Load data from Kafka to Hadoop

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

1. **Downloaded the Jar file using:**

https://search.maven.org/artifact/org.apache.spark/spark-sql-kafka-0-10_2.11/2.4.6/jar

2. **Run to get data from kafka to local**

```
# Import Dependencies
import os
import sys
from pyspark.sql import SparkSession
from pyspark.sql.functions import *
from pyspark.sql.types import *
from pyspark.sql.functions import from_json
from pyspark.sql.window import Window

spark = SparkSession \
        .builder \
        .appName("Kafka-to-local") \
        .getOrCreate()
spark.sparkContext.setLogLevel('ERROR')

# Read Input
kafka_read = spark  \
        .readStream  \
        .format("kafka")  \
        .option("kafka.bootstrap.servers","18.211.252.152:9092")  \
        .option("subscribe","de-capstone3")  \
        .option("startingOffsets", "earliest")  \
        .load()

updated_data= kafka_read \
      .withColumn('value_str',kafka_read['value'].cast('string').alias('key_str')).drop('value') \
      .drop('key','topic','partition','offset','timestamp','timestampType')

clickstream_data = updated_data.writeStream \
  .format("json") \
  .outputMode("append") \
  .option("truncate", "false") \
  .option("path", "clickstream_data/") \
```

```
  .option("checkpointLocation", "clickstream_data/cp/") \
  .start()

clickstream_data.awaitTermination()
```

3. **Commands used to run the Spark Submit job:**

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.6  spark_kafka_to_local.py
```

<Steps to load the data into Hadoop>

1. **Clean data using spark_local_flatten.py**

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import *

spark=SparkSession.builder.appName("Kafka-to-HDFS").master("local").getOrCreate()
spark

#Reading data from hdfs
df=spark.read.json("clickstream.json")

df.show(10,truncate=False)
#Selecting the columns from the clickstream data set
df=df.select(get_json_object(df['value_str'],"$.customer_id").alias("customer_id"),
            get_json_object(df['value_str'],"$.app_version").alias("app_version"),
            get_json_object(df['value_str'],"$.OS_version").alias("OS_version"),
            get_json_object(df['value_str'],"$.lat").alias("lat"),
            get_json_object(df['value_str'],"$.lon").alias("lon"),
            get_json_object(df['value_str'],"$.page_id").alias("page_id"),
            get_json_object(df['value_str'],"$.button_id").alias("button_id"),
            get_json_object(df['value_str'],"$.is_button_click").alias("is_button_click"),
            get_json_object(df['value_str'],"$.is_page_view").alias("is_page_view"),
            get_json_object(df['value_str'],"$.is_scroll_up").alias("is_scroll_up"),
            get_json_object(df['value_str'],"$.is_scroll_down").alias("is_scroll_down"),
            get_json_object(df['value_str'],"$.timestamp").alias("timestamp")
             )

df.printSchema()

df.show(10)
```

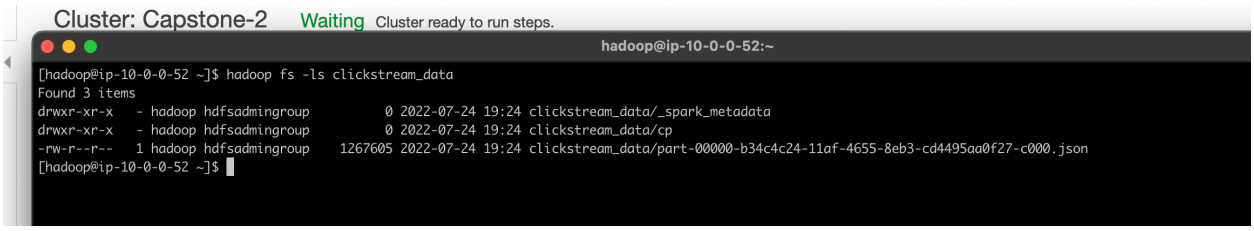
#Writing the dataset to hdfs
df.coalesce(1).write.format('csv').mode('overwrite').save('/user/hadoop/clickstream_flattened',header='true')

