Design Theory : תרגיל 5

תאריך הגשה : 23:55, 29.05.22.

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</u> (25 נקודות)

נחזור וניזכר במידול מידע על זכיות האוסקר מתרגיל הבית 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 זכיות האוסקר היו גם ערכי null. מכיוון שלא דיברנו על טיפול ב null בתיאוריית התכנון, ניתן להניח שכל השדות תמיד מקבלים ערך שאינו null. כמו כן, ניתן להניח ששדות בלשון יחיד (כמו Name או Studio) מקבלים ערך יחיד.

ענו על הסעיפים הבאים :

א. נתון היחס Oscars עם הסכמה הבאה :

Oscars (ID, Title, Year, Studio, Award, releaseYear, Duration,
Genres, imdbRating, imdbVotes, contentRating, Directors, Authors, Actors)

הסבירו מדוע היחס איננו בצורה נורמלית ראשונה וציינו שתי בעיות שיכולות להיגרם מכך.

ישנן שדות שהן מערכים – במאים, כותבים, ז'אנרים ושחקנים. במצב כזה עלולות להיווצר אנומליות (הכנסה, עדכון, או מחיקה) – כלומר מצבים בהן הדאטה איננו קונסיסטנטי – למשל אותו במאי יכול להופיע בשמות שונים מעט. בנוסף, ישנה כפילות גדולה במידע – אותם שחקנים יופיעו ביחס הרבה פעמים.

כעת הוחלט להסיר את השדות Directors, Authors, Actors, Genres מהסכמה. להלן הסכמה החדשה וקבוצת תלויות פונקציונליות מעליה. **יש להתייחס רק לסכמה זו בכל הסעיפים הבאים.**

Oscars (ID, Title, Year, Studio, Award, releaseYear, Duration, imdbRating, imdbVotes, contentRating)

$F = \{$ ID → Title , Studio, releaseYear, Duration
Title, Studio → Year, Award
Title, Duration, releaseYear → imdbRating, imdbVotes, contentRating
Title, releaseYear → ID, Studio, Year
Title, Year → Duration $\}$

ב. מצאו כיסוי מינימאלי ל-F:

$F = \{$ ID → Title
ID → Studio
ID → releaseYear
Title, Studio → Year
Title, Studio → Award
Title, releaseYear → imdbRating
Title, releaseYear → imdbVotes
Title, releaseYear → contentRating
Title, releaseYear → ID
Title, Year → Duration $\}$

ג. מצאו את כל המפתחות של F מעל היחס Oscars. מהי הצורה הנורמלית של יחס זה?
המפתחות הם:
1. ID
2. Title, releaseYear

הצורה הנורמלית היא לא 3NF ולא BCNF.

ד. נתון פירוק של Oscars לתתי סכמות :

R_1 = (ID, Title, Duration, Award)
R_2 = (Title, releaseYear, imdbRating, imdbVotes, contentRating)
R_3 = (ID, Studio, Year) .

האם הפירוק הוא ללא אובדן ? נמקו! לא. להלן הטבלה בסיום ריצת האלגוריתם:

	ID	Title	Year	Studio	Award	releaseYear	Duration	imdbR	imdbV	ContentR
R1	a1	a2	a3	a4	a5	b36	a7	b38	b39	b310
R2	b21	a2	b23	b24	b25	a6	b27	a8	a9	a10
R3	a1	a2	a3	a4	a5	b36	a7	b38	b39	b310

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

$R_1 = (\text{ID}, \text{Title})$ BCNF
$R_2 = (\text{ID}, \text{Studio})$ BCNF
$R_3 = (\text{ID}, \text{releaseYear})$ BCNF
$R_4 = (\text{Title}, \text{Studio}, \text{Year})$ BCNF
$R_5 = (\text{Title}, \text{Studio}, \text{Award})$ BCNF
$R_6 = (\text{Title}, \text{releaseYear}, \text{imdbRating})$ BCNF
$R_7 = (\text{Title}, \text{releaseYear}, \text{imdbVotes})$ BCNF
$R_8 = (\text{Title}, \text{releaseYear}, \text{contentRating})$ BCNF
$R_9 = (\text{Title}, \text{releaseYear}, \text{ID})$ BCNF
$R_{10} = (\text{Title}, \text{Year}, \text{Duration})$ BCNF

