CSE441 DATABASE SYSTEMS

ASSIGNMENT 3

Question: To perform duplicate elimination for a given relation.

FUNCTION PROTOTYPE : distinct (R, M, type_of_index)

R is a name of relation.
n the number of attributes. (n<=4 , and each attribute value will be < 1000)
M is the number of blocks. Note that B(R) >M and M>2.
Type_of_index takes two values: hash or btree.

Write a program to remove the duplicates. You need to write three routine **open()**, **Getnext()** and **close()** . The program should call above routines to eliminate the duplicates. You can create indexes as a part of open().

To search whether the record is duplicate or not, use B+tree and hashing main memory structures for inserting and checking. The space required for B+Tree and hash is not a part of M (you can use OS buffers).

Your program should expect the number of main memory blocks and a filename which contains the sequence of integers (first n represents first row, and the next n represents the next row and so on).
Your program can specify minimum and maximum size of the blocks.

Instructions :

Out of the M buffers, M-1 Buffers will be used as input buffers (which will hold the records from the input file). 1 buffer will be used as output buffer (holds the distinct records). If the output buffer gets filled, it should be flushed to the output file. If the input buffers get empty, next chunk of records should be read from the input file.

Getnext() function when called should always return one of the following :
1. Record : This needs to be forwarded to the output buffer (Its a distinct record)
2. Null : The input file is completely processed. Proceed for close() routine.
Output: Vary from M >= B(R) to M= (3/4)(B(R)) and calculate the execution time by employing B+tree and hashing for inserting and checking duplicate entry.

Generation of R: Generate R of 1GB size .Generate R using random number function with r% duplication (r is an integer). After generating every 100 tuples, copy of any r tuples generated so far by selecting the same in the random manner.

Plot the values of M versus execution times for B+tree , Hashing and upload the image file.
Allowed Languages : C,C++,Java,python.

Any sought of copying from internet or from friends will lead to straight zero in all the assignments.

Upload Format :
1. Create a folder with your rollnumber.
2. Put all the code files & ReadMe.txt in to the folder created in 1.
3. Tar.gz the folder and name the archive as Rollnumber_Assignment3.tar.gz

Deadline : 1 October 2017 09:00PM