Load data from Kafka to Hadoop

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

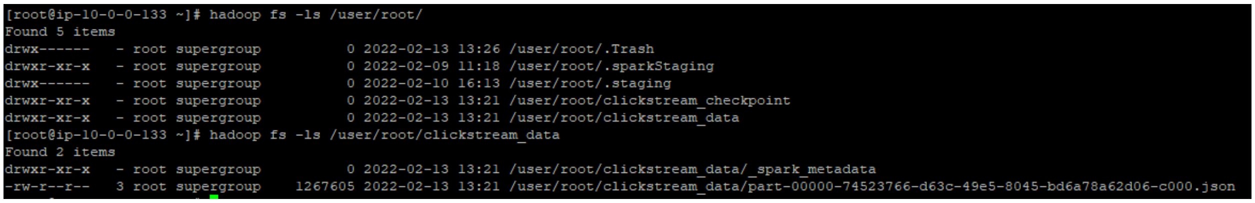
1. Connect to the ec2 instance
2. Switch the user from ec2 to root using. (**sudo –i**)
3. Create a python file with code that can ingest real-time clickstream data from Kafka Server and save it to the local directory. (vi **spark_kafka_to_local.py)**
4. Run the command - **export SPARK_KAFKA_VERSION=0.10**
5. Submit the spark job using 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**)
6. Create another python file for cleaning the Kafka loaded data and structuring the data and saving it in csv file format. (vi **spark_local_flatten.py**)
7. 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>

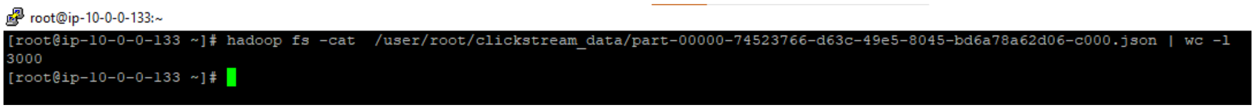
In the above cleaning python file we have specified the default path for the creation of the folder where the Kafka data would be stored in the structured format and screenshot of the code and the folder is provided below.

df1.coalesce(1).write.format('com.databricks.spark.csv').mode('overwrite').save('user/root /clickstream_data_flatten', header = 'true')

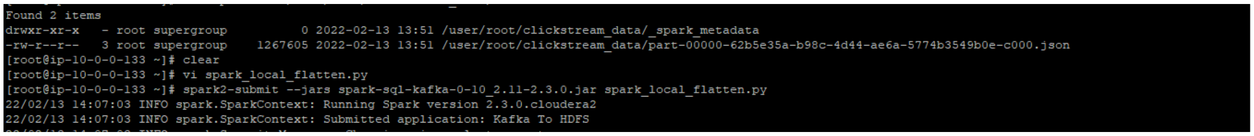
<Screenshot of the data>



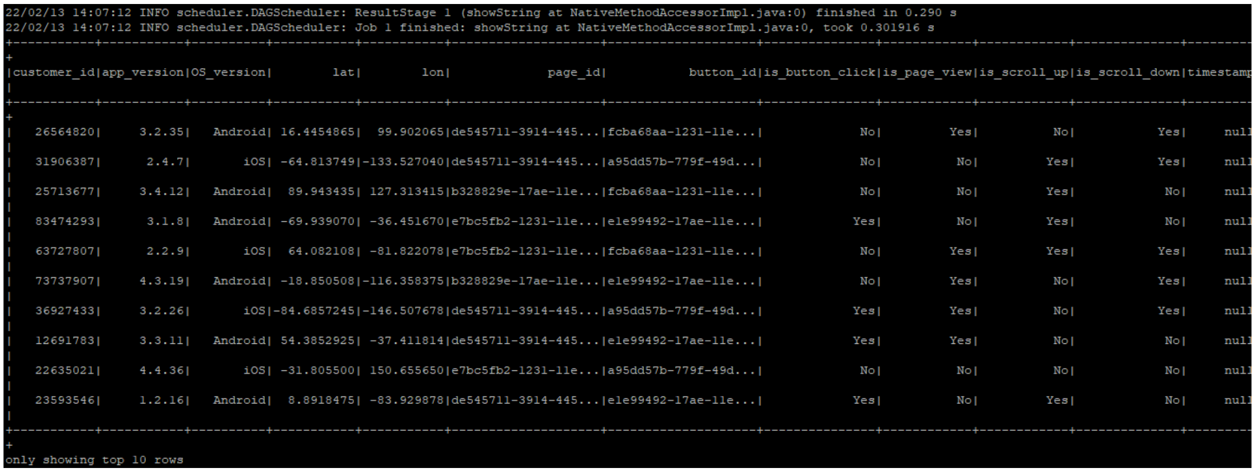
```
[root@ip-10-0-0-133 ~]# hadoop fs -ls /user/root/
Found 5 items
drwx------   - root supergroup          0 2022-02-13 13:26 /user/root/.Trash
drwxr-xr-x   - root supergroup          0 2022-02-09 11:18 /user/root/.sparkStaging
drwx------   - root supergroup          0 2022-02-10 16:13 /user/root/.staging
drwxr-xr-x   - root supergroup          0 2022-02-13 13:21 /user/root/clickstream_checkpoint
drwxr-xr-x   - root supergroup          0 2022-02-13 13:21 /user/root/clickstream_data
[root@ip-10-0-0-133 ~]# hadoop fs -ls /user/root/clickstream_data
Found 2 items
drwxr-xr-x   - root supergroup          0 2022-02-13 13:21 /user/root/clickstream_data/_spark_metadata
-rw-r--r--   3 root supergroup    1267605 2022-02-13 13:21 /user/root/clickstream_data/part-00000-74523766-d63c-49e5-8045-bd6a78a62d06-c000.json
```



