

## ממן 12

קורס מימוש מערכות בסיסי נתונים, 20574

מגיש טל גלנצמן, 302800354

תאריך 05/04/2021 סמסטר 2021ב

### שאלה 1

#### סעיף א

בדף יש 10 רשומות של האינדקס E1 כלומר, כאשר סורקים את האינדקס נידרשת גישה אחת לדיסק עבור איתור כל 10 רשומות ביחס P. לכן, על מנת לאתר את הרשומה עבורה id=567 נצטרך

$$\lceil \frac{567}{10} \rceil = 57$$

גישות לדיסק. רק נציין, שקריאת הרשומה עצמה דורשת כמובן גישה נוספת.

סה"כ, נדרשות 57 גישות לדיסק על מנת לאתר את הרשומה המבוקשת

#### סעיף ב

רשומה של האינדקס E2 מצביעה על רשומה ראשונה של האינדקס E1 בדף ובכך למעשה מצביעה ל-10 רשומות רצופות של E1. על מנת לאתר את הרשומה המבוקשת ביחס P, נצטרך לגשת לדף ה-57 באינדקס E1 ולאתור דף זה נצטרך לגשת לדף החמישי באינדקס E2 עפ"י החישוב

$$\lceil \frac{57}{10} \rceil = 6$$

משמע, נדרשות 7=6+1 גישות לדיסק על מנת לאתר את הרשומה המבוקשת

#### סעיף ג

עבור השדה salary יש 50 ערכים שונים המתפלגים אחיד - נסיק אם כן כי עבור כל ערך  $s \in salary$  קיימות בממוצע  $\frac{1000}{50} = 20$  רשומות עבורן השדה salary מקבל את הערך s. נסיק אם כן כי 2 הדפים האחרונים של האינדקס E3 מכילים רשומות הצבעה לרשומות של P אותן אנו מחפשים.

אם נבצע סריקה על האינדקס מהערך הקטן לערך הגדול נצטרך למעשה לסרוק את כל האינדקס מהשידורש 100 גישות לדיסק - עבור קריאה של 100 דפים.

גם, באופן מעשי, אומנם התפלגות הערכים אחידה, אבל עדיין קיימת הסתברות לא אפסית שבה כל הרשומות מקבלות אותו ערך, לכן גם כאן, במקרה הגרוע, נצטרך לגשת לכל דפי האינדקס מהשידורש 100 גישות לדיסק.

**הערה** אם ננצל את העובדה שהאינדקס ממורן, ואנו מחפשים את הערך המקסימלי של salary ונבצע סריקה מהסוף להתחלה נצטרך לבצע  $\frac{20}{10} = 2$  גישות לדיסק בממוצע

## סעיף ד

הרמה הראשונה של האינדקס E4 תכלול רשומת מצביע עבור כל ערך אפשרי של salary, כלומר 50 רשומות הצבעה מה שיידרוש  $\frac{50}{10} = 5$  דפים. המצביע הרמה השנייה של האינדקס ימצא בדף החמישי ולכן כדי להגיע למצביע זה יידרשו 5 גישות לדיסק.

כפי שראינו בסעיף ג', עבור כל ערך של salary יהיו בממוצע 20 שורות המקבלות ערך זה, ולכן, ברמה השנייה של האינדקס יידרשו שני דפים להכיל את רשומות המצביעים לרשומות אלו.

בפרט זה נכון עבור הערך המקסימלי של salary, לכן יידרשו 2 גישות לדיסק על מנת למצוא את הרשומות הדרושות כאשר נתון לנו כבר גפי הרמה השנייה.

לסיכום נדרשות 7 גישות לדיסק - 5 גישות לדיסק על מנת למצוא את הדף הראשון בשרשרת הדפים, ועוד שתי גישות לדיסק על מנת לקרוא את שני הדפים הרלוונטיים המכילים את המצביעים לרשומות עצמן

## שאלה 2

### סעיף א

$$N = 10^8, S = 4, P = 4096$$

נפח האחסון של עמודה  $A_1$  הוא  $NS$  בתים, כלומר

$$\lceil \frac{NS}{P} \rceil = 97657$$

דפים.

עבור כל אחת מהעמודות  $A_{2,3,4,5}$  נדרשים  $2NS$  בתים שכן זו מחזיקה גם את ערך העמודה וגם את ערך המפתח, כלומר

$$\lceil \frac{2NS}{P} \rceil = 195313$$

דפים.

