סעיף ג – מה עושים כשהנתונים מגיעים בזרימה (stream)

במקום שהנתונים יהיו שמורים בקובץ, לפעמים הם מגיעים כל הזמן בשידור חי – לדוגמה, מחיישנים או מערכת שמעדכנת כל כמה שניות.

כדי לחשב ממוצע שעתי בזמן אמת, אני צריכה לשמור בזיכרון:

- כמה ערכים הגיעו בכל שעה
- מה הסכום של הערכים האלה

בכל פעם שמגיע ערך חדש, אני מחשבת את השעה שלו (למשל 15:00), מוסיפה את הערך הזה לסכום, ומגדילה את המונה באחד.

בכל רגע אני יכולה לחשב את הממוצע של אותה שעה לפי הנוסחה: **ממוצע = סכום חלקי כמות**

זו שיטה פשוטה, בלי לשמור את כל הערכים, אבל צריך לשים לב לא לשמור יותר מדי שעות אחורה כדי לא לתפוס הרבה זיכרון.

סעיף ד – למה כדאי לפעמים להשתמש בפורמט Parquet במקום CSV

CSV הוא קובץ פשוט שכל אחד יכול לפתוח, אבל לפעמים הוא לא מספיק טוב כשיש הרבה נתונים.
פורמט Parquet הוא קובץ יותר חכם, במיוחד כשיש הרבה שורות או עמודות.

היתרונות של Parquet:

- הקובץ שוקל פחות
- אפשר לקרוא רק חלק מהעמודות במקום את כולן
- שומר על הסוג של כל עמודה (כמו תאריך, מספר, וכו')
- עובד טוב עם כלים של "ביג דאטה" (כמו Spark)
- מהיר יותר לקריאה כשיש הרבה מידע

בדרך כלל נעדיף Parquet כשיש לנו הרבה מידע, או כשאנחנו רוצים לעבוד בצורה מהירה ומסודרת