תרגיל 4 : Join Algorithms

תאריך הגשה: 23:55, 4.06.23.

הוראות הגשה:
בתרגיל זה אתם נדרשים להגיש קובץ zip בודד שיכלול את הקבצים הבאים:

- ex4.pdf עם התשובות מפורטות לשאלות. יש לפרט חישובים לא רק תשובה סופית!
- improved.sql עבור התשובה לשאלה 4 סעיף א חלק 1.
- README שמכיל שורה בודדת ובו ה-login של הסטודנט שמגיש את התרגיל. אם התרגיל מוגש בזוגות, על שורה זאת להכיל את שני ה-login מופרדים בפסיק.

תזכורת: יש להגיש תרגיל מוקלד בלבד.

<u>שאלה 1 (40 נקודות):</u>

נתונים שני היחסים הבאים מתוך מסד נתונים של הכנסת (זהים ליחסים בתרגיל 2):

members (<u>uid</u>, name, occupation, birthPlace, gender, educatedAt, language, birthYear)

memberInKnesset (<u>number</u>, <u>uid</u>, <u>party</u>)

נניח:

- השדות הנומריים uid, birthYear, number תופסים כל אחד 4 בייט.
- השדות הטקסטואליים: name, occupation, birthPlace, educatedAt, language, party תופסים כל אחד 10 בייט.
- השדה gender תופס בייט אחד.
- בטבלה members יש 1,000 שורות.
- בטבלה memberInKnesset יש 3,000 שורות.
- גודל בלוק הוא 256 בייט.
- גודל החוצץ (buffer) הוא 20 בלוקים.

נרצה לחשב עלות של צירוף (join) של הטבלאות $members \bowtie memberInKnesset$.

1. מה תהיה עלות החישוב של הביטוי לפי כל אחד מהאלגוריתמים הבאים?
אם החישוב לא אפשרי, הסבירו למה.

   א. Block-nested-loops?

   ב. Hash-join?

הוסיפו את החישוב של בדיקת התנאי לחישוב האלגוריתם לתשובתכם.

ג. Sort-merge-join?
הוסיפו את החישוב של בדיקת התנאי לחישוב האלגוריתם לתשובתכם, וציינו איזו גרסה של האלגוריתם אפשרית אם בכלל.

2. כעת הניחו שגודל החוצץ הוא 30, איך הייתה משתנה העלות שחישבתם בסעיף 1?

א. Block-nested-loops?

ב. Hash-join?

ג. Sort-merge-join?

3. מה גודל החוצץ המינימלי הנדרש כדי שיהיה ניתן להפעיל כל אחד מהאלגוריתמים?

א. Block-nested-loops?

ב. Hash-join?

ג. Sort-merge-join?

ד. Sort-merge join בשימוש באופטימיזציה שמאפשרת חישוב יעיל יותר (הנמנעת ממיון מלא של היחסים)?

<u>שאלה 2 (25 נקודות):</u>

רוצים לחשב את הביטוי $\sigma_{B=100 \wedge C<45} (R(A,B) \bowtie S(A,C))$.
גודלי היחסים הם B(S)=200 , B(R)=2,000. בכל בלוק של R יש 150 שורות, ובכל בלוק של S יש 30 שורות. ליחס R יש שני אינדקסים עם עלות גישה זניחה: אחד על אטריבוט A ואחד על אטריבוט B. כמו כן, ידוע שA הוא מפתח ביחס R, וכן V(R,B)=20 ,V(S,A)=1,000. בחוצץ (buffer) יש 10 בלוקים.

הערה: הכוונה ב"עלות גישה זניחה" היא שעלות הגישה לאינדקס - הירידה בו וטיול על העלים - זניחה, ולכן עלות השימוש באינדקס הוא שליפה של בלוקים מהטבלה בלבד. זה מתאים מאד למקרה בו מסד הנתונים שומר את מבנה האינדקס בזיכרון המרכזי.

א. העריכו את גודל התוצאה **בבלוקים** של הביטוי $\sigma_{C<45} S(A,C)$

ב. העריכו את גודל התוצאה **בבלוקים** של הביטוי $\sigma_{B=100} R(A,B)$

ג. העריכו את מספר **השורות** בתוצאה של הביטוי כולו $\sigma_{B=100 \wedge C<45} (R(A,B) \bowtie S(A,C))$

ד. מהו האלגוריתם הכי יעיל לחישוב התוצאה? ציירו את עץ ה-query plan.

ה. מה עלות החישוב היעיל ביותר?

