

ה א ו נ י ב ר ס י ט ה ה פ ת ו ח ה

20574

מימוש מערכות
בסיסי נתונים
חוברת הקורס – אביב 2021ב

כתבה : יעל ארד

פברואר 2021 – סמסטר אביב – תשפ"א

פנימי – לא להפצה.

© כל הזכויות שמורות לאוניברסיטה הפתוחה.

תוכן העניינים

א	אל הסטודנטים
ג	1. לוח זמנים ופעילויות
ה	2. תיאור המטלות
ה	3. חומר הלימוד הנדרש לפתרון המטלות
ה	4. התנאים לקבלת נקודות זכות בקורס
1	ממ"ח 01
9	ממ"ן 11
13	ממ"ח 02
19	ממ"ן 12
21	ממ"ן 13

אל הסטודנטים,

עם הצטרפותכם לקורס "מימוש מערכות בסיסי נתונים", אני מאחלת לכם הצלחה רבה, ומקווה שתמצאו בו עניין ותועלת.

החוברת שלפניכם כוללת את לוח הזמנים של הקורס, תנאים לקבלת נקודות זכות ואת המטלות.

לקורס קיים אתר באינטרנט בו תמצאו חומרי למידה נוספים.
בנוסף, האתר מהווה עבורכם ערוץ תקשורת עם צוות ההוראה ועם סטודנטים אחרים בקורס.
פרטים על למידה מתוקשבת ואתר הקורס, תמצאו באתר שה"ם בכתובת:

<http://telem.openu.ac.il>

מידע על שירותי ספרייה ומקורות מידע שהאוניברסיטה מעמידה לרשותכם, תמצאו באתר הספרייה באינטרנט www.openu.ac.il/Library.

אפשר לפנות אלי בכל שאלה בימי ב' בין השעות 11:00-12:00, בטלפון 02-5795343 או באימייל rael@openu.ac.il. במידת הצורך אפשר לתאם פגישה.

לתשומת לב הסטודנטים הלומדים בחו"ל:

למרות הריחוק הפיסי הגדול, נשתדל לשמור אתכם על קשרים הדוקים ולעמוד לרשותכם ככל האפשר.

הפרטים החיוניים על הקורס נכללים בחוברת הקורס וכן באתר הקורס.
מומלץ מאוד להשתמש באתר הקורס ובכל אמצעי העזר שבו וכמובן לפנות אלינו במידת הצורך.

בברכה,

יעל ארד

מרכזת ההוראה בקורס.

1. לוח זמנים ופעילויות (20574 / ב'2021)

שבוע לימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	תאריך אחרון למשלוח ממ"ן (למנחה)
1	05.03.2021-28.02.2021	פרק 1	
2	12.03.2021-07.03.2021	פרק 2	
3	19.03.2021-14.03.2021	פרק 3	
4	26.03.2021-21.03.2021	פרק 3	ממ"ח 01 26.3.2021
5	02.04.2021-28.03.2021 (א-ו פסח)	פרק 4	
6	09.04.2021-04.04.2021 (ה יום הזכרון לשואה)	פרק 4	ממ"ן 11 9.4.2021
7	16.04.2021-11.04.2021 (ד יום הזיכרון, ה יום העצמאות)	פרק 4	
8	23.04.2021-18.04.2021	פרק 5	
9	30.04.2021-25.04.2021 (ו ל"ג בעומר)	פרק 5	ממ"ח 02 30.4.2021

* התאריכים המדויקים של המפגשים הקבוצתיים מופיעים ב"לוח מפגשים ומנחים".

לוח זמנים ופעילויות - המשך

שבוע הלימוד	תאריכי שבוע הלימוד	יחידת הלימוד המומלצת	תאריך אחרון למשלוח הממ"ן (למנחה)
10	07.05.2021-02.05.2021	פרק 6	
11	14.05.2021-09.05.2021	פרק 6 / פרק 7	
12	21.05.2021-16.05.2021 (ב שבועות)	פרק 7	ממ"ן 12 21.5.2021
13	28.05.2021-23.05.2021	פרק 7 / פרק 8	
14	04.06.2021-30.05.2021	פרק 8	
15	11.06.2021-06.06.2021	חזרה	ממ"ן 13 11.6.2021

מועדי בחינות הגמר יפורסמו בנפרד

* התאריכים המדויקים של המפגשים הקבוצתיים מופיעים ב"לוח מפגשים ומנחים".

2. תיאור המטלות

בקורס 5 מטלות. 3 ממ"נים ו-2 ממ"חים.
הממ"חים מתרגלים מגוון רחב של נושאים ומושגים מתוך פרקי הלימוד הרלוונטיים.
הממ"נים מכילים שאלות גדולות ומורכבות יותר, שמאפיינות שאלות מבחינות.
משקל כל מטלה הוא 5 נקודות. חובה להגיש מטלות במשקל כולל של 15 נקודות לפחות. אפשר להגיש תמהיל כלשהו של ממ"נים וממ"חים.

ללא צבירת 15 נקודות בהגשת המטלות לא ניתן יהיה לגשת לבחינת הגמר.

3. חומר הלימוד הדרוש לפתרון המטלות

המטלות מלוות את פרקי הלימוד בקורס. להלן פירוט המטלות והפרקים שאליהם מתייחסת כל מטלה. בחלק מהמטלות תופענה גם שאלות המתייחסות לפרקים קודמים, שכבר נלמדו.

<u>חומר הלימוד</u>	<u>ממ"ח / ממ"ן</u>
פרקים 1,2,3	ממ"ח 01 ; ממ"ן 11
פרקים 3,4,5	ממ"ח 02 ; ממ"ן 12
פרקים 5,6,7,8	ממ"ן 13

4. התנאים לקבלת נקודות זכות בקורס

- א. צבירת 15 נקודות לפחות במטלות.
- ב. ציון של 60 לפחות בבחינת הגמר.
- ג. ציון סופי בקורס של 60 נקודות לפחות.

הציון הסופי בקורס מורכב מציון הבחינה בשיעור 85%, וממוצע ציוני המטלות בשיעור 15%.
משקל המטלות יכול לעלות עד ל-25% מציון הקורס, אם אינו מוריד את הממוצע שחושב לפי השיעורים לעיל.

