Concordia University
Computer Science and Software Engineering
COMP 6521: Advanced Database Technology and Applications
Winter 2021

Total Points: 10
Codes & Reports due: Sunday, February 28th @ 23:55
Lab Demos due: Wednesday, March 17th @ 23:55

Project Description: We have two large tables T1 and T2 which records information about orders to a retail business with branches B1 and B2, respectively. The schema of these tables are identical and defined as follows – additional info to some attributes are provided in (…).

Client-ID INT(8)
Name CHAR(30)
Gender INT(1) (1 for male and 0 for female)
Social Security Number INT(9)
Item-ID-Ordered CHAR(45)
Quantity-Ordered INT(7)
Order-Date CHAR(10) (format: 'YYYY-MM-DD')

Clients order items to branches B1 and B2, which are recorded in tables T1 and T2, respectively. The general manager of the retail store needs your team to merge T1 and T2 such that for every client C, the "merged" table T would include the records with the information: Client-ID, Name, Gender, SSN, Item-ID-Ordered, Total-Quantity-Ordered, Date-of-Last-Order, and No-of-Orders, where the last attribute indicates the number of orders in T1 and T2 made by client C. Originally, T1 and T2 are stored and maintained separately and independently as data blocks of size 1024 bytes each, with 9 records per block. The remaining 34 bytes in a block is unused by the application (it may be used by OS to store block header info).
Implement the TPMMS sort-based algorithm discussed in the lecture to scan T1 and T2 from disk, block by block and follow steps to produce the desired "merged" table T.
To provide the input tables T1 and T2 you may either use a DBMS that holds these tables or use Java or Python to develop an entirely file processing system to perform the task. The main part of this project is development of the come that implements TPMMS, evaluate and improve its performance under the following conditions about the available main memory to this process. If you use Java or Python, assume that each record in T1 and T2 is on a single line terminated with the end-of-line character, and there is no delimiter separating the values of different attributes on a record line. An example of a tuple in these files is shown below 2 lines which is actually a single line:

```
12345678John Smith                    1123456789Heat pump, model X2000
                          65212021-01-2701
```

More details: There is no null values in the inputs. Report the number of

records in the resulting table T, the number of blocks used to hold records in T (pack 9 output records in each block), the total number of disk I/O's performed to produce T, as well as the execution time in minutes.

You need to evaluate the performance of your implementation using large instances of T1 and T2. The PODs will help create instances of T1 and T2 which you can use to run and evaluate the performance. Your report should properly present the test results and analyses. You may include additional test results using instances you created and used for performance evaluation.

To further evaluate the performance of your implementation, consider instances of tables T1 and T2 with half a million and a million records, respectively. Study the performance of your implementation in the following two situations of restricted main memory available: (1) 10MB and (2) 20MB. Run experiments in case (1) and (2), and report the following issues:

- Compare the number of disk I/O's and the execution time, in seconds, for the sort operations given in each case of the two main memory sizes. This should include the time to write the sorted data back to the disk.

- Compare the number of disk I/O's and the execution time for performing the whole task in each case of main memory sizes.

Project Report: We suggest the following structure and content of your report:

1. Description of your code and architecture designs
2. Highlights of your implementation features
3. Performance results and analysis

Please use any standard format to present your report. For example, use double-column, single-spaced, 11pt font size, with "reasonable" margins.

It is important to properly acknowledge any ideas, work, tools, code, etc. borrowed in your work.

What tools you should use?

Use VM argument Xmx5m in Eclipse to restrict the main memory sizes in Java Virtual Machine.

What to submit by the due date?

Through Moodle, submit your report (in a single PDF format) and the source codes (all as a single zip file). This also includes instruction to compile and run your code.

Demos: Book a time slot with the lab assistants for the demo of your project on March 17th. You and your team member must be present during your project demo.

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