**שם הפרויוקט Real Time Flights Information :**

**מטרת הפרוייקט ותיאור כללי**

מטרת הפרוייקט היא לקבל מידע בזמן אמת, לפי פרמטרים שונים, כמו שדי תעופה של המוצאה והיעד, חברות תעופה, ערים וכדו', ולייצר מהם לוח טיסות שמתעדכן בזמן אמת.

**API**

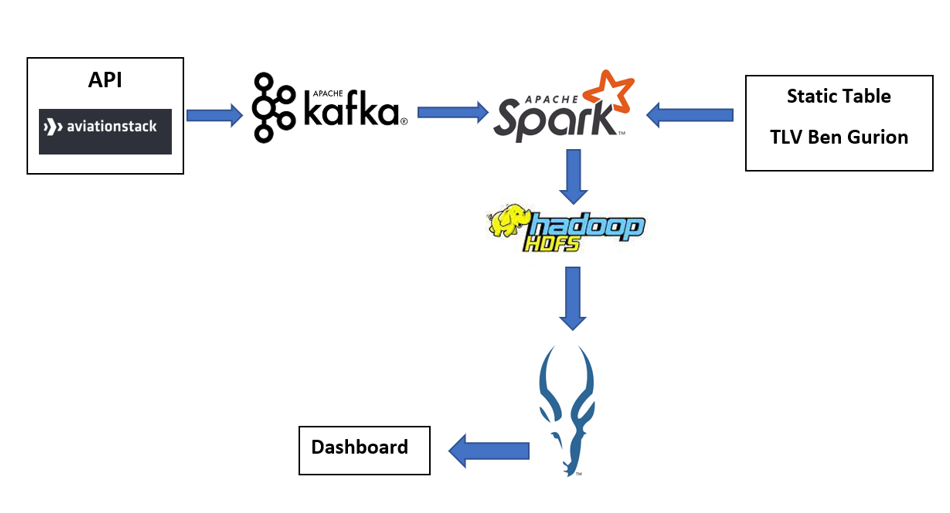
לקחנו מידע מאתר:

**https://aviationstack.com/**

המידע מגיע בJSON מכונון, היינו צריכים לשטח אותו.

צירפנו טבלה סטטית עם מידע על שדי התעופה ועשינו איתה JOIN באימפלה.