נחשב את נפח האחסון של האינדקס

כל עלה בעץ האינדקס מחזיק ערך של  $A_1$  ועוד 4 מצביעים, כלומר 20 בתים ולכן בדף יהיו  $\lfloor \frac{4096}{20} \rfloor = 204$  עלים. נדרשים לנו  $10^8$  עלים ולכן העלים לבדם יתפסו  $\lceil \frac{10^8}{204} \rceil = 490197$  דפים לאחסון העלים.

בנוסף, בהנחה שנרצה שגודל צומת פנימי בעץ יתפוס כמה שיותר מגודל דף, נחשב את ה-n האופטימלי ע"י אי-השוויון

$$(n-1) \cdot 4 + 4n \leq 4096$$

$$8n \leq 4100$$

$$n \leq 512.5$$

אז n אופטימלי הוא 512. על מנת לחשב את תוכן העץ בממוצע, ונניח תפוסה של 85%, ולכן כל צומת ביניים יצביע על  $0.85 \cdot 435 = 370$  ילדים.

נחשב את סך הצמתים, מהשכבה התחונה לעליונה

$$1.100000000 \text{ עלים}$$

$$2. \left\lceil \frac{100000000}{435} \right\rceil = 229886 \text{ צמתי ביניים}$$

$$3. \left\lceil \frac{229886}{435} \right\rceil = 529 \text{ צמתי ביניים}$$

$$4. \left\lceil \frac{529}{435} \right\rceil = 2 \text{ צמתי ביניים}$$

5. שורש

אם כן, בעץ יש  $10^8$  עלים ובנוסף  $229886 + 529 + 2 + 1 = 230418$  צמתי ביניים. מהגדרת  $n$  צומת ביניים יחיד תופס דף שלם ולכן, התפוסה בדפים של כל האינדקס היא  $720615\$ = 230418 + 490197$  דפים.

לבסוף, סה"כ נפח איחסון הקובץ בדפים הוא סכום

• 720615 דפים לאיחסון האינדקס

• 97657 דפים לאיחסון העמודה  $A_1$

• 781252 דפים לאיחסון העמודות  $A_{2,3,4,5}$

כלומר 1599524 דפים.

## סעיף ב

נפרק את השאלתא לשלבי ביצוע

1. מציאת הערך המינימלי של  $A_4$

2. מציאת הערך המינימלי של  $A_2$  אשר גדול ממש מהערך המינימלי של  $A_4$

3. סריקת  $A_2$  החל מהערך שהתקבל בשלב הקודם

תחילה נציין שנפח האיחסון עבור כל עמודה הוא ידוע ולכן ניתן לגשת באופן ישיר לכל אחת מהרשומות - הכוונה לפי סדר, לא לפי ערך.

נסמן  $K = 195313$  מספר הדפים לאחסון עמודות ערכים

**שלב 1** מציאת הערך המינימלי של  $A_4$  מתבצעת פשוט ע"י קריאת הדף הראשון של  $A_4$  ושליפת הערך הראשון, כיוון שהעמודה ממוינת, זהו הערך המינימלי.

תידרש כאן גישה אחת לדיסק.

**שלב 2** כעת נבצע חיפוש בינארי על  $A_2$  למציאת הערך המינימלי של  $A_2$  המינימלי אשר גדול מהערך המינימלי של  $A_4$  אשר קיבלנו בשלב הקודם.

בממוצע תהליך זה ייקח  $\lceil \log_2 K \rceil = 18$  גישות לדיסק

**שלב 3** בהנחת בתפלגות אחידה, נניח שהערך של  $A_2$  שהתקבל בשלב הקודם הוא בחצי העמודה.

בממוצע יידרשו  $\lceil \frac{K}{2} \rceil = 97657$  גישות לדיסק - אבל, בשלב הקודם, לפי שיטת החיפוש, כחצי הגישות בוצעו על דפים שבהם הערך גדול מהערך שחופש ולכן 9 דפים הוטמנו בדפי החוצץ. כאן אנחנו ניגשים לפחות מ-100000 דפים לכן סביר להניח ש9 הדפים האלו עדיין מוטמנים.

