תרגיל 4

תאריך הגשה : 23:55, 03.07.24.

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

- **ex4.pdf** עם התשובות לשאלות. בכל הסעיפים יש לכתוב הסבר לדרך הפתרון, ולהדגיש את התוצאה הסופית של כל חישוב!
- **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.

בתרגיל זה נשתמש באותה סכמה של טבלאות מתרגיל 2 ובהנחות להלן : (בכל חלק א ובחלק ב שאלה 1)

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

Actors (actorId, name, byear)

PlaysIn (movieId, actorId, character)

הנחות :

- גודל בלוק הוא 1,000 בייטים.
- השדות הנומריים : movieId, rating, year, duration, actorId, byear, תופסים כל אחד 8 בייט.
- השדות הטקסטואליים : title, genre, name, character תופסים כל אחד 10 בייט.
- מצביע תופס 8 בייט.
- הערכים בduration בטבלה Movies מתפלגים אחיד בטווח [1,200]
- הערכים בgenre בטבלה מחולקים ל4 קטגוריות באופן אחיד.
- בטבלה Movies יש 10,000 שורות.
- בטבלה Actors יש 50,000 שורות.
- בטבלה PlaysIn יש 100,000 שורות.
- גודל החוצץ (buffer) הוא 52 בלוקים.

חלק א -אינדקסים

א. נתונה השאילתה הבאה :

```
SELECT avg (duration)
FROM Movies
WHERE duration > 100
```

1. מה עלות חישוב השאילתה בהנחה שאין אינדקסים על הטבלה?

כעת, נתון האינדקס הבא על הטבלה:

```
CREATE index on movies(duration)
```

2. מה תהיה דרגת הפיצול האופטימלית של האינדקס?

3. מה תהיה עלות חישוב השאילתה באמצעות האינדקס, בהנחה שדרגת הפיצול היא זו שחושבה בסעיף הקודם?

ב. נתונה השאילתה הבאה:

```
SELECT avg(duration)
FROM Movies
WHERE genre = 'Drama'
```

ונתון האינדקס הבא על הטבלה:

```
create index on movies(genre)
```

1. מה תהיה דרגת הפיצול האופטימלית של האינדקס?
2. מה תהיה עלות חישוב השאילתה באמצעות האינדקס, בהנחה שדרגת הפיצול היא זו שחושבה בסעיף הקודם?

ג. נתונה אותה שאילתה כמו בסעיף ב, אבל כעת, נתון האינדקס הבא על הטבלה:

```
create index on movies(genre, duration)
```

1. מה תהיה דרגת הפיצול האופטימלית של האינדקס?
2. מה תהיה עלות חישוב השאילתה באמצעות האינדקס, בהנחה שדרגת הפיצול היא זו שחושבה בסעיף הקודם?

חלק ב – Query Plans

שאלה 1 :

בשאלה זו נשתמש שוב בסכמת היחסים ובהנחות המופיעות במלבן הכחול בתחילת התרגיל.

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

א. מה תהיה עלות החישוב של הביטוי בעזרת אלגוריתם *Block-nested-loops*?

ב. מה גודל החוצץ המינימלי הנדרש כדי לקבל את עלות החישוב שקיבלתם בסעיף א?

שאלה 2:

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. ליחס R יש אינדקס על האטריבוט B עם עלות גישה זניחה. כמו כן V(S,A)=1000,V(R,B)=200 וידוע שA הוא מפתח ביחס R. בחוצץ (buffer) יש 70 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. 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 3:

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

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

1. 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.

2. 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*.

(ניתן למצוא את הקובץ גם במערכת המחשבים במעבדה בתיקיה ~ db2/data/ex4/

נרצה לחשב את השאילתה הבאה:

```
select distinct movieId, title, duration
from Movies M1
where duration = (select min(duration)
                  from Movies M2
                  where M2.year = M1.year)
order by movieId, title, duration;
```

כאשר מריצים את השאילתה, היא רצה יותר מ-2 דקות. (מוזמנים לנסות בעצמכם...)

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

```
                                    QUERY PLAN
-----------------------------------------------------------------------------------
 Unique  (cost=4206590.88..4206591.53 rows=37 width=480)
   →  Sort  (cost=4206590.88..4206590.98 rows=37 width=480)
         Sort Key: m1.movieid, m1.title, m1.rating, m1.year, m1.duration, m1.genre
         →  Seq Scan on movies m1  (cost=0.00..4206589.92 rows=37 width=480)
               Filter: (duration = (SubPlan 1))
               SubPlan 1
                 →  Aggregate  (cost=561.69..561.70 rows=1 width=4)
                       →  Seq Scan on movies m2  (cost=0.00..561.60 rows=37 width=4)
                             Filter: (year = m1.year)
 JIT:
   Functions: 10
   Options: Inlining true, Optimization true, Expressions true, Deforming true
(12 rows)
```

כעת ענו על השאלות הבאות:

<u>הערה:</u> 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 ה-*query 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 ה*query 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)

טיפ - כאשר אתם רוצים לנסות אינדקסים שונים מומלץ לתת שם לאינדקס בפקודת היצירה למשל:

```
CREATE INDEX <index_name> ON R (A);
```

כדי שתוכלו למחוק את האינדקס לפני שתנסו ליצור אינדקס חדש. מחיקה מתבצע ע״י הפקודה:

```
DROP INDEX <index_name>;
```

בהצלחה!