לתשומת לבכם!

כדי לעודדכם להגיש לבדיקה מספר רב של מטלות הנהגנו את ההקלה שלהלן :

אם הגשתם מטלות מעל למשקל המינימלי הנדרש בקורס, **המטלות** בציון הנמוך ביותר, שציוניהן נמוכים מציון הבחינה (**עד שתי מטלות**), לא יילקחו בחשבון בעת שקלול הציון הסופי.

זאת בתנאי שמטלות אלה **אינן חלק מדרישות החובה בקורס** ושהמשקל הצבור של המטלות האחרות שהוגשו, מגיע למינימום הנדרש.

זכרו! ציון סופי מחושב רק לסטודנטים שעברו את בחינת הגמר בציון 60 ומעלה והגישו מטלות כנדרש באותו קורס.

מטלת מחשב (ממ"ח) 01

הקורס: 20574 - מימוש מערכות בסיסי נתונים

חומר הלימוד למטלה: פרקים 1-3 - אחסון מידע ואינדקסים

משקל המטלה: 5 נקודות

מספר השאלות: 10

מועד אחרון להגשה: 26.3.2021

סמסטר: 2021ב

את התשובות לממ"ח יש לשלוח באמצעות מערכת שאלת"א

בכתובת <http://www.openu.ac.il/sheilta/>

הממ"ח נבדק בצורה ממוחשבת. אין לשלוח את פתרון הממ"ח למנחה!

שימו לב, המנחה לא יכול לדחות לכם את תאריך הגשת הממ"ח. זה אוטומטי והוא לא יכול להשפיע על כך. לכן אין טעם לפנות למנחה בבקשות כאלו.

זכרו לשמור את האסמכתא שאתם מקבלים מהמחשב לאחר הגשת הממ"ח. אם לא קיבלתם מספר אסמכתא כאילו לא הגשתם את הממ"ח

מטלות מחשב – ממ"ח

- הממ"ח הוא "מבחן רב-ברירה" (מבחן אמריקאי) הנבדק באמצעות מחשב.
- יש להקפיד לשלוח את התשובות לממ"ח במועד שנקבע. אל תקדימו במשלוח התשובות יותר משבוע לפני התאריך הנקוב בלוח הזמנים לאותו ממ"ח.
- בתוך שלושה שבועות מהתאריך האחרון, המצוין בלוח הזמנים, תקבלו לבתיכם הודעה שתכלול:
- התשובות הנכונות לממ"ח לעומת תשובותיכם.
 - הערות (אם תהיינה כאלה) המתייחסות לתשובותיכם.
 - ציונכם בממ"ח ומשקלו של ממ"ח זה בחישוב הציון הסופי בקורס.

הנחיות לפתרון הממ"ח

יש לקרוא כל שאלה פעמים מספר ולהתייחס לכל מלה בה. קריאה זהירה והבנה מדויקת של משמעות כל משפט בשאלה הן תנאי ראשון להצלחתכם בממ"ח.

לכל שאלה יש רק תשובה נכונה אחת. קראו תחילה את כל האפשרויות הנתונות, החליטו מהי האפשרות הנכונה ביותר מבין כל האפשרויות ואז סמנו אפשרות זו.

אם נדמה לכם שיש לשאלה אחת שתי תשובות נכונות, או אף שלוש, ייתכן כי תגלו, לאחר קריאת כל התשובות, תשובה אחת האומרת "שלוש התשובות הקודמות נכונות". במקרה כזה, מובן שתסמנו תשובה זו ואותה בלבד כנכונה. אם לא מופיע משפט מסוג זה, הרי רק אחת התשובות

נכונה. קיימת גם אפשרות שאין כל תשובה נכונה, ובמקרה כזה תינתן לכם אפשרות לסמן כנכונה את התשובה: "אין אף תשובה נכונה."

משלוח הממ"ח

יש לשלוח את התשובות לממ"ח באמצעות מערכת **שאלתא** (שירותים אינטראקטיביים לסטודנטים באמצעות תקשורת ואינטרנט). הסבר על המערכת ניתן למצוא בחוברת הקורס וכן באתר האו"פ באינטרנט בכתובת: www.openu.ac.il/sheilta במערכת ניתן לראות את תוצאות בדיקת הממ"ח מיד עם פרסומן.

הוראות למילוי תשובות ומשלוח ממ"ח באמצעות מערכת שאלתא

1. היכנסו למערכת שאלתא. (הכניסה היא מאתר הבית של האו"פ בכתובת www.openu.ac.il/sheilta באמצעות שם המשתמש והסיסמה שנשלחה אליך).
2. היכנסו לתפריט **קורסים**.
3. בדף הקורסים, בחרו ב**פירוט הקורס המבוקש**.
4. בפירוט הקורס, היכנסו ל**קישור מטלת מחשב**.
5. בחרו בממ"ח שברצונכם לשלוח ע"י הקלקה על הכפתור שמימין לממ"ח ולחצו על **הזנת תשובות**.
6. הזנו את התשובות לכל השאלות. (לבחירת התשובה לחצו על החץ שבכל תיבה).
7. שלחו את תשובותיך על-ידי לחיצה על **לחצן שלח**.
8. בתפריט **פניות** תוכלו לראות את פרטי הממ"ח ששלחתם.

ערעור על ציון בממ"ח

ערעור על ציון שקיבלתם בממ"ח יוגש למרכז ההישגים הלימודיים תוך שבוע מיום קבלת תוצאות הממ"ח, ובצירוף ההודעה על הציון שקיבלתם מהמחשב (או צילומה). אין ערעור נוסף על ההחלטה בערעור זה.

שאלה 1 – דיסק מגנטי

נתון דיסק מגנטי עם הנתונים הבאים :

- הדיסק מסתובב במהירות של 6000 סיבובים בדקה (6000 RPM).
- יש 100 גלילים, הממוספרים מ-1 עד 100.
- הזזת הראש הקורא על-פני t גלילים אורכת $(1+t/100)$ מילישניות.
- גודל גוש (בלוק) – 32 מגה-בתים.
- קצב העברה הוא 16 גיגה בתים לשניה.

