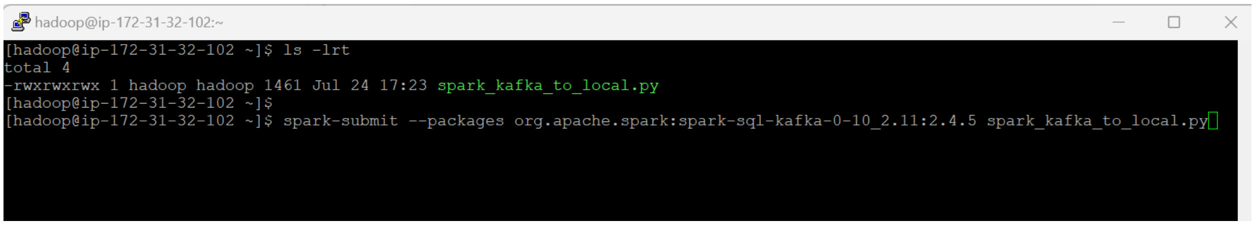
# Load data from Kafka to Hadoop

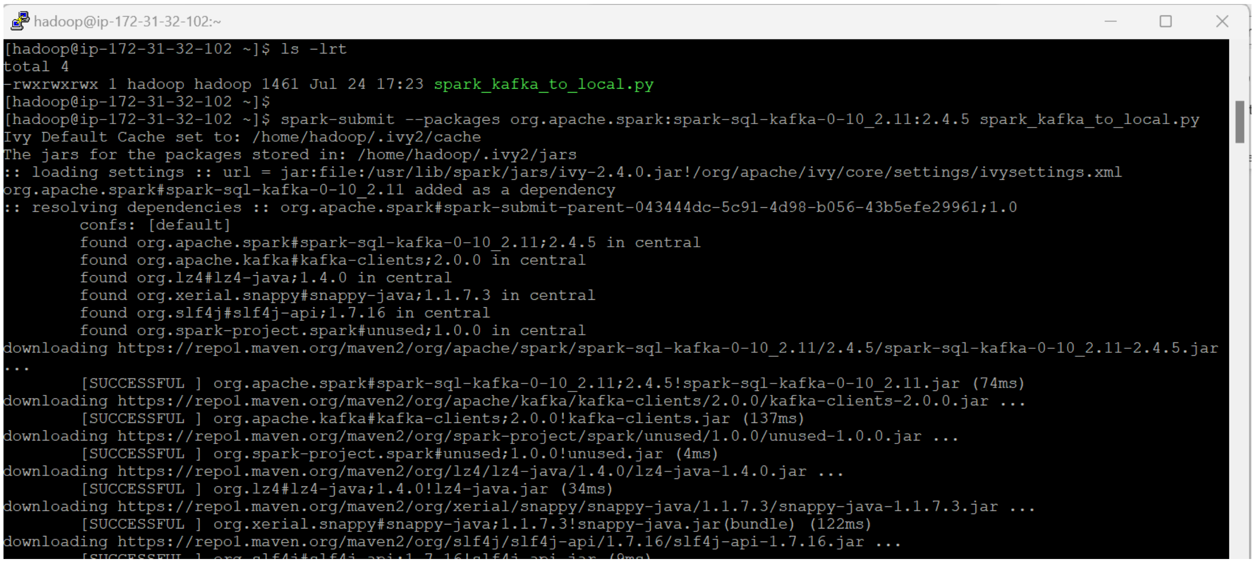
## Steps to run the python file to load data from Kafka

1. **Running the file ‘spark_kafka_to_local.py’ using ‘spark-submit’ command to load the data from Kafka topic to HDFS:**

