תרגיל 4 : Join Algorithms

תאריך הגשה : 20.12.20 ,23: 55.

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!**
- 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.

שימו לב:

- נא לקרוא על הדרישות המנהליות של הקורס בלינק באתר הקורס כדי למלא אחר ההוראות להגשה של קבצים סרוקים!
- תרגיל מוקלד יזכה ב- 2 נקודות בונוס!

שאלה 1 (40 נקודות):

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

Movies (movieId, title, rating, year, duration, genre)

Actors (actorId, name, byear, dyear)

PlaysIn (movieId, actorId, character)

נניח :

- השדות הנומריים : movieId, rating, year, duration, actorId, byear, dyear תופסים כל אחד 4 בייט.
- השדות הטקסטואליים : title, genre, name, character תופסים כל אחד 10 בייט.
- בטבלה Movies יש 10,000 שורות.
- בטבלה Actors יש 50,000 שורות.
- בטבלה PlaysIn יש 100,000 שורות.
- גודל בלוק הוא 8192 בייט.
- גודל החוצץ (buffer) הוא 15 בלוקים.

נרצה לחשב עלות של צירוף (join) של הטבלאות $Movies \bowtie PlaysIn$.

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*?

ג. *Sort-merge-join*?

2. כעת הניחו שגודל החוצץ הוא 16, 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 לחשב 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 $\sigma_{A<10 \wedge C=8}\ (R(A,B) \bowtie S(B,C))$.
גודלי היחסים הם B(S)=1,000 , B(R)=300. בכל בלוק של R יש 100 שורות, ובכל בלוק של S יש 50 שורות.
ליחס S יש שני אינדקסים עם עלות גישה זניחה: אחד על אטריבוט C ואחד על אטריבוט B. כמו כן, ידוע שB הוא מפתח ביחס S, וכן V(R,B)=100, V(S,C)=200. בחוצץ (buffer) יש 10 בלוקים.

<u>הערה</u>: 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 <u>בבלוקים</u> של הביטוי $\sigma_{C=8}\ S(B,C)$

ב. העריכו את גודל התוצאה <u>בבלוקים</u> של הביטוי $\sigma_{A<10}\ R(A,B)$

ג. העריכו את <u>מספר השורות</u> בתוצאה של הביטוי כולו $\sigma_{A<10 \wedge C=8}\ (R(A,B) \bowtie S(B,C))$

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

ה. מה עלות החישוב היעיל ביותר?

<u>שאלה 3 (20 נקודות):</u>

רוצים לחשב את הביטוי $\pi_{A,D}\sigma_{B=20 \wedge D<5}\ (R(A,B) \bowtie S(A,C,D))$. ההטלה היא ללא מחיקת כפילויות.
גודלי היחסים הם B(S)=1,200, B(R)=4,000 . גודל כל אחד מהאטריבוטים הוא 10 bytes וגודל בלוק הוא

2,000 bytes. אין אינדקסים ואסור לבנות אותם. כמו כן V(S,A)=1000,V(R,B)=10 וידוע שA הוא מפתח ביחס R. בחוצץ (buffer) יש 70 בלוקים.

א. מה יהיה מספר השורות 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?

שאלה 4 (15 נקודות):

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

Movies (movieId, title, rating, year, duration, genre)

רוצים לחשב את השאילתה הבאה:

```
select distinct *
from Movies M1
where duration = (select min(duration)
                  from Movies M2
                  where M2.year = M1.year);
```

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

- היכנסו למסד הנתונים (psql -h dbcourse public) והשתמשו בפקודה הבאה ליצירת הטבלה:

```
create table movies(
    movieId integer primary key,
    title varchar(150) not null,
    rating numeric check (rating>=0 and rating <=10),
    year integer check (year>0),
    duration integer check (duration>0),
    genre varchar(50));
```

הערה: אם עדיין קיימת הטבלה משימוש בתרגילים קודמים, מומלץ למחוק אותה (ואת שאר הטבלאות) באמצעות הקובץ *drop.sql* וליצור מחדש.

- צאו ממסד הנתונים, והריצו את הפקודה הבאה:

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

(כאשר Movies-file-path הוא שם התיקייה שבה מיקמת את הקובץ moviesBig.csv.)

- חזרו לתוך מסד הנתונים.

כעת ענו על השאלות הבאות:

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

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

ב. נסו לשפר את זמן הריצה ע"י שינוי בתחביר השאילתה.
כתבו את השאילתה החדשה, וכמה זמן לקח להריץ אותה.
הריצו את השאילתה עם פקודת explain analyze, שמראה את ה*query plan* של השאילתה החדשה, צרפו אותה לתשובות.
נסו להסביר מה גרם לשיפור בזמן הריצה.

ג. האם אפשר לשפר את זמן הריצה ע"י הוספת אינדקס? בדקו אפשרויות שונות לאינדקס.
כתבו איזה אפשרות של אינדקס שבניתם היה הכי יעיל, ואת זמן הריצה החדש. הריצו את השאילתה עם פקודת *explain analyse*, שמראה את ה*query plan* של השאילתה, צרפו אותה לתשובות.
נסו להסביר את השינוי בזמן הריצה.

בהצלחה!