נניח שסיימנו לקרוא בלוק בגליל מספר 80, ואנחנו רוצים עכשיו לקרוא בלוק בגליל מספר 10. כמה זמן יידרש עד שהבלוק יופיע בזיכרון? (נניח שהמרכיבים היחידים שנתייחס אליהם הם זמן חיפוש, זמן השהייה וזמן העברה, ונתעלם מעיכובים מכניים אחרים).

א. 10.5

ב. 8.7

ג. 7

ד. 8.5

שאלה 2 – RAID

דיסק מגנטי מנוהל עם יתירות מסוג RAID 1 ללא stripping. כלומר, מוחזק דיסק נוסף, זהה בתוכנו לחלוטין. נתון כי שיעור התקלות (failure rate) בכל אחד מהדיסקים הוא 20% בשנה, כלומר משך הזמן בממוצע עד שתתרחש תקלה (Mean Time to Failure - MTTF) בכל אחד מהם הוא 5 שנים. כמו כן נניח כי לוקח 12 שעות להחליף דיסק שהתקלקל ולהעתיק אליו את המידע מן הדיסק התקין. (כל הנתונים הללו הרבה יותר גרועים מהמציאות...)

מהו משך הזמן הממוצע עד שתתרחש תקלה עם אובדן מידע? כלומר: משך הזמן הממוצע עד שתתרחש תקלה בדיסק השני לפני השלמת החלפת הדיסק התקול.

שימו לב שאין צורך לבדוק לשם כך את ההסתברות ששני הדיסקים יתקלקלו בבת אחת, אלא רק את ההסתברות של דיסק אחד להתקלקל בפרק זמן של 12 שעות מסוימות בשנה כלשהי.

א. 18,250 שנים

ב. 60 שנים

ג. 3,650 שנים

ד. 10 שנים

שאלה 3 – אחסון עמודות

יחס $r(A,B,C,D)$ מכיל מיליון רשומות. השדה A ברוחב 8, השדה B ברוחב 12, השדה C ברוחב 10 והשדה D ברוחב 15. גודל דף 4,000 בתים. מאחסנים את היחס באחסון עמודות, כל עמודה מאוחסנת ברצף. אין מפצלים ערך בודד בין דפים ואין מערבבים בדף אחד ערכים של עמודות שונות. בכמה דפים יאוחסן היחס כולו?

א. 11,279 דפים.

ב. 11,235 דפים.

ג. 11,264 דפים.

ד. 10,764 דפים.

שאלה 4 – זיכרון הבזק

נניח שמשתמשים בזיכרון הבזק עם שכבת תרגום (FTL), הכוללת טבלת תרגום (translation table). הטבלה ממפה דפים לוגיים לדפים פיזיים בזיכרון ההבזק, ובכך מאפשרת לנהל החלפה של דפים פיזיים העשויה להידרש כדי לטפל בשחיקת הדפים הפיזיים (ראו בעמ' 440 בספר הקורס). הטבלה מוחזקת בזיכרון הפנימי.

נפח האחסון הכולל בהתקן ההבזק הוא 64 ג'יגה בתים, וכל דף הוא בן 4,096 בתים. כתובת של דף היא באורך 32 סיביות. טבלת התרגום מאוחסנת כמערך בזיכרון. מה יהיה נפח הטבלה?

א. 64 מגה בתים

ב. 64 ג'יגה בתים

ג. 16 ג'יגה בתים

ד. 16 מגה בתים

שאלה 5 – מדיניות החלפת דפים

נתון כי יש 3 דפי חוצץ. סורקים קובץ במבנה ערימה בן 10 דפים פעמיים מתחילתו ועד סופו ופעם שניה מהסוף להתחלה. אם משתמשים במדיניות LRU – בכמה מבין 20 הדפים שנקרא הדף יהיה כבר בדיסק ולא נצטרך לקרוא אותו בשנית ("hit")?

א. 1

ב. 2

ג. 3

ד. 0

שאלה 6 – ארגון רשומות בדפים

על דיסק מאוחסן קובץ employee עם 50,000 רשומות באורך קבוע, ואין מפצלים רשומות בין דפים. גודל דף במערכת 4,000 בתים. בכל רשומה השדות הבאים :

- ת.ז. - 9 בתים
- שם משפחה - 20 בתים
- שם פרטי - 20 בתים
- כתובת - 40 בתים
- טלפון - 14 בתים
- תאריך לידה - 10 בתים
- ת.ז. ממונה - 9 בתים
- מחלקה - 4 בתים
- קוד תפקיד - 4 בתים

כמה בתים סה"כ 'מבוזבזים' עקב כך שלא מפצלים רשומות בין דפים? הנפח המבוזבז כולל גם את יתרת הדף האחרון, אם אינו מלא לגמרי ברשומות של הקובץ.

א. 166,600

ב. 166,700

ג. 168,000

ד. 168,600

שאלה 7 – עצי B⁺

לעץ B⁺ עם n=5 מכניסים בזה אחר זה את הערכים הבאים (משמאל לימין) :

5, 11, 85, 32, 14, 7, 53, 69, 46, 19, 28, 30

כמה פיצולים יהיו במהלך סדרת ההכנסות, בהנחה שמפצלים 2 שמאלה ו-3 ימינה?

א. 1

ב. 2

ג. 3

ד. 0

שאלה 8 – אינדקס בגיבוב מתרחב

בונים אינדקס עבור שדה בקובץ בן 1,000 רשומות. השדה מכיל את 100 הערכים המספריים השלמים בין 1 ל-100. נתון כי לאחסון ערך כזה דרושים 2 בתים, לאחסון מצביע (כתובת) במערכת זו דרוש בית אחד, וכל דף מכיל 25 בתים.

האינדקס נבנה בגיבוב מתרחב, כאשר רשומות אינדקס לאותו ערך מופיעות כמספר ההופעות של הערך. אם אין מקום בסל המתאים לכל החזרות (לא עוזר להרחיב את הפונקציה..) – יוקצה דף גלישה.

כמה דפים יידרשו לאחסון האינדקס? יש לכלול בחישוב נפח האחסון גם את דפי הגלישה וגם את נפח טבלת כתובות הסלים.

