

קורס מפתחי BI

מסמך מסכם לפרויקט

**מגישים:**

**רונן צור, נליה סגל, דבי סופר**

יוני 2021

1. על הפרויקט

הפרויקט מדמה סביבה ארגונית של חברת סלולר גדולה בארץ.

הפרויקט המוגש מדמה מערכת BI עבור חברת הסלולר אשר מסייעת למנהלי החברה בקבלת החלטות קריטיות כדוגמת:

* תמחור נכון של שירותים שונים
* העברת משאבים לשירות מסוים על חשבון אחר
* מציאת לקוחות הנוטים לנטוש את החברה ולתת מענה לפני הנטישה
* הקצאת משאבים נכונה כדי לחסוך בהוצאות

1. שלב התכנון ויישום סביבת העבודה
   1. אפיון הסביבה העסקית

חברת הסלולר (The Voice) מספקת שירותי מידע וקול ללקוחותיה בארץ ובעולם ולחברה מערכות תפעוליות רבות המשמשות את החברה במתן הפתרונות הרבים אותם מצפים לקוחותיה לקבל לפי רמת השירות הגבוהה מאוד המקובלת בשוק. המערכת התפעולית מורכבת מטבלת רשומות שיחה ענקית (CDR) ממנה ניתן לגזור את כל המידע הרצוי בתוספת של טבלאות מידע עסקי כגון:

* לקוחות
* שרותים
* אזורים
* מקורות

מקורות המידע של הפרויקט הינם בסיס הנתונים התפעולי של החברה + קבצי אקסל וקבצי CSV  שונים.

החברה מבקשת ליצור פרויקט BI גדול מאוד אשר ישמש מודל שיעזור למשתמשים בקבוצות שונות וישמש כמערכת – DSS – Data Support System כדי לעזור להם בהחלטות הנדרשות מהם:

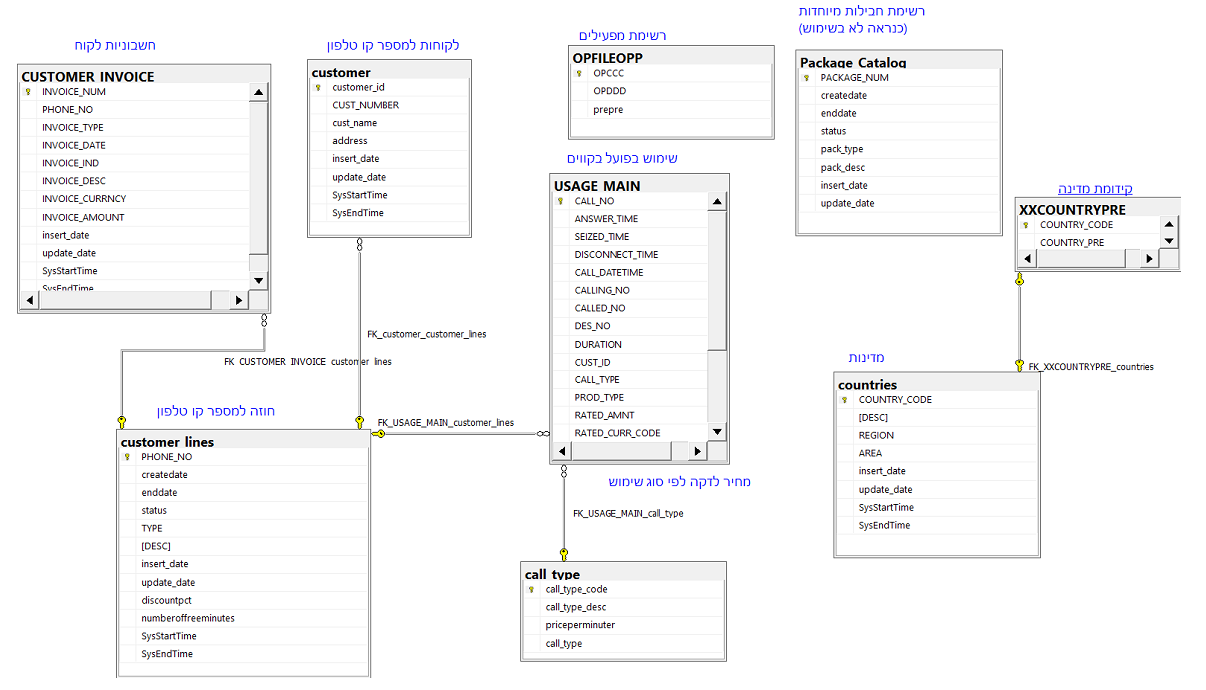
המודל יספק מענה לשאלות עסקיות הבאות:

* ניתוח של סה"כ תנועה של שימושי קול לפי מדינה
* ניתוח שימושים שונים לפי סוגי שיחות – call types
* ניתוח דקות לפי מספר הטלפון
* ניתוח שימוש duration לפי חבילה (לפני ואחרי הנחה)
* ניתוח של מספר שיחות ומספר לקוחות ביום ושעה
* ניתוח לקוחות רדומים – שאינם משתמשים בשרות
* ניתוח תנועה של דקות יוצאות ברמה יומית/ חודשית/ שנתית

1. מבנה הנתונים ודיאגרמה של המערכת התפעולית
   1. מערכת המתגים הארגונית:

|  |  |  |  |
| --- | --- | --- | --- |
|  | מקור – ישויות (טבלאות)במערכת התפעולית | תיאור | דוגמא |
| .1 | –OPFILEOPP קובץ מפעילים על פי קידומת | מתאר את המפעילים הקיימים בארץ על פי קידומת | 050 – פלאפון |
| .2 | קובץ –XXCOUNTRYPRE  קידומות מדינה | מתאר את המדינות  בעולם עלפי קידומת. | ישראל – 972 |
| .3 | CUSTOMER\_INVOICE | טבלה המתארת את חשבוניות הלקוח | חשבונית עבור לקוח 972520002375+ , עבור חודש דצמבר 2013 , ניתנה בינואר 2014 , חשבון לחיוב 400 שקלים |
| .4 | Package\_Catalog | טבלת קטלוג חבילות | חבילה מספר 1 – חבילת משפחה, חבילה 2 – מסלול לחיילים לקוח שמספרו 972525154837+ שייך לחבילת משפחה, סטטוס פעיל, מקבל הנחה של 35% בשיחות בתוך המשפחה |
| .5 | customer\_lines | טבלת קווי לקוח |  |
| .6 | countries | קודי מדינה בעולם | מדינה: אנגליה , יבשת: אירופה |
| .7 | call\_type | סוגי שיחה/שירות - | Call - Cellular Call ,TEXT - Text Messaging |
| .8 | Customer | לקוחות האירגון | לקוח מספר 1 , מספ טלפון  972541185146+ , אלברט אינשטיין , תאריך הצטרפות לחברה 5.2.2010 |
| .9 | USAGE\_MAIN | מכילה את כל התנועה האירגונית |  |

* 1. תרשים ERD של המערכת הארגונית



1. הנחות בפרויקט

* **קידומת מספר טלפון:** קידומת בת שלוש ספרות של מספר הטלפון מהוה אינדיקציה חד ערכית לשם החברה (לדוגמא 005 = פלאפון), אנו נתעלם מאופציית שיוך מספרים הקיימת בחברות אחרות.
* **קידומת מדינה:** קידומת מדינה היא עד כדי שלוש ספרות (לדוגמא 1 = אמריקה, 972 = ישראל)
* **מספר טלפון של אדם מסויים:** מספר הטלפון של אדם מסוים מהוה יחוס חד חד ערכי למספרו כלקוח, כלומר לקוח מזוהה ע"י מס' טלפון אחד. במידה ובטבלאות יש יותר מאחד, יש לקחת את המעודכן ביותר בלבד.
* **הנחות נוספות כפי שהובנו על-ידנו:**
  + נדרשנו לשנות את אורך ה- שדה של מספר הטלפון בטבלת USAGE ל- Varchar(20) כדי לייצר קשרים בין הטבלאות.
  + נדרשנו לשנות את אורך השדה call\_type בטבלת usage\_main ל- varchar(10) כדי לחבר אותו לטבלת call\_types.

1. סכמת הכוכב של מחסן הנתונים
2. שלבים בפיתוח הפרויקט
   1. תיקון בעיות נרמול בבסיס הנתונים התפעולי
   2. פיתוח בשפת T-SQL

להלן השלבים שהוכנו תוך שימוש בשפת SQL:

* + - 1. מחסן הנתונים - סקריפט מלא ליצירת כלל האובייקטים – DB , טבלאות וכו'

קובץ: \_\_\_\_\_\_\_\_ בתיקיית \_\_\_\_\_\_\_\_

* + - 1. ג'ובים - סקריפט מלא ליצירת כלל הג'ובים הנדרשים לתהליכי ה ETL

קובץ: \_\_\_\_\_\_\_\_ בתיקיית \_\_\_\_\_\_\_\_

* + - 1. לולאה – סקריפט מלא ליצירת רשומות חדשות בטבלת USAGE\_MAIN

קובץ: \_\_\_\_\_\_\_\_ בתיקיית \_\_\_\_\_\_\_\_

הרשומות להיות תואמות לשאר הרשומות בטבלאות האחרות נוצר Data עבור שנה אחת תחת שדות ה TIME וה-DATE

* + 1. כלי ה ETL של מייקרוסופט – SSIS

תהליך ה- ETL מעביר את הנתונים מהמערכת התפעולית למחסן הנתונים במנגנון מתוזמן.