**ארכטקטורה**



**מרכיבי הPipeline**

**Kafka Produscer**

**Kafka\_consumer\_to\_file**

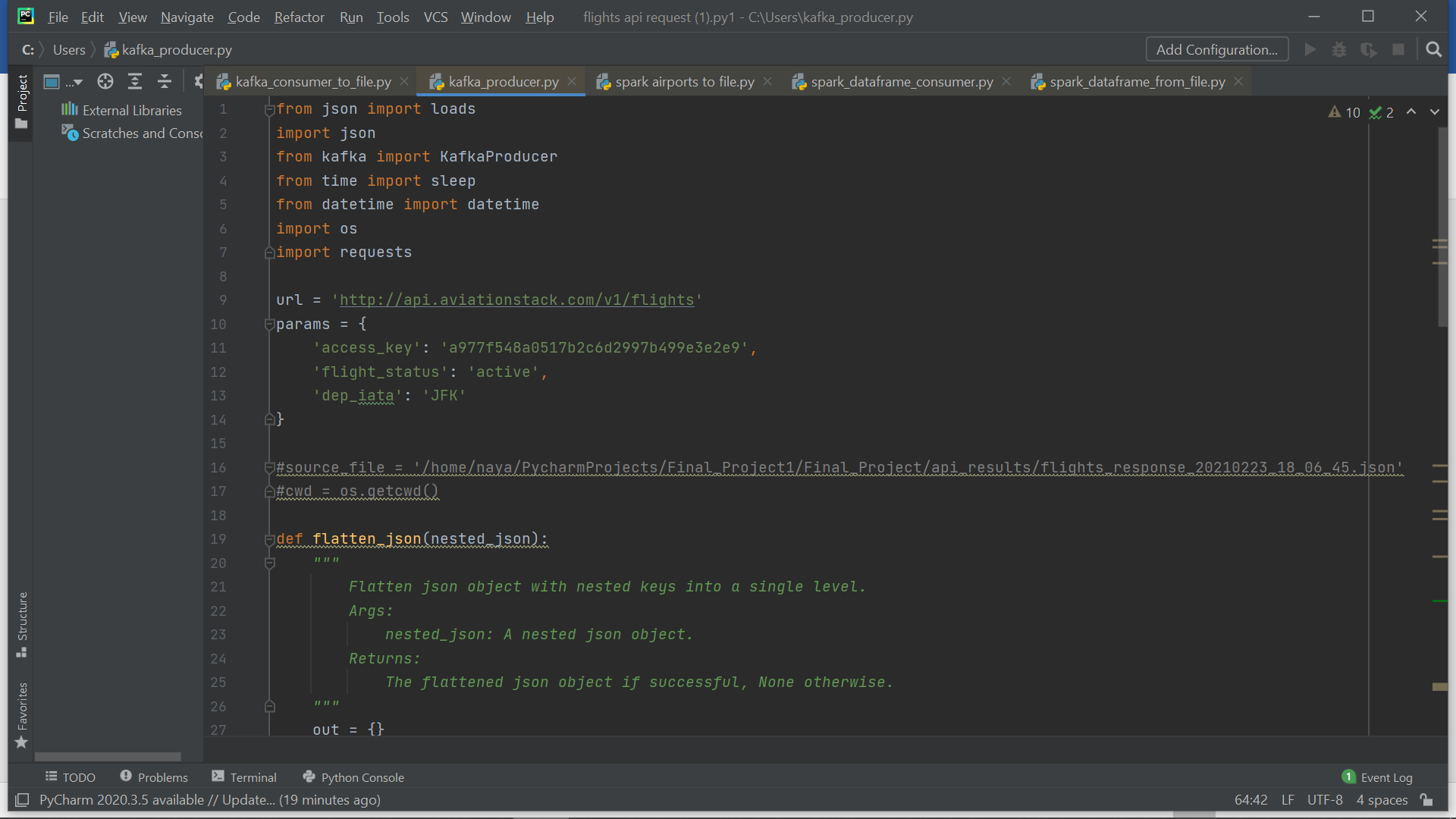
**Spark airoports to file**

**Spark\_dataframe\_consumer**

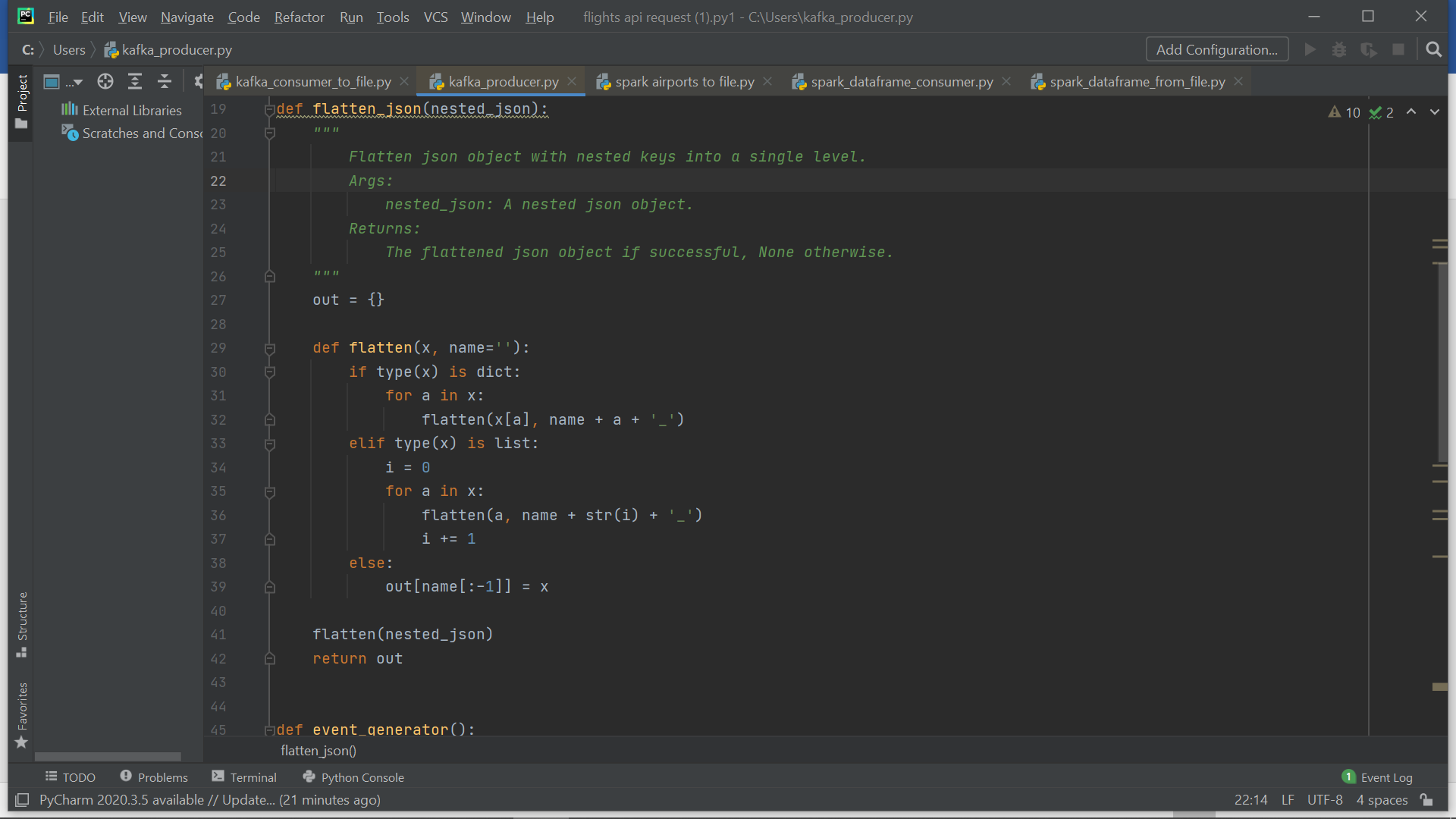
**Ptython scripts**

**Kafka Produscer**

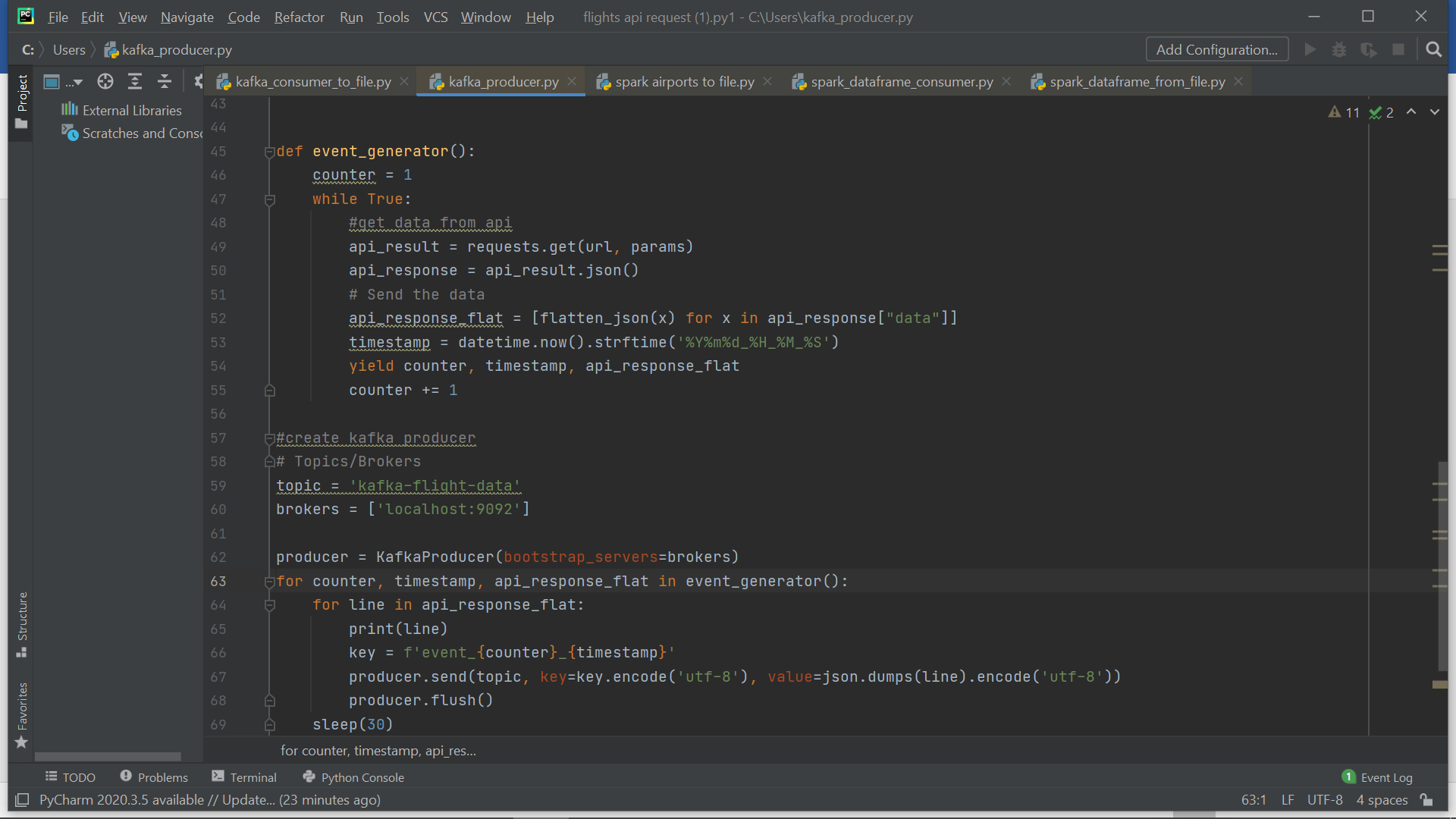
חיבור ל-API



מתקבל קובץ JSON, הפונקציה flatten\_json משטחת אותו:



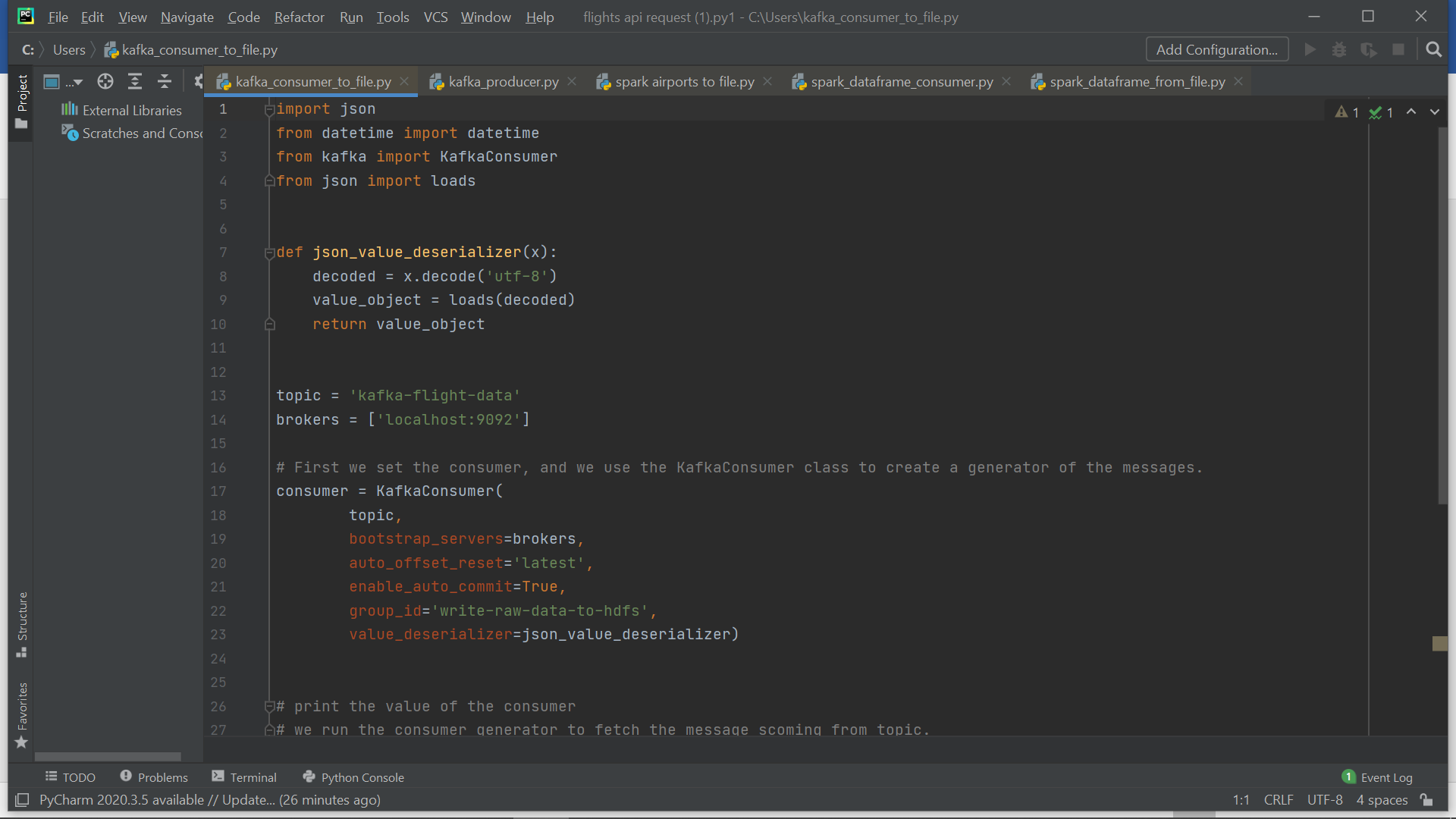
גנרטור שמקבל את המידע מהקובץ ושלוח לפרודוסר:



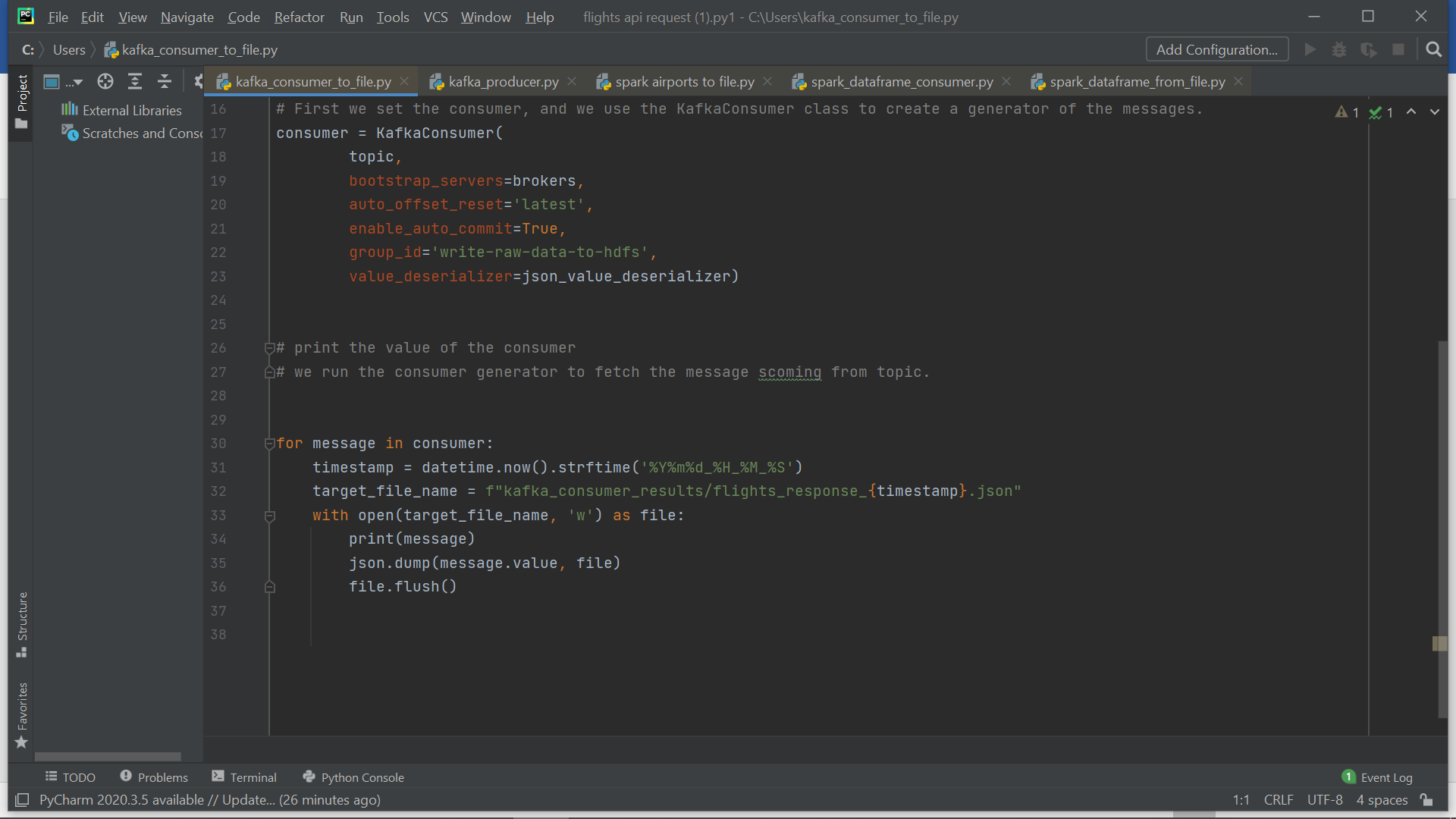
**Kafka\_consumer\_to\_file**

Json\_value\_deserializer-גיבוי של המידע לקובץ

Topic ו- brokers

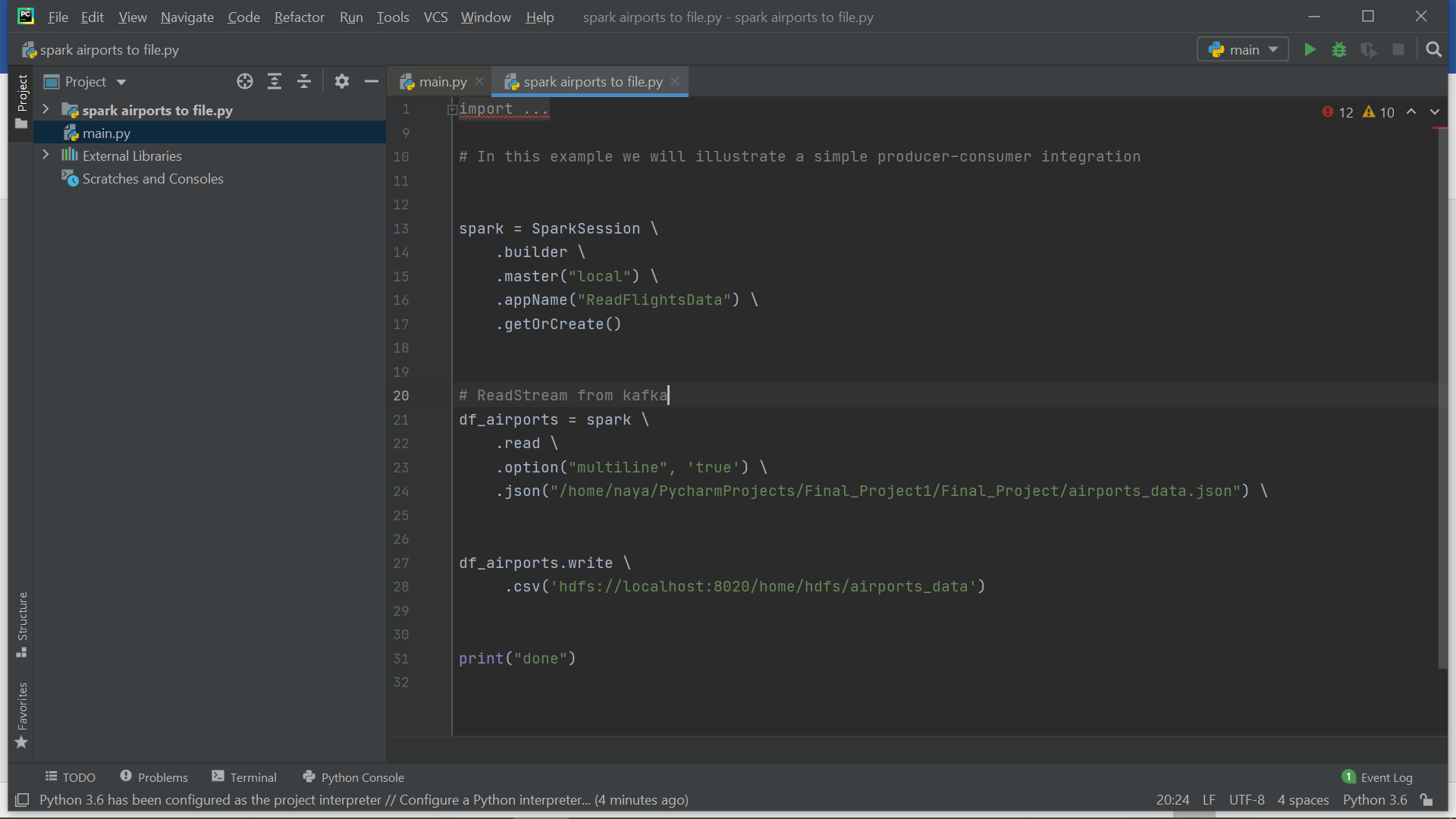


Kafka Consumer



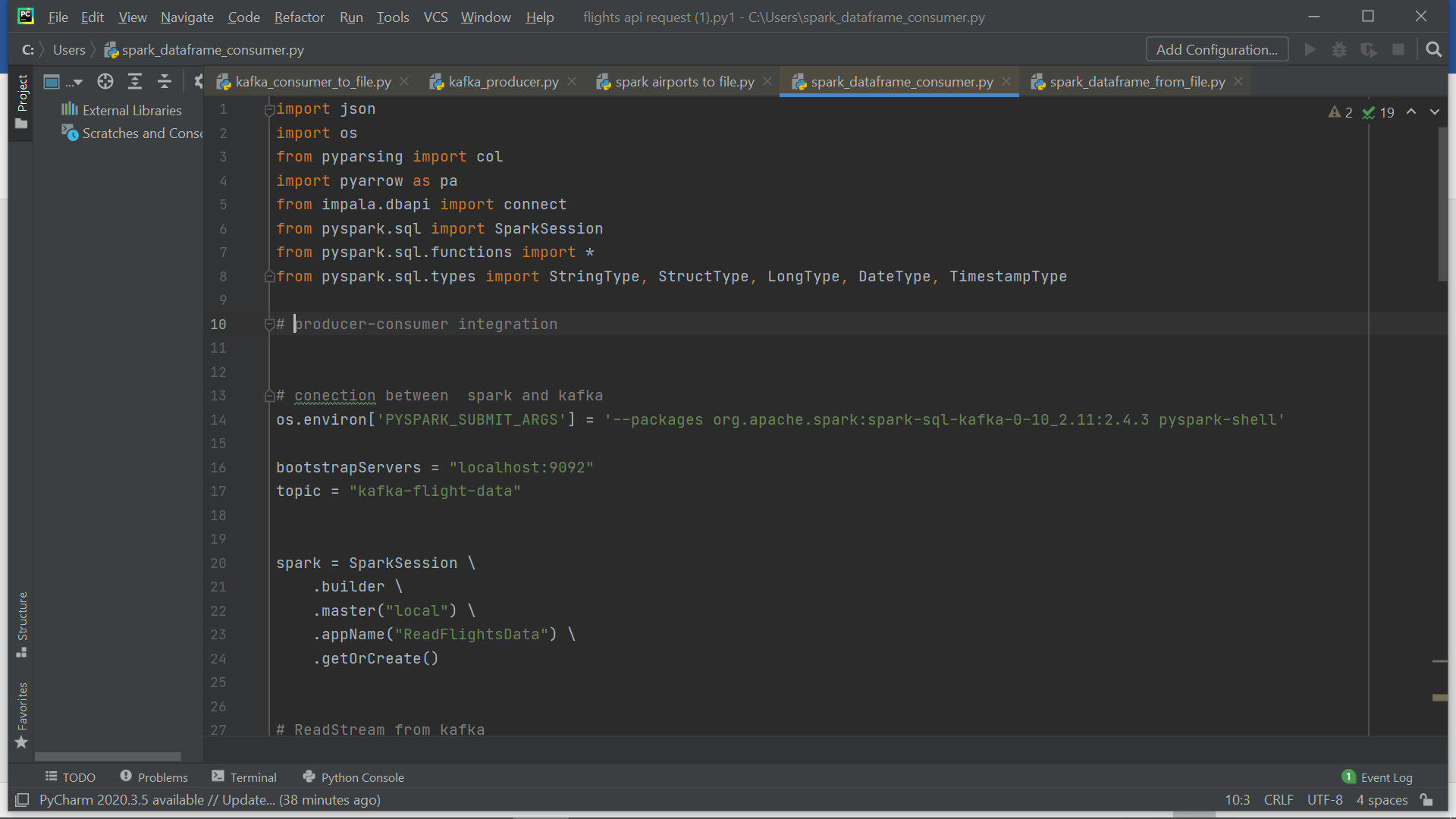
**Spark airoports to file**

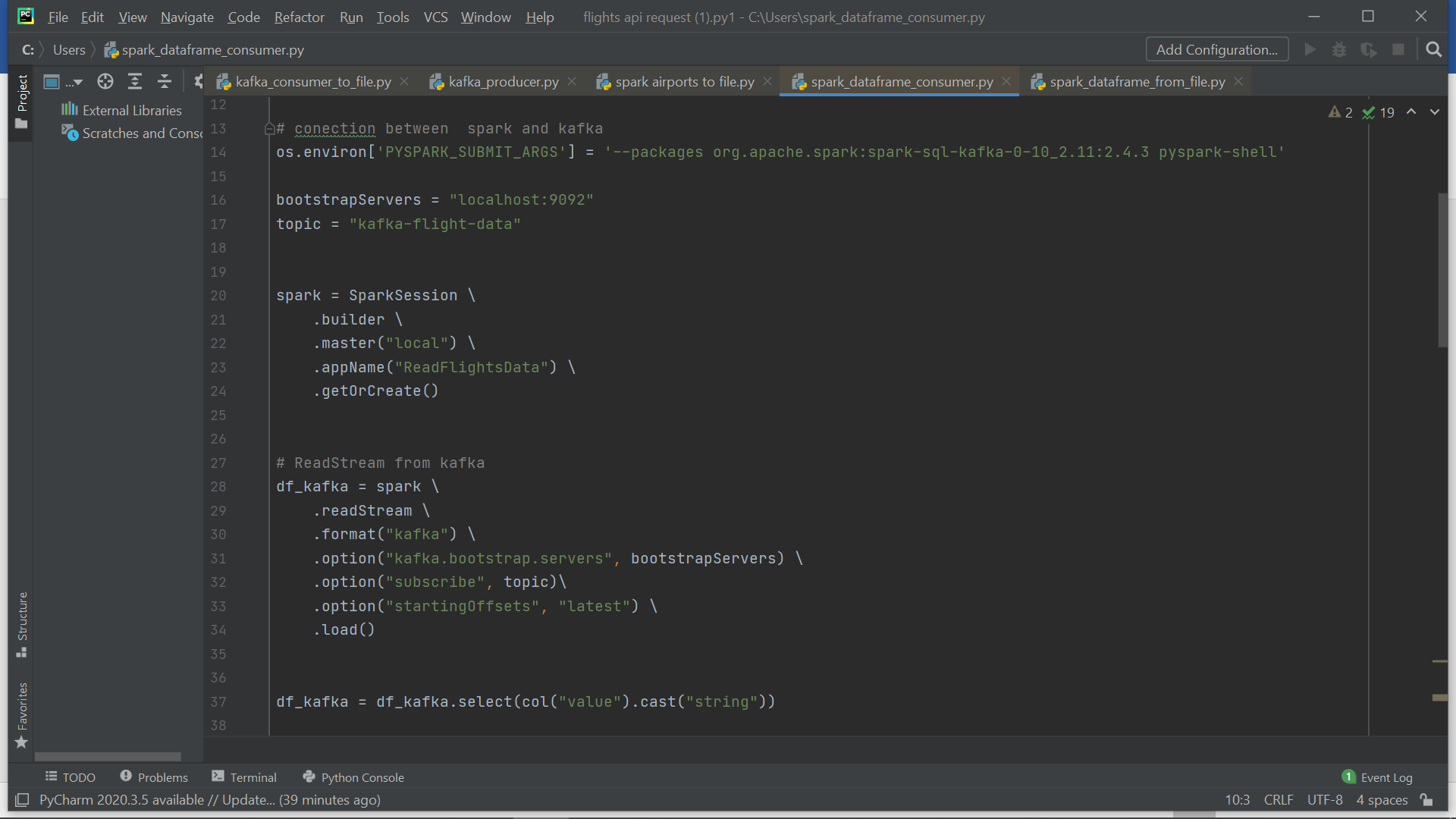
**מקבל את המידע מהפרודוסר ובספרק מייצר סכמה**