ו. מצאו פירוק של היחס Oscars ל-$BCNF$ על פי האלגוריתם הנלמד בכיתה. האם הפירוק שמצאתם משמר תלויות? נמקו

$R_1 = (\text{Title}, \text{Year}, Duration)$
$R_2 = (\text{Title}, \text{Studio}, \text{Year}, \text{Award})$
$R_3 = (ID, \text{Title}, \text{Studio}, \text{releaseYear}, \text{imdbRating}, \text{imdbVotes}, \text{contentRating})$

אכן הפירוק משמר תלויות, כל אחת מהתלויות נשמרת באחת מתתי הסכמות.

<u>שאלה 2</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 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 במאגר מידע על רכישות מ-kaggle (תוכלו לעיין במאגר המקורי כאן אך אנחנו נעבוד עם מידע מעובד ומצומצם יותר המסופק באתר הקורס). קובץ ה-csv מצורף לתרגיל באתר המודל. טבלה זו מכילה מידע על רכישות בחנות אונליין לונדונית של מתנות בשנים 2018-2019. הטבלה מכילה את העמודות הבאות:

- TranscationNo - מספר מזהה ייחודי של הרכישה, אם לפני המספר מופיעה האות C – המשמעות היא ביטול רכישה קיימת.
- Date - תאריך הרכישה.
- ProductNo – מספר מזהה ייחודי של המוצר.
- ProductName – שם המוצר.
- Price – מחיר ליחידה של המוצר בפאונד.
- Quantity – מספר היחידות מהמוצר ברכישה (מספר שלילי עבור ביטול).
- CustomerNo – מספר מזהה ייחודי ללקוח.
- Country – אזור המגרים של הלקוח.

לפי התיאור של המאגר, אמורים להתקיים ההנחות הבאות:

1. בכל רשומה מתוארת חלק מרכישה (מוצר מסוים בכמות מסוימת מתוך הרכישה).
2. לרכישה יש בדיוק תאריך אחד בו בוצעה, ולקוח אחד שביצע אותה.
3. באותו רכישה יכולים להופיע מספר פעמים אותו המוצר בכמויות ובמחירים שונים.
4. למוצר יש בדיוק שם אחד, אך מחירו יכול להשתנות לאורך זמן.
5. ללקוח יש בדיוק מדינת מגורים 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.

$$F = \{ \begin{array}{l} TranscationNo \rightarrow Date \\ TranscationNo \rightarrow CustomerNo \\ ProductNo \rightarrow ProductName \\ CustomerNo \rightarrow Country \} \end{array}$$

ב. מה המפתח של הטבלה? אם יש מספר מפתחות, ציינו את כולם.
המפתח הוא TransactionNo, ProductNo, Price, Quantity

ג. מה הצורה הנורמלית של הטבלה? נמקו. לא BCNF ולא 3NF – בתלות הראשונה למשל צד שמאל איננו מפתח וצד ימין לא נמצא במפתח.

ד. בסעיף זה נבחן אלו תלויות פונקציונליות מתקיימות בפועל במופע של הטבלה שקיבלתם ואלו לא מתקיימות. כלומר, אתם תגלו את בעיות העקביות של הנתונים. כדי לעשות זאת:
1. כתבו קובץ create.sql שמייצר טבלה בשם **sales** עם כל העמודות הנתונות **וללא** אילוצים בכלל. ניתן, למשל, להגדיר את כל העמודות להיות עמודות של טקסט (varchar).
2. טענו את הנתונים מהקובץ שסופק עם התרגיל לתוך הטבלה (בצורה הרגילה, שתוארה בתרגילים קודמים). **שימו לב:** יש להשתמש רק בקובץ שסופק באתר הקורס.
3. כתבו שאילתת SQL בקובץ contradictions.sql שמחזירה את כל השורות המעורבות בסתירה של תלות פונקציונלית. על השאילתה להחזיר רק את העמודות מספר הרכישה ומספר המוצר, ממוינים לפי מספר הרכישה, ואח"כ מספר המוצר, בסדר עולה. יש למחוק כפילויות.
4. כתבו קובץ drop.sql שמוחק את הטבלה.

