

תרגיל 5 : Design Theory

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!

שאלה 1 (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.

הערה: 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.

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)

$F = \{ \text{ID} \rightarrow \text{Title, Studio, releaseYear, Duration}$
 $\text{Title, Studio} \rightarrow \text{Year, Award}$

Title, Duration, releaseYear \rightarrow imdbRating, imdbVotes, contentRating
 Title, releaseYear \rightarrow ID, Studio, Year
 Title, Year \rightarrow Duration }

- ב. מצאו כיסוי מינימאלי ל-F.
- ג. מצאו את כל המפתחות של F מעל היחס Oscars. מהי הצורה הנורמלית של יחס זה?
- 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!

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

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

שאלה 2 (35 נקודות)

בשיעור למדנו ששמירת נתונים בטבלה בצורה נורמלית גבוהה (3NF או BCNF) 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.

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.
- ב. מה המפתח של הטבלה? אם יש מספר מפתחות, ציינו את כולם.
- ג. מה הצורה הנורמלית של הטבלה? נמקו.
- 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 ו**ללא** 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.

שאלה 3 (20 נקודות)

נתונים הסכמה וקבוצת התלויות:

$$R = (A, B, C, D, E)$$

$$F_R = \{CD \rightarrow AB, A \rightarrow E, E \rightarrow D, ACE \rightarrow B\}$$

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.
- ב. האם R ב-3NF? האם הוא ב-BCNF? נמקו
- ג. האם הפירוק ללא אובדן?
- ד. אלו תלויות נשמרות על ידי הפירוק ואלו אינן נשמרות?
- ה. מצאו כיסוי מינימלי של התלויות FR_1 . האם R_1 ב-3NF? האם R_1 ב-BCNF? נמקו.
- ו. מצאו כיסוי מינימלי של התלויות FR_2 . האם R_2 ב-3NF? האם R_2 ב-BCNF? נמקו.

שאלה 4 (20 נקודות)

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