

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:

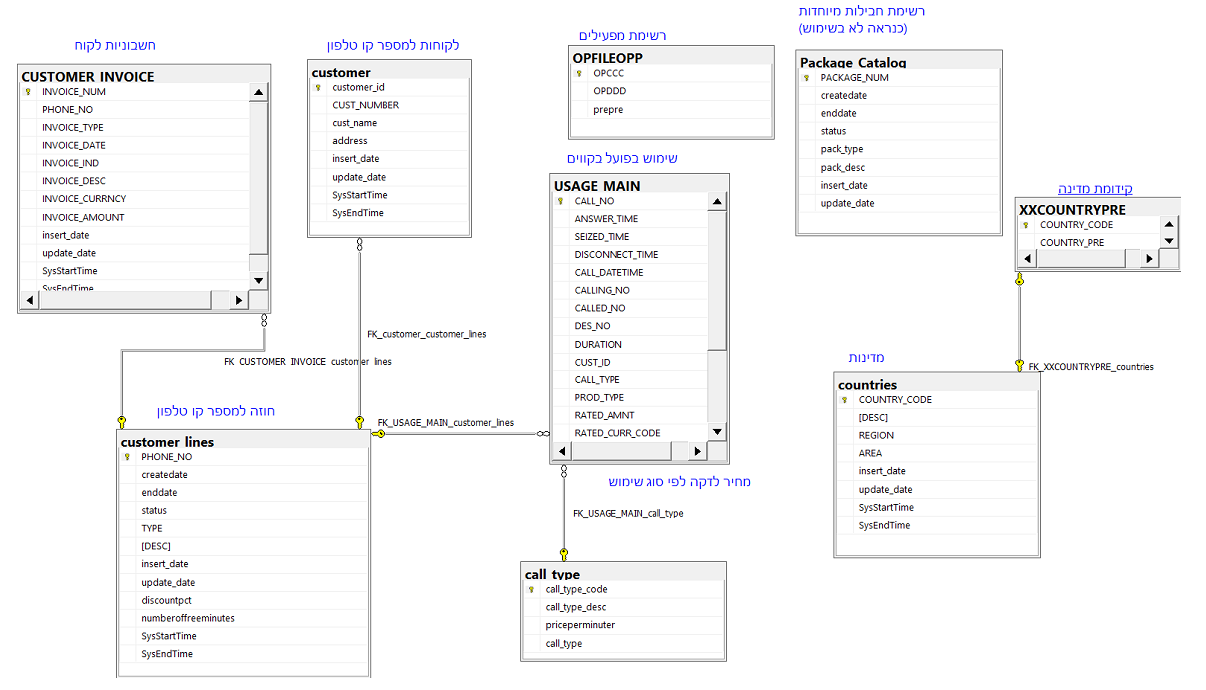
המודל יספק מענה לשאלות עסקיות הבאות:

* 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 |  |

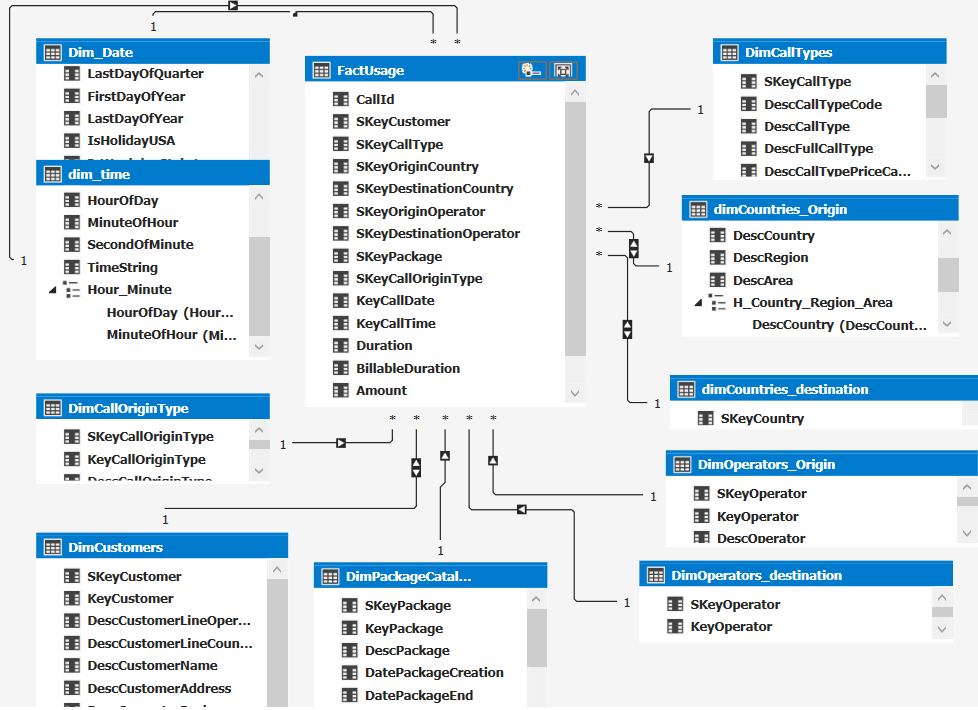
* 1. תרשים ERD של המערכת הארגונית



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.

1. סכמת הכוכב של מחסן הנתונים



1. שלבים בפיתוח הפרויקט
   1. תיקון בעיות נרמול בבסיס הנתונים התפעולי

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. פיתוח בשפת T-SQL

להלן השלבים שהוכנו תוך שימוש בשפת SQL (כל הקבצים הועלו ל- GIT כנדרש):

* + - 1. בניית סקריפט מלא ליצירת כלל האובייקטים – DB , טבלאות וכו'
      2. סקריפט מלא ליצירת כלל הג'ובים הנדרשים לתהליכי ה ETL
      3. לולאה – סקריפט מלא ליצירת רשומות חדשות בטבלת USAGE\_MAIN
  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.

בניית מחסן הנתונים נעשתה בהתאם לדרישות לבנית מחסן הנתונים המפורטות בקובץ האקסל: - 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 DONE.

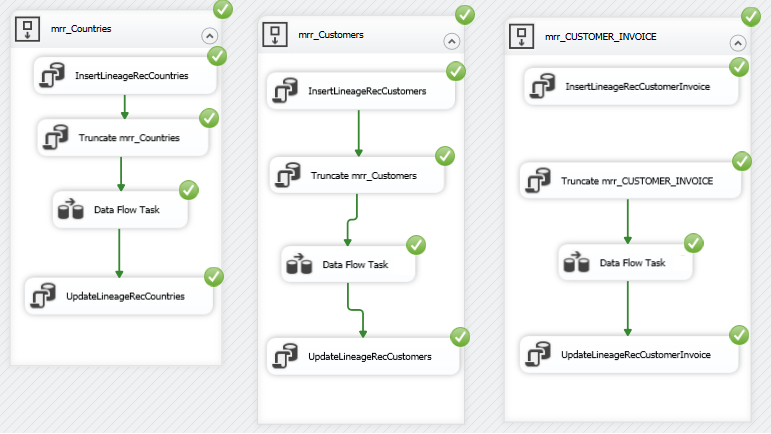
* קובץ OPFILEOPP מכיל את ה opfileopp כמחרוזת.
* קובץ call\_type מכיל את ה-  call\_type כמחרוזת.
  + - 1. STG – ניקוי והמרות של הנתונים, חישוב עמודות Data Cleansing

