ניתוח איומי תקינות נתונים – יולי 2026

סכמת אירועי ירי משטרתיים : officers, subjects, incidents

officers

case_number	varchar
race	char
gender	char
last_name	varchar
first_name	varchar
full_name	varchar

subjects

case_number	varchar
race	char
gender	char
last_name	varchar
first_name	varchar
full_name	varchar

incidents

case_number	varchar
date	date
location	varchar
subject_statuses	varchar
subject_weapon	varchar
subjects	varchar
subject_count	int
officers	varchar
officer_count	int
grand_jury_disposition	varchar
attorney_general_forms_url	varchar
summary_url	varchar
summary_text	varchar
latitude	double
longitude	double

Commented [1]: עבודה יפה ומקיפה, כתבתי הערות

Commented [2]: הוסף את מועד המבחן

בסכמה קיימות שלוש טבלאות: incidents (אירועי ירי), officers (שוטרים המעורבים באירוע) ו-subjects (המעורבים שנורו).
על כל שדה נזהה את האיום, כיצד הוא יתבטא בנתונים, מה תהיה ההשפעה, ומה הפתרון.

לפני הירידה לרמת השדה הבודד נציג מספר איומים החוצים את כל הסכמה. איומים אלו חמורים במיוחד משום שאינם נובעים משדה מסוים אלא ממבנה הסכמה עצמו, ולכן לא ניתן לפתור אותם בהוספת אילוץ נקודתי. הבעיה המרכזית שבהם, שממנה נגזרות רבות מן האחרות, היא שאין בסכמה ייצוג עצמאי לאנשים.

חלק א': איומים ברמת הסכמה

היעדר ייצוג עצמאי לאדם

איום: השוטר והמעורב הם ישויות יסוד בסכמה, ואף על פי כן אין להם ייצוג עצמאי. הם מופיעים אך ורק בהקשר של תיק מסוים, כשורה בטבלת officers או בטבלת subjects, ואין בסכמה שום מקום שבו מאוחסנים פרטיו של אדם כשלעצמו. זהו האיום המרכזי בסכמה, ורוב הבעיות שיפורטו בהמשך נגזרות ממנו.

השפעה: לא ניתן לייצג אדם ללא אירוע, ולכן פרטיו האישיים משוכפלים בכל אירוע שבו הוא מופיע. שוטר המעורב בשלושה אירועים נרשם בשלוש שורות נפרדות שאין ביניהן שום קשר פורמלי, ואם פרטיו מתעדכנים בשורה אחת בלבד נוצרות במסד שתי גרסאות סותרות של אותו אדם. יש להדגיש שאין כאן הפרה של כלל נורמליזציה. כללי הנורמליזציה עוסקים בתלויות בין שדות בתוך טבלה, ולא בשאלה אילו ישויות ראוי לייצג במערכת.

הפתרון: הפרדת פרטי האדם לטבלה עצמאית בעלת מפתח ראשי משלה, ושימוש במפתח זר אליה מתוך הטבלאות המתארות את השתתפותו בתיק. קיימת כאן דילמה עיצובית: אפשר לרכז את כל האנשים בטבלה אחת, ואפשר להפריד בין שוטרים למעורבים. מאחר שהמשמעות של שוטר ושל מעורב שונה, ומאחר שסביר שנרצה לשמור על כל אחד מהם פרטים אחרים, לדוגמה דרגה ויחידה עבור שוטר לעומת מצב ונשק עבור מעורב, יש עדיפות להפרדה לשתי ישויות. מנגד, אילו לאדם עצמו לא הייתה חשיבות במערכת, היה היגיון בשמירת שמו בלבד, כפי שנהוג לגבי שם מדינה. עצם העובדה שהסכמה שומרת מין וגזע מלמדת שלא זה המקרה כאן.

היעדר מזהה ייחודי לאדם

איום: בהמשך לסעיף הקודם, בטבלאות officers ו-subjects אין שום שדה המזהה אדם מסוים. הזיהוי היחיד הזמין הוא השם, ובפרט השדה full_name, ושם אינו מזהה ייחודי. אין בסכמה מספר תעודת זהות או מזהה מנהלי אחר.

השפעה: מעקב אחר אדם בין אירועים אפשרי בפועל על בסיס השם, אך הוא נשבר דווקא במקרים המעניינים: שני שוטרים שונים בעלי אותו שם ימוזגו לאדם אחד, ואילו אותו אדם שנרשם בשני כתיבים שונים יתפצל לשניים. מכיוון שהשאלה המחקרית המרכזית בדאטהסט מסוג זה היא האם קיימים שוטרים החוזרים על עצמם במספר אירועי ירי, מדובר בפגיעה ישירה בשימוש העיקרי בנתונים. יש לציין שהיעדר תעודת זהות אינו בהכרח מחדל תכנוני, אלא נובע לעיתים מאופי העבודה המשטרתית: כאשר מאתרים גופה לא תמיד ידועה זהותה, ולעיתים גם שמה אינו ידוע.

