# פרויקט קורס בינה עסקית

## חלק א' – ניתוח המערכת

**<u>תיאור סביבה עסקית</u>**

חברת ***** עוזרת לבעלי אתרים לייצר רווח מתוכן הווידאו שלהם על ידי הצעת פתרונות מונטיזציה פרוגרמתיים של מודעות.
הפלטפורמה שלהם מספקת יכולות פרסום מתקדמות, המאפשרות לבעלי אתרים להציג מודעות וידאו ממוקדות ורלוונטיות לקהל שלהם. החברה שמה דגש על אלגוריתמי האופטימיזציה ויכולות למידת המכונה שלה כדי לשפר את ביצועי המודעות ולמקסם את ההכנסה עבור בעלי אתרים.

הלקוחות של החברה הם אתרי הפרסום, שמרוויחים כסף מכל צפייה של USER בפרסומת שהם מציגים לו באותו אתר.
החברה מתמודדת עם כמויות אדירות של DATA ,שנאגר מידי יום ומגיע למילוני רשומות .
הרשומות במאגר המידע מייצגות הזדמנות פרסום שיש לנו בעת כניסה של יוזר מסוים לאתר מסוים. במאגר הנתונים מתועדים כל האירועים שמתרחשים וזה גם מה שלמעשה החברה מוכרת לאתרי הפרסום האלו – את ההזדמנות לפרסם, ואת הדרך הטובה ביותר לעשות את זה.

הניתוחים שנבצע על ה DATA הם המהות של העשייה של החברה, כיוון שעל בסיס הבנה מעמיקה של התנהגות היוזרים באותם אתרים, ועל בסיס פרמטרים כמו אזור או מכשיר הפעלה – נוכל לאפטם את סגנון הפרסומות לאוכלוסיות מסוימות באזורים מסוימים, וגם את התזמון המתאים ביותר להצגת הפרסומים .

השונות בנתונים והמגוון הרחב של היוזרים והתוכן הקיים שניתן להתאים לכל יוזר, הובילו אותנו להבין את הצורך המתבקש בדשבורד - ויזואליזציה של הנתונים שמשתנה על בסיס יומי – שתסייע באופן משמעותי לקבל החלטות מבוססות נתונים, שמטרתם למקסם את ההכנסות של אותם אתרי פרסום ולשפר את ביצועי המודעות עבור בעלי האתרים.

**<u>השאלות העסקיות שבהן נעסוק</u>**

<u>שאלות עסקיות בנושא הכנסות</u>

1. כמה הכנסה בדולרים הצלחנו להפיק ?
2. כמה כסף הכנסנו מתחילת החודש האחרון? וכמה כסף הכנסנו מתחילת השנה?
3. מהי התפלגות ההכנסות לפי שנים ?
4. לצורך תכנון עסקי ותקציבי, נרצה לדעת מהי התפלגות ההכנסות לפי שנים וכך גם לזהות טרנדים ודפוסים בהתנהגות הלקוחות והשוק. לצורך כך, הצגנו גרף עמודות.
5. מהם אחוזי הגידול והקיטון לפי חודשים ?
6. מהו ממוצע ההכנסות למשתמש?
7. מהו ממוצע ההכנסה לכל Event של כניסת משתמש לאתר?
8. מיהם עשרת המפרסמים עם נתח ההכנסה המשמעותי ביותר? ומיהו המפרסם המוביל נכון לעכשיו ?
9. כמה כל מפרסם משלם לנו עבור 1000 פרסומות(CPM) ?

<u>שאלות עסקיות בנושא פעילות משתמשים</u>

1. כמה משתמשים פעילים יש במערכת וכמה מתוכם גברים/נשים?
2. מהי כמות משתמשים לפי קבוצות גיל?
3. מהי התפלגות הצפיות לפי גיל ומגדר? ומהי שכבת הגיל שהכי מובילה בצפיות?
3. מהי התפלגות ההזדמנויות לפרסום לפי שעה במהלך היום?
4. האם באמת קפצה פרסומת כאשר הייתה לנו הזדמנות לפרסם ?
5. מהם אחוזי ההתרשמות בפועל של המשתמשים מהתוכן המפורסם מתוך הזדמנויות הפרסום?

6. מיהם ה publishers שיש להם הכי הרבה צפיות?
7. מיהם ה Publishers שאחוזי ה FR שלהם גבוהים מ 100?%
8. האם הלקוח סיים לצפות בפרסומת?

