(67329) DB | תרגיל 4

שם: רונאל חרדים, עומרי טויטו | ת"ז:208917641,208432361

תחילה נחשב חישובים כללים על הטבלאות:

מספר בייטים בשורה בטבלה $authors$:

$$3 \cdot 10 + 3 \cdot 4 = 42$$

מספר הבלוקים בטבלה $authors$**:**

$$\frac{42 \cdot 165000}{8192} = \frac{6930000}{8192} = 847$$

מספר בייטים בשורה בטבלה $conferences$:

$$3 \cdot 10 = 30$$

מספר הבלוקים בטבלה $conferences$**:**

$$\frac{30 \cdot 20000}{8192} = \frac{600000}{8192} = 74$$

שאלה 1

(א)

נחשב את עלות הצירוף של שתי הטבלאות לפי האלגוריתמים הנתונים:

1. **עלות החישוב לפי** BNL**:**

נרצה שהיחס החיצוני יהיה היחס בקטן יותר לכן נבחר את $conferences$ להיות היחס החיצוני ונחשב לפי הנוסחה:

$$74 + 847 \cdot \left\lceil \frac{74}{25-2} \right\rceil = 74 + 847 \cdot 4 = 3462$$

2. **עלות החישוב לפי** $Hash\ join$**:**

תחילה נבדוק כי התנאי מתקיים:

$$\left\lceil \frac{74}{25-1} \right\rceil = 4 \leq 23$$

לכן ניתן להריץ את האלגוריתם, כעת נחשב את העלות:

$$3 \cdot (74 + 847) = 2763$$

3. נוכל להשתמש באלגוריתם $SORT\ MARGE$ רק אם מתקיים התנאי:

$$\left\lceil \frac{B(R)}{M} \right\rceil < M, \left\lceil \frac{B(S)}{M} \right\rceil < M$$

נבדוק אם התנאי מתקיים:

$$\left\lceil \frac{847}{25} \right\rceil = 34 > M = 25$$

התנאי לא מתקיים ולכן לא ניתן להריץ את האלגוריתם.

(ב)

נחשב את עלות הצירוף של שתי הטבלאות לפי האלגוריתמים הנתונים עם גודל חוצץ = 30:

1. **עלות החישוב לפי** BNL**:**

נרצה שהיחס החיצוני יהיה היחס בקטן יותר לכן נבחר את $conferences$ להיות היחס החיצוני ונחשב לפי הנוסחה:

$$74 + 847 \cdot \left\lceil \frac{74}{30-2} \right\rceil = 74 + 847 \cdot 3 = 2615$$

2. **עלות החישוב לפי** $Hash\ join$**:**

תחילה נבדוק כי התנאי מתקיים:

$$\left\lceil \frac{74}{25-1} \right\rceil = 4 \leq 30$$

לכן ניתן להריץ את האלגוריתם, כעת נחשב את העלות:

$$3 \cdot (74 + 847) = 2763$$

3. נוכל להשתמש באלגוריתם $SORT\ MARGE$ רק אם מתקיים התנאי:

$$\left\lceil \frac{B(R)}{M} \right\rceil < M \ and \ \left\lceil \frac{B(S)}{M} \right\rceil < M$$

נבדוק אם התנאי מתקיים:

$$\left\lceil \frac{74}{30} \right\rceil = 3 < 30, \left\lceil \frac{847}{30} \right\rceil = 29 < 30$$

התנאי מתקיים לכן נוכל להריץ את האלגוריתם (לא ביעילות כי התנאי על החיבור שקטן מהבאר לא מתקיים)

$$5 \cdot (74 + 847) = 4605$$

(ג)

נחשב את גודל החוצץ המינימלי:

1. **גודל החוצץ המינימלי לאלגוריתם** BNL**:**

נרצה שהיחס החיצוני יהיה היחס בקטן יותר לכן נבחר את $conferences$ להיות היחס החיצוני ונחשב לפי הנוסחה:

$$74 + 847 \cdot \left\lceil \frac{74}{M-2} \right\rceil =$$

נשים לב כי כל מספר בלוקים המקיים $M - 2 > 0$ אפשרי, לכן נצטרך **3 בלוקים.**

2. **גודל החוצץ המינימלי לאלגוריתם** $Hash\ join$**:**

התנאי שצריך להתקיים הוא תקיים:

$$\left\lceil \frac{74}{M-1} \right\rceil \leq M - 2 \Rightarrow 74 \leq (M-1)(M-2) \Rightarrow 74 \leq M^2 - 3M + 2 \Rightarrow$$

$$M^2 - 3M - 72 \geq 0$$

קיבלנו כי $M = 10.11$ הוא פתרון אי השוויון, נבדוק עבור $M = 10$

$$\left\lceil \frac{74}{10-1} \right\rceil = 9 > 10 - 2 = 8$$

עבור $M = 11$ **התנאי מתקיים:**

$$\left\lceil \frac{74}{11-1} \right\rceil = 8 \leq 11 - 2 = 9$$

3. **גודל החוצץ המינימלי לאלגוריתם** $SORT\ MARGE$:

$$\left\lceil \frac{B(R)}{M} \right\rceil < M, \left\lceil \frac{B(S)}{M} \right\rceil < M$$

נבדוק מתי התנאי מתקיים על הטבלה הגדולה יותר:

$$\left\lceil \frac{847}{M} \right\rceil < M \Rightarrow 847 < M^2 \Rightarrow \sqrt{847} = \lceil 29.1 \rceil = 30 = M$$

נבדוק עבור $M = 30$:

$$\left\lceil \frac{847}{30} \right\rceil = 29 < M = 30$$

לכן $M = 30$ **הוא המינימלי.**

4. **נבדוק מהו הבאפר המינימלי עבור אלגוריתם** $SORT\ MARGE$ **אופטימלי:**

$$\left\lceil \frac{74}{M} \right\rceil + \left\lceil \frac{847}{M} \right\rceil < M \Rightarrow 74 + 847 < M^2 \Rightarrow \sqrt{921} = 30.34$$

נבדוק עבור $M = 31$:

$$\left\lceil \frac{74}{31} \right\rceil + \left\lceil \frac{847}{31} \right\rceil 3 + 28 = 31 = M$$

נבדוק עבור $M = 32$:

$$\left\lceil \frac{74}{32} \right\rceil + \left\lceil \frac{847}{32} \right\rceil 3 + 28 = 31 < M = 32$$

לכן $M = 32$ **מקיים את הדרוש.**

שאלה 2

נחשב את מספרי השורות ביחסים:

מספר השורות ביחס S**:**

$$200 \cdot 150 = 30,000$$

מספר השורות ביחס R**:**

$$1500 \cdot 60 = 90000$$

(א)

גודל התוצאה בבלוקים של הביטוי $\sigma_{B=6}S(B,C)$ **הוא:**

אנו יודעים כי $V(S,B)=250$ כלומר יש 250 ערכים שונים באטרביוט B נניח התפלגות אחידה ולכן מספר השורות בביטוי:

$$\frac{30000}{250}=120$$

בכל בלוק יש 150 שורות לכן

סה"כ 1 בלוק.

(ב)

גודל התוצא בבלוקים של הביטוי $\sigma_{A<25}R(A,B)$**:**

אנו מניחים התפלגות של שליש ולכן מבספר השורות בתוצאה:

$$\frac{90000}{3}=30,000$$

נמצא את מספר הבלוקים ע"י חלוקה במספר השורות בבלוק:

$$\frac{30,000}{60}=500$$

(ג)

מספר השורות בביטוי $\sigma_{A<25\wedge B=6}(R(A,C)\bowtie S(B,C))$ **שווה ל:**

$$\frac{\frac{30000}{250}\cdot\frac{90000}{3}}{max\left\{250,30,000\right\}}=\frac{120\cdot 30,000}{30,000}=120$$

(ד)

נמצא את האלגוריתם היעיל ביותר לצירוף:

יש לנו שתי אופציות ל BNL ושתי אופציות ל INL (איזה יחס יהיה החיצוני ואיזה הפנימי) ועוד אופציה אחת עבור $hash$.

