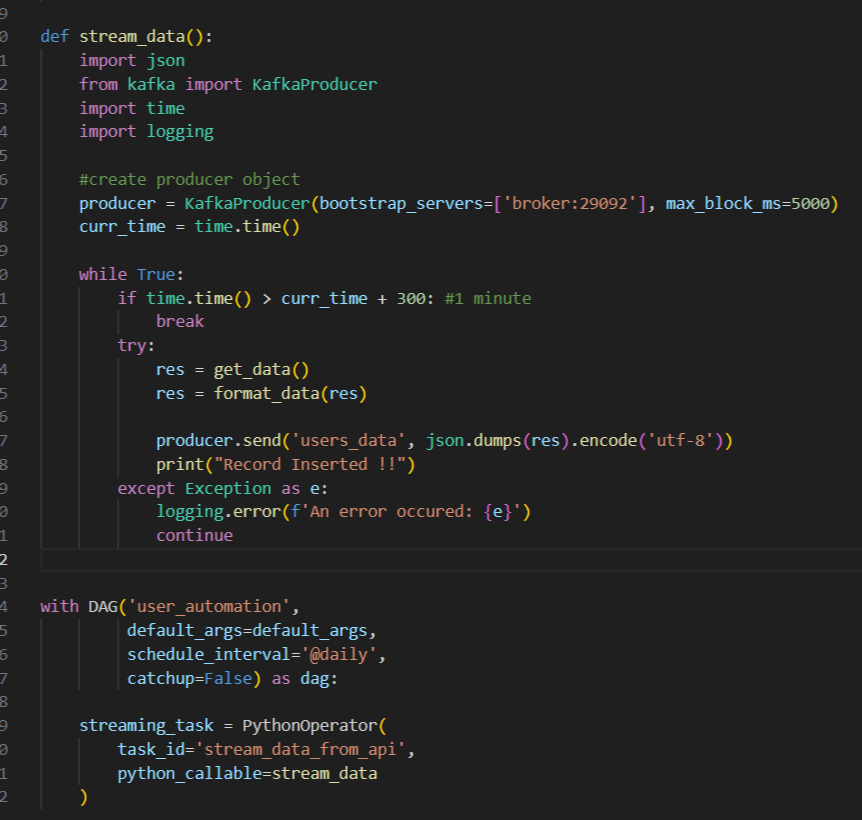
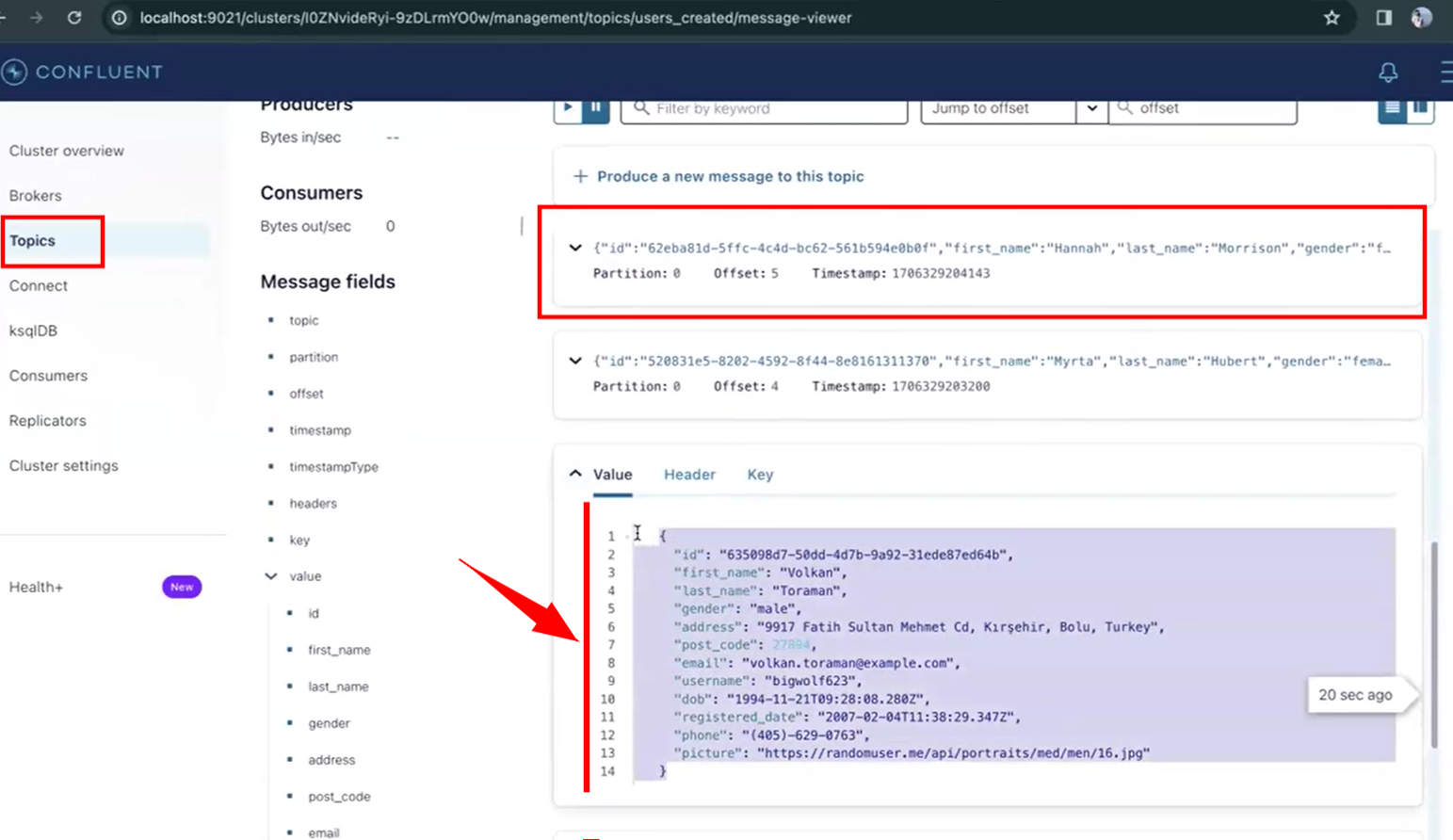
Users Streaming Data Ingestion Using Spark

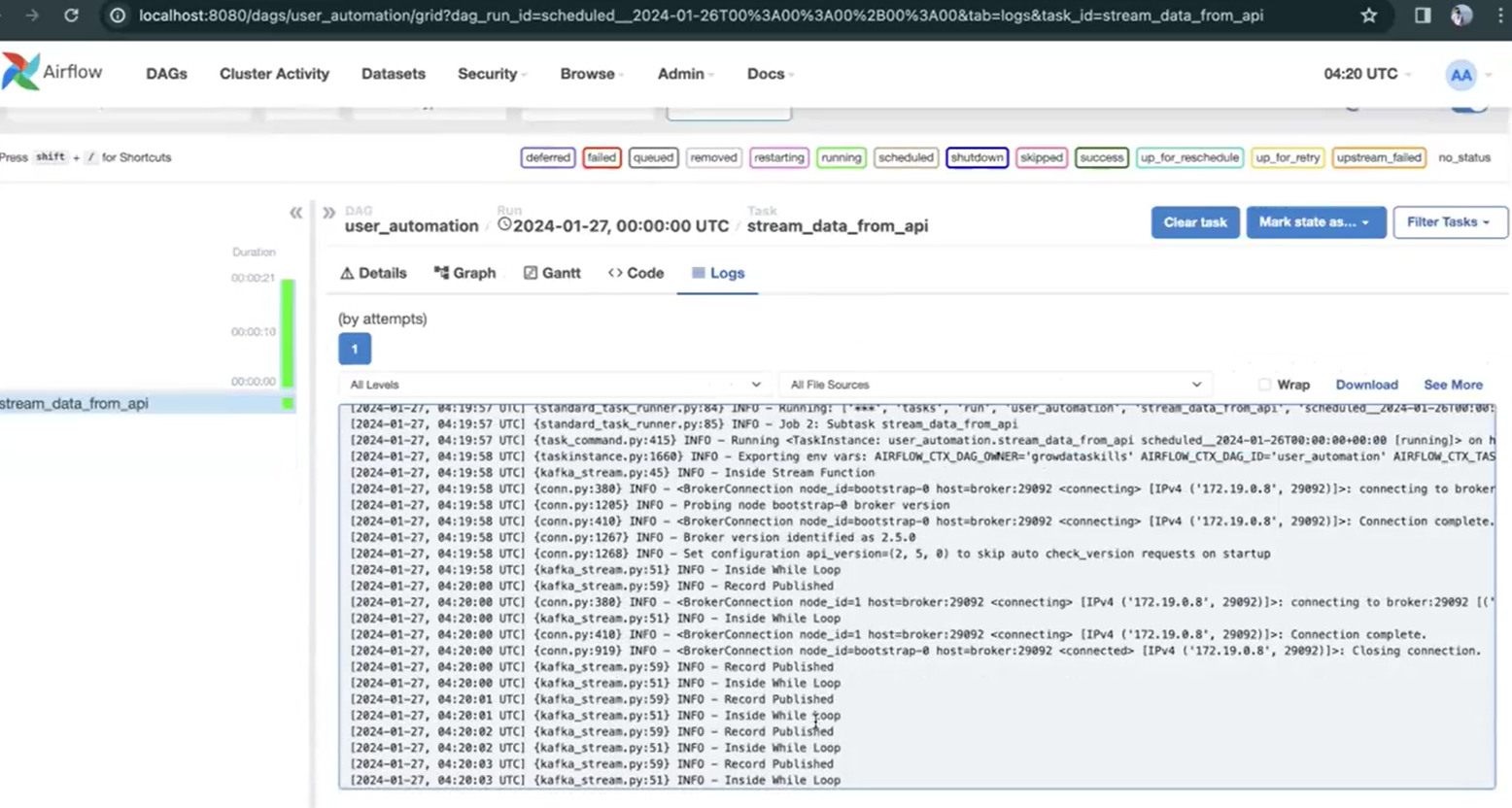
1. installed confluent kafka & airflow locally following the doc notes. Opened both kafka & airflow dashboard on browser locally
2. we were publishing data to kafka topic using the airflow script



1. so we create kafka topic 'users\_data' in confluent kafka and schedule the dag file where records started publishing to kafka topic.

Checked both dag file logs and kafka topic messages:





1. Next the data was streaming from kafka and ingesting into cassandra. For this wrote a spark\_stream script where first I established a spark connection. Inside spark connection we included all the jar files using which spark can communicate with both kafka and cassandra. Used jar files based on our pyspark version.



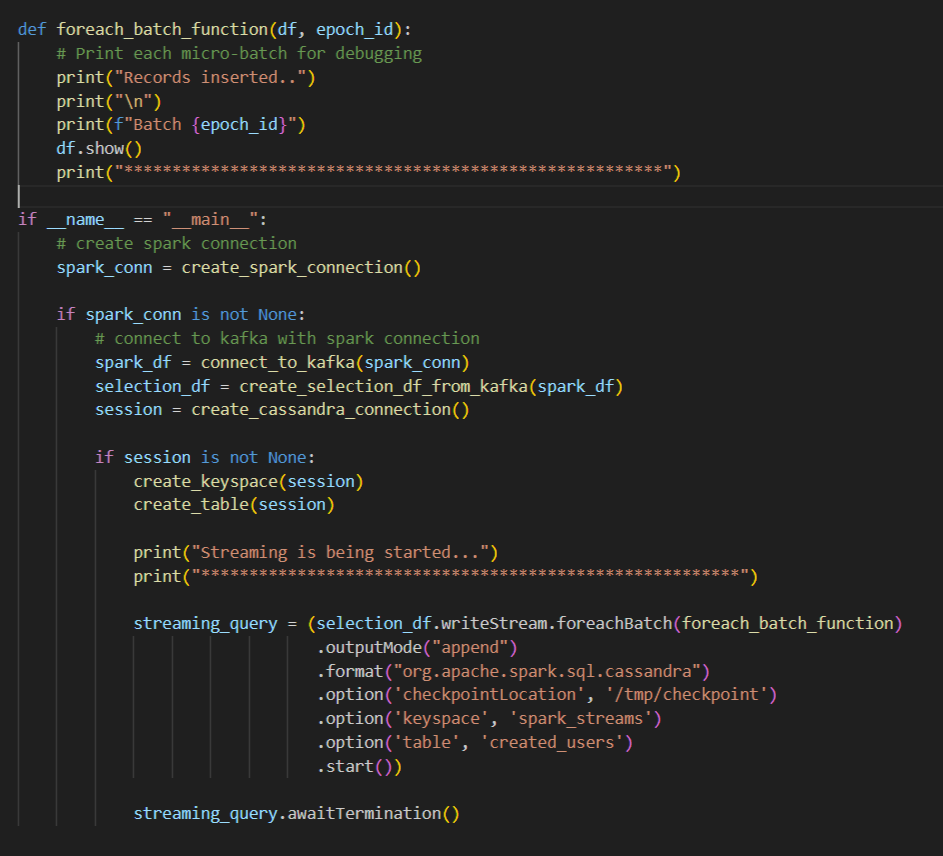
1. Once spark connection is done, next we created dataframe by connecting & reading data from kafka topic. So we were getting data in the form of key value pair where we were using custom schema to get a clean row columnar based dataframe.



