

**מבחן במסדי נתונים (ב')**

**ד"ר עמוס עזריה**

**7028010**

סמסטר ב' מועד ב' **ט"ו באב** התשע"ז, 7.8.2017

**הנחיות כלליות:**

* משך הבחינה: 180 דקות.
* **יש לענות בגוף השאלון!** המחברת תשמש כטיוטא בלבד. על מענה במחברת יורדו נקודות!
* אין להכניס שום חומר עזר.
* השימוש במחשבון **אסור**.
* מומלץ לקרוא את ההוראות באנגלית ולפנות לעברית רק במקרה של חוסר הבנה.
* בסיום הבחינה - נא למסור את השאלון ואת המחברת.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Total** | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |  |
| **103** | 14 | 5 | 9 | 10 | 7 | 10 | 8 | 12 | 13 | 15 | Max points |
|  |  |  |  |  |  |  |  |  |  |  | Grade |

**ב ה צ ל ח ה!**

1. SQL (15 pt)

נתונות הטבלאות הבאות לייצוג מרפאת שיניים:

Patients (patientId, firstName, lastName, YearOfBirth, gender)

TreatmentTypes (treatmentTypeId, treatmentName, price)

TreatmentForPatient (patientId, treatmentTypeId, dayAsInt)

ניתן להניח שdayAsInt הוא מספר שלם המייצג את יום העסקים מאז המרפאה החלה לפעול.

כיתבו שאילתא שמחזירה את מספר הגברים (gender=0) שטופלו בכל יום ובתנאי שביום זה ההכנסה של המרפאה הייתה לפחות 1000 ש"ח.

Select tp.dayAsInt, count(p.gender = 0), sum(tt.price) as revenue

from Patients as p join TreatmentForPatient as tp

On(p.patientId = tp.patientId)

join TreatmentTypes as tt

on(tt.treatmentTypeId = tp.treatmentTypeId)

group by (tp.dayAsInt) having(revenue >= 1000)

2. Normalization (13 pt):

Given the following relation:

נתונה הרלציה הבאה והתלויות שלה:

R(A, B, C, D)

And the following dependencies:

{A,B}->C

{B} ->D

{C, D}->A

{C, D} -> B

a. What are the sets of candidate keys (5 pt):

מה הם מפתחות הקנדידייט?

(A,B), (C,D),(B,C)

b. What are the **prime** attributes? (1 pt.):

אילו מהתכונות הן פריים?

\_A,B,C,D\_\_\_\_\_\_\_\_\_\_

c. What level of NF (normal form) does the relation adhere to? (1NF, 2NF, 3NF, BCNF(3.5NF), 4NF)? Show why the relation does not adhere to any higher NF (7 pt).

באיזה נורמל פורם הרלציה? נמק.

1NF – assuming true?

2NF,3NF – all keys are prime, so trivially holds.

3.5 – doesn’t hold since B->D but B is a subset of candi-key: (A,B)

3. ERD

להלן מודל ERD עבור ספריית היישוב נווה הדרים:

# מס' סידורי

\* שם ספר

\* שם סופר

\* ז'אנר

\* שנת הוצאה

ספרים

# מס' הוצאה

\* שם הוצאה

\* טלפון

הוצאת ספרים

# מס' עותק

\* סטטוס ספר

O תאריך השאלה

ספרים בספריה

**שימו לב שהמודל שומר על כללי הנרמול לפחות עד 3.5NF. יש לשמור על נירמול זה בכל סעיף בו מתבצע שינוי למודל.**

1. מהם מפתחות ה-foreign key בכל אחד מהטבלאות שניצור עבור המודל (4 נק')?

מספר סידורי, מספר הוצאה. הסבר: לכל הוצאה יש כמה ספרים, מכל ספר יש כמה עותקים בספרייה.

1. הגיע ספר חדש בהוצאה משותפת של "ידיעות ספרים" ומשרד הביטחון. האם מודל ה-ERD הקיים תומך בו (4 נק')?

אם כן, נמקו

אם לא, תקנו את השירטוט הקיים.

