תרגיל 3 - מסדי נתונים

מיכל לוי - 313573743 , רועי בן יוסף - 307920116

2 בדצמבר 2020

שאלה 1:

א.

```
EXPLAIN ANALYZE
SELECT DISTINCT P.actorId
FROM PlaysIn P
WHERE character = 'Sheriff'
```

ב.

```
                                                    QUERY PLAN
-----------------------------------------------------------------------------------------------------------------------
 Unique  (cost=634.39..634.64 rows=50 width=4) (actual time=3.850..3.951 rows=44 loops=1)
   ->  Sort  (cost=634.39..634.51 rows=50 width=4) (actual time=3.847..3.879 rows=50 loops=1)
         Sort Key: actorid
         Sort Method: quicksort  Memory: 27kB
         ->  Seq Scan on playsin p  (cost=0.00..632.98 rows=50 width=4)
                                    (actual time=0.662..3.781 rows=50 loops=1)
               Filter: (("character")::text = 'Sheriff'::text)
               Rows Removed by Filter: 32602
 Planning Time: 0.614 ms
 Execution Time: 4.084 ms
(9 rows)
```

פוסטגרס מיין לפי $actorId$ בשיטת $quicksort$. ולאחר מכן הריץ סריקה מלאה על כל איברי הטבלה $playsIn$ למציאת הרשומות בהן $character =' sherrif'$

ג.

```
CREATE INDEX char
ON PlaysIn(character)
```

ד.

השאילתא מיפתה את הערכים לטבלת גיבוב ולאחר מכן יצרה מפת ביטים (כל תא עדי ביטי) , לפי שמות הדמויות, וסרקה לפי האינדקס המתאים לייצוג הביטי של שם הדמות $sheriff$

שאלה 2:

א.

.1
נחשב את מספר הבלוקים הדרושים לכל הטבלה:

$$\left\lfloor \frac{1000}{150} \right\rfloor = 6$$

כלומר 6 שורות בבלוק. לכן אנו צריכים

$$\left\lceil \frac{10000}{6} \right\rceil = 1667$$

בלוקים בסך הכל. מכיוון שנידרש לעבור על כולם (ללא אינדקס) יתבצעו
1667 פעולות I/O
הערה: מכיוון שערכי $duration$ מתפלגים באופן אחיד, נידרש לעבור לכל היותר על חצי מהערכים (כלומר 834) בשביל למצוא ערך $duration$ גדול מ־100 . אך מבדיקה שערכנו במסד הנתונים, ה־$DISTINCT$ לא שינה, ולא שיפר את היעילות.

.2
נחשב את דרגת הפיצול האופטימלית:

$$8 \cdot d + 8(d-1) = 1000$$

$$16d = 1008$$

$$d = 63$$

.3
לחישוב עלות באמצעות אינדקס נחשב תחילה את גובה העץ המקסימלי:

$$\left\lceil log_{\lceil \frac{d}{2} \rceil}(10,000) \right\rceil = \lceil log_{32}(10,000) \rceil = 3$$

שליפת עלה בודד תדרוש פעולת IO נוספת ולכן בסך הכל להרצת השאילתא נידרש ל4 פעולות IO.

ב.

1. בדומה לשאלה הקודמת, גם כאן נסרוק את כל הטבלה ולכן נידרש ל1667 פעולות IO
2. אנו משתמשים באותו אינדקס בדיוק, דרגת הפיצול האופטימלית היא עדיין $d = 63$
3. מכיוון ש$duration$ מתפלג באופן אחיד על $[1, 200]$ ולכן מספר השורות המתאימות לתנאי הוא חצי מכל השורות:

$$\frac{10000}{2} = 5000$$

ובהתאם לנלמד בשיעור, לאחר שהגענו לעלה הראשון המתאים לתנאי ב־3 פעולות IO, נסרוק ימינה את שאר העלים בעלות של:

$$\left\lceil \frac{5000}{\lceil \frac{d}{2} \rceil - 1} \right\rceil = \left\lceil \frac{5000}{31} \right\rceil = 162$$

ולכן בסך הכל נידרש ל $162 + 3 = 165$ פעולות IO

ג.

