

## 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 python file to 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>

```
[hdfs@ip-10-0-0-218 ~]$ hadoop fs -ls /user/root/clickstream_data_flatten
Found 2 items
-rw-r--r--  3 root supergroup      0 2021-03-11 15:13 /user/root/clickstream_data_flatten/_SUCCESS
-rw-r--r--  3 root supergroup 361629 2021-03-11 15:13 /user/root/clickstream_data_flatten/part-00000-2f2214f7-80c2-4286-90b9-b53bb9e4d91c-c000.csv
[hdfs@ip-10-0-0-218 ~]$ hadoop fs -cat /user/root/clickstream_data_flatten/part-00000-2f2214f7-80c2-4286-90b9-b53bb9e4d91c-c000.csv | wc -l
2973
[hdfs@ip-10-0-0-218 ~]$
```