**Homework 2 – Business Intelligence**

**חלק 2 – הגדרת Data Warehouse**

1. מתוך הסכמות הקיימות ובהתבסס על טבלת הנתונים שבחרנו במטלה הקודמת סיכמת כוכבים תהיה המתאימה ביותר, מכיוון שהיא מספקת דרך פשוטה ואינטואיטיבית לארגן ולנתח את הנתונים שלנו. על ידי ארגון הנתונים שלנו בצורה זו, נוכל לבצע בקלות שאילתות כדי לנתח את הקשרים בין הממדים והמדדים. לדוגמה, נוכל לנתח בקלות כיצד התמחור משפיע על דירוגי משתמשים עבור תעשייה או קטגוריה ספציפית, או כיצד מספר המשתמשים משתנה לפי קנה המידה העסקי וסוג המשתמש.
2. להלן פירוט של האופן שבו ניתן לארגן את טבלת הנתונים לפי סכמת כוכבים:
3. Fact Table:

* OS key
* Industry key
* Product id (measure)
* Mobile apps (measure)
* Rating (measure)

1. Dimension Tables:

* OS Dimension:
* OS Key
* Operating System Name
* Industry Dimension:
* Industry Key
* Industry Category

1. דוגמא לuse case בשאלת המחקר שלנו – " מדידת נתח השוק של מערכת Mac בכל תעשייה- נבצע בדיקה על ידי סכימה של השימוש במערכת Mac ובכך נוכל להסיק על נתח השוק של המערכת ואף להציע תוכנית לייעול על מנת שהמערכת תהיה בעלת שוק נרחב."

על מנת שנוכל לבצע ניתוח כזה על נתח השוק של המערכת, נצטרך לעשות איחוד בין טבלאות שונות (Rating ו- OS). בעזרת שימוש ב""Over נבצע הפחתת שימוש בJoin מאחר והפונקציה מאפשרת שליפה מהירה של מידע וקבלת פלט, חוסך זיכרון (שמירת המידע המקורי ללא שינוי) ומקצר את זמן ההמתנה של המשתמש. דבר זה יכול לעזור לחברת התוכנה לקבל החלטות מושכלות לגבי פיתוח מוצר, אסטרטגיות תמחור ומיקוד לקוחות, מה שמוביל לתוצאות עסקיות משופרות.

**חלק 3 – הגדרה ומימוש ETL**

1. תהליך ETL (Extract, Transform, Load) משמש לאיסוף נתונים ממקורות שונים, להפוך אותם לפורמט רצוי ולטעון אותם למחסן נתונים או למערכת יעד אחרת. הנה דוגמה לתהליך ETL לאיסוף הנתונים עם עמודות OS, Industry, Product id, Mobile apps, Rating:

הוצאה:

השלב הראשון בתהליך ה-ETL הוא לחלץ את הנתונים מהמקור שלו. במקרה זה, הנתונים עשויים להיות מאוחסנים במקורות שונים, כגון גיליונות אלקטרוניים, מסדי נתונים או קבצי CSV. הנתונים נשלפים ומועברים לאזור היערכות בו ניתן לנקות ולהפוך אותם.

טרנספורמציה:

בשלב הטרנספורמציה הנתונים עוברים ניקוי, סטנדרטיזציה והפיכה לפורמט המתאים לניתוח.   
ניתן לבצע את ההמרות הבאות על הנתונים:

* סטנדרטיזציה של מזהה המוצר על ידי הסרת תווים או רווחים מיוחדים.
* סיווג התעשייה בהתבסס על קטגוריה.
* סטנדרטיזציה של סוג הפריסה ל-on-premise, ענן או היברידי.
* סיווג התמחור בהתבסס על קטגוריה.
* סטנדרטיזציה של סולם הדירוג מ-1-3 ו מ-3.5-5.

טוען:

ברגע שהנתונים עוברים טרנספורמציה, הם נטענים למערכת היעד, כגון מחסן נתונים, לצורך ניתוח.   
יש לבצע אופטימיזציה של מערכת היעד לסוגי השאילתות והניתוח שיבוצעו על הנתונים.  
במהלך שלב הטעינה, ניתן לנקות, לסנן או לצבור את הנתונים כדי לייעל את השימושיות שלהם. לדוגמה, ניתן להסיר רשומות כפולות, לזקוף ערכים חסרים או לצבור נתונים ברמה גבוהה יותר, כגון חודשי או רבעוני.

לסיכום, תהליך ה-ETL לאיסוף הנתונים עם לאיסוף הנתונים עם עמודות OS, Industry, Product id, Mobile apps, Rating כרוך בחילוץ הנתונים מהמקור שלהם, הפיכתם לפורמט סטנדרטי המתאים עבור ניתוח, וטעינתו למערכת יעד המותאמת לשאילתות וניתוח.