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

התלויות המופרות הן:

$$ProductNo \rightarrow ProductName$$
$$CustomerNo \rightarrow Country$$

הפירוק המוצע:

R1=(<u>TranscationNo, ProductNo, Quantity, Price</u>, Date, CustomerNo)
R2=(<u>ProductNo,</u> ProductName)
R2=(<u>CustomerNo,</u> Country)

בפירוק זה, מכיוון שProductNo הוא מפתח בR2, ו-CustomerNo הוא מפתח ב-R3 לא יהיה ניתן להכניס מספר שמות לאותו מוצר או מספר מדינות מגורים לאותו הלקוח. נשים לב שהפירוק משמר תלויות, ללא אובדן, וכן כל היחסים לאחר הפירוק בצורה נורמלית BCNF.

<u>שאלה 3</u> (20 נקודות)

נתונים הסכמה וקבוצת התלויות:

$$R = (A, B, C, D, E)$$
$$F_R = \{CD \to AB, A \to E, E \to D, ACE \to B\}$$

וכן נתון פירוק של R לתתי סכמות:

$$R_1 = (A, C, D, E)$$
$$R_2 = (B, C, D, E)$$

א. כתבו את כל המפתחות של R.

AC,CD,CE

ב. האם R ב-3NF? האם הוא ב-BCNF? נמקו

3NF

ג. האם הפירוק ללא אובדן?

כן

ד. אלו תלויות נשמרות על ידי הפירוק ואלו אינן נשמרות?

כל התלויות נשמרות.

ה. מצאו כיסוי מינימלי של התלויות FR_1. האם R_1 ב-3NF ? האם R_1 ב- BCNF ? נמקו.

הכיסוי המינימלי הוא

$$F_{R1} = \{A \to E, E \to D, CD \to A\}$$

והמפתחות שלו זהים לאלו שבסעיף א' ולכן 3NF.

ו. מצאו כיסוי מינימלי של התלויות FR_2. האם R_2 ב-3NF ? האם R_2 ב- BCNF ? נמקו.

הכיסוי המינימלי הוא

$$F_{R2} = \{E \to D, CD \to B, CD \to E\} \text{ or } F_{R2} = \{E \to D, CE \to B, CD \to E\}$$

ובגלל התלות הראשונה הוא ב-3NF.

שאלה 4 (20 נקודות)

א. נתון יחס R וקבוצת תלויות פונקציונליות F מעל R. נתון שF כיסוי מינימאלי של עצמו, כלומר אין תלויות או אטריביוטים מיותרים ב-F. יהי X→A תלות ב-F. הוכח או תן דוגמה נגדית פשוטה: תת הסכמה X היא בהכרח ב BCNF.

The claim is correct. Suppose, by way of contradiction, that X is not in BCNF. Then, there is a nontrivial dependency Y→Z in F_X such that Y is not a key of the subschema X. Observe that both Y and Z are contained in X, by definition. Also, there is some attribute in Z that is not in Y (since the dependency is nontrivial). Consider the set X' = X- Z U Y. Clearly X' is strictly contained in X. In addition, Z is in the closure of X'. Hence, (X')+ = X. Therefore, the dependency X'→A follows from F. However, this contradicts the minimality of X→A.

ב. נתון יחס R וקבוצת תלויות פונקציונליות F מעל R. נתון שF כיסוי מינימאלי של עצמו. יהי X→A תלות ב-F. הוכח או תן דוגמה נגדית פשוטה: תת הסכמה XA היא בהכרח ב BCNF.

The claim is incorrect. Suppose R = (A,B,C), and F = {BC→A, A→B}. Clearly, F is a minimal cover. The subschema defined by the first dependency BC→A is not in BCNF (but rather in 3NF).