# תיעוד המערכת – סדנה במסדי נתונים (89679)

<u>מגישים</u>: נועם רוט (212450431), טל סיגמן (322888389), יאיר לוי (322884834)

#### רעיון המערכת

אנשים רבים מתעניינים בשינויים היומיים והמגמות השונות של התפשטות מגפת הקורונה בכל העולם. עם זאת, קיים קושי בהצגת כלל המידע על ההתפשטות המושפעת ממאפיינים שונים בכל מדינה (גיל חציוני, צפיפות אוכלוסין, גודל האוכלוסיה מצב כלכלי ועוד), באופן ברור ונגיש.

לכן, קיים הצורך בהעברת המידע באופן נגיש, ויזואלי (ע"י גרפים, מפות ועוד) וידידותי למשתמש, שיציג את התמונה המלאה על המגפה במקומות שונים. המערכת תציג את הנתונים ותאפשר השוואה של נתוני הקורונה בין מדינות, יבשות שונות ואף עבור אותה מדינה בפרקי זמן שונים, וכן הצגה עפ"י מאפיינים רלוונטיים של המדינות.

#### המידע שעליו בנויה המערכת

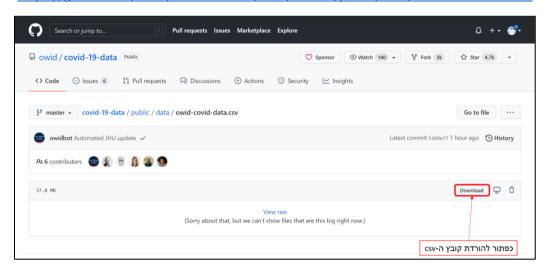
הנתונים שיוצגו במערכת הם נתונים הנאספו במשך שנה וחצי (החל מ-22.01.2020) על מגפת הקורונה, ומתעדכנים מספר פעמים ביום (מספרי נדבקים ומתים – יומיים ומצטברים) לפי מדינות שונות ונתונים שלהן (יבשת, צפיפות אוכלוסין, גיל חציוני וכו').

ה-dataset שמהנתונים שלו תתבסס המערכת נלקח מהאתר OurWorldInData:

https://ourworldindata.org/coronavirus-source-data

שמקור המידע שלו הוא Johns Hopkins University, וממנו נבחר קובץ ה-csv שמתוכו נבחרו העמודות הרלוונטיות למימוש הפרויקט שלנו:

https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv

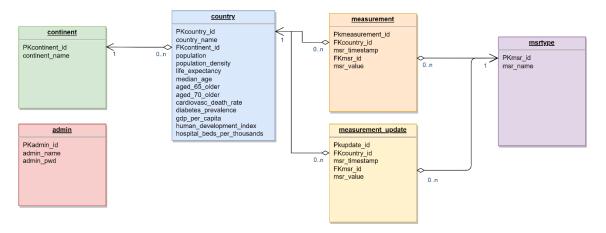


ב-dataset זה ישנן למעלה מ-100K רשומות עם יותר מ-25 עמודות שונות בתחומים הרלוונטיים.

כל המידע המופיע בפרויקט נבחר מטבלה זו (ששונתה במבנה, אך לא בתוכן בהתאם לרעיון הפרויקט).

### סכמת בסיס הנתונים של המערכת

מבנה מסד הנתונים של המערכת ייבנה באופן המאפשר הרחבת מודל הנתונים בעתיד ותחזוקתיות קלה של המערכת (החיצים שבין הטבלאות מקשרים ביניהן לפי הקשרים בין הישויות אותן מייצגות: one-to-many):



כעת, נתאר את העמודות בכל טבלה ואת סוג המידע (מחרוזת, ערך נומרי וכו') שמכילה כל עמודה, ואת הקשרים שבין הטבלאות (וסוג הקשר ביניהן).