נבדוק מי היעיל ביותר:

ראשית נוריד את האופציות מ BNL ו INL בהם היחס החיצוני הוא הגדול יותר ונישאר עם האופציה שהיחס הקטן הוא היחס החיצוני:

עבור BNL **עם היחס** S **חיצוני:**

$$cost\ with\ BNL=200+1500\cdot\left\lceil\frac{}{10-2}\right\rceil=$$

נצייר עץ $query\ plan$**:**

(ה)

עלות החישוב היעיל ביותר היא:

$$= 1620$$

שאלה 3

נבצע מספר חישובי עזר:

1. ראשית נשים לב כי מספר הערכים B ביחס R הם 5, נניח התפלגות אחידה ולכן מספר הבלוקים בביטוי $\pi_{A,C}\sigma_{B=15}(R)$ הוא:

תחילה נחשב את מספר השורות בביטוי:

$$T\left(\pi_{A,C}\sigma_{B=15}(R)\right) = \left\lceil \frac{T(R)}{V(R,B)} \right\rceil = \left\lceil \frac{B(R)\left\lceil \frac{1,500}{30} \right\rceil}{5} \right\rceil = \frac{1000 \cdot 50}{5} = 10,000$$

נחלק במספר השורות לבלוק $= 75$, ונקבל:

$$\left\lceil \frac{10,000}{75} \right\rceil = 134$$

2. כמו כן עבור הפעולה $\sigma_{D<4}$ מספר הבלוקים בביטוי $\sigma_{D<4}(S)$ בטבלה S

$$B\left(\sigma_{D<4}(S)\right) = \frac{T\left(\sigma_{D<4}(S)\right)}{75} = \frac{75,000}{75} = 1000$$

3. **נחשב את מספר השורות בטבלה** S**:** מספר הבלוקים ביחס · מספר הבייטים לבלוק , חלקי מספר הבייטים בשורה

$$T(S) = \frac{3000 \cdot 1500}{20} = 225,000$$

(א)

נחשב את מספר השורות בתוצאה:

מכיוון של C הוא מפתח ב R אזי:

$$\frac{T(S)}{5 \cdot 3} = \frac{225,000}{15} = 15,000$$

(ב)

נחשב את גודל התוצאה בבלוקים:

יש לנו 15,000 שורות בתוצאה, גודל כל שורה 20 בייטים (שני אטרביוטים), וגודל כל בלוק 1500 בייטים לכן:

$$\frac{15,000 \cdot 20}{1500} = 200$$

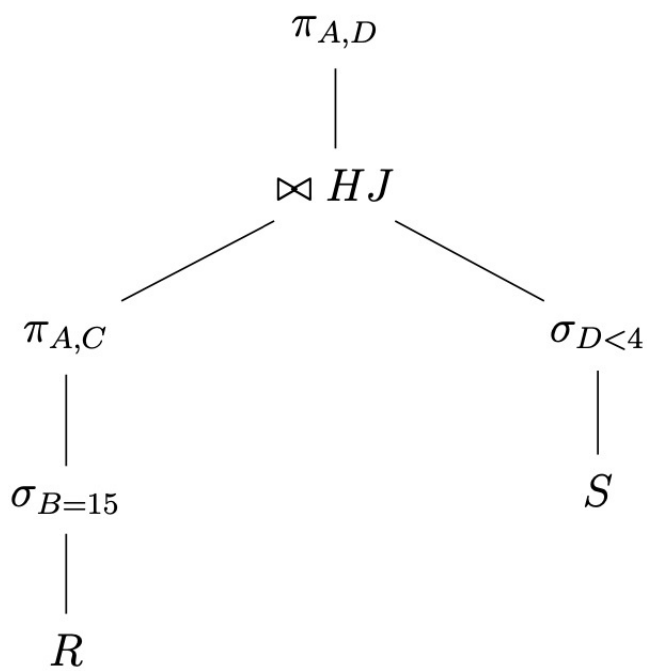
סה"כ 200 בלוקים בתוצאה.

