**פרויקט קורס בינה עסקית**

**קורס מס' 40205**

**מטלה מספר 2**

**מרצה: מר אור פרץ**



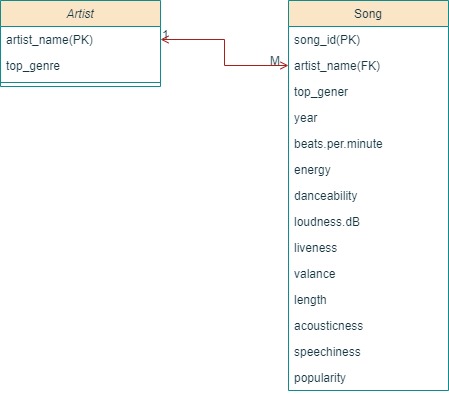
**שמות מגישים:**  
ים הדס - 318810553  
אפריים לדר - 316012236  
חן בשארי – 315341180  
גל ברדוגו - 318434107

**חלק ב'- Warehouse Data**

**שאלה מספר 1:**בחרנו בסכמת STAR. כדי שנוכל לענות שאלת המחקר שלנו אנו נדרשים לייצר טבלה מרכזית אחת המכילה בתוכה את הממדים והמדדים. בנוסף נציין כי הסכמה היא קלה להבנה ופשוטה לדירוג היררכי.

**שאלה מספר 2:**

תיאור הdata warehouse באמצעות ERD:



**שאלה מספר 3:**

חברת ההפקות המוזיקלית מעוניינת להפיק אלבום חדש ורוצה לדעת כיצד המאפיינים של ,energy אורך השיר, loudness משפיעים על הפופולריות שלו. STAR היא סכמה שיכולה לעזור לחברת ההפקות בשליפה מהירה של המאפיינים הרצויים ללא ריבוי של .joins

**חלק ג' - הגדרה ומימוש ETL**

**1. הגדרת תהליך הETL- עבור אוסף הנתונים:**

Extraction – בשלב זה נחלץ את הנתונים מטבלאות המקור artist ו- .songs

Transformation – בשלב זה העברנו את הנתונים שחילצנו למודל טבלאי אחד. כעת נבצע סינון של העמודות הרלוונטיות בהתאם לשאלה העסקית.

Load – בשלב זה נטעין את הנתונים שחילצנו ל.DW

**2. מימוש ETL ע"י STTM:  
  
שלב 1:** Reference data בשלב זה נגדיר את סט הנתונים המורכב מ2 טבלאות: ARTIST וSONG, אותן נגדיר באמצעות השדות הרלוונטיים.  
  
**שלב 2:** Extract from data reference בשלב זה חילוץ של הנתונים ע"י קובץ CSV.  
  
**שלב 3:** Data validation בשלב זה נוודא כי הנתונים הקיימים מתאימים למטרת הפרויקט, כלומר נוודא שקיימים המאפיינים של סגנון השיר, BPM, אורך השיר ומידת הפופולריות שלו.  
  
**שלב 4:** Transformation data בשלב זה נבצע אימות של שילוב המידע מתוך הטבלאות בהן השתמשנו. שלב זה כולל ניקוי של הדאטה, במקרה שלנו לא הופיעו בנתונים ערכים חסרים או חריגים ולכן לא היה צורך בכך.  
מחקנו עמודות לא רלוונטיות מתוך סט הנתונים:  
  
מטבלת ARTIST : ARTIST, YEAR.  
  
מטבלת SONGS: TITLE, ENERGY , DANCEABILITY , LOUDNESS , LIVENESS,VALANCE , AOUSTICNESS , SPEECHINESS. **שלב 5:** Stage : שלב זה הוא שלב ביניים בו כל הנתונים נמצאים באזור ה STAGING בתוכנות כמו TABLEAU או AIRFLOW, דרכן הנתונים עוברים.  
  
**שלב 6:** Publish to data warehouse בשלב זה העברנו את הנתונים הרלוונטיים לאחר עיבודם למחסן נתונים שמורכב מטבלה אחת בה יהיו המפתחות של השירים והאמנים.

**3. הרצת ובדיקת ETL:**

import pandas as pd  
  
data = pd.read\_csv('Top 100 most Streamed - Sheet1.csv')  
data = data.rename(columns={  
 'beats.per.minute': 'bpm',  
 'loudness.dB': 'loudness'  
})  
data = data.drop(['title', 'year', 'bpm', 'energy', 'danceability', 'loudness', 'liveness', 'valance', 'length', 'acousticness','speechiness', 'popularity'], axis=1, inplace=True)  
  
data.to\_csv('Artists on top 100 songs.csv', index=False)  
  
  
  
  
  
class SpotifyTop100ETL:  
 def \_\_init\_\_(self):  
 self.Songs\_data = None  
 self.Artists\_data = None  
  
 def extract\_Songs\_data(self):  
 self.Songs\_data = pd.read\_csv('Top 100 most Streamed - Sheet1.csv', encoding='latin-1')  
 self.Songs\_data.dropna(inplace=True)  
 print("Songs data:")  
 print(self.Songs\_data.columns)  
  
 def extract\_Artists\_data(self):  
 self.Artists\_data = pd.read\_csv('Artists on top 100 songs.csv', encoding='latin-1')  
 self.Artists\_data.dropna(inplace=True)  
 print("Artist data:")  
 print(self.Artists\_data.columns)  
  
 def transform\_data(self):  
 merged\_data = pd.merge(self.Songs\_data, self.Artists\_data, on='artist')  
 profitable\_program = merged\_data.groupby('artist')['Total Songs by Artist'].idxmax()  
 transformed\_data = merged\_data.loc[profitable\_program, ['Artist', 'Total Songs by Artist']]  
 return transformed\_data  
  
 def load\_data(self):  
 transformed\_data = self.transform\_data()  
 print("data successfully loaded to the database.")  
 print(transformed\_data)  
  
 def run\_etl(self):  
 self.extract\_Songs\_data()  
 self.extract\_Artists\_data()  
 self.load\_data()  
  
  
SpotifyTop100ETL.run\_etl()