תרגיל 3 : SQL מתקדם ואינדקסים

תאריך הגשה : 01.05.22 ,23:55.

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

- ex3.pdf עם התשובות לשאלות בחלק ב : אינדקסים.
- q1.sql
- q2.sql
- q3.sql
- q4.sql
- q5.sql
- q6.sql
- 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!

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

authors (name, conference, year, institution, count, adjustedcount)

conferences (conference, area, subarea)

institutions (institution, region, country)

הערות :

- בטבלה של מחברים (authors) יש את המידע על פרסומים של מחברים בכנסים שונים :
  - name – שם המחבר.
  - conference – שם הכנס שבו הוא פרסם.
  - year – השנה שבה פורסם המאמר בכנס.
  - institution – שם המוסד האקדמי של המחבר.
  - count – מספר המאמרים שהמחבר פרסם באותו הכנס.
  - adjustedcount – מספר הפרסומים היחסי של המחבר בכנס. למשל אם פרסם מאמר אחד והיה אחד משני כותבים הספירה היחסית תהיה 0.5. אם היו שלשה כותבים למאמר ...0.33 וכו'.
- בטבלה של הכנסים (conferences) יש את המידע לגבי הכנסים :
  - conference – שם הכנס
  - area – תחום המחקר של הפרסומים בכנס
  - Subarea – תת-תחום המחקר.
- בטבלה של מוסדות (institutions) יש מידע על מוסדות אקדמים :
  - institution – שם המוסד
  - region – אזור גיאוגרפי בעולם.
  - country – המדינה בה נמצא המוסד מיוצגת בתקציר ע"י שני אותיות. למשל ישראל היא il.

באתר הקורס יש קובץ create.sql המכיל הגדרות עבור הטבלאות וקובץ drop.sql המכיל פקודות המוחקות את הטבלאות.
כמו כן, נתונים הקבצים:

- generated-author-info.csv
- conferences.csv
- country-info.csv

הקבצים מכילים מידע על מחברים, פרסומים, כנסים ומוסדות אקדמים. המידע משמש את האתר http://csrankings.org/ לדירוג מוסדות אקדמים בתחום מדעי המחשב. כל הקבצים זהים לאלו שסופקו בתרגיל 2.

את המידע המלא ניתן למצוא בלינקים הבאים:

https://raw.githubusercontent.com/cohensara/csrankings/main/conferences.csv
https://raw.githubusercontent.com/emeryberger/CSrankings/gh-pages/generated-author-info.csv
https://raw.githubusercontent.com/emeryberger/CSrankings/gh-pages/country-info.csv

ניתן למצוא את הקבצים גם במערכת המחשבים במעבדה בתיקיה (הקבצים זהים לאלו של תרגיל 2):

~ db/data/ex2/

ניתן להעתיק אותם לתיקיה שלכם.

על מנת לבדוק את התרגיל שלכם, יש ליצור את הטבלאות בעזרת create.sql, ולטעון לתוכן נתונים בעזרת הפקודות

```
cat generated-author-info.csv | psql -h dbcourse public -c "copy authors from STDIN DELIMITER ',' CSV HEADER"
cat conferences.csv | psql -h dbcourse public -c "copy conferences from STDIN DELIMITER ',' CSV HEADER"
cat country-info.csv | psql -h dbcourse public -c "copy institutions from STDIN DELIMITER ',' CSV HEADER"
```

חלק א: שאילתות SQL (40 נקודות):

כתבו את השאילתות הבאות בSQL. שם הקובץ שבו צריכה להופיע התשובה לכל שאלה נמצא בתחילת השאלה.

בכל התשובות לשאלות בחלק זה:

- השתמשו ב SELECT DISTINCT כדי למנוע כפילויות בתשובות (אם כפילויות עלולות להוויצר בתשובה).
- **שימו לב:** בכל סעיף כתוב באיזה סדר למיין את התוצאות וכן את שמות העמודות בתוצאה.

1. **(q1.sql)** לכל אזור גיאוגרפי החזר את מספר המדינות בו. לעמודת מספר המדינות יש לקרוא בשם countryCount. יש להחזיר טבלה עם העמודות region, countryCount ממויינת לפי region.

2. **(q2.sql)** לכל אזור גיאוגרפי החזר את המספר הממוצע של מוסדות למדינה בו (כלומר אם יש באזור גיאוגרפי מסוים 2 מדינות ו-6 מוסדות נקבל 3 מוסדות בממוצע למדינה באזור זה). לעמודת הממוצע יש לקרוא בשם insAvg. יש להחזיר טבלה עם העמודות region, insAvg ממויינת לפי region.

3. **(q3.sql)** נאמר שמחבר הוא מומחה בתחום systems אם פרסם לפחות שני מאמרים (על פי העמודה count) בכנסים (אחד או יותר) בתחום זה. נאמר שמומחה ל-system הוא עדכני אם הוא פרסם לפחות מאמר אחד בתחום ה-systems החל מ-2014. החזר את שמות כל המומחים העדכניים בתחום ה-systems. יש להחזיר טבלה עם העמודה name ממויינת.

