# Load data from Kafka to Hadoop

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

1. Creating a python file to consume data from kafka using vi editor

   vi spark_kafka_to_local.py

2. Spark job command to be run

   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5
   spark_kafka_to_local.py 18.211.252.152 9092 de-capstone5

3. Creating one more python file cleaning the data to a more structured format using vi editor

   vi spark_local_flatten.py

4. Running the spark job command

   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5
   spark_local_flatten.py

**<Steps to load the data into Hadoop>**

1.Creating a directory in hadoop file sysytem

   hadoop fs -mkdir clickstream_data_flatten

2. Load the data from local file system to hadoop file system

   hadoop fs- put ~/clickstream_data_flatten clickstream_data_flatten

3. Verifying the data file in hadoop

   hadoop fs -ls clickstream_data_flatten

   hadoop fs -cat clickstream_data_flatten/clickstream_data_flatten/part-00000-bb423f13-4963-4dd7-8afb-0630877df998-c000.csv | wc -l

**<Screenshot of the data>**

```
[hadoop@ip-172-31-70-125 ~]$ hadoop fs -ls clickstream_data_flatten
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-04-26 19:12 clickstream_data_flatten/_SUCCESS
-rw-r--r--   1 hadoop hadoop     376742 2024-04-26 19:12 clickstream_data_flatten/part-00000-bb423f13-4963-4dd7-8afb-0630877df998-c000.csv
[hadoop@ip-172-31-70-125 ~]$ hadoop fs -cat clickstream_data_flatten/part-00000-bb423f13-4963-4dd7-8afb-0630877df998-c000.csv | wc -l
2454
[hadoop@ip-172-31-70-125 ~]$
```