פרויקט מעשי קורס DATA ANALYST

א. הפרויקט

מטרת הפרויקט הינה לעזור לעריית ניו יורק לנתח את פעילות אכיפת עבירות החניה בשטחה. לשם כך, עליכם להשתמש בכלים השונים שלמדתם במסגרת מסלול ההכשרה, על מנת שתוכלו לתת תמונה כוללת של הפעילות.

בעיריית NY ניתנים מדי שנה יותר מ-10 מיליון (!) דוחות חנייה בגין עבירות חניה שונות. על מנת להקל על עבודתכם האנליטית, נבחר באופן רנדומלי אחוז אחד של הנתונים. הנתונים נלקחו מאתר- NYC Open data.

ישנן דרכים שונות לביצוע פרויקט מסוג זה. בחרו בדרך הכי נכונה לביצוע המשימה - דרך המשלבת ההיבטים השונים של עבודת אנליסט בעולם ה-BI. הפרויקט המעשי המסכם, מדמה את עבודת האנליסט המתחיל בפרויקט חדש של ניתוח הנתונים ומעוניין גם להקים בסיס נתונים מסודר לצורך ניתוחים נוספים בארגון.

המשימות בפרויקט הנן כדלהלן:

- הקמת בסיס נתונים בשרת SQL SERVER
- טעינת הנתונים לבסיס הנתונים באמצעות כלי ETL (SSIS)
- הכנת הנתונים ובניית ויזואליזציות באמצעות ה-Power BI

*מילה על איכות הנתונים: בניגוד לבסיסי הנתונים ששימשו אתכם במהלך הקורס, הנתונים בפרויקט מעשי זה הינם נתונים אמתיים על כל מה שמשתמע מכך: ערכים חסרים, קודים שאין עבורם תרגום, קודים שגויים. עליכם להתמודד עם מצבים אלה במהלך העבודה על הפרויקט.

ב. המשימות בפרויקט

1. מקורות הנתונים ומודל הנתונים של הפרויקט

- מודל הנתונים של הפרויקט נבנה על פי העקרונות של המידול הממדי במבנה מסוג פתית שלג. ראו נספח א. על מנת להקל על עבודתכם, הוגדרו לכם הטבלאות במודל הנתונים.
- נתוני המקור לפרויקט הנם בפורמט CSV. ראו נספח א. מקורות הנתונים לפרויקט.

2. הקמת הטבלאות בתוך ה-DB

- הטבלאות ייבנו בבסיס הנתונים ייעודי בתוך שרת ה-SQL SERVER – שם בסיס הנתונים: DWH_DATA_ANALYST
- יש להשתמש בפקודות SQL להקמת הטבלאות.
- לכל טבלה יוגדר Primary Key (חובה). לשם הנוחות נבחר להשתמש בחלק מהטבלאות במפתחות שאינן מסוג INTEGER.

תוצר השלב: תוכנית SQL ליצירת ה-DB ולהקמת הטבלאות ב-DB

3. בניית תהליכי SSIS לאכלוס וטיוב הנתונים בבסיס הנתונים

יש לבנות תהליך מקצועי של טעינת הנתונים באמצעות תוכנת ה-SSIS.

מאפיינים כלליים לתהליך אינטגרציית הנתונים:

- יש לבנות פרויקט חדש ב-SQL SERVER DATA TOOLS (Visual Studio) בשם **DWH_PROJECT_SSIS**
- בפרויקט יוגדרו מספר Packages בהתאם למספר הטבלאות לאכלוס ב-DWH.
- חובה לאפשר חזרה על תהליך הטעינה באמצעות מחיקה אוטומטית של הרשומות מהטבלה לפני הטעינה (על ידי שימוש ב-Execute SQL Task ב-Control Flow).
- רוב הטבלאות נטענות אחת לאחת מקבצי המקור בפורמט ה-CSV (ראה נספח א.) אלא אם כן צוין אחרת.
- נתוני המקור לטבלת ה-FACT מחולקים לשלושה קבצים. יש לחברם יחד לטבלה אחת.
- **ממד המיקום**: קובץ ה-CSV אינו מכיל את כל השדות הנדרשים למילוי הנתונים. יש למלא את השדות החסרים (שם הרחוב, עיר, וקוד המדינה):
  - שמות הרחובות ישלפו מקובץ Street_codes.csv. שימו לב למפתח החיבור (רובע וקוד רחוב).
  - שם העיר והמדינה אחידים בכל הטבלה (New York City ו-NY).

תוצרי השלב:

- פתרון SSIS לאכלוס וטיוב הטבלאות בבסיס הנתונים
- בסיס נתונים מאוכלס בנתונים

4. הכנת הנתונים ובניית דוחות ב- Power BI for Desktop

a. שלב הכנת הנתונים

מקור הנתונים העיקרי לדוחות הינו בסיס הנתונים DWH_DATA_ANALYST שנבנה בשלב הראשון של הפרויקט. יש לטעון את הנתונים מבסיס הנתונים לתוך מודל הנתונים של ה-Power BI.

