# Load data from Kafka to Hadoop

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

1. Create a python file to consume data from kafka

   ```
   vi spark_kafka_to_local.py
   ```

2. Run spark submit  command

   ```
   export SPARK_KAFKA_VERSION=0.10
   ```

   ```
   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-capstone3
   ```

3. Create another python file to clean the loaded Kafka data to a more structured format

   ```
   vi spark_local_flatten.py
   ```

4. Run spark submit  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. Make a directory using mkdir command

   ```
   hadoop fs -mkdir clickstream_data_flatten
   ```

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

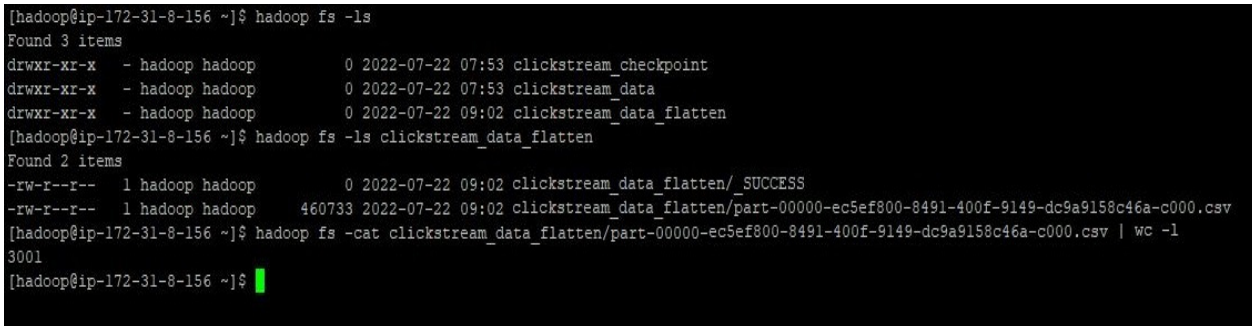
   ```
   hadoop fs- put ~/clickstream_data_flatten clickstream_data_flatten
   ```

3. Checking the data file in hadoop

   ```
   hadoop fs -ls clickstream_data_flatten
   ```

   ```
   hadoop fs -cat clickstream_data_flatten/ part-00000-ec5ef800-8491-400f-9149-
   dc9a9158c46a-c000.csv | wc -l
   ```

## <Screenshot of the data>



```
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -ls
Found 3 items
drwxr-xr-x   - hadoop hadoop          0 2022-07-22 07:53 clickstream_checkpoint
drwxr-xr-x   - hadoop hadoop          0 2022-07-22 07:53 clickstream_data
drwxr-xr-x   - hadoop hadoop          0 2022-07-22 09:02 clickstream_data_flatten
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -ls clickstream_data_flatten
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2022-07-22 09:02 clickstream_data_flatten/_SUCCESS
-rw-r--r--   1 hadoop hadoop     460733 2022-07-22 09:02 clickstream_data_flatten/part-00000-ec5ef800-8491-400f-9149-dc9a9158c46a-c000.csv
[hadoop@ip-172-31-8-156 ~]$ hadoop fs -cat clickstream_data_flatten/part-00000-ec5ef800-8491-400f-9149-dc9a9158c46a-c000.csv | wc -l
3001
[hadoop@ip-172-31-8-156 ~]$
```