## טבלת continent

טבלה המכילה מידע על היבשות השונות:

הערות	טיפוס	שם עמודה	עמודה
מספר הזיהוי של היבשת	TINYINT	PKcontinent_id	.1
שם היבשת	VARCHAR(45)	continent_name	.2

#### טבלת country

טבלה המכילה מידע סטטי (לא מדידה תלוית זמן) על המדינות השונות:

הערות	טיפוס	שם עמודה	עמודה
מספר הזיהוי של המדינה	INT	PKcountry_id	.1
שם המדינה	VARCHAR(45)	country_name	.2
מספר הזיהוי של היבשת בה נמצאת המדינה –	TINYINT	FKcontinent_id	.3
foreign key			
גודל האוכלוסייה	INT	population	.4
צפיפות האוכלוסייה	FLOAT	population_density	.5
תוחלת חיים ממוצעת	FLOAT	life_expectancy	.6
גיל חציוני	FLOAT	median_age	.7
אחוז מהאוכלוסייה מעל גיל 65	FLOAT	aged_65_older	.8
אחוז מהאוכלוסייה מעל גיל 70	FLOAT	aged_70_older	.9
אחוז מהאוכלוסייה עם מחלות לב	FLOAT	cardiovasc_death_rate	.10
אחוז מהאוכלוסייה עם סוכרת	FLOAT	diabetes_prevalence	.11
התמ"ג של המדינה	DOUBLE	gdp_per_capita	.12
מדד הפיתוח של המדינה	FLOAT	human_development_index	.13
מספר מיטות בבתי חולים ביחס לגודל	FLOAT	hospital_beds_per_thousand	.14
האוכלוסייה			

### measurement טבלת

טבלת הנתונים המרכזית - טבלה המכילה מידע על המדידות השונות (מדידות של נתוני קורונה):

הערות	טיפוס	שם עמודה	עמודה
מספר הזיהוי של המדידה	INT	PKmeasurement_id	.1
מספר הזיהוי של המדינה שבה נערכה המדידה	INT	FKcountry_id	.2
foreign key –			
זמן קיום המדידה	DATE	msr_timestamp	.3
foreign key – מספר הזיהוי של סוג המדידה	INT	FKmsr_id	.4
ערך המדידה	FLOAT	msr_value	.5

#### מבנה נתונים זה מאפשר:

- 1. הוספה עתידית קלה של מדדים נוספים
- 2. מאפשר קליטת מדדים חלקיים ממדינות ובתקופות מסויימות
  - 3. שונות בין המדדים הנאספים מכל מדינה

## שבלת msrtype

טבלה המכילה מידע על **סוגי** המדידות השונות (מדידות של נתוני קורונה):

הערות	טיפוס	שם עמודה	עמודה
מספר הזיהוי של סוג המדידה	INT	PKmsr_id	.1
שם קטגוריית המדידה	VARCHAR(45)	msr_name	.2

## measurement update טבלת

טבלת עדכון הנתונים - טבלה המכילה מידע על עדכונים של מדידות שונות של נתוני קורונה:

הערות	טיפוס	שם עמודה	עמודה
מָספר הזיהוִי של המדידה המעודכנת	INT	PKupdate_id	.1
(המתווספת)			
מספר הזיהוי של המדינה שבה נערכה המדידה	INT	FKcountry_id	.2
foreign key – המעודכנת			
זמן קיום המדידה המעודכנת	DATE	msr_timestamp	.3
מספר הזיהוי של סוג המדידה המעודכנת –	INT	FKmsr_id	.4
foreign key			
ערך המדידה המעודכנת	DOUBLE	msr_value	.5

