|RAIL| gifts cajole carefully.|
33|61336|8855|1|31|40217.23|0.09|0.04|A|F|1993-10-29|1993-12-19|1993-11-08|COLLECT COD|TRUCK|ng to the furiously ironic package|
33|60519|5532|2|32|47344.32|0.02|0.05|A|F|1993-12-09|1994-01-04|1993-12-28|COLLECT COD|MAIL|gular theodolites|
33|137469|9983|3|5|7532.30|0.05|0.03|A|F|1993-12-09|1993-12-25|1993-12-23|TAKE BACK RETURN|AIR|. stealthily bold exc|
33|33918|3919|4|41|75928.31|0.09|0.00|R|F|1993-11-09|1994-01-24|1993-11-11|TAKE BACK RETURN|MAIL|unusual packages doubt caref|
```

testinput.in



```
4
1
4
10
32
```

testoutput.out



```
1|2132|4633|4|28|28955.64|0.09|0.06|N|O|1996-04-21|1996-03-30|1996-05-16|NONE|AIR|lites. fluffily even de|
1|15635|638|6|32|49620.16|0.07|0.02|N|O|1996-01-30|1996-02-07|1996-02-03|DELIVER IN PERSON|MAIL|arefully slyly ex|
1|24027|1534|5|24|22824.48|0.10|0.04|N|O|1996-03-30|1996-03-14|1996-04-01|NONE|FOB| pending foxes. slyly re|
1|63700|3701|3|8|13309.60|0.10|0.02|N|O|1996-01-29|1996-03-05|1996-01-31|TAKE BACK RETURN|REG AIR|riously. regular, express dep|
1|67310|7311|2|36|45983.16|0.09|0.06|N|O|1996-04-12|1996-02-28|1996-04-20|TAKE BACK RETURN|MAIL|ly final dependencies: slyly bold |
1|155190|7706|1|17|21168.23|0.04|0.02|N|O|1996-03-13|1996-02-12|1996-03-22|DELIVER IN PERSON|TRUCK|egular courts above the|
-1
4|88035|5560|1|30|30690.90|0.03|0.08|N|O|1996-01-10|1995-12-14|1996-01-18|DELIVER IN PERSON|REG AIR|- quickly regular packages sleep. idly|
-1
-1
32|2743|7744|4|4|6582.96|0.09|0.03|N|O|1995-08-04|1995-10-01|1995-09-03|NONE|REG AIR|e slyly final pac|
32|11615|4117|6|6|9159.66|0.04|0.03|N|O|1995-07-21|1995-09-23|1995-07-25|COLLECT COD|RAIL| gifts cajole carefully.|
32|44161|6666|3|2|2210.32|0.09|0.02|N|O|1995-08-07|1995-10-07|1995-08-23|DELIVER IN PERSON|AIR| express accounts wake according to the|
32|82704|7721|1|28|47227.60|0.05|0.08|N|O|1995-10-23|1995-08-27|1995-10-26|TAKE BACK RETURN|TRUCK|sleep quickly. req|
32|85811|8320|5|44|79059.64|0.05|0.06|N|O|1995-08-28|1995-08-20|1995-09-14|DELIVER IN PERSON|AIR|symptotes nag according to the ironic depo|
32|197921|441|2|32|64605.44|0.02|0.00|N|O|1995-08-14|1995-10-07|1995-08-27|COLLECT COD|AIR|lithely regular deposits. fluffily |
-1
```

7. 提交说明

请将所有要提交的内容放入一个压缩文件内，命名格式为“组号_组长姓名_实现平台”，例如“01_张三_linux”。组号信息请向学委咨询。要提交的内容如下：

(1) README

写明所提交的内容和小组信息。小组信息包括组员学号、姓名、班级。

(2) 源代码

放入”src”文件夹内。

(3) 可执行程序

放入”bin”文件夹内。

这里需要提交四个可执行程序。由于程序有两个参数，一个是页的数量，一个是哈希的扩展方式，根据排列组合，有 4 种方案。根据方案中参数的不同分别命名为 least_128, most_128, least_8, most_8。并附上 README 说明每个程序完成的功能和如何运行程序。

(4) 实验报告

可用中文或英文写。实验报告必须为 pdf 格式，请以“组号_report”的方式命名，例如”01_report”。

实验报告必须包含以下内容：

a) 实验环境、工具说明
b) 实验思路
c) 数据结构设计
d) 实验过程设计
e) 代码结构
f) 实验结果、不同参数下性能的比较（[5. 实验细节](1)和(5)中提到的比较内容）
g) 心得体会

以上部分内容可根据需要使用简单的伪代码、流程图等表示。但切勿大篇幅粘贴代码！

可参考以下实验报告：

http://ss.sysu.edu.cn/~fjl/sun.doc
http://ss.sysu.edu.cn/~fjl/ruan.pdf

8. 注意

(1) 务必将文件路径作为启动参数传入程序。
(2) 请不要改动任何输入文件，即 lineitem.tbl 和 testinput.in。
(3) 请严格按照要求进行输入输出和提交课程设计。
(4) 请注意实验细节中提到的所有要求，请按照要求实现算法。对于没有明确要求的地方，可自行决定，并在实验报告中说明。
(5) 建议不要使用 fscanf 和 fprintf，因为读写文件速度不理想。按照页的大小读入一定字节的数据会有利于提高 I/O 速度。
(6) 虽然没有提供查询测试数据，但可自行设计查询测试数据来比较查询速度等。
(7) 组内所有成员课程设计部分的得分都相同。
(8) 如果某同学期末考试卷面分数低于 40 分，那么就可能将其课程设计分数乘以卷面分数，然后除以 100，作为最终的课程设计分数。

课程设计提交截止时间：2012 年 5 月 13 日 11:00 PM