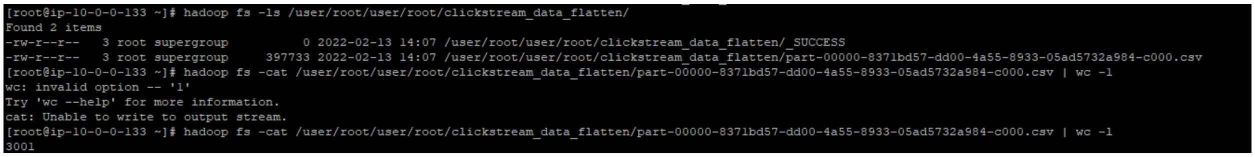
```
[root@ip-10-0-0-133 ~]# hadoop fs -cat  /user/root/clickstream_data/part-00000-74523766-d63c-49e5-8045-bd6a78a62d06-c000.json | wc -l
3000
[root@ip-10-0-0-133 ~]#
```



```
Found 2 items
drwxr-xr-x   - root supergroup          0 2022-02-13 13:51 /user/root/clickstream_data/_spark_metadata
-rw-r--r--   3 root supergroup    1267605 2022-02-13 13:51 /user/root/clickstream_data/part-00000-62b5e35a-b98c-4d44-ae6a-5774b3549b0e-c000.json
[root@ip-10-0-0-133 ~]# clear
[root@ip-10-0-0-133 ~]# vi spark_local_flatten.py
[root@ip-10-0-0-133 ~]# spark2-submit --jars spark-sql-kafka-0-10_2.11-2.3.0.jar spark_local_flatten.py
22/02/13 14:07:03 INFO spark.SparkContext: Running Spark version 2.3.0.cloudera2
22/02/13 14:07:03 INFO spark.SparkContext: Submitted application: Kafka To HDFS
```



```
22/02/13 14:07:12 INFO scheduler.DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 0.290 s
22/02/13 14:07:12 INFO scheduler.DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 0.301916 s
+-----------+-----------+----------+-----------+-----------+--------------------+--------------------+---------------+------------+------------+--------------+---------
+
|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|de545711-3914-445...|fcba68aa-1231-11e...|             No|         Yes|          No|           Yes|     null
|
|   31906387|      2.4.7|       iOS| -64.813749|-133.527040|de545711-3914-445...|a95dd57b-779f-49d...|             No|          No|         Yes|           Yes|     null
|
|   25713677|     3.4.12|   Android|  89.943435| 127.313415|b328829e-17ae-11e...|fcba68aa-1231-11e...|             No|          No|         Yes|            No|     null
|
|   83474293|      3.1.8|   Android| -69.939070| -36.451670|e7bc5fb2-1231-11e...|e1e99492-17ae-11e...|            Yes|          No|         Yes|            No|     null
|
|   63727807|      2.2.9|       iOS|  64.082108| -81.822078|e7bc5fb2-1231-11e...|fcba68aa-1231-11e...|             No|         Yes|         Yes|           Yes|     null
|
|   73737907|     4.3.19|   Android| -18.850508|-116.358375|b328829e-17ae-11e...|e1e99492-17ae-11e...|             No|         Yes|          No|           Yes|     null
|
|   36927433|     3.2.26|       iOS|-84.6857245|-146.507678|de545711-3914-445...|a95dd57b-779f-49d...|            Yes|         Yes|          No|           Yes|     null
|
|   12691783|     3.3.11|   Android| 54.3852925| -37.411814|de545711-3914-445...|e1e99492-17ae-11e...|            Yes|         Yes|          No|            No|     null
|
|   22635021|     4.4.36|       iOS| -31.805500| 150.655650|e7bc5fb2-1231-11e...|a95dd57b-779f-49d...|             No|          No|          No|            No|     null
|
|   23593546|     1.2.16|   Android|  8.8918475| -83.929878|de545711-3914-445...|e1e99492-17ae-11e...|            Yes|          No|         Yes|            No|     null
|
+-----------+-----------+----------+-----------+-----------+--------------------+--------------------+---------------+------------+------------+--------------+---------
+
only showing top 10 rows
```



```
[root@ip-10-0-0-133 ~]# hadoop fs -ls /user/root/user/root/clickstream_data_flatten/
Found 2 items
-rw-r--r--   3 root supergroup          0 2022-02-13 14:07 /user/root/user/root/clickstream_data_flatten/_SUCCESS
-rw-r--r--   3 root supergroup     397733 2022-02-13 14:07 /user/root/user/root/clickstream_data_flatten/part-00000-8371bd57-dd00-4a55-8933-05ad5732a984-c000.csv
[root@ip-10-0-0-133 ~]# hadoop fs -cat /user/root/user/root/clickstream_data_flatten/part-00000-8371bd57-dd00-4a55-8933-05ad5732a984-c000.csv | wc -1
wc: invalid option -- '1'
Try 'wc --help' for more information.
cat: Unable to write to output stream.
[root@ip-10-0-0-133 ~]# hadoop fs -cat /user/root/user/root/clickstream_data_flatten/part-00000-8371bd57-dd00-4a55-8933-05ad5732a984-c000.csv | wc -l
3001
```