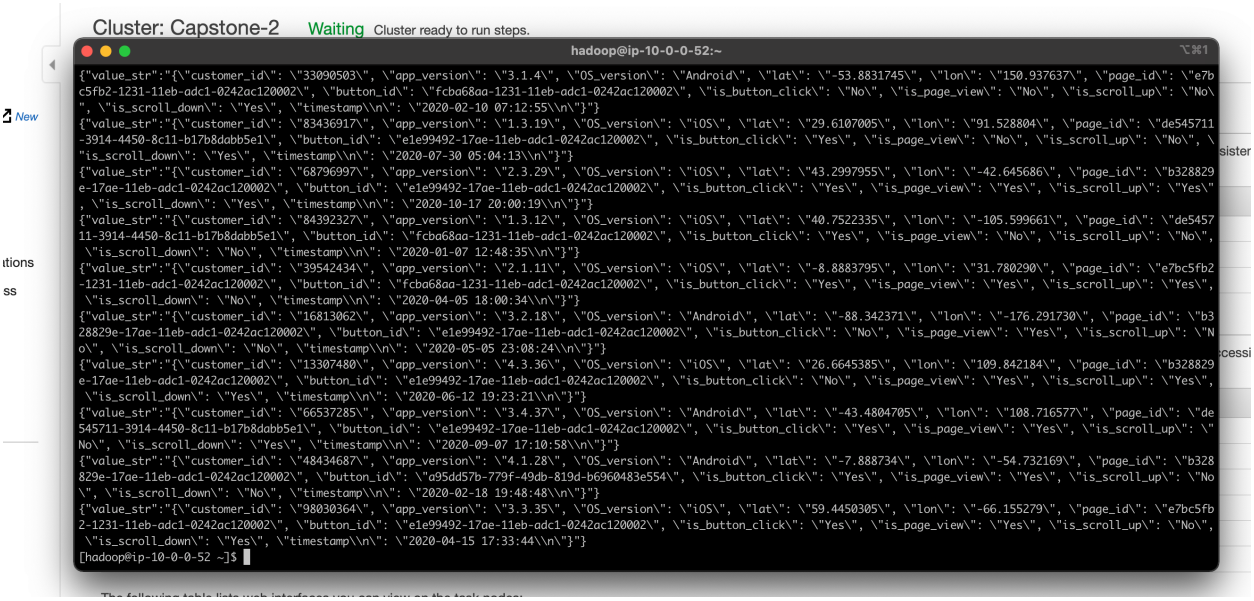
2. Run Spark Submit Job

spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.6 spark_local_flatten.py

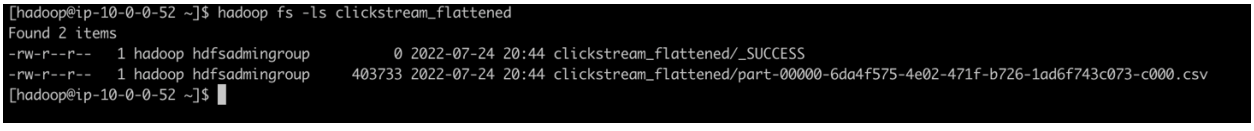
<Screenshot of the data>

1. Kafka to Local





2. Local to HDFS



```
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|value_str
                                                                                                                                                                                |
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|{"customer_id": "26564820", "app_version": "3.2.35", "OS_version": "Android", "lat": "16.4454865", "lon": "99.902065", "page_id": "de545711-3914-4450-8c11-b17b8dabb5e1", "button_id": "fcba68aa-1231-11eb-adc1-0242ac120
002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "Yes", "timestamp\n": "2020-09-14 09:59:07\n"}  |
|{"customer_id": "31906387", "app_version": "2.4.7", "OS_version": "iOS", "lat": "-64.813749", "lon": "-133.527040", "page_id": "de545711-3914-4450-8c11-b17b8dabb5e1", "button_id": "a95dd57b-779f-49db-819d-b6960483e554
", "is_button_click": "No", "is_page_view": "No", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp\n": "2020-05-16 16:30:21\n"}      |
|{"customer_id": "25713677", "app_version": "3.4.12", "OS_version": "Android", "lat": "89.943435", "lon": "127.313415", "page_id": "b328829e-17ae-11eb-adc1-0242ac120002", "button_id": "fcba68aa-1231-11eb-adc1-0242ac120
002", "is_button_click": "No", "is_page_view": "No", "is_scroll_up": "Yes", "is_scroll_down": "No", "timestamp\n": "2020-02-09 00:52:13\n"}   |
|{"customer_id": "83474293", "app_version": "3.1.8", "OS_version": "Android", "lat": "-69.939070", "lon": "-36.451670", "page_id": "e7bc5fb2-1231-11eb-adc1-0242ac120002", "button_id": "e1e99492-17ae-11eb-adc1-0242ac120
002", "is_button_click": "Yes", "is_page_view": "No", "is_scroll_up": "Yes", "is_scroll_down": "No", "timestamp\n": "2020-06-17 10:42:50\n"}  |
