

מבחן במערכות בסיסי נתונים

פרופ' טובה מילוא

סמסטר ב' מועד א'

27 ביוני, 2011

סמסטר ב' מועד ב'

12 בספטמבר, 2011

חומר סגור, שלוש שעות

סה"כ 100 נקודות

נא לכתוב את כל התשובות על הדפים
(המבחן כולל 13 דפים כולל דף זה)

מס מחברת: _____ מס סטודנט: _____

חלק	ציון
1	
2	
3	
סה"כ	

1. [40 נקודות]

חברת מזרונ'י ים מחזיקה מסד נתונים עם הסכמה הבאה:

Transactions(tid, date, uid)

Sales(tid, pid, quantity)

Returns(rid, tid, pid)

בכל קניה נפתחת טרנזקציה (מזהה הטרנזקציה, מזהה הלקוח והתאריך נשמרים). בטרנזקציה ניתן לקנות מספר מוצרים ובכמויות שונות. בנוסף המערכת מחזיקה טבלת "החזרות" עבור פריטים שהוחזרו (שימו לב שנשמר מזהה החזרה, וכל החזרה היא עבור פריט בודד).

א. כתבו שאילתת SQL המחזירה את ה UID של הלקוחות ה"טובים" - אלה אשר מעולם לא החזירו מוצר כלשהו.

ב. כתבו את אותה שאילתה באלגברה רלציונית.

ג. לקוח רמאי הוא לקוח אשר החזיר כמות גדולה יותר של מוצרים (של מוצר ספציפי) ממה שקנה ביחס לאותה טרנזקציה. כתבו שאילתת SQL המחזירה את ה UID של הרמאים.

ד. התבוננו בזוגות Q1, Q2 של השאלות הבאות. לכל זוג סמנו מי מהמשפטים הבאים מתקיים:

- (A) שתי השאלות מחזירות תמיד בדיוק אותה תשובה
(B) התשובות המחוזרות ע"י Q1 מוחזרות גם ע"י Q2, אבל יתכן ש Q2 תחזיר ערכים נוספים או כפילויות של ערכים.
(C) התשובות המחוזרות ע"י Q2 מוחזרות גם ע"י Q1, אבל יתכן ש Q1 תחזיר ערכים נוספים או כפילויות של ערכים.
(D) אף אחד מהנ"ל

דוגמה 1:

Q1 = **SELECT * FROM** products **WHERE** cost > 100

Q2 = **SELECT * FROM** products **WHERE** cost > 50

עליכם לענות (B)

דוגמה 2:

Q1 = **SELECT** name **FROM** customers

Q2 = **SELECT DISTINCT** name **FROM** customers

עליכם לענות (C)

i.

Q1 = **SELECT** S1.tid, S2.quantity
FROM Sales S1, Sales S2
WHERE S1.tid = S2.tid

Q2 = **SELECT** S1.tid, S2.quantity
FROM Sales S1, Sales S2
WHERE S1.tid = S2.tid **AND**
S1.pid = S2.pid

(A)

(B)

(C)

(D)

ii.

Q1 = **SELECT** **DISTINCT** S.tid
FROM Sales S, Returns R
WHERE S.quantity = 1 **AND**
S.tid = R.tid **AND**
S.pid = R.pid

Q2 = **SELECT** S.tid
FROM Sales S
WHERE S.quantity = 1 **AND**
S.quantity =
(**SELECT** count(*)
FROM Returns R
WHERE R.tid = S.tid **AND** R.pid = S.pid)

(A)

(B)

(C)

(D)

iii.

Q1 = **SELECT** tid
FROM Transaction
ORDER BY uid

Q2 = **SELECT** tid
FROM Returns
GROUP BY tid, pid

(A)

(B)

(C)

(D)

2. [30 נקודות]

א. נתונה הטבלה $R(A, B, C, D, E, F)$ עם התלויות הפונקציונאליות הבאות:

$$A \rightarrow B$$

$$A, B \rightarrow D$$

$$C, D \rightarrow F$$

$$F \rightarrow A$$

i. מצאו את כל המפתחות (המינימאליים) של טבלה R

ii. פרקו את הטבלה לטבלאות BCNF ונמקו את תשובתכם. התשובה צריכה להכיל:

- רשימת טבלאות והשדות בכל טבלה
- המפתחות בכל אחת מהטבלאות

iii. עבור כל טבלה שקיבלתם (בפרוק הסופי) ציינו האם היא במקרה גם ב 3NF

ב. עבור קבוצה X של שדות נגיד ש X סגורה (ביחס לקבוצה נתונה של קבוצות פונקציונאליות) אם $X + X = X$. שימו לב שמידע שקבוצת שדות מסוימת הינה סגורה נותן לנו גם ידע מסוים על הצורה האפשרית של התלויות הפונקציונאליות הקיימות.