- יש ליצור קשרי גומלין מתאימים בין כל הטבלאות במודל (ראה נספח ג')
- נתוני דוחות החנייה יסוננו בשלב הטעינה למודל על מנת לשמור את הדוחות שניתנו בשנת 2015 ובשנת 2016 בלבד.
- במודל שנבנה חסר ממד תאריכים. יש להגדיר טבלה חדשה במודל עבור ממד זה באמצעות בניית טבלה מחושבת ושדות מחושבים. ראו דוגמה בקישור הבא:

http://www.mssqltips.com/sqlservertip/4857/creating-a-date-dimension-table-in-power-bi/

- על מנת להעשיר את הנתונים לדוחות, יש לטעון ל-Power BI נתוני תיאור עבירת החנייה מקובץ ה-CSV. יש לשמור על סטנדרט שמות הטבלאות והשדות כפי שמתואר בנספח ג'. לאחר טעינת הטבלה למודל הנתונים, יש ליצור קשר גומלין עם הטבלה הרלוונטית.
- התאריכים בטבלאות הינם מסוג טקסט. יש להפוך את כל התאריכים בטבלאות לשדות מסוג DATE.
- כמו כן, יש להפוך שדה הזמן לשדה מסוג TIME.

b. שלב בניית הדוחות

יש להשקיע מאמצים במראה המקצועי לדוחות באמצעות ויזואליזציות מתאימות וברורות לצופה בהן, שימוש בכותרות, בעיצוב השדות (כולל עיצובים מותנים).

דוח 1 – ניתוח סוגי עבירות החנייה

הדוח יציג ניתוח כולל של סוגי עבירות החנייה וייתן מענה לשאלות העסקיות הבאות:

- מה הם 5 סוגי עבירות החנייה הנפוצות ביותר לאורך השנים? ומה אנו יודעים על כל שנה בנפרד? האם חל שינוי משנה לשנה?
- באיזה יום בשבוע יש יותר עבירות חניה? האם יש הבדל בין הרובעים השונים בעיר?
- באיזה שעות של היום (בפרקי זמן של שעתיים) יש יותר עבירות חניה? האם זה תלוי ברובע?
- הציגו השוואת מספר הדוחות משנה לשנה וקצב הגידול השנתי.
- הציגו טבלה המציגה את מספר המצטבר של הדוחות לאורך השנה (YTD) לפי רובעים.

- **שאלת בונוס:** הציגו ויזואליזציה המראה כמה רכבים ביצעו יותר מ-10 עבירות חניה, כמה בין 5 ל-9 וכמה מתחת ל-5 עבירות.

דוח 2 - ניתוח סוגי הרכבים המעורבים בעבירות החנייה

דוח זה אמור לספק תובנות עסקיות לגבי סוגי הרכבים המעורבים בעבירות חנייה.

- מה סוג הרכב (Body Type) המקבל הכי הרבה דוחות חנייה ב-NY?
  *תחשבו אם לא כדאי לקבץ את סוגי הרכב לקטגוריות.
- מה הוא גובה הקנס הממוצע לכל סוג רכב (או לקבוצת סוגי הרכב שיצרתם)?
- האם יש צבע רכב דומיננטי?
- מאיפה באים רוב הרכבים המעורבים בעבירות חניה? האם העירייה צריכה לשפר את ההסבר למי שאינו תושב מדינת NY (או המדינות הסמוכות לה)?
- **שאלת בונוס**: יש לייבא מהאינטרנט נתונים על רכבים בארה"ב ולנסות לקבוע האם ישנה נטייה לבעלי רכבים מחברות מסוימות לקבל יותר דוחות חנייה (כמובן ביחס להערכת מספר הרכבים מסוג זה בארה"ב).

דוח 3 - ניתוח גאוגרפי של עבירות החנייה של משאיות במנהטן

העירייה מעוניינת להבין איפה ומתי רוב עבירות החניה של משאיות (Delivery Trucks) מתבצעות ברובע מנהטן.

- הציגו על מפה גיאוגרפית את עבירות החניה שביצעו משאיות ברובע מנהטן עם חלוקה של שעות היום והלילה (תחשבו על חלוקה הגיונית של השעות). האם אפשר לבודד בצורה ברורה אזורים ו/או חלקי יום בעייתיים?
- מה הם עשרת הרחובות במנהטן שיש בהם הכי הרבה דוחות חניה עבור משאיות המספקות סחורה? אולי העירייה צריכה לחשוב על פתרונות חניה באזורים אלו?
- מה גובה הקנס הממוצע שמקבלים בעלי רכב אלו?

טיפ: על מנת להציג את נתוני המיקום על גבי המפה, יש לבנות טור מחושב המקבץ את מספר הבית, שם הרחוב, שם העיר, ושם המדינה.

