Homework #4 (Query Execution)

Release: March 8, 2024

Due: 11:59pm, Friday, March 22, 2024

Points: 100

1. [External Sorting, 60 points]

   Consider the following table:

   product(name, price)

   Suppose the table is stored in a CSV file product.csv in the following format:

   iphone15,1500
   iphone13,1000
   t450s,2000
   …

   Consider a query:

   Select *
   from product
   order by name

   Suppose the table is big and order by is implemented using external sorting.
   Assume that a page/block **can only hold 2 rows of the table and there are 3 pages available for the external sorting process (which includes sorting phase and potentially multiple merging phases**). Like the example in class, we assume that only one page is used for the sorting, while all 3 pages are used in the merging process.

   Write a Python program ext_sort.py that takes the csv file and produces a version of the file sorted by name of product.

   Execution format:

   python3 ext_sort.py product.csv product_sorted.csv

   Note that your code should not load the entire CSV file into the memory.

2. [Simple Sort-Based Join, 40 points] Now consider another table maker:

   maker(product, company)

   Suppose the table is stored in a CSV file maker.csv in the following format:

   iphone15,apple
   iphone13,apple
   t450s,Lenovo
   …

Consider the following query:

Select name,price,maker

From product join maker on product.name = maker.product

Write a Python program ssb_join.py that uses the **simple sort-based join** algorithm to implement the join. Recall that this join algorithm completely sorts each table first, then merges the runs, one for each table, to find the join tuple.

Execution format:

python3 ssb_join.py product_sorted.csv maker_sorted.csv joined_data.csv

It should **write** output to the **output file [joined_data.csv]** in the following format:

iphone15,1500,apple

iphone13,1000,apple

t450s,2000,Lenovo

…

[Q2] Requirements:

- You should use the ext_sort.py program to produce sorted tables.
- Similar to Task 1, you should assume that there are only 3 pages of main memory available for the program.

Allowed libraries: sys,csv,os

Resources:

1. https://docs.python.org/3/library/csv.html
2. https://docs.python.org/3/library/functions.html#next

Submission Instructions:

1. Submit 2 .py files - ext_sort.py and ssb_join.py
2. Do not modify any contents in the template. Just fill the template by reading the comments.
3. You will get 0 points if the code breaks for any syntax errors or any other problems. Please test the code thoroughly before submitting.
4. Only 50% of the entire credit will be given if any other modules are inserted other than the one specified above.
5. **The logic of the algorithms should be from the course lectures. No points will be awarded if any other logics or algorithms are used.**
6. More than output correctness this assignment will be evaluated based on the process of algorithm implementation.