

מבחן במערכות בסיסי נתונים

פרופ' טובה מילוא

סמסטר א' מועד א'
04 בפברואר, 2009
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 משחקי ווידאו מחזיקה מסד נתונים עם הסכמה הבאה :

Customers(uid, name, address, birthday)

Products(pid, name, cost)

Sales(pid, uid, date, quantity)

לכל לקוח יש מזהה, שם, כתובת ותאריך לידה. לכל מוצר יש מזהה, שם ומחיר. כדי לנהל מעקב אחר המכירות, נעזרים בטבלת המכירות: כל מכירה מזהה על ידי המוצר, הלקוח, התאריך והכמות הרצויה (למשל 2 שלטי PlayStation נקנו על ידי דביר ב 01-01-2009)

א. מנהל החנות רוצה להפטר ממלאי המיקרופונים למשחק ה ROCK-BAND הפופולרי. הקונים הפוטנציאליים שלהם אפשר למכור את המיקרופון הם אלא שקנו גם את קונסולת ה XBOX (pid=151) וגם את משחק ה ROCK-BAND (pid=204). עזרו לבעל החנות לכתוב שאילתה שתחזיר את פרטי הלקוחות הרלוונטיים (שם וכתובת).

ב. כתבו את אותה שאילתה באלגברה רלציונית .

ג. מנהל החנות רוצה לברר את הכנסות החנות לפי חתך של שנים. הכנסה של שנה ספציפית היא סכום כל המכירות באותה בשנה , כאשר כל מכירה מחושבת לפי מחיר הפריט כפול כמות הפריטים . עזרו למנהל לכתוב את השאילתה הנחוצה . התשובה צריכה להחזיר זוגות (year, SumOfSales) ממוינים לפי סכום ההכנסות בסדר יורד .

הערה: כדי לבדוד את השנה משדה תאריך , ניתן להשתמש בפונקצית `getYear(date)`. לדוגמה `SELECT DISTINCT getYear(date) FROM Sales` תחזיר את רשימת של כל השנים שבהם נקנו מוצרים .

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** **DISTINCT** C.uid
 FROM Customers C, Sales S
 WHERE S.uid = C.uid **AND**
 C.birthday = '01-01-1982'

Q2 = **SELECT** uid
 FROM Customers
 WHERE birthday = '01-01-1982' **AND**
 uid **IN**
 (**SELECT** C.uid
 FROM Customers C, Sales S
 WHERE S.uid = C.uid)

(A)

(B)

(C)

(D)

ii.

Q1 = **SELECT** pid
FROM Sales
GROUP BY pid
HAVING SUM(quantity) > 10

Q2 = **SELECT** pid
FROM Sales
WHERE quantity > 10

(A)

(B)

(C)

(D)

iii.

Q1 = **SELECT** pid
FROM Sales
GROUP BY pid
HAVING SUM(quantity) > 10

Q2 = **SELECT** **DISTINCT** pid
FROM Sales
WHERE quantity > 10

(A)

(B)

(C)

(D)

2. [30 נקודות]

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

$A \rightarrow B$

$C \rightarrow A$

$D B \rightarrow E$

i. פרקו את הטבלה לטבלאות 3NF ונמקו את תשובתכם. התשובה צריכה להכיל:

- רשימת טבלאות והשדות בכל טבלה
- המפתחות בכל אחת מהטבלאות

ii. עבור כל טבלה שקיבלתם (בפרוק הסופי) ציינו האם היא במקרה גם ב BCNF

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 יד ע 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 , תנו קבוצה אפשרית של תלויות פונקציונאליות שעבודה הטענה מתקיימת .

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

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

3. [30 נקודות]

נתונות שתי טבלאות $R(\underline{A}, B, C, D)$ ו $S(\underline{E}, F)$ ושאלת ה SQL הבאה:

```
SELECT  R.A, S.E, S.F
FROM    R, S
WHERE   R.B = '12' AND R.C = '45' AND R.D = S.E
```

נתונים האינדקסים הבאים:

- על השדה R.A (מסוג unclustered)
- על השדה R.B (מסוג clustered)
- על השדה R.C (מסוג unclustered)
- על השדה R.D (מסוג unclustered)

נניח ששתי הטבלאות גדולות ולא ניתן להחזיק אותן בזיכרון, אבל האינדקסים קטנים וכן יושבים בזיכרון בשלמותם.

עבור השאלתה הר שומה, נתונות תוכניות הביצוע הלוגיות הבאות:

```
P1 = selectB=12( selectC=45( S joinE=D R ) )
P2 = S joinE=D ( selectB=12( selectC=45(R)))
P3 = S joinE=D ( selectC=45( selectB=12(R)))
```

ענו על השאלות הבאות (נמקו בקצרה – שתי שורות לכל היותר!)

א. מבין תוכניות הביצוע $\{P2, P3\}$, האם אפשר לשאר שאחת תסתיים מהר יותר מהשנייה?

ב. האם העברת אינדקס ה cluster משדה R.B לשדה R.C (כלומר, כעת R.B הוא unclustered | R.C הוא clustered) תשפר את ביצועי P1?

ג. נתון המידע הסטטיסטי הבא :

(מספר הדפים בטבלה R)	$B(R) = 100$
(מספר הרשומות בטבלה R)	$T(R) = 2000$
(מספר הערכים השונים בשדה B בטבלה R)	$V(R,B) = 20$
(מספר הערכים השונים בשדה C בטבלה R)	$V(R,C) = 25$
(מספר הדפים בטבלה S)	$B(S) = 20$
(מספר הרשומות בטבלה S)	$T(S) = 1000$
(גודל הזכרון בדפים)	$M=80$

נניח שכל אחת מתוכנית הביצוע $\{P1, P2, P3\}$ מחושבת ב pipeline, כלומר תוצאת אופרטור אחד ניתנת ישירות לאופרטור הבא (בלי להיכתב לדיסק בדרך).

יש שני אופרטורים בסיסים :

- JOIN, נניח שממומש על ידי HASH-JOIN

- SELECT, נניח שכאשר הוא האופרטור הראשון בתוכנית (מבוצע ישירות על הטבלה R), הוא ממומש בעזרת האינדקס. כאשר הוא לא האופרטור הראשון, הוא ניתן לביצוע ישירות על ה tuple שהוא מקבל כקלט (בלי תשלום של I/O).

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)

A. $COST(P1) =$

B. $COST(P2) =$

C. $COST(P3) =$

ד. ציינו מהי התוכנית הזולה ביותר לחישוב השאילתה , ואת מחירה (מבוטא כמספר)