CS 186 Fall 2024

Introduction to Database Systems
Alvin Cheung

DIS 6

1 Assorted Joins

- Companies: (company_id, industry, ipo_date)
- NYSE: (company_id, date, trade, quantity)

We have 20 pages of memory, and we want to join two tables Companies and NYSE on C.company_id = N.company_id. Attribute company_id is the primary key for Companies. For every tuple in Companies, assume there are 4 matching tuples in NYSE.

NYSE contains [N] = 100 pages, NYSE holds $p_N = 100$ tuples per page.

Companies contains [C] = 50 pages, C holds $p_C = 50$ tuples per page.

There are alternative 3 unclustered B+ tree indexes of height 1 on C.company_id and N.company_id. Throughout the problem, do not assume any caching of index nodes.

(a) How many disk I/Os are needed to perform a simple nested loop join?

(b) How many disk I/Os are needed to perform a block nested loop join?

(c) How many disk I/Os are needed to perform an index nested loop join?

(d) For this part only, assume the index on NYSE.company_id is clustered. What is the cost of an index nested loop join using companies as the outer relation?

(e) In the average case, how many disk I/Os are needed to perform a sort merge join without optimization? If we can perform the sort merge join optimization, how many disk I/Os are needed with optimization?

2 Grace Hash Join

We have 2 tables – Catalog and Transactions.

Catalog has a total of 100 pages and 20 tuples per page. Transactions has a total of 50 pages and 50 tuples per page. Assume the hash functions uniformly distribute the data for both tables.

(a) If we had 10 buffer pages, how many partitioning phases would we require for grace hash join? Consider which table we should build the hash table in the probing phase on.

(b) What is the I/O cost for the grace hash join then?

(c) For the above question, if we only had 8 buffer pages, how many partitioning phases would there be?

(d) What will be the I/O cost?

3 Relational Algebra

Consider the schema:

- `Songs(SONG_ID, song_name, album_id, weeks_in_top_40)`
- `Artists(ARTIST_ID, artist_name, first_yr_active)`
- `Albums(ALBUM_ID, album_name, artist_id, yr_released, genre)`

Write relational algebra expressions for the following queries:

(a) Find the names of the artists who have albums with a genre of either ’pop’ or ’rock’.

(b) Find the names of the artists who have albums of genre ’pop’ and ’rock’.

(c) Find the id of the artists who have albums of genre ’pop’ or have spent over 10 weeks in the top 40.

(d) Find the names of the artists who do not have any albums