**פרויקט קורס בינה עסקית**

**קורס מס' 40205**

**מטלה מספר 2**

**מרצה: מר אור פרץ**

**שמות מגישים:**  
שחר בן ישי - 312272818  
אלון ששון – 208625699

חן דנינו – 316393057

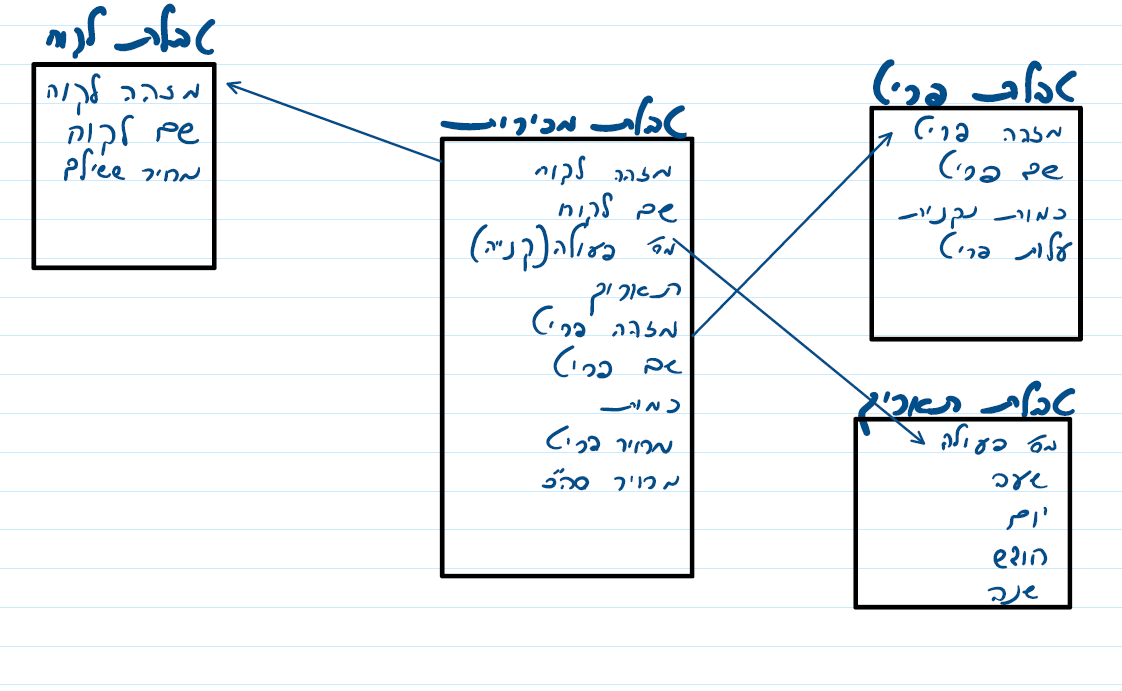
נועם פרידשטין - 205946239

**חלק ב'- Warehouse Data**

**שאלה מספר 1:**בחרנו לעבוד עם סכמת כוכב מכיון שהנתונים שלנו – מערך קניות של לקוחות במינימרקט, מתממשים הכי טוב עם סכמיה זו כאשר יש לנו את נתוני הקניה של כל לקוח באמצע ומסביבו אפשר לפרט על המוצר, כמות המוצר, הלקוח והזמן שנקנה המוצא

**שאלה מספר 2:**

תיאור הdata warehouse באמצעות ERD:

****

**שאלה מספר 3:**

מנהל של חברת מינימרקטים רוצה לדעת מהו טווח השעות הכי רווחי של הסופר וכיצד מאפיינים כמו סוג מוצר משפיעים על שעת המכירה.

בחרנו בסכמת הכוכב מכיוון שהשליפה מהירה ע"ב מודל טבלאי אחוד.

**חלק ג' - הגדרה ומימוש ETL**

**1. הגדרת תהליך הETL- עבור אוסף הנתונים:**

Extraction – נחלץ את הנתונים מטבלאות המקור CUSTOMER ו-PURCHASES .

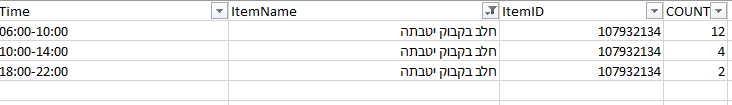
Transformation – נעביר את הנתונים שחולצו למודל טבלאי אחוד ונבצע סינון של העמודות הרלוונטיות בהתאם לשאלה העסקית שהגדרנו בסעיף 3.

Load – נטעין את הנתונים שחילצנו ל.DW

**2. מימוש ETL ע"י STTM:  
  
שלב 1:** Reference data בשלב זה נגדיר את סט הנתונים המורכב מ2 טבלאות: CUSTOMER ו-PURCHASES, אותן נגדיר באמצעות השדות הרלוונטיים.  
  
**שלב 2:** Extract from data reference - נחלץ את הנתונים ע"י קובץ XLSX.  
  
**שלב 3:** Data validation - נוודא כי הנתונים שחילצנו עונים על השאלה העסקית ומתאימים למטרת הפרויקט.  
  
**שלב 4:** Transformation data – נזקק את המידע הרלוונטי לפרויקט כלומר ננקה את הDATA – במסד נתונים שלנו לא היו נתונים ריקים/חסרים או נתונים חריגים ולכן לא היה צורך לזקק את המידע. **שלב 5:** Stage : שלב זה הוא שלב ביניים בו כל הנתונים נמצאים באזור ה STAGING בתוכנות כמו TABLEAU או AIRFLOW, דרכן הנתונים עוברים.  
  
**שלב 6:** Publish to data warehouse נעביר את הנתונים הרלוונטיים שביצענו עליהם עיבוד לDATABASE אחוד וממנו קיבלנו את טווחי הזמן שבהם נעשו הכי הרבה קניות ואיזה מוצר הכי נמכר.

**3. דוגמא לתוצרי הרצת ובדיקת ETL:**

פילוח שעות קניה וכמויות קניה של מוצר מסוג "חלב בקבוק יטבתה"

****

A screenshot of a computer

Description automatically generated with low confidenceתצוגת הנתונים לפי פילוח שעות