<u>שאלות עסקיות בנושא תקלות</u>

1. מהו אחוז ה EWF הכולל?
2. מהו אחוז התקלות החודשי ביחס לשנה שעברה?
3. באילו אזורים גיאוגרפיים יש הכי הרבה תקלות? ומהם אחוזי התקלות בכל איזור?
4. הצגת תקלות לפי סוג מכשיר .
5. באילו מדינות הנתח הגדול ביותר של התקלות?
6. באילו שעות ביום יש הכי הרבה תקלות? הצגת התפלגות התקלות לפי שעות ביום.
7. מהו אחוז התקלות במגמה חודשית ומהי החריגה מיעד התקלות שהוגדר ?
8. מהו אחוז התקלות החודשי ביחס לשנה שעברה?

שאלות אלו יוכלו לעזור למנהלים לקבל החלטות עסקיות מבוססות נתונים וליישם אסטרטגיות לשיפור הביצועים.

**<u>המדדים שיתנו מענה לשאלות העסקיות</u>**

- **Dim_Inventory** - כמות הפעמים שהמערכת טוענת את הנגן, כל ריענון של העמוד או כניסה של משתמש לאתר Inventory .מייצג את הזדמנות הפרסום.
  זה מדד כמותי המחושב לפי התנהגות המשתמשים באתר.
  אופן החישוב: סכום כניסות לאתר/ריענון של הדף .

- **Dim_Impression** - הופעה שמתייחסת למקרה בודד של הצגת פרסומת למשתמש בדף האינטרנט / בתוך אפליקציה. אופן חישוב: ספירת כמות הפרסומות שנוגנו לפחות 2 שניות.

- **Dim_Complete** - כמות הפעמים שהיוזר סיים לצפות בפרסומת. אופן החישוב: ספירת כמות הפעמים שהיוזר סיים לצפות בפרסומת באופן מלא.

- **Dim_Revenue** -כמות הרווח בדולרים ששולם מצד מפרסם הפרסומת לאותו אתר פרסום.
  אופן החישוב: סכום הרווחים שנצברו מהפרסומות המוצגות באתר.

- **Empty Water Fall (EWT)** - כמות הפעמים שהנגן נטען אך לא נמצאה פרסומת מתאימה עבור התכונות של אותו יוזר שנכנס .

**מדדים מחושבים–**

- **Fill Rate-** אחוז ההשמה - מתוך ההזדמנויות להציג פרסומת (כמות הפעמים שנטען הנגן) כמה פעמים הצלחנו לממש ולהשיג Impressions
מחושב באופן הבא:

$$\text{Fill Rate} = \frac{\text{Impression}}{\text{Inventory}}$$

- **CPM -** מדד קבוע בתעשיית ה - Adetch העלות ל-1,000 חשיפות של מודעה באתר .
כאשר Revenue זה הרווח מצד מפרסם הפרסומת ו- Impression זה כמות ההופעות של הפרסומת, כלומר כמות הפעמים שהוצגה הפרסומת. המדד מחושב באופן הבא:

$$1000 \times \left(\frac{Revenue}{Impression}\right) = CPM$$

**<u>סכמת הנתונים של מחסן הנתונים</u>**

<u>אפיון הFact-Tables</u>

הרשומות בטבלת ה Fact-Tables כפי שציינו, מאופיינים כ- Events ומייצגים כניסה בודדת של יוזר לאתר מסוים. בכל רשומה מפורט איזה יוזר נכנס, מתי הוא נכנס, לאיזה אתר, מאיזה מכשיר הפעלה, מי המפרסם הפוטנציאלי באותו אתר ואיזה מדינה זה קרה. את הפרטים האלו מייצגים טבלאות הממדים והם מקושרות אל הFact-Table באמצעות מפתחות זרים (key constrains-foreign). הprimary key של הFact-Tables הוא EventID, והוא מספר סידורי של הרשומות. המפתח הראשי של הFact-Table הוא שילוב של כל המפתחות הזרים שנמצאים בה.

בנוסף, ברמת ה Fact-Tables מפורט – בכל פעם שהיה אירוע של משתמש X יוצג לו בשדה Complete את כמות הפעמים שסיים לצפות בפרסומת, בשדה Inventory יופיעו כמות הפעמים שנכנס לאתר או רענן את האתר (הזדמנויות פרסום), בשדה EWF יופיע האם הייתה תקלה ולא עלה תוכן שמתאים ליוזר, בשדה Impression תוצג כמות הפעמים שהיוזר צפה יותר מ-2 שניות בפרסומת. השדות כולם Fully-Adative ומפורטים ברמת ה Event הבודד, מה שמאפשר ניתוח מעמיק של התנהגות היוזרים .

