שאלה 2 – תקינות נתונים (2025, מועד ב׳)

שדה **bank_to** בטבלת **order** הוא מסוג **varchar**. האם כדאי להחליפו בשדה המפוענח כ **key foreign** מטבלה חדשה של בנקים? שימו לב כי גם בטבלת **transactions** יש שדה **bank**. ציינו יתרונות וחסרונות ליצוג הנוכחי וליצוג בטבלת בנקים. אנא התייחסו ל**:**

1. חשיבות השדה

השדה מזהה בנק יעד להעברות/חיובים, מה שהופך אותו לבעל חשיבות גבוהה ביותר.

2. משמעות הטעות ואופן הטיפול בה

שם של בנק שאינו קיים ימנע את ביצוע הפעולה הבנקאית באופן תקין.

3. אפשרויות אחרות למנוע שגיאות

מכיוון שרשימת הערכים התקינים לא ידוע לנו ולא יכולה להיות ידועה לנו, עלינו להשתמש במנגנון חיצוני.

4. סט הערכים האפשרי

הערכים האפשריים התקינים הם בנקים בעולם. כאשר משתמשים בטקסט חופשי, כל ערך אפשרי, כמו שגיאות כתיב, שיובילו לערכים לא תקינים. בנוסף, הופעת השדה בשתי טבלאות מובילה לדרישת תאימות.

הבעיה היא שלנו לא ידועים השמות כל הבנקים בעולם (לדוגמה, בנקים במיקרונזיה, בנק חדש שיוקם עוד שנתיים).

הפתרון למצב נובע מכך שזו בעיה המשותפת לכלל הבנקים ועדיין המערכת הבנקאית מתפקדת וכסף לא עובר לבנקים לא קיימים.

לא ידוע לנו כיצד פתרון זה פועל. אם מסופק רק API מולו ניתן לבדוק שם בנק, אפשר להאיר יצוג במחרוזת ולבדוק תקינות. אם ניתן לקבל את רשימת הבנקים רצוי להשתמש בטבלת בנקים, מהשיקולים הרגילים. יש לשים לב שבנקים נפתחים ונסגרים כך שיש לדאוג למנגנון עידכון.

---

b. קשר בין **client** ל־**loan**

i. סוג הקשר בין **loan** ל־**client**
הקשר בין **loan** ל־**client** הוא דרך **account** ו־**disp**. ההלוואה משויכת לחשבון (**loan.account_id**),

והחשבון משויך ללקוחות דרך טבלת **disp** (שם מוגדר סוג הקשר, למשל **OWNER**). מבחינת הסכמה, זהו קשר רבים־לרבים: ללקוח יכולים להיות כמה חשבונות, וכל חשבון יכול להיות שייך לכמה לקוחות. בפועל, במערכת בנקאית רוב החשבונות שייכים ללקוח יחיד, אבל הסכמה מאפשרת גם בעלות משותפת ולכן הקשר מתואר כ־**N:N**.

ii. שליפת סכום ההלוואות של כל לקוח
כדי לחשב את סכום ההלוואות של כל לקוח, יש לעבור מהלקוח אל החשבון דרך **disp**, ומשם אל ההלוואות. השאילתה:

```
SELECT c.client_id, SUM(l.amount) AS total_loans

FROM client c

JOIN disp d ON c.client_id = d.client_id

JOIN account a ON d.account_id = a.account_id

JOIN loan l ON a.account_id = l.account_id

GROUP BY c.client_id;
```

שאילתה זו מקבצת לפי מזהה הלקוח ומחזירה לכל לקוח את סכום ההלוואות של כלל החשבונות שלו.

iii. ייצוג הלוואות ישירות של לקוח
אופציה אחרת היא להוסיף לטבלת loan עמודה client_id שמפנה ישירות לטבלת client. במבנה כזה כל הלוואה משויכת ללקוח בלי צורך לעבור דרך החשבון.

iv. יתרון של הייצוג הנוכחי
הייצוג הנוכחי מונע אי־תאימות בין נתונים. אם הלקוח היה נשמר גם בהלוואה עצמה וגם בחשבון לו היא מקושרת, היה יכול להיות שיוך ללקוח אחר בכל אחד מהם.

---

c.

`account.frequency` הוא VARCHAR

i. מה לדעתכם משמעות השדה? (כל תשובה קבילה עם נימוק)

בפועל, השדה מכיל את תדירות החיוב של הלקוח, כתוב בצ'כית.

כמובן שהיה קשה לדעת זאת.

שימו לב שתשובה שמייחסת לשדה את תדירות הפעילות מניחה שבוני הסכמה יצרו כפילות שיכולה להוביל לאי תאימות והתעלמות מכך שתדירות פעולות לקוח לרוב משתנה כך שלא ניתן ליצגה בערך קבוע.

ii. מה לדעתכם הערכים בשדה?

יצוג אפשרי במחרוזת הוא יחידת זמן בה פעולה מתבצעת, ההופכי של תדירות. כלומר, "שבועית", "חודשית" וכדומה.

iii. כיצד הייתם מייצגים כדי לחשב ממוצעי תדירויות?

