תרגיל 4 – MapReduce (חלק ב')

בתרגיל זה נמשיך לתרגל שימוש ב MapReduce.

תורידו מ Moodle קובץ targil4.zip. הוא מכיל שני קבצים של נתונים עבור פרסומים של מוצרים בערוצים שונים.

הקובץ ads.txt מכיל זוגות מורכבות משם מוצר ושם ערוץ בו המוצר מתפרסם, מופרדים ע"י פסיק. הנה חלק הקובץ של ads.txt:

Amazon Kindle,VBE

Amazon Kindle,CPI

Amazon Kindle,LYE

Apple AirPort,QWW

Apple AirPort,HWX

Apple AirPort,QZR

Apple AirPort,YBJ

הקובץ channels.txt מכיל זוגות מורכבות משם ערוץ וכמות הצופים שלו. הנה ההתחלה של channels.txt:

GID,14312

PAY,71628

BRG,71971

YCR,96656

FRV,44468

JEA,50557

IAZ,5080

**השאלה:** (בתרגיל זה יש שאלה אחת בלבד)

עליכם לפתח תהליך או שרשור תהליכי MapReduce ש- יחשב, עבור כל מוצר, את מספר הצופים הפוטנציאליים של הפרסומת שלו (כלומר, סה"כ צופים של כל הערוצים בהם המוצר מתפרסם).

במושגים של בסיסי נתונים, יש לנו שתי טבלאות שקיים ביניהם קשר רבים-ל-יחיד:

ads: product, channel

channels: channel, nViewers

עליכם להשתמש בשיטת MapReduce כדי לבצע השאילתה:

SELECT product, SUM(nViewers)

FROM ads

INNER JOIN channels ON channels.channel = ads.channel

GROUP BY product

שוב נשתמש ב lmr המוכר לכם מהתרגיל הקודם. יש לבחור בערך מתאים עבור -k של lmr. עבור הקבצים הנתונים, צריכים להיות כ 10 mappers. בנוסף, כל תהליך MapReduce ישתמש ב- 2 reducers לכל הפחות.

אפשר (אין חובה) לכתוב תסריט (script) בעזרתו יהיה ניתן להריץ את החישוב כל פעם מחדש.

הערה טכנית: כשקוראים את הקלט שימו לב שבסוף שורה ישנו תו של "שורה חדשה" וצריך לנקות אותו. למשל, ניתן לעבוד עם line[:-1] במקום עם line.

**מקרים לבדיקה (Test Cases):**

הנה ההתחלה והסוף של רשימת המוצרים מסודרת לפי מספר הצופים (השתמשתי בפקודת column כדי ליצור פלט בפורמט הזה). עליכם לקבל מוצרים ומספרים זהים לפלט זה (הפורמט אינו חובה):

Sony Mavica MVC-FD5 2250393

Texas Instruments SR-10 2224058

Nikon F 2218498

Digital Thermometer 2201838

Hamilton Pulsar 2159439

Epson HX-20 2153635

Dyson Air Multiplier 2127679

**…**

Diamond Rio PMP300 355059

Apple iPad 345245

Sony Walkman TPS-L2 335335

Hearing Aids 256392

Samsung Uproar 223154

**רמזים:**

1. ניתן לחשוב על פתירת INNER JOIN פשוט של שתי הטבלאות (כמו שתיארנו בהרצאה) כשלב ראשון.
2. מכיוון שיש רווחים בשמות המוצרים, השתמשתי ב \t כדי להבדיל בין העמודות. ואז הפקודות sort ו column נראות כך:

sort -grk 2 **-t $'\t'** | column -t -s $'\t'

# אופן הגשה:

1. יש להגיש למודל קובץ zip עם כל התוכניות שכתבתם. ניתן ליצור את ה zip כך (כאשר תיקיית הורה היא התיקייה הנוכחית):

zip –r targil4-solution.zip targil4

1. יש למלא את הבוחן.