הפתרון: הוספת מזהה טכני ייחודי לכל אדם בטבלת האנשים העצמאית שתוארה לעיל. יש להבחין בין שתי בעיות נפרדות: מניעת שורות כפולות זהות באותו אירוע היא בעיה אחת, ואותה אפשר לאכוף גם ללא מזהה, באמצעות אילוץ ייחודיות על צירוף מספר התיק והשם. מעקב אחר אדם בין אירועים הוא בעיה אחרת, והוא מחייב מזהה יציב. עד ליישום השינוי המבני ניתן לאתר חשד לערבוב בין אנשים שונים בעלי אותו שם באמצעות שאילתת בקרה המחפשת שם מלא זהה המופיע עם מין או גזע שונים:

```
SELECT a.full_name, a.case_number, b.case_number FROM officers AS a
JOIN officers AS b ON a.full_name = b.full_name WHERE a.race <>
b.race OR a.gender <> b.gender
```

היעדר הגדרה מפורשת של מפתחות ושל טבלאות קוד

איום: הסכמה שלפנינו מוצגת כתרשים המפרט שמות שדות וטיפוסים בלבד, ואינה כוללת הגדרת מפתחות. לכן לא ניתן לדעת בוודאות אילו אילוצים מוגדרים בפועל, וזו נקודה שיש לברר מול בעלי המערכת ולא לקבוע אותה

כבעיה ידועה. עם זאת סביר להניח שהשדה case_number משמש כמפתח הראשי של טבלת האירועים, והחיצים בתרשים מייצגים קשר לוגי שלא ברור אם הוא נאכף כמפתח זר. לעומת זאת, עובדה אחת ניתנת לצפייה ישירה בתרשים: אין בסכמה טבלאות קוד למין ולגזע, ולכן ממילא אין מנגנון האוכף סט ערכים חוקי בשדות אלו.

השפעה: אם אכן אין מפתח ראשי, אותו אירוע עלול להופיע פעמיים בטבלת האירועים עם אותו מספר תיק, וכל ספירה סטטיסטית תהיה מנופחת. אם אין מפתח זר, ניתן להכניס שורת שוטר עם מספר תיק שאינו קיים בטבלת האירועים, ותיווצר רשומה יתומה שתיעלם מכל שאילתה המבצעת JOIN, כך שהבעיה תישאר סמויה: הנתון קיים במסד אך אינו נספר בשום דוח. היעדר טבלאות הקוד מאפשר להזין בשדות המין והגזע כל ערך שהוא, ללא כל בקרה.

הפתרון: ראשית יש לברר מול בעלי המערכת אילו מפתחות מוגדרים. אם אינם מוגדרים, יש להגדיר PRIMARY KEY על case_number בטבלת האירועים, ומפתח זר מטבלאות המשתתפים אליה יחד עם NOT NULL. במקביל יש להוסיף טבלאות קוד למין ולגזע ולקשר אליהן בעזרת מפתח זר.

<u>שכפול מידע בין incidents לבין הטבלאות המקושרות</u>

איום: השדות subjects, officers, subject_count ו-officer_count בטבלת האירועים מכילים מידע הקיים כבר בטבלאות subjects ו-officers. זו כפילות מכוונת מסוג pre-compute.

השפעה: המידע מאוחסן פעמיים, ולכן קיים איום של אי התאמה: אם מתווספת שורה לטבלת officers מבלי לעדכן את officer_count ואת רשימת השמות, שני המקורות יסתרו זה את זה ולא נדע מי מהם הנכון. בנוסף מבוזבז מעט שטח אחסון. בדאטהסט רגיש כמו אירועי ירי משטרתיים, אי התאמה כזו אינה רק בעיה טכנית, שכן היא עלולה להוביל לפרסום נתון שגוי על מספר המעורבים באירוע. יש לציין שאין כאן הפרה של כלל נורמליזציה, שכן כללי הנורמליזציה עוסקים בתלויות בתוך טבלה ולא בקשרים בין טבלאות שונות.

הפתרון: יש למחוק שדות אלו ולחשב אותם בזמן אמת בעזרת COUNT ו-GROUP_CONCAT. התועלת האפשרית בעיצוב הקיים היא האצת שליפה של מקרה בודד, אך במקרה זה קשה להגן על ההחלטה: גם בעיר עתירת אירועים כמו דאלאס, מספר אירועי הירי הוא בסדר גודל שאינו יוצר שום בעיית ביצועים, וספירה ישירה על הטבלאות המקושרות מהירה לחלוטין. אם בכל זאת יוחלט לשמר את השדות, חובה לאכוף את העדכון בעזרת TRIGGER על הוספה ומחיקה בטבלאות המקושרות, ולא להשאיר את העדכון באחריות קוד האפליקציה. בנוסף מומלץ להריץ שאילתת בקרה תקופתית המשווה בין הערך השמור לבין הספירה בפועל.

<u>שדות רב ערכיים המאוחסנים בשדה יחיד</u>

איום: השדות subjects, officers, subject_statuses ו-subject_weapon הם שדות varchar בודדים המכילים רשימה של ערכים. עצם הריבוי ניכר בשמות השדות עצמם בלשון רבים. אין לנו מידע על צורת הייצוג בפועל, וסביר להניח שנעשה שימוש ב-JSON או במבנה דומה ולא בהצמדה חופשית של מחרוזות.

