## 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>**





