חלק א'

<u>סעיף א' – 5:</u>

נגדיר:

n – מספר השורות בקובץ.

e – מספר קודי השגיאה.

ניתוח זמן ריצה:

- מעבר על כל השורות וחילוץ קוד השגיאה – O(n).
- קיבוץ קודי השגיאה ב-defaultdict – O(e).
- מיון קודי השגיאה בסדר יורד – (eloge).
- החזרת N קודי השגיאה השכיחים ביותר – O(N).

כיוון שברוב המקרים e<n זמן הריצה יהיה O(n).

במקרה הגרוע (שבו n=e) זמן הריצה יהיה O(n + eloge).

ניתוח סיבוכיות מקום:

- conter – מילון שמכיל e נתונים
- defaultdict – מילון שמכיל e נתונים
- result – מערך שמכיל N נתונים

ולכן סיבוכיות המקום היא: O(e + N).

<u>סעיף ב' – 3:</u>

כאשר הנתונים אינם מגיעים מקובץ אלא בזרימה (stream) יש אפשרות לתכנן את הפתרון כך שנוכל לחשב את הממוצעים השעתיים באופן מיידי.

כך נעשה זאת:

במקום לשמור את כל הנתונים שהתקבלו מהזרם, נשמור עבור כל שעה את הנתונים הבאים:

- סכום הערכים שהתקבלו עבור אותה שעה.
- מספר הערכים שהתקבלו לאותה שעה.

באמצעות שני הנתונים האלו ניתן לחשב את הממוצע בקלות ע"י הנוסחה:
סכום הערכים/מספר הערכים = ממוצע.

כך נעשה זאת בפועל:

עבור כל רשומה שמגיעה נחלץ את השעה העגולה מתוך העמודה timestamp. נשתמש במבנה נתונים שישמור עבור כל שעה את סכום הערכים וכמות הערכים וכשנרצה נוכל לחשב את הממוצע של כל שעה בקלות.

סעיף ב' – 4:

פורמט parquet מציע יתרונות חשובים לאחסון נתונים, במיוחד כשמדובר בקבצי מידע גדולים במיוחד:

- **אחסון עמודות**

Parquet מאחסן את הנתונים לפי עמודות ולא שורות, וזה מאפשר:

  - **דחיסה טובה יותר** - כיוון שהנתונים בעמודה הם בדרך כלל מאותו סוג, מה שמפחית את הגודל של הקובץ.
  - **שאילתות יותר מהירות** - אפשר לקרוא רק את העמודות שצריך, כך שמפחיתים את כמות הנתונים שצריך לעבד.

- **דחיסה טובה**

Parquet משתמש בדחיסה מתקדמת, מה שמפחית את גודל הקובץ וגורם לביצועים טובים יותר כשקוראים את הנתונים.

- **תמיכה במבני נתונים מורכבים**

Parquet יכול לשמור נתונים מורכבים, כמו נתונים עם מבנים מקוננים, כך שניתן לעבוד בצורה נוחה עם נתונים מסובכים.

- **גמישות בסכימה**

Parquet מאפשר לשנות את הסכימה בקלות, מבלי לפגוע בנתונים הקיימים.

- **יעילות בקריאה**

אפשר לקרוא רק את העמודות שמעניינות אותנו, מה שגורם לביצועים יותר טובים.

לסיכום parquet הוא פורמט מאוד טוב לאחסון נתונים גדולים כי הוא חוסך מקום, מאפשר קריאה מהירה וקל לעבוד איתו גם עם נתונים מורכבים.