

איפיון שלב 3 – Refactoring יחד עם ETL מחזורי מבוסס `insertion_time`

3.1 מטרה

בשלב זה יש להרחיב את המערכת כך שתתמוך בתהליך ETL מחזורי ממונע למסד נתונים רלציוני (MySQL), על בסיס שדה זמן הוספה (`insertion_time`) שיתווסף בעת קליטת ההודעה.

השלב כולל:

- עדכון קוד קיים (Refactoring)
 - כולל הוספת שדה מטא-דאטה בעת כתיבה למונע
- פיתוח שירות חדש (ETL Service)
- שיטוח JSON מקונן
- טעינת נתונים ל-MySQL הכוללת הגדרה של מודל רלציוני

3.2 Refactoring לשירות ה-Consumer הקיים

יש לעדכן את ה-Consumer הקיים (משלב 2) כך שלפני שמירת המסמך במונע:

1. יתווסף שדה חדש בשם `insertion_time`.
2. השדה יכיל את זמן ההכנסה בפועל למסד הנתונים.
3. הערך חייב להיווצר בצד ה-Consumer (לא מגיע מה-Producer).
4. השדה יישמר כחלק מה-document במונע.

הערה:

זהו שינוי בקוד קיים. אין ליצור 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 ממונוגו.
- לבחור רשומות על בסיס `insertion_time`.
- לשטח את ה-JSON המקוון.
- להזין את הנתונים למסד רלציוני (MySQL).

3.4 אופן הפעולה של שירות ה-ETL

השירות ירוץ בלולאה מחזורית.

בכל מחזור:

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

אין לבצע טעינה מלאה של כל הדאטה שנמצא במונוגו בכל ריצה, כלומר **על השירות למנוע טעינה כפולה של רשומות** למסד הנתונים MySQL.

מנגנון שמירת מצב ההתקדמות (state) נתון לבחירתכם.

3.5 שיטוח הנתונים

עבור כל רשומה ממונָגו:

1. לזהות את הישות הראשית.
 2. לזהות את המערך המקוּן.
 3. ליצור רשומה עבור הישות הראשית.
 4. ליצור רשומות עבור כל איבר במערך.
 5. להבטיח קשר תקין בין הרשומות של המערך לרשומה של הישות הראשית במסד הרלציוני.
- שימו לב:** יש לגזור את המבנה הרלציוני מתוך מבנה ה-JSON (כלומר קשר של PK-FK).

נדרש:

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

3.6 ארכיטקטורה בשלב זה

Docker Compose יכלול:

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

3.7 קריטריוני הצלחה

השלב ייחשב תקין אם:

- ה-Consumer טוען את המסמכים במונוגו עם `insertion_time`.
- שירות ה-ETL רץ בצורה מחזורית.
- לא מתבצעת טעינה מלאה בכל ריצה.
- לא נוצרות כפילויות במסד היעד.
- ניתן להריץ שאילתות JOIN תקינות על הנתונים.
- ניתן לאמת שכל הרשומות שהוזרמו למונוגו עובדו והוזנו למסד הרלציוני.
- ניתן להוסיף נתונים חדשים למערכת ותהליך ה-ETL יקלוט ויטעין רק אותם.

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