- א. 100
- ב. 105
- ג. 200
- ד. 206

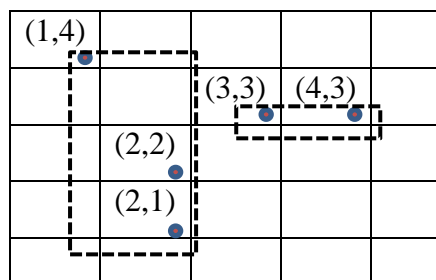
שאלה 9 – אינדקסים במפות סיביות

ננתוני השאלה הקודמת, נניח עתה שבונים את האינדקס לשדה X במפות סיביות. מה יהיה נפח האחסון הכולל של האינדקס?

- א. 500
- ב. 600
- ג. 1,000
- ד. 100

שאלה 10 – עצי R

נתונות נקודות ומלבנים מקיפים שמאוחסנים בשני עלים במבנה של עץ R.



איזה מבין המלבנים הבאים – המתוארים באמצעות הפינה השמאלית התחתונה והפינה הימנית העליונה – יכול להופיע בצומת ההורה של עלים אלה?

א. $[(2,1),(5,5)]$

ב. $[(1,1),(4,3)]$

ג. $[(1,4),(4,4)]$

ד. $[(1,1),(5,5)]$

מטלת מנחה (ממ"ן) 11

הקורס: 20574 - מימוש מערכות בסיסי נתונים

חומר הלימוד למטלה: פרקים 1-3 - אחסון מידע ואינדקסים

משקל המטלה: 5 נקודות

מספר השאלות: 5

מועד אחרון להגשה: 9.4.2021

סמסטר: 2021ב

קיימות שתי חלופות להגשת מטלות:

שליחת מטלות באמצעות מערכת המטלות המקוונת באתר הבית של הקורס
שליחת מטלות באמצעות הדואר או הגשה ישירה למנחה במפגשי ההנחיה
הסבר מפורט ב"נוהל הגשת מטלות מנחה"

שאלה 1 – אינדקסים (24%)

בסיס נתונים כולל יחס $P(id, name, age, salary)$. ביחס 1,000 שורות, והערכים בעמודת המפתח id הם הערכים 1, 2, 3, ..., 1,000. היחס מאוחסן באופן ממוין לפי id , 5 רשומות בדף.

א. בונים אינדקס צפוף פשוט (לא במבנה עץ) על השדה id , $E1$. האינדקס מאוחסן באופן ממוין, 10 רשומות אינדקס בדף. חיפוש באינדקס מבוצע על-ידי סריקה סדרתית של רשומות האינדקס, החל מהרשומה הראשונה (הערך הקטן ביותר). כמה גישות לדיסק יידרשו כדי לאתר בעזרת האינדקס רשומה עם הערך $id=567$?

ב. כדי לשפר את ביצועי האינדקס הקודם, בונים אינדקס דליל $E2$, שכל רשומת אינדקס בו מצביעה לרשומה הראשונה של דף מבין דפי $E1$, וכך לכל דפי $E1$. כמה גישות לדיסק יידרשו עתה כדי לאתר את הרשומה עם הערך $id=567$?

ג. עתה נגדיר אינדקס צפוף $E3$, על השדה $salary$, 10 רשומות אינדקס בדף. נניח שיש 50 ערכים שונים בשדה $salary$, שמתפלגים באופן אחיד בין רשומות הקובץ P . כמו כן נניח כי ערכים חוזרים מופיעים באינדקס כמספר המופעים שלהם. כלומר, תהיה רשומת אינדקס נפרדת לכל מופע של הערך עם מצביע לרשומה שבה הוא מופיע. כמה גישות לדיסק יידרשו במקרה הגרוע כדי לאתר את כל רשומות P עם הערך המקסימלי בשדה $salary$?

ד. הפעם נתייחס לאינדקס E4, שאף הוא אינדקס צפוף על השדה $salary$, אלא שבו כל ערך מופיע פעם אחת בלבד, והמצביע ברשומה מצביע לדף שמכיל מצביעים לכל הרשומות שיש בהן את הערך הזה. אם דף אחד אינו מספיק לאחסון כל המצביעים – המצביע מרשומת האינדקס יהיה לדף הראשון בשרשרת דפים שמכילים את כל המצביעים הדרושים. כמה גישות לדיסק יידרשו לאיתור כל רשומות P עם הערך המקסימלי בשדה $salary$?

בכל הסעיפים (אם יש צורך) אפשר להניח שרוחב מצביע זהה לרוחב של id . בכל הסעיפים, חיפוש באינדקס הוא בסריקה סדרתית החל מהערך הקטן ביותר.

שאלה 2 – אחסון עמודות (20%)

נתון יחס $r(A_1, A_2, \dots, A_5)$ בן 100 מיליון רשומות. התכונה A_1 היא מספר מזהה של הרשומה ורוחבה 4 בתים. גם שאר התכונות הן ברוחב 4 בתים כל אחת. מספר הערכים השונים זה מזה בכל אחת מהעמודות שאינה מפתח הוא $1/100$ ממספר הרשומות. רוחב דף הוא 4,096 בתים, רוחב מצביע 4 בתים, יש 100,000 דפי חוצץ לרשות החישובים.

מאחסנים את היחס באחסון עמודות, באופן הבא:

- כל עמודה מאוחסנת בצורה ממוינת לפי ערכי העמודה, ולצד כל ערך מופיעה רשימת כל ערכי A_1 שמופיעים עם ערך זה ביחס המקורי.
- יש אינדקס בעץ B^+ על השדה A_1 , כך שבעלים, לצד כל ערך של A_1 יש 4 מצביעים, אחד לכל אחת מהעמודות הנוספות, כשהמצביע מוביל לדף שבו מאוחסן ערך זה באחסון העמודה הרלוונטית. בסוף העלה יש מצביע לעלה הבא. הצמתים הפנימיים בנויים כמו בכל עץ B^+ .

