Databases for data analytics

מועד ב 2026

מבוא

משך הבחינה: 3 שעות
חומר עזר:

1. דף נוסחאות המגיע עם המבחן
2. מילון, לצרכי סטודנטים הרוצים להימנע משגיאות כתיב. שימו לב, **זהו מבחן חשיבה**. לא יורדו נקודות על שגיאות כתיב.

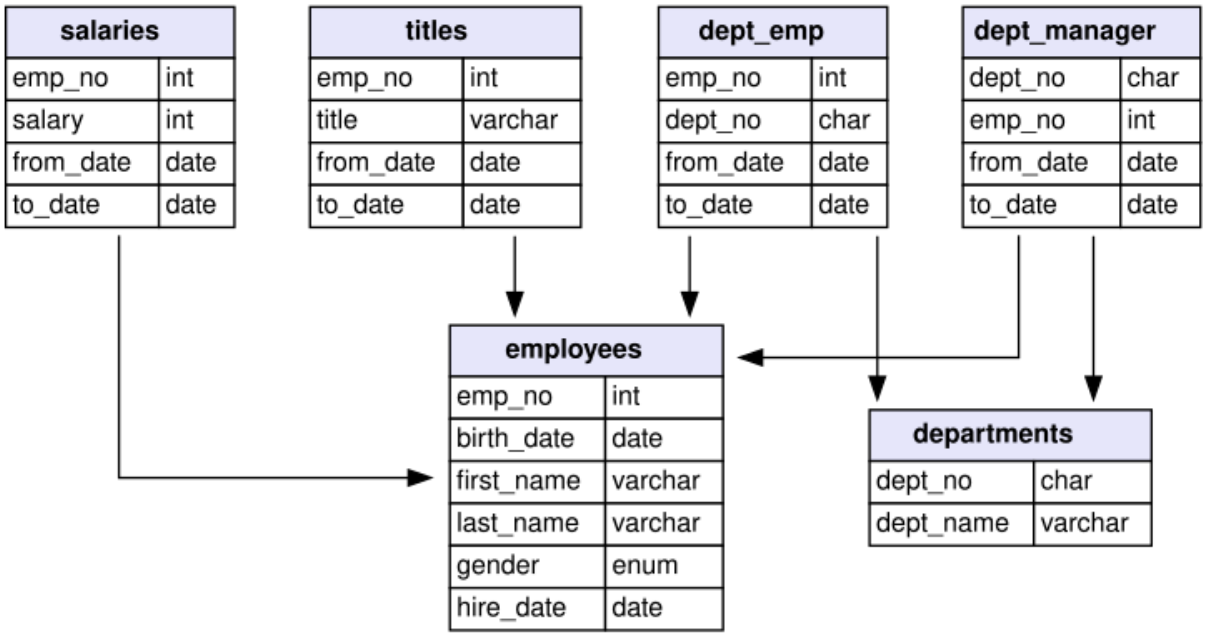
ישנם כמה סעיפי בונוס של חמש נקודות. הבונוסים הם מבוססים על הנלמד בקורס אבל ברמה גבוהה מהרגיל.
מומלץ לנסות לענות על הבונוסים אולם **רק לאחר שפתרתם את כל הבחינה**.

בהצלחה!

סכמה

השאלות בבחינה תתייחסנה לסכמה של Employee, מבנה ארגוני של חברה, המובאת בתרשים למטה.

1. ניתן להניח ששדות עם סיומת no הם מפתחות בטבלה, או קשורים למפתח בטבלה אחרת.
2. פרשנות שמות השדה היא לרוב מילולית. אם אתם לא בטוחים בפרשנות, ציינו את הפרשנות שבחרתם מפורשות ונמקו אותה.



1. **<u>שליפות (30 נקודות)</u>**

במהלך הבחינה מותר, ואפילו רצוי, להשתמש בשאילתות שכתבתם קודם.
כדאי לעשות זאת אנא השתמשו ב views וציינו בשאילתה המשתמשת את המקום בו הגדרתם אותה.

