הגשה 3

Shir Sneh 312177231

Shahar Tefler 213991029

1) הסבר high level על חישוב הסטטיסטיקות:
יצרנו producer בפייתון לפי ה-tutorial. השתמשנו בחלוקת streams מויקיפדיה ל 2 חלקים: עריכות מrecentChange ויצירות דפים מpageCreate. . כמו כן השתמשנו בכמה מחלקות שמתארות את הפעולות שקרו לצורך חישוב הממדים הנדרשים, והשתמשנו בפייתון בסיסי כדי להמיר את המחלקות לJson.

ה-producer:
הגדרנו מחלקה wikiStatsTopoligy שבה מתבצעים חישובי הסטטיסטיקות. כמו כן יש לנו מתודות שאחראיות לחישוב הסטטיסטיקות המבוקשות:
countPages מחזירה את מספרים הדפים החדשים/המעודכנים לפי המבוקש.
mostActiveUsers מחזירה לנו את 5 המשתמשים הפעילים ביותר.
mostActivePages מחזירה לנו את 5 הדפים הפעילים ביותר.

יש את הסטרים הראשי של ה-events. עבור ספירות של פעולות (יצירה/עריכה) יצרנו פונקציה שמקבלת סטרים וסופרת בו את מספר ה-events. ואז בעצם עבור כל חיתוך שנרצה, רק צריך להגדיר את הסטרים כך שהאירועים יהיו מסווגים לפי המפתח שלהם.

עבור חיתוך לפי זמנים, יצרנו מפה שמחזיקה את ה-Duration לכל חלון, ושלחנו את הסטרים אחרי שעשינו לו filter רק עבור האירועים בכל חלון זמן (שעה אחרונה, שבוע אחרון וכו').

עבור חיתוך לפי סוג משתמש (בוט או לא) מיפינו את האירועים למפתחות של סוג משתמש (בוט ומשתמש רגיל). ובאופן דומה, עבור שפה המפתח היה השפה עצמה.

עבור ה-top 5, יצרנו אגרגציה שכוללת מחלקה ששומרת אצלה כול הזמן את ה-top5.

2) הוראות להרצת התרגיל:
בבקשה תקרא את הreadMe המצורף בקישור הבא:
https://github.com/shahar0809/WikipediaStats

3) כאשר אנו יוצרים topic אנחנו צריכים להגדיר את כמות המחיצות (partition) ואת פקטור השכפול (replication factor). יש חשיבות רבה לבחירה נכונה של ערכים אלו כי הם משפיעים על יעילות המערכת.
ישנם מספר שיקולים לבחירת מספר המחיצות הנדרש (ערך ה-partition):
כמות גדולה יותר של מחיצות תורמת למקביליות ולתפוקה של התוכנית. בנוסף כמות גדולה יותר של מחיצות תאפשר לנו להריץ יותר consumers בקבוצה ותאפשר לנו לעבוד עם יותר broker – ים. שיקולים אלו יתרמו ליעילות התוכנית אך יש לשים לב גם לחסרונות בבחירת כמות גדולה של מחיצות – ה-zookeeper ממנה לכל מחיצה

partition leader ולכן אם יהיו לנו הרבה מחיצות נצטרך גם הרבה partition leader. בנוסף יותר קבצים יצטרכו להיות פתוחים במקביל ע"י kafka.
לכן בבחירת כמות המחיצות נשים לב לדברים הבאים:

- גודל הcluster – אם יש לנו cluster קטן המכיל פחות מ-6 brokers, נגדיר 3x מחיצות כאשר x הוא מספר הברוקרים.לפי ההסבר המופיע בקישור המצורף למטה צריך פי 3 מחיצות ממספר הברוקרים מכיוון שאם יתווספו ברוקרים נוספים עם הזמן יהיו לנו מחיצות בשבילם. אם יש לנו cluster גדול, נגדיר 2x מחיצות כאשר x הוא מספר הברוקרים.
- מספר הconsumers - נשים לב שיש לנו מספיק מחיצות ל consumers השונים.
- תפוקת הproducer- נשים לב שעלינו להתחשב בתפוקת הproducer – אם התפוקה גדולה יש להגדיר את כמות המחיצות פי 3 ממספר הברוקרים.

בנוסף עלינו לבחור את הreplication factor. לצורך שמירה על אמינות ועמידות התוכנית יש צורך לאפשר שכפול של מחיצות. ההמלצה היא להגדיר את ה replication factor ל-3 מכיוון שזה מקנה איזון מיטבי בין יעילות התוכנית לבין עמידותה בפני שגיאות. הסבר נוסף הוא שלרוב שירותי ענן שונים מספקים 3 data centers.
היתרון ב-replication factor גדול יותר הוא שכושר ההתאוששות (ה resilience) של התוכנית תהיה טובה יותר. אם הreplication factor הוא N אז לכל היותר N-1 ברוקרים יכולים להיכשל בלי לפגוע בavailability של התוכנית.
החסרונות הם:

- Higher latency בעבודה עם הproducers בגלל הזמן שיקח למידע להשתכפל לכל הreplicated brokers.
- יותר מקום בזכרון ידרש.

מקור עזר לשאלה 3

4) נוכל להשיג את המדדים הללו בגישה offline ע"י הורדת הdatabase של ויקיפדיה (ספציפית את התיקיות של Wikimedia statistics) וביצוע שאילתות עליו.