

איפיון שלב 3 – Refactoring יחד עם ETL מוחזרי insertion_time מבווע

3.1 מטרה

בשלב זה יש להרחיב את המערכת כך שתתמוך בתהליכי ETL מוחזרי ממנגו למסד נתונים רלציוני (MySQL), על בסיס שדה זמן הוסף (`insertion_time`) שיתווסף בעת קליטת ההודעה. השלב כולל:

- עדכון קוד קיימ (Refactoring)
 - כולל הוספה שדה מטא-דאטה בעת כתיבה למונגו
- פיתוח שירות חדש (ETL Service)
- שיטוח JSON מקובן
- טעינת נתונים ל-MySQL הכוללת הגדרה של מודול רלציוני

3.2 לשירות ה-Consumer Refactoring הקיים

יש לעדכן את השירות ה-Consumer הקיים (שלב 2) כך שלפני שמירת המסמך במנגו:

1. יתווסף שדה חדש בשם `insertion_time`.
2. השדה יוכל את זמן ההכנסה בפועל למסד הנתונים.
3. הערך חיב להיווצר לצד השירות Consumer (לא מגיע מה-Producer.).
4. השדה ישמר כחלק מה-document docuemnt במנגו.

הערה:

זהו שינוי בקוד קיימ. אין ליזור Consumer חדש — יש לבצע Refactoring לקוד הקיים.

3.3 שירות ETL חדש (Service שלישי)

יש ליצור שירות חדש נוסף ב-Docker Compose.

תפקיד השירות:

- לקרוא נתונים ממנגו.
- לבחור רשומות על בסיס `.insertion_time`.
- לשטח את ה-JSON המפון.
- להזין את הנתונים למסד רלציוני (MySQL).

3.4 אופן הפעולה של שירות ה-ETL

השירות ירעץ בלולאה מחזורת.

בכל מחזור:

1. לקרוא ממנגו מקבץ (batch) של רשומות עדכניות.
2. הבחירה תבסס על שדה `.insertion_time`.
3. יש לעבד רק רשומות שטרם עובדו בעבר.
4. לאחר עיבוד מוצלח — יש לעדכן את נקודת ההתקדמות של התהיליך.

אין לבצע טעינה מלאה של כל הדטה שנמצא במונגו בכל ריצה, כמו **רעיון השירות למניע טעינה כפולה של רשומות** למסד הנתונים MySQL.

מנגנו שמירת מצב ההתקדמות (state) נתון לבחירתכם.

3.5 שיטוח הנתונים

עבור כל רשומה ממונגה:

1. לזהות את הישות הראשית.
2. לזהות את המערך המקורי.
3. ליצור רשומה עבור הישות הראשית.
4. ליצור רשומות עבור כל איבר במערך.
5. להבטיח קשר תקין בין הרשומות של המערך לרשותה של הישות הראשית במסד הרלציווי.

שימוש לב: יש לגזר את המבנה הרלציווי מຕוך מבנה ה-JSON (כלומר קשר של PK-FK).

נדרש:

- שתי טבלאות.
- Primary Keys
- Foreign Key
- Data Types מתאימים.
- מנגן למניעת רשומות כפולות.

3.6 ארכיטקטורה בשלב זה

INCLUDE Docker Compose

- api (Producer)
- consumer (Kafka → Mongo, **כלל insertion_time**)
 - etl-service (Mongo → MySQL)
 - kafka
 - mongodb
 - mysql

3.7 קритריוני הצלחה

השלב ייחשב תקין אם:

- ה-Consumer טוען את המסמכים במנגנון עם `.insertion_time`.
- שירות ה-ETL רץ בצורה מחזורית.
- לא מתבצעת טעינה מלאה בכל ריצה.
- לא נוצרות כפליות במסד המידע.
- ניתן להריץ שאלות JOIN תקינות על הנתונים.
- ניתן לאמת שכל הרשומות שהוזרמו למנהגו עובדו והוזנו למסד הרלצוני.
- ניתן להוסיף נתונים חדשים למערכת ותהליך ETL יקלוט ויתען רק אותם.

לא ניתן מספר רשומות צפוי מראש, עליוכם לבצע ולידציה עצמאית של **שלמות הנתונים בין ה-source ל-target**.