נסיק כי יידרשו כאן כ- 97648 גישות לדיסק

**בסה"כ** בממוצע יידרשו כ-  $97648 + 18 + 1 = 97667$  גישות לדיסק

### שאלה 3

#### סעיף א

מפת סיביות עבור brand

	11-111-11	22-222-22	33-333-33	44-444-44
opel	1	1	0	0
peugeot	0	0	1	0
bmw	0	0	0	1

מפת סיביות עבור color

	11-111-11	22-222-22	33-333-33	44-444-44
grey	1	0	0	0
red	0	1	0	0
black	0	0	1	1

#### סעיף ב

- ניקח את הערך המתאים של opel ממפת הסיביות של brand ונסמן  $x = 1100_2$
- ניקח את הערך המתאים של red ממפת הסיביות של color ונסמן  $y = 0100_2$
- נספור את כמות הסיביות הדולקות בערך  $x \& y = 0100_2$ , במקרה הזה 1

### שאלה 4

#### סעיף א

מהנתונים

- גודל סל יהיה כגודל דף, כלומר 4010 בתים
- גודל רשומה הוא 40 בתים

לכן מספר הרשומות שניתן לאחסן בסל הוא לכל היותר

$$\lfloor \frac{4010}{40} \rfloor = 100$$

על מנת לאחסן  $10^8$  רשומות נצטרך  $\lceil \frac{10^8}{100} \rceil = 10^6$  סלים, לכן דרושות לנו 20 סיביות לכל הפחות על מנת לייצג את כל הסלים האפשריים. באותם סימונים בחוברת, ובהנחה שרזולוציית הערכים שלנו היא בבתים, נקבע את הערך של b להיות 24, 3 בתים.

הרי

- גודל מצביע, וכן גם גודל כניסה במדריך 10 בתים
- גודל דף 4010

בדף נכנסות 401 כניסות של המדריך.

כיוון ש-  $b = 24$  יהיו במדריך  $2^{24} = 16777216$  כניסות מה שמצריך מקום של  $\frac{16777216}{401} = 41838.44$  דפים. עדיין לא נעגל כאן לתפוסת דף שלמה שכן יש עוד מידע עבור המדריך ואלי נוכל להשתמש בשארית הדף לאחסון מידע נוסף.

המדריך מחזיר את הרוחב הגלובל וגם את הרוחבים הלוקלים של הסלים התואמים כל כניסה. לכן סהכ יהיו  $1 + 2^{24}$  ערכים כאלה. בדף ניתן לאחסן  $\lfloor \frac{4010}{3} \rfloor = 1336$  ערכים כאלה מה שאומר שאת הערכים האלו נצטרך לאחסן על פני  $\frac{2^{24}+1}{1336} = 12557.79$

בסה"כ עבור המדריך נצטרך אם כן  $54397 = \lceil 54396.23 \rceil = \lceil 41838.44 + 12557.79 \rceil$  דפים.

לכן סה"כ הדפים הדרושים לקובץ, כולל הסלים וטבלת הסלים הוא  $10^6 + 54397 = 1054397$

## סעיף ב

מהנתונים

- רוחב מצביע הוא 10 בתים
- רוחב ערך  $A_2$  הוא 30 בתים

אם כן, כאשר  $n$  פרמטר העץ, בצומת ביניים יהיו לכל היותר  $10n + 30(n-1)$  בתים. בהנחה שגודל צומת פנימי מתפרש על דף,  $n$  אופטימלי כזה נמצא ע"י הא"ש

$$\begin{aligned} 30(n-1) + 10n &\leq 4010 \\ 40n &\leq 4040 \\ n &\leq 101 \end{aligned}$$

נסיק כי  $n=101$ . לכן, בתפוסה של 80%, צומת ביניים יצביע ל- 80 ילדים.

כיוון שגודל עלה כגודל דף - 4010 בתים, וגודל רשומה הוא 40 בתים, נכנסות 100 רשומות בעלה. סה"כ הרשומות ביחס הוא  $10^8$  לכן יהיו באינדקס  $10^6 = \frac{10^8}{10^2}$  עלים.

סך הצמתים בעץ, מרמת העלים תחילה

$$\begin{aligned} 1. & 10^6 \\ 2. & \lceil \frac{10^6}{80} \rceil = 12500 \\ 3. & \lceil \frac{12500}{80} \rceil = 157 \\ 4. & \lceil \frac{157}{80} \rceil = 2 \\ 5. & \lceil \frac{2}{80} \rceil = 1 \end{aligned}$$

כלומר,  $1012660 = 10^6 + 12500 + 157 + 2 + 1$  צמתים. כל צומת תופס דף, ולכן זהו מספר הדפים של האינדקס.

את הרשומות אנו מאחסנים במלואן בעלי האינדקס, לכן זהו גודל הקובץ כולו.