

Load data from Kafka to Hadoop

Steps to run the python file to load data from Kafka

1. Use WinScp to transfer the file from local computer to the cluster.

2. Spark submit the python file

```
>> 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
```

4. Use 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

```
>> 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

The data is already loaded into Hdfs using the Python file.

Screenshot of the data

Data loaded in .json format from kafka stream.

```
[hadoop@ip-172-31-32-203 ~]$ hadoop fs -ls /home/hadoop/
Found 3 items
drwxr-xr-x - hadoop hadoop          0 2023-10-23 20:55 /home/hadoop/clickstream_checkpoint
drwxr-xr-x - hadoop hadoop          0 2023-10-23 20:55 /home/hadoop/clickstream_data
drwxr-xr-x - hadoop hadoop          0 2023-10-23 20:58 /home/hadoop/clickstream_data_flatten

[hadoop@ip-172-31-32-203 ~]$ hadoop fs -ls /home/hadoop/clickstream_data
Found 2 items
drwxr-xr-x - hadoop hadoop          0 2023-10-23 20:55 /home/hadoop/clickstream_data/_spark metadata
-rw-r--r-- 1 hadoop hadoop    1267706 2023-10-23 20:55 /home/hadoop/clickstream_data/part-00000-e65e868d-b14a-44d5-b5e3-bd36bafbbdd6-c000.json
[hadoop@ip-172-31-32-203 ~]$ hadoop fs -ls /home/hadoop/clickstream_data/part-00000-e65e868d-b14a-44d5-b5e3-bd36bafbbdd6-c000.json |wc -l
1
[hadoop@ip-172-31-32-203 ~]$ hadoop fs -cat /home/hadoop/clickstream_data/part-00000-e65e868d-b14a-44d5-b5e3-bd36bafbbdd6-c000.json |wc -l
3003
[hadoop@ip-172-31-32-203 ~]$
```

20°C Clear Search [Taskbar Icons: File Explorer, Edge, Chrome, etc.] ENG IN 02:25 24-10-2023

Clean Data file.

```
[hadoop@ip-172-31-32-203 ~]$ hadoop fs -cat /home/hadoop/clickstream_data_flatten/part-00000-c0f48642-94f9-4f9a-a473-c444e5d96456-c000.csv | wc -l
3003
```

customer_id	app_version	OS_version	lat	lon	page_id	button_id	is_button_click	is_page_view	is_scroll_up	is_scroll_down	click_timestamp
265648201	3.2.35	Android	16.4454865	99.902065	de545711-3914-445... fcb68aa-1231-11e...		No	Yes	No	Yes	2020-09-14 09:59:07
31906387	2.4.7	iOS	-64.813749	-133.527040	de545711-3914-445... a95dd57b-779f-49d...		No	No	Yes	Yes	2020-05-16 16:30:21
25713677	3.4.12	Android	89.943435	127.313415	b328029e-17ae-11e... fcb68aa-1231-11e...		No	No	Yes	No	2020-02-09 00:52:13
83474293	3.1.8	Android	-69.939070	-36.451670	e7bc5fb2-1231-11e... e1e99492-17ae-11e...		Yes	No	Yes	No	2020-06-17 10:42:50
63727807	2.2.9	iOS	64.082108	-81.822078	e7bc5fb2-1231-11e... fcb68aa-1231-11e...		No	Yes	Yes	Yes	2020-07-06 02:51:53
73737907	4.3.19	Android	-18.850508	-116.358375	b328029e-17ae-11e... e1e99492-17ae-11e...		No	Yes	No	Yes	2020-04-26 06:18:16
36927433	3.2.26	iOS	-84.6857245	-146.507678	de545711-3914-445... a95dd57b-779f-49d...		Yes	Yes	No	Yes	2020-02-06 10:21:18
12691783	3.3.11	Android	54.3852925	-37.411814	de545711-3914-445... e1e99492-17ae-11e...		Yes	Yes	No	No	2020-08-08 04:23:56
22635021	4.4.36	iOS	-31.805500	50.455650	e7bc5fb2-1231-11e... a95dd57b-779f-49d...		No	No	No	No	2020-08-02 00:33:50
23593546	1.2.16	Android	8.8918475	-83.929878	de545711-3914-445... e1e99492-17ae-11e...		Yes	No	Yes	No	2020-07-23 23:59:19