מכיוון שלא ניתן לחשב ממוצעים על "שבועית" וכדומה, היה ניתן ליצג במספר את פרק הזמן העובר בין פעלות (7, 30).

iv. כיצד הייתם מייצגים כדי לאפשר תדירות שעתית, יומית ושבועית

בסעיף זה נרצה גם להמנע מתדירות של 4.5 וגם ליחס משמעות סמנטית לערכים. ניתן לבצע זאת על ידי טבלת תדירויות ולפענח ממנה את המשמעות הסמנטית.

v. כיצד הייתם מייצגים כדי לאפשר את שני הסעיפים הקודמים יחד

נשמור טבלה של ערכי תדירות אפשרים. בטבלה נשמור שם סמנטי ואת פרק הזמן ליחידה, עליו נחשב ממוצעים.

---

d. בונוס – GS זוגות סרטים וציון

i. האם צריך להזין גם **AB** וגם **BA?**

אם הקשר בין זוגות סימטרי תמיד,שמירה לא סימטרית תכיל את כל המידע. במקרה כזה מספיק לשמור כיוון אחד בלבד. יתכן ונרצה לשמור את שני הכיוונים כדי להקל על השימוש (לדוגמה, לבצע join בלי התאמה לשתי אפשרויות היצוג). אם לא כל האפשריות סימטריות, נוכל להניח את הסימטריה כברירת המחדל ולשמור רק חריגות.

---

ii. בזוגות לא־סימטריים (כמו סדרת סרטי המשך) – האם צריך להזין את שני הכיוונים?

כן, נהיה חייבים להזין את שני הכיוונים. ללא הזנתם לא נדע באיזה כיוון ההמלצה הטובה ובאיזה כיוון ההמלצה הלא טובה.

Commented [1]: נתת הסבר למה תדירות החיוב, אבל בפועל ענית על הסעיף ב2.

Commented [2]: הייתי מנסח - אם בחרתם שמשמעות הייצוג תהיה תדירות הפעילות עליכם לקחת בחשבון שהיא עשויה להשתנות ושהיא עלולה להוביל לכפילות נתונים ושזו הנחה שיוצרי הסכמה לקחו בחשבון.

Commented [3]: אני התייחסתי לזה בתור דרגות של פעילות - פעיל מאוד - לא פעיל וכדומה, תדירות שבועית חודשית זה פתח לבעיות, כי אתה לא משווה בין כל הלקוחות לפי אותה יחידת זמן. אם כבר לציין את היחידה השם השדה - נגיד תדירות שבועית - וערך כלשהו. הצעה בלבד

Commented [4]: תדירות של 4.5 - לא ברור למה התכוונת. הייתי אומר שהיינו רוצים להימנע מתדירות שאינה שלמה, ולהחזיק בצד טבלת ערכים שלפי נגדיר את המשמעות הסמנטית של תדירות לדוגמה (4.5 נחשבת תדירות גבוהה). תוספת - אפשר ע"י נרמול של כל התדירויות הקיימות (הלקוח הכי תדיר הוא 10 או 100) או על פי קביעת סף (נגיד תדירות של 10 נחשבת גבוהה).

Commented [5]: התשובה מנוסחת בצורה תמציתית וברורה, אך ייתכן שכדאי להרחיב מעט ולהדגים איך טבלת ערכי התדירות תתחבר בפועל לנתוני המשתמשים. בנוסף, ליישום אמיתי אפשר לשקול גם עם חותמות זמן של כניסות. כך (log) טבלת פעולות ניתן יהיה לחשב ישירות את משך הקיום והתדירות בפועל, בעוד שטבלת יחידות הזמן תשמש כמילון אחיד של פרקי זמן. החיבור בין שתי הטבלאות מבהיר את ההיבט התאורטי וגם את היישום המעשי.

iii. הצעה לאיתור זוגות "שוברי סימטריה" ומימוש

בהנתן ה gold standard ניתן לאחד את הכיוון הכיוון ההפוך לו ולמצוא המלצות סותרות.

אם נרצה למצוא זוגות כאלו שלא תויגו, נוכל לחפש צמדים קשורים (לדוגמה, סרטים עם אותו במאי ושחקנים) בהם יש סרט אחד אהוב וסרט לא אהוב. במקרה כזה סביר שהמלצה מהלא אהוב לאהוב תהיה טובה ובכיוון השני רעה.

מידע על מידת אהבה לסרט מופיע כאגגרגציה ב rank ובשיוך למדרג ב collaborative filtering.

כדי לזהות זוגות שבהם כיוון אחד חזק והשני חלש, אפשר לבדוק מקרים שבהם ציון ההמלצה בכיוון AB גבוה, אך בכיוון BA נמוך (או להפך).

לדוגמה:

```
SELECT m1.a_id, m1.b_id, m1.score AS high_score, m2.score AS low_score

FROM movies_reccomendations AS m1

JOIN movies_reccomendations AS m2

  ON m2.a_id = m1.b_id

 AND m2.b_id = m1.a_id

WHERE m1.score >= 8   -- ציון גבוה בכיוון אחד

  AND m2.score <= 3;  -- ציון נמוך בכיוון ההפוך
```

כך אנחנו מזהים זוגות שבהם אחד הכיוונים מקבל המלצה חיובית והשני שלילית, כלומר חוסר סימטריה אמיתי.

---

iv. מה החשיבות של זוגות אלו?

בהמלצות אישיות יש כמות גבוהה מאד של Null records, צמדים שאינם המלצות טובות ושלרוב לא ינתנו כהמלצות טובות.

זוגות בהן אין סימטריה הם כאלו בהם יש קשר חזק (מספיק לכיוון אחד) אך עדיין לא טובות בכיוון השני.

ככאלה, זוגות שוברי סימטריה הם מקרי בדיקה טובים ובנוסף כופים על המודל לייצג קשרים לא סימטרים.