בוצע תהליך Data Cleansing מלא לרבות המרת ערכי null

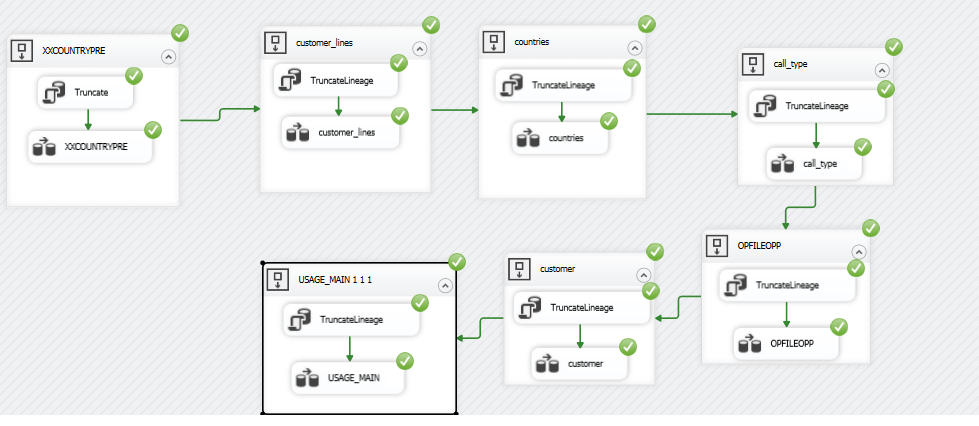
* + - 1. DWH – בוצעה חלוקה לפי DIM\FACT
  1. פיתוח קובייה או מודל בעזרת כלי הSSAS

ראה פירוט להלן בסעיף 7.2.

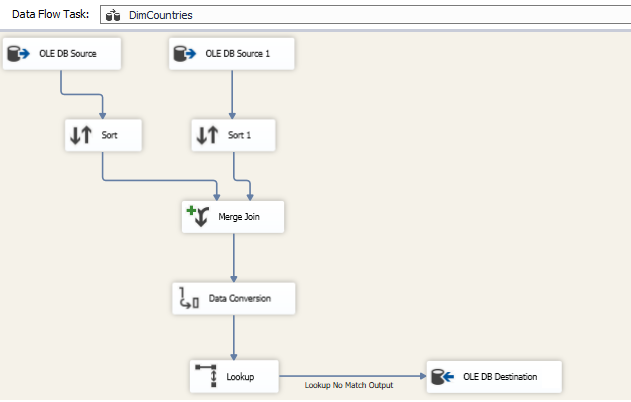
1. צילומי מסך מתהליך בניית הפרויקט
   1. SSIS
      1. מסך לדוגמא של שלב ה- MRR