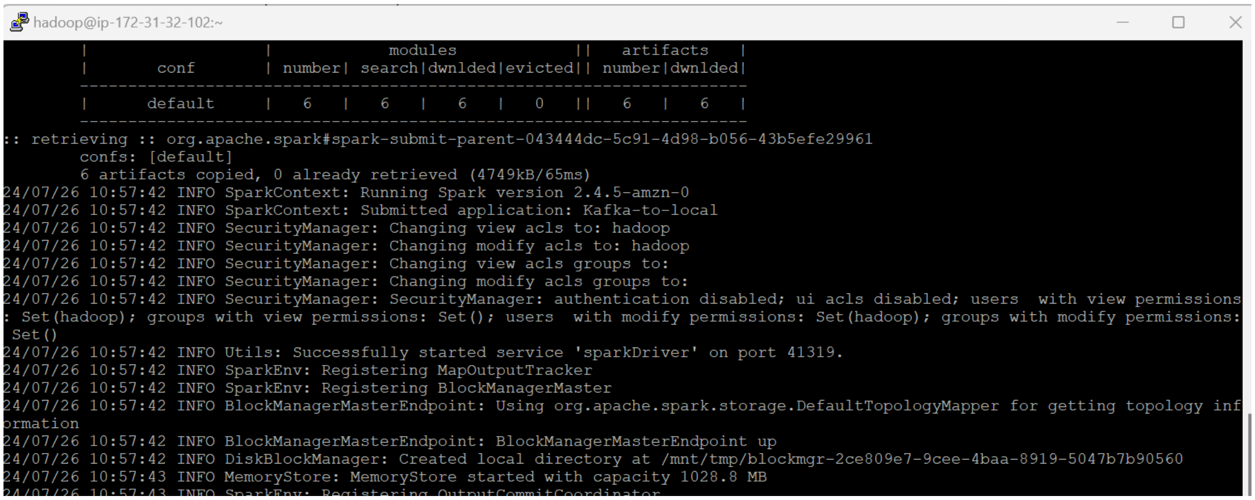
export SPARK_KAFKA_VERSION=0.10
spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_kafka_to_local.py



```
[hadoop@ip-172-31-32-102 ~]$ ls -lrt
total 4
-rwxrwxrwx 1 hadoop hadoop 1461 Jul 24 17:23 spark_kafka_to_local.py
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_kafka_to_local.py
```



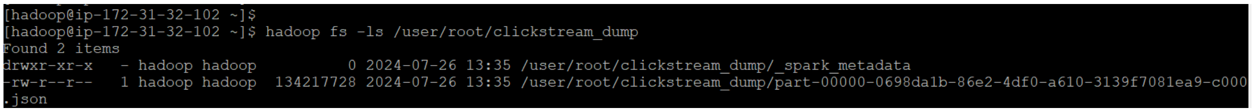
```
[hadoop@ip-172-31-32-102 ~]$ ls -lrt
total 4
-rwxrwxrwx 1 hadoop hadoop 1461 Jul 24 17:23 spark_kafka_to_local.py
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_kafka_to_local.py
Ivy Default Cache set to: /home/hadoop/.ivy2/cache
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-043444dc-5c91-4d98-b056-43b5efe29961;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
downloading https://repo1.maven.org/maven2/org/apache/spark/spark-sql-kafka-0-10_2.11/2.4.5/spark-sql-kafka-0-10_2.11-2.4.5.jar
...
        [SUCCESSFUL ] org.apache.spark#spark-sql-kafka-0-10_2.11;2.4.5!spark-sql-kafka-0-10_2.11.jar (74ms)
downloading https://repo1.maven.org/maven2/org/apache/kafka/kafka-clients/2.0.0/kafka-clients-2.0.0.jar ...
        [SUCCESSFUL ] org.apache.kafka#kafka-clients;2.0.0!kafka-clients.jar (137ms)
downloading https://repo1.maven.org/maven2/org/spark-project/spark/unused/1.0.0/unused-1.0.0.jar ...
        [SUCCESSFUL ] org.spark-project.spark#unused;1.0.0!unused.jar (4ms)
downloading https://repo1.maven.org/maven2/org/lz4/lz4-java/1.4.0/lz4-java-1.4.0.jar ...
        [SUCCESSFUL ] org.lz4#lz4-java;1.4.0!lz4-java.jar (34ms)
downloading https://repo1.maven.org/maven2/org/xerial/snappy/snappy-java/1.1.7.3/snappy-java-1.1.7.3.jar ...
        [SUCCESSFUL ] org.xerial.snappy#snappy-java;1.1.7.3!snappy-java.jar(bundle) (122ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.16/slf4j-api-1.7.16.jar ...
        [SUCCESSFUL ] org.slf4j#slf4j-api;1.7.16!slf4j-api.jar (9ms)
```



```
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   6   |   6   |   6   |   0   ||   6   |   6   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-parent-043444dc-5c91-4d98-b056-43b5efe29961
        confs: [default]
        6 artifacts copied, 0 already retrieved (4749kB/65ms)
24/07/26 10:57:42 INFO SparkContext: Running Spark version 2.4.5-amzn-0
24/07/26 10:57:42 INFO SparkContext: Submitted application: Kafka-to-local
24/07/26 10:57:42 INFO SecurityManager: Changing view acls to: hadoop
24/07/26 10:57:42 INFO SecurityManager: Changing modify acls to: hadoop
24/07/26 10:57:42 INFO SecurityManager: Changing view acls groups to:
24/07/26 10:57:42 INFO SecurityManager: Changing modify acls groups to:
24/07/26 10:57:42 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions
: Set(hadoop); groups with view permissions: Set(); users  with modify permissions: Set(hadoop); groups with modify permissions:
 Set()
24/07/26 10:57:42 INFO Utils: Successfully started service 'sparkDriver' on port 41319.
24/07/26 10:57:42 INFO SparkEnv: Registering MapOutputTracker
24/07/26 10:57:42 INFO SparkEnv: Registering BlockManagerMaster
24/07/26 10:57:42 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology inf
ormation
24/07/26 10:57:42 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
24/07/26 10:57:42 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-2ce809e7-9cee-4baa-8919-5047b7b90560
24/07/26 10:57:43 INFO MemoryStore: MemoryStore started with capacity 1028.8 MB
24/07/26 10:57:43 INFO SparkEnv: Registering OutputCommitCoordinator
```

