מחסני נתונים - תרגיל בית: ניתוחי וירוס הקורונה

מרכז הבקרה ומניעת המגיפות בדרום קוריאה שיתף נתונים לגבי בדיקות לוירוס הקורונה, חולי COVID-19 (המחלקה שנגרמת מהוירוס), ונתונים אפידימולוגיים כלליים[1][2].

לרשותכם (ב-Moodle) קבצי נתונים מעובדים שמכילים מידע אודות חולים, בדיקות, מסלולי החולים, ומידע דמוגרפי כללי על אזורים בדרום קוריאה. עליכם לבנות מחסן נתונים סביב תהליך גרעיני.

אוסף הנתונים ניתן בפורמט CSV במספר קבצים:

- patient – נתונים אפידמיולוגיים של חולי COVID-19.
- route – מסלולי החולים ב-COVID-19 בדרום קוריאה (מקומות שהם ביקרו).
- time - נתוני סדרות זמן על מצב COVID-19 בדרום קוריאה.
- region - מקומות ונתונים סטטיסטיים של האזורים בדרום קוריאה.

תיעוד מלא של מבנה הקבצים ניתן למצוא בנספחים.

אתם מתבקשים לתכנן ולבנות מחסן נתונים ולבצע תהליך ETL מקבצי הנתונים הקיימים.

[1] https://www.kaggle.com/kimjihoo/coronavirusdataset
[2] https://github.com/jihoo-kim/Coronavirus-Dataset

חלק א: תכנון מחסן נתונים

אתם מתבקשים לתכנן מחסן נתונים בהתאם לניתוחים שאנו מעוניינים לבצע:

1. כמות הנדבקים לפי ימים עבור שלושת הערים עם מספר ההדבקות הגבוה ביותר בדרום קוריאה. נדבק מזוהה לפי שורה בטבלה patient.
2. כמות המחלימים לפי ימים עבור שלושת הערים עם מספר ההדבקות הגבוה ביותר בדרום קוריאה. מחלים מזוהה לפי state בטבלת patient.
3. כמות המקומות שנדבק ביקר בהם עד שלושה ימים, עד יומיים ועד יום לפני כניסה לבידוד.
4. כמות האנשים שבעיר סיאול מתחת לגיל 30 שנדבקו ושהדביקו אדם אחר מגיל 30 ומעלה.
5. צרו מידע שיאפשר לקבל החלטה לגבי אילו ערים בדרום קוריאה כדאי לסגור כדי לצמצם את ההפצה של הוירוס.

דרישות:

אפיינו מחסן נתונים בסכמת **כוכב** יחיד (לא פתית שלג) לצורך הניתוחים שאנו מעוניינים לבצע תוך שימוש בשלבים שנלמדו בכיתה:

א. זיהוי התהליך בו מחסן הנתונים מתמקד

ב. בחירת הגרעין (Grain)

ג. בחירת מימדי מחסן הנתונים (לפחות 3 מימדים שונים)

ד. זיהוי העובדות

שימו לב - בחלק א׳ לא צריך לענות ממש על השאילתות, אלא רק לתכנן מחסן נתונים **שיאפשר** לענות על השאילתות בצורה פשוטה.

תפוקות (בקובץ Word):

- פירוט של העיצוב מחסן הנתונים, תרשים קונספטואלי של מחסן הנתונים, כולל הגדרת השדות בפורמט המתואר בדוגמה:

דוגמה - טבלת מימד תאריך (Date)

שדה	טיפוס נתונים	מקור נתונים	הערות
date_id	int	שדה Auto Increment	מפתח ראשי, ייחודי
Date	date	מיוצר באופן אוטומטי. מתחיל מתאריך MIN ומסתיים MAX מקובץ time עמודה date	תאריך
Tests	int	מחושב ע״י MAX(test) לאותו היום.	מהווה תמונת מצב מצטברת עד אותו היום, לגבי מספר בדיקות שבוצעו בסוף היום

חלק ב׳: תהליך ETL וניתוח הנתונים

בנו את מחסן הנתונים שהגדרתם בחלק א׳ ובצעו תהליך ETL לנתונים הקיים: עליכם לבנות תהליך ETL בעזרת Python המייצר מסלול קריאת הנתונים (Pipeline) מקבצי הקלט (Extract), עיבוד הנתונים (Transform) וטעינתם לבסיס הנתונים MySQL (Load). לאחר מכן, צרו את הניתוחים מחלק א׳.

דרישות:

1. בנו תהליך ETL אוטומטי. נדרש להשתמש בחבילת Pandas לצורך ביצוע תהליך ETL.
2. נהלו ערכים חסרים (NULL).
3. נהלו ערכים חריגים תוך שימוש ב-outlier detection.
4. בצעו בדיקה לתקינות בסוף תהליך ה-ETL.
5. תעדו היטב את קוד ה-Python וה-SQL.
6. ענו על השאילתות בחלק א׳ באמצעות SQL תוך שימוש במחסן הנתונים.

תפוקות נדרשות (בקובץ Jupyter notebook):

- פירוט קוד Python ושאילתות SQL (בקובץ Jupyter notebook), כולל תיעוד הפעולות והשיקולים שנלקחו במהלך העבודה (נסחו בצורה ברורה בעברית או באנגלית).
- מידע על הנתונים בכל טבלה במחסן הנתונים – כמות שדות, כמות רשומות, טווחי ערכים רלוונטיים (למשל טווח תאריכים). יש לצרף דוגמית של נתונים (כ- 5-10 שורות) מכל טבלה.
- עבור השאילתות מחלק א׳ יש לצרף דוגמית של פלט (5-10 שורות).

נהלים והנחיות כלליות

1. ציון יחושב באופן הבא:
   - אפיון ועיצוב מחסן הנתונים
   - תהליך ה-ETL
   - נכונות שלבי טעינת המידע
   - רמת האוטומציה בתהליך
   - רמת התיעוד של תהליך ETL
   - איכות הקוד הטעינה
   - ניהול ערכים חסרים וחריגים
   - בדיקה לתקינות בסוף תהליך ה-ETL
   - מענה בצורה פשוטה על הניתוחים באמצעות שאילתות ב-SQL
2. הגשה בקבוצות של 3 סטודנטים.
3. מועד הגשה מפורסם באתר הקורס.
4. ההשגה תיעשה על ידי סטודנט אחד מחברי הקבוצה בקובץ ZIP הכולל את:
   א. קובץ Word המכיל את תוצרי (חלק א׳).
   ב. קובץ Jupyter Notebook המכיל את קוד ETL (חלק ב׳).
5. כל הקבצים צריכים להכיל את ת.ז של חברי הקבוצה ומס׳ הקבוצה. נא לא לצרף את קבצי הנתונים.
6. חריגה מפורמט ההגשה (בפרט תיעוד לקוי, קוד שלא ניתן להרצה או מחסור בקבצי התקנה נלווים) כמו גם איחור במועד ההגשה יובילו להורדת ציון.
7. המסמך יהיה כתוב בגופן David, גודל 12, עם מרווח של שורה וחצי.

בהצלחה !

נספחים

נספח א': patient.csv

נתונים אפידמיולוגיים של חולי COVID-19 בדרום קוריאה.

רשימת כל השדות הקיימים בקובץ:

Field	Description
[3]patient_id	the ID of the patient (n-th confirmed patient)
global_num	the number given by KCDC
sex	the sex of the patient
birth_year	the birth year of the patient
age	the age of the patient
country	the country of the patient
city	the city of the patient
disease	0: no disease / 1: underlying disease
infection_case	the collective infection
infection_order	the order of infection
infected_by	the parient_id of who has infected the patient
contact_number	the number of contacts with people
symptom_onset_date	the date of symptom onset
confirmed_date	the date of confirmation
released_date	the date of discharge
deceased_date	the date of decease
state	isolated / released / deceased

[3] לצורך התרגיל נניח שחולה קורונה יכול לחלות שוב. כלומר, patient_id יכול לחזור בטבלת patient.

נספח ב׳: route.csv

מסלולי החולים COVID-19 בדרום קוריאה (מקומות שהם ביקרו).

Field	Description
patient_id	the ID of the patient (n-th confirmed patient)
global_num	the number given by KCDC
date	Year-Month-Day
province	Special City / Metropolitan City / Province(-do)
city	City(-si) / Country (-gun) / District (-gu)
latitude	the latitude of the visit (WGS84)
longitude	the longitude of the visit (WGS84)

נספח ג׳: time.csv

נתוני סדרות זמן על מצב COVID-19 בדרום קוריאה.

Field	Description
date	Year-Month-Day
time	Time (0 = AM 12:00 / 16 = PM 04:00)
test	the accumulated number of tests
negative	the accumulated number of negative results
confirmed	the accumulated number of positive results
released	the accumulated number of releases
deceased	the accumulated number of deceases

נספח ד׳: region.csv

מקומות ונתונים סטטיסטיים של האזורים בדרום קוריאה.

Field	**Description**
code	the code of the region
province	Special City / Metropolitan City / Province(-do)
city	City(-si) / Country (-gun) / District (-gu)
latitude	the latitude of the visit (WGS84)
longitude	the longitude of the visit (WGS84)
elementary_school_count	the number of elementary schools
kindergarten_count	the number of kindergartens
university_count	the number of universities
academy_ratio	the ratio of academies
elderly_population_ratio	the ratio of the elderly population
elderly_alone_ratio	the ratio of elderly households living alone
nursing_home_count	the number of nursing homes