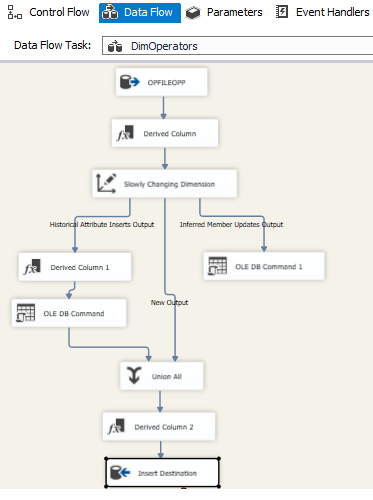


* + 1. מסך לדוגמא של שלב ה- STG



* + 1. מסך לדוגמא של שלב ה- DWH

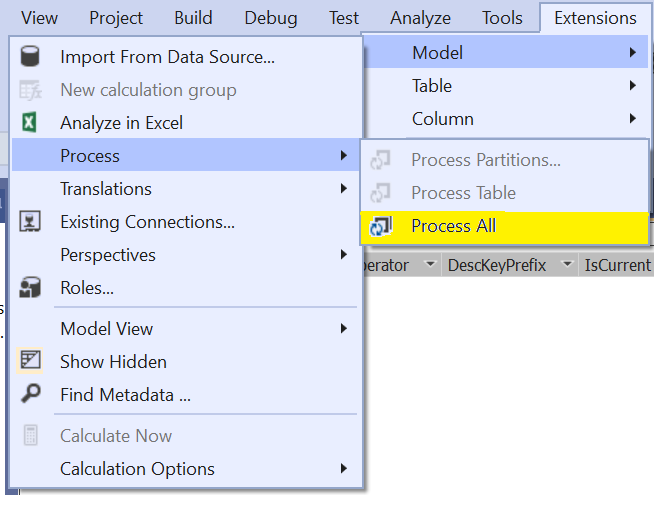




* 1. פיתוח Tabular Model מעל DW

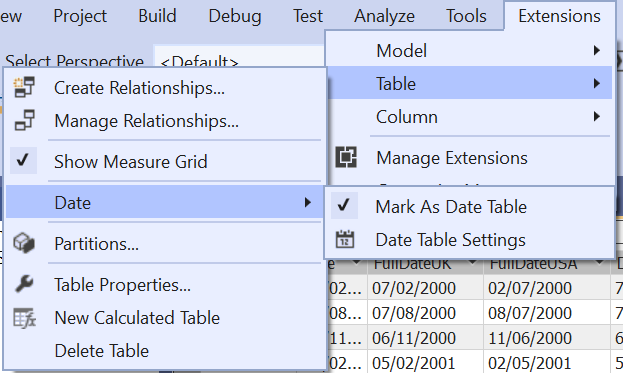
מודל "הטבולר" הוא למעשה החלק המאפשר גישה מהירה למשתמשי הקצה של מערכת ה- BI על מנת שיקבלו תשובות לכל השאלות העסקיות של הארגון במהירות רבה תוך הצגה גרפית יפה על מנת לקבל החלטות נכונות לעתיד החברה.

רענון הקשר בין המודל לבין בסיס הנתונים:

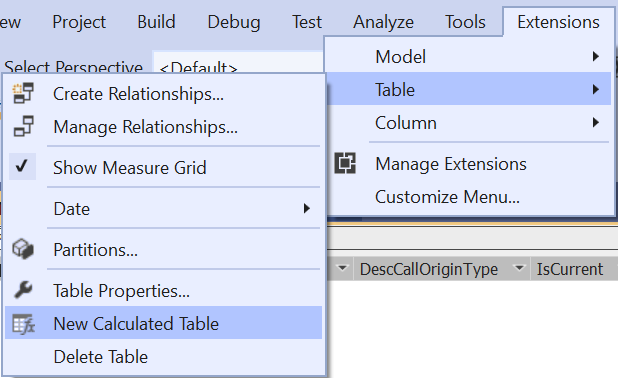


* + 1. להלן הסעיפים שמומשו:

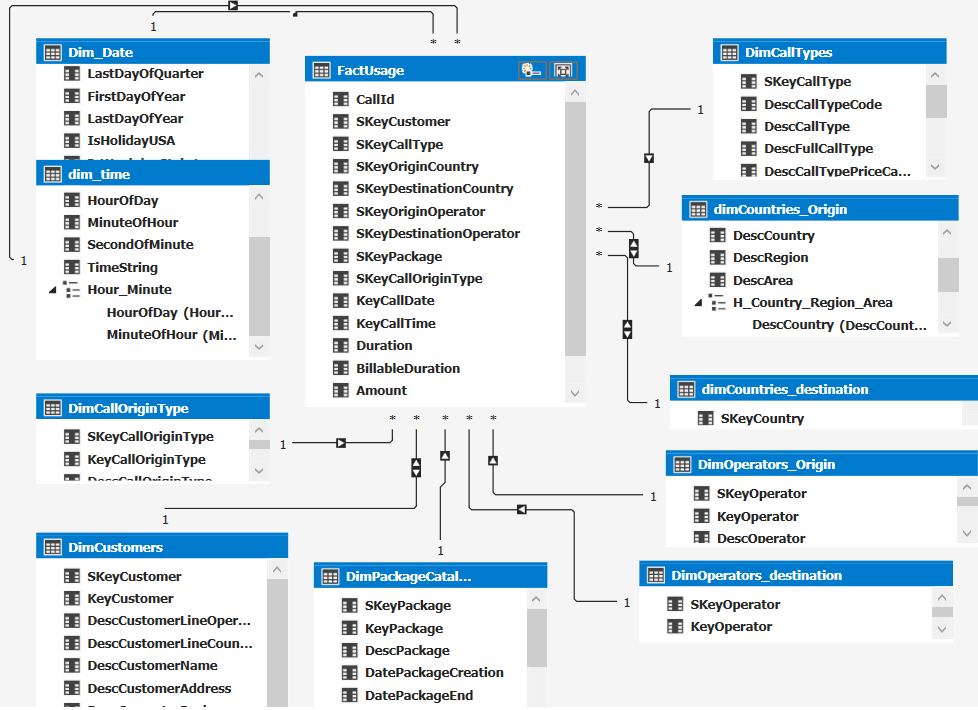
1. Rename שמות עמודות וטבלאות
2. הגדרת טבלת Table Date - בוצע על טבלת Dim\_Date באמצעות תפריט: Extentions



