**Department of Computer Science**

**The University of Hong Kong**

**CSIS0323/COMP3323: Advanced Database Systems**

**Assignment 1 (10% of total marks)**

**Due date: Saturday, October 8, 2016, 5:00pm**

**Q1 [30%]**. Consider two relations *r(A,B)* and *s(C,D)*. Relation *r* has 50000 tuples, and *s* has 6000 tuples. 25 tuples of *r* fit in one block, 40 tuples of *s* fit in one block. Assume that the system has a memory of *M* =100 blocks.

(i) [20%] Estimate the **minimum** cost (in number of block accesses) of the equi-join with *r.B*=*s.C* using the following algorithms:

(a) nested-loops join, (5%)

(b) block nested-loops join, (5%)

(c) merge-join (assume that relations *r* and *s* are not initially sorted), (5%)

(d) hash-join (assume no overflow occurs and no recursive partitioning occurs.) (5%)

(ii) [10%] Assume that the block size is 4096 bytes. Suppose there are *B+*-tree indexes on both *r.B* and *s.C*. Neither of these two B+-tree indexes is primary index. For each block in the two *B+*-tree indexes, 96 bytes space is reserved for block header information (e.g., *is\_leaf* flag and number of entries). Except one sibling access pointer, only key values and record-ids are stored in leaf nodes. Both key values, *r.B* and *r.C*, are 12 bytes long. A record-id and a block-id each occupies 12 bytes. Please estimate the **minimum** cost (in number of block accesses) of running the index nested loops equi-join algorithm with *r.B* =*s.C*. (Assume that the join result contains N=30000 tuples.)

**Q2 [30%]**. SQL query:

SELECT \*

FROM R, S

WHERE (R.c=S.c) AND (71 <= R.a AND R.a < 80) AND (S.b=5)

Given the above SQL query, please compute the cost (num of block access) and the output size (number of tuples) produced by the following query plan. We have the following information:

* The relation schemas are R(id,a,c) and S(id,b,c). All id,a,b,c are integers.
* R has 5000 records and S has 10000 records
* The domain of the attribute R.a is [1,200].
* The domain of the attribute S.b is [1,10].
* The domain of the attributes R.c and S.c is [1,5].
* Values of non-key attributes are uniformly distributed.
* All values in the attribute domains appear in the relations.
* Each block can hold 10 records
* The smaller temporary table (between T1 and T2) is used as the outer relation of the block nested loop join.

Show your steps clearly.



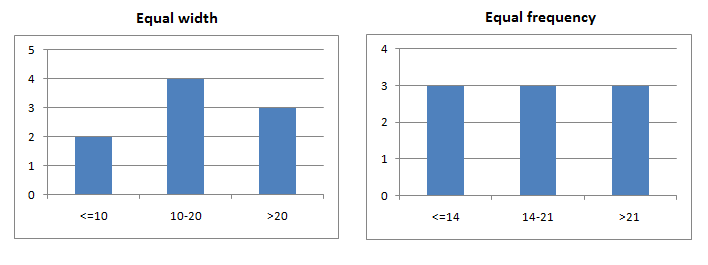
**Q3 [40%]**. **Histograms (Programming)**

Histograms are used by database systems to approximately capture the distribution of the values of an attribute in a table. They are mainly used for the selectivity estimation of selections. A histogram divides the domain of the attribute into ranges (bins) and for each range it stores the average number of records that have a value in this range. In this exercise you will write a program in C++ or Java which will create histograms and use them to estimate the results of queries.

Browse to <https://kdd.ics.uci.edu/databases/internet_usage/internet_usage.html> and download files final\_general.dat.gz and final\_general.col . Uncompress the first file. The first file includes the tuples of a table, whereas the second file includes the table’s schema (i.e., its attributes and their domain types). You are asked to extract the values of the **2nd field (Age)** and store them in an internal representation of your program (e.g., array, vector). Note that the range of age in this dataset is [0, 80].

**Part 1**: You are asked to create an *equi-width histogram* and an *equi-depth histogram* for attribute Age. In the equi-width histogram all bins have the same width (i.e., all ranges cover the same number of attribute values), for example [0,10), [10,20), [20,30), etc. In the equi-depth histogram (a.k.a. equal frequency histogram) the ranges are determined such that every range has the same sum of records in it. In our test case, the range bounds of the equi-depth histogram is such that the number of records from final\_general.dat that fall in each range is the same. As a result some ranges are wider than others.

The representation of a equi-width histogram is a vector of numbers (i.e., one number for each range of values), whereas the representation of a equi-depth histogram is a vector of range separators. The next figure shows examples of an equi-width and an equi-depth histogram.



Write a program that reads final\_general.dat, generates the two histograms, and writes them in a file. The histograms that you will generate should have **8** bins each. For the equi-width histogram, the eight intervals are: [0, 10), [10, 20), ```, [60, 70), [70, 80], so you need to write a file with 8 lines, with each line containing a single integer representing the frequency of the bin. For the equi-depth histogram, you need to decide 7 thresholds a1, a2, ```, a7, so that the 8 intervals will be: [0, a1), [a1, a2), ```, [a6, a7), [a7, 80]. You need to write a file with 8 lines; the first line should have the total number of records in the relation, whereas each line that follows containing a single integer representing the threshold.

**Part 2**: In the second part of this exercise you will test the accuracy of the histograms created in the first part for range selection queries expressed by closed intervals (e.g. [α, β]). In other words, we can use a histogram to estimate how many tuples are in the result of a selection on the attribute Age, for which the condition is α <= Age <= β. If α (or β) are not exactly the begin (or end) point of the bin ranges, then we will have to estimate the result based on the percentage of the bin ranges that are covered by [α, β]. For example, if a bin covers sub-domain [10, 20) in which 300 tuples fall, then the query interval [17, 64] covers [17, 20) from this bin’s range, i.e., 30% of the bin, therefore we estimate that 90 tuples (=300\*30%) from this bin are part of the result.

Write a program that takes as input a range [α, β], reads the histograms from the file and outputs: (1) the estimated result using the equi-width histogram (2) the estimated result using the equi-depth histogram (3) the real result, using the actual data. Through experimentation, conclude whether the equi-width histogram is superior/inferior to the equi-depth histogram, for different widths of the query range [α, β].

*Notes:*

*Please submit two files for this assignment to Moodle:*

1. *A docx (or PDF) file, containing your answers for Q1, Q2 and a report for Q3.*
2. *A zip file, containing a folder of either C++ or Java project.*

*Please DO NOT submit any other more materials.*