4. **(q4.sql)** לכל אזור גיאוגרפי החזר את המדינה ממנה פורסמו הכי הרבה מאמרים באותו אזור (על פי count), ואת מספר המאמרים שפורסמו בה (אם ישנן כמה מדינות מהם פורסם אותו מספר מקסימלי החזר את כולן). יש להחזיר טבלה עם העמודות region, country, totalCount ממויינת לפי region, ואז country. הצעה – היעזרו ב-with על מנת להגדיר טבלת עזר ובה מספר המאמרים שפורסמו בכל מדינה, והאזור הגיאוגרפי בו היא שוכנת.

5. **(q5.sql)** נאמר שכנס הוא ותיק אם התקיים בלפחות עשר שנים שונות (לאו דווקא ברצף). החזר את שמות המחברים שהשתתפו רק בכנסים ותיקים (גם אם הכנס טרם היה ותיק כשפרסמו בו מאמר). יש להחזיר טבלה עם עמודה בודדת ממויינת של שמות המחברים הנקראת name.

6. **(q6.sql)** נאמר שהמרחק בין שני מחברים, a1 ו-a2, הוא 1 אם שניהם פרסמו מאמר (לפחות אחד) באותו הכנס באותה שנה. בהמשך לכך, אם מחבר a3 פרסם מאמר באותו הכנס ובאותה שנה עם a2, נוכל לומר שהמחברים a1 ו-a3 הם מחברים במרחק 2. כתבו שאילתה רקורסיבית אשר מחזירה את כל המחברים במרחק עד 2 (כולל) מהמחבר "Noam Nisan" . שימו לב, כל מחבר הוא במרחק 0 מעצמו. יש להחזיר טבלה עם עמודה אחת ממויינת בשם name ובה שמות המחברים.

חלק ב: אינדקסים (60 נקודות): (להגשה בכתב בקובץ ex3.pdf)

בחלק זה של התרגיל אנחנו עדיין נשתמש בסכמה המובאת פה שוב לייתר נוחות :

authors (name, conference, year, institution, count, adjustedcount)

conferences (conference, area, subarea)

institutions (institution, region, country)

שאלה 1:

א. כתבו שאילתה בSQL המחזירה את השנים בהם פרסם מחבר מהאוניברסיטה העברית (Hebrew University of Jerusalem) מאמר, בלי כפילויות.

ב. הריצו את השאילתה עם פקודת explain analyse, שמראה את הquery plan של השאילתה, צרפו אותה לתשובות. כתבו כמה זמן לקח להריץ את השאילתה, והסבירו את אופן חישוב השאילתה. אם אתם לא מבינים לגמרי את ה query plan חפשו באינטרנט דוקומנטציה שתעזור לכם להסביר.

ג. כיתבו פקודה אשר תייצר אינדקס **על שדה בודד** שישפר את זמן הריצה של השאילתה.

ד. הריצו את פקודת הבנייה של האינדקס ואת השאילתה עם פקודת explain analyse, שמראה את הquery plan של השאילתה, צרפו אותה לתשובות.
כיתבו כמה זמן לקח להריץ את השאילתה, והסבירו את אופן חישוב השאילתה.
אם אתם לא מבינים לגמרי את ה query plan חפשו באינטרנט דוקומנטציה שתעזור לכם להסביר.

שאלה 2:

בסעיפים הבאים, יש לכתוב הסבר לדרך הפתרון, ולהדגיש את התוצאה הסופית של כל חישוב!

הנחות:

- גודל בלוק הוא 2,000 בייטים.
- בטבלה authors יש 12,000 שורות,
- כל שורה תופסת 180 בייטים.
- התכונה conference תופסת 6 בייט.
- התכונה name תופסת 16 בייט.

- התכונה count תופסת 8 בייט.
- התכונה year תופסת 4 בייט.
- מצביע תופס 8 בייט.
- הערכים ב-count בטבלה authors הם מספרים שלמים המתפלגים אחיד בטווח [1,20]
- הערכים ב-conference בטבלה מחולקים ל80 קטגוריות באופן אחיד.

א. נתונה השאילתה הבאה, אשר הפלט עבורה הוא yes כמספר המחברים שפרסמו שני מאמרים בכנס ובשנה כלשהם:

```
SELECT "yes"
FROM authors
WHERE count = 2
```

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

כעת, נתון האינדקס הבא על הטבלה:

```
CREATE index on authors(count)
```

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 name
FROM authors
WHERE count = 2
```

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

כעת, נתון האינדקס הבא על הטבלה:

```
CREATE index on authors(count)
```

2. מה תהיה דרגת הפיצול האופטימלית של האינדקס?

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

ג. נתונה השאילתה הבאה:

```
SELECT "yes"
FROM authors
WHERE name='x' and conference='y' and year=1999
```

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

כעת, נתון האינדקס הבא על הטבלה :

```
CREATE index on authors(name, conference, year)
```

2. מה תהיה דרגת הפיצול האופטימלית של האינדקס?

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

ד. נתונה השאילתה הבאה:

```
SELECT name
FROM authors
WHERE conference = ‘x’ or conference = 'y'
```

כעת, נתון האינדקס הבא על הטבלה:

```
create index on authors(conference, name)
```

1. מה תהיה דרגת הפיצול האופטימלית של האינדקס?

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

ה. נתונה השאילתה הבאה:

```
SELECT name
FROM authors
WHERE count > 1
```

כעת, נתון האינדקס הבא על הטבלה:

```
create index on authors(count)
```

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

בהצלחה!