1. יצירת יחסים בין הממדים וטבלאות ה Fact
   1. לצורך יצירת הקשרים היינו צריכים לייצר מספר טבלאות משוכפלות ב- INMEMMORY. כך למשל עבור מדינות מקור ליעד שכפלנו בזיכרון את טבלת המדינות. כך גם לגבי מפעילים. האופן שבו הדבר בוצע הוא כדלקמן:
      * ראשית שינינו את שם הטבלה אותה רצינו לשכפל (הוספנו לה את הסיומת Origin.
      * לאחר מכן עברנו לתצוגת ה- Model ובאמצעות תפריט: Extentions יצרנו את הטבלה המשוכפלת בזכרון:



* 1. בדרך זו התאפשר לנו לייצר קשרים בין מספר שדות בטבלת ה- FACT לבין "אותה" טבלה בטבלאות ה- DIM.
  2. חשוב לציין שהטבלה איננה באמת משוכפלת אלא רק נכנסת מספר פעמים לזכרון.
  3. משפט DAX רבלנטי ליצירת טבלת זכרון משוכפלת:  
      = CALCULATETABLE(DimOperators\_Origin)
  4. תרשים הקשרים בין הטבלאות:



* 1. עבור כל שדה SK המופיע בטבלת FACT הוספנו עמודות ל- FACT בהן שתלנו את נוסחת: Related המפנה לכל השדות הרלבנטיים מטבלת ה- DIM.
  2. לאחר שהוספנו את כל השדות כמופיע בסעיף הראשון החבאנו את שדות ה- SK באמצעות לחצן ימני: "Hide From Client Tools".

1. למודל הוגדרו Measures המתבססים על טבלת ה :fact
   1. Count

NumberOfCalls:=count(FactUsage[CallId])

* 1. Count distinct on customer

NumberOfCustomersMakingCalls:=DISTINCTCOUNT(FactUsage[SKeyCustomer])

* 1. Call Duration

SumofCallduration:=sum(FactUsage[Duration])

* 1. Billable Call Duration

SumOfBillableDuration:=sum(FactUsage[BillableDuration])

* 1. Amount

SumOfAmount:=sum(FactUsage[Amount])

* 1. Billable Amount

SumOfBillableAmount:=sum(FactUsage[BillableAmount])

* 1. Call Duration Average

CallDurationAverage:=AVERAGE(FactUsage[Duration])

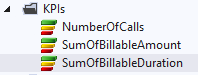
* 1. :Calculated Columns יוגדרו למודל.
  2. Billable Amount – Amount :Discount

=FactUsage[Amount]-FactUsage[BillableAmount]

* 1. Discount / Amount :Discount Percentage

=if(FactUsage[Amount]<>0, FactUsage[Discount]/FactUsage[Amount],0)

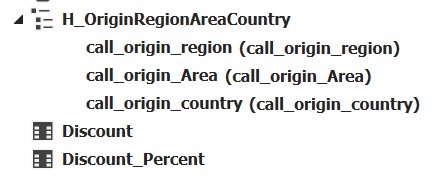
1. הוגדרו 3 KPI וערכי מטרה



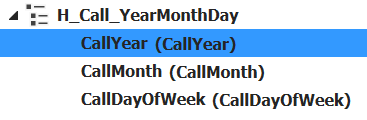
1. הגדרנו 3 פרספקטיבות ובהן כל המימדים והמדדים הרלוונטיים



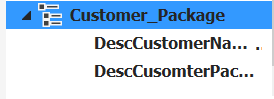
1. נבנו המימדים הבאים:
2. מימד מדינות  - היררכיה ו- attributes:
   1. region->area->country
   2. Areas
   3. Countries
   4. Regions



1. מימד זמן - היררכיה ו- attributes:
   1. Year -> month -> Date
   2. Months
   3. Days In Week