- חשבו מהו נפח האחסון הכולל של הקובץ בצורת אחסון זו.
- תארו כיצד אפשר לחשב במערכת זו את השאילתה הבאה, והעריכו את עלות החישוב במספר גישות לדיסק:

```
select A2
from r
where A2 > (select min (A4)
             from r)
```

שאלה 3 – אינדקסים במפות סיביות (16%)

נתון יחס המכיל מידע על מכוניות :

<i>license-no</i>	<i>brand</i>	<i>model</i>	<i>color</i>	<i>risk</i>
11-111-11	Opel	corssa	grey	low
22-222-22	Opel	corssa	red	medium
33-333-33	Peugeot	206	black	medium
44-444-44	BMW	A	black	high

- א. בנו אינדקסים במפות סיביות עבור השדות *brand* ו-*color* (כל אחד בנפרד).
- ב. הסבירו כיצד לחשב בעזרת אינדקסים אלה את השאילתה "מצאו כמה מכוניות אדומות של Opel יש עם דרגת סיכון בינונית".

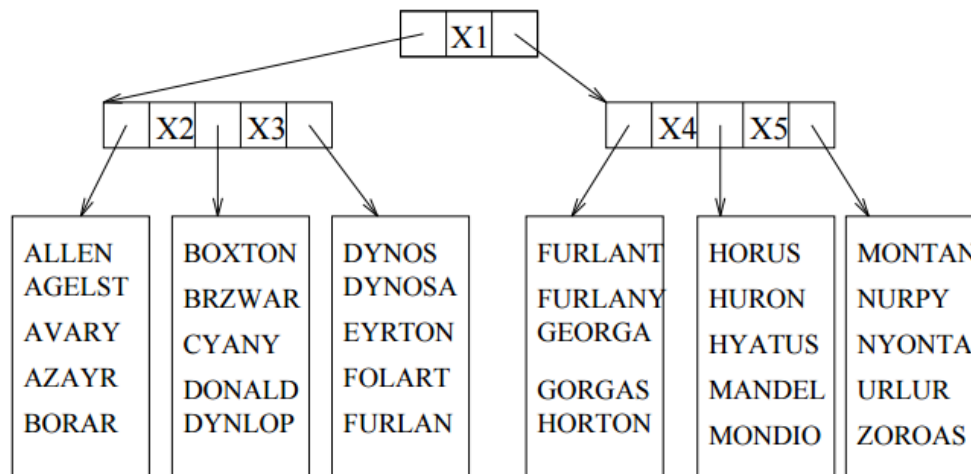
שאלה 4 (20%) – אינדקסים ואחסון

נתון יחס $r(A_1, A_2)$ בן 100 מיליון רשומות. התכונה A_1 היא מזהה של הרשומה, ורוחבה 10 בתים. רוחב השדה A_2 30 בתים. רוחב כתובת (מצביע) – 10 בתים. רוחב דף הוא 4,010 בתים.

- א. רשומות היחס מאוחסנות בשלמותן בסלים של מבנה גיבוב מתרחב, שבנוי לפי השדה A_1 . מהו הנפח הכולל הנדרש לאחסון הקובץ, בדפים? הנפח הכולל פירושו גם הסלים וגם הדפים הדרושים לאחסון טבלת כתובות הסלים.
- ב. בשלב מסוים מסתבר שיש מספר רב של שאילתות עם תנאי טווח המתייחסות דווקא לשדה A_2 , ולכן רוצים לשנות את המבנה, כך שהרשומות יאוחסנו בעלים של עץ B^+ שיהיה בנוי לפי השדה A_2 , תוך שמירה על תפוסה ממוצעת של 80% בעלים. אין צורך לטפל באופן מיוחד בתפוסת הצמתים הפנימיים. את המעבר מבצעים בטעינה באצווה. מהו הנפח הכולל של הקובץ במבנה החדש?

שאלה 5 - עצי B^+ , עץ תחיליות (20%)

- א. (8 נק') נתון אינדקס במבנה של עץ תחיליות, כמתואר בשרטוט. שדה החיפוש הוא שמות משפחה של סטודנטים. הערכים (התחיליות) בצמתים הפנימיים לא רשומים במפורש, אלא מיוצגים באמצעות המשתנים $X1...X5$. כתבו מהי התחילית שמייצג כל אחד מחמשת המשתנים.



ב. (12 נק') הניחו כי שדה החיפוש של האינדקס לעיל מאוחסן ברוחב קבוע של 20 בתים לכל שם, וכי רוחב מצביע הוא 4 בתים. עתה נניח כי עלי העץ הנתון בשרטוט הם עלים בעץ B^+ "רגיל" (לא עץ תחיליות), וכי תפוסתם היא $5/8$ (62.5%).

- מהו גודל דף על-פי נתונים אלה? (זה לא צריך להיות דומה לגודל דף מקובל.)
- כמה ערכים צריכים להיות בצומת לכל הפחות וכמה לכל היותר?
- מה יהיו הצמתים הפנימיים של העץ ה"רגיל" על-פי הנ"ל?

מטלת מחשב (ממ"ח) 02

הקורס: 20574 - מימוש מערכות בסיסי נתונים

חומר הלימוד למטלה: פרקים 4,5 - עיבוד ואופטימיזציה של שאילתות

משקל המטלה: נקודות

מספר השאלות: 10

מועד אחרון להגשה: 30.4.2021

סמסטר: 2021ב

קיימות שתי חלופות להגשת מטלות:

שליחת מטלות באמצעות מערכת המטלות המקוונת באתר הבית של הקורס
שליחת מטלות באמצעות הדואר או הגשה ישירה למנחה במפגשי ההנחה
הסבר מפורט ב"נוהל הגשת מטלות מנחה"

שאלה 1 - שקילויות אלגבריות

נתון בסיס נתונים עם שני יחסים: $R(A,B)$ ו- $S(B,C,D)$. נתונה השאילתה הבאה באלגברה של יחסים המופנית לבסיס נתונים זה:

$$\pi_{AD}[\sigma_{C=5}(R \bowtie S)]$$

יש לכתוב את הביטוי באופן שקול כך שהטלות ובחירות יחושבו במיקום הפנימי ביותר האפשרי. איזה מן הביטויים הבאים עומד בתנאים אלה:

א. $\pi_D[\sigma_{C=5} S] \bowtie \pi_A(R)$

ב. $\pi_{AD}([\sigma_{C=5} S] \bowtie \pi_B R)$

ג. $\pi_{AD}(\pi_{BD}[\sigma_{C=5} S] \bowtie R)$

ד. $\pi_{AD}[\sigma_{C=5} S] \bowtie \pi_{AD}(R)$

שאלה 2 - עיבוד פעולות בחירה

נתון יחס $R(A,B,C,D)$. כל השדות באותו רוחב. ביחס 100,000 רשומות, 10 בדף. ידוע כי יש 500 דפי חוצץ לרשות החישובים. היחס אינו ממור, אך מסתבר שהוא חולק למנות ממוינות (כל אחת בגודל 500 דפים) בחישוב קודם ויש אפשרות לגשת למנות אלה בנפרד. רוצים למצוא את כל הרשומות שבהן $A=20$. לשם כך מבצעים חיפוש בינרי בתוך כל אחת מהמנות הממוינות.

