מטלה 2 חלק 2

1. סכמת הנתונים שנבחר הוא STAR

הנתונים שלנו בנויים מטבלה אחת מרכזית FACT ואין צורך בטבלאות קטנות.

1. ה-Data Warehouse שלנו יראה זהה לData המקורי יציג את העמודות לאחר ביצוע ETL העמודות שיוצגו הם: title, selftext, over\_18, diagnosis
2. הצגת USE-CASE:

תמונה שמכילה תרשים

התיאור נוצר באופן אוטומטי

מה שנרצה לראות נמצא בטבלה אחת מרכזית וכך מתבצע שליפה מהירה יותר.

חלק 3- הגדרה ומימוש ETL

1. בשלב הראשון Extract- אסיפת הנתונים בעזרת קובץ csv שאותו המרנו לקובץל-json .
2. בשלב השני Transform- סינון עמודות, לקיחת רק העמודות שנרצה לעבוד איתם בהמשך. ניקוי ערכי null מהטבלה, ניקוי מחרוזות שאינם מילים. הפחתת שורות אם ישנם ערכים זהים בכל השורה. אותיות גדולות לעמודת subreddit.

בשלב השלישי Load- ניקח את כל הנתונים שעבדנו עליהם ונוסיף אותם למחסן הנתונים שלנו, Data Warehouse .