**מטלה 2 – בינה עסקית:**

אביב רביד 316018175

נועה דלאל 316045335

עדן בן דוד 206957797

**חלק 1 – הגדרת Data Warehouse:**

1. בחרנו בסכמת ה Star כיוון שהנתונים שלנו מתייחסים לדירות, ולכל דירה יש פרטים המאפיינים אותה. כל שורה תואמת לדירה מסוימת ועל כן הקשר הינו יחיד לרבים בין טבלת הfact לטבלאות הdimensions.
2. נציג את תרשים הERD המתאר את ה Data Warehouse.

|  |
| --- |
| **ds\_Houses\_Prices** |
| houses\_prices\_id |
| price |
| bedrooms |
| bathrooms |

|  |
| --- |
| **Houses\_FACT** |
| houses\_prices\_id |
| houses\_data\_id |
| houses\_location\_id |
| price |
| bedrooms |
| bathrooms |
| floors |
| sqm\_living |
| yr\_built |
| yr\_renovated |
| city |

|  |
| --- |
| **ds\_Houses\_Data** |
| houses\_data\_id |
| sqm\_living |
| sqm\_lot |
| sqm\_above |
| sqm\_basement |
| floors |
| yr\_built |
| yr\_renovated |

|  |
| --- |
| **ds\_Houses\_Location** |
| houses\_location\_id |
| street |
| city |
| statezip |
| country |

1. הטבלאות מומשו, והקוד sql שלהן הינו:

CREATE TABLE **ds\_Houses\_Prices**

(

houses\_prices\_id INTEGER NOT NULL,

price INTEGER,

bedrooms INTEGER,

bathrooms DECIMAL(5,2)

);

CREATE TABLE **ds\_Houses\_Data**

(

houses\_data\_id INTEGER NOT NULL,

sqm\_living DECIMAL(12,2),

sqm\_lot DECIMAL(12,2),

sqm\_above DECIMAL(12,2),

sqm\_basement DECIMAL(12,2),

floors DECIMAL(2,1),

yr\_built INTEGER,

yr\_renovated INTEGER

);

CREATE TABLE **ds\_Houses\_Location**

(

houses\_location\_id INTEGER NOT NULL,

street varchar(150),

city varchar(100),

statezip varchar(100),

country varchar(10)

);

CREATE TABLE **Houses\_FACT**

(

houses\_prices\_id INTEGER NOT NULL,

houses\_data\_id INTEGER NOT NULL,

houses\_location\_id INTEGER NOT NULL,

price INTEGER,

bedrooms INTEGER,

bathrooms DECIMAL(12,2),

floors DECIMAL(2,1),

sqm\_living DECIMAL(12,2),

yr\_built INTEGER,

yr\_renovated INTEGER,

city varchar(100)

);

1. קיימות עמודות בטבלת הfact המגיעות מטבלאות שונות. במידה ולא היינו משתמשות במבנה של Data Warehouseהזה, כדי להגיע לממוצעים היה צורך לבצע בכל פעםjoin בין כמה טבלאות שונות, דבר שהיה גורם לחישובים חוזרים, וזמן ריצה ארוך.

לדוגמא – עבור חישוב ממוצע שטח מגורים לדירת 3 חדרים היינו צריכות לבצע Join בין 2 טבלאות ולאחר מכן לבצע את חישוב הממוצע בכל שאילתה מחדש, וזאת במקום לבצע windows function על טבלת הfact בלבד.

1. הטבלאות מצורפות בצורת excel.

**חלק 2 – הגדרה ומימוש ETL:**

1. תהליך ETL –

בשלב הE-extract ניקח את העמודות הרלוונטיות שקיבלנו מטבלאות הלקוח (במקרה שלנו כולן משום בשאלה הלא מונחית אנחנו מתייחסות לפרמטרים שמאפיינים את הבית).

בשלב הtransform-T נמיר את גודל הבית בfeet לגודל במטר, שאר העמודות יהיו ללא שינוי.

בשלב הL-Load נטען את הטבלאות החדשות לData warehouse שלנו, לטבלאות הממדים וטבלת הfact שיצרנו.

1. את תהליך הETL Pipeline נבצע באמצעות טבלת אקסל המפרטת את השינוי בין הטבלאות של הלקוח לdata warehouse שלנו.
2. מצורף קובץ אקסל STTM.