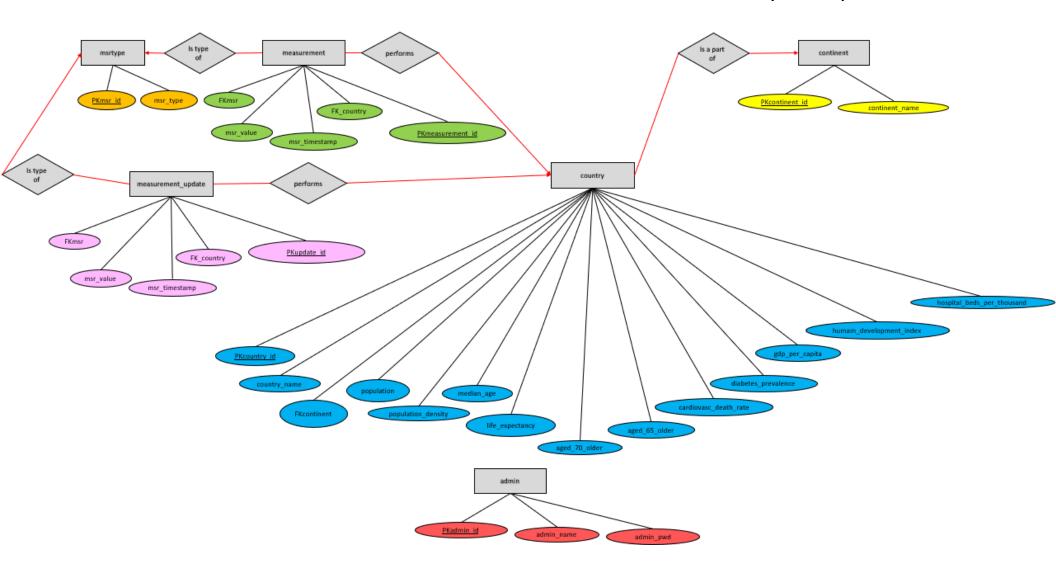
## admin טבלת

טבלה המכילה מידע על מנהלי המערכת (שיכולים לאשר/ לדחות עדכונים המוצעים ע"י משתמשים):

הערות	טיפוס	שם עמודה	עמודה
מספר הזיהוי של מנהל המערכת	INT	PKadmin_id	.1
שם המנהל	VARCHAR(45)	admin_name	.2
סיסמת המנהל	VARCHAR(45)	admin_pwd	.3

