Load data from Kafka to Hadoop

Reading Kafka stream and writing to json file

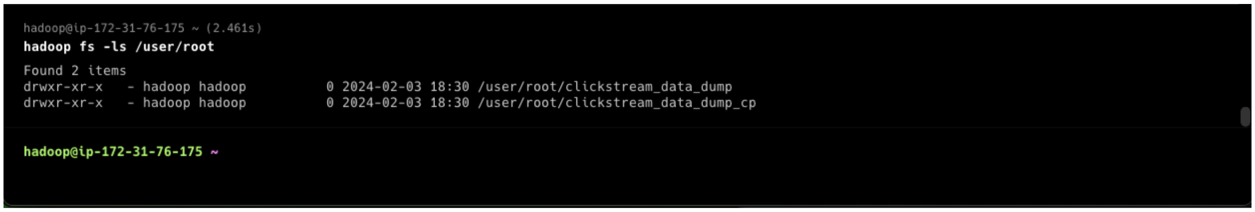
1. pyspark file "spark_kafka_to_local.py" created to read data from kafka and write to hdfs using spark

2. Run the job using below command.

3.
```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5
s3://aws-logs-147098151990-us-east-1/elasticmapreduce/j-
3P5AA3ER77VJN/capstone/spark_kafka_to_local.py
```

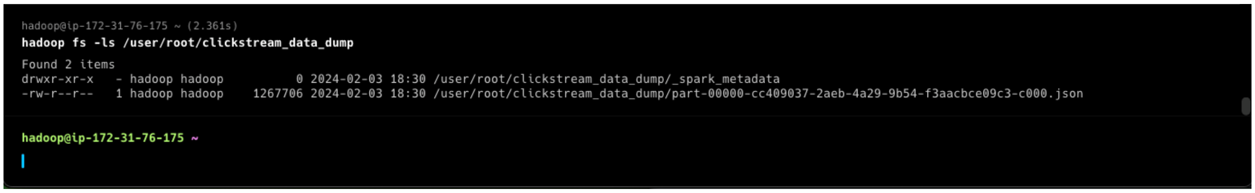
4. Verify Files are created using command.

```
hadoop fs -ls /user/root
```



```
hadoop@ip-172-31-76-175 ~ (2.461s)
hadoop fs -ls /user/root
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2024-02-03 18:30 /user/root/clickstream_data_dump
drwxr-xr-x   - hadoop hadoop          0 2024-02-03 18:30 /user/root/clickstream_data_dump_cp

hadoop@ip-172-31-76-175 ~
```



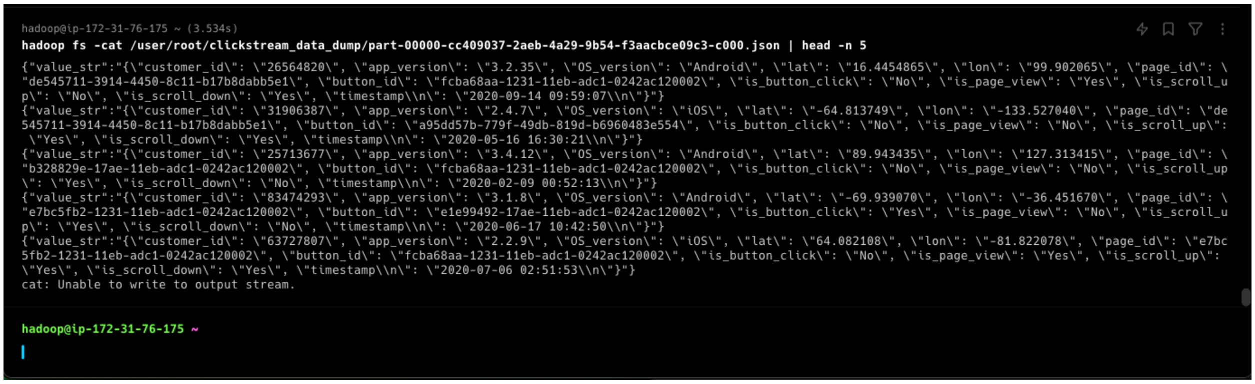
```
hadoop@ip-172-31-76-175 ~ (2.361s)
hadoop fs -ls /user/root/clickstream_data_dump
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2024-02-03 18:30 /user/root/clickstream_data_dump/_spark_metadata
-rw-r--r--   1 hadoop hadoop    1267706 2024-02-03 18:30 /user/root/clickstream_data_dump/part-00000-cc409037-2aeb-4a29-9b54-f3aacbce09c3-c000.json

hadoop@ip-172-31-76-175 ~
```

