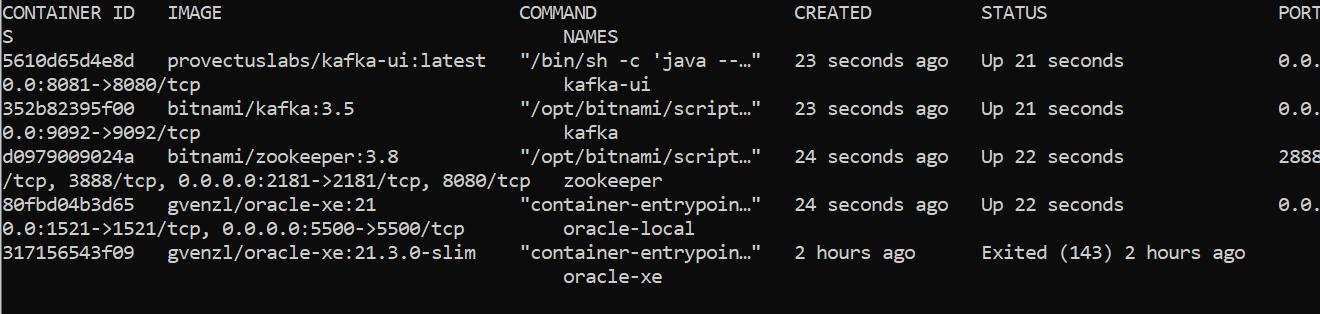
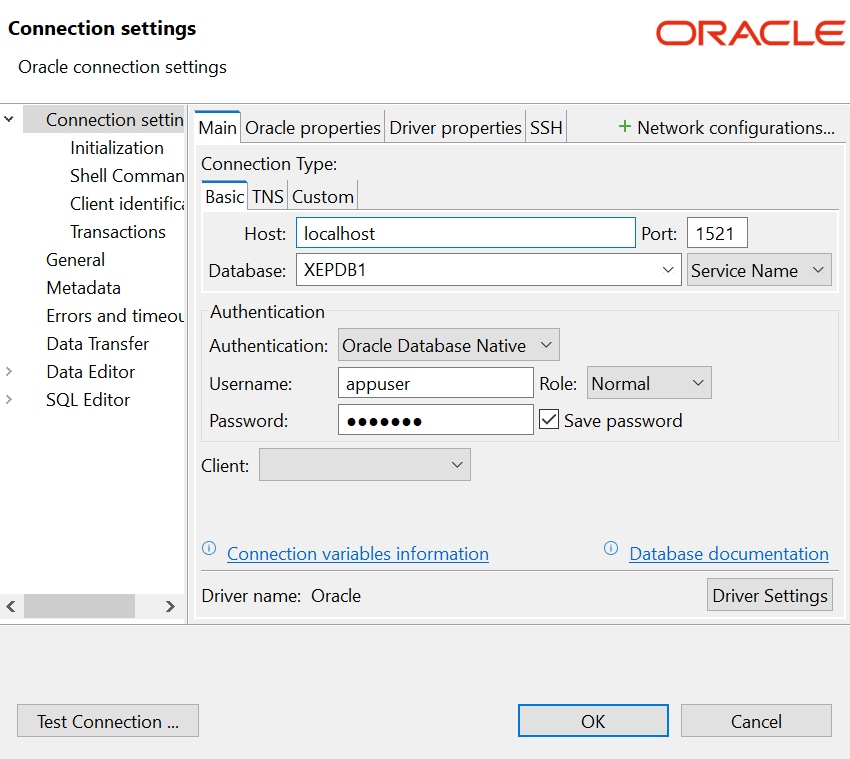
**דרישות פרויקט ספארק:**  
חלק א-   
  
**התקנה:**  
יש להריץ דרך ה cmd בתקייה שבה נמצא ה docker את הפקודות הבאות:

1. יש להריץ את הפקודה - docker-compose up -d
2. יש להריץ את הפקודה - docker ps -a ולבדוק שיש את ה output הבא: 
3. יש להתחבר לאורקל דרך ה cmd או דרך ui – אני ממליצה על dbeaver  
   פרטי החיבור:  
   
4. יש לבדוק את ה kafka-ui או פקודות הרצה ב cmd-   
   docker exec -it kafka bash   
   ואחר כך - kafka-topics.sh --list --bootstrap-server localhost:9092  
   ולראות שאכן קיים הטופיק: output\_results  
   http://localhost:8081/  
     
   A screenshot of a computer

   AI-generated content may be incorrect.

**בסיום שלב זה אתם מוכנים להתחיל את הפרויקט.**

**Input:**

1. קריאת ה data מקובץ parquet מקומי
2. יש להסתכל על תוכן הקובץ – הקובץ מתאר קבוצת חומרים מסוימת של אנשים, רכבים קניונים והתכונות שלהם
3. יש לקרוא את הקובץ ולהפריד את הרשומות לפי resource\_id ולשמור בקבצי parquet נפרדים.
4. יש לעבור לקרוא כל קובץ
5. יש לפרמט אותו לעמודות ב data frame של ספארק עם סכמה שמורכבת מהשדות של כל resource\_id שנמצאים בטבלהprops\_data לפי הסוג הפרימיטיבי. יש לשמור ב broadcast את הטבלה -לסנן כל פעם עבור אותו resource כדי שנוכל לבצע את הלוגיקות שיש לפי field\_special\_type

בסיום שלב זה- אמור להיות DataSet של ספארק עם העמודות של כל קובץ resource לפי הסכמה שנבנה באופן דינמי.   
לבדיקת שלב שהתבצע נכון- יש לעשות show(false) על הDS ולראות את התוצאה. לשמור צילום מסך של הפלט.

**Logics**:

לאחר קריאת ה data. עבור resource אנשים מ US יש לבצע את הלוגיקות הבאות:

1. יש לעגל את הגובה height למעלה
2. יש להוסיף עמודה לאזרחים שהם לא ישראלים שנקראת visa\_approve אם מטרת הביקור שלהם היא “work”
3. יש לסנן אנשים שהם ביזבזו פחות מ 10 שקלים
4. יש להעשיר את הטבלה בקידומות של כל מדינה לפי join עם העמודה - citizenship\_loccaly ועם העמודה - country\_code ולהוסיף את העמודות הבאות ל DF - phone\_prefix, number\_of\_cities, full\_name
5. יש לבצע קנוניזציה על הטלפון – להוריד קידומת 05 ולהחליף בקידומת של phone\_prefix אם העמודה של הטלפון ב DS לא מכילה כולה מספרים יש לשים null. את הכל יש לשמור בעמודה חדשה שנקראת [שם העמודה המקורית]\_formatted

**Output:**יש להפוך ל json שטוח את כל העמודות, יש לבחור את העמודות שלא בוצעו עליהם לוגיקות ואת העמודות לאחר הלוגיקות formatter ולהחליף את השם של העמודות המקוריות ב [שם העמודה ]\_ orig  
יש לשלוח את ההודעות לטופיק – output\_results  
יש לבצע צילום מסך של ההודעה הסופית בטופיק.