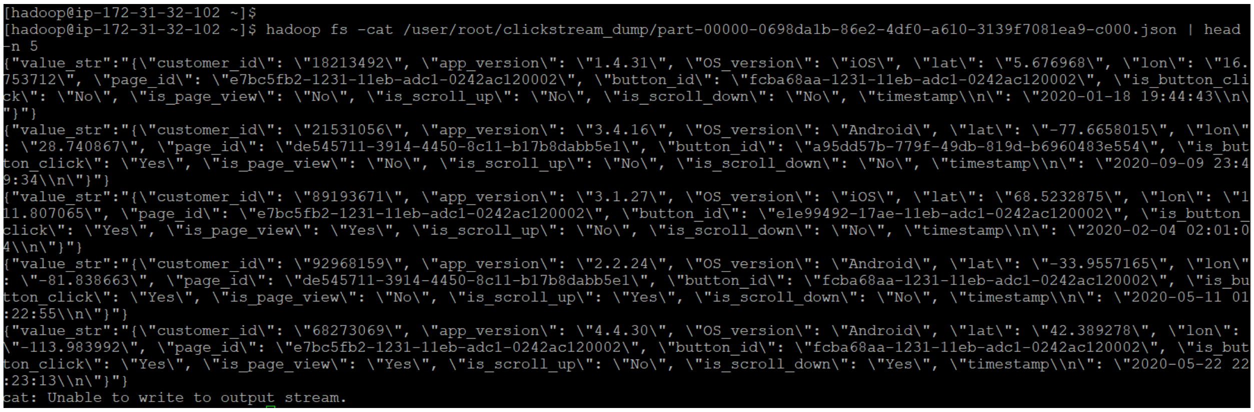
## 2. Validating the directory and contents of files created by Pyspark file in HDFS:

hadoop fs -ls /user/root/clickstream_dump



```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/clickstream_dump
Found 2 items
drwxr-xr-x   - hadoop hadoop          0 2024-07-26 13:35 /user/root/clickstream_dump/_spark_metadata
-rw-r--r--   1 hadoop hadoop  134217728 2024-07-26 13:35 /user/root/clickstream_dump/part-00000-0698da1b-86e2-4df0-a610-3139f7081ea9-c000
.json
```

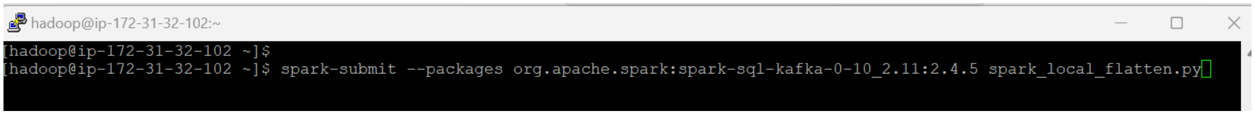
hadoop fs -cat /user/root/clickstream_dump/part-00000-0698da1b-86e2-4df0-a610-3139f7081ea9-c000.json | head -n 5