דוח 4 (בונוס) - ניתוח כדאיות הכלכלית של פעילות האכיפה

מעבר לצורך לאכוף את עבירות החניה בעיר, עיריית NY מעוניינת לבחון את הכדאיות הכלכלית של פעילות אכיפת עבירות חניה בעיר. ההכנסות מהפעילות יחושבו באמצעות הנתונים במודל על גובה הקנסות. עלות הפעילות תחושב לפי עלות המשכורת של הפקחים (issuer) לפי השתייכותם למוסדות האמונים על האכיפה (Issuing Agencies). יש להניח תקורה של 5$ לדוח חניה. במידה ולא רשום בנתונים קוד זיהוי של פקח החניה, יש להניח עלות של 15 $ לעבירת החניה.

הרווח יהיה מחושב כהפרש בין הכנסות לעלויות. הדוח ייתן מענה לשאלות העסקיות הבאות:

- מה הרווח הממוצע ושולי הרווח של עיריית NY מאכיפת חוקי החניה בכל שנה?
- מה הרובע (borough) הכי רווחי בעיר NYC?
- האם היו שינויים מהותיים לאורך החודשים ברווח מדוחות חנייה?
- מה הם עשרת סוגי הקנסות הרווחיים ביותר בשנת 2016?

c. שלב העברת הדוחות ל-Power BI Service

לאחר פיתוח הדוחות יש לבצע את המשימות הבאות:

1. פרסום הדוחות ל-Power BI service
2. בניית דאשבורד המרכז את התצוגות המעניינות ביותר.

תוצרי השלב: קובץ PBIX עם הדוחות ודוחות ודאשבורד ב-Power BI Service.

בהצלחה!

נספח א. מקורות הנתונים לפרויקט

טבלת היעד בבסיס הנתונים	File Name	מהות הנתונים
DimBorough	Borough_code.csv	שמות הרובעים
	DOF_Parking_Violation_Codes.csv	קודי עבירת תנועה ועלות קנס
DimIssuingAgency	ISSUING AGENCY.csv	קוד המוסד
DimPlateType	PLATE TYPE.csv	קוד סוגי לוחות זיהוי
DimLocation	STREET_CODES.csv	שמות רחובות
DimLocation	Dimlocation.csv	ממד המיקום
DimState	USSatesCode.csv	שמות המדינות בארה"ב
DimBodyType	VEHICLE BODY TYPE.csv	קוד סוגי הרכבים
DimColor	VEHICLE COLOR CODE.csv	קוד צבעי רכב
FactParkingViolation	Parking Violation 2015.csv	נתוני דוחות חנייה שנת 2015
FactParkingViolation	Parking Violation 2016.csv	נתוני דוחות חנייה שנת 2016
FactParkingViolation	Parking Violation 2017.csv	נתוני דוחות חנייה שנת 2017
DimVehicle	DimVehicle.csv	ממד רכבים
DimIssuer	DimIssuer.csv	ממד פקח חניה
	trafrule.pdf	חוקי התנועה של עיריית ניו יורק

נספח ב. שמות שדות וטבלאות בבסיס הנתונים

FactParkingViolation

Column Name	Condensed T
ParkingViolationKey	int
[Summons Number]	varchar(50)
[Issue Date]	varchar(50)
[Violation Code]	varchar(50)
[Violation Time]	varchar(50)
[Violation In Front Of Or Opposite]	varchar(50)
VehicleKey	int
IssuerKey	int
LocationKey	int

DimPlateType

Column Name	Condensed Type
PlateTypeCode	varchar(3)
PlateTypeName	varchar(40)

DimState

Column Name	Condensed Type
StateCode	varchar(2)
StateName	varchar(40)

DimBodyType

Column Name	Condensed Type
BodyTypeCode	varchar(5)
BodyTypeName	varchar(40)

DimLocation

Column Name	Condensed Type
LocationKey	int
BoroughCode	varchar(2)
StreetCode	varchar(10)
StreetName	varchar(10)
HouseNumber	varchar(20)
City	varchar(20)
StateCode	varchar(2)

DimVehicle

Column Name	Condensed
VehicleKey	int
PlateID	varchar(10)
RegistrationStateCode	varchar(2)
PlateTypeCode	varchar(3)
BodyTypeCode	varchar(5)
BodyMakeName	varchar(5)
VehicleColorCode	varchar(10)
vehicleYear	varchar(4)
ExpirationDate	varchar(8)
ExpirationDateKey	int

DimIssuer

Column Name	Condensed Typ
IssuerKey	int
IssuerID	varchar(20)
IssuingAgencyCode	varchar(5)

DimColor

Column Name	Condensed Type
ColorCode	varchar(10)
ColorName	varchar(40)

DimIssuingAgency

Column Name	Condensed Type
IssuingAgencyCode	varchar(5)
IssuingAgencyName	varchar(100)
IssuingAgencyAverageYearlySalary	money

DimBorough

Column Name	Condensed Type
BoroughCode	varchar(2)
BoroughName	varchar(20)

נספח ג' מודל הנתונים ב-Power BI