Yonathan Bettan

302279138

yonibettan@gmail.com

Alon kwart

201025228

Alon.kwart@gmail.com

מבנה הפתרון שבחרנו מכיל 3 פאזות:

פאזה 1:

Map:

$$(lineNum,"prod1,prod2") \rightarrow (prod1,prod2)$$
$$(prod2,prod1)$$

- הקלט הינו מהצורה $(key,val)=(lineNum,line)$ כאשר $line$ מכיל 2 מוצרים המופרדים ע"י התו ",".
- דבר זה נעשה על מנת לבטל את הסדר בין האיברים, כלומר עבור לקוח שקנה מברג וגם פטיש יש צורך לכלול את הפטיש כמוצר נלווה למברג אך גם את המברג כמוצר נלווה לפטיש.

Reduce:

$$(p,[p_1,p_2,...p_n]) \rightarrow (p,p_1\$p_2\$...\$p_n)$$

- מקבלת רשימה של כל המוצרים הנלווים עבור מוצר ספציפי
- מתרגמת רשימה זו למחרוזת עם תו הפרדה "$" בין המוצרים הנלווים.

פאזה 2:

Map:

$$(p,p_1\$p_2\$...\$p_n) \rightarrow (p\$n\$p_1,1)$$
$$(p\$n\$p_2,1)$$
$$...$$
$$(p\$n\$p_n,1)$$

- סופרת את מספר המופעים הכולל של מוצר מסוים ומשרשרת אותו ל- key
- מייצרת צמדים שכל אחד מהם מכיל מופע יחיד של מוצר נלווה לצורך ספירה בשלב ה- reduce

Reduce:

$$(p\$n\$p_k,[1,1,...,1]) \rightarrow (p\$n\$p_k,count\_p_k)$$

- מונה את מספר הפעמים שמוצר הופיע עם מוצר נלווה כלשהו.

פאזה 3:

Map:

$$\left(p\$n\$p_k, count\_p_k\right) \rightarrow \left(p\$\left(1-\frac{count\_p_k}{n}\right)\$p_k, none\right)$$

- מחשבת את הרלוונטיות של כל מוצר נלווה p_k למוצר ספציפי בעל n מופעים ע"י $\frac{count\_p_k}{n}$.
- אינה מייצרת איברים אשר אינם עומדים בתנאי $R_{p_1}(p2) \geq THERSHOLD \equiv 0.025$ כאשר $R_{p_1}(p_2)$ מייצג את הרלוונטיות של מוצר נלווה p_2 להימכר עם מוצר p_1
- המוצרים שיועברו ל- reduce הינם ממוינים לפי שם מוצר ותת מיון לפי דרגת רלוונטיות למוצר הנלווה **בסדר יורד**

- עבור מספרים עשרוניים קטנים או שווים ל-1 מיון לקסיקוגרפי ומיון לפי ערך אמתי של המספר נותנים את אותה תוצאה.
- מכיוון ש- MapReduce ממיין לפי סדר עולה הכנסנו למפתח את המשלים של הרלוונטיות על מנת שהרלוונטיות תמוין בסדר יורד (כי המשלים ממוין בסדר עולה)

Reduce:

$$\left(p\$\left(1-\frac{count\_p_k}{n}\right)\$p_k, none\right) \rightarrow \left(p \qquad p_k \qquad \frac{count\_p_k}{n}, none\right)$$

- מחזיר את הרלוונטיות (לעומת המשלים)
- מסדר את התוצאה בהתאם לפורמט המבוקש
- מתחזק counter סטטי אשר סופר את מספר המוצרים הנלווים עבור מוצר ספציפי ומונע הדפסה של יותר מ- k מוצרים נלווים בהתאם לקלט התכנית

- המוצרים הנלווים כבר ממוינים מהרלוונטי ביותר לפחות רלוונטי ולכן ה- k הראשונים שיודפסו הם ה- k הרלוונטים ביותר

נצילות זיכרון:

- נעשה שימוש במשתנים סטטים עבור שדות מהותיים של מחלקות לניהול יעיל יותר ל ה- memory

בברכת בדיקה מהנה