Homework 2 – Business Intelligence

מגישים:

תום בייבוץ – 311450910

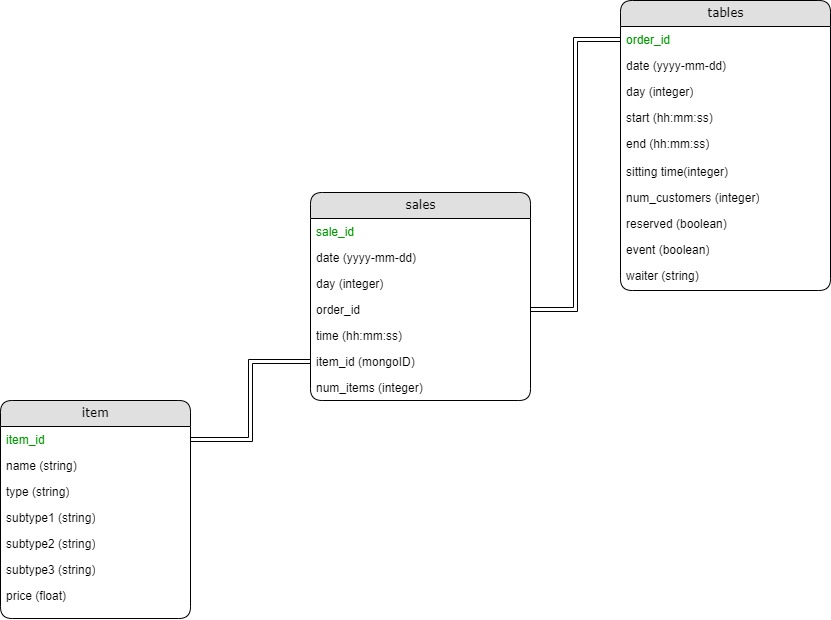
ענת בליאבין – 316613355

יהונתן כהן – 209799576

בעז אביסרור – 206338212

אייל לשסקו – 312116957

**חלק 2**

1. **The Snowflake Scheme**  
   הסכימה המתאימה לבסיס נתונים שלנו הינה סכימת הפתית שלג. זאת מכיוון שהטבלאות מאורגנות בצורה היררכית, כאשר טבלת המכירות משמשת כגשר בין טבלאות ההזמנות והפריטים, ויוצרות מבנה מסועף הדומה לפתית שלג.
2. **ERD**
3. **Use-Case**סכמת Snowflake מאופיינת ביכולת שלה להתמודד עם טבלאות מנורמלות המסודרות בצורה היררכית.

אחד היתרונות הגדולים של הסכמה הזאת הוא שהיא מאפשרת למשוך מידע בצורה מאוד ממוקדת מכל טבלה ובכך גם להפחית את כמות הזיכרון הנמצא בשימוש בעת ביצוע ETL.

Use case לדוגמה שבו סכמת Snowflake תהיה יעילה הוא חישוב הסכום הסופי שכל שולחן במסעדה שישב במסעדה בתקופה מסוימת שילם.

לאחר משיכת השורות המייצגות את השולחנות שלהם נרצה לחשב את הסכום הסופי, נוכל למשוך מטבלת sales רק את השורות המייצגות מכירות המקושרות לשולחנות האלו, ומתוך טבלת items רק את הפריטים המופיעים במכירות. אמנם ביצענו join לשלוש טבלאות, אבל כמות הנתונים שמשכנו היא מאוד קטנה ולכן לא מכבידה על המערכת

**חלק 3**

1. **תהליך ה-ETL**

שלב ה-Extraction:

בשלב זה קיבלנו ממנהל מסעדה גישה למסד הנתונים של המסעדה בטאביט.  
ממסד הנתונים הוצאנו את הטבלאות הבאות:

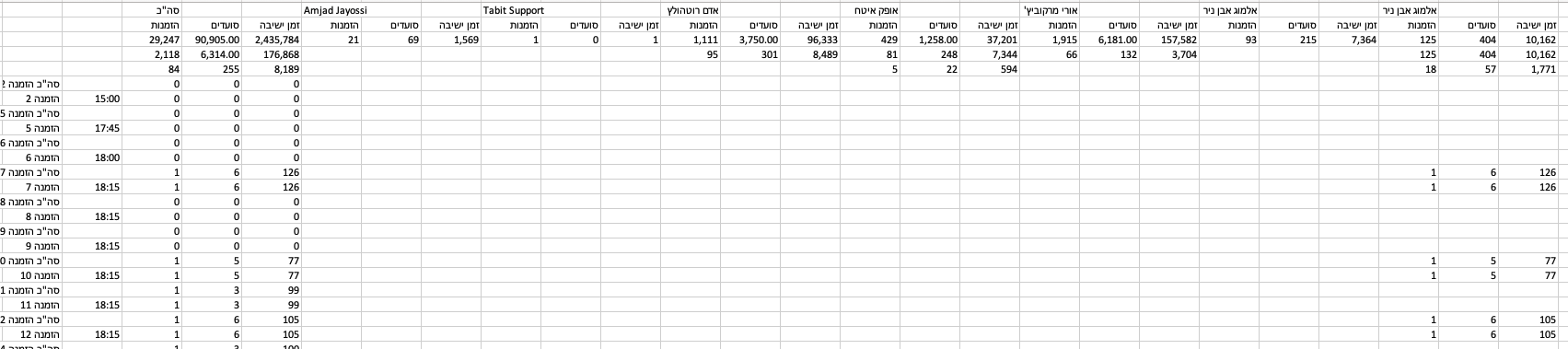
1. מכירות לפי תאריך, שעה פריט ומספר הזמנה.
2. מידע מרוכז על כל הזמנה: תאריך ושעה, כמות סועדים, זמן ישיבה ,האם השולחן היה מוזמן מראש או מזדמן ופרטי המלצר שהיה אחראי על השולחן.
3. פריטי התפריט: מחיר וקטגוריות

שלב ה-Transform:

תהליכים עיקריים בעיבוד הנתונים:

1. העברת כל נתוני הזמן לפורמט אחיד על מנת לאפשר גמישות באגרגציה של המידע: yyyy-mm-dd hh:mm
2. תרגום שמות כל המנות לאנגלית
3. מחיקת כפילויות - לדוגמה: פריטים שהוכנסו למערכת פעמיים עם שינויים קלים בשם (סלט קישואים, ס.קישואים)
4. ניקיון של שורות מטבלת המכירות שמכילים פריטים כלליים הקשורים להתנהלות השוטפת של המסעדה (הערה למטבח, טעימות יין וכו׳)
5. תהליך הETL לא יכול להתבצע בצורה של STTM מאחר והדאטה שקיבלנו מהמסעדה מגיע בפורמט מבולגן של Excel, מה שדרש מאיתנו לשבת ולבנות סקריפט שינקה את הדאטה ויעביר אותו לפורמט אחיד בצורה מאוד מסוימת.

דוגמה לפורמט המבולגן שבו קיבלנו את הנתונים מהדאטה בייס של המסעדה בטאביט הם הנתונים על שולחנות שישבו במסעדה:



הקבצים כוללים המון שורות ועמודות ריקות והמון עמודות סיכום מיותרות. בין היתר, הטבלה מסודרת בצורה כזאת שכל 3 עמודות (בהתאמה) מייצגות פעילות של מלצר יחיד ביום מסוים.

ריכזנו את הסקריפטים שבנינו על מנת לנקות את הדאטה ולהביא אותו לצורה טבלאית, ובנוסף מוצגים שלושת קבצי ה-CSV שיצאו כתוצאה מהסקריפטים האלה.

דוגמה לתוצאה של טבלת tables אחרי עיבוד:

