תרגיל 5 : Design Theory

תאריך הגשה : 25.06.23 ,23: 55

הוראות הגשה:

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

- ex5.pdf עם התשובות מפורטות לשאלות.
- create.sql המתאים לשאלה 3 סעיף ד.1.
- contradictions.sql המתאים לשאלה 3 סעיף ד.3.
- drop.sql המתאים לשאלה 3 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.

תזכורת: יש להגיש תרגיל מוקלד בלבד.

שאלה 1 (30 נקודות)

תהי R סכמה עם קבוצת תלויות פונקציונליות F. מוסיפים ל R אטריבוט מבלי לבצע אף שינוי של F.
נתונות הטענות הבאות. לכל טענה תנו דוגמה שמראה היתכנות של הטענה או הוכיחו שהטענה שגויה.

א. טענה : יתכן שלפני הוספת האטריבוט, R היה ב-BCNF וגם לאחר השינוי R נשאר ב BCNF.

ב. טענה :יתכן שלפני הוספת האטריבוט R ,היה לפחות ב-3NF (כלומר ב-3NF או ב-BCNF) ולאחר השינוי R כבר אינו ב-3NF ואינו ב-BCNF

ג. טענה :יתכן שלפני הוספת האטריבוט R ,היה ב-3NF אבל לא ב-BCNF, ולאחר השינוי R ב-BCNF.

שאלה 2 (35 נקודות)

נחזור וניזכר במידול מידע על ההרשמה לאוניברסיטאות 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.

הערה : בטבלה המקורית של מידע ההרשמה היו גם ערכי null. מכיוון שלא דיברנו על טיפול ב null בתיאוריית התכנון, ניתן להניח שכל השדות תמיד מקבלות ערך שאינו null.

נתון היחס enrollment עם הסכמה הבאה : (היחס דומה לטבלה מתרגיל 1 אך הורדנו כמה שדות לשם הפשטות של התרגיל)

enrollment (country, countrycode, region, incomegroup, iau_id1, eng_name, orig_name, foundedyr, latitude, longitude, year, students5_estimated)

נתונה קבוצת התלויות הפונקציונליות הבאה מעל הסכמה של enrollment :

$$enrollment(\overbrace{country}^{x1}, \overbrace{countrycode}^{x2}, \overbrace{region}^{x3}, \overbrace{incomegroup}^{x4}, \overbrace{iau_id1}^{x5}, \overbrace{\text{eng_name}}^{x6}, \underbrace{\text{orig_name}}_{x7}, \underbrace{foundedyr}_{x8}, \underbrace{latitude}_{x9}, \underbrace{longitude}_{x10}, \underbrace{year}_{x11}, \underbrace{\text{student5_estimated}}_{x12})$$

$$
\begin{aligned}
F = \{ \quad & x1 \rightarrow x2, x4 \\
& x2 \rightarrow x1, x3, x4 \\
& x5 \rightarrow x6, x7, x8 \\
& x6 \rightarrow x7, x5, x1 \\
& x7 \rightarrow x6, x9, x10 \\
& x5, x11 \rightarrow x12, x8 \}
\end{aligned}
$$

F = { country → countrycode, incomegroup
countrycode → country, region, incomegroup
iau_id1 → eng_name, orig_name, foundedyr
eng_name → orig_name, iau_id1, country
orig_name → eng_name, latitude, longitude
iau_id1, year → students5_estimated, foundedyr }

ענו על הסעיפים הבאים:

א. מצאו את כל המפתחות של הטבלה enrollment. (אין צורך לפרט את החישובים)

ב. מה הצורה הנורמלית של הטבלה enrollment?

ג. נתון פירוק של enrollment לתתי סכמות:

R_1 = (iau_id1,eng_name,orig_name,country)
R_2 = (iau_id1, year, students5_estimated, foundedyr)
R_3 = (country ,countrycode, region, incomegroup)
R_4 = (eng_name, latitude, longitude)

האם הפירוק הוא ללא אובדן ?
נמקו בעזרת תוצאת האלגוריתם שנלמד בכיתה.

ד. מצאו כיסוי מינימאלי ל-F. (אין צורך לפרט את החישובים)

ה. מצאו פירוק של enrollment ל-$3NF$ על פי האלגוריתם הנלמד בכיתה.
לכל אחד מתת הסכמות בפירוק, כתבו מה הצורה הנורמלית.

ו. מצאו פירוק של enrollment ל-$BCNF$ על פי האלגוריתם הנלמד בכיתה.
פרטו את השלבים של האלגוריתם, את התלות על פיה פירקתם בכל שלב, והדגישו את התשובה הסופית.

ז. האם הפירוק שמצאתם בסעיף הקודם (ו) משמר תלויות?
אם לא, פרטו אילו תלויות אינן נשמרות.

<u>שאלה 3</u> (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 הבאות:

- 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 User_Rating, Reviews, Price נלקחו מAmazon פעם אחת, בתאריך 13/10/20.
5. ספר יכול להופיע ברשימת רבי המכר ביותר משנה אחת.

<u>ענו על השאלות הבאות:</u>

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.

ב. מה המפתח של הטבלה? אם יש מספר מפתחות, ציינו את כולם.

ג. מה הצורה הנורמלית של הטבלה? נמקו.

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?

ו. תנו פירוק מומלץ של הטבלה לתתי יחסים והסבירו איך שמירת הנתונים בפירוק היתה מונעת הכנסת שורות לא קונסיסטנטיות.

בהצלחה!