# 4. Spark Streaming. Sinks

ДЗ - повторить действия как на уроке, только со своими данными, использовать свою схему, свой топик в кафке, попробовать как складываются файлы в паркет, в csv, изменить на json загружать в кафку, использовать другие режимы апдате или комплит, не аппенд. Посмотреть каким ещё образом можно складывать файлы паркет, при этом остановить поток а потом запустить его ещё раз.

Скопируем подготовленный файл «drake_data.json» на удаленный сервер с помощью команды `scp`. Эта команда запускается на локальном компьютере

scp -i ~/.ssh/id_rsa_student898_2 -r drake_data.json student898_2@37.139.41.176:~/for_stream

Подключаемся и проверяем, что файл drake_data.json загрузился.

ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176

ls for_stream

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
igor@igor-MS-7808:~$ scp -i ~/.ssh/id_rsa_student898_2 -r drake_data.json student898_2@37.139.41.176:~/for_stream
drake_data.json                                                                                    100%  827KB   9.5MB/s   00:00
igor@igor-MS-7808:~$ ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176
Last login: Fri Jan 21 20:30:08 2022 from 109-252-19-10.nat.spd-mgts.ru
[student898_2@bigdataanalytics-worker-3 ~]$ ls  for_stream
archive.csv  data.json    drake_data.json  iris.json          product_list2.csv  product_list4.csv
data.csv     dataset.csv  file1.json       product_list1.csv  product_list3.csv  product_list.csv
[student898_2@bigdataanalytics-worker-3 ~]$
```

less for_stream/drake_data.json

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[{"album": "Certified Lover Boy", "lyrics_title": "Certified Lover Boy* Lyrics", "lyrics_url": "https://genius.com/Drake-certified-lover-boy-lyrics",
"lyrics": "Lyrics from CLB Merch\n\n[Verse]\nPut my feelings on ice\nAlways been a gem\nCertified lover boy, somehow still heartless\nHeart is only ge
ttin' colder", "track_views": "8.7K"}, {"album": "Certified Lover Boy", "lyrics_title": "Like I\u2019m Supposed To/Do Things Lyrics", "lyrics_url": "h
ttps://genius.com/Drake-like-im-supposed-to-do-things-lyrics", "lyrics": "[Verse]\nHands are tied\nSomeone's in my ear from the other side\nTellin' me
 that I should pay you no mind\nWanted you to not be with me all night\nWanted you to not stay with me all night\nI know, you know, who that person is
 to me\nDoesn't really change things\n\n[Chorus]\nI know you're scared of dating, falling for me\nShorty, surely you know me\nRight here for you alway
s\nYou know, I don't ever change\nRight here for you always\nYou know I don't ever change\nRight here for you\n\n[Bridge]\nIn mind you make me want to
 do things, love you\nLike I'm supposed to\nYou make me want to love you\nLike I'm supposed to\nYou make me want to love you\nLike I'm supposed to, re
:
```

cat for_stream/drake_data.json

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
 Tiller]\nAye, 2 A.M. with my departure\nJug out to Miami, flew in charter\nStay one hunnid, look at me I've been in charge (Charge)\nOh, you know I'm
 'bout to sauce on you (Sauce)\nSuggest you lighten up, I might go dark on you (Sauce)\nHey, pull up, pull up, skrt, vallet park on you\nAyy, King Kon
g, climbing up the charts on you\nOh, goddamn, say my life is full of drama\nHot boy, I feel like Dwayne Michael Carter\n(Sononous on the beat)\nLife
been good since I became a father\nThought you said a kid would make it harder\nNo-no-no, trust me it just made me smarter\nI cut some niggas off, put
 some real ones on the roster\nNever mix the real niggas with the impostors\nKeep them niggas far from us, I say, \"Fuck 'em all\" (I)\nI say, \"Fuck
'em all\"\nI just checked all my accounts and my money sittin' tall\nNigga, shut your fuckin' mouth, ain't no need to air you out\nNiggas know what yo
u about, heard you goin' through a drought\nMust be why you talkin' down\nBoy I know\n\n[Chorus: Drake with Bryson Tiller]\nMy life is full of drama\n
I just want the top, don't wanna charm her\nTalk about the boy and you get karma\nBad karma, that's a sad story\n\n[Outro]\n(Sonorous on the beat)", "
track_views": "50.6K"}][student898_2@bigdataanalytics-worker-3 ~]$
```

hdfs dfs -ls

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
track_views": "50.6K"}][student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 8 items
drwx------   - student898_2 student898_2          0 2022-01-21 18:12 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 20:33 checkpoints
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-17 13:39 input_csv_for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-17 20:24 my_parquet_sink
drwxr-xr-x   - student898_2 student898_2          0 2022-01-17 20:20 tolstykov_les4_file_checkpoint
drwxr-xr-x   - student898_2 student898_2          0 2022-01-17 20:49 tolstykov_les4_kafka_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$
```

Удаляю свои старые файлы

hdfs dfs -rm -f -r tolstykov_les4_file_checkpoint

hdfs dfs -rm -f -r tolstykov_les4_kafka_checkpoint

hdfs dfs -rm -f -r input_csv_for_stream

hdfs dfs -rm -f -r my_parquet_sink

hdfs dfs -ls

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -f -r my_parquet_sink
22/01/21 21:37:34 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/my_parquet_sink' to trash at: hd
fs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/my_parquet_sink
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 4 items
drwx------   - student898_2 student898_2          0 2022-01-21 18:12 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 20:33 checkpoints
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
[student898_2@bigdataanalytics-worker-3 ~]$
```

Создадим папку `input_csv_for_stream` на HDFS, из которой стрим будет читать файлы
hdfs dfs -mkdir input_csv_for_stream
hdfs dfs -ls

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -mkdir input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 5 items
drwx------   - student898_2 student898_2          0 2022-01-21 18:12 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 20:33 checkpoints
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 21:38 input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$
```

Запускаем Spark
export SPARK_KAFKA_VERSION=0.10
/opt/spark-2.4.8/bin/pyspark --packages org.apache.spark:spark-sql-kafka-0-10_2.11:2.4.5 --driver-memory 512m --master local[1]

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
22/01/21 21:39:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/01/21 21:39:49 WARN util.Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 4041.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.8
      /_/

Using Python version 2.7.5 (default, Nov 16 2020 22:23:17)
SparkSession available as 'spark'.
>>>
```

Подключены все зависимости. Форич-бач в прошлой версии не работал.
В другом терминале, смотрим лист топиков
ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176
/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --zookeeper bigdataanalytics-worker-3:2181 --list

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
orders_json
s559_6
shadrin_iris
shadrin_iris_sink
student559_12
student559_8_lesson2
test-lesson2
test_lesson2_1
test_lesson_2_sapr
tolstykov_les4
tolstykov_les4_sink
us_navy
[student898_2@bigdataanalytics-worker-3 ~]$
```

Удаляем старые топики

/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --delete --topic tolstykov_les4 --zookeeper bigdataanalytics-worker-3:2181
/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --delete --topic tolstykov_les4_sink --zookeeper bigdataanalytics-worker-3:2181
/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --zookeeper bigdataanalytics-worker-3:2181 --list

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
oganesyan_les2
order_items
orders_json
s559_6
shadrin_iris
shadrin_iris_sink
student559_12
student559_8_lesson2
test-lesson2
test_lesson2_1
test_lesson_2_sapr
us_navy
[student898_2@bigdataanalytics-worker-3 ~]$
```

Создаю топик tolstykov_les4
/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --create --topic tolstykov_les4 --zookeeper bigdataanalytics-worker-3:2181 --partitions 1 --replication-factor 1

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
order_items
orders_json
s559_6
shadrin_iris
shadrin_iris_sink
student559_12
student559_8_lesson2
test-lesson2
test_lesson2_1
test_lesson_2_sapr
tolstykov_les4
us_navy
[student898_2@bigdataanalytics-worker-3 ~]$
```

Загрузить файл в топик
/usr/hdp/current/kafka-broker/bin/kafka-console-producer.sh --broker-list bigdataanalytics-worker-3:6667 --topic tolstykov_les4 < for_stream/drake_data.json

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
s559_6
shadrin_iris
shadrin_iris_sink
student559_12
student559_8_lesson2
test-lesson2
test_lesson2_1
test_lesson_2_sapr
tolstykov_les4
us_navy
[student898_2@bigdataanalytics-worker-3 ~]$ /usr/hdp/current/kafka-broker/bin/kafka-console-producer.sh --broker-list bigdataanalytics-worker-3:6667 -
-topic tolstykov_les4 < for_stream/drake_data.json
>>[student898_2@bigdataanalytics-worker-3 ~]$
```

Прочитать топик  tolstykov_les4
/usr/hdp/current/kafka-broker/bin/kafka-console-consumer.sh --topic tolstykov_les4 --from-beginning --bootstrap-server bigdataanalytics-worker-3:6667 --max-messages 15

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
see a hunnid, times a hunnid\nFor the guys, for the city and the gang\nI might appear if you keep calling on my name\nPut a curse up on you, things ar
e getting strange\nYou ain't lie when you told me that I changed\nAnd I know\n\n[Chorus: Drake]\n(Sonorous on the beat)\nMy life is full of drama\nI j
ust want the top, don't wanna charm her\nTalk about the boy and you get karma\nBad karma, that's a sad story, yeah, real bad karma\n\n[Verse 2: Bryson
 Tiller]\nAye, 2 A.M. with my departure\nJug out to Miami, flew in charter\nStay one hunnid, look at me I've been in charge (Charge)\nOh, you know I'm
 'bout to sauce on you (Sauce)\nSuggest you lighten up, I might go dark on you (Sauce)\nHey, pull up, pull up, skrt, vallet park on you\nAyy, King Kon
g, climbing up the charts on you\nOh, goddamn, say my life is full of drama\nHot boy, I feel like Dwayne Michael Carter\n(Sononous on the beat)\nLife
been good since I became a father\nThought you said a kid would make it harder\nNo-no-no, trust me it just made me smarter\nI cut some niggas off, put
 some real ones on the roster\nNever mix the real niggas with the impostors\nKeep them niggas far from us, I say, \"Fuck 'em all\" (I)\nI say, \"Fuck
'em all\"\nI just checked all my accounts and my money sittin' tall\nNigga, shut your fuckin' mouth, ain't no need to air you out\nNiggas know what yo
u about, heard you goin' through a drought\nMust be why you talkin' down\nBoy I know\n\n[Chorus: Drake with Bryson Tiller]\nMy life is full of drama\n
I just want the top, don't wanna charm her\nTalk about the boy and you get karma\nBad karma, that's a sad story\n\n[Outro]\n(Sonorous on the beat)", "
track_views": "50.6K"}]
```

В терминале со спарк

```
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType, FloatType
kafka_brokers = "bigdataanalytics-worker-3:6667"

raw_data = spark.readStream. \
    format("kafka"). \
    option("kafka.bootstrap.servers", kafka_brokers). \
    option("subscribe", "tolstykov_les4"). \
    option("startingOffsets", "earliest"). \
    option("maxOffsetsPerTrigger", "5"). \
    load()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> from pyspark.sql import functions as F
>>> from pyspark.sql.types import StructType, StringType, FloatType
>>> kafka_brokers = "bigdataanalytics-worker-3:6667"
>>> raw_data = spark.readStream. \
...     format("kafka"). \
...     option("kafka.bootstrap.servers", kafka_brokers). \
...     option("subscribe", "tolstykov_les4"). \
...     option("startingOffsets", "earliest"). \
...     option("maxOffsetsPerTrigger", "5"). \
...     load()
>>>
```

Определяем схему данных нашего исходного датасета.

```
schema = StructType() \
    .add("album", StringType()) \
    .add("lyrics_title", StringType()) \
    .add("lyrics_url", StringType()) \
    .add("lyrics", StringType()) \
    .add("track_views", StringType())
```

Сделаем преобразование в плоскую структуру

```
parsed_data = raw_data \
        .select(F.from_json(F.col("value").cast("String"), schema).alias("value"), "offset") \
        .select("value.*", "offset")
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> schema = StructType() \
...     .add("album", StringType()) \
...     .add("lyrics_title", StringType()) \
...     .add("lyrics_url", StringType()) \
...     .add("lyrics", StringType()) \
...     .add("track_views", StringType())
>>> parsed_data = raw_data \
...     .select(F.from_json(F.col("value").cast("String"), schema).alias("value"), "offset") \
...     .select("value.*", "offset")
>>>
```

```
parsed_data.printSchema()
raw_data.printSchema()
```

```
>>> parsed_data.printSchema()
root
 |-- album: string (nullable = true)
 |-- lyrics_title: string (nullable = true)
 |-- lyrics_url: string (nullable = true)
 |-- lyrics: string (nullable = true)
 |-- track_views: string (nullable = true)
 |-- offset: long (nullable = true)

>>> raw_data.printSchema()
root
 |-- key: binary (nullable = true)
 |-- value: binary (nullable = true)
 |-- topic: string (nullable = true)
 |-- partition: integer (nullable = true)
 |-- offset: long (nullable = true)
 |-- timestamp: timestamp (nullable = true)
 |-- timestampType: integer (nullable = true)

>>>
```

Чекпоинт

```
def console_output(df, freq):
    return df.writeStream \
        .format("console") \
        .trigger(processingTime='%s seconds' % freq) \
        .option("truncate",False) \
        .start()
```

```
>>> def console_output(df, freq):
...     return df.writeStream \
...         .format("console") \
...         .trigger(processingTime='%s seconds' % freq) \
...         .option("truncate",False) \
...         .start()
...
>>>
```

```
out = console_output(parsed_data, 5)
out.stop()
```

```
...
>>> out = console_output(parsed_data, 5)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
>>> +-----+------------+----------+------+-----------+------+
|album|lyrics_title|lyrics_url|lyrics|track_views|offset|
+-----+------------+----------+------+-----------+------+
|null |null        |null      |null  |null       |0     |
+-----+------------+----------+------+-----------+------+

>>>
```

Данные не читаются, сравниваю структуру файлов darke_data.json и iris.json

## drake_data.json (846.62 kB)

### About this file

JSON contains lyrics, song title, album title, url, view count (at this time)

This preview is truncated due to the large file size. The number of JSON items and individual items might be might be truncated. Create a Notebook or download this file to see the full content.

Download  Create Notebook


```
"root" : [ 36 items
  0 : { 5 items
    "album" : string "Certified Lover Boy"
    "lyrics_title" : string "Certified Lover Boy* Lyrics"
    "lyrics_url" : string "https://genius.com/Drake-certified-lover-boy-lyrics"
    "lyrics" :
    string "Lyrics from CLB Merch [Verse] Put my feelings on ice Always been a gem Certified lover boy, somehow still heartless Heart is only gettin' colder"
    "track_views" : string "8.7K"
  }
  1 : {...} 5 items
```


Iris Dataset (JSON Version) | Kaggle

отер ▾ Обучение ▾ Работа ▾ GB ▾ Потоковая обработка ▾

Activity  Metadata  Download (16 kB)  New Notebook

## iris.json (15.8 kB)

### About this file

Keys: sepalLength, sepalWidth, petalLength, petalWidth and species.


```
"root" : 150 items
  [ 100 items
    0 : { 5 items
      "sepalLength" : float 5.1
      "sepalWidth" : float 3.5
      "petalLength" : float 1.4
      "petalWidth" : float 0.2
      "species" : string "setosa"
    }
    1 : {...} 5 items
```

Запись потока в память

```
def memory_sink(df, freq):
        return df.writeStream.format("memory") \
                .queryName("my_memory_sink_table") \
                .trigger(processingTime='%s seconds' % freq) \
                .start()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> def memory_sink(df, freq):
...     return df.writeStream.format("memory") \
...             .queryName("my_memory_sink_table") \
...             .trigger(processingTime='%s seconds' % freq) \
...             .start()
...
>>>
```

```
stream = memory_sink(parsed_data, 15)
```

Что бы считать из памяти

```
spark.sql("select * from my_memory_sink_table").show()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> stream = memory_sink(parsed_data, 15)
>>> spark.sql("select * from my_memory_sink_table").show()
+-----+------------+----------+------+-----------+------+
|album|lyrics_title|lyrics_url|lyrics|track_views|offset|
+-----+------------+----------+------+-----------+------+
| null|        null|      null|  null|       null|     0|
+-----+------------+----------+------+-----------+------+

>>>
```

```
spark.sql('select count(*) from my_memory_sink_table').show()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> spark.sql('select count(*) from my_memory_sink_table').show()
+--------+
|count(1)|
+--------+
|       1|
+--------+

>>>
```

Запись файла в формат parquet

```
def file_sink(df, freq):
        return df.writeStream.format("parquet") \
                .trigger(processingTime='%s seconds' % freq) \
                .option("path", "my_parquet_sink") \
                .option("checkpointLocation", "tolstykov_les4_file_checkpoint") \
                .start()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> def file_sink(df, freq):
...     return df.writeStream.format("parquet") \
...             .trigger(processingTime='%s seconds' % freq) \
...             .option("path", "my_parquet_sink") \
...             .option("checkpointLocation", "tolstykov_les4_file_checkpoint") \
...             .start()
...
>>>
```

В другом терминале

```
hdfs dfs -ls
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 5 items
drwx------   - student898_2 student898_2          0 2022-01-21 18:12 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 20:33 checkpoints
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 21:38 input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$
```

В первом терминале
stream = file_sink(parsed_data, 5)

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> def file_sink(df, freq):
...     return df.writeStream.format("parquet") \
...             .trigger(processingTime='%s seconds' % freq) \
...             .option("path", "my_parquet_sink") \
...             .option("checkpointLocation", "tolstykov_les4_file_checkpoint") \
...             .start()
...
>>> stream = file_sink(parsed_data, 5)
>>>
```

Во втором терминале
hdfs dfs -ls

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 7 items
drwx------   - student898_2 student898_2          0 2022-01-21 18:12 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 20:33 checkpoints
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 21:38 input_csv_for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 22:24 my_parquet_sink
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 22:24 tolstykov_les4_file_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$
```

В первом окне останавливаем стрим
stream.stop()
Во втором окне смотрим, что внутри папок
hdfs dfs -ls my_parquet_sink

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 21:38 input_csv_for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 22:24 my_parquet_sink
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 22:24 tolstykov_les4_file_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls my_parquet_sink
Found 2 items
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 22:24 my_parquet_sink/_spark_metadata
-rw-r--r--   2 student898_2 student898_2       1299 2022-01-21 22:24 my_parquet_sink/part-00000-283ff4a1-e54b-4c0d-bc6c-4705e03cfdf4-c000.snappy.parqu
et
[student898_2@bigdataanalytics-worker-3 ~]$
```

Метод записи из kafka делаем структуру key - value

```
def kafka_sink(df, freq):
    return df.selectExpr("CAST(null AS STRING) as key", "CAST(struct(*) AS STRING) as value") \
        .writeStream \
        .format("kafka") \
        .trigger(processingTime='%s seconds' % freq) \
        .option("topic", "tolstykov_les4") \
        .option("kafka.bootstrap.servers", kafka_brokers) \
        .option("checkpointLocation", "tolstykov_les4_kafka_checkpoint") \
        .start()
```

```
>>> def kafka_sink(df, freq):
...     return df.selectExpr("CAST(null AS STRING) as key", "CAST(struct(*) AS STRING) as value") \
...             .writeStream \
...             .format("kafka") \
...             .trigger(processingTime='%s seconds' % freq) \
...             .option("topic", "tolstykov_les4") \
...             .option("kafka.bootstrap.servers", kafka_brokers) \
...             .option("checkpointLocation", "tolstykov_les4_kafka_checkpoint") \
...             .start()
...
>>>
```

Во втором окне терминала создадим топик tolstykov_les4_sink
/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --create --topic tolstykov_les4_sink --zookeeper bigdataanalytics-worker-3:2181 --partitions 3 --replication-factor 2 --config retention.ms=-1

```
Found 2 items
drwxr-xr-x   - student898_2 student898_2          0 2022-01-21 22:24 my_parquet_sink/_spark_metadata
-rw-r--r--   2 student898_2 student898_2       1299 2022-01-21 22:24 my_parquet_sink/part-00000-283ff4a1-e54b-4c0d-bc6c-4705e03cfdf4-c000.snappy.parqu
et
[student898_2@bigdataanalytics-worker-3 ~]$ /usr/hdp/current/kafka-broker/bin/kafka-topics.sh --create --topic tolstykov_les4_sink --zookeeper bigdata
analytics-worker-3:2181 --partitions 3 --replication-factor 2 --config retention.ms=-1
WARNING: Due to limitations in metric names, topics with a period ('.') or underscore ('_') could collide. To avoid issues it is best to use either, b
ut not both.
Created topic "tolstykov_les4_sink".
[student898_2@bigdataanalytics-worker-3 ~]$
```

/usr/hdp/current/kafka-broker/bin/kafka-topics.sh --zookeeper bigdataanalytics-worker-3:2181 —list

```
