Loading data from Kafka to Hadoop

Reading kafka stream and writing steaming data as .json file in HDFS local directory

1. **Pyspark file "spark_kafka_to_local.py" created to read data from the kafka server "kafka.bootstrap.servers".**

2. **Executing this file to save streaming data into HDFS:**

   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 s3://do lat- s3/CapstoneFiles/spark_kafka_to_local.py

```
[hadoop@ip-172-31-33-215 ~]$ hadoop fs -ls /user/root
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2023-05-23 12:12 /user/root/bookings
drwxr-xr-x   - hadoop hadoop          0 2023-05-23 13:00 /user/root/datewise_bookings_agg
[hadoop@ip-172-31-33-215 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 s3://dolat-s3/CapstoneFiles/spark_kafka_to_local.py
```


3. **Verifying Streaming data files are created in HDFS:**

   hadoop fs -ls /user/root/clickstream_data_dump

```
[hadoop@ip-172-31-33-215 ~]$ hadoop fs -ls /user/root
Found 4 items
drwxr-xr-x   - hadoop hadoop          0 2023-05-23 12:12 /user/root/bookings
drwxr-xr-x   - hadoop hadoop          0 2023-05-23 13:31 /user/root/clickstream_data_dump
drwxr-xr-x   - hadoop hadoop          0 2023-05-23 13:31 /user/root/clickstream_data_dump_cp
drwxr-xr-x   - hadoop hadoop          0 2023-05-23 13:00 /user/root/datewise_bookings_agg
[hadoop@ip-172-31-33-215 ~]$ hadoop fs -ls /user/root/clickstream_data_dump
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2023-05-23 13:31 /user/root/clickstream_data_dump/_spark_metadata
-rw-r--r--   1 hadoop hadoop    1267706 2023-05-23 13:31 /user/root/clickstream_data_dump/part-00000-e08eb778-7b21-4c2b-84cd-9e2c6467c5ea-c000.json
[hadoop@ip-172-31-33-215 ~]$
```

4. **Verifying top few records in created json file.**

   hadoop fs -cat /user/root/clickstream_data_dump/part-00000-e08eb778-7b21-4c2 -84cd- 9e2c6467c5ea-c000.json | head -n 5

```
[hadoop@ip-172-31-33-215 ~]$ hadoop fs -cat /user/root/clickstream_data_dump/part-00000-e08eb778-7b21-4c2b-84cd-9e2c6467c5ea-c000.json | head -n 5
{"value_str":"{\"customer_id\": \"26564820\", \"app_version\": \"3.2.35\", \"OS_version\": \"Android\", \"lat\": \"16.4454865\", \"lon\": \"99.902065
de545711-3914-4450-8c11-b17b8dabb5e1\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_click\": \"No\", \"is_page_view\": \"Yes
": \"No\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-09-14 09:59:07\\n\"}"}
{"value_str":"{\"customer_id\": \"31906387\", \"app_version\": \"2.4.7\", \"OS_version\": \"iOS\", \"lat\": \"-64.813749\", \"lon\": \"-133.527040\",
45711-3914-4450-8c11-b17b8dabb5e1\", \"button_id\": \"a95dd57b-779f-49db-819d-b6960483e554\", \"is_button_click\": \"No\", \"is_page_view\": \"No\",
"Yes\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-05-16 16:30:21\\n\"}"}
{"value_str":"{\"customer_id\": \"25713677\", \"app_version\": \"3.4.12\", \"OS_version\": \"Android\", \"lat\": \"89.943435\", \"lon\": \"127.313415
b328829e-17ae-11eb-adc1-0242ac120002\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_click\": \"No\", \"is_page_view\": \"No\
: \"Yes\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-02-09 00:52:13\\n\"}"}
{"value_str":"{\"customer_id\": \"83474293\", \"app_version\": \"3.1.8\", \"OS_version\": \"Android\", \"lat\": \"-69.939070\", \"lon\": \"-36.451670
e7bc5fb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"No
": \"Yes\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-06-17 10:42:50\\n\"}"}
{"value_str":"{\"customer_id\": \"63727807\", \"app_version\": \"2.2.9\", \"OS_version\": \"iOS\", \"lat\": \"64.082108\", \"lon\": \"-81.822078\", \
fb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_click\": \"No\", \"is_page_view\": \"Yes\",
Yes\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-07-06 02:51:53\\n\"}"}
cat: Unable to write to output stream.
[hadoop@ip-172-31-33-215 ~]$
```