כמה פעולות קלט/פלט לכל היותר נבצע בשיטה זו עד לאיתור כל הרשומות המבוקשות?

א. 500

ב. 13

ג. 180

ד. 9

שאלה 3 – מיון חיצוני

במערכת מסוימת יש 50 דפי חוצץ. מהו גודל הקובץ המקסימלי (בדפים) שאפשר למיין בתנאים אלה בשני מעברים (מעבר מיון אחד ומעבר מיזוג אחד)?

א. 2,450

ב. 2,401

ג. 2,500

ד. 2,501

שאלה 4 – צירוף בלולאה מקוננת בגושים

נתונים שני יחסים $R(X,Y)$ ו- $S(Y,Z)$. ב- R 100,000 רשומות, 40 רשומות בדף. ב- S 20,000 רשומות, 50 רשומות בדף. הקבצים אינם ממוינים ולא ידוע על אינדקסים כלשהם. יש 102 דפי חוצץ לרשות החישובים. מחשבים צירוף טבעי בין שני היחסים בלולאה מקוננת בגושים, באופן היעיל ביותר. כמה פעולות קלט/פלט יידרשו להשלמת החישוב (לא לוקחים בחשבון את עלות כתיבת התוצאה)?

א. 10,800

ב. 10,400

ג. 12,500

ד. 11,000

שאלה 5 – צירוף-מיון-מיזוג

עבור אותם נתונים שבשאלה הקודמת (שאלה 4), נניח עתה שמחשבים את הצירוף באמצעות צירוף-מיון-מיזוג, באופן הבא: יוצרים מ- R מנות ממוינות לפי Y ככל שמאפשר גודל החוצץ. יוצרים מ- S מנות ממוינות לפי Y ככל שמאפשר גודל החוצץ. עתה קוראים דף תורן מכל אחת מהמנות הממוינות של שני הקבצים הממוינים לצורך המיזוג המשלים את המיון. תוך כדי המיזוג

מחשבים את הצירוף ומוציאים את התוצאה לפלט. מה יהיה מספר פעולות הקלט/פלט הכולל לצורך השלמת חישוב זה (ללא כתיבת התוצאה הסופית לדיסק)?

א. 8,700

ב. 11,600

ג. 9,900

ד. 14,500

שאלה 6 – צירוף-גיבוב

עבור אותם נתונים שבשאלה 4 – נחשב עתה את הצירוף באמצעות צירוף-גיבוב היברידי. מהו מספר הסלים המינימלי שאפשר לגבב אליו את שני היחסים כדי שנוכל לנצל את נפח החוצץ בצירוף-גיבוב היברידי?

א. 4

ב. 5

ג. 25

ד. 24

שאלה 7 – הערכת גודל תוצאה

נתונים שני יחסים, $R(A,B,C)$ ו- $S(B,C,D)$. כל השדות מאותו טיפוס נתונים, ברוחב 10 בתים כל אחד.

ב- R 100,000 רשומות; $V(R, A) = 100$; $V(R, B) = 50$; $V(R, C) = 20$.

ב- S 10,000 רשומות; $V(S, B) = 20$; $V(S, C) = 50$; $V(S, D) = 10$.

רוחב דף במערכת 4,000 בתים. אין מפצלים רשומות בין דפים.

כמה דפים נצפה שיהיו בתוצאת הביטוי: $R \bowtie (\sigma_{B=5}(S))$? הניחו כי אין חוזרים על העמודות המשותפות בתבנית התוצאה.

א. 200

ב. 40

ג. 2

ד. 20

שאלה 8 – היסטוגרמות

מבצעים צירוף של שני יחסים: $r(A,B) \bowtie s(B,C)$. להלן היסטוגרמות הנותנות את השכיחות (מספר הרשומות שבהן מופיע כל ערך) של חמשת הערכים הנפוצים ביותר בעמודה B בשני היחסים. יש סה"כ 10 ערכי B שונים זה מזה בשני היחסים. לגבי הערכים הפחות שכיחים – אנו מניחים הנחת התפלגות אחידה:

	0	1	2	3	4	Others
r.B	500	60	400	150	90	800
s.B	100	200	300	250	150	500

כמה רשומות יהיו בתוצאת הצירוף לפי נתונים אלה?

- א. 633,000
- ב. 300,000
- ג. 313,000
- ד. 233,000

שאלה 9 – עיבוד שאילתות – index only

נתון יחס $R(A,B,C)$ בן מיליון דפים. בנפרד מוגדר אינדקס על A קטן בהרבה. איזו מן השאילתות הבאות ניתן לחשב בעזרת האינדקס בלבד, בלי לגשת לקובץ?

- א. `select * from R where A=2`
- ב. `select sum(A) from R where A>2`
- ג. `select A from R where B=2`
- ד. `select A from R where (A=2 or B=5)`

שאלה 10 – ביטול כפילויות

ביחס x יש 10,000 רשומות, 10 רשומות בדף. בעמודה A ביחס זה יש 200 ערכים שונים זה מזה, והם מפוזרים בהתפלגות אחידה בין הרשומות. רוחב העמודה A מחצית מרוחב רשומה. רוצים לחשב את השאילתה:

`select A from x group by A`

החישוב מבוצע באמצעות מיון. יש 20 דפי חוצץ לרשות החישוב. את הרשומות הנקראות במעבר הראשון אפשר לצמצם לערך של A בלבד, ואף לבטל כפילויות. אפשר כך לקרוא רשומות נוספות

לפני שכותבים מנה ממוינת לדיסק. כותבים מנות ממוינות המכילות רק ערכים של A. כמה פעולות קלט/פלט לכל הפחות יידרשו לחישוב (ללא כתיבת התוצאה הסופית)?

א. 1,000

ב. 2,000

ג. 3,000

ד. 1,010

מטלת מנחה (ממ"ן) 12

הקורס: 20574 - מימוש מערכות בסיסי נתונים

חומר הלימוד למטלה: פרקים 4,5 - עיבוד ואופטימיזציה של שאילתות

משקל המטלה: 5 נקודות

מספר השאלות: 5

מועד אחרון להגשה: 21.5.2021

סמסטר: 2021ב

קיימות שתי חלופות להגשת מטלות:

- שליחת מטלות באמצעות מערכת המטלות המקוונת באתר הבית של הקורס
 - שליחת מטלות באמצעות הדואר או הגשה ישירה למנחה במפגשי ההנחה
- הסבר מפורט ב"נוהל הגשת מטלות מנחה"

שאלה 1 - עיבוד יעיל (16%)

בשאלה 4 בממ"ן 11 מתוארות שתי צורות אחסון לקובץ, בשני סעיפי השאלה. בסעיף הראשון מתואר אחסון במבנה של גיבוב מתרחב, ובסעיף השני מתואר מעבר מצורת אחסון זו לאחסון בעץ B+.

תארו בפירוט איך לבצע את המעבר מצורת האחסון הראשונה לשניה בצורה היעילה ביותר. (חלק מניקוד הסעיף יוקדש ליעילות). הניחו כי יש 250,000 דפי חוצץ לרשות החישובים. חשבו את עלות החישוב שהצעתם במספר גישות לדיסק. זכרו כי בסופו של דבר יש לכתוב כל העץ עם הרשומות שבעלים לדיסק, כאחסון קבוע של הנתונים.

שאלה 2 - סדר פעולות צירוף (24%)

נתונים שלושה יחסים: $r(A,B)$ בן N דפים, $s(B,C)$ בן $10N$ דפים ו- $q(C,D)$ בן $100N$ דפים. רוצים לחשב צירוף טבעי של כל שלושת היחסים ויש $0.4N$ דפי חוצץ לרשות החישובים.

- הציגו את כל סדרי החישוב שהם עצים לינאריים שמאליים. (החלפה בין יחס חיצוני ופנימי באותו צירוף נחשב סדר חישוב שונה).
- נניח שכל התכונות הן מאותו טיפוס נתונים, המפתח של כל אחד מהיחסים מסומן בתבניות היחסים לעיל. העריכו את גודל התוצאה בדפים. לצורך הערכה זו אפשר להניח כי בתוצאת כל צירוף חוזרת העמודה המשותפת פעמיים (כמו בצירוף לפי תנאי כלשהו).
- בהנחה שמשמשים רק באלגוריתם לולאה מקוננת בגושים לחישוב צירופים, מה יהיה סדר הפעולות הזול ביותר מבחינת מספר גישות לדיסק?

שאלה 3 – עיבוד אופרטורים (20%)

נגדיר פעולה בינרית חדשה בין יחסים תואמים (בעלי אותה תבנית), שנסמנה \otimes . הפעולה מחזירה את כל השורות שנמצאות באחד משני היחסים אך לא בשניהם.

- א. (5%) כתבו ביטוי אלגברי שקול לפעולה זו באמצעות פעולות אלגבריות מוכרות.
- ב. (15%) נתונים שני יחסים r, s שתבניתם (A, B) . בכל אחד מהם 100,000 רשומות ב-10.000 דפים. יש 1,000 דפי חוצץ לרשות החישובים. הציעו כיצד לחשב את הפעולה $s \otimes r$ בצורה יעילה ככל האפשר עבור יחסים אלה.

שאלה 4 – מיון חיצוני ויעילות (20%)

- לרשות מיון חיצוני עומדים 200KB של זיכרון פנימי. גודל דף במערכת הוא 4K.
- א. (6%) תארו כיצד למיין במערכת זו קובץ בן 300KB בדרך היעילה ביותר וחשבו את עלותה במספר פעולות קריאה וכתיבה.
- ב. (6%) מהו גודל הקובץ המקסימלי שאפשר למיין בשני מעברים במערכת זו? מהי עלות המיון (ללא כתיבת התוצאה הסופית)?
- ג. (8%) סעיף זה קשה יותר: נתון שעלות מיון קובץ מסוים היתה 10,050 גישות, כשבצעו אותה בצורה באופן היעיל ביותר. מהו היה גודל הקובץ?

שאלה 5 - אלגוריתמי צירוף (20%)

לכל אחת מהטענות הבאות קבעו אם היא נכונה או לא ונמקו היטב.

- א. אם אפשר לחשב צירוף בין שני יחסים r ו- s שגדלם P ו- Q דפים, בהתאמה, בעלות $3(P+Q)$ באמצעות צירוף-מיון-מיזוג, בהכרח אפשר גם לחשב את הצירוף באמצעות צירוף-גיבוב (רגיל) באותה עלות.
- ב. אם אפשר לחשב את הצירוף באמצעות צירוף-גיבוב (רגיל) בעלות $3(P+Q)$, בהכרח אפשר גם לחשב אותו באמצעות צירוף-מיון-מיזוג באותה עלות.
- ג. אם בצירוף-גיבוב נדרש לבצע rehash לאחד היחסים, הכרחי לבצע את אותו rehash גם ליחס השני לצורך השלמת החישוב.
- ד. אם בצירוף-מיון-מיזוג נדרש למיין את אחד היחסים בשלושה מעברים – הכרחי למיין גם את היחס השני בשלושה מעברים לשם השלמת החישוב.

מטלת מנחה (ממ"ן) 13

הקורס: 20574 - מימוש מערכות בסיסי נתונים

חומר הלימוד למטלה: פרקים 6-8 - ניהול תנועות, בו-זמניות, התאוששות.

משקל המטלה: 5

מספר השאלות: 5

מועד אחרון להגשה: 11.6.2021

סמסטר: 2021ב

קיימות שתי חלופות להגשת מטלות:

- שליחת מטלות באמצעות מערכת המטלות המקוונת באתר הבית של הקורס
 - שליחת מטלות באמצעות הדואר או הגשה ישירה למנחה במפגשי ההנחה
- הסבר מפורט ב"נוהל הגשת מטלות מנחה"

מטלה תיאורטית : פרקים 5-7 - ניהול תנועות, בו-זמניות, התאוששות.

שאלה 1 – ניהול תנועות ; בו-זמניות (20%)

להלן נתונים 5 תזמונים של תנועות. לכל אחד מהתזמונים שרטטו את גרף הקדימויות המתאים לו. אם הוא שווה סדרתי בקונפליקט - קבעו את כל התזמונים הסדרתיים השקולים.

1. T1: start; T2: start; T1: r(A); T2: r(A); T1: w(B); T2: w(B)
2. T1: start; T2: start; T2: r(A); T2: commit; T1: r(A); T1: w(A)
3. T1: start; T2: start; T3: start; T1: r(A); T3: w(A); T3: commit; T2: r(B); T2: w(A)
4. T1: start; T2: start; T3: start; T1: r(A); T1: w(A); T2: r(A)
5. T1: start; T2: start; T3: start; T1: r(A); T2: w(A); T3: w(A); T2: r(A)

שאלה 2 – פרוטוקול סידור לפי חותמות זמן (20%)

חמשת התזמונים משאלה 1 מורצים במערכת המפקחת על פעילות בו-זמנית של תנועות באמצעות סידור לפי חותמות זמן, כולל שימוש בכלל הכתיבה של תומס.

הניחו כי אם בתזמון כלשהו לא מופיע אירוע התחייבות בתוך סדרת הפעולות - הכוונה היא שהוא אמור להופיע במקום כלשהו לאחר סיומה.

המתזמן משייך חותמות זמן לתנועות לפי סדר אירועי ההתחלה שלהן.

א. לכל אחד מהתזמונים עליכם לקבוע כיצד תטופל הפעולה האחרונה. עליכם לבחור בכל מקרה אחת משלוש האפשרויות הבאות ולנמק היטב:

- המתזמן יאשר את הפעולה
- המתזמן יתעלם מהפעולה
- הפעולה תתבטל ותגולגל לאחור

ב. אם התזמון הסדרתי השקול שהתקבל שונה ממה שמצאתם בשאלה 1 סעיף ב - הסבירו מדוע.

שאלה 3 – ניהול תנועות; בו-זמניות (20%)

נתונות שתי תנועות:

T1: R(A), R(B), W(A), W(B)

T2: R(C), W(C), R(B), W(B)

א. האם קיים תזמון שווה סדרתי (שאינו סדרתי) של שתי תנועות אלה ששקול לתזמון הסדרתי $T1 \rightarrow T2$? אם כן הציעו תזמון כזה. אם לא – נמקו היטב מדוע לא.

ב. האם קיים תזמון שווה סדרתי (שאינו סדרתי) של שתי תנועות אלה ששקול לתזמון הסדרתי $T2 \rightarrow T1$? אם כן הציעו תזמון כזה. אם לא – נמקו היטב מדוע לא.

ג. הציעו תזמון משולב (לא סדרתי) של שתי התנועות עם נעילות בהתאם לפרוטוקול 2PL שנוצר בו מצב קיפאון, בהנחה שלא מפעילים שיטה כלשהי למניעת קיפאון.

ד. הציעו תזמון משולב (לא סדרתי) של שתי התנועות כך שאם מפעילים עליו את הפרוטוקול של סידור לפי חותמות זמן מתקבל תזמון סדרתי שקול הפוך מסדר התחלת התנועות. כלומר, אם בתזמון T1 מתחילה את הפעולה הראשונה בתזמון, התזמון הסדרתי השקול בסופו של דבר יהיה כזה שבו T2 לפי T1, ולהיפך.

שאלה 4 – התאוששות (20%)

לכל אחת מהטענות הבאות קבעו אם היא נכונה או לא ונמקו היטב. לכל טענה שאינה נכונה – תנו דוגמה שבה הטענה לא נכונה.

א. אם ברגע נתון במהלך פעולה סדירה של המערכת (ללא נפילה), יש הבדל בין התוכן של דף כפי שהוא על הדיסק לבין תוכנו בחוצץ (כלומר הדף "מלוכלך") – מובטח שיש ביומן פעולה המעדכנת לערך העדכני ביותר.

ב. אם ברגע נתון במהלך תהליך התאוששות יש הבדל בין התוכן של דף כפי שהוא על הדיסק לבין תוכנו בחוצץ – מובטח שהתוכן בחוצץ עדכני יותר.

- ג. אם ברגע נתון במהלך פעולה סדירה בין כתיבת begin checkpoint לבין end checkpoint יש הבדל בין התוכן של דף כפי שהוא על הדיסק לבין תוכנו בחוצץ – מובטח שהערך על הדיסק עדכני יותר.
- ד. אם ברגע נתון במהלך תהליך ההתאוששות כל הדפים בחוצץ זהים להעתקיהם על הדיסק – בהכרח רגע זה הוא סיום שלב ה-redo ולפני שלב ה-undo.

שאלה 5 – התאוששות (20%)

4 תנועות מבצעות את הסדרה הבאה של פעולות (מספרי השורות רשומים לנוחיות):

1. T1 writes page 1
2. T2 writes page 2
3. T2 writes page 1
4. T3 writes page 2
5. begin checkpoint
6. T3 writes page 3
7. T4 writes page 4
8. end checkpoint
9. T1 writes page 4
10. T1 commit
11. T2 commit
12. T4 writes page 3

א. רשמו את היומן המתקבל מסדרת פעולות זו (לפי אלגוריתם ההתאוששות הרגיל שנלמד, לא ARIES). הניחו כי כל דף מכיל ערך יחיד; הערכים ההתחלתיים של כל הדפים הם 0; T1 כותבת תמיד את הערך 1 לדף שהיא כותבת, T2 את הערך 2, T3 את הערך 3 ו-T4 את הערך 4.

ב. נניח שהמערכת נפלה לאחר שהיומן הנ"ל נכתב לדיסק. תארו את שלבי ההתאוששות:

- אם יתוספו שורות ליומן במהלך ההתאוששות כתבו אותן.
- ציינו אילו תנועות יגולגלו לאחור בסופו של דבר, ומה יהיו הערכים הסופיים של כל אחד מהדפים.