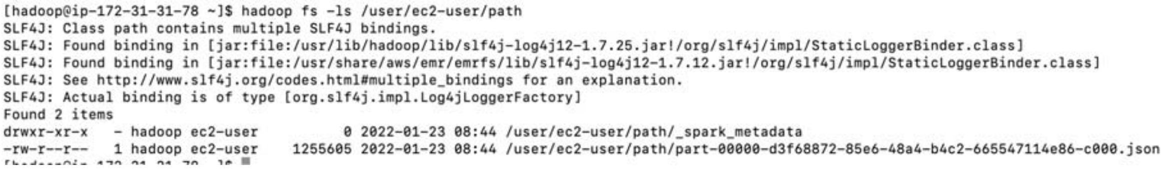
Load data from Kafka to Hadoop

Task: Write a job to consume clickstream data from Kafka and ingest to Hadoop

Steps to run the python file to load data from Kafka

1. Log into the machine where spark is installed.
2. Change user from ec2-user to root.
    a. Command: `sudo -i`
3. Write a python script that *consumes the clickstream data from Kafka and saves it locally*.
4. Command: `vi spark_kafka_to_local.py` (This file is added in the submission folder with the same name)
5. Run the submit-spark command with the above mentioned file
    a. Command: `spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2 spark_kafka_to_local.py`
    b. Note (everywhere): `org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2` version is used because the machine where the code is running is having spark 3.1.2, if you're using some other version, then use respective version
6. Verify the data stored in hadoop file system
    a. Command: `hadoop fs -ls /user/ec2-user/path`



```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls /user/ec2-user/path
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
drwxr-xr-x   - hadoop ec2-user          0 2022-01-23 08:44 /user/ec2-user/path/_spark_metadata
-rw-r--r--   1 hadoop ec2-user    1255605 2022-01-23 08:44 /user/ec2-user/path/part-00000-d3f68872-85e6-48a4-b4c2-665547114e86-c000.json
[hadoop@ip-172-31-31-78 ~]$
```

7. Write a python script that converts the clickstream json data to CSV format
    a. Command: `vi spark_local_flatten.py` (This file is added in the submission folder with the same name)
8. Run the submit-spark command with above script
    a. Command: `spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2 spark_local_flatten.py`
9. Verify that the CSV file is generated
    a. Command: `hadoop fs -ls user/ec2-user/ClickStreamData`
10. Cat the generated file for verification
    a. Command: `hadoop fs -cat user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv`

Steps to load the data into Hadoop

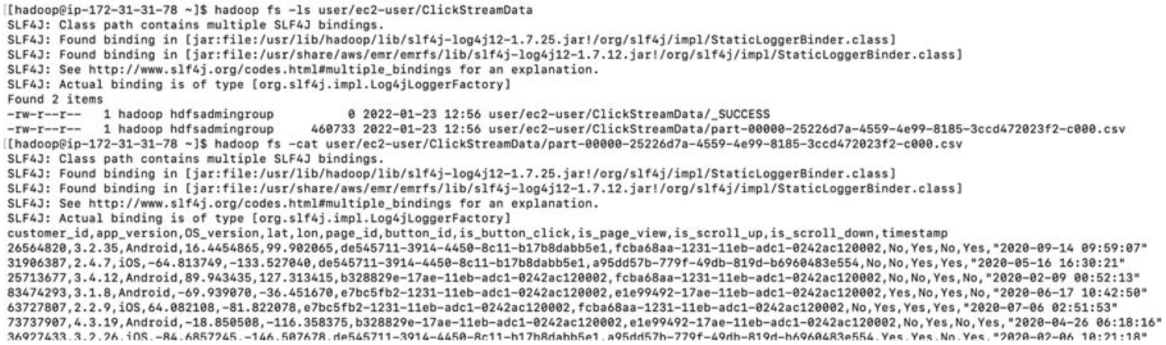
1. Transfer the data from local fs to hdfs can be done through following command:
    a. hdfs dfs -put $LOCAL_PATH/ClickStreamData /user/root/ClickStreamData

Note: Since the outputs of the above scripts are directly stored

Screenshot of the data

Output of command:

- hadoop fs -ls user/ec2-user/ClickStreamData
    - Output file can be seen
- hadoop fs -cat user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv
    - CSV ClickStream data can be seen



```
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -ls user/ec2-user/ClickStreamData
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2022-01-23 12:56 user/ec2-user/ClickStreamData/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup     460733 2022-01-23 12:56 user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv
[hadoop@ip-172-31-31-78 ~]$ hadoop fs -cat user/ec2-user/ClickStreamData/part-00000-25226d7a-4559-4e99-8185-3ccd472023f2-c000.csv
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/share/aws/emr/emrfs/lib/slf4j-log4j12-1.7.12.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
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,"2020-09-14 09:59:07"
31906387,2.4.7,iOS,-64.813749,-133.527040,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,No,No,Yes,Yes,"2020-05-16 16:30:21"
25713677,3.4.12,Android,89.943435,127.313415,b328829e-17ae-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,No,Yes,No,"2020-02-09 00:52:13"
83474293,3.1.8,Android,-69.939070,-36.451670,e7bc5fb2-1231-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,Yes,No,Yes,No,"2020-06-17 10:42:50"
63727807,2.2.9,iOS,64.082108,-81.822078,e7bc5fb2-1231-11eb-adc1-0242ac120002,fcba68aa-1231-11eb-adc1-0242ac120002,No,Yes,Yes,Yes,"2020-07-06 02:51:53"
73737907,4.3.19,Android,-18.850508,-116.358375,b328829e-17ae-11eb-adc1-0242ac120002,e1e99492-17ae-11eb-adc1-0242ac120002,No,Yes,No,Yes,"2020-04-26 06:18:16"
36927433,3.2.26,iOS,-84.6857245,-146.507678,de545711-3914-4450-8c11-b17b8dabb5e1,a95dd57b-779f-49db-819d-b6960483e554,Yes,Yes,No,Yes,"2020-02-06 10:21:18"
```