2DB3 Assignment 5 Part 2

Yisi Liu
400494848
liu1857

March 14, 2025

1 Efficiency of queries

1.1 P5

The cross product of Book (4000 rows) and BookCopy (5 copies per books, 20000 rows) will be 80000000 rows.

The cross product of Review (20 reviews per user, 1000 users, 20000 rows total) and Loan (4 per user, 4000 rows) will be 80000000 rows.

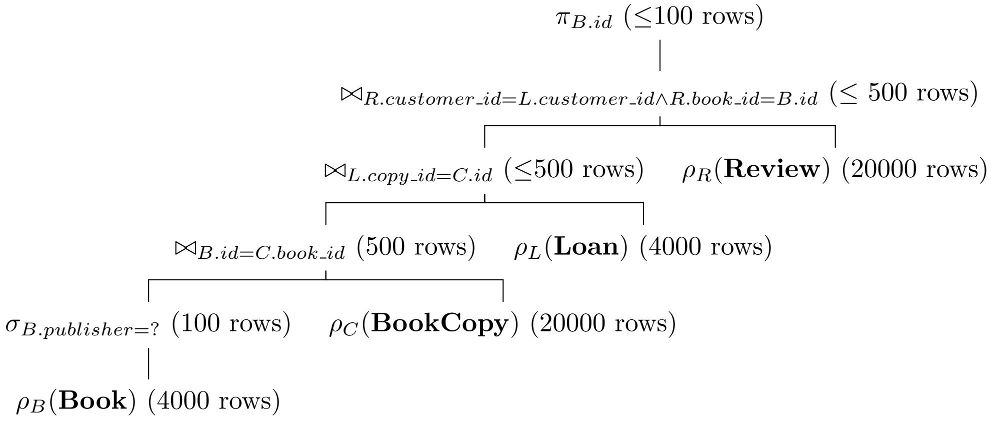
The cross product of the cross products will be 6400000000000000 rows.

The first σ operation will produce at most 4000 rows because the condition between Book and BookCopy restricts the rows to BookCopy, and the condition between BookCopy and Loan restricts the rows to Loan, and the conditions between Review and Loan and Review and Book restrict Review to the rows of the rest of the conglomerate. The max number of rows in Loan is 4000.

The second σ operation selecting a specific publisher restricts the number of rows to at most 500 because each publisher has 100 books and each book has 5 copies.

The final π operation selecting distinct book IDs can have at most 100 rows because a single publisher can have at most 100 books, even if none of them were weeded out at the previous steps.

1.2 P6


$\pi_{B.id}$ ($\leq$100 rows)
$\bowtie_{R.customer\_id=L.customer\_id \wedge R.book\_id=B.id}$ ($\leq$ 500 rows)
$\bowtie_{L.copy\_id=C.id}$ ($\leq$500 rows)
ρ_R(**Review**) (20000 rows)
$\bowtie_{B.id=C.book\_id}$ (500 rows)
ρ_L(**Loan**) (4000 rows)
$\sigma_{B.publisher=?}$ (100 rows)
ρ_C(**BookCopy**) (20000 rows)
ρ_B(**Book**) (4000 rows)

It is a good query execution plan because it removed a redundancy, L.returned = false, since all Loans are assumed to be active, and because the max number of rows for an intermediate step is 500, which is way less than the 6400000000000000 of the original query execution plan.

1.3 P7

The size estimates are already marked on the query execution plan tree, but here their reasoning will be explained.

Starting from the bottom, the σ operation on Book that select a specific publisher narrows it down to 100 rows because there are 40 publishers that each published the same number of books.

Next, the inner join with BookCopy creates 500 rows because each book has 5 copies.

Next, the inner join with Loan creates at most 500 rows because each copy of a book is unique on the same level for BookCopy and Loan, and if every copy of every book by that publisher all happen to be on loan, there is a maximum of 500 books on loan.

Next, the inner join with Review creates at most 500 rows because each unique pair of book_id and customer_id gets a review (if that customer has reviewed that book) so unless the same person reviews the same book multiple times, which is not possible since the schema for Review only has book_id and customer_id as its primary key, the number of rows cannot increase.

Finally, the π operator drops the max possible number of rows down to 100 because it gets rid of the copies of the books by selecting the distinct book IDs.

1.4 P8

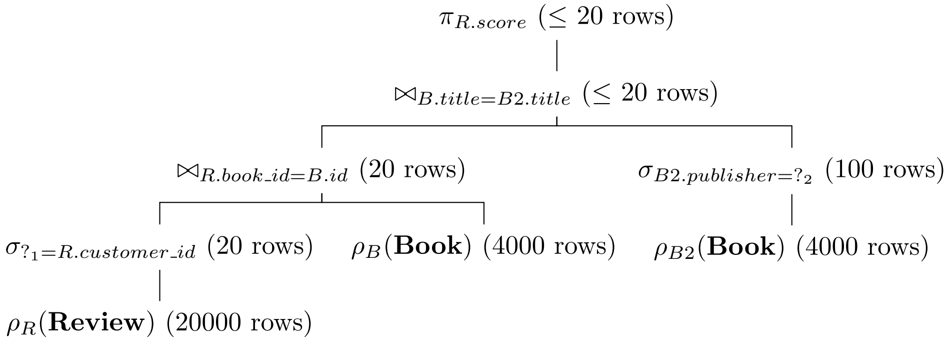
SELECT DISTINCT B.id
FROM Book B, BookCopy C, Review R, Loan L
WHERE B.id=C.book_id AND C.id=L.copy_id AND R.book_id = B.id AND L.customer_id=R.customer_id
AND L.returned=false AND B.publisher=? ;

1.5 P9

$\pi_{R.score}(((\sigma_{?_1=R.customer\_id}(\rho_R(\textbf{Review})) \bowtie_{R.book\_id=B.id} (\rho_B(\textbf{Book}))) \bowtie_{B.title=B2.title} (\sigma_{B2.publisher=?_2}(\rho_{B2}(\textbf{Book})))$

The query is slightly simplified. Customer is unnecessary because Review also has a customer ID, and when representing the IN keyword with an inner join, I found it unnecessary to select the title because it's in the join condition. I did have to name the second Book table to B2 even though the SQL query did not have to.

1.6 P10


$\pi_{R.score}$ ($\leq$ 20 rows)
$\bowtie_{B.title=B2.title}$ ($\leq$ 20 rows)
$\bowtie_{R.book\_id=B.id}$ (20 rows)
$\sigma_{B2.publisher=?_2}$ (100 rows)
$\sigma_{?_1=R.customer\_id}$ (20 rows)
$\rho_B(\textbf{Book})$ (4000 rows)
$\rho_{B2}(\textbf{Book})$ (4000 rows)
$\rho_R(\textbf{Review})$ (20000 rows)

My query execution plan is good because at most an intermediate step will generate 100 rows, which is not very many.

1.7 P11

Starting from the bottom, since a customer writes 20 reviews on average, first the σ operator selecting a specific customer cuts the rows down to around 20.

Next, the join between the smaller Review and Book is still around 20 rows because each review has one book they are reviewing.

Next, jumping to the right side, the B2 Book table is 100 rows after σ a specific publisher because 40 publishers with 4000 books is 100 books each.

Next, the join between the two sides is around 20 rows or less because a book may or may not fit both criteria.

Finally, the π operator is also around 20 or less because there may be duplicate scores.