תרגיל בית 3 - Kafka

הקדמה

כמו שירותים רבים באינטרנט, גם ויקיפדיה בנויה על בסיס קפקא כמערכת עיבוד וניהול אירועים. ויקיפדיה גם מאפשרים לצרוך חיצונית את האירועים מהתורות שלהם, וכך לבנות על בסיסם שירותים וניתוחים אחרים.

התרגיל

בתרגיל זה נבנה מערכת מבוססת Kafka Streams ליצירת סטטיסטיקות מהעדכונים בויקיפדיה.
ויקיפדיה מפרסמים את כל אירועי העדכון במערכת בשירות ה-Event Platform. ראו תיעוד להלן:
https://wikitech.wikimedia.org/wiki/Event_Platform/EventStreams
אנחנו מעוניינים לצרוך את כל האירועים בהם נוצר או נערך דף בויקיפדיה ולהפיק מהם מספר סטטיסטיקות:

1. כמות הפעולות:
    a. כמה דפים נוצרים?
    b. כמה דפים משתנים?
    c. כמה פעולות revert מתבצעות?
2. מי הם המשתמשים הפעילים ביותר?
3. מי הם הדפים הפעילים ביותר?

נרצה גם להיות מסוגלים להפיק את הסטטיסטיקות בחיתוכים הבאים:

1. בחודש האחרון, בשבוע האחרון, ביממה האחרונה ובשעה האחרונה
2. בחלוקה לפי בוטים ומשתמשים
    a. נרצה להפיק גם מדדים נפרדים לבוטים ומשתמשים, וגם את החלק היחסי של כל אחד מהם עבור כמות הפעולות
3. בחלוקה לפי שפות שונות
    a. ניתן להסיק זאת משדה ה-url
    b. הכוונה להפיק מדדים נפרדים לכל שפה, וכן במדדי כמות הפעולות להפיק מדד של החלק היחסי של כל שפה בעדכונים

על מנת לבצע את התרגיל יש:

1. להרים קלאסטר קפקא
2. להגדיר נושאים לפי בחירתכם
3. לייצר producer שיצרוך את האירועים מויקיפדיה ויזין אותם לתוך הנושאים
4. להרים קלאסטר של קפקא סטרימס
5. לייצר stream workers שייצרו את המדדים שהוגדרו למעלה

כנקודת התחלה כדאי להסתכל על הפוסט הזה, ניתן להסתמך על הקוד שמופיע בו:
https://towardsdatascience.com/introduction-to-apache-kafka-with-wikipedias-eventstreams-service-d06d4628e8d9

למשתמשים בווינדווס:

1. השתמשו במדריך הזה כדי להפעיל קפקא מקומית:
    https://www.loginradius.com/blog/engineering/quick-kafka-installation/

2. הדגל –zookeeper כבר אינו בשימוש, הוראות להחלפתו:
https://stackoverflow.com/questions/53428903/zookeeper-is-not-a-recognized-option-when-executing-kafka-console-consumer-sh

הגשת התרגיל

בהגשת התרגיל יש לכלול:

1. קובץ zip עם כל הקוד שכתבתם, בפרט:
   a. הקוד של ה producer
   b. הקוד של ה streams worker
   c. סקריפט המייצר את הקלאסטר ומקנפג את ה topics
   d. כלי כלשהו שמאפשר לקרוא את מצב הסטטיסטיקות הנוכחי ולהדפיס אותם למסך
2. קובץ pdf עם הסבר על התרגיל שיכלול:
   a. הסבר ב-high level על הגישה לחישוב הסטטיסטיקות
   b. הוראות להרצת התרגיל והפקת המדדים למערכת הפעלה שבה השתמשתם.
      i. אין לכלול צעדים מורכבים, יש לעטוף הכל בסקריפטים.
   c. נניח כי אנחנו מעוניינים להריץ את המערכת בסביבת production אמיתית. כיצד תבחרו ערכים מתאימים לדגלים partitions ו-replication-factor שסיפקתם בעת יצירת ה-topic?
      i. מה השיקולים לבחירת הערכים, יתרונות וחסרונות?
   d. הציעו גישה offline ליצירת מדדים כאלה ותארו אותה בקווים כלליים.
      i. מה היתרונות והחסרונות לגישת offline ביחס לגישה שנקטנו בתרגיל?

הערות ועצות

- שימו לב שקצב האירועים בויקיפדיה יכול להיות מאוד גבוה וגם אינו בשליטתכם. הכניסו לתוכנה דגלים שיאפשרו לכם לשלוט בכך. מומלץ להוסיף לפחות:
  - מספר אירועים שאחריהם ה-producer יכבה
  - מגבלה על מספר האירועים שה-producer מייצר בשניה
- ישנו אתר שמספק סטטיסטיקות דומות וניתן להשתמש בו להשראה:https://codepen.io/Krinkle/pen/BwEKgW?editors=1010
- ניתן להפיק אירועים מהתור החי של ויקיפדיה על מנת להבין את התוכן הצפוי: https://stream.wikimedia.org/v2/ui/#/
- יש גם תיעוד של ה-:API
https://stream.wikimedia.org/?doc
-