השפעה: ייצוג זה שובר את הצורה הנורמלית הראשונה. גם אם התוכן תקין, לא ניתן לבצע JOIN לפי הערכים ולא ניתן לאנדקס אותם, וכל שימוש בהם מחייב פירוק המבנה בזמן השליפה. חשוב לציין שמסדי נתונים מודרניים מספקים פונקציות גישה ל-JSON, ולכן שליפה של ערך מתוך המבנה אפשרית ואינה מחייבת בהכרח שימוש ב-LIKE. עם זאת, אם הייצוג הוא מחרוזת חופשית ולא JSON, החיפוש ייעשה בפועל באמצעות LIKE ועלול לתפוס התאמות חלקיות שגויות, ובנוסף הופעת תו ההפרדה בתוך ערך תשבש את הפירוק עצמו.

הפתרון: פירוק לטבלאות מקושרות. עבור subjects ו-officers הטבלאות כבר קיימות, ולכן יש למחוק את שדות הטקסט. עבור subject_statuses ו-subject_weapon יש להעביר את השדות לטבלת המעורבים, שכן הם מתארים תכונה של מעורב מסוים ולא של האירוע כולו, כמפורט בהמשך. אם בכל זאת נדרש לשמור מבנה מרובה ערכים בשדה יחיד, לדוגמה לנתונים דלילים או משתנים, עדיף להגדיר את השדה כטיפוס JSON. טיפוס זה אוכף לפחות את תקינות הפורמט ומונע הזנת טקסט שאינו מתאים לו, אך הוא אינו תחליף להעברת השדות לטבלאות של האנשים.

<u>חלק ב': ניתוח לפי טבלה ושדה</u>

<u>officers</u>

טבלת השוטרים המעורבים באירוע. הטבלה זהה במבנה שלה לטבלת subjects, עובדה שתידון בסיכום. מכיוון שזו ההופעה הראשונה של מרבית השדות, נפרט כאן את האיומים במלואם, ובטבלת subjects נסתפק בהפניה אליהם.

case_number

איום : השדה מקשר את שורת השוטר לאירוע. אין בידינו מידע האם מוגדרים עליו NOT NULL או מפתח זר לטבלת האירועים, ולכן זו נקודה לבירור. נעיר כי בחירת הטיפוס varchar עבור מספר תיק אינה בעיה כשלעצמה : היא אינה חושפת לשגיאות כתיב יותר מטיפוס מספרי, וההגנה על תקינות הערך מגיעה ממפתח זר ולא מבחירת הטיפוס.

השפעה : אם אין אילוץ, ניתן להזין מספר תיק ריק או מספר שאינו קיים בטבלת האירועים, וייווצרו רשומות שוטר יתומות. שורות אלו לא יופיעו בשום דוח המבוסס על JOIN לאירועים, ולכן הבעיה תישאר בלתי מזוהה לאורך זמן. המחיר האמיתי של מפתח טקסטואלי אינו בתקינות אלא בביצועים : מפתח ארוך מגדיל את האינדקסים ומאט את פעולות החיבור בין הטבלאות.

הפתרון : NOT NULL יחד עם FOREIGN KEY לטבלת האירועים. אין להוסיף אילוץ CHECK על מבנה המזהה, שכן מספר התיק נקבע ככל הנראה במערכת חיצונית, ולהניח לגביו פורמט קבוע אינה הנחה סבירה. אילוץ כזה ראוי רק לאחר אישור מפורש של הפורמט מצד בעלי המערכת. לשיפור ביצועים ניתן לשקול מפתח מספרי פנימי לטבלת האירועים ולהפנות אליו מטבלאות המשתתפים.

race

איום : השדה מוגדר כטיפוס char. אין בידינו את רשימת הערכים המותרת ואיננו יודעים האם קיים אילוץ המגביל אותה, ולכן יש להציג זאת כנקודה לבירור ולא כבעיה ידועה. מה שכן ניתן לראות בסכמה הוא שאין טבלת קוד לגזע שאליה השדה מפנה.

השפעה : שני איומים נפרדים. ראשית, בהיעדר טבלת קוד או רשימת ערכים סגורה, ערכים לא תקניים או ריקים ייכנסו ללא התנגדות. שנית, ומשמעותי יותר בהקשר של דאטהסט זה, השדה מקודד לפי מוסכמה שאינה מתועדת בסכמה עצמה. תו בודד אינו מספיק ברור, ובלי מילון ערכים אי אפשר לדעת בוודאות מה משמעות כל קוד. בניתוח סטטיסטי של אירועי ירי לפי מוצא, פרשנות שגויה של קוד היא שגיאה קריטית.

הפתרון : יצירת טבלת מילון races(race_code, race_description) והגדרת השדה כמפתח זר אליה. פתרון חלופי ופחות מומלץ הוא CHECK עם רשימת ערכים סגורה. בכל מקרה יש לתעד את משמעות כל קוד, ולשקול קוד באורך גדול מתו בודד כדי שהערך יהיה קריא גם ללא מילון.

gender

איום : האיומים בשדה gender זהים במהותם לאלו שתוארו בשדה race, ולכן נציג אותם כאן בקצרה : טיפוס char, ללא טבלת קוד וללא רשימת ערכים ידועה.