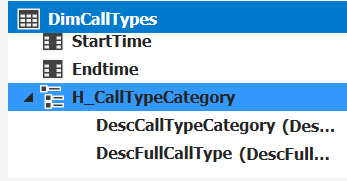


1. מימד לקוחות  - היררכיה ו- attributes:
   1. Packages -> Customers

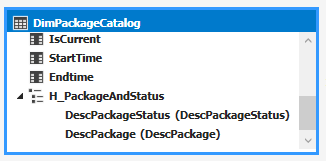


* 1. Country -> Customers
  2. Operator -> Customers
  3. Customers

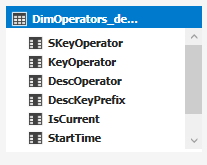
1. מימד סוג שיחה  - היררכיה ו- attributes:
   1. Call Type Category -> Call Types



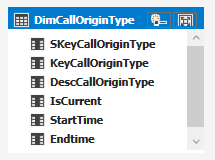
1. מימד חבילה  - היררכיה ו- attributes:
   1. Package Status -> Packages



1. מימד מפעיל  - היררכיה ו- attributes:

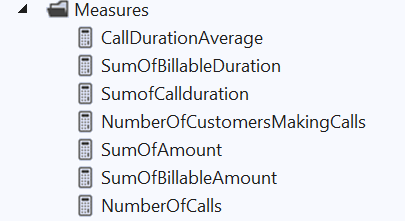


1. מימד מקור שיחות - היררכיה ו- attributes:
   1. Call Origin Type

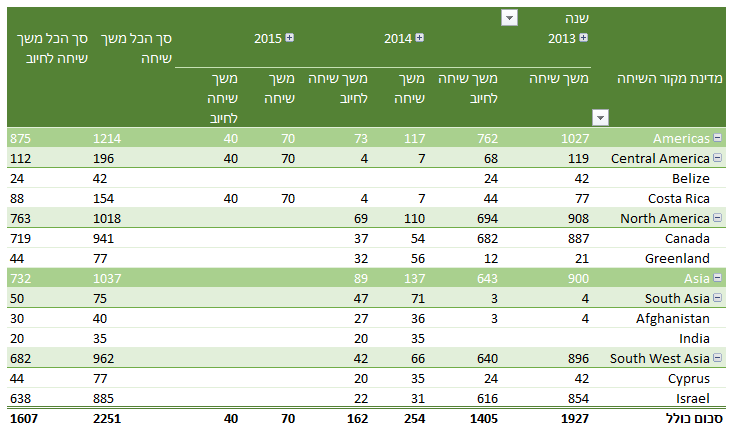


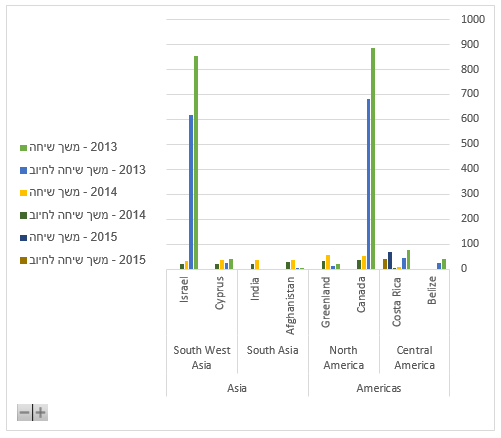
בטבלת ה- FACT מופיעים רק תיארים.

1. בנית הקובייה
2. במסמך " The Voice Source To Target " ישנה הגדרה של cube dimension usage
3. לקובייה יש measure group אחד, שיתבסס על טבלת ה fact. ה measures שנבנו:

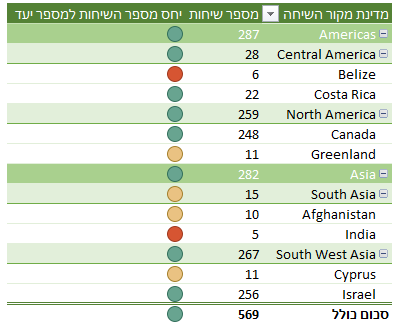


1. דוחות וניתוחים באקסל המתבססים על הקוביה:
   1. ניתוח משכי שיחות ומשכים לחיוב בחלוקה לפי שנים ואזורים גיאוגרפיים שונים:





* 1. השוואת מספר שיחות מכל מדינה ואזור ליעד מספר השיחות הרצוי



* 1. נבנו דוחות -PBI להלן צילומי מסך נבחרים:

