**תשובות לתרגיל בית – הדסים 5**

**חלק א'**

**סעיף א:**

**5. סיבוכיות זמן:**

1. קריאת קובץ Excel והמרתו לקובץ טקסט: קריאת הקובץ והמרת הנתונים לפורמט טקסט נעשית באמצעות הספרייה pandas סיבוכיות הזמן היא O(n) כאשר nהוא מספר השורות בקובץ. כל שורה נקראת ומומרת לפורמט טקסט.
2. פיצול הקובץ לחלקים: הקובץ פוצל לחלקים קטנים בגודל קבוע (100,000 שורות). כל שורה נקראת ומפוצלת בהתאם לגודל החלק. סיבוכיות הזמן היא O(n) כאשר n הוא מספר השורות בקובץ, כיוון שכל שורה נקראת ונכתבת מחדש.
3. ספירת קודי שגיאה בכל חלק: חיפוש קוד שגיאה בכל שורה בעזרת ביטוי רגולרי (re.search). לכל חלק, סיבוכיות הזמן היא O(k) כאשר k הוא מספר השורות בחלק. כלומר, כל שורה בקובץ נבדקת לאחר פיצול הקובץ.
4. חיפוש השגיאות הנפוצות ביותר: לאחר ספירת השגיאות, מבוצע איחוד בין כל הספירות ומיון השגיאות הנפוצות ביותר בעזרת collections.Counter.most\_common סיבוכיות הזמן עבור מיון היא O(mlogN) כאשר mהוא מספר השגיאות השונות וN הוא מספר השגיאות הנפוצות ביותר שנרצה להציג.

סיבוכיות זמן כוללת:

O(n+mlogN) אם לא ההיתי ממירה לקובץ טקסט היינו מקבלים: O(mlogN)

n הוא מספר השורות בקובץ.

m הוא מספר השגיאות השונות בקובץ.

N הוא מספר השגיאות הנפוצות ביותר שנרצה להציג.

**סיבוכיות מקום:**

1. אחסון הנתונים בעת קריאת קובץ ה־ Excel ב־ DataFrame דורש זיכרון בגודל O(n) לאחר שמירת הנתונים לקובץ טקסט ה־ DataFrame נמחק מהזיכרון.
2. עיבוד קובצי הטקסט וה־ chunks נעשה שורה שורה ואינו דורש שמירה של כל הנתונים בזיכרון.
3. ספירת קודי השגיאה נשמרת במבנה collections.Counter ששומר זוגות של קוד שגיאה ומספר מופעים. גודל המבנה הוא O(m) כלומר תלוי במספר קודי השגיאה השונים ולא במספר השורות הכולל.
4. שליפת התוצאה הסופית של N השגיאות הנפוצות ביותר דורשת מקום נוסף בגודל O(N)

סיבוכיות מקום כוללת:

O(m+N)

m הוא מספר השגיאות השונות בקובץ.

N הוא מספר השגיאות הנפוצות ביותר שנרצה להציג.

**סעיף ב:**

**1.א.** בדיקה נוספת: האם יש בקובץ ערכים חסרים.

**3.** כדי לעדכן את הממוצעים השעתיים בזמן אמת כשנתונים זורמים כל הזמן, אפשר לשמור זיכרון זמני למשל buffer שבו אנחנו שומרים את הסכום של הערכים ואת הכמות שלהם עבור כל שעה. כל פעם שמגיע ערך חדש, בודקים לאיזו שעה הוא שייך, מוסיפים את הערך לסכום של אותה שעה, ומעלים ב־1 את הספירה. ככה, בכל רגע, אפשר לחשב את הממוצע של אותה שעה על ידי חלוקת הסכום במספר הערכים שיש. השיטה הזו מאפשרת לחשב ממוצעים שמתעדכנים מיד בכל פעם שמגיע נתון חדש.

**4.** **יתרונות אחסון המידע בפורמט Parquet**

1. יעיל מאוד באחסון – דוחס את הנתונים ובכך חוסך הרבה מקום.
2. מבנה הפורמט הוא לפי עמודות – שומר את הנתונים לפי עמודות, מה שמאפשר קריאה מהירה יותר כי קוראים רק מה שצריך ולא את כל הטבלה.
3. מהיר לביצוע שאילתות – במיוחד כששואלים רק על חלק מהעמודות, זה רץ יותר מהר.
4. תמיכה במבנים מורכבים – תומך גם בטבלאות עם עמודות מקוננות, לא רק טבלאות פשוטות.
5. עובד מעולה עם Big Data – מיועד לעבודה עם כמויות נתונים גדולות במיוחד.

**חלק ב'**

מצורף קובץ בשם tree\_family.db

**חלק ג'**

1. תשובה: 1.

2. a. השידור מהשלט למזגן נעשה בדרך כלל או דרך אינפרא אדום, או דרך גלי רדיו. כשאנחנו לוחצים על כפתור בשלט, הוא שולח סוג של הודעה למזגן בעזרת אחת מהשיטות האלה.

b. רכיבים בצד השלט:

1. חלק שאחראי לשלוח את הפקודות.
2. או משדר אינפרא אדום או משדר גלי רדיו.
3. סוללה.
4. כפתורים כדי לבחור פקודות שרוצים שהמזגן יבצע.

רכיבים בצד המזגן:

1. קולט אינפרא אדום או קולט גלי רדיו.
2. חלק שקורא ומבין את הפקודות מהשלט.
3. מערכת שמבצעת את הפקודות שהתבקשו.

c. אפשרויות לייצוג מקשים ש​ונים בשידור מ​השלט למזגן:

1. לכל כפתור יש קוד משלו – למשל, כפתור "חימום" שולח קוד אחד וכפתור "קירור" קוד אחר.
2. השלט שולח סדרת פקודות והמזגן מזהה בתוך הסדרה איזו פקודה נשלחה לפי הכפתור שנלחץ.
3. שינוי בתדר השידור -כל כפתור יכול לשלוח בתדר שונה וככה המזגן יודע מה ביקשנו ממנו.

**חלק ד'**

תיקיה בשם client\_server