תרגיל מסכם

**תקראו הכל לפני שאתם מתחילים.**

**תודה.**

# עלילה

* אתם חיילים באצ"ל אשר מנסים להאזין למ.ק של האויב (הבריטים, כי זה באנגלית).
* בידיכם חיישני ל"א **חכמים** היכולים לפרסר את תקשורת האויב לכתב.
* על מה האויב מדבר? ביקורות לאוכל, סרטים והזמנות מאמזון.
* עלינו לקבל תשדורת מהחיישן ולחזות האם מדובר בביקורת **חיובית** או **שלילית**.
* כאמור, החיישנים חכמים ועל כן התשדורות עוברות בקלאסטר של **Apache Kafka**.
* מאידך, החיישנים לא מושלמים, ולא בכל המקרים הם יצליחו לפרסר את ההודעה.

# דרישות

* pipeline אשר מעביר את המידע מהחיישנים וחוזה האם מדובר בביקורת חיובית או שלילית.
* ניקוי של המידע רק לשדות הרלוונטיים.
* שמירה של המידע בשלבי העיבוד השונים על מנת אנליזה עתידית.
  + מידע גולמי - בדיוק כפי שקיבלנו מהחיישנים.
  + מידע מעובד - אחרי ניקוי שדות ושמירת שפה אחידה.
  + תחזיות - הוספה של שדה החיזוי להודעה.
  + הערה: המידע צריך להישמר בפורמט parquet.
* המידע בקפקא יעבור בפורמט Avro בעזרת Scheme Registry.
* על ה-pipeline עצמו להיות:
  + fault-tolerant
  + near real time
* התחזיות יעשו ע"י מודל ML שאומן על מידע דומה בעבר.

# ארכיטקטורה

* ראשית, מכיוון שאנחנו בונים pipeline די מסובך שצריך להיות רובסטי, נשתמש ב-**NiFi**. **🎉**
* נצרוך מה-topic ששמה "speech-sensors" את ההודעות.
* נעביר את ההודעות לשפה אחידה - בחלק מההודעות שדה ה-timestamp הוא בפורמט unix timestamp ובחלק הוא בפורמט קריא לבני אדם - תעבירו בבקשה שהכל יהיה קריא לבני אדם.
* אחרי העיבוד לעיל נשים את ההודעות ב-S3 ב-bucket ששמו "*speech-sensors-RAW*" ובנוסיף ל-topic בקפקא אשר נושאת את אותו שם.
* נאזין ב-NiFi ל-topic שהרגע יצרנו ונבצע את השלבים הבאים:
  + אם שדה "parsing\_status" שלילי (כלומר, החיישן לא הצליח לפרסר), נפלטר את ההודעות החוצה.
  + נוריד את השדות הלא רלוונטיים למשימה: "reception" ו- "parsing\_status".
  + ושוב, נשים ב-S3 ב-bucket ששמו *speech-sensors-ORG* וכנ"ל לקפקא.
* מה-topic שיצרנו עכשיו, נשתמש ב-**Spark Streaming** ובמודל המאומן שלנו בשביל לחזות האם "parsed\_sentence" הוא ביקורת חיובית או שלילית.
  + פרטים על המודל בהמשך.
  + את התחזיות מ-spark streaming נעביר ל-topic חדש בקפקא "speech-sensors-PREDICTIONS".
* מה-topic שיצרנו, בעזרת NiFi נשמור את המידע ב-S3 ב-bucket בעל אותו שם.

## שאלה למחשבה (נדבר עליה בסוף היום): הייתם בונים את הארכיטקטורה באופן שונה? מה הייתם משנים / משפרים?

# המודל המאומן

נאמן את המודל בעזרת pyspark וספריית ML של ספארק, על מנת שנוכל לתרגל עיבוד של dataset לחילוץ פיצ'רים.  
  
קיימת מחברת jupyter שחצי-מדריכה איך לבצע את תהליך האימון.

לחלק הזה תוקצב שעה. אם לא סיימתם, ניתן להשתמש במודל מאומן שהכינותי מראש.

# סיימתם?

עזרו לחבריכם :)