Reading created .json file from HDFS and transforming into structured format as .csv file

1. **Pyspark file "spark_local_flatten.py" created to read .json file and transform into more structured format as .csv file:**

   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 s3://do lat- s3/CapstoneFiles/spark_local_flatten.py

```
[hadoop@ip-172-31-33-215 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 s3://dolat-s3/CapstoneFiles/spark_local_flatten.py
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
The jars for the packages stored in: /home/hadoop/.ivy2/jars
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apache/ivy/core/settings/ivysettings.xml
org.apache.spark#spark-sql-kafka-0-10_2.11 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-c88c7eb4-949f-4a65-a8dd-c3de08117f30;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 in central
        found org.apache.kafka#kafka-clients;2.0.0 in central
        found org.lz4#lz4-java;1.4.0 in central
        found org.xerial.snappy#snappy-java;1.1.7.3 in central
        found org.slf4j#slf4j-api;1.7.16 in central
        found org.spark-project.spark#unused;1.0.0 in central
:: resolution report :: resolve 413ms :: artifacts dl 12ms
        :: modules in use:
        org.apache.kafka#kafka-clients;2.0.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5 from central in [default]
        org.lz4#lz4-java;1.4.0 from central in [default]
        org.slf4j#slf4j-api;1.7.16 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.7.3 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   6   |   0   |   0   |   0   ||   6   |   0   |
        ---------------------------------------------------------------------
```


2. **Verifying created csv file in specified directory**

   hadoop fs -ls /user/root/clickstream_flattened

```
[hadoop@ip-172-31-33-215 ~]$ hadoop fs -ls /user/root/clickstream_flattened
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2023-05-23 14:03 /user/root/clickstream_flattened/_SUCCESS
-rw-r--r--   1 hadoop hadoop     403841 2023-05-23 14:03 /user/root/clickstream_flattened/part-00000-199eb4e5-0cac-44a1-a819-151d442489df-c000.csv
[hadoop@ip-172-31-33-215 ~]$
```


3. **Printing top few records from this csv file to make sure that transformation was successful:**

   hadoop fs -cat /user/root/clickstream_flattened/part-00000-199eb4e5-0cac-44a1-a819-151d442489df- c000.csv | head -n 10

```
[hadoop@ip-172-31-33-215 ~]$ hadoop fs -cat  /user/root/clickstream_flattened/part-00000-199eb4e5-0cac-44a1-a819-151d442489df-c000.csv | head -n 10
customer_id,app_version,OS_version,lat,lon,page_id,button_id,is_button_click,is_page_view,is_scroll_up,is_scroll_down,timestamp
26564820,3.2.35,Android,16.4454865,99.902065,de545711-3914-4450-8c11-b17b8dabb5e1,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,No,Yes,""
31906387,2.4.7,iOS,-64.813749,-133.527040,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,No,No,Yes,Yes,""
25713677,3.4.12,Android,89.943435,127.313415,b328829e-17ae-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,Yes,No,""
83474293,3.1.8,Android,-69.939070,-36.451670,e7bc5fb2-1231-11eb-adc1-0242ac120002,ele99492-17ae-11eb-adc1-0242ac120002,Yes,No,Yes,No,""
63727807,2.2.9,iOS,64.082108,-81.822078,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,Yes,Yes,""
73737907,4.3.19,Android,-18.850508,-116.358375,b328829e-17ae-11eb-adc1-0242ac120002,ele99492-17ae-11eb-adc1-0242ac120002,No,Yes,No,Yes,""
36927433,3.2.26,iOS,-84.6857245,-146.507678,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,Yes,Yes,No,Yes,""
12691783,3.3.11,Android,54.3852925,-37.411814,de545711-3914-4450-8c11-b17b8dabb5e1,ele99492-17ae-11eb-adc1-0242ac120002,Yes,Yes,No,No,""
22635021,4.4.36,iOS,-31.805500,150.655650,e7bc5fb2-1231-11eb-adc1-0242ac120002,a95dd57b-779f-49db-819d-b6960483e554,No,No,No,No,""
cat: Unable to write to output stream.
[hadoop@ip-172-31-33-215 ~]$
```