תרגיל 4 : Query Processing

תאריך הגשה: 23:55, 30.12.19.

הוראות הגשה:
בתרגיל זה אתם נדרשים להגיש קובץ zip בודד עם השם "ex4.zip". הקובץ יכלול את הקובץ ex4.pdf המכיל את התשובות לשאלות 1-5. כמו כן קובץ הzip יכלול קובץ README עם הפרטים הבאים: IDNumber loginName .

שאלה 1 (20 נקודות):

נתון מסד הנתונים הבא:

Patient (pid, pname, bmi, gender)
Visit (did, pid, vdate, fee)

הנחות:

- גודל בלוק הוא 100 בייטים.
- בטבלה Patient יש 40,000 שורות, בכל בלוק 12 שורות.
- בטבלה Visit יש 500,000 שורות, בכל בלוק 40 שורות.
- קיים אינדקס מסדר 10 על תכונת המפתח pid בטבלה Patient.
- קיים אינדקס מסדר 10 על התכונה fee בטבלה Visit.
- הערכים בfee בטבלה Visit מתפלגים אחיד בטווח [1,1000]

בסעיפים הבאים, בכל סעיף, חשבו את עלות השאילתה עם ובלי שימוש באינדקסים.

א.
```
SELECT DISTINCT "exists"
FROM Visit
WHERE fee > 990
```

ב.
```
SELECT sum(fee)
FROM Visit
WHERE fee > 990
```

ג.
```
SELECT pid
FROM Visit
WHERE fee > 990
```

ד.
```
SELECT pname
FROM Patient
WHERE pid = 111111
```

שאלה 2 (20 נקודות):

נתונה מערכת בעלת המאפיינים הבאים:

- בטבלה R(A,B) יש 2,000,000 שורות. כל בלוק של R מכיל 100 שורות.
- בטבלה S(B,C,D) יש 5,000 שורות, כל בלוק של S מכיל 20 שורות.
- גודל החוצץ (buffer) הוא 102 בלוקים.

נרצה לחשב עלות של צירוף (join) של הטבלאות $R \bowtie S$.

1. מה תהיה עלות החישוב של $R \bowtie S$ לפי כל אחד מהאלגוריתמים הבאים? אם החישוב לא אפשרי, הסבירו למה.

   א. *Block-nested-loops*?

   ב. *Sort-merge-join*?

   ג. Hash-join?

2. כעת הניחי שגודל החוצץ הוא 300, איך הייתה משתנה העלות שחישבת בסעיף 1?

   א. *Block-nested-loops*?

   ב. *Sort-merge-join*?

   ג. Hash-join?

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

   א. *Block-nested-loops*?

   ב. *Sort-merge-join*?

   ג. *Hash-join*?

שאלה 3 (25 נקודות):

רוצים לחשב את הביטוי $\sigma_{A=11 \wedge C<3} (R(A,B) \bowtie S(B,C))$.
גודלי היחסים הם B(S)=300 , B(R)=5,000. בכל בלוק של R יש 10 רשומות, ובכל בלוק של S יש 5 רשומות. ליחס R יש שני אינדקסים עם עלות גישה זניחה: אחד על אטריבוט A ואחד על אטריבוט B. כמו כן, ידוע שB הוא מפתח ביחס R, וכן V(S,B)=20, V(R,A)=100. בחוצץ (buffer) יש 10 בלוקים.

א. תעריכי את גודל התוצאה בבלוקים של הביטוי $\sigma_{C<3} S(B,C)$

ב. תעריכי את גודל התוצאה בבלוקים של הביטוי $\sigma_{A=11} R(A,B)$

ג. תעריכי את מספר השורות בתוצאה של הביטוי כולו $\sigma_{A=11 \wedge C<3} (R(A,B) \bowtie S(B,C))$

ד. מהו האלגוריתם הכי יעיל לחישוב התוצאה? ציירו את עץ הquery plan.

ה. מה עלות החישוב היעיל ביותר?

שאלה 4 (25 נקודות):

רוצים לחשב את הביטוי $\sigma_{A<10 \wedge D<5}\pi_{A,D}\ (R(A,B,\mathrm{C}) \bowtie S(B,D))$. ההטלה היא ללא מחיקת כפילויות. גודלי היחסים הם B(S)=90, B(R)=1,000 . גודל כל אחד מהאטריבוטים הוא 10 bytes וגודל בלוק הוא 3,000 bytes. אין אינדקסים ואסור לבנות אותם. כמו כן, V(S,B)=100 וידוע שB הוא מפתח ביחס R. בחוצץ (buffer) יש 22 בלוקים.

א. מה יהיה מספר הרשומות בתוצאה?

ב. מה יהיה גודל התוצאה בבלוקים?

ג. מהו האלגוריתם הכי יעיל לחישוב התוצאה? ציירו את עץ הquery plan.

ד. מה עלות החישוב היעיל ביותר?

ה. מה תהיה עלות החישוב היעיל ביותר אם B(S)=60.

שאלה 5 (10 נקודות):

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

נתון היחס

People(id, name, phonenumber, city,country,bdate).

ורוצים לחשב את השאילתה הבאה:

```
select distinct *
from People P1
where bdate = (select min(bdate)
                from People P2
               where P2.country = P1.country);
```

לצורך מענה על הסעיפים הבאים, יש לטעון את הנתונים מהקובץ *People.csv* הנמצא באתר הקורס לתוך מסד הנתונים במחשב לפי ההוראות הבאות:

1. היכנסי למסד הנתונים (*psql -h dbcourse public*) והשתמשי בפקודה הבאה ליצירת הטבלה:

```
create table People(
    id integer,
    name varchar,
    phone_number varchar,
    city varchar,
    country varchar,
    job_title varchar,
    bdate date
);
```

2. צאי ממסד הנתונים, והריצי את הפקודה הבאה :

cat People-file-path/people.csv |
psql -hdbcourse public -c "copy People FROM STDIN DELIMITER ',' CSV HEADER"

כאשר *People-file-path* הוא שם התיקייה שבה מיקמת את הקובץ *people.csv*.

3. חזרי לתוך מסד הנתונים.

כעת עני על השאלות הבאות:

א. הריצי את השאילתה. כמה זמן לקח להריץ?
(אם לוקח יותר משתי דקות, אפשר להפסיק את ההרצה ולענות: יותר מ2 דקות).
הריצי פקודת *explain*, שמראה את ה*query plan* של השאילתה וצרפי אותה לתשובות.

ב. נסי לשפר את זמן הריצה ע"י שינוי בתחביר השאילתה.
כתבי את השאילתה החדשה, וכמה זמן לקח להריץ אותה.
הריצי את השאילתה עם פקודת *explain analyse*, שמראה את ה*query plan* של השאילתה החדשה, צרפי אותה לתשובות.
נסי לשער מה גרם לשיפור בזמן הריצה.

ג. האם אפשר לשפר את זמן הריצה ע"י הוספת אינדקס?
בדקי אפשרויות שונות לאינדקס.
כתבי איזה אפשרות של אינדקס שבנית היה הכי יעיל,
כתבי את זמן הריצה החדש, הריצי את השאילתה עם פקודת *explain analyse*, שמראה את ה*query plan* של השאילתה, צרפי אותה לתשובות.
נסי להסביר את השינוי בזמן הריצה.

בהצלחה!