**ניעזר גם במדדים חיצוניים כמו השעות החופשיות של קהל היעד. לדוגמא -באיזה שעות עובדים מסיימים לעבוד/ימי חופשות/חגים. כיוון שהנתונים שלנו מגיעים מארצות שונות - ניקח בחשבון גם את הזמנים המשתנים מסביב לעולם.**

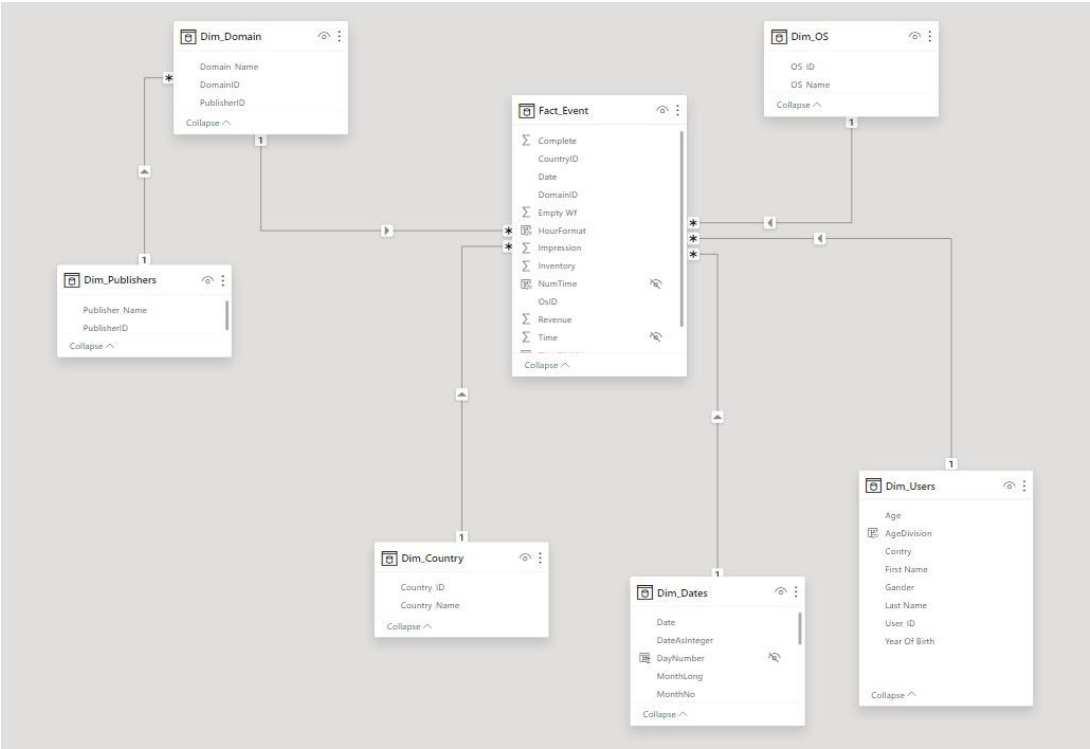
<u>אפיון הממדים</u>

- **Users** - טבלת המשתמשים. המפתח הראשי של הטבלה הוא: UserID – מספר סידורי שכל משתמש מקבל. השדות שהטבלה מכילה City , YearofBirth , UserName
- **Country -** טבלת המדינות שבהם גולשים המשתמשים. השדות של הטבלה: CountryID – המפתח הראשי, ו-CountryName.
- **Date -** טבלה שמכילה את פרטי הזמן שבו בוצע אותו
Event. FullDate , MonthDate , YearDate , DayOfWeek , QurterDate , hour.
הסיבה שירדנו לרזולוציות כל כך נמוכות בממד הזמן היא שהזמן משחק תפקיד חשוב מאוד באופטימיזציה של פרסום. התנהגות הגלישה של היוזרים באתרים ברמת השעה במהלך היום, היום בשבוע ואפילו התקופה בשנה, עשויה להוות תפקיד מכריע בקבלת ההחלטות של המפרסמים.
- **Publisher** – חברות או גופים שמנהלים המון אתרים (Domains). המפתח הראשי בטבלה הוא- PublisherID. והשדה הנוסף הוא PublisherName.
- **Domain** - האתרים שהיוזרים גולשים בהם ושמפרסמים בהם את הפרסומות. המפתח הראשי – DomainID. הטבלה מכילה מפתח זר של PublisherID, על מנת לקשר אותה עם טבלת האב שלה – Publisher. את הקשר הזה יצרנו על מנת להריץ בקלות וביעילות שאילתות כמו איזה Domain שייך לאיזה Publisher.

- **OS** – מכשיר ההפעלה של אותו המשתמש. המפתח הראשי של הטבלה - OSID. והשדה הנוסף הוא OSName.

המפתחות הראשיים של הממדים כולם מספרים סידוריים. הסיבה שבחרנו במספרים סידוריים היא על מנת לוודא את החד-חד ערכיות שלהם. כמות ה DATA שמתקבלת כל יום היא עצומה ומספרים סידוריים זו דרך קלאסית לאכוף את החד-חד ערכיות שלהם .

<u>הצגת סכמת הנתונים</u>




הסכמה שלנו מסודרת באמצעות שיטת פתית שלג. בזכות היתרונות של סכמת פתית שלג, ניתן לבצע שאילתות מורכבות. נוכל לבצע ניתוחים מעניינים על פי מדינה או מכשיר בקלות, במקום להכליל את המידע לטבלת העובדות. בשיטת סכמת פתית שלג, ניתן לבצע שאילתות מורכבות בצורה מהירה ויעילה. ניתן לשמור על מבנה מופרד לטבלאות ה Dimensions ולטבלת העובדות. בנוסף, זה מקל על תחזוקה וניהול של המערכת.

במרכז נמצאת טבלת העובדות, ומסביבה טבלאות הממד. הסידור הזה מייעל את הביצועים של המערכת שלנו כיוון שכמות ה Joinשהמערכת צריכה לעשות בשביל לענות על שאלות שמאפיינות את הנתונים שלנו, כמו – כמה Domain יש לכל ,Publisher או כמה Users מתחברים מכל מדינה או משתמשים בכל מכשיר הפעלה וכן הלאה, מצטמצמים ל Join-של מקסימום 3 טבלאות. דבר זה מייעל משמעותית את הביצועים של המערכת.

בזכות המבנה של סכמת פתית שלג, הוספת טבלאות חדשות או שינויים עתידיים במבנה הנתונים יהיו קלים לביצוע.

כיוון שהנתונים שלנו בנויים מאירועים שהם סוג של טראנזקציות, אנחנו מנהלים את המידע ממש ברמת הרשומה הבודדת, ואנחנו למעשה מסתכנים בבעיות מסוג Anomaly הכנסה של רשומות חדשות, עדכון של רשומות קיימות ומחיקה של רשומות עלולות לסבך את העדכון ולפגום בנתונים .

מסיבה זו בחרנו לנרמל את המודל של הנתונים ולעשות דה-נורמליזציה לסכמת פתית שלג, ולפרק את הנתונים לטבלאות קטנות. בנוסף, סכמה זו מבטיחה לנו Integrity של הנתונים - מבטיחה אחידות והתאמה מדויקת בין הנתונים, מורידה את הכפילויות ומונעת התנגשויות בנתונים.
בסך הכל סכמת פתית שלג מתאימה בצורה מושלמת לדרישות שלנו כארגון.

מדיניות העדכון

בבעיית ה- Slowly Changing Dimensions, הממדים משתנים עם הזמן ומתרחשים עדכונים בטבלאות הממד, כמו פרטים על היוזרים או על המפרסמים – נמיין לפי השיטות הנפוצות ביותר את הטבלאות שלנו (Type 1/Type 2):

**Users : Type 2**

מידע על היוזרים כולל את ההיסטוריה שלהם ואת השינויים במאפיינים שלהם. כאשרר יש שינויים בנתונים שלהם, נוסיף שורה חדשה עם המידע המעודכן.

**Country : Type 0**

אין צורך לעדכן את המידע כיוון שמדובר בשמות של מדינות.

**Date : Type 0**

אין צורך בעדכון של התאריכים.

**Domain: Type 1**

ברוב המקרים נתוני הדומיין קבועים ואינם משתנים, לכן אנו נבחר לעדכן את הנתונים ישירות במקרים בהם הם משתנים.

**Publishers : Type 2**

במידה ויש שינוי במאפיינים או בהיסטוריה של המפרסמים נוסיף שורה חדשה לטבלה עם המידע המעודכן.

**OS : Type 2**

במידה ויהיה שינוי במכשיר ההפעלה של המשתמש, נבחר לייצר גרסה חדשה של הרשומה כולה.