Load data from Kafka to Hadoop

Reading clickstream data from Kafka and writing the data as .json file in HDFS local directory.

1. **Python file "spark_kafka_to_local.py" created to read data from kafka server and write to HDFS as .json file. The python file is saved in S3 bucket.**

2. **Executing the file to save streaming data into HDFS.**

   spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.0 s3://sudhesh-elasticmapreduce/capstone_project/spark_kafka_to_local.py

```
[hadoop@ip-172-31-44-42 ~]$ hadoop fs -ls /user/hadoop/
Found 1 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2025-02-12 05:23 /user/hadoop/bookings
[hadoop@ip-172-31-44-42 ~]$ spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.0 s3://sudhesh-elasticmapreduce/capstone_project/spark_kafka_to_local.py
Feb 12, 2025 5:42:46 AM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/Log4jHotPatchFat.jar does not exist at the configured location
```

3. **Verifying streaming data files on HDFS.**

   hadoop fs -ls /user/hadoop/kafka_stream/kafka_clickstream

```
[hadoop@ip-172-31-44-42 ~]$ hadoop fs -ls /user/hadoop/kafka_stream/kafka_clickstream
Found 26 items
drwxr-xr-x   - hadoop hdfsadmingroup          0 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/_spark_metadata
-rw-r--r--   1 hadoop hdfsadmingroup     668149 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-0010cb59-2aaa-4e89-b83b-ae3113c3b9b0-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     143237 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-03131073-7bdc-4183-bf65-23971f3deed6-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup    1073973 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-0769e910-726c-45b2-859b-6652b2fee93d-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     104730 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-08965eae-e27c-401d-8734-304d3e3f0ccb-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     102631 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-0a67d8ee-b906-48be-bfe9-4af2a3064073-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     856124 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-204b2a63-6d8b-41de-996c-7f8949303c2b-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     381690 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-21157386-2bd2-4bb1-905b-1d2165acb3d2-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     194031 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-29681636-b0cc-4793-99ff-55b86a9d5ff5-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     113643 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-2a03c738-4bcc-4848-b27b-65a5df9b9f82-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     123819 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-2e6e7051-815d-4ed0-bbfb-2103a3f28da4-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     661379 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-3886f75e-312a-47cc-ab06-111ceb75cea5-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup    1013180 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-47f9f2a6-32ce-4ac9-8fdd-5b7d09dfa262-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     385316 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-53ed2dc1-c516-4a8b-b476-31be324e5f84-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     401468 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-6104aebc-d09d-4a48-a938-96240adeb3bb-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     363685 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-66d9decf-3c1f-43cd-8551-fe899f79a70e-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup    1055430 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-7380ef5d-f1c9-4074-9352-9c13ff947af8-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup   24695100 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-79b5a6b0-e805-48aa-9195-fb95c855efa3-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     629500 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-7b8504ef-af51-4ac2-b815-9f67d299651d-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     159275 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-b71bcfc6-3f1c-4146-9c9f-00a4d89b58e9-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     128476 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-b73ef601-29e4-41ca-9db6-83a9e7ab8043-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     267793 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-b9e45e96-66e7-40cd-b895-6452b5f0f021-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     298748 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-baaf75f1-8f3a-4160-97d3-a143f7fcccf6-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     386228 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-efd5bec9-a25a-4410-a700-f07002a71729-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup 1330899666 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-f71f7827-b260-4cc3-9375-0f68f1c9519a-c000.json
-rw-r--r--   1 hadoop hdfsadmingroup     578298 2025-02-12 05:54 /user/hadoop/kafka_stream/kafka_clickstream/part-00000-fb328e44-5b3e-4db3-99fc-0f2927ac1920-c000.json
[hadoop@ip-172-31-44-42 ~]$
```

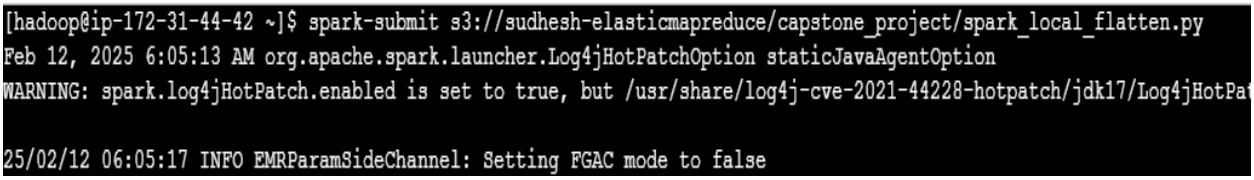
4. **Verifying records in json file.**

hadoop fs -cat /user/hadoop/kafka_stream/kafka_clickstream/part-00000-03131073-7bdc-4183-bf65-23971f3deed6-c000.json | head -5

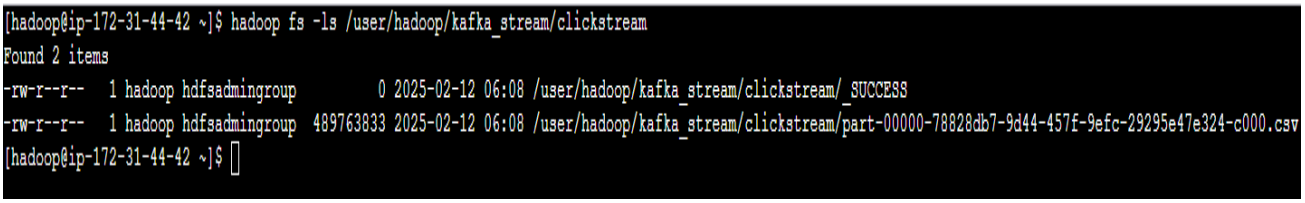
Reading the .json files from HDFS and transforming into a structured .csv file and saving in HDFS.

1. **Python file "spark_local_flatten.py" to read .json file from HDFS, transforming and saving as .csv file. The python file is saved in S3 bucket.**

2. **Executing the file to save .csv file into HDFS.**
   spark-submit s3://sudhesh-elasticmapreduce/capstone_project/spark_local_flatten.py



3. **Verifying the .csv file in specified directory.**
   hadoop fs -ls /user/hadoop/kafka_stream/clickstream



4. **Verifying records in .csv file.**
   hadoop fs -cat /user/hadoop/kafka_stream/clickstream/part-00000-78828db7-9d44-457f-9efc-29295e47e324-c000.csv | head -5