Concordia University
Dept. of Computer Science and Software Engineering
COMP 6521: Advanced Database Technology and Applications
Winter 2018

Mini-Project 1
==========
>> Total Points: 7
>> Project Report: Friday February 16th
>> Project Demo: Wednesday March 7th

Project Description: Consider two relations T1 and T2 both with the same schema defined as follows:

Student ID:	int (08)
First Name:	char(10)
Last Name:	char(10)
Department:	int (03)
Program:	int (03)
SIN Number:	int (09)
Address:	char(57)

Assume that every 40 tuples are stored in one block of 4K bytes. Suppose these relations are stored on the disk, and that each relation can have duplicated tuples. You may consider that each tuple appears on a separate line of the input file, stored on consecutive disk blocks. Note that there will be no separator (character) between the attribute values in a tuple and hence the length of the attributes is used to identify and extract the values. Examples of possible two records (lines) are as follows:

12345678John Smiths 111222999999999 1455 Maisonneuve West, Montreal, QC, H3G 1M8
88888888Roselyn Shiri 222444888888888 1515 Saint-Catherine West, Montreal, QC, H3G 1M7

You and your team are required to implement a sort-based bag difference *bd*(T1,T2), such that if a tuple t appears m times in T1 and n times in T2, the program returns (t,k), where k=m-n, if m>n, and returns 0 otherwise. Note that *bd* is commutative. Your program should report th e total number of tuples in the output and the number of blocks used to store the output on the disk, packing 40 tuples in each block. Implement the TPMMS method to sort the input relations and use it as a step to produce the result. Also report the number for disk I/Os to produce the results and write it back on the disk.

You need to evaluate the performance of your implementation on large instances of relations T1 and T2. The lab instructors will create sample instances of T1 and T2 which you can use to evaluate your work. Report the results of your experiments and evaluations obtained using these instances. They will be made available soon later.

To evaluate your work, create and use instances of T1 and T2 at around 1,000,000 and 1,000,000 tuples, respectively.

To better understand the impact of the amount of available main memory, limit the main memory available to two cases:
(1) 5 MB and
(2) 10 MB
Run your experiments in each of these two cases and report the following:

(1) Compare the number of disk I/Os and the execution time (in seconds) for sorting the files only using each of the main memory sizes above.

(2) Compare the number of disk I/Os and the execution time (in seconds) for performing the whole operation (sort and computing the bag difference) for each of the above main memory sizes.

What tools you should use?

Use VM argument Xmx5m in Eclipse to restrict the main memory usage of Java Virtual Machine. The lab assistants can help you with using Xmx5m.

What to submit on due date:

Submit your project report and the source codes through MOODLE. Please include instruction to compile and run your code. Make sure your program compiles and runs on the lab computers.

Book a time slot with the lab assistants for the demo of your project on March 7th. Both lab instructors will be present for evaluating your program. Every member of your team MUST be present during your project demo.

Bonus: The lab instructors may recommend additional 2 points for the implementation with best performance and additional 1 point for the next best.