* 1. העברת נתונים מבסיס הנתונים התפעולי למחסן הנתונים
     1. תהליך ה - ETL

מפת הגזירה כוללת את כל השלבים MRR: STG: DWH

בתוך בסיס הנתונים של מחסן הנתונים ישנן 3 סכמות: MRR, STG , DW בנוסף בסיס הנתונים מכיל סכמות לחלוקה בין טבלאות Integration  וטבלאות Dim ו- Fact.

קבצי ה- Csv נטענים באופן דינאמי ורשומות חדשות נטענות אוטומטית לבסיס הנתונים של ה- MRR. עם תום הטעינה הקובץ שטופל מועבר לתיקית קבצים שכבר נטענו.

תהליך ה- ETL מורץ באופן יומי באמצעות SCHEDULER.

נוצרה טבלת תיעוד טעינות בשם lineage ובה, מועד הריצה האחרון, שם הטבלה, שם המחשב  המריץ, שם משתמש,  שם ה package שהופעל.

שלפנו רק DELTA מטבלאות שהם Temporal Tables.ד

בניית מחסן הנתונים נעשתה בהתאם לדרישות לבנית מחסן הנתונים המפורטות בקובץ ה- EXCEL המצורף The Voice Source to Target.

* + - 1. MRR – העתקה מלאה של הנתונים מבסיס הנתונים התפעולי וקבצי האקסל וה- CSV אל סכמת ה- MRR שנמצאת ב- DWH

טבלאות המופיעות בסכמה של ה- MRR:

|  |  |  |  |
| --- | --- | --- | --- |
| TABLE\_CATALOG | TABLE\_SCHEMA | TABLE\_NAME | TABLE\_TYPE |
| TheVoice\_MRR\_STG\_DWH | MRR | Countries | BASE TABLE |
| TheVoice\_MRR\_STG\_DWH | MRR | Customer | BASE TABLE |
| TheVoice\_MRR\_STG\_DWH | MRR | CUSTOMER\_INVOICE | BASE TABLE |
| TheVoice\_MRR\_STG\_DWH | MRR | customer\_lines | BASE TABLE |
| TheVoice\_MRR\_STG\_DWH | MRR | Package\_Catalog | BASE TABLE |
| TheVoice\_MRR\_STG\_DWH | MRR | USAGE\_MAIN | BASE TABLE |
| TheVoice\_MRR\_STG\_DWH | MRR | XXCOUNTRYPRE | BASE TABLE |

* + - 1. טעינה דינאמית CSV

ישנם 2 סוגי קבצי CSV. הראשון OPFILEOPP והשני call\_type. בשלב ה- MRR נבנה תהליך ETL דינאמי עבור כל סוג קובץ אשר נטען לטבלה מתאימה תהליך הטעינה הינו דינאמי ובוחן את הרשומות שמתקבלות בכל קובץ ומוסיף רק רשומות חדשות. בסיום טעינה הקבצים מועתקים לספריית \_\_\_\_\_\_\_. בנוסף שם הקובץ משתנה לשם מקורי בתוספת חתימה של תאריך ושעה. בתחילת התהליך נוצר סקריפט להכנת כל סביבת העבודה מוכנה. כולל יצירת ספריות, והעתקת קבצים.

* כל קובץ נטען לתיקייה ספציפית על השרת בו מותקן server sql - יש לבצע משיכה יומית משם.
* שמות הקבצים שונים מפעם לפעם.
* קובץ OPFILEOPP מכיל את ה opfileopp כמחרוזת.
* קובץ call\_type מכיל את ה-  call\_type כמחרוזת.
  + - 1. STG – ניקוי והמרות של הנתונים, חישוב עמודות Data Cleansing

בוצע תהליך Data Cleansing מלא לרבות המרת ערכי null

אחרי כל ,lookup במקרה שלא נמצא ערך יש להשלים עם 1- או unknown

* + - 1. DWH - חלוקה לפי DIM\FACT
  1. פיתוח קובייה או מודל בעזרת כלי הSSAS
  2. דוחות מנהלים בעזרת מחולל דוחות
     1. Reporting Services
     2. Power Bi
  3. הקמת פורטל מקומי לאתר הדוחות

הגעה לפורטל באמצעות הקישור הבא:

1. צילומי מסך מתהליך בניית הפרויקט
   1. SSIS
   2. פיתוח Tabular Model מעל DW

מודל "הטבולר" הוא למעשה החלק המאפשר גישה מהירה למשתמשי הקצה של מערכת ה- BI על מנת שיקבלו תשובות לכל השאלות העסקיות של הארגון במהירות רבה תוך הצגה גרפית יפה על מנת לקבל החלטות נכונות לעתיד החברה. בניית המודל, במסמך Target" To Source "TheVoice ישנה הגדרה של:

* + 1. usage dimension cube

נדרש לממש את הסעיפים שלהלן:

1. Rename שמות עמודות וטבלאות
2. הגדרת טבלת Table Date
3. יצירת יחסים בין הממדים וטבלאות ה Fact
4. וודא Type Data ו – Format מתאים לעמודות – יש לתקן במידת הצורך.
5. בצע Column By Sort בעמודות הנדרשות
6. "החבא" עמודות במידת הנדרש
7. למודל יוגדרו Measures שיתבססו על טבלת ה :fact
   1. Count
   2. Count distinct on customer
   3. Call Duration
   4. Billable Call Duration
   5. Amount
   6. Billable Amount
   7. Call Duration Average
   8. :Calculated Columns יוגדרו למודל.
   9. Billable Amount – Amount :Discount
   10. Discount / Amount :Discount Percentage
8. יוגדרו 3 KPI וערכי מטרה דינמיים
9. יש ליצור 3 פרספקטיבות ובהן כל המימדים והמדדים הרלוונטיים
10. בניית המימדים הבאים: ניתן לוותר על היררכיה ולבנות כל מימד בטבלה אחת.
11. מימד מדינות  - היררכיה ו- attributes:
    1. region->area->country
    2. Areas
    3. Countries
    4. Regions
12. מימד זמן - היררכיה ו- attributes:
    1. Year -> month -> Date
    2. Months
    3. Days In Week
13. מימד שעות - היררכיה ו- attributes:
    1. Hours -> Minutes
    2. Hours
    3. Minutes
14. מימד לקוחות  - היררכיה ו- attributes:
    1. Packages -> Customers
    2. Country -> Customers
    3. Operator -> Customers
    4. Customers
15. מימד סוג שיחה  - היררכיה ו- attributes:
    1. Call Type Category -> Call Types
    2. Call Types
16. מימד חבילה  - היררכיה ו- attributes:
    1. Package Status -> Packages
    2. Packages
17. מימד מפעיל  - היררכיה ו- attributes:
    1. Operators
18. מימד מקור שיחות - היררכיה ו- attributes:
    1. Call Origin Type

יש להקפיד להביא את התיאור (Description) ולא להציג למשתמש קצה קוד או מספר.

1. בנית הקובייה
2. במסמך " The Voice Source To Target " ישנה הגדרה של cube dimension usage
3. לקובייה יהיה measure group אחד, שיתבסס על טבלת ה fact. ה measures שיש להביא:
   1. Count
   2. Count distinct on customer
   3. Call Duration
   4. Billable Call Duration
   5. Amount
   6. Billable Amount
   7. Reporting Services

יש ליצור לפחות שישה דוחות שונים לבחירה מתוך בסיס השאלות העסקיות הבאות:

1. ניתוח של סה"כ תנועה של שימושי קול  לפי מדינה
2. ניתוח שימושים שונים לפי סוגי שיחות types( )call
3. ניתוח דקות לפי מספר הטלפון – SSRS
4. ניתוח שימוש duration לפני ואחרי ההנחה
5. ניתוח של כמות שיחות וכמות לקוחות ביום ושעה
6. ניתוח לקוחות רדומים – שאינם משתמשים בשרות
7. ניתוח תנועה של דקות יוצאות ברמה יומית/ חודשית/ שנתית

את כל הדוחות יש לפרוס (deploy) בשרת הדוחות – צילומי מסך ישלחו בהתאם במסמך המקדים

* 1. דגשים לדוחות ה SSRS/PBI

יש ליישם את הפונקציונאליות שלהלן:

הקמת Server Report BI Power – קונפיגורציה מלאה

יישום האובייקטים הבאים בשרת הדוחות: KPI דינאמי  ,   Source ,Data Set ,Data SLICER

פרמטרים לדו"ח  מתוך רשימת בחירה

עבור הסעיפים הקשורים לשרת הדוחות server report BI Power - נדרשים צילומי מסך מלאים במסמך המקדים הכוללים כתובת URL

* + 1. Dax Calculation Formulas

יש להדגים שימוש של תשאול המודל בעזרת שימוש ב- dax queries

עבור הסעיפים הקשורים ל -  DaxStudio נדרשים צילומי מסך מלאים במסמך המקדים

* 1. PBI

1. בעיות שנתקלנו בהן בפרויקט ודרכי ההתמודדות עמן