|{"customer_id": "63727807", "app_version": "2.2.9", "OS_version": "iOS", "lat": "64.082108", "lon": "-81.822078", "page_id": "e7bc5fb2-1231-11eb-adc1-0242ac120002", "button_id": "fcba68aa-1231-11eb-adc1-0242ac120002",
 "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "Yes", "is_scroll_down": "Yes", "timestamp\n": "2020-07-06 02:51:53\n"}      |
|{"customer_id": "73737907", "app_version": "4.3.19", "OS_version": "Android", "lat": "-18.850508", "lon": "-116.358375", "page_id": "b328829e-17ae-11eb-adc1-0242ac120002", "button_id": "e1e99492-17ae-11eb-adc1-0242ac1
20002", "is_button_click": "No", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "Yes", "timestamp\n": "2020-04-26 06:18:16\n"}|
|{"customer_id": "36927433", "app_version": "3.2.26", "OS_version": "iOS", "lat": "-84.6857245", "lon": "-146.507678", "page_id": "de545711-3914-4450-8c11-b17b8dabb5e1", "button_id": "a95dd57b-779f-49db-819d-b6960483e5
54", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "Yes", "timestamp\n": "2020-02-06 10:21:18\n"}  |
|{"customer_id": "12691783", "app_version": "3.3.11", "OS_version": "Android", "lat": "54.3852925", "lon": "-37.411814", "page_id": "de545711-3914-4450-8c11-b17b8dabb5e1", "button_id": "e1e99492-17ae-11eb-adc1-0242ac12
0002", "is_button_click": "Yes", "is_page_view": "Yes", "is_scroll_up": "No", "is_scroll_down": "No", "timestamp\n": "2020-08-08 04:23:56\n"} |
|{"customer_id": "22635021", "app_version": "4.4.36", "OS_version": "iOS", "lat": "-31.805500", "lon": "150.655650", "page_id": "e7bc5fb2-1231-11eb-adc1-0242ac120002", "button_id": "a95dd57b-779f-49db-819d-b6960483e554
", "is_button_click": "No", "is_page_view": "No", "is_scroll_up": "No", "is_scroll_down": "No", "timestamp\n": "2020-08-02 00:33:50\n"}        |
|{"customer_id": "23593546", "app_version": "1.2.16", "OS_version": "Android", "lat": "8.8918475", "lon": "-83.929878", "page_id": "de545711-3914-4450-8c11-b17b8dabb5e1", "button_id": "e1e99492-17ae-11eb-adc1-0242ac120
002", "is_button_click": "Yes", "is_page_view": "No", "is_scroll_up": "Yes", "is_scroll_down": "No", "timestamp\n": "2020-07-23 23:59:19\n"}  |
+--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```