לא. צריך לשנות את הקשר בין הוצאת ספרים לספרים לקשר דו כיווני של many to many

1. עקב התרחבות היישוב והקמת שכונות חדשות, הוחלט להקים ספריות נוספות. כיצד נתקן את המודל כך שיתמוך בכמה ספריות? תקנו את השרטוט לעיל **באופן מלא** (4 נק')

4. XML and XSD (8 pt):

State all the reasons why the following XML is not valid according to the given XSD:

כתבו את כל הסיבות מדוע האקס אם אל הבא לא תקף לפי האקס אס די הבא:

XML:

<computer>

 <keyboard>105 keys</keyboard>

 <ramSlot> 4GB </ramSlot>

 <mouse>3 button</mouse>

 <cpu> intel i5 </cpu>

 <monitor>vga</monitor>

</computer>

XSD:

<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">

 <xs:element name="computer">

   <xs:complexType>

     <xs:sequence>

       <xs:element name="keyboard"/>

       <xs:element name="ramSlot" minOccurs="2" maxOccurs="16"/>

       <xs:element name="mouse" />

       <xs:element name="monitor" />

       <xs:element name="cpu" />

       <xs:element name="joystick" minOccurs="0"/>

       <xs:element name="hardDrive" maxOccurs="unbounded"/>

     </xs:sequence>

   </xs:complexType>

 </xs:element>

</xs:schema>

1. there’s only 1 occurance of “ramSlot”.
2. Monitor occurs in the wrong order
3. No hardDrive element.

5. NoSQL (10 pt.):

אתם מעוניינים לכתוב תוכנה שתעבוד עם דאטא רב. את הדאטא תרצו לאחסן במחשבים רבים ומעוניינים בשליפה ואחסון מהירים. לכל אחד מהמקרים הבאים ציינו באיזה מבסיסי הנתונים תשתמשו (תשובות אפשריות: Redis, MongoDB, Cassandra, ElasticSearch, Neo4J). שימו לב: למרות שהתשובה אינה בהכרח חד משמעית (בדר"כ ניתן להשתמש ביותר מבסיס נתונים אחד בדרכים שונות), אף על פי כן, ישנו בסיס נתונים אחד בלבד שהוא המתאים ביותר. נמקו!

1. אתם רוצים בסיס נתונים אשר יאחסן סיפורים, ויתן לכם אפשרות לקבל את הסיפורים הרלוונטיים ביותר לפי מילות מפתח.

elasticSearch מאפשר חיפוש מהיר עם שיטת ניקוד לתוצאות שונות במסד הנתונים

1. מבחינת התוכנה שאתם כותבים, תפקידו של בסיס הנתונים אשר הוא לאחסן מסמך לפי id ולאחר מכן לשלוף את **כל** תוכן המסמך. בחלוף הזמן, תרצו להגדיר חלק מהמסכמים ככאלה שאתם לא צריכים עוד, אך לשמור אותם לעוד חודש ליתר בטחון.

Redis מאפשר מנגנון של תפוגה בנוסף מאפשר שליפת מסמך לפי מפתח בצורה יעילה

1. אתם רוצים בסיס נתונים אשר יתן לכם לבטא קשרים בין ישויות שונות בסביבה בה יש קשרים מרובים. מעניין אתכם מי מקושר למי גם אם הקשר עקיף (למשל דרך ישות נוספת).

Neo4j

מאפשר עבודה עם מבנה שעובד ע״ג גרף, כך ששליפה עקיפה מתבצעת ע״י אלגוריתמים של גרפים בצורה היעילה והמתאימה ביותר למתואר

1. אתם רוצים שהדאטא שלכם יהיה מאורגן בדומה למערכת קבצים. כדי לגשת למידע, אתם מעוניינים לתת את הערכים המדוייקים של כל הattributes הקודמים. אתם מעוניינים בשפת שאילתא הדומה לזו של relational databases.

Cassandra

מאפשר גישה למסמכים רק ע״י תיאור מדוייק של המפתחות המובילים אליהם, השפה מאוד מזכירה

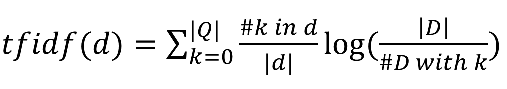
mysql

1. הדאטא שלכם בנוי ממסמכים בהם יש פרטים מספריים רבים. אתם מעוניינים במסד נתונים שיאפשר לכם שאילתות עם חישובים מורכבים על פרטים מתוך המסמכים.

Mongodb מאפשר חישובים מורכבים ע״י aggregation and mapReduce שמיועד להתבזר על מספר מכונות ולבצע חישובים במקביל

6. TF-IDF (7 pt)

**Rank** the following documents according to their TF-IDF score given the following query (no calculator is allowed or required); enter the most relevant document first:

Recall that the TF-IDF formula is:

**דרגו** את המשפטים הבאים לפי ניקוד הטי אף איי די אף.

Q: elephants with horns

D1: Mary went on a date **with** Larry

D2: Google will lock **horns** **with** Microsoft

D3: I never knew **elephants** had **horns**

D4: I saw **elephants** **with** big trunks

D5: Bears **with** blue ears also have **horns**

First (most relevant): \_\_D3\_

Second: \_\_D4\_\_\_

Third: \_D2\_\_\_\_

Fourth: \_\_D5\_\_\_\_\_  
Fifth (least relevant): \_D1\_\_\_

7 Cassandra (10 pt)

Given the following two tables:

נתונות שתי הטבלאות הבאות:

T1 defined as: CREATE TABLE T1 (A INT, B INT, C INT, D INT, PRIMARY KEY((A, B), C, D)));

And:

T2 defined as: CREATE TABLE T2 (A INT, B INT, C INT, D INT, PRIMARY KEY(A, B, C));

For each of the following CQL queries, determine whether it is legal, illegal (e.g. requires ALLOW\_FILTERING) or syntax error (e.g. “EAT C WITH T1”):

לכל אחת מהשאילתות הבאות כיתבו האם היא חוקית, לא חוקית, או בכלל לא בסינטאקס חוקי לCQL.

1. SELECT \* FROM T1 WHERE A=37 \_\_illegal\_
2. SELECT \* FROM T2 WHERE A=345 INNER JOIN T1 ON T1.C=T2.C \_\_Syntax\_\_
3. SELECT \* FROM T2 WHERE A>54 \_\_illegal\_\_\_\_
4. SELECT \* FROM T1 WHERE A=145 AND B=356 AND D=43 \_\_\_illegal\_
5. SELECT \* FROM T1 WHERE C<679 AND A=465 AND B=195 \_\_legal\_\_\_

8. Mongo DB (9 pt):

Consider the car collection which contains the following documents;

db.car.insert([

{car\_id:"c1", name:"Audi", color:"Black", current\_speed:50},

{car\_id:"c2", name:"Polo", color:"White", current\_speed:65},

{car\_id:"c3", name:"Alto", color:"White", current\_speed:75},

{car\_id:"c4", name:"Santro", color:"Black", current\_speed:150},

{car\_id:"c5", name:"Subaru", color:"Black", current\_speed:100},

{car\_id:"c6", name:"Zen", color:"Blue", current\_speed:97} ] )

What will be the content of my\_out after the following mapreduce call (i.e. what will we get when we type: db.my\_out.find()):

db.car.mapReduce(

function (){

if ( this.current\_speed < 120 ) {

emit(this.color, this.current\_speed);

}

},

function(key, val1) {

var total =0;

for (var i = 0; i < val1.length; i++) {

total += val1[i];

}

return total / val1.length;

},

{out: "my\_out"});

מה תכיל my\_out לאחר הרצת הקוד לעיל (כלומר, מה נקבל כשנקליד db.my\_out.find()):

my\_out:[

{\_id:’Black’,Value:75},

{\_id:’White’,value:70},

{\_id:’Blue’,value:97}]

9 JAVA Streams (5 pt):

What will be the output of the following Java streams program:

List<String> myList = Arrays.asList("camel", "zebra", "you", "apple", "banana", "me");

myList.stream().map(a->a+a).filter(a->a.length()>=7).sorted().forEach(System.out::println);

מה יהיה הפלט של התוכנית?

appleapple

bananabanana

camelcamel

zebrazebra

10 Linear regression (14 pt)

Recall that for linear regression, we have defined the following loss function

https://lh6.googleusercontent.com/3XpcoJ4CjAchf3zlaoBtbvtlgsTaT7XRBnua7zwOoQnEJlPjhdKJxuDh7kGJ8BHP0y9nuD_nRYtasoBXciRjU4c_PvwtADrhhDbnh2kJcwvYrmF9c2giMMan6_-FBZauV6pgHM5E

Recall that the hypothesis in linear regression is: h(x) = xw + b

Given the following data: x = [3, 2, 0, -4] and y = [2, 1, 1, -2]

(Assume batch gradient descent).

1. What are the values of the **gradient** (both according to w and b) when w=0 and b=0 (provide numerical values) (7pt)?

מה **הגראדינט** ב w=0,b=0 (יש לתת ערך מספרי)?

1. Assume the learning rate (alpha) is 0.1, what will be the value of w and b in the next iteration (3pt)? Use the gradient calculated in the section a.

הניחו שאלפא=0.1, מה יהיו הערכים של w וb באיטרציה הבאה? (השתמשו בגרדיאנט שחישבתם בסעיף א')

1. According to the values you found for w and b, what will be the prediction for x=3? When is the prediction for x=3 more accurate, with w=0, b=0 or with the values you found in the previous section? (show your calculations) (4pt).

לפי הערכים שמצאתם בסעיף הקודם, מה יהיה הניבוי לx=3? מתי הניבוי לx=3 מדוייק יותר כשw=0,b=0 או עם הערכים שמצאתם בסעיף הקודם? (יש להראות חישוב)

* The updated values of w and b give a better prediction for <3,2>