**עבודה מסכמת**

אלון ששון

חן דנינו

נועם פרידשטיין

שחר בן ישי

**חלק 1**

**שאלה עסקית**

מנהל של חברת מינימרקטים רוצה לדעת על נתוני קניית הלקוחות במקום. ממוצע קניית הלקוחות כדי לדעת מאיזה מחיר לעשות מבצעים, אילו לקוחות קונים הכי יותר ומהם המוצרים הנמכרים ביותר.

**KPI'S**

1. ערך הזמנה ממוצע – ערך ממוצע של רכישה של כל לקוח בחנות, עוזר להעריך את הרגלי ההוצאות וכוח הקנייה של הלקוחות.

חישוב =

1. שיעור רכישה חוזרת: מודד אחוז הלקוחות שמבצעים רכישות חוזרות בחנות לתקופה כלשהי. זה עוזר להעריך את נאמנות הלקוחות ואת היעילות של אסטרטגיות שימור.

חישוב =

**חלק 2**

1. בחרנו בסכימת הנתונים Star עבור מערך הנתונים שלנו. סכימת Star מאפשרת אחזור נתונים יעיל ומהיר. הוא תאים במיוחד למודלים שבהם יש טבלה מרכזית (טבלת עובדות) המכילה את הנתונים העיקריים, וטבלאות ממדים קטנות יותר המתייחסות לנתונים המרכזיים. בסכימת Star, לטבלה המרכזית יש מפתח ראשי המשמש לשאילתה ולקישור עם טבלאות הממדים. זה מאפשר אחזור נתונים מהיר וגישה יעילה למידע המרכזי.
2. מחסן הנתונים לאחר הסכימה:

מחסן הנתונים משלב את סכימת Star ומשפר את הביצועים והיעילות באחזור נתונים. הוא משמש כאחסון מרכזי המאגד מידע רלוונטי מטבלאות שונות, ומאפשר שאילתות מהירות וביצוע יעיל של שאילתות מורכבות. מחסן הנתונים מספק מבנה מאורגן ומותאם לניתוח נתונים ויכול לשמש ככלי יעיל לניהול, תכנון וקבלת החלטות מונחה נתונים בארגון.

2.3 Use case:

A picture containing diagram, text, technical drawing, plan

Description automatically generated

A picture containing diagram, line, text, technical drawing

Description automatically generated

**חלק 3**

**תהליך הETL**

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

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

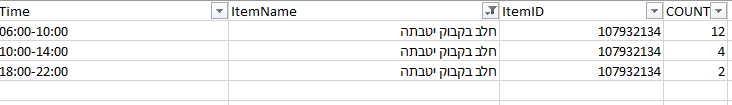
Transformation – נעביר את הנתונים שחולצו למודל טבלאי אחוד ונבצע סינון של העמודות הרלוונטיות בהתאם לשאלה העסקית שהגדרנו בסעיף 3.

Load – נטעין את הנתונים שחילצנו ל.DW

1. **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

**חלק 4**

/\*1 total price for each buy\*/

SELECT Customer\_Id, name, total\_Price,

SUM(total\_Price) OVER (PARTITION BY Customer\_Id ORDER BY date) AS running\_total

FROM dataset;

/\*2 most selled items\*/

SELECT itemId, item\_Name, SUM(amount) AS total\_quantity,

RANK() OVER (ORDER BY SUM(amount) DESC) AS rank

FROM dataset

GROUP BY itemId, item\_Name

ORDER BY total\_quantity DESC;

/\*3 avg spent vs all avg spent\*/

SELECT Customer\_Id, name, AVG(total\_Price) AS avg\_amount\_spent,

AVG(total\_Price) OVER () AS avg\_amount\_spent\_all\_customers

FROM dataset

GROUP BY Customer\_Id, name;

/\*4 customers who spent more than avg\*/

SELECT Customer\_Id, name, total\_Price

FROM (

SELECT Customer\_Id, name, total\_Price,

AVG(total\_Price) OVER () AS avg\_amount\_spent

FROM dataset

GROUP BY Customer\_Id,item\_name

) AS subquery

WHERE total\_Price > avg\_amount\_spent;

/\*5 customers total spending\*/

SELECT Customer\_Id, name, SUM(total\_Price) AS total\_spent

FROM dataset

GROUP BY Customer\_Id, name;

/\*6 ranking costumers in tiers(10)\*/

SELECT Customer\_Id, name, total\_spent,

NTILE(10) OVER (ORDER BY total\_spent DESC) AS tier

FROM (

SELECT Customer\_Id, name, SUM(total\_Price) AS total\_spent

FROM dataset

GROUP BY Customer\_Id, name

) AS subquery

ORDER BY tier ASC;

/\*7 number of unique times they made purchases\*/

SELECT Customer\_Id, name, COUNT(DISTINCT date) AS num\_unique\_times

FROM dataset

GROUP BY Customer\_Id, name

ORDER BY num\_unique\_times DESC;

/\*8\*/

SELECT Customer\_Id, name, total\_Price

FROM (

SELECT Customer\_Id, name, total\_Price,

AVG(total\_Price) OVER () AS avg\_amount\_spent

FROM dataset

GROUP BY Customer\_Id,item\_name

) AS subquery

WHERE total\_Price > avg\_amount\_spent;

/\*9 items that were purchased by all customers\*/

SELECT itemId, item\_Name

FROM dataset

WHERE itemId IN (

SELECT itemId

FROM dataset

GROUP BY itemId

HAVING COUNT(DISTINCT Customer\_Id) = (

SELECT COUNT(DISTINCT Customer\_Id)

FROM dataset

)

);

/\*10 customers who spent the most on a single purchase\*/

SELECT Customer\_Id, name, date, total\_Price

FROM (

SELECT Customer\_Id, name, date, total\_Price,

RANK() OVER (PARTITION BY Customer\_Id ORDER BY total\_Price DESC) AS rank

FROM dataset

) AS subquery

WHERE rank = 1;

**חלק 5**

**המסקנות עבור חלק 4:**

1. המוצרים הנמכרים ביותר הם מוצרי החלב (מחלקה כללית הוא מוצר שסריקת הברקוד לא עבר) ובתוכם הכי נקנים חלב, קוטג' 5% ודנונה יוגורט 3% לאחר מכן משקאות קרים ומתוקים.
2. ממוצע קניית לקוח הוא 20.74 ₪.
3. חמשת הלקוחות הקונים ביותר הם: רייכמן, שרגא זילברמי, אילנה קרן, מאיה ברק וסאן זרח
4. הלקוחות שהגיעו הכי הרבה פעמים לקנות הם: יעל סנטור, דרור דוברונסקי, רונית גדרון, דניאלה דניאלה ואלטר עדי