(ג)

נמצא את האלגוריתם ביעיל ביותר:

נשים לב כי אנו האלגוריתמים BNL, INL לא עמדים בתאנאי הסף על גודל הדליים לכן האלגוריתם היעיל הוא HJ.

נצייר את עץ ה $query\ plan$**:**



כשהמעבר $R \Rightarrow \sigma_{B=15}$ והמעבר $S \Rightarrow \sigma_{D<4}$ נעשים עם $full\ table\ scan$.

(ד)

עלות החישוב היעיל ביותר:

מספר השורות ב R, S + פעמיים מספר הבלוקים ב R, S

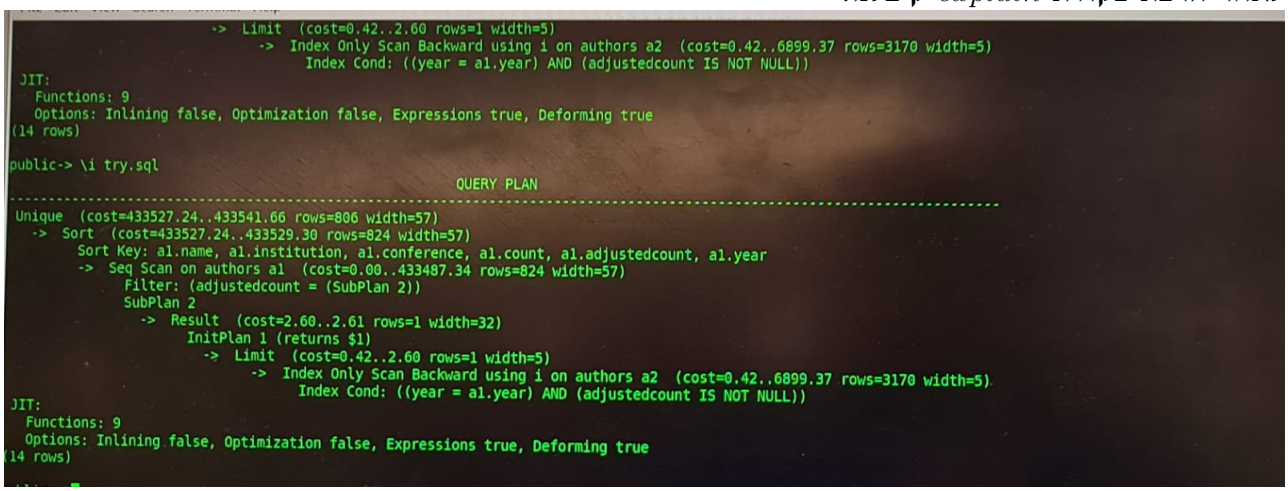
$$1000 + 3000 + 2 \cdot (134 + 1000) = 6264$$

שאלה 4

(א)

קיבלנו $time\ out$ לאחר מספר דקות.

לאחר הרצת פקודת $explain$ **קיבלנו:**



```
                    ->  Limit  (cost=0.42..2.60 rows=1 width=5)
                          ->  Index Only Scan Backward using i on authors a2  (cost=0.42..6899.37 rows=3170 width=5)
                                Index Cond: ((year = a1.year) AND (adjustedcount IS NOT NULL))
JIT:
  Functions: 9
  Options: Inlining false, Optimization false, Expressions true, Deforming true
(14 rows)

public-> \i try.sql
                                                    QUERY PLAN
-----------------------------------------------------------------------------------------------------------------------
Unique  (cost=433527.24..433541.66 rows=806 width=57)
  ->  Sort  (cost=433527.24..433529.30 rows=824 width=57)
        Sort Key: a1.name, a1.institution, a1.conference, a1.count, a1.adjustedcount, a1.year
        ->  Seq Scan on authors a1  (cost=0.00..433487.34 rows=824 width=57)
              Filter: (adjustedcount = (SubPlan 2))
              SubPlan 2
                ->  Result  (cost=2.60..2.61 rows=1 width=32)
                      InitPlan 1 (returns $1)
                        ->  Limit  (cost=0.42..2.60 rows=1 width=5)
                              ->  Index Only Scan Backward using i on authors a2  (cost=0.42..6899.37 rows=3170 width=5)
                                    Index Cond: ((year = a1.year) AND (adjustedcount IS NOT NULL))
JIT:
  Functions: 9
  Options: Inlining false, Optimization false, Expressions true, Deforming true
(14 rows)
```

