

פרויקט: מנתח לוגים של תעבורת רשת

שלב 1 - טעינה וסינון בסיסי

רקע

קובץ לוג (Log File) הוא קובץ שמתעד פעילות במערכת. בפרויקט שלנו נעבוד עם לוג של תעבורת רשת - כל שורה מייצגת חבילת מידע שעברה ברשת. המטרה: לבנות מערכת שמזהה תעבורה חשודה ומדווחת עליה.

מבנה הנתונים

קובץ הלוג הוא קובץ CSV (Comma Separated Values) - קובץ טקסט שבו כל שורה מכילה ערכים מופרדים בפסיקים. כל שורה בקובץ בנויה כך:

```
timestamp, source_ip, dest_ip, port, protocol, size
```

שדה	משמעות	דוגמה
timestamp	תאריך ושעה	08:23:45 2024-01-15
source_ip	כתובת IP מקור (מי שלח)	192.168.1.100
dest_ip	כתובת IP יעד (למי נשלח)	10.0.0.5

443	פורט - מספר שמציין סוג שירות	port
HTTPS	פרוטוקול התקשורת	protocol
1024	גודל החבילה בבייטים	size

דוגמה לשורה:

```
2024-01-15 08:23:45,192.168.1.100,10.0.0.5,443,HTTPS,1024
```

הגדרת תעבורה חשודה

במערכת שלנו נגדיר 4 סוגי חשדות:

קוד	סוג חשד	הגדרה
EXTERNAL_IP	IP חיצוני	כתובת מקור שלא מתחילה ב- 192.168 או 10.
SENSITIVE_PORT	פורט רגיש	פורטים: 22 (SSH), 3389, 23 (Telnet), (RDP)
LARGE_PACKET	גודל חריג	חבילה מעל 5000 בייט
NIGHT_ACTIVITY	שעות לילה	פעילות בין 00:00 ל-06:00

שימו לב: המערכת צריכה להיות דינאמית - כלומר קל להוסיף סוגי חשדות נוספים בעתיד בלי לשנות הרבה קוד.

דרישות שלב 1

בשלב זה נתמקד בטעינת הנתונים וסינון בסיסי באמצעות **List Comprehension**.

לכל דרישה יש לכתוב פונקציה נפרדת.

1 טעינת הקובץ

כתבו פונקציה שמקבלת נתיב לקובץ CSV ומחזירה רשימה של רשימות - כל שורה כרשימה של שדות.

2 חילוץ כתובות IP חיצוניות

כתבו פונקציה שמקבלת את הנתונים ומחזירה רשימה של כתובות IP מקור חיצוניות בלבד.

למה חיצוניות? כתובות שמתחילות ב- 192.168 או 10. הן כתובות פנימיות (פרטיות) - הן שייכות לרשת המקומית שלנו. כל כתובת אחרת היא חיצונית - מגיעה מהאינטרנט, ולכן עשויה להיות מקור לאיום.

3 סינון לפי פורט רגיש

כתבו פונקציה שמקבלת את הנתונים ומחזירה רשימה של כל השורות עם פורט רגיש (22, 23 או 3389).

למה פורטים אלה רגישים?

- פורט 22 (SSH) - מאפשר שליטה מרחוק במחשב דרך שורת פקודה
- פורט 23 (Telnet) - פרוטוקול ישן לשליטה מרחוק, לא מוצפן ולכן מסוכן
- פורט 3389 (RDP) - מאפשר שליטה מלאה במחשב Windows מרחוק

4 סינון לפי גודל

כתבו פונקציה שמקבלת את הנתונים ומחזירה רשימה של כל השורות עם חבילות מעל 5000 בייט.

5 תיוג תעבורה

כתבו פונקציה שמקבלת את הנתונים ומחזירה רשימה שבה כל שורה מתויגת: "LARGE" אם הגודל מעל 5000, אחרת "NORMAL".

דוגמה לפלט הסופי (לשלבים הבאים)

כך ייראה הפלט בסוף הפרויקט:

```
{ "45.33.32.156": ["EXTERNAL_IP", "LARGE_PACKET"],  
  "87.120.5.22": ["EXTERNAL_IP", "NIGHT_ACTIVITY",  
  "SENSITIVE_PORT"], "203.0.113.50": ["EXTERNAL_IP"] }
```

כל IP חשוד ממופה לרשימת סוגי החשדות שלו.