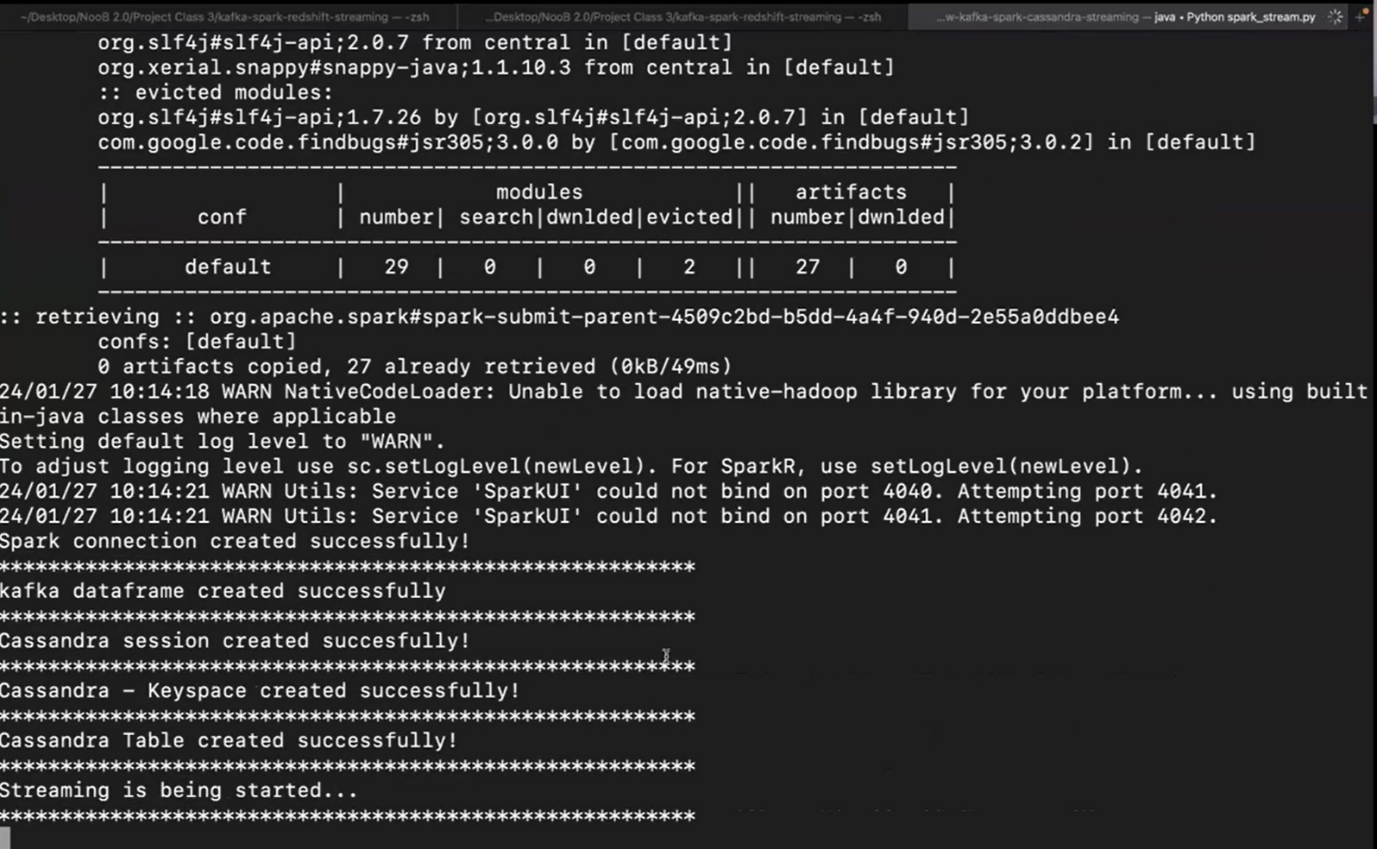
1. Next we created cassandra connection. We created keyspace and table using connection.



1. Then we were writing dataframe into cassandra using spark writeStream



1. Checked spark streaming app running successfully



1. At last, checked Cassandra table records using cql command

