Design Theory : תרגיל 5

תאריך הגשה : 23:55, 03.01.21.

הוראות הגשה:

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

- ex5.pdf עם התשובות מפורטות לשאלות. **יש לפרט חישובים לא רק תשובה סופית!**
- create.sql המתאים לשאלה 2 סעיף ד.1.
- contradictions.sql המתאים לשאלה 2 סעיף ד.3.
- drop.sql המתאים לשאלה 2 סעיף ד.4.
- 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!

<u>שאלה 1 (35 נקודות)</u>

נחזור וניזכר במידול מידע על האולימפיאדה מהתרגיל בית הראשון. הפעם, במקום למדל בעזרת דיאגרמת ישויות קשרים, נשתמש בגישת תיאוריית התכנון על מנת להבין איך יש להפריד טבלה אחת גדולה לתתי טבלאות.

<u>הערה</u> : בטבלה המקורית של מידע אולימפי היו גם ערכי null. מכיוון שלא דיברנו על טיפול ב null בתיאוריית התכנון, ניתן להניח שכל השדות תמיד מקבלות ערך שאינו null. כמו כן, שינינו מעט את הסכמה על מנת לפשט את השאלה. (הטבלה המקורית אפילו לא הייתה בצורה נורמלית ראשונה.)

נתון היחס athlete_events עם הסכמה הבאה :

athlete_events (ID, Name, Sex, Age, Height, Weight, Team,
NOC, Year, Season, City, Sport, Event, Medal)

נתונה קבוצת התלויות הפונקציונליות הבאה מעל הסכמה של athlete_events :

$$
\begin{aligned}
F = \{ \quad & ID \rightarrow Name, Sex \\
& Year, Season \rightarrow City \\
& ID, Year, Season, City \rightarrow Name, Sex, Age, Height, Weight, Team, NOC \\
& Event \rightarrow Sport \\
& Team \rightarrow NOC \\
& NOC, Year \rightarrow Team \\
& ID, Year, Season, Team, NOC, Event \rightarrow Sport, Medal \quad \}
\end{aligned}
$$

ענו על הסעיפים הבאים :

1. מצאו את כל המפתחות של הטבלה athlete_events. יש רק מפתח אחד ID, Year, Season,Event

2. מה הצורה הנורמלית של הטבלה athlete_events? לא 3NF ולא BCNF

3. נתון פירוק של athlete_events לתתי סכמות :

$R_1 = (\text{ID, Year, Season, Name, Sex, Age, Height, Weight})$
$R_2 = (\text{ID, Year, Season, City, Team, Event, Sport, Medal})$
$R_3 = (\text{Team, NOC})$.

האם הפירוק הוא ללא אובדן ? נמקו! כן

4. מצאו כיסוי מינימאלי ל-F.

$$
\begin{aligned}
F = \{ \quad & \text{ID} \rightarrow \text{Name} \\
& \text{ID} \rightarrow \text{Sex} \\
& \text{Year, Season} \rightarrow \text{City} \\
& \text{ID, Year, Season} \rightarrow \text{Age} \\
& \text{ID, Year, Season} \rightarrow \text{Height} \\
& \text{ID, Year, Season} \rightarrow \text{Weight} \\
& \text{ID, Year, Season} \rightarrow \text{Team} \\
& \text{Event} \rightarrow \text{Sport} \\
& \text{Team} \rightarrow \text{NOC} \\
& \text{NOC, Year} \rightarrow \text{Team} \\
& \text{ID, Year, Season, Event} \rightarrow \text{Medal} \quad \}
\end{aligned}
$$

5. מצאו פירוק של athlete_events ל-$3NF$ על פי האלגוריתם הנלמד בכיתה.
לכל אחד מתת הסכמות בפירוק, כיתבו מה הצורה הנורמלית.

$R_1 = (\text{ID, Name})$ BCNF
$R_2 = (\text{ID, Sex})$ BCNF
$R_3 = (\text{Year, Season, City})$ BCNF
$R_4 = (\text{ID, Year, Season, Age})$ BCNF
$R_5 = (\text{ID, Year, Season, Height})$ BCNF
$R_6 = (\text{ID, Year, Season, Weight})$ BCNF
$R_7 = (\text{ID, Year, Seasn, Team})$ BCNF
$R_8 = (\text{Event, Sport})$ BCNF
$R_9 = (\text{NOC, Year, Team})$ 3NF
$R_{10} = (\text{ID, Year, Season, Event , Medal})$ BCNF

6. מצאו פירוק של athlete_events ל-$BCNF$ על פי האלגוריתם הנלמד בכיתה.

$R_1 = (\text{ID, Name, Sex})$
$R_2 = (\text{Year, Season, City})$

$R_3 = (\text{ID, Year, Season, Age, Height, Weight, Team})$
$R_4 = (\text{Event, Sport})$
$R_5 = (\text{NOC, Team})$
$R_6 = (\text{ID, Year, Season, Event , Medal})$

7. האם הפירוק שמצאתם בסעיף הקודם (ח) משמר תלויות? נמקו.
לא, התלות NOC,Year→ Team לא נשמרת.

שאלה 2 (35 נקודות)

בשיעור למדנו ששמירת נתונים בטבלה בצורה נורמלית גבוהה (BCNF או 3NF) הוא חשוב, על מנת למנוע הכנסה לטבלה של נתונים שאינם עקביים. בשאלה זו אתם תתנסו בהתמודדות עם מידע אמיתי שלא נשמר בצורה נורמלית טובה. כאשר מעוניינים לבצע אנליזה על מאגר מידע נתון, שלב חשוב בתחילת התהליך הוא ניקוי המידע מהשגיאות שנמצאות בו.

