Load data from Kafka to Hadoop

<Steps to run the python file to load data from Kafka>

1. Go to root user (sudo –i)
2. Create a python file which contains code to consume clickstream data from Kafka and save it to local directory. (vi spark_kafka_to_local.py)
3. Spark submit the python file with spark jar file (spark2-submit --jars "spark-sql-kafka-0-10_2.11-2.3.0.jar" spark_kafka_to_local.py)
4. Create another pyhton file o clean the loaded Kafka data to a more structured format as csv file format.(spark_local_flatten.py).
5. Spark submit the python file with spark jar file (spark2-submit --jars "spark-sql-kafka-0-10_2.11-2.3.0.jar" spark_local_flatten.py)

<Steps to load the data into Hadoop>

1. Create a input directory (hadoop fs -mkdir /user/root/)
2. Transfer and store a data file from local systems to the Hadoop file system using the put command. (hdfs dfs -put home/clickstream_data_flatten /user/root/clickstream_data_flatten)

<Screenshot of the data>