נניח שנתונה טבלה $R(A, B, C, D)$ שקבוצת התלויות הפונקציונאליות שלה אינה ידועה. לכל אחת מהטענות הבאות, החליטו האם הטענה אפשרית: אם כן, תנו קבוצה אפשרית של תלויות פונקציונאליות שעבודה הטענה מתקיימת. אם לא אין צורך ציינו מדוע (לא יותר מ 2 משפטים).

(i) הקבוצות היחידות שסגורות הן $\{ \}, \{A\}, \{C, D\}, \{A, B, C, D\}$

(ii) הקבוצות היחידות שסגורות הן $\{ \}, \{A, B\}, \{B, C, D\}, \{A, B, C, D\}$

3. [30 נקודות]

נתונות שתי טבלאות $R(\underline{A}, B, C)$ ו $S(\underline{D}, E)$ ותוכנית הביצוע הלוגית הבאה

$$P = (\sigma_{B=12}(R)) \bowtie_{C=E} S$$

הערה: שימו לב שאופרטור ה SELECT מתבצע ראשון בתוכנית הביצוע.

אופרטור ה JOIN יכול להיות ממומש בכמה דרכים שונות. התוכניות הבאות מייצגות את ההבדלים הללו:

$P(1) = P$ by "Index nested loop" join

$P(2) = P$ by "Sort-Merge" join

$P(3) = P$ by "Hash-join" join

לדוגמה, $P(2)$ היא מימוש תוכנית הביצוע P על ידי ביצוע Sort-Merge

כאלגוריתם ה JOIN.

אופרטור ה SELECT ממומש בעזרת האינדקס כאשר נוכח, ובקריאה סדרתית כאשר אינו קיים.

א. [12]

עבור כל אחת משלושת התוכניות $\{P(1), P(2), P(3)\}$ הסבירי (בקצרה, שלוש שורות לכל היותר) האם כדאי/צריך להוסיף אינדקס לאחד משדות הטבלה או לא. אם כן, לאיזה שדה והאם כדאי שהוא יהיה CLUSTERED או לא.

(i) עבור $P(1)$:

(ii) עבור $P(2)$:

(iii) עבור $P(3)$:

ב. [16]

נתון המידע הסטטיסטי הבא:

$B(R) = 100$ (מספר הדפים בטבלה R)

$T(R) = 2000$ (מספר הרשומות בטבלה R)

$V(R,B) = 20$ (מספר הערכים השונים בשדה B בטבלה R)

$B(S) = 20$ (מספר הדפים בטבלה S)

$T(S) = 1000$ (מספר הרשומות בטבלה S)

שליפת רשומה בעזרת האינדקס עולה $I(R)=1.2$ ו $I(S)=1.1$ פעולות IO מ R ו S בהתאמה.

נניח שכל אחת מתוכנית הביצוע $\{P(1), P(2), P(3)\}$ מחושבת ב pipeline, כלומר תוצאת אופרטור אחד ניתנת ישירות לאופרטור הבא (בלי להיכתב לדיסק בדרך). האופרטורים ממומשים כפי שהוגדרו בתחילת השאלה.

חשבו את מחיר כל אחת מתוכנית הביצוע במונחים של פעולות I/O (לא כולל כתיבת התוצאה הסופית לדיסק). בטאו כל אחת מהתשובות כפונקציה של הסטטיסטיקות הנ"ל. לדוגמה: $COST(P1) = B(R)B(S)/V(R,A)$ (לא התשובה האמיתית!)

הסבירו כל תשובה בקצרה (2 שורות מקסימום)

(i) עבור $P(1)$ כאשר יש אינדקס על B (clustered) I E .

(ii) עבור $P(1)$ כאשר יש אינדקס על B, C I E .

(iii) עבור $P(3)$ כאשר יש אינדקס על C I E .

(iv) עבור $P(3)$ כאשר יש אינדקס על B (clustered) I (clustered) E .

ג. [2]

ציינו מהי התוכנית (סעיפים i-iv) הזולה ביותר לחישוב השאילתה, ואת מחירה (מבוטא כמספר)