

נורמליזציה

- נורמליזציה (או נירמול) - Normalization. הוא תהליך של הבאת מסד הנתונים למצב שהנתונים נישמרים עם כמה שפחות כפילויות.

- לרוב מדברים בתיכנון של בסיס נתונים על 3 צורות של נירמול (1NF, 2NF, 3NF). הראשונה מתיחסת לכפילויות מידע. טבלה היא מבנה דו-מימדי (שורות ועמודות). כל שורה מזהה 'דבר מה' שונה מכל שורה אחרת. למשל טבלה שבה יש עמודה עבור שם ועמודה עבור תאריך לידה. מטבע הדברים יתכנו שורות כפולות - זה אינו בצורה נורמלית ראשונה. פיתרון אפשרי להוסיף עמודה עבור מספר זהות.

כל עמודה צריכה להכיל ערך מסוג מסוים. אם בעמודה של תאריך הלידה פעם מופיע חודש ושנה ועבור שורה אחרת מופיעים יום, חודש ושנה - זה אינו בצורה נורמלית ראשונה. כמו כן באותו תא בטבלה אסור שיהיו מספר ערכים, למשל אם הייתה עמודה עבור עיר לידה, עבור כל אדם יהיה רק ערך אחד, לא יכול להיות משהו כמו: London, New York. עוד סוג הפרת מצב של 1NF, אם ישנן עמודות בטבלה בעלות אותה משמעות (חוזרות), לדוגמה אם ניצור טבלה **בעלת עמודות חוזרות** בכל אחד ילד של האדם שמתואר בשורה. מה החיסרון במצב כזה?

- המצב הזה אינו יציב, כי לכל אדם תיאורטית יהיה מספר שונה של עמודות. אם נניח שלא יכולים ליות יותר מ 20 ילדים, גם נזבז עמודות מיותרות על הרבה אנשים וגם אולי נחמיץ איזה בדואי שנשוי ל 4 נשים ויש לו 28 ילדים (או את גואל רצון)

- הפיתרון במקרה זה יהיה לחלק את הטבלה לשתי טבלאות. באחת נתונים של האדם בלבד, ובשנייה נתונים שמשייכים אדם לילדיו, כאשר בכל שורה מופיע האדם עם ילד אחר בטבלה השניה.

לפני הנירמול:

<u>Person_ID</u>	<u>name</u>	<u>email</u>	<u>phone</u>	<u>child1</u>	<u>child2</u>	<u>child3</u>
0001	Itamar	it@gmail.com	0521111111	Sam	Dina	Maya
0002	Dimon	dd@yahoo.com	0502222222	Riva	Shani	Leeam

אחרי

Persons (table)

Person_ID	name	email	phone
0001	Itamar	it@gmail.com	0521111111
0002	Dimon	dd@yahoo.com	0502222222

PersonChildren

Person_ID	Sequence	Child
0001	01	Sam
0001	02	Dina
0001	03	Maya
0002	01	Riva
0002	02	Shani
0002	03	Leeam

היתרונות של הצורה הנורמלית הראשונה ברורים. אין צורך לשנות את מבנה הטבלה ולהוסיף עמודות אם לאדם יש יותר ילדים משלושה, אין ביזבוז של זיכרון במידה ולאדם ישנם פחות משלושה ילדים. כמוכן כפילות הנתונים היא מינימלית, (PersonID)

טבלה שנימצאת בצורה נורמלית ראשונה ומפירה את הצורה הנורמלית השניה, היא טבלה שבה יש **תלות בין חלק מהמפתח הראשי ועמודות אחרות**. למשל טבלה שבה יש שם עובד, כישורים וכתובת עבודה. המפתח הראשי מורכב **משם + כישורים**, אבל כתובת העבודה תלויה רק בשם העובד. דוגמה:

Employee skills

Name	Skill	Work Address
Brown	programmer	61 Main St, NY , NY
Tony	MD	14 E63rd St, Boston, USA
Brown	QA	61 Main St, NY , NY
Brown	Systems analysy	61 Main St, NY , NY

שימו לב שעבור Brown ישנה כפילות, מחזיקים את כתובת מקום עבודתו מספר פעמים.