# תרשים ישויות והקשרים ביניהן

# תרשים הישויות והקשרים ביניהן



### שאילתות פשוטות המופיעות בפרויקט (בקובץ cop\_queries.py)

קריאת הנתונים הבסיסיים (יבשת ואוכלוסייה) של כל מדינה להצגה במפה (בפונקציה get countries): SELECT continent.continent name, country.country name, country.population FROM continent, country WHERE continent. PKcontinent id = country. FKcontinent id רשימת המשתנים הדינאמיים האפשריים למדינה (בפונקציה get variables): SELECT msr\_name FROM msrtype קריאת התאריך המאוחר ביותר בו נאספו נתונים (בפונקציה get dates): SELECT MAX(msr timestamp) FROM measurement קריאת התאריך המוקדם ביותר בו נאספו נתונים (בפונקציה get\_dates): SELECT MIN(msr timestamp) FROM measurement בתאריך הכי מאוחר (new\_cases, total\_cases, new\_deaths, total\_deaths) בתאריך הכי מאוחר בטווח - [מהתאריך המוקדם ביותר ב-DB עד התאריך שקיבלנו] שבו כל מדינה מדדה את המדידה של המשתנה הרלוונטי. נשים לב כי לא מובטח שבכל תאריך בהכרח כל המדינות ערכו את כל הבדיקות. (get info of variable בפונקציה): SELECT SUM(msr\_value) FROM measurement AS m1 USE INDEX(mapIndex), SELECT FKcountry\_id AS 'country\_id', MAX(msr\_timestamp) AS 'max\_timestamp' FROM measurement USE INDEX(searchIndex) WHERE msr timestamp BETWEEN '{0}' AND '{1}' AND FKmsr\_id = (SELECT PKmsr\_id FROM msrtype WHERE msr\_name = '{2}') GROUP BY FKcountry id) AS m2 WHERE m1.msr timestamp = m2.max timestamp AND m1.FKmsr id = (SELECT PKmsr id FROM msrtype WHERE msr name = '{3}') AND m1.FKcountry id = m2.country id סכימת ערך המאומתים/המתים במדינה מסוימת בתאריך נתון (get\_info\_of\_variable\_in\_specific\_date): SELECT sum(msr\_value) FROM measurement USE INDEX(searchIndex) WHERE msr timestamp = '{}' AND FKmsr id = (select PKmsr id FROM msrtype WHERE msr name = '{}')

```
קריאת הנתונים בכל תאריך עבור מדינה נתונה וסוג מדידה נתון (בפונקציה get data for scatter line graph):
SELECT msr timestamp, msr value
FROM measurement
WHERE FKcountry_id = (SELECT PKcountry_id FROM country WHERE country_name LIKE "{0}")
AND FKmsr_id = (ELECT PKmsr_id FROM msrtype WHERE msr_name LIKE "{1}")
                                החזרת ערך משתנה סטטי של מדינה נתונה (בפונקציה get_static_data):
SELECT {0}
FROM country
WHERE PKcountry id = (SELECT PKcountry id FROM country WHERE country name LIKE "{1}")
           רשימת כל המדינות כדי שהמשתמש יוכל לבחור מדינה לעדכן בה נתונים (בפונקציה user update):
SELECT PKcountry id FROM country WHERE country name LIKE '{0}'
     רשימת כל המדידות כדי שהמשתמש יוכל לבחור מדידה שהנתונים שלה יעודכנו (בפונקציה user update):
SELECT PKmsr id FROM msrtype WHERE msr name = '{0}'
                                                          הכנסת הבקשה לעדכון לרשימת העדכונים
  (user_update | update_measurements_table | check_for_adding_another_measurement בפונקציות):
INSERT INTO measurement update ({0}) VALUES (%s, %s, %s, %s)
                          בדיקה האם המשתמש והסיסמה של ה-admin נכונים (בפונקציה check admin):
SELECT EXISTS(
SELECT admin name, admin pwd FROM admin WHERE admin name = '{0}' AND admin pwd = '{1}')
                             יצירת סוג מדידה דינאמית חדש (בפונקציה add new measurement type):
INSERT INTO msrtype ({0}) VALUES (%s)
                          אישור עדכון מרשימת העדכונים בידי admin (בפונקציה):
SELECT FKcountry_id, msr_timestamp, FKmsr_id, msr_value
FROM measurement_update
WHERE FKcountry_id = (SELECT PKcountry_id FROM country WHERE country_name = '{0}')
AND msr timestamp = '{1}'
AND FKmsr id = (SELECT PKmsr id FROM msrtype WHERE msr name = '{2}')
AND msr value = \{3\}
                     קבלת שם סוג המדידה בהינתן הקוד שלה (בפונקציה update measurements table):
SELECT msr name FROM msrtype WHERE PKmsr id = {}
```

```
מחזיר את המדידה הקיימת (במידה וקיימת) המקיימת את אותם נתוני מדידה (בפונקציה
                                                                  :(update measurements table
SELECT EXISTS (
       SELECT 1
       FROM measurement
       WHERE msr timestamp = \{0\} AND FKcountry id = \{1\} AND FKmsr id = \{2\})
       קבלת ערך המדידה עבור המדידה עם המאפיינים הנתונים (בפונקציה update measurements table):
SELECT msr value
FROM measurement
WHERE FKcountry id = {0} AND FKmsr id = {1} AND msr timestamp = '{2}'
            ערכון ערך המדידה הקיימת לערך מטבלת העדכונים (בפונקציה update measurements table):
UPDATE measurement SET msr value = {0}
WHERE msr timestamp = '{1}' AND FKcountry id = {2} AND FKmsr id = {3}
          קבלת קוד סוג המדידה עבור סוג מדידה "total_cases" (בפונקציה update_measurements_table):
SELECT PKmsr id FROM msrtype WHERE msr name = 'total cases'
         קבלת קוד סוג המדידה עבור סוג מדידה "total deaths" (בפונקציה update measurements table):
SELECT PKmsr id FROM msrtype WHERE msr name = 'total deaths'
                     מציאת התאריך הבא ביחס לתאריך הנתון (בפונקציה update_measurements_table):
SELECT DATE ADD('{}', INTERVAL 1 DAY)
                  עדכון ערך במדידות העוקבות למדידה הקיימת (בפונקציה update next measurements):
UPDATE measurement SET msr value = {0}
WHERE FKcountry id = {1} AND FKmsr id = {2} AND msr timestamp BETWEEN '{3}' AND '{4}'
                                 oeירת מספר העדכונים הקיימים (בפונקציה get updates for display):
SELECT count(*) FROM measurement_update
                               החזרת העדכונים מטבלת העדכונים (בפונקציה get updates for display):
SELECT DISTINCT country.country name, measurement update.msr timestamp, msrtype.msr name,
measurement update.msr value
FROM measurement update, country, msrtype
WHERE country.PKcountry id = measurement update.FKcountry id
AND measurement update.FKmsr id = msrtype.PKmsr id
LIMIT {0}
```

```
קבלת התאריך המאוחר ביותר למדידה במדינה והסוג הנתונים (בפונקציה
                                                      :(check for adding another measurement
SELECT MAX(msr timestamp) FROM measurement WHERE FKcountry id = {0} AND FKmsr id = {1}
         קבלת ערך המדידה עבור הנתונים שניתנים (בפונקציה check for adding another measurement):
SELECT msr value FROM measurement
WHERE FKcountry_id = {0} AND FKmsr_id = {1} AND msr_timestamp = '{2}'
                       החזרת ערכי הרשומות למחיקה מטבלת העדכונים (בפונקציה reject user update):
SELECT FKcountry id, msr timestamp, FKmsr id, msr value
FROM measurement_update
WHERE FKcountry_id = (
       SELECT PKcountry_id FROM measurement_update WHERE country_name = '{0}')
AND msr timestamp = '{1}' AND FKmsr id = (SELECT PKmsr id FROM msrtype WHERE msr name = '{2}')
AND msr_value = {3}
                                      מחיקת עדכון מטבלת העדכונים (בפונקציה reject user update):
DELETE FROM measurement update
WHERE FKcountry_id = {0} AND msr_timestamp = '{1}' AND FKmsr_id = {2} AND msr_value = {3}
```