השפעה : כמתואר בשדה race. נוסיף רק כי אי עקביות בקידוד אינה נמנעת גם כאשר קיימת טבלת קוד. טבלת קוד אוכפת שהערך חוקי, אך אינה מבטיחה שאותה קבוצה תקודד באותו אופן לאורך זמן ובין מקורות שונים, ולכן אותה קבוצה עלולה להיספר פעמיים תחת שני קודים שונים וכל פילוח לפי מגדר יהיה שגוי.

הפתרון : כמתואר בשדה race, טבלת מילון ומפתח זר אליה. בנוסף יש להחליט מפורשות כיצד מיוצג ערך לא ידוע, האם כערך NULL או כקוד ייעודי, ולתעד את ההחלטה.

first_name, last_name

איום : אין NOT NULL, אין אילוץ אורך מינימלי, ואין תקן כתיב.

השפעה : מחרוזת ריקה תתקבל כשם חוקי, ותיווצר שורת שוטר ללא זיהוי כלל. בנוסף, אותו אדם עלול להופיע בכתיבים שונים ברשומות שונות, ומאחר שאין מזהה ייחודי לאדם, אין שום דרך לזהות שמדובר באותו אדם.

הפתרון : NOT NULL יחד עם CHECK על אורך מינימלי, למשל CHECK (char_length(last_name) >= 1), כדי למנוע מחרוזת ריקה. יש להדגיש ש-NOT NULL לבדו אינו מספיק, שכן מחרוזת ריקה אינה NULL.

full_name

Commented [3]: הרבה סטודנטים חשבו שעצם פה הוא בעיה והדבר לא נכון. אין varchar השימוש ב חשיפה גדולה יותר לשגיאות כתיב מבמספר. ההגנה שאינו רלוונטי במפתח ראשי. כן תהיה FK באה מ בעית ביצועים ושימוש באינדקסים. סביר להניח שיש אילוץ חיצוני על מספר התיק

Commented [4R3]: חשבתי שפה זו טבלת המקרים. היא המרכזית אז עדיף להתחיל בה.

Commented [5]: הבעיות במין וגזע זהות. אחד את ההערות לאחד מהן וציין בשני הפניה. למה דווקא פה אתה מתייחס לטפסים שונים? זו בעיה גם אם יש טבלאות קוד.

Commented [6]: ראה תשובה לבמחן, התייחס לבעית הנורמליזציה שמתעוררת פה

Commented [7]: התייחס גם להעדר תעודת זהות כבעיה. יתכן שחלק מהסיבה לכך היא העבודה המשטרתית. אם מוצאים גופה, לא יודעים מה תעודת הזהות שלה (אבל כנראה גם לא את השם).

איום : השדה מכיל את אותה עובדה המאוחסנת כבר בשדות first_name ו-last_name, כלומר אותו מידע מיוצג פעמיים באותה שורה. יש להוסיף לכך שהשם הוא המזהה היחיד של אדם בסכמה, שכן אין בה תעודת זהות או מזהה מנהלי אחר, כמתואר בסעיף היעדר מזהה ייחודי לאדם.

השפעה : ראשית, איום של אי התאמה : עדכון שם המשפחה מבלי לעדכן את full_name ייצור שורה הסותרת את עצמה, ושתי שאילתות לגיטימיות יחזירו שתי תשובות שונות לאותה שאלה. בנוסף מבוזבז מעט שטח אחסון. שנית, ובכך העיקר מבחינת נורמליזציה, קיימת כאן תלות בין שדות שאינה תלות במפתח : מתוך השם המלא ניתן לגזור את השם הפרטי ואת שם המשפחה, ואילו הכיוון ההפוך אינו ודאי, שכן לא ידוע כיצד מורכב השם המלא. שם מלא בן שלושה חלקים אינו מאפשר לדעת אם החלק המרכזי הוא שם שני או שם נעורים.

הפתרון : עדיף לשמור את חלקי השם בלבד ולהרכיב מהם את השם המלא בעת השליפה, וכך נמנעת אי הוודאות שבפירוק שם מלא. אם נדרש לשמור את השדה, יש להגדירו כעמודה מחושבת (Generated Column), כך שמסד הנתונים אחראי לעדכונו והשכפול אינו יכול לסתור את המקור. יש לציין שלעיצוב הקיים יש גם צד תומך : שמות מורכבים, כדוגמת ג'ורג' בוש הבן או הנרי השמיני, אינם נפרקים בקלות לשם פרטי ולשם משפחה, וייתכן שקיימת התממשקות למערכות חיצוניות שכל אחת מהן מייצגת שם בצורה אחרת.

subjects

טבלת המעורבים שנורו באירוע. מבנה הטבלה זהה לחלוטין לטבלת officers, ולכן כל האיומים שפורטו שם חלים גם כאן במלואם : case_number ללא מפתח זר ידוע, race ו-gender ללא טבלת קוד, שמות ללא אילוץ, ו-full_name משוכפל. להלן שני איומים הייחודיים לטבלה זו.

