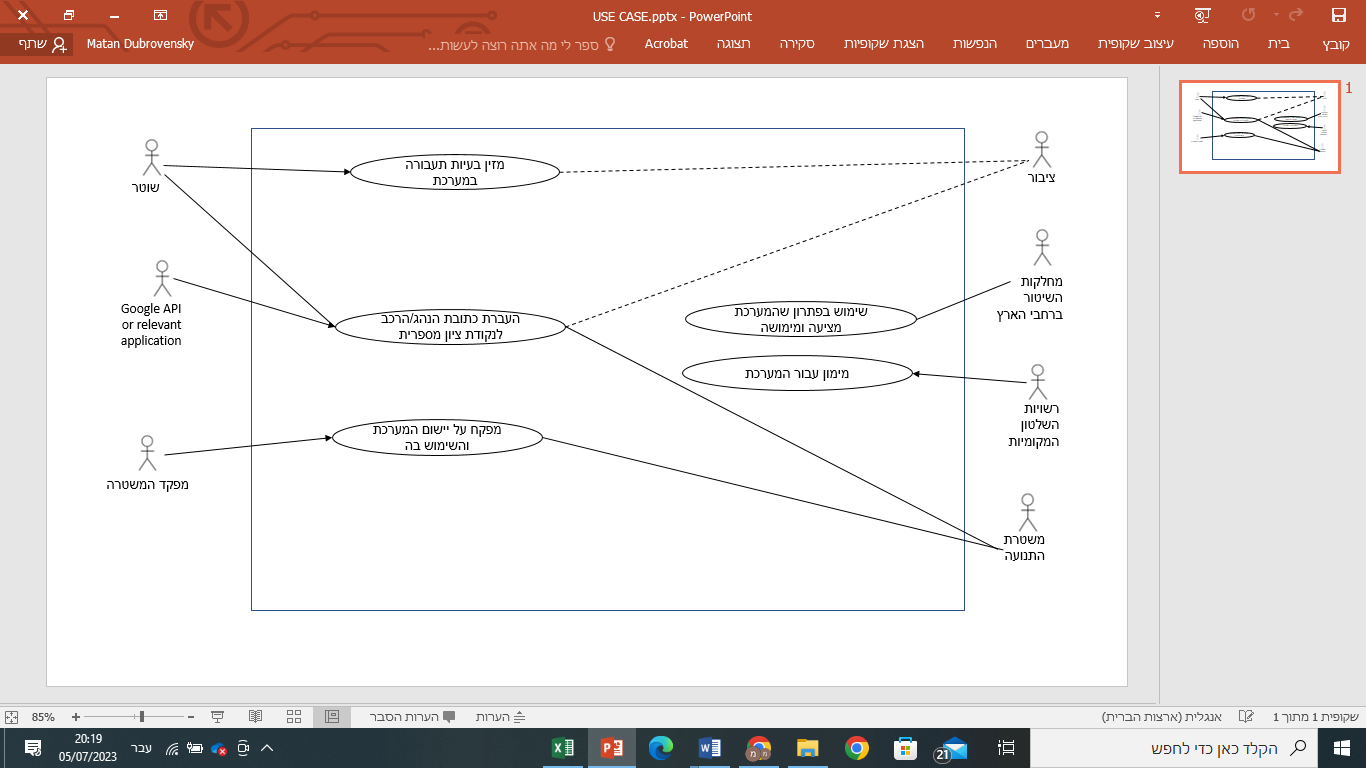
Homework 2 – Business Intelligence

**הגדרת Data-Warehouse**

בחרנו בסכימת Star מכמה סיבות:

פשטות: סכימת Star מציעה עיצוב פשוט ואינטואיטיבי. היא כוללת טבלת עובדות מרכזית, שמכילה את המדדים המדויקים לדוגמה contraband found is arrestedd, המוקפת טבלות ממד שמספקות אפיונים תיאוריים לדוגמה, police department, driver gender, driver age, driver race, violation

ביצועי חיפוש מיטביים: סכימת Star מתאימה היטב לשאילתות וניתוח יעילים. המבנה הדינמי שלה מפחית את מספר השימוש ב join בשאילתות, בעזרת סכמת הכוכב, המשתמש יוכל לפנות לטבלה ייחודית ומשפר את ביצועי החיפוש. ניתן לבצע פונקציות אגרגטיביות וסינון בקלות בטבלאות ממדים, בעוד שטבלת העובדות מכילה את מדדי המפתח, מה שמאפשר אחזור נתונים מהיר יותר.



* + חלק 3
* במקרה שלנו, הדאטה סט נאסף מאתר "קאגל", בצורת קובץ אקסל שמכיל נתונים גולמיים הקשורים לעולם בעיות התעבורה.
* Extract -שלב החילוץ כולל את תהליך ייבוא הנתונים מהמקורות השונים. זה כולל תקשורת עם מקורות הנתונים ואת השליפה של הנתונים עצמם. המקורות יכולים להיות מסדי נתונים, קבצים מקומיים, מערכות אפליקציות ועוד. בשלב זה, הנתונים מועברים ממקורם המקורי לשלב הבא בתהליך.
* Transform-עיבוד ותיקון: שלב העיבוד והתיקון מתרחש לאחר שהנתונים מיובאים ממקורם. בשלב זה, הנתונים מועברים דרך סדרת שלבים שונים על מנת לטפל בתהליך. זה כולל:
  + ניקוי הנתונים מפגמים כגון ערכים חסרים, ערכים לא תקינים, ערכים מותאמים אישית וכו'.
  + פילוח הנתונים לפי מילון שדות ופרמטרים מוגדרים מראש.
  + שינוי פורמט הנתונים לתקן מסוים או פורמט שמתאים למטרות העיבוד והאנליזה המבוקשות.
  + הוספת חישובים, תקופות זמן, שינויי מבנה ועוד.
  + לסיכום שלב העיבוד והתיקון מתחיל עם הנתונים הגולמיים שנאספו בשלב החילוץ ומתסיים עם יצירת נתונים מועילים ומקוננים בצורה מתואמת ומוכנה לטעינה.
* Load–טעינה: שלב הטעינה מתרחש לאחר השלמת שלב העיבוד והתיקון. בשלב זה, הנתונים מועברים למערכת היעד שבה נשמרים ומנוהלים. יתרון הטעינה היא שניתן ליישוב הנתונים ממקורות שונים באותה מסד הנתונים, או לטעון אותם למסדי נתונים שונים לפי הצורך. בשלב זה, ניתן גם לייצר אינדקסים, מפתחות עיקריים ותנאים עבור הנתונים כדי לשפר את ביצועי השאילתות והפעולות האנליטיות על הנתונים.
  + תהליך ETL חשוב כדי לאפשר עיבוד וניתוח יעיל של הנתונים. התהליך מתבצע במערכות ניהול בסיסי נתונים ובכלים מתקדמים שמטרתם לייבא, לעבד ולטעון נתונים במצב שאינו נקי ולהפוך אותם לנתונים מועילים ומבוקרים שניתן להשתמש בהם למטרות עסקיות ואנליטיות