לצורך התרגיל, אנחנו נשתמש במידע שנמצא בקישור: Amazon Top 50 Bestselling Books 2009 - 2019 | Kaggle. טבלה זו מכילה מידע על ספרים שהיו רבי מכר ב Amazon בשנים 2009 עד 2019. ניתן גם להוריד את המידע הזה מאתר הקורס. הטבלה מכילה את העמודות הבאים:

- Name, שם הספר.
- Author, שם המחבר.
- User Rating, דירוג הקוראים בAmazon
- Reviews, מספר הביקורות של הספר בAmazon
- Price, מחיר הספר בדולרים
- **Year, השנה בו דורג הספר להיות רב מכר**
- Genre, הז'אנר של הספר (fiction או nonfiction)

לפי התיאור של המאגר, אמורים להתקיים ההנחות הבאות:
1. לספרים שונים יש שמות שונים.
2. לכל ספר יש בדיוק מחבר אחד.
3. כל ספר שייך לז'אנר אחד.
4. הנתונים Rating, Num_Reviews, Price נלקחו מAmazon פעם אחת, בתאריך 13/10/20.
5. ספר יכול להופיע ברשימת רבי המכר ביותר משנה אחת.

ענו על השאלות הבאות:

א. כתבו את קבוצת התלויות הפונקציונליות שאמורות להתקיים בטבלה לפי כל ההנחות הנ"ל.
כתבו את התלויות בצורה אטומית, כלומר שבצד ימין של כל תלות יופיע רק שדה אחד. אין לציין תלויות טריוויאליים.

$$F = \{ \quad \text{Name} \rightarrow \text{Author}$$
$$\text{Name} \rightarrow \text{Rating}$$
$$\text{Name} \rightarrow \text{Num\_Reviews}$$

Name → Price
Name → Genre }

ב. מה המפתח של הטבלה? אם יש מספר מפתחות, ציינו את כולם. המפתח הוא Name,Year
ג. מה הצורה הנורמלית של הטבלה? נמקו. המפתח הוא Name,Year ולכן הצורה הנורמלית היא לא 3NF

ד. בסעיף זה נבחן אילו תלויות פונקציונליות מתקיימות בפועל במופע של הטבלה באתר kaggle ואילו לא מתקיימות. כלומר, אנחנו נגלה את בעיות העקביות של הנתונים. כדי לעשות זאת:

1. כתבו קובץ create.sql שמייצר טבלה בשם **bestsellers** עם כל העמודות הנתונות **וללא** אילוצים בכלל.
2. טענו את הנתונים מהקובץ לתוך הטבלה (בצורה הרגילה, שתוארה בתרגילים קודמים). **שימו לב:** ההכנסה תהיה פשוטה יותר אם לפני כן תורידו את כל הפסיקים שמופיעים בשמות הספרים. בגרסה באתר הקורס הורדנו את הפסיקים עבורכם.
3. כתבו שאילתת SQL בקובץ contradictions.sql שמחזירה את כל השורות המעורבות בסתירה של תלות פונקציונלית. על השאילתה להחזיר רק את העמודות שם ספר, מחבר הספר ושנה, ממוינים לפי שם ספר, ואח״כ שנה, בסדר עולה וכן, להחזיר כל שורה רק פעם אחת.
4. כתבו קובץ drop.sql שמוחק את הטבלה.

ה. אלו תלויות פונקציונליות שכתבתם בסעיף א מתקיימות בנתונים, ואילו תלויות מופרות? תנו פירוק מומלץ של הטבלה לתתי יחסים והסבירו איך שמירת הנתונים בפירוק היה מונע הכנסת שורות לא קונסיסטנטיות.

התלויות המופרות הן:

Name → Rating
Name → Num_Reviews
Name → Price

הפירוק המוצע: R1=(Name, Author, Rating, Reviews, Price, Genre) , R2=(Name, Year)
בפירוק זה, מכיוון שName הוא מפתח בR1, לא יהיה ניתן להכניס מידע שונה לאותו ספר כי לכל ספר תהיה שורה אחת בלבד.

שאלה 3 (30 נקודות)

א. תנו דוגמה פשוטה ליחס, תלויות פונקציונליים ופירוק כך שהפירוק משמר תלויות אך עם אובדן. אם לא קיימת דוגמה כזאת, הסבירו.

R = (A,B,C), F = A->B
R1 = (A,B), R2 = (C)

ב. תנו דוגמה פשוטה ליחס, תלויות פונקציונליים ופירוק כך שהפירוק אינו משמר תלויות אך ללא אובדן. אם לא קיימת דוגמה כזאת, הסבירו.

R = (A,B,C), F = A->B, CB->A
R1 = (A,B), R2 = (A,C)

ג. נתון יחס R=(A, B, C, D, E) וקבוצה F של תלויות פונקציונליים F. ידוע **<u>שצד ימין</u>** של **<u>כל אחד</u>** מהתלויות ב-F הוא בדיוק האטריבט A. הוכח או תן דוגמה נגדית: R ב-BCNF אם ורק אם R ב-3NF.

Correct.

Case 1: All the dependencies in F are trivial. Then R is both in 3NF and in BCNF

Case 2: There is some non-trivial dependency in F. Since B,C,D,E for not appear on the right side of any dependency, they must be in every key. Clearly, A is in the closure of BCDE, due to the non-trivial dependency. Therefore, the only key is BCDE. We have shown already that any relation with a single key is in BCNF if and only if it is in 3NF.