תכונות של המעורב שאינן יושבות בישות שלו

איום : מדובר במקרה פרטי של בעיה כללית יותר : תכונה המתארת אובייקט מסוים אינה מאוחסנת בישות של אותו אובייקט. השדות subject_statuses ו-subject_weapon מתארים את מצבו של מעורב ואת נשקו, אך הם מוגדרים בטבלת האירועים ולא בטבלת המעורבים.

השפעה : באירוע שבו שלושה מעורבים יהיו שלוש רשומות בטבלת subjects, ואילו הנשק יופיע ברשומת האירוע הבודדת בלבד. מכאן שאין שום מידע פורמלי הקושר סטטוס או נשק לאדם מסוים, וכאשר יש יותר ממעורב אחד לא ניתן כלל לדעת למי מהם שייך הנשק. השיוך נשען על הנחה שהסדר ברשימות זהה לסדר ברשימת השמות, הנחה שאינה נאכפת בשום צורה ואינה ניתנת לאימות. אם באחת הרשימות חסר ערך, כל השיוך נסדק והנתונים משויכים לאנשים הלא נכונים.

הפתרון : העברת שני השדות לטבלת subjects כתכונות של השורה, כלומר status ו-weapon לכל מעורב בנפרד, עם מפתח זר לטבלאות מילון מתאימות. זהו התיקון החשוב ביותר בסכמה מבחינת נכונות המידע, שכן הוא היחיד המבטיח שהמידע משויך לאדם הנכון.

היעדר קשר בין הופעותיו של אותו אדם בשני התפקידים

איום : מכיוון ש-officers ו-subjects הן שתי טבלאות נפרדות בעלות מבנה זהה, ומכיוון שאין ייצוג עצמאי לאדם, אותו אדם עלול להופיע בשתיהן ללא כל קשר פורמלי ביניהן.

השפעה : יש להדגיש שהופעה בשני התפקידים אינה בהכרח שגיאה. שוטר שנורה באירוע הוא מקרה לגיטימי לחלוטין, והסכמה הנוכחית פשוט אינה יודעת לייצג אותו כאדם אחד. התוצאה כפולה : פרטיו האישיים משוכפלים בשני מקומות, ואם יעודכנו בטבלה אחת בלבד ייווצרו במסד שתי גרסאות סותרות של אותו אדם. בנוסף לא ניתן לענות על שאלות החוצות את שתי הקבוצות.

הפתרון : כפי שתואר בסעיף היעדר ייצוג עצמאי לאדם, יש להפריד את זהות האדם מהשתתפותו באירוע ולנהל טבלת השתתפות המבחינה בין התפקידים. חלופה של טבלת קישור אחת עם שדה role אינה מספקת, שכן אדם עשוי למלא את שני התפקידים, ולעיתים אף באותו אירוע, ושדה יחיד יאלץ לבחור אחד מהם בלבד.

incidents

טבלת האירועים, ליבת הסכמה. זוהי גם הטבלה הצוברת את מרבית האיומים, בין היתר משום שהוכנסו אליה שדות שמקומם בטבלאות אחרות.

case_number

איום : השדה משמש בפועל כמזהה האירוע. כאמור בחלק א', הסכמה אינה מגלה אילו מפתחות מוגדרים, ולכן יש לברר האם הוגדר עליו PRIMARY KEY. כפי שצוין בשדה המקביל בטבלת officers, בחירת הטיפוס varchar אינה בעיה כשלעצמה.

השפעה : אם אין מפתח ראשי, ניתן להכניס שני אירועים שונים עם אותו מספר תיק, וכן להכניס אירוע ללא מספר תיק כלל. במקרה הראשון כל JOIN לטבלאות officers ו-subjects יכפיל שורות ויעוות כל ספירה. במקרה השני האירוע יהיה מנותק לחלוטין מהשוטרים והמעורבים שלו.

הפתרון : PRIMARY KEY, הכולל ממילא NOT NULL וייחודיות. אין להוסיף אילוץ CHECK על מבנה המזהה, שכן הפורמט נקבע ככל הנראה במערכת חיצונית וההנחה שהוא קבוע אינה סבירה.

date

איום : כאן קיימות שתי בעיות נפרדות. הראשונה, אין אילוץ על טווח הערכים. השנייה, השדה מכיל תאריך בלבד וללא רכיב שעה. נציין שטיפוס התאריך עצמו תקין ומהווה נקודת חוזק בסכמה.

השפעה : לגבי הבעיה הראשונה, ניתן להזין תאריך עתידי או תאריך רחוק מדי בעבר, שניהם ערכים שאינם אפשריים לאירוע שהתרחש בפועל, ושגיאת הקלדה בשנה תיצור אירוע בעתיד שיעוות כל ניתוח מגמות לאורך זמן. לגבי הבעיה השנייה, היעדר שעה מונע ניתוח דפוסים לפי שעות היממה ומונע הבחנה בין שני אירועים שהתרחשו באותו יום. בהקשר זה ראוי לשים לב שגם משמעות המועד אינה חד משמעית : מועד הירי ומועד גילויו אינם בהכרח זהים, וייתכנו מקרים שבהם השעה אינה ידועה כלל ונדרש ייצוג מפורש למצב זה.

הפתרון : CHECK (date <= CURRENT_DATE) למניעת תאריכים עתידיים, יחד עם NOT NULL. אם נדרש ניתוח לפי שעה, יש לשקול שינוי הטיפוס ל-DATETIME. כמו כן יש להגדיר ולתעד איזה מועד בדיוק נשמר בשדה.

location

איום : טקסט חופשי ללא תקן, ובמקביל קיימים בטבלה גם latitude ו-longitude המתארים את אותו נתון עצמו.

השפעה : שתי בעיות נפרדות. ראשית, אותו מקום ייכתב בווריאציות שונות ברשומות שונות, וכל ניסיון לקבץ אירועים לפי מיקום ייכשל או יפצל קבוצה אחת לכמה. שנית, וזהו האיום המשמעותי יותר, המיקום מיוצג פעמיים בשתי צורות שונות ואין שום מנגנון המוודא שהן מסכימות זו עם זו. שגיאת הקלדה בשם רחוב לא תתגלה מול הקואורדינטות, ולהפך.

הפתרון : יש לציין תחילה שלעיצוב הקיים יש היגיון ממשי, ולכן אין למחוק אף אחד משני הייצוגים. תיאור מילולי ונקודה גאוגרפית עונים על צרכים שונים : אם אדם נורה בפארק הירקון, נרצה גם את הנקודה המדויקת לצורך ניתוח מרחבי וגם את הידיעה שמדובר בפארק לצורך ניתוח לפי סוג מקום, ואת המידע השני קשה מאוד להפיק מקואורדינטות בלבד. התיקון הנדרש הוא אפוא הסדרת היחס בין השניים : פירוק הכתובת לשדות מובנים, כלומר עיר, רחוב ומספר, או מפתח זר לטבלת מקומות מנורמלת, קביעה מפורשת איזה מקור הוא העליון, והוספת בקרת עקביות בין הכתובת לקואורדינטות.

subject_statuses

איום : רשימה טקסטואלית של מצבי המעורבים, המאחסנת מספר ערכים בשדה יחיד וללא מילון ערכים.

השפעה : כפי שפורט בטבלת subjects, אין שיוך פורמלי בין סטטוס לאדם, וזו הבעיה החמורה כאן. מעבר לכך, מדובר בשדה שערכיו ניתנים למנייה מראש ובכל זאת אינו מוגבל לרשימת ערכים סגורה, כך שאותו מצב עלול להיכתב בכמה נוסחים. כל פילוח סטטיסטי לפי תוצאות האירוע, שהוא הפילוח המרכזי בדאטהסט מסוג זה, יהיה שגוי.

הפתרון : העברת השדה לטבלת subjects כשדה יחיד לכל מעורב, עם מפתח זר לטבלת מילון subject_statuses(status_id, description) בעלת רשימת ערכים סגורה. אם בכל זאת מעדיפים לשמור מבנה מרובה ערכים, עדיף להגדירו כטיפוס JSON כדי לאכוף לפחות את תקינות הפורמט, אך זהו פתרון נחות : הוא אינו פותר את הכפילות מול טבלת המעורבים ואינו יוצר את השיוך לאדם.

subject_weapon

איום : השם בלשון יחיד אך האירוע כולל מספר מעורבים, כך שקיימת סתירה בין שם השדה למשמעותו בפועל. בנוסף, השדה הוא טקסט חופשי ללא מילון.

Commented [8]: ציין את הסיבות בעד העיצוב. אם מישהו נורה בפארק הירקון, נרצה גם את המיקום במדוייק אבל גם לדעת שזה בפארק.

Commented [9]: יופי

השפעה: אי בהירות מהותית: לא ברור אם השדה מתאר את נשקו של מעורב אחד, רשימה של כל הנשקים באירוע, או את הנשק המשמעותי ביותר. פרשנות שגויה של השדה תוביל למסקנות שגויות. בנוסף, אין הגדרה ברורה כיצד מיוצג מצב שבו למעורב לא היה נשק כלל, האם כ-NULL, כמחרוזת ריקה, או כערך מפורש, ושלוש האפשרויות ייספרו אחרת בשאילתה.

הפתרון: העברה לטבלת subjects כתכונה של מעורב בודד, עם מפתח זר לטבלת מילון נשקים. יש לקבוע ערך מפורש עבור היעדר נשק ולתעד אותו, במקום להסתמך על NULL שמשמעותו דו משמעית.

subjects

איום: רשימת שמות המשכפלת את תוכן טבלת subjects.

השפעה: האיום המרכזי הוא אי התאמה מול הטבלה המקורית. קיימת גם בעיית פענוח, אם כי סביר להניח שהייצוג בפועל הוא JSON או מבנה דומה ולא הצמדת מחרוזות חופשית, ולכן חומרתה מוגבלת. עם זאת, מכיוון שהרשימה מכילה שמות בלבד וללא מזהים, לא ניתן לאמת את תוכנה מול טבלת המקור באופן חד משמעי.

הפתרון: מחיקת השדה. הרשימה תופק בעת הצורך באמצעות JOIN לטבלת subjects יחד עם GROUP_CONCAT.

subject_count

איום: ערך מחושב מראש של מספר המעורבים, המשכפל ספירה שניתן לבצע ישירות על טבלת subjects. בנוסף אין אילוץ על תחום הערכים.

השפעה: השדה עלול לסתור את מספר השורות בפועל, ואין מנגנון המונע זאת. במקביל, ניתן להזין ערך שלילי או אפס, ערכים שאינם אפשריים באירוע ירי שבו מעורב לפחות אדם אחד. במקרה של דאטהסט המשמש לדיווח ציבורי, פער בין השדה לבין הנתונים בפועל הוא כשל.

הפתרון: עקרונית, מחיקת השדה וחישובו בזמן אמת. אם נשמר, נדרשים CHECK (subject_count >= 1) יחד עם TRIGGER המסנכרן אותו אוטומטית, ושאילתת בקרה תקופתית המשווה בינו לבין COUNT בפועל.

officers

איום: רשימת שמות המשכפלת את תוכן טבלת officers. זוהי אותה בעיה שתוארה בפירוט בשדה subjects שלמעלה.

השפעה: כמתואר שם. נוסיף רק כי כאן הבעיה חמורה במיוחד, משום שזיהוי שוטר החוזר על עצמו בין אירועים הוא שאלת המחקר המרכזית, ורשימת שמות בשדה יחיד היא הייצוג הגרוע ביותר לצורך זה.

הפתרון: מחיקת השדה והפקת הרשימה באמצעות JOIN לטבלת officers.

officer_count

איום: ערך מחושב מראש של מספר השוטרים, המשכפל ספירה על טבלת officers, ללא אילוץ תחום.

השפעה: זהה ל-subject_count: אפשרות לסתירה מול הנתונים בפועל, ואפשרות להזנת ערך שלילי או אפס שאינו הגיוני באירוע ירי משטרתי.

הפתרון: מחיקה וחישוב בזמן אמת, או לחלופין CHECK (officer_count >= 1) יחד עם TRIGGER מסנכרן ובקרה תקופתית.

grand_jury_disposition

איום: טקסט חופשי לתיאור החלטת חבר המושבעים, למרות שמדובר בשדה בעל מספר סופי וידוע של ערכים אפשריים.

השפעה: אותה החלטה משפטית עלולה להיכתב בכמה נוסחים שונים, וכל פילוח לפי תוצאה משפטית יפצל קבוצה אחת לכמה קבוצות מדומות. בנוסף, אין הבחנה ברורה בין אירוע שההחלטה בעניינו טרם התקבלה לבין אירוע שכלל לא הועבר לחבר מושבעים, שני מצבים שונים מהותית שעלולים שניהם להיות מיוצגים כ-NULL.

הפתרון: מפתח זר לטבלת מילון dispositions בעלת רשימת ערכים סגורה. יש להוסיף ערך מפורש למצב 'טרם הוכרע' ולהבחין בינו לבין 'לא רלוונטי', במקום להסתמך על NULL.

Commented [10]: כשיש מקרים של אותה בעיה, פרט בהתייחסות הראשונה והתייחס אליה במקרה השני. ככה יהיה יותר ברור שזו אותה הבעיה.

Commented [11]: יופי

attorney_general_forms_url, summary_url

איום: שני שדות כתובת אינטרנט ללא אימות פורמט וללא הגדרת משמעות ל-NULL.

השפעה: ניתן להזין מחרוזת שאינה כתובת תקינה כלל, והשגיאה תתגלה רק כאשר משתמש ינסה לפתוח את הקישור. מעבר לכך, כתובת השמורה במסד אינה מבטיחה שהיעד קיים: ייתכן שהמסמך מעולם לא פורסם, וייתכן שהיה קיים והוסר מן האתר, ושני המצבים אינם ניתנים לזיהוי מתוך המסד עצמו. בנוסף, ערך ריק עלול להתפרש בשתי משמעויות שונות, שהמסמך אינו קיים או שהוא קיים אך טרם קושר, ואין דרך להבחין ביניהן.

הפתרון: CHECK על תבנית הכתובת, למשל בדיקה שהערך מתחיל בקידומת פרוטוקול תקינה, יחד עם CHECK על אורך מינימלי למניעת מחרוזת ריקה. יש לתעד מפורשות את משמעות הערך NULL בשדות אלו. מכיוון שמדובר במשאב חיצוני שאינו בשליטתנו, מומלץ גם לשמור את מועד הבדיקה האחרון של הקישור, ולשקול שמירה מקומית של המסמך עצמו.

summary_text

איום: השדה מוגדר כ-varchar, שהוא טיפוס בעל אורך מרבי, בעוד שתוכנו הוא תיאור מילולי חופשי של האירוע שאורכו אינו ידוע מראש. בנוסף, תיאור זה עשוי לחפוף לסיכום המתפרסם בכתובת שבשדה summary_url.

השפעה: תיאור החורג מהאורך שהוגדר יידחה ברוב מסדי הנתונים, ולכן הסיכון המעשי הוא כשל בטעינה או איבוד רשומה שלמה, ואין מקום להניח חיתוך שקט. הבעיה המשמעותית יותר היא הכפילות מול הסיכום המתפרסם באתר: אותו תוכן מיוצג בשני מקומות, ואם אחד מהם מתעדכן והשני לא, לא נדע איזה מהם הנכון. בדאטהסט שבו התיאור המילולי הוא לעיתים המקור היחיד למידע על נסיבות האירוע, לפער כזה יש משמעות ממשית.

הפתרון: שינוי הטיפוס ל-TEXT, שאינו מגביל את האורך בפועל. במקביל יש להחליט מהו מקור האמת לתיאור האירוע, המסד או האתר, ולתעד את ההחלטה.

latitude, longitude

איום: שני שדות מטיפוס double ללא אילוץ טווח, ובמקביל הם משכפלים את המידע שבשדה location.

השפעה: שלושה איומים נפרדים. ראשית, ניתן להזין ערכים מחוץ לטווח החוקי, וקואורדינטה כזו תמוקם על מפה במקום שרירותי או תגרום לכשל. שנית, ערך אפס בשני השדות הוא נקודה חוקית מבחינה טכנית, אך היא נמצאת באוקיינוס האטלנטי ורחוקה מאוד מדאלאס, ולכן היא מייצגת כמעט בוודאות נתון חסר שנרשם כאפס במקום כערך NULL, וכל ניתוח מרחבי יכלול נקודות מדומות. שלישית, טיפוס double הוא נקודה צפה, ולכן השוואות שוויון מדויקות בין קואורדינטות אינן אמינות.

הפתרון: CHECK (latitude BETWEEN -90 AND 90) ו-CHECK (longitude BETWEEN -180 AND 180). יש לוודא שנתון חסר נרשם כ-NULL ולא כאפס, ולשקול אילוץ המונע את הצירוף (0,0). לצורך השוואות מדויקות עדיף DECIMAL עם דיוק קבוע על פני double.

חלק ג': סיכום והצעת סכמה מתוקנת

האיומים שזוהו מתחלקים לשלוש רמות חומרה.

- ברמה הראשונה, היעדר ייצוג עצמאי לאדם. זהו כשל מבני שממנו נגזרות רוב הבעיות האחרות, ובכללן היעדר מזהה ייחודי, שכפול הפרטים האישיים בכל אירוע, וחוסר היכולת לעקוב אחר אדם בין אירועים.
- ברמה השנייה, שכפול מידע בין טבלת האירועים לטבלאות המקושרות, שדות רב ערכיים המאוחסנים בשדה יחיד, ותכונות של המעורב היושבות ברשומת האירוע במקום ברשומתו שלו.
- ברמה השלישית, שדות בודדים ללא טבלת קוד או ללא אילוץ תחום, וכן נקודות שאינן נגזרות מן התרשים וטעונות בירור מול בעלי המערכת, ובראשן הגדרת המפתחות.

התיקון המשמעותי ביותר אינו הוספת אילוצים לסכמה הקיימת אלא שינוי מבני, כלומר הפרדת זהות האדם מהשתתפותו באירוע. הפרדה זו פותרת בבת אחת את בעיית הכפילות, את בעיית המעקב בין אירועים ואת שכפול הפרטים האישיים. מאחר שהמשמעות של שוטר ושל מעורב שונה, ומאחר שנרצה לשמור על כל אחד מהם פרטים אחרים, המבנה המוצע מפריד ביניהם לשתי ישויות ולא מאחד אותם לטבלת אנשים אחת. במקביל, העברת

Commented [12]: ויתכן שבכלל אין אתר כזה או שהיה וירד. הסתכל בתשובות לניתוחי סכמות אחרים

subject_statuses ו-subject_weapon לרשומת המעורב היא התיקון החשוב ביותר מבחינת נכונות המידע, שכן היא היחידה המבטיחה שהמידע משויך לאדם הנכון.

מבנה מוצע:

```
officers(officer_id PK, first_name, last_name, race_code FK,
gender_code FK)
subjects(subject_id PK, first_name, last_name, race_code FK,
gender_code FK)
incidents(case_number PK, incident_datetime, location_id FK,
disposition_id FK, latitude, longitude, attorney_general_forms_url,
summary_url, summary_text)
incident_officers(case_number FK, officer_id FK)
incident_subjects(case_number FK, subject_id FK, status_id FK,
weapon_id FK)
```

- בשתי טבלאות הקישור המפתח הראשי מורכב, ומוגדר על צירוף שני השדות יחד.
- בנוסף נדרשות טבלאות מילון עבור גזע, מין, מצב המעורב, סוג נשק, החלטת חבר המושבעים ומקומות.
- אם יידרש בעתיד לזהות אדם המופיע בשני התפקידים, לדוגמה שוטר שנורה, ניתן להוסיף מעל שתי הטבלאות טבלת אנשים משותפת שאליה יפנו שתיהן, מבלי לוותר על ההפרדה בין התפקידים.