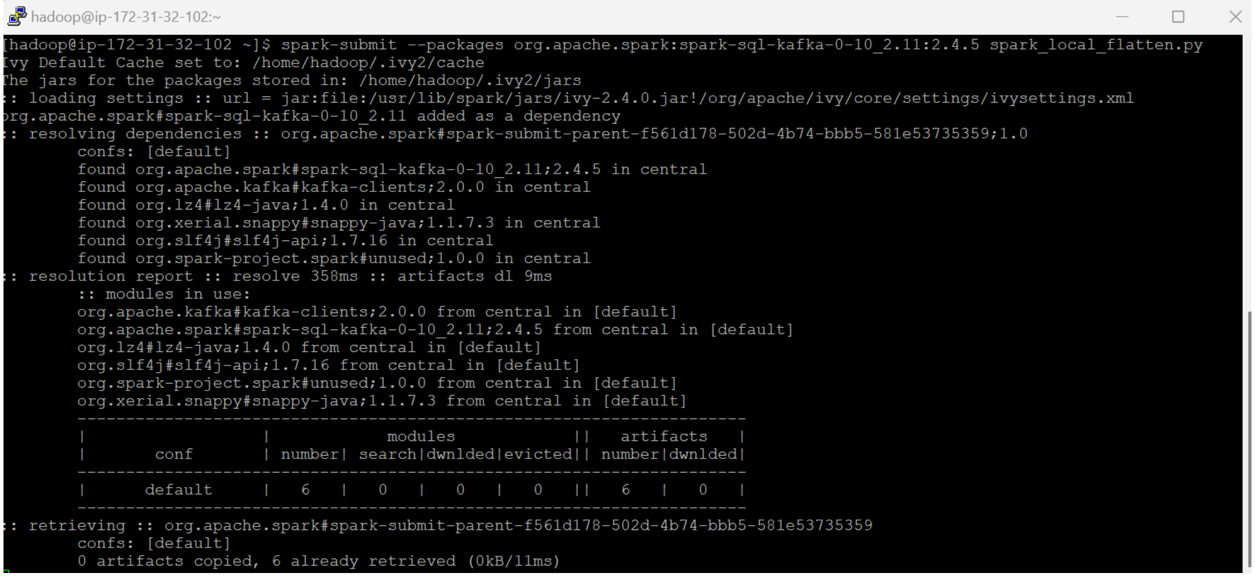
```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -cat /user/root/clickstream_dump/part-00000-0698da1b-86e2-4df0-a610-3139f7081ea9-c000.json | head
-n 5
{"value_str":"{\"customer_id\": \"18213492\", \"app_version\": \"1.4.31\", \"OS_version\": \"iOS\", \"lat\": \"5.676968\", \"lon\": \"16.
753712\", \"page_id\": \"e7bc5fb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_button_cli
ck\": \"No\", \"is_page_view\": \"No\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-01-18 19:44:43\\n\
"}"}
{"value_str":"{\"customer_id\": \"21531056\", \"app_version\": \"3.4.16\", \"OS_version\": \"Android\", \"lat\": \"-77.6658015\", \"lon\"
: \"28.740867\", \"page_id\": \"de545711-3914-4450-8c11-b17b8dabb5e1\", \"button_id\": \"a95dd57b-779f-49db-819d-b6960483e554\", \"is_but
ton_click\": \"Yes\", \"is_page_view\": \"No\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-09-09 23:4
9:34\\n\"}"}
{"value_str":"{\"customer_id\": \"89193671\", \"app_version\": \"3.1.27\", \"OS_version\": \"iOS\", \"lat\": \"68.5232875\", \"lon\": \"1
11.807065\", \"page_id\": \"e7bc5fb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"e1e99492-17ae-11eb-adc1-0242ac120002\", \"is_button_
click\": \"Yes\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-02-04 02:01:0
4\\n\"}"}
{"value_str":"{\"customer_id\": \"92968159\", \"app_version\": \"2.2.24\", \"OS_version\": \"Android\", \"lat\": \"-33.9557165\", \"lon\"
: \"-81.838663\", \"page_id\": \"de545711-3914-4450-8c11-b17b8dabb5e1\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_bu
tton_click\": \"Yes\", \"is_page_view\": \"No\", \"is_scroll_up\": \"Yes\", \"is_scroll_down\": \"No\", \"timestamp\\n\": \"2020-05-11 01
:22:55\\n\"}"}
{"value_str":"{\"customer_id\": \"68273069\", \"app_version\": \"4.4.30\", \"OS_version\": \"Android\", \"lat\": \"42.389278\", \"lon\":
\"-113.983992\", \"page_id\": \"e7bc5fb2-1231-11eb-adc1-0242ac120002\", \"button_id\": \"fcba68aa-1231-11eb-adc1-0242ac120002\", \"is_but
ton_click\": \"Yes\", \"is_page_view\": \"Yes\", \"is_scroll_up\": \"No\", \"is_scroll_down\": \"Yes\", \"timestamp\\n\": \"2020-05-22 22
:23:13\\n\"}"}
cat: Unable to write to output stream.
```