כדי לפתור זאת נפרק את הטבלה לשתיים. באחת נישמור שם וכישורים (2 עמודות) ובשניה שם וכתובת (2 עמודות). כלומר צורה נורמלית שניה מכילה גם את הראשונה פלוס מה שהזכרנו.

Employees

Name	Work Address
Brown	61 Main St, NY , NY
Tony	14 E63rd St, Boston, USA

Employee skills

Name	Skill
Brown	Programmer
Brown	AQ
Brown	Systems analysy
Tony	MD

עבור צורה נורמלית שלישית ניראה דוגמה לטבלה שנימצאת בצורה נורמלית שניה ולא שלישית:

Tournament winners

Tournament	year	Winner	Winner DOB
Indiana Invitational	2010	Al Fredrickson	July 15, 1985
NY Open	2012	Bob Albertson	August 1, 1992
US Closed	2013	Al Fredrickson	July 15, 1985
NY Open	2014	Chip Masterson	November 3, 1990

כל שורה בטבלה זו צריכה לומר לנו מי ניצח טורניר מסוים ולכן עמודות שהן יכולות להיות מפתח למשל, הינן: **Tournament + year** (במינימום כדי להבטיח יחודיות)

העמודה של Winner DOB תלויה ב-Winner - עובדה שיכולה לגרום עידכונים לא עיקביים, כי תאריך הלידה יחזור על עצמו אם אותו אדם זכה בטורניר זהה בשנים שונות או באותה שנה בטורנירים שונים.

- זהו דפקט קצת קשה יותר לאבחנה, **אחד השדות תלוי בשדה אחר שאינו מפתח ראשי (כלומר תלות עקיפה במפתח הראשי)** - לשם כך נפצל את הטבלה לשתיים בצורה שאחת תכיל את המפתח הראשי (טורניר ושנה) ואת אחד השדות (המנצח) וטבלה שניה שמייחסת את המנצח לתאריך לידתו.

Tournament winners

Tournament	year	winner
Indiana Invitational	2010	Al Fredrickson
NY Open	2012	Bob Albertson
US Closed	2013	Al Fredrickson
NY Open	2014	Chicp Mastersin

Winner DOB

Winner	DOB
Al Fredrickson	July 15, 1985
Bob Albertson	August 1, 1992
Chicp Mastersin	November 3, 1990

הרעיון של הנירמול זה להפחית כפילות נתונים גם מבחינת שימורם וגם מבחינת עידכונם ופישוט מקסימלי למבנה הטבלאות, אך ורק לנתונים ההכרחיים בכל טבלה ע"מ לשלוף כל סוג של נתון מהטבלאות. כל זה שייך לתחום של תיכנון בסיסי נתונים שבחלקו הינו מדע ובחלקו גם אמנות. יש סיכוי שנעבוד בחברות Startup ולכן כדאי לדעת מעט על נושא זה, גם אם זה לא העיקר של הלימוד.

ההגדרה ה"יבשה" של צורה שלישית נורמלית היא מתמטית אוטורית, אבל בקירוב אפשר לומר כמו בבית משפט שכל עמודה שאיננה מפתח ראשי, חייבת לומר משהו על המפתח, על כל המפתח וכלום מלבד המפתח.

כיוון שהבחור מ.י.ב.מ. שהמציא את המודל היחסי קוראים לו: CODD, ישנה בדיחת קרש שבאנגלית נישמעת טוב: (עמודה מכונה גם: attribute)

Bill Kent: "[Every] non-key [attribute] must provide a fact about the key, the whole key, and nothing but the key."^[7] - so help me Codd"^[8].