### שאילתות מורכבות המופיעות בפרויקט (בקובץ oop\_queries.py)

נרצה לסכום את מספר המאומתים בכל יבשת (סכימת ה-total\_cases של כל המדינות המשתייכות לאותה היבשת) בתאריך הכי מאוחר בטווח שבו כל מדינה מדדה את המדידה של מספר המאומתים המצטבר. נשים לב כי לא מובטח שבכל תאריך בהכרח כל המדינות ערכו את כל הבדיקות (ונבחר לכל מדינה תאריך משלה).

נרצה גם לסכום את גודל האוכלוסייה בכל יבשת (סכימת האוכלוסייה של כל המדינות המשתייכות לאותה היבשת) כדי לחשב את האחוזים המתאימים מהאוכלוסייה שאומתו.

כלומר: נסכום לכל יבשת את הערכים המאוחרים ביותר של מדידות במדינות המתאימות ונסכום את האוכלוסייה לכל יבשת (ערך סטטי) ונחלק את התוצאות אלו באלו.

אחוז המאומתים המצטבר בכל יבשת ביחס לגודל האוכלוסיה בתאריך המאוחר ביותר (בפונקציה (percentage\_cases\_out\_of\_total\_population\_in\_each\_continent

SELECT continent\_name, SUM(msr\_value) AS total\_cases, continental\_population AS total\_population, 
100 \* SUM(msr\_value) / continental\_population AS cases\_percentage
FROM (WITH last\_msr AS (

SELECT \*, ROW\_NUMBER () OVER (PARTITION BY FKcountry\_id ORDER BY msr\_timestamp DESC) rn

FROM measurement WHERE FKmsr\_id = (SELECT PKmsr\_id

FROM msrtype WHERE msr\_name = 'total\_cases')

) SELECT last\_msr.\*, country.FKcontinent\_id FROM last\_msr, country

WHERE rn = 1 AND PKcountry id = FKcountry id) AS m, continent, (

SELECT PKcontinent\_id, SUM(population) AS continental\_population

FROM country, continent

WHERE continent.PKcontinent\_id = country.FKcontinent\_id

GROUP BY PKcontinent\_id) AS continent\_population

WHERE continent.PKcontinent\_id = m.FKcontinent\_id

AND continent.PKcontinent\_id = continent\_population.PKcontinent\_id

GROUP BY m.FKcontinent\_id

ORDER BY cases\_percentage DESC

נרצה למצוא את מספר המתים בכל אחת מהמדינות בתאריך הכי מאוחר בטווח שבו כל מדינה מדדה את המדידה של מספר המתים המצטבר. נשים לב כי לא מובטח שבכל תאריך בהכרח כל המדינות ערכו בדיקה זו (ונבחר לכל מדינה תאריך משלה).

נרצה גם למצוא את מספר המאומתים בכל אחת מהמדינות בתאריך הכי מאוחר בטווח שבו כל מדינה מדדה את המדידה של מספר המאומתים המצטבר. נשים לב כי לא מובטח שבכל תאריך בהכרח כל המדינות ערכו בדיקה זו (ונבחר לכל מדינה תאריך משלה).

כלומר: נסכום לכל אחת מ-5 המדינות את הערכים המאוחרים ביותר של מדידות המתים והמאומתים ונחלק את התוצאות אלו באלו.

אחוז המתים מכלל המאומתים בכל מדינה מבין 5 המדינות עם אחוז האוכלוסיה מעל גיל 70 הגבוה ביותר אחוז המתים מכלל המאומתים בכל מדינה מבין 5 המדינות עם אחוז האוכלוסיה מעל גיל 70 הגבוה ביותר (percentage of verified deaths out of total cases):

```
SELECT DISTINCT country name, aged.aged 70 older, total deaths, population,
total cases, 100*total deaths/total cases AS deaths precentage
FROM (
       SELECT FKcountry_id, aged_70_older, msr_value AS total_deaths
       FROM (WITH last msr AS (
              SELECT *, ROW_NUMBER () OVER (
                      PARTITION BY FKcountry id ORDER BY msr timestamp DESC) rn
              FROM measurement
               WHERE FKmsr id = (
                      SELECT PKmsr id FROM msrtype WHERE msr name = 'total deaths')
       ) SELECT last_msr.*, country.aged_70_older FROM last_msr, country
       WHERE rn = 1 AND PKcountry id = FKcountry id) AS m
       ORDER BY aged_70_older DESC
       LIMIT 5) AS aged, (
       SELECT FKcountry id, msr value AS total cases
       FROM (WITH last_msr AS (
       SELECT *, ROW NUMBER () OVER (
              PARTITION BY FKcountry_id ORDER BY msr_timestamp DESC) rn
       FROM measurement
       WHERE FKmsr id = (SELECT PKmsr id FROM msrtype WHERE msr name = 'total cases')
       ) SELECT last msr.* FROM last msr WHERE rn = 1) AS m) AS cases, country
WHERE aged.FKcountry id = cases.FKcountry id AND aged.FKcountry id = country.PKcountry id
ORDER BY deaths_precentage DESC
```

נרצה למצוא את מספר המאומתים בכל יבשת בתאריך הכי מאוחר בטווח שבו כל מדינה ביבשת מדדה את המדידה של מספר המאומתים המצטבר. נשים לב כי לא מובטח שבכל תאריך בהכרח כל המדינות ערכו את כל הבדיקות (ונבחר לכל מדינה תאריך משלה).

נרצה למצוא את מספר המאומתים העולמי הכולל בתאריך הכי מאוחר בטווח שבו כל מדינה מדדה את המדידה של מספר המאומתים המצטבר. נשים לב כי לא מובטח שבכל תאריך בהכרח כל המדינות ערכו את כל הבדיקות (ונבחר לכל מדינה תאריך משלה). לבסוף נסכום את כל הערכים הסופיים לכדי המספר העולמי הכולל.

כלומר: נסכום לכל אחת מהיבשות את הערכים המאוחרים ביותר של מדידות המאומתים ונחלק את התוצאות בערך העולמי הכולל שחושב.

אחוז המאומתים בכל יבשת מכלל המאומתים הגלובלי בתאריך המאוחר ביותר (בפונקציה (percentage of verified cases out of all global verified cases for each continent

```
SELECT cases per continent.continent name, 100*total cases continent/global total cases AS
percentage
FROM (
       SELECT continent name, SUM(msr value) AS total cases continent
       FROM (WITH last msr AS (
                      SELECT *, ROW NUMBER () OVER (
                             PARTITION BY FKcountry id ORDER BY msr timestamp DESC) rn
                      FROM measurement
                      WHERE FKmsr id = (
                             SELECT PKmsr_id FROM msrtype WHERE msr_name = 'total_cases')
              ) SELECT last msr.*, country.FKcontinent id FROM last msr, country
              WHERE rn = 1 AND PKcountry_id = FKcountry_id) AS m, continent
       WHERE continent. PK continent id = m. FK continent id
       GROUP BY continent name) AS cases per continent, (
       SELECT SUM(msr value) AS global total cases
       FROM (WITH last msr AS (
              SELECT *, ROW_NUMBER () OVER (
                      PARTITION BY FKcountry id ORDER BY msr timestamp DESC) rn
              FROM measurement
              WHERE FKmsr id = (SELECT PKmsr id FROM msrtype WHERE msr name = 'total cases')
       ) SELECT last_msr.*, country.FKcontinent_id FROM last_msr, country
       WHERE rn = 1 AND PKcountry id = FKcountry id) AS m) AS total continental cases
ORDER BY percentage DESC
```

#### index-שימוש ב-trigger

#### measurement טבלת

הוספנו אל הטבלה הזו trigger כדי שברגע העלאת הנתונים לטבלה, במידה ואחד הערכים הוא שלילי, אז ה-trigger יהפוך אותו אל 0. כמו כן, הוספנו שני אינדקסים (בעת יצירת הטבלה) כדי לייעל את החיפוש:

```
def add_indices_to_measurements_table(my_cursor, my_connection):
    my_cursor.execute("create index searchIndex on measurement(msr_timestamp, FKmsr_id);")
    my_cursor.execute("""create index mapIndex on measurement(msr_timestamp, FKmsr_id, FKcountry_id);""")
    my_connection.commit()
```

המטרה של שני האינדקסים היא לייעל את זמן השאילתה שנמצאת בפונקציה get\_map\_variables בקובץ cop\_queries.py. מטרת השאילתה היא שבהינתן שני תאריכים נוכל למצוא לכל מדינה את ערך המדידה של new\_cases שנמדד בתאריך הכי מאוחר (בטווח שבין שני התאריכים שקיבלנו) שבו כל מדינה מדדה (לא מובטח שהמדידה האחרונה של new\_cases בהכרח תהיה באותה התאריך לכל המדינות). מקטע מהשאילתא הנ"ל:

```
SELECT SUM(msr_value)

FROM measurement AS m1 USE INDEX(mapIndex),

(SELECT FKcountry_id AS 'country_id', MAX(msr_timestamp) AS 'max_timestamp'

FROM measurement USE INDEX(searchIndex)

WHERE msr_timestamp BETWEEN '{4}' AND '{5}'

AND FKmsr_id = (SELECT PKmsr_id FROM msrtype WHERE msr_name = 'new_deaths')

GROUP BY FKcountry_id) AS m2

WHERE m1.msr_timestamp = m2.max_timestamp AND m1.FKcountry_id = m2.country_id

AND m1.FKmsr_id = (SELECT PKmsr_id FROM msrtype WHERE msr_name = 'new_deaths')
```

מטרת searchIndex - שיפור זמן ריצת חיפוש הרשומות לפי msr\_timestamp - שיפור זמן ריצת חיפוש הרשומות לפי הsearchIndex - שיפור זמן בשאילתה החיצונית, נשתמש ב-msr\_timestamp, FKmsr\_id לחיפוש לפי החיצונית, נשתמש ב-msr\_timestamp, FKmsr\_id לחיפוש לפי

### טבלת country

הוספנו אל הטבלה הזו trigger כדי שברגע העלאת הנתונים לטבלה, במידה ואחד הערכים הוא NULL, אז ה-trigger יהפוך אותו אל 0.