ניהול מידע מבוזר - פרויקט חלק ב'

מגישים:

ירין בן שטרית 206230021

רועי פאפו 316327451

תומר פרץ 318295029

חלק 1 - Extract, Transform, Load:

הדאטה שמגיע מהשרת נשמר עם הסכמה:

```
StructType([StructField('StationId', StringType(), False),
            StructField('Date', IntegerType(), False),
            StructField('Variable', StringType(), False),
            StructField('Value', IntegerType(), False),
            StructField('M_Flag', StringType(), True),
            StructField('Q_Flag', StringType(), True),
            StructField('S_Flag', StringType(), True),
            StructField('ObsTime', StringType(), True)])
```

בנוסף, על ה- DataFrame מופעלות הטרנספורמציות הבאות:

1. הסרת העמודות M_Flag, S_Flag, ObsTime: החלטנו שעמודות אלו אינן מוסיפות ערך מוסף לצורך הניתוח אותו אנו מבצעים בשלבים הבאים.
2. הסרת הרשומות בהן הערך בעמודה Q_Flag אינו Null: זאת על מנת להבטיח שכל הרשומות בהן נשתמש לצורך ניתוח ולמידה יכילו ערכים שהתקבלו ממדידות תקינות בלבד.
3. הסרת העמודה Q_Flag: לאחר הסרת הרשומות הבעייתיות בשלב הקודם, אין צורך בעמודה זו, שכן כל הרשומות שנותרו יכילו ערך Q_Flag = None.
4. הוספת עמודת fullDate שמתרגמת את עמודת Date הנתונה לפורמט תאריך בעזרת הפונקציה to_date() של Spark, זאת כדי לשלוף בצורה נוחה את התאריך של כל רשומה.

האבחנות עליהן עבדנו בחלק 2 דרשו לבצע ניתוח של הנתונים לפי מדינות ויבשות. לשם כך, יצרנו קובץ CSV בשם GeoData בעל 3 עמודות: עמודת FIPS המכילה את קוד ה- FIPS של המדינה, עמודת Continent המכילה את היבשת אליה שייכת המדינה ועמודת CountryName המכילה את שם המדינה. בחרנו להשתמש בקובץ זה כדי לאתר ולשלוף את הנתונים הרלוונטיים לאבחנות שלנו בקלות.

שתי פונקציות עזר בהן השתמשנו לצורך תקשורת עם השרת הם:

1. writeToServer(): כותבת spark.DataFrame לטבלה קיימת בשרת.
2. read_df_from_sql_server(): קורא טבלה קיימת בשרת ומחזיר spark.DataFrame עם המידע.

בחרנו להשתמש בסטרימינג דינאמי המורכב מ- readStream ו- writeStream. המידע מגיע ב- Batch'ים בגודל של 5,000,000 רשומות, ועבור כל אחד מתבצעת הטרנספורמציה לעיל. לאחר בדיקה אמפירית גילינו שלאחר פרק זמן של שעתיים (7200 שניות) מתקבלות מעל 100,000,000 רשומות כנדרש.

ביצענו פרגמנטציה של שרת ה-Kafka כך שכל אבחנה תתבסס על חלק מעמודות הטבלה המקורית בפרגמנט יחיד (לפי המונחים שנלמדו בקורס נוכל לומר שביצענו חלוקות אופקית ואנכיות לסירוגין).