

Streaming homework.

First, we create a kafka topic:

```
[maria_dev@sandbox-hdp bin]$ ./kafka-topics.sh --create --zookeeper sandbox-hdp.hortonworks.com --partitions 1 --replication-factor 1 --topic trainTopic
Created topic "trainTopic".
[maria_dev@sandbox-hdp bin]$
```

Then we start our daemon:

```
[maria_dev@sandbox-hdp bin]$ ./spark-submit --master yarn-client --driver-memory 4G --number-executors 4 --executor-memory 4G /home/maria_dev/streamingProducer.jar -topic trainTopic -url sandbox-hdp.hortonworks.com:6667 -nThreads 4 -filePath /home/maria_dev/train.csv
```

I have implemented a custom MessageCallback that gets triggered each time a line gets send into a topic by a KafkaProducer. Here's sample output:

```
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2639, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2640, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2641, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2642, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2643, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2644, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2645, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2646, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2647, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2648, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2649, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2650, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2651, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2652, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2653, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2654, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2655, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2656, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2657, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2658, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2659, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2660, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2661, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2662, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2663, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2664, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2665, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2666, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2667, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2668, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2669, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2670, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2671, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2672, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2673, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2674, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2675, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2676, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2677, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2678, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2679, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2680, Partition : 0
18/11/21 16:12:46 INFO MessageCallback$: Topic : trainTopic, Offset : 2681, Partition : 0
18/11/21 16:12:46 INFO MessageC
```

Next we need to read from kafka topic and write to HDFS, here's the command to do it without batching:

```
[maria_dev@sandbox-hdp bin]$ ./spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.2 --master yarn-client --driver-memory 1G --num-executors 4 --executor-memory 1G /home/maria_dev/streamingConsumer.jar -topic trainTopic -url sandbox-hdp.hortonworks.com:6667 -filePath /user/maria_dev/consumerFolder -fileFormat csv -doBatch false
```

And with batching:

```
[maria_dev@sandbox-hdp bin]$ ./spark-submit --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.3.2 --master yarn-client --driver-memory 1G --num-executors 4 --executor-memory 1G /home/maria_dev/streamingConsumer.jar -topic trainTopic -url sandbox-hdp.hortonworks.com:6667 -filePath /user/maria_dev/consumerFolder -fileFormat csv -doBatch true
```

I have tested both variants and they perform roughly the same. On a 100 mb dataset both took about 10 seconds.




On a 2.6 GB dataset both took about 91 seconds. Here's the batching consumer result screenshot:

```
18/11/22 08:57:24 INFO ContextCleaner: Cleaned accumulator 15
18/11/22 08:57:24 INFO ContextCleaner: Cleaned accumulator 11
18/11/22 08:57:24 INFO ContextCleaner: Cleaned accumulator 28
18/11/22 08:57:24 INFO ContextCleaner: Cleaned accumulator 30
18/11/22 08:57:24 INFO Consumers$: Elapsed: 91.366698948
18/11/22 08:57:24 INFO AbstractConnector: Stopped Spark@6de0f580{HTTP/1.1,[http/1.1]}{0.0.0.0:4040}
18/11/22 08:57:24 INFO SparkUI: Stopped Spark web UI at http://sandbox-hdp.hortonworks.com:4040
18/11/22 08:57:24 INFO BlockManagerInfo: Removed broadcast 0 piece0 on sandbox-hdp.hortonworks.com:4040
```

And here's the streaming consumer result screenshot:

```
spark-kafka-source-91a453f2-6b99-4bef-9833-543162e91053--1622573403-driver-0] Resetting offset for partition trainTopic-0 to offset 25706789.
18/11/22 09:04:39 INFO Consumers$: Elapsed: 90.093134952
18/11/22 09:04:39 INFO MicroBatchExecutor: Streaming query made progress
```

Here's the dataset written to hdfs:

		
 _SUCCESS	0.1 kB	2018-11-22 11:57
 part-00000-972d2ade-4fcf-4e05-997d-04...	2.6 GB	2018-11-22 11:57