1. בדומה לשאלה הקודמת, גם כאן נסרוק את כל הטבלה ולכן נידרש ל1667 פעולות IO
2. כיוון ש־$movieId$ ו־$duration$ שניהם בגודל 8 בתים. חישוב d יערך באותה צורה וייתן את אותה תוצאה של $d = 63$
3. גובה העץ יהיה זהה, כלומר 3. ולכן נידרש ל־3 פעולות ע"מ להגיע לעלה הרלוונטי, ועוד פעולה אחת ע"מ לשלוף את הבלוק של הרשומה. כלומר בסך הכל 4 פעולות IO
מכיוון ש־$movieId$ הוא מפתח ראשי, קיים רק אחד כזה שערכו 200, ולכן לא נידרש ליותר מבלוק אחד.

ד.

1. בדומה לשאלה הקודמת, גם כאן נסרוק את כל הטבלה ולכן נידרש ל1667 פעולות IO
2. במקרה זה $genre$ הוא בגודל 10 בתים ולכן נחשב מחדש:

$$8 \cdot d + 10(d - 1) = 1000$$

$$18d = 1010$$

$$\lfloor d \rfloor = 56$$

.3
לחישוב עלות באמצעות אינדקס נחשב תחילה את גובה העץ המקסימלי:

$$\left\lceil log_{\lceil \frac{d}{2} \rceil}(10,000) \right\rceil = \lceil log_{28}(10,000) \rceil = 3$$

מכיוון ש$genre$ מתפלג באופן אחיד על 4 ערכים. מספר השורות המתאימות לתנאי הוא רבע מכל השורות:

$$\frac{10000}{4} = 2500$$

ובהתאם לנלמד בשיעור, לאחר שהגענו לעלה הראשון המתאים לתנאי ב־3 פעולות IO, נסרוק ימינה את שאר העלים בעלות של:

$$\left\lceil \frac{2500}{\lceil \frac{d}{2} \rceil - 1} \right\rceil = \left\lceil \frac{2500}{27} \right\rceil = 93$$

לכן בסך הכל 96 פעולות IO.
בנוסף, שליפת 2500 ערכי $duration$ המתאימים בטבלה המקורית תדרוש מעבר על כל 1667 הבלוקים שבה (שכן הערכים המתאימים לא בהכרח מרוכזים באותם הבלוקים)
לכן בסף הכל נדרש ל 1763 פעולות IO

ה.

1. בדומה לשאלה הקודמת, גם כאן נסרוק את כל הטבלה ולכן נידרש ל1667 פעולות IO
2. במקרה זה $(genre, duration)$ הוא בגודל 18 בתים ולכן נחשב מחדש:

$$8 \cdot d + 18(d-1) = 1000$$

$$26d = 1018$$

$$\lfloor d \rfloor = 39$$

.3
לחישוב עלות באמצעות אינדקס נחשב תחילה את גובה העץ המקסימלי:

$$\left\lceil log_{\lceil \frac{d}{2} \rceil}(10,000) \right\rceil = \lceil log_{20}(10,000) \rceil = 4$$

מכיוון ש$genre$ מתפלג באופן אחיד על 4 ערכים. מספר השורות המתאימות לתנאי הוא רבע מכל השורות:

$$\frac{10000}{4} = 2500$$

ובהתאם לנלמד בשיעור, לאחר שהגענו לעלה הראשון המתאים לתנאי ב־4 פעולות IO, נסרוק ימינה את שאר העלים בעלות של:

$$\left\lceil \frac{2500}{\lceil \frac{d}{2} \rceil - 1} \right\rceil = \left\lceil \frac{2500}{19} \right\rceil = 132$$

לכן בסך הכל 136 פעולות IO.