Load data from Kafka to Hadoop

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

1. CREATING AN EDITABLE PYTHON FILE TO LOAD KAFKA DATA

   vi spark_kafka_to_local.py

2. SPARK SUBMIT

   spark2-submit --jars "spark-sql-kafka-0-10_2.11-2.3.0.jar" spark_kafka_to_local.py

3. CREATING ANOTHER PYTHON FILE TO CLEAN LOADED KAFKA DATA

   vi spark_local_flatten.py

4. SPARK SUBMIT

   spark2-submit --jars "spark-sql-kafka-0-10_2.11-2.3.0.jar" spark_kafka_to_local.py

<Steps to load the data into Hadoop>

1. MAKE A DIRECTORY

   hadoop fs -mkdir clickstream_data_flatten

2. LOADING THE DATA FROM LOCAL FILE SYSTEM TO HDFS

   hadoop fs- put ~/clickstream_data_flatten clickstream_data_flatten

3. VALIDATING THE DATA FILE IN HDFS

   hadoop fs -ls clickstream_data_flatten hadoop fs -cat clickstream_data_flatten/part-00000-e2081929-45dc-d88a-40dd-9532a4c2b628-fc4c.csv | wc -1

<Screenshot of the data>