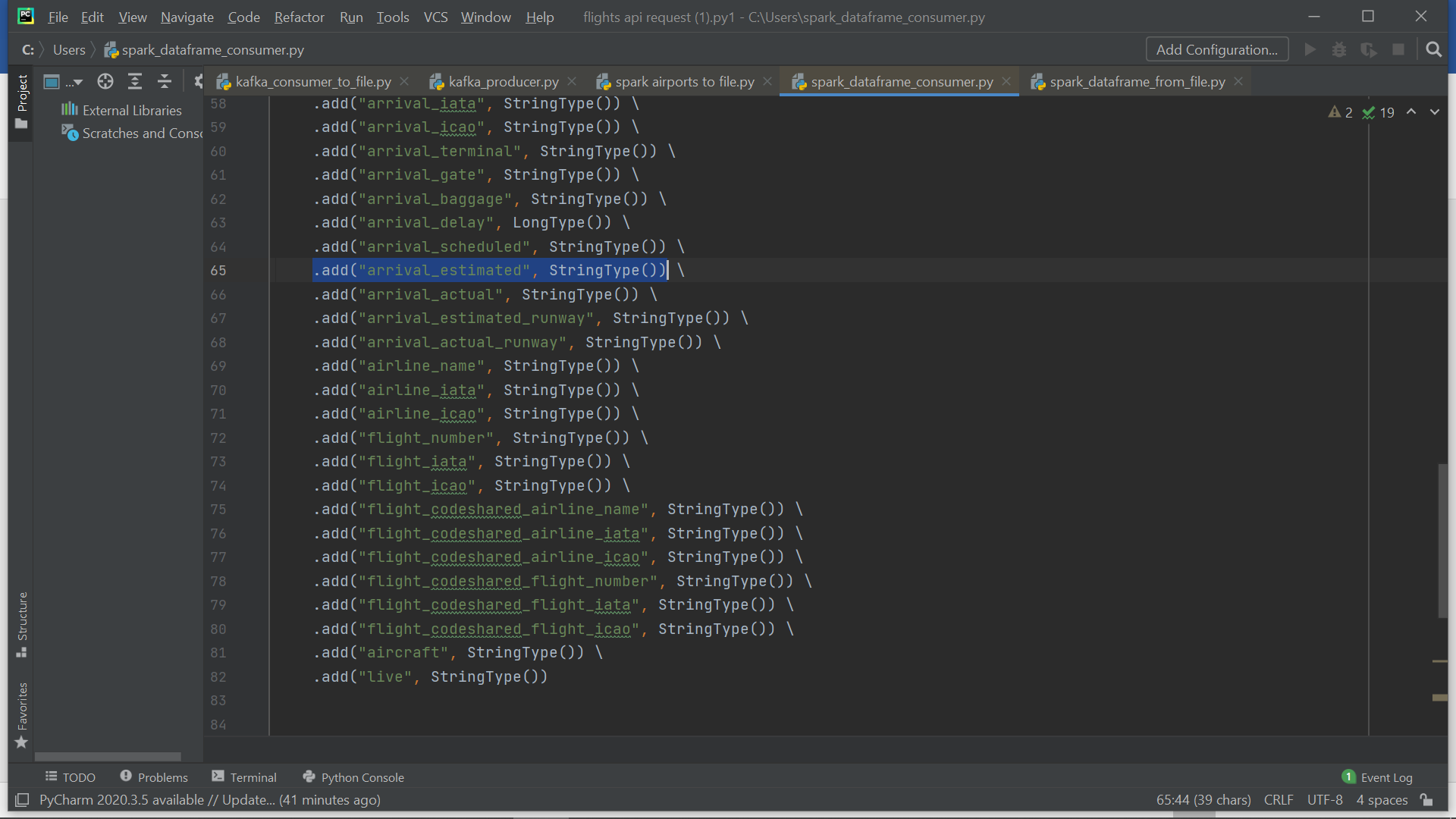


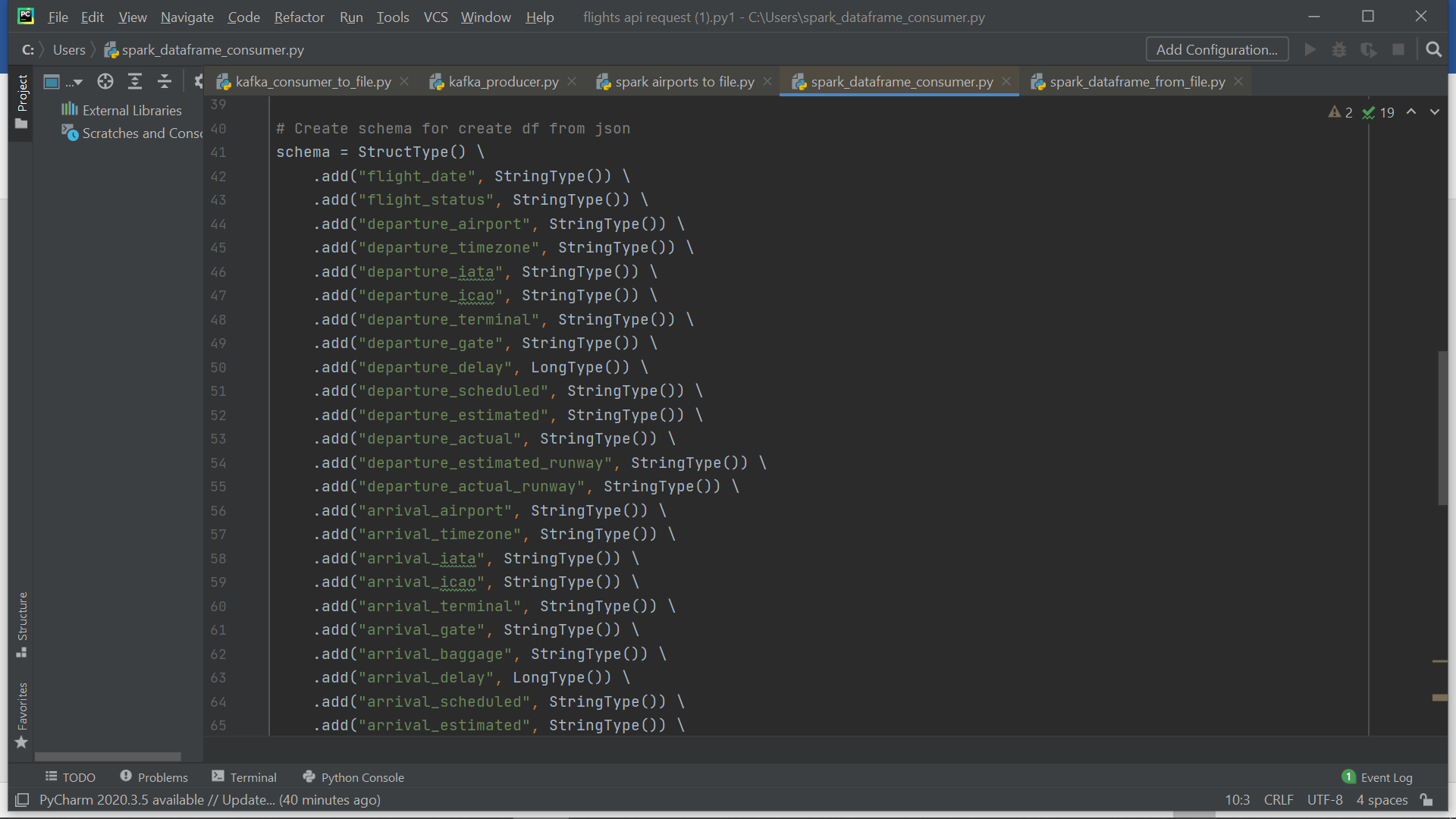
**Spark\_dataframe\_consumer**

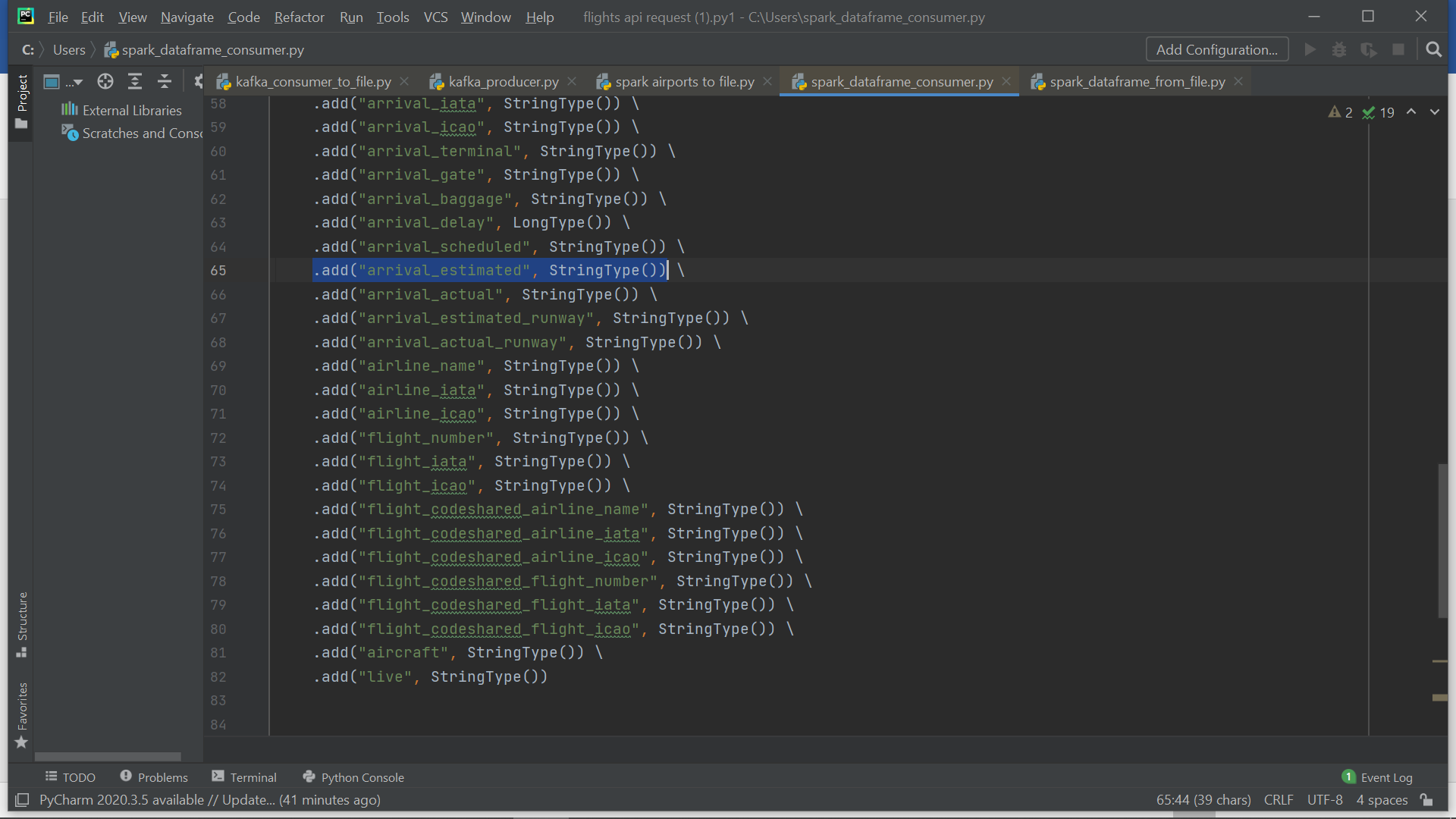
מתחבר לטופיק, יוצר סכמה ומעביר את המידע ל HDFS

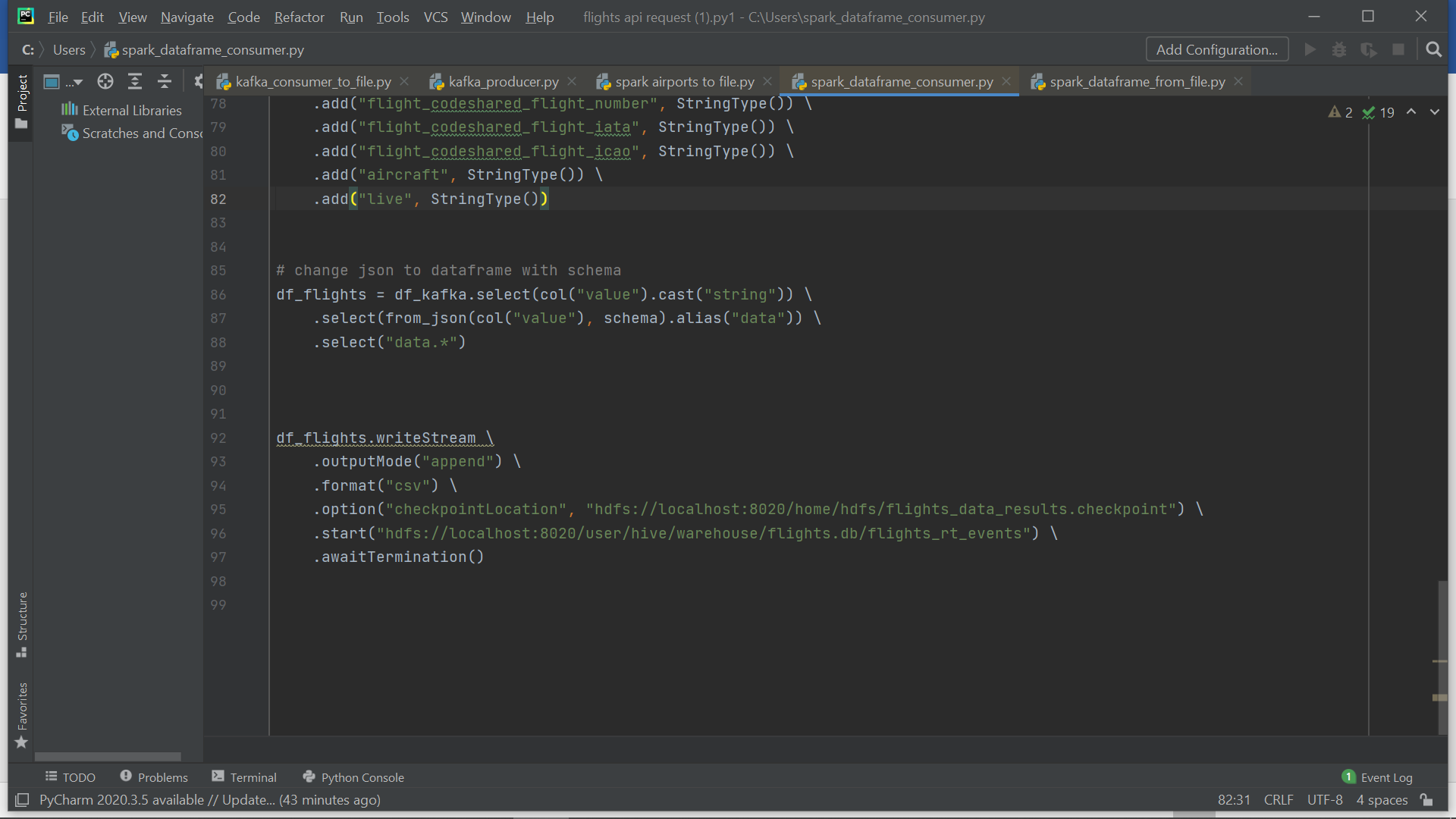










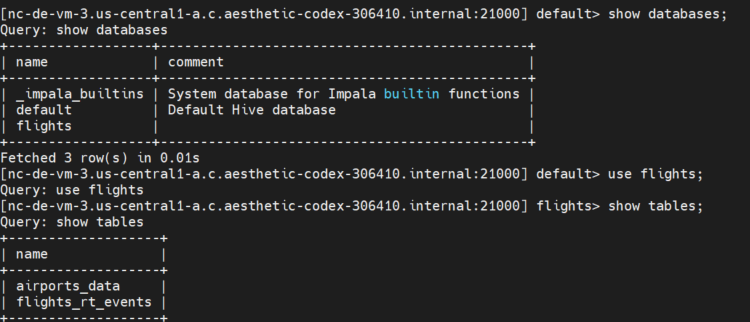


**מקבלים טבלה ומתשאלים אותה בImpala**

באימפלה יש DB flights

ובתוכה יש טבלה airports\_data- זו הטבלה הסטטית עם המידע על שדי תעופה

ויש טבלת המידע שמתקבל – flights\_rt\_events



פותחים את הטבלה flights\_rt\_events