(ב)

נכתוב שאילתה חדשה:

```
explain analyse
SELECT distinct *
FROM authors
WHERE (year,adjustedcount) IN
( SELECT year, MAX(adjustedcount)
  FROM authors
  GROUP BY year
)
```

הזמן שלקח להריץ את השאילתה החדשה: $02232 + 261ms$

לאחר הרצת פקודת $explain$ **קיבלנו:**

```
                                                    QUERY PLAN
----------------------------------------------------------------------------------------------------------------------------------
 Unique  (cost=5886.68..5893.19 rows=372 width=57) (actual time=260.655..260.859 rows=98 loops=1)
   ->  Sort  (cost=5886.68..5887.61 rows=372 width=57) (actual time=260.653..260.715 rows=98 loops=1)
         Sort Key: authors.name, authors.institution, authors.conference, authors.count, authors.adjustedcount, authors.year
         Sort Method: quicksort  Memory: 38kB
         ->  Nested Loop  (cost=4314.78..5870.80 rows=372 width=57) (actual time=259.792..260.369 rows=98 loops=1)
               ->  HashAggregate  (cost=4314.36..4314.88 rows=52 width=36) (actual time=259.765..259.807 rows=53 loops=1)
                     Group Key: authors_1.year
                     ->  Seq Scan on authors authors_1  (cost=0.00..3490.24 rows=164824 width=9) (actual time=0.012..113.451 rows=164824 loops=1
               ->  Index Scan using i on authors  (cost=0.42..29.84 rows=7 width=57) (actual time=0.005..0.007 rows=2 loops=53)
                     Index Cond: ((year = authors_1.year) AND (adjustedcount = (max(authors_1.adjustedcount))))
 Planning Time: 0.232 ms
 Execution Time: 260.992 ms
(12 rows)

public->
```

מה גרם לשיפור: בשאילתה המקורית השוונו את השנה עבור כל ערך, ובשאילתה המשופרת לקחנו את הערך המקסימלי בלבד - מיינו מראש.

(ג)

האינדקס היעיל ביותר היה האינדקס הבא : $index(year, adjustedcount)$.

זמן הריצה הוא: $7.267ms$

$quary\ plan$ **של השאילתה:**

```
public-> /i drop index i
public-> \i try.sql
psql:try.sql:1: ERROR:  relation "i" already exists
                                                    QUERY PLAN
----------------------------------------------------------------------------------------------------------------------------------
 Seq Scan on authors a1  (cost=0.00..433487.34 rows=824 width=57) (actual time=35.318..1926.931 rows=98 loops=1)
   Filter: (adjustedcount = (SubPlan 2))
   Rows Removed by Filter: 164726
   SubPlan 2
     ->  Result  (cost=2.60..2.61 rows=1 width=32) (actual time=0.009..0.010 rows=1 loops=164824)
           InitPlan 1 (returns $1)
             ->  Limit  (cost=0.42..2.60 rows=1 width=5) (actual time=0.007..0.007 rows=1 loops=164824)
                   ->  Index Only Scan Backward using i on authors a2  (cost=0.42..6899.37 rows=3170 width=5) (actual time=0.005..0.005 rows=1 l
                         Index Cond: ((year = a1.year) AND (adjustedcount IS NOT NULL))
                         Heap Fetches: 164824
 Planning Time: 0.137 ms
 JIT:
   Functions: 9
   Options: Inlining false, Optimization false, Expressions true, Deforming true
   Timing: Generation 0.973 ms, Inlining 0.000 ms, Optimization 0.275 ms, Emission 6.019 ms, Total 7.267 ms
 Execution Time: 1928.069 ms
(16 rows)

public->
```

השינוי בזמן הריצה: נובע מכך שלא עברנו על כל הטבלה, אלא סיננו קודם את כל השדות הרלוונטים ורק אחכ ביצענו את השאילתה.