שאלה 3 (20 נקודות):

רוצים לחשב את הביטוי $\pi_{A,D}\sigma_{A=30 \wedge D<17}(R(A,B) \bowtie S(B,C,D))$. ההטלה היא ללא מחיקת כפילויות.
גודלי היחסים הם B(S)=3,000, B(R)=6,000 . גודל כל אחד מהאטריבוטים B,A הוא 8 bytes, גודל כל אחד מהאטריבוטים D,C הוא 10, וגודל בלוק הוא 2048 bytes. אין אינדקסים ואסור לבנות אותם.
כמו כן V(R,B)=20,V(R,A)=100 וידוע שB הוא מפתח ביחס S. בחוצץ (buffer) יש 50 בלוקים.

א. מה יהיה מספר השורות בתוצאה?

ב. מה יהיה גודל התוצאה בבלוקים?

ג. מהו האלגוריתם הכי יעיל לחישוב התוצאה? ציירו את עץ ה-query plan.

ד. מה עלות החישוב היעיל ביותר?

שאלה 4 (15 נקודות):

מטרת שאלה זו היא התנסות עם כתיבה יעילה של שאילתות ושימוש באינדקס להתייעלות.

נתון היחס הבא:
(מתאים לטבלה המקורית שקיבלתם בתרגיל 1 לפני שפירקתם אותה לפי הדיאגרמה)

enrollment (country, countrycode, region, incomegroup, iau_id1, eng_name, orig_name, foundedyr, yrclosed, private01, latitude, longitude, phd_granting, divisions, specialized, year, students5_estimated)

רוצים לחשב את השאילתה הבאה המחזירה לכל שנה את האוניברסיטה עם מספר הנרשמים הגדול ביותר:

```
select distinct e1.year, e1.eng_name
from enrollment e1
where  students5_estimated = (select max(students5_estimated)
                              from enrollment e2
                              where e2.year = e1.year)

order by year, eng_name;
```

כאשר מריצים את השאילתה, היא רצה יותר מ5 דקות ובסוף מתקבלת השגיאה הבאה:

ERROR: canceling statement due to statement timeout

(מוזמנים לנסות בעצמכם...)

כאשר הרצנו את השאילתה עם פקודת explain (שבשונה מפקודת explain analyse לא מריצה את השאילתה, רק מציגה את הquery plan) קיבלנו את הפלט הבא:

```
                                    QUERY PLAN
-----------------------------------------------------------------------------------------
 Unique  (cost=738474125.70..738474130.83 rows=668 width=39)
   →  Sort  (cost=738474125.70..738474127.41 rows=683 width=39)
         Sort Key: e1.year, e1.eng_name
         →  Seq Scan on enrollment e1  (cost=0.00..738474093.55 rows=683 width=39)
               Filter: ((students5_estimated)::text = (SubPlan 1))
               SubPlan 1
                 →  Aggregate  (cost=5406.22..5406.23 rows=1 width=32)
                       →  Seq Scan on enrollment e2  (cost=0.00..5383.45 rows=9106 width=4)
                             Filter: ((year)::text = (e1.year)::text)
 JIT:
   Functions: 12
   Options: Inlining true, Optimization true, Expressions true, Deforming true
(12 rows)
```

לצורך מענה על הסעיפים הבאים, יש להשתמש בטבלה enrollment שהוגדרה בתרגיל 1.
(אם מחקתם כבר את הטבלה, בבקשה צרו אותה מחדש וטענו את הנתונים לפי ההוראות בתרגיל 1 שאלה 3 סעיף ג.)

הערה: כדי למדוד זמן ריצה של שאילתה, יש להריץ אותה עם פקודת explain analyse וזמן הריצה המבוקש הוא זמן התכנון + זמן הביצוע.

א. נסו לשפר את זמן הריצה ע"י שינוי בתחביר השאילתה.

1. כתבו את השאילתה החדשה בקובץ בשם improved.sql.
2. הריצו את השאילתה עם פקודת explain analyse, שמראה את הquery plan של השאילתה החדשה, צרפו צילום מסך של התוצאה לתשובות (בדומה לצילום בתחילת השאלה), וכתבו מה זמן הריצה החדש.
3. נסו לשער מה גרם לשיפור בזמן הריצה.
(איך צורך להסביר את כל הפרטים של ה-query plan רק מה גרם לשיפור)

ב. האם אפשר לשפר את זמן הריצה של השאילתה המקורית (לפני השינוי מסעיף ב') ע"י הוספת אינדקס? בדקו אפשרויות שונות.

1. מצאו אינדקס אשר משפר את זמן ריצת השאילתה כך שהיא תרוץ בפחות מ30 שניות. כתבו בתשובה לסעיף זה את הפקודה לבניית האינדקס.
2. בנו את האינדקס והריצו את השאילתה עם פקודת explain analyse, שמראה את הquery plan של השאילתה, צרפו אותה לתשובות, וכתבו את זמן הריצה החדש.
3. נסו להסביר את השינוי בזמן הריצה.
(איך צורך להסביר את כל הפרטים של ה-query plan רק מה גרם לשיפור)

בהצלחה!