

מבחן במערכות בסיסי נתונים

רובי בויס

סמסטר א' מועד א'
12 בפברואר, 2012

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 סלטי חומוס מחזיקה מסד נתונים עם הסכמה הבאה:

Product (pid, pname, price)

Sales (sid, startdate, enddate, salename)

SaleProducts (sid, pid, discount)

טבלת Products מחזיקה את כל המוצרים ולהם מזהה, שם ומחיר. עבור מבצע המערכת שומרת תאריך התחלה וסיום, שם המבצע, רשימת המוצרים המשתתפים ואחוז ההנחה הניתן לכל מוצר ומוצר. (נתונים אלה נשמרים בטבלאות Sales ו SaleProducts).

הערות:

- מוצר ספציפי יכול להיות בזמן נתון תחת יותר ממבצע אחד, או בלי בכלל.
- כמובן ש sid ו pid בטבלת SaleProducts הם foreign keys (כלומר לא יכולים להופיע שם ערכים שלא קיימים בשתי הטבלאות האחרות).

א. [8] כתבו שאילתת SQL המחזירה את השם והמזהה של כל המוצרים

שאינם נמצאים בהנחה בחודש פברואר 2012 (01.02.2012 –

01.03.2012).

ב. [8] כתבו את אותה שאילתה באלגברה רלציונית.

ג. [9] כתוב שאילתת SQL המחזירה את מספר המוצרים המשתתפים כיום
(12.02.2012) ביותר ממבצע אחד.

ד. [15] התבוננו בזוגות Q1, Q2 של השאילתות הבאות. לכל זוג סמנו מי מהמשפטים הבאים מתקיים:

- (A) שתי השאילתות 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** pid
FROM Products

Q2 = **SELECT** SP.pid
FROM Sales S, SaleProducts SP
WHERE S.sid = SP.sid

(A)

(B)

(C)

(D)

ii.

Q1 = **SELECT** SP.pid,
max(P.price) * max(SP.discount) AS 'real-price'
FROM Products P, SaleProducts SP
WHERE P.pid = SP.pid
GROUP BY SP.pid

Q2 = **SELECT** P.pid, P.price * T.fdiscount AS 'real-price'
FROM Products P,
(**SELECT** pid, max(discount) as 'fdiscount'
FROM SaleProduct
GROUP BY pid) T
WHERE P.pid = T.pid

(A)

(B)

(C)

(D)

iii.

Q1 = **SELECT** S.startdate, avg(SP.discount)
FROM SaleProducts SP, Sales S
WHERE SP.sid = S.sid
GROUP BY S.stardate

Q2 = **SELECT** S.startdate, avg(SP.discount)
FROM SaleProducts SP, Sales S
WHERE SP.sid = S.sid AND
S.enddate < '01.06.2012'
GROUP BY S.stardate

(A)

(B)

(C)

(D)

2. [30 נקודות]

א. [18] נתונה הטבלה $R(A, B, C, D, E)$ עם התלויות הפונקציונאליות הבאות:

$A \rightarrow C$

$B, C \rightarrow D$

$C \rightarrow E$

i. [4] מצאו את המפתח (המינימאלי) של טבלה R

ii. [12] פרקו את הטבלה לטבלאות BCNF ונמקו את תשובתכם.
התשובה צריכה להכיל:

- רשימת טבלאות והשדות בכל טבלה
- המפתחות בכל אחת מהטבלאות

iii. [2] עבור כל טבלה שקיבלתם (בפרוק הסופי) ציינו האם היא במקרה גם ב 3NF

ב. [12] נסמן ב X^+ את הסגור של קבוצת השדות X .
לכל אחת מהטענות הבאות ציינו האם היא נכונה או לא. אין צורך לנמק.

(i) עבור כל שתי קבוצות של שדות X ו Y :
 $(X \cup Y)^+ = X^+ \cup Y^+$

(ii) עבור כל שתי קבוצות של שדות X ו Y :
אם $X \subseteq Y$ אז גם $X^+ \subseteq Y^+$

(iii) עבור כל שתי קבוצות של שדות X ו Y :
אם $X \subset Y$ אז גם $X^+ \subset Y^+$

(iv) עבור כל שתי קבוצות של שדות X ו Y :
אם $X^+ = X$ וגם $Y^+ = Y$ אז גם
 $(X \cap Y)^+ = X \cap Y$

3. [30 נקודות]

א. [5]

עבור כל שתי טבלאות S ו R , אשר גודלן $B(S)$ ו $B(R)$ בהתאמה (בדפים), מה היא דרישת הזיכרון המינימלית לביצוע אלגוריתם Hash-Join בצורה הסטנדרטים כפי שנלמד בכיתה? הסבירי בקצרה (שלוש שורות לכל היותר)

ב. [5]

כאשר לא נתון מידע סטטיסטי על התפלגויות הערכים בטבלה, האם ניתן לשנות את האינדקס מ CLUSTERED ל "רגיל" באלגוריתם Index nested loop join מבלי לפגוע בביצועי המערכת? אם כן מתי, אם לא מדוע. הסבירי בקצרה (שלוש שורות לכל היותר)

ג. [20]

נתונות שלוש טבלאות $R(\underline{A}, B, C)$ $S(\underline{D}, E)$ $Q(\underline{F}, G, H)$ עם האינדקסים הבאים:

- שדה $R.C$ (unclustered)

- שדה $S.E$ (unclustered)

- שדה $Q.G$ (unclustered)

כמו כן נתונות תוכניות הביצוע הלוגיות הבאות:

$$P(1) = \sigma_{H=12} ((R \bowtie_{C=E} S) \bowtie_{B=G} Q) \quad (\text{Sort-Merge Join})$$

$$P(2) = \sigma_{H=12} ((R \bowtie_{C=E} S) \bowtie_{E=G} Q) \quad (\text{Sort-Merge Join})$$

$$P(3) = (\sigma_{B=12} R) \bowtie_{C=E} S \quad (\text{Index-Nested Loop Join})$$

הערות:

- לכל תוכנית מצוין גם אלגוריתם מימוש ה JOIN
- שימו לב שכל תוכניות הביצוע שונות ($P(1)$ ו $P(2)$ נבדלות ב JOIN החיצוני)
- במימוש Index-Nested Loop, הטבלה שמצד שמאל היא בלולאה החיצונית
- הניחו שמיון טבלה (מכל גודל) ניתן למימוש בשני מעברים (מעבר = קריאה + כתיבה), כפי שנלמד בכיתה

נניח שכל הטבלאות גדולות ולא ניתן להחזיק אותן בזיכרון, אבל האינדקסים קטנים וכן יושבים בזיכרון בשלמותם. כמו כן נתון המידע הסטטיסטי הבא:

$B(R) = 100$	(מספר הדפים בטבלה R)
$T(R) = 2000$	(מספר הרשומות בטבלה R)
$V(R,B) = 10$	(מספר הערכים השונים בשדה B בטבלה R)
$V(R,C) = 20$	(מספר הערכים השונים בשדה C בטבלה R)
$B(S) = 20$	(מספר הדפים בטבלה S)
$T(S) = 1000$	(מספר הרשומות בטבלה S)
$V(S,E) = 50$	(מספר הערכים השונים בשדה E בטבלה S)

$$\begin{array}{ll} B(Q) = 20 & \text{(מספר הדפים בטבלה Q)} \\ T(Q) = 1000 & \text{(מספר הרשומות בטבלה Q)} \end{array}$$

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) [6] עבור $P(1)$

עבור $P(2)$ [6] (ii)

עבור $P(3)$ [6] (iii)

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