5. Verify top 5 records created in hdfs

```
hadoop fs -cat /user/root/clickstream_data_dump/part-00000-38bbb9c2-20ac-
43a8-a27f-7f875beb8011-c000.json | head -n 5
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

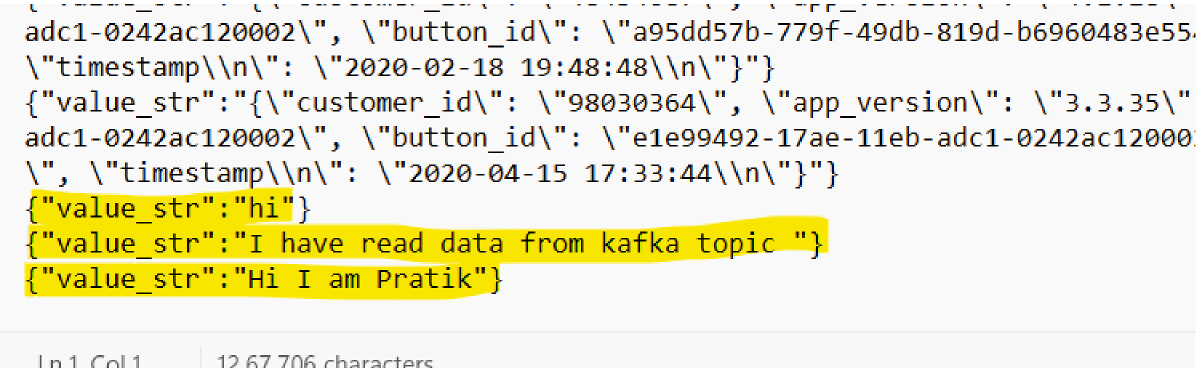


```
hadoop@ip-172-31-76-175 ~ (3.534s)
hadoop fs -cat /user/root/clickstream_data_dump/part-00000-cc409037-2aeb-4a29-9b54-f3aacbce09c3-c000.json | head -n 5
{"value_str":"{\"customer_id\": \"26564820\", \"app_version\": \"3.2.35\", \"OS_version\": \"Android\", \"lat\": \"16.4454865\", \"lon\": \"99.902065\", \"page_id\": \
"de545711-3914-4450-8c11-b17b8dabb5e1\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_click\": \"No\", \"is_page_view\": \"Yes\", \"is_scroll_u
p\": \"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\", \"page_id\": \"de
545711-3914-4450-8c11-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\"}"}
{"value_str":"{\"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-adc1-0242ac120002\", \"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\"}"}
{"value_str":"{\"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-adc1-0242ac120002\", \"is_button_click\": \"Yes\", \"is_page_view\": \"No\", \"is_scroll_u
p\": \"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\", \"page_id\": \"e7bc
5fb2-1231-11eb-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\"}"}
cat: Unable to write to output stream.

hadoop@ip-172-31-76-175 ~
```

Reading json file from hdfs and transforming to a csv file

1. Invalid data need cleanup in python to take valid customer info



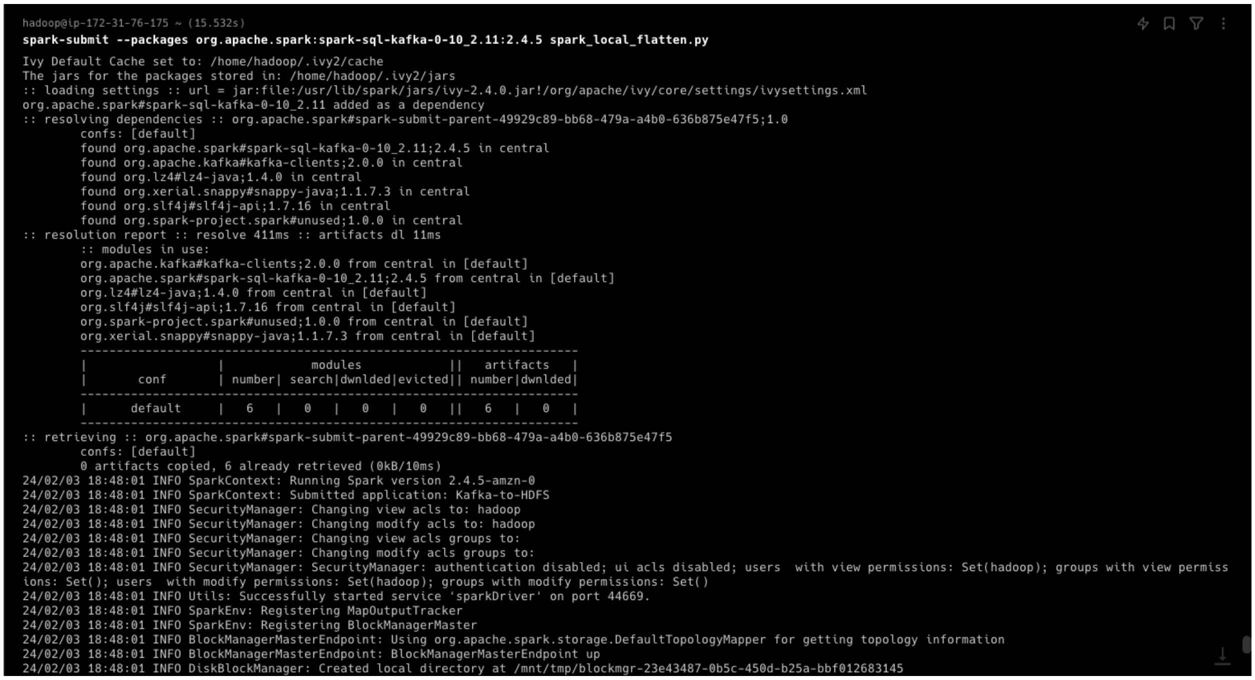
```
adc1-0242ac120002\", \"button_id\": \"a95dd57b-779f-49db-819d-b6960483e55
\"timestamp\\n\": \"2020-02-18 19:48:48\\n\"}"}
{"value_str":"{\"customer_id\": \"98030364\", \"app_version\": \"3.3.35\"
adc1-0242ac120002\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac12000
\", \"timestamp\\n\": \"2020-04-15 17:33:44\\n\"}"}
{"value_str":"hi"}
{"value_str":"I have read data from kafka topic "}
{"value_str":"Hi I am Pratik"}
```

Ln 1, Col 1 | 12,67,706 characters

2. pyspark file "spark_local_flatten.py" created to read json data and convert to csv format

```
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5
spark_local_flatten.py
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved

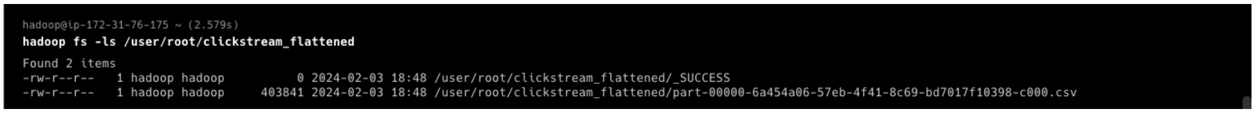


```
hadoop@ip-172-31-76-175 ~ (15.532s)
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_local_flatten.py

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-49929c89-bb68-479a-a4b0-636b875e47f5;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 411ms :: artifacts dl 11ms
        :: 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   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-49929c89-bb68-479a-a4b0-636b875e47f5
        confs: [default]
        0 artifacts copied, 6 already retrieved (0kB/10ms)
24/02/03 18:48:01 INFO SparkContext: Running Spark version 2.4.5-amzn-0
24/02/03 18:48:01 INFO SparkContext: Submitted application: Kafka-to-HDFS
24/02/03 18:48:01 INFO SecurityManager: Changing view acls to: hadoop
24/02/03 18:48:01 INFO SecurityManager: Changing modify acls to: hadoop
24/02/03 18:48:01 INFO SecurityManager: Changing view acls groups to:
24/02/03 18:48:01 INFO SecurityManager: Changing modify acls groups to:
24/02/03 18:48:01 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with view permiss
ions: Set(); users  with modify permissions: Set(hadoop); groups with modify permissions: Set()
24/02/03 18:48:01 INFO Utils: Successfully started service 'sparkDriver' on port 44669.
24/02/03 18:48:01 INFO SparkEnv: Registering MapOutputTracker
24/02/03 18:48:01 INFO SparkEnv: Registering BlockManagerMaster
24/02/03 18:48:01 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
24/02/03 18:48:01 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/02/03 18:48:01 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-23e43487-0b5c-450d-b25a-bbf012683145
```

3. Verify csv file is created

```
hadoop fs -ls /user/root/clickstream_flattened
```

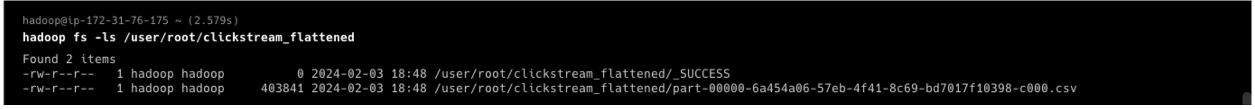


```
hadoop@ip-172-31-76-175 ~ (2.579s)
hadoop fs -ls /user/root/clickstream_flattened

Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-02-03 18:48 /user/root/clickstream_flattened/_SUCCESS
-rw-r--r--   1 hadoop hadoop     403841 2024-02-03 18:48 /user/root/clickstream_flattened/part-00000-6a454a06-57eb-4f41-8c69-bd7017f10398-c000.csv
```

4. Verify top 5 records created in hdfs

```
hadoop fs -cat clickstream_flattened/part-00000-65875598-48b5-4f6b-b2df-e8e32020a98a-c000.csv | head -n 5
```



```
hadoop@ip-172-31-76-175 ~ (2.579s)
hadoop fs -ls /user/root/clickstream_flattened

Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-02-03 18:48 /user/root/clickstream_flattened/_SUCCESS
-rw-r--r--   1 hadoop hadoop     403841 2024-02-03 18:48 /user/root/clickstream_flattened/part-00000-6a454a06-57eb-4f41-8c69-bd7017f10398-c000.csv
```

© Copyright 2020. upGrad Education Pvt. Ltd. All rights reserved