## 3. Executing Python file ‘spark_local_flatten.py’ to clean the loaded Kafka data (i.e. json file in HDFS) to a more structured format (.csv):

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



```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 spark_local_flatten.py
```

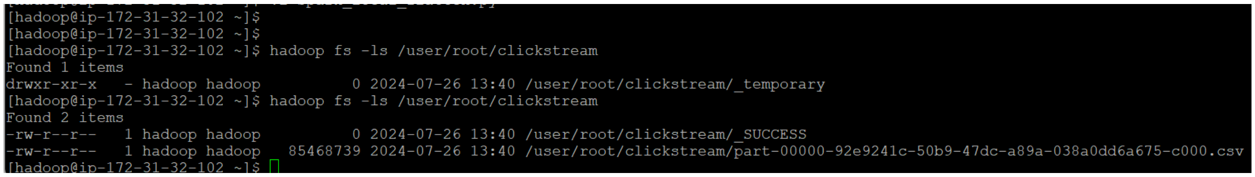


hadoop@ip-172-31-32-102:~

```
[hadoop@ip-172-31-32-102 ~]$ 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-f561d178-502d-4b74-bbb5-581e53735359;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 358ms :: artifacts dl 9ms
        :: 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-f561d178-502d-4b74-bbb5-581e53735359
        confs: [default]
        0 artifacts copied, 6 already retrieved (0kB/11ms)
```

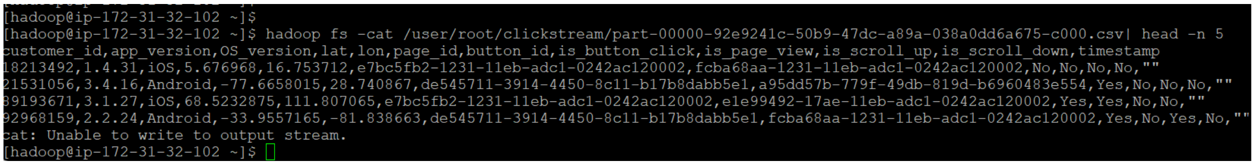
## 4. Validating the directory and contents of flattened file(.csv from .json) created by Pyspark file in HDFS:

hadoop fs -ls /user/root/clickstream



```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/clickstream
Found 1 items
drwxr-xr-x   - hadoop hadoop          0 2024-07-26 13:40 /user/root/clickstream/_temporary
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -ls /user/root/clickstream
Found 2 items
-rw-r--r--   1 hadoop hadoop          0 2024-07-26 13:40 /user/root/clickstream/_SUCCESS
-rw-r--r--   1 hadoop hadoop   85468739 2024-07-26 13:40 /user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv
[hadoop@ip-172-31-32-102 ~]$
```

hadoop fs -cat /user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv| head -n 5



```
[hadoop@ip-172-31-32-102 ~]$
[hadoop@ip-172-31-32-102 ~]$ hadoop fs -cat /user/root/clickstream/part-00000-92e9241c-50b9-47dc-a89a-038a0dd6a675-c000.csv| head -n 5
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
18213492,1.4.31,iOS,5.676968,16.753712,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,No,No,""
21531056,3.4.16,Android,-77.6658015,28.740867,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,Yes,No,No,No,""
89193671,3.1.27,iOS,68.5232875,111.807065,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,Yes,No,No,""
92968159,2.2.24,Android,-33.9557165,-81.838663,de545711-3914-4450-8c11-b17b8dabb5e1,fcba68aa-1231-11eb-adc1-0242ac120002,Yes,No,Yes,No,""
cat: Unable to write to output stream.
[hadoop@ip-172-31-32-102 ~]$
```