במידה ויש סעיף שהתקשתם בפיתרונו, אתם יכולים לציין שם של view שפותר אותו, לכתוב מה השדות בו, ולהשתמש בו בסעיפים אחרים.

a. שילפו את ה hire_date של העובד החדש ביותר בחברה
b. שילפו את כל זוגות המחלקות שיותר משלושה עובדים עבדו בשתיהן. **ביחרו ונמקו:** האם השאילתא צריכה להיות רפלקסיבית? סימטרית?
c. שילפו עובדים שיש להם שני title שונים באותה התקופה. ביחרו ונמקו את השדות הנשלפים.
d. שילפו את העובדים שמשכורתם ירדה. עליכם להחזיר את מספר העובד והתאריך בו המשכורת ירדה. Window functions יכולות להועיל.
e. שילפו את התפלגות מספר העובדים פר מחלקה. כלומר, מספר הפעמים שבמחלקה יש 0, 1, 2, 3 וכדומה עובדים ממוינים בסדר עולה. חשבו גם את ההסתברות לכל מספר.

פלט לדוגמה: מתוך עשר מחלקות יש אחת בלי עובדים, הסיכוי הוא עשירית. יש שלוש מחלקות יש עובד יחיד, סיכוי 30% וכן הלאה.

f. שאלת תקינות שאילתה, ללא קשר לסכמה זו. ציינו לפחות 3 סיבות בגללן השאילתה הבאה עשויה להכשל:

```
delete from customers where name = 'Dan';
```

g. **בונוס:**

i. שילפו את ה cumulative distribution של מספר העובדים במחלקה. כלומר, הסיכוי שבמחלקה **לא יותר** מ 0 עובדים, הסיכוי שבמחלקה **לא יותר** מעובד 1, וכן הלאה.

ii. **המלצה:** זו שאלה קשה. גשו אליה רק אחרי שעניתם ובדקתם היטב את המבחן.

1. **<u>תקינות נתונים (30 נקודות)</u>**

a. בשאלה הקודמת הוצגה סכמה של מבנה ארגוני. בשאלה זו עליכם לבחור את אופן השימוש בסכמה והחליט איך המידע הארגוני ימופה אליה. ניתן לבחור לאכלס את הטבלאות בצורות שונות. אנא ביחרו דרך לאכלס את הטבלאות כך שניתן יהיה ליצג את המקרים הבאים. ציינו חסרונות ויתרונות של הדרך שבחרתם ונמקו את בחירתכם. מותר להוסיף טבלאות ונתונים במידת הצורך.

i. המחלקה של המנכ"ל

ii. השייך הארגוני של סמנכ"לים ובהכללה מנהלים. האם סמנכ"ל משוייך להנהלה הכללית או לארגון אותו הוא מנהל.

iii. יצוג נוח של רמות שונות בארגון. לדוגמה, מחלקת מכירות המכילה מחלקת מכירות באירופה, המכילה מחלקת מכירות בצרפת.

b. שדה ה title של עובד מיוצג כמחרוזת. ציינו חסרונות ויתרונות. הציעו יצוג חלופי. התייחסו לנקודות הבאות.

i. תפקידים בעלי שמות שונים

ii. היררכיה של תפקידים (X זוטר, X בכיר, X אלוף אולימפי)

iii. תפקידים שלא מתאימים למחלקת העובד

c. ישנם יחסים שצריכים להתקיים בין תאריכים שונים במערכת (לדוגמה, מומלץ שעובד יוולד לפני תחילת העבודה). ביחרו לפחות שלושה יחסים. עבור כל יחס ציינו:

1. מה היחס?
2. כיצד הפרת היחס עשויה לפגוע במערכת?
3. כיצד ניתן להתמודד עם הבעיה? התייחסו למספר הטבלאות המעורבות.

ii. **בונוס:** נניח שבטעות הוזנו לבסיס הנתונים של IMDB בו השתמשנו בקורס סרטים מסויימים באופן כפול. כלומר, הסרט על כל נגררותיו מופיע באופן תקין אך מספר פעמים. כל אחת מההופעות היא עם id שונה. כמו כן נניח שלא

הבחנו בכך במהלך הפרוייקט וסטודנטים דירגו את את כלל המופעים של כל סרט.

1. האם יצוג סרט בכמה מופעים הוא בעית נתונים? נמקו וציינו השלכות.
2. בחלק מהמקרים סטודנטים נתנו דירוגים שונים ומופעים שונים של אותו הסרט. האם זו בעית הנתונים? האם ניתן למנוע אותה ללא מודעות לכפילות המופעים. נמקו וציינו השלכות.
3. כיצד הייתם מייצגים את המופעים של כל סרט? כיצד הייתם משתמשים בטבלה כדי לחשב את יציבות הדירוג של סרט?

3. עיצוב (30 נקודות)

אתר Spotify מספק שירותי האזנה למוזיקה. המאזין נרשם, ויכול לבחור שירים בצורות שונות. אמנים יכולים להעלות שירים ולהחשף לקהלים בעולם.

עצבו בסיס נתונים עבור Spotify, עבור השימושים למאזין.
אין צורך להתייחס לשימוש האמן.
המערכת צריכה לתמוך בשימושים הבאים.

1. ייצוג מאזין (חישבו ונמקו את הפרטים שאתם שומרים בייצוג).
2. איתור אמנים, בשמות שונים (עברי, לועזי, שם משפחה בלבד, כינויים).
3. איתור אמנים בשגיאות כתיב. עליכם לנמק האם יש ליצג זאת בבסיס הנתונים. תשובה שלילית אפשרית.
4. איתור שירים לפי שם, סוג, ואמן. שיר יכול להיות מבוצע בדואטים וכדומה.
5. יצוג playlists.
6. יש מידע רב שאינו נדרש להאזנה ישירה אך נאסף לצרכי טיוב החוויה. דוגמאות הן: Likes, search queries and selection, listening to a song, aborting listening to a song. ביחרו פריט מידע שכזה.

1. כיתבו קוד ליצירת הטבלאות הנדרשות. הקפידו על נרמול.
2. פרטו ונמקו את כל ההחלטות העיצוביות שקיבלתם:
   a. ציינו מה כל הישויות שאתם מייצגים, מה התרחיש בשבילו הם נדרשים, ומה היצוג שנבחר. אם יש ישויות רלוונטיות שאתם לא מייצגים ציינו והסבירו למה.
   b. הסבירו את השדות שבחרתם, התרחיש לו הם נדרשים, והטיפוס שבחרתם להם.
   c. נמקו את המפתחות שבחרתם, האם הם טבעיים (קיימים בעולם האמיתי) או פנימיים (לדוגמה auto increment)
   d. ציינו מה היחסים בין הישויות (1:1, n:1, או n:n) בעולם האמיתי. אם בחרתם ליצג יחס אחר, ציינו ונמקו. מספיק לציין יחסים רק בין ישויות שכנות, שיש בניהן קשר ישיר.
   e. פרטו מה ה constraints (לדוגמה: unique, FK, not null, check) שבחרתם ומה ההגנה שהם מספקים. אם יש נקודות שלא מוגנות על ידי בסיס הנתונים ציינו אותן ונמקו איך להגן.
3. הסבירו כיצד תשתמשו במידע שבחרתם כדי לשפר את חווית השימוש במוצר.
4. במודל Boston Consulting Group מחלקים מוצרים לארבע קבוצות. אחת הקבוצות היא כוכבים - כאלו שזכו להצלחה וממשיכים לצמוח. ממשו על בסיס הנתונים שיצרתם שאילתה שמאתרת כוכבים.

5. האם המערכת צריכה ליצג playlists שתכנם זהה? נמקו.
6. **בונוס:** מערכת המלצות לשירים מאד דומה לפרוייקט המלצות הסרטים שאתם עושים. הציגו מודל שאפשר לממש ב spotify ואי אפשר לממש על IMDB.

4. שיפור ביצועים (10%)

השאילתה הבאה נכתבה לבסיס הנתונים IMDB מהקורס.
מטרת השאילתה היא לשלוף את התפקידים שהופיעו יותר מ 100 פעמים בז'אנר ספציפי.

```
drop table if exists role_per_genere;

create table
role_per_genere
as
select
genre
, r.role
, count(*) as cnt
from
movies_genres as mg
join
movies as m
on
mg.movie_id = m.id
join
roles as r
on
m.id = r.movie_id
group by
genre, r.role
having
count(*) > 100
order by
genre, r.role, cnt
;
```

הציעו לפחות שלושה תיקונים ושיפורים לשאילתה. אתם יכולים ליצור טבלאות נוספות ולהגדיר אינדקסים במידה שיועילו לכם.

1. עבור כל שינוי
   a. ציינו מה הבעיה שהוא מתקן.
   b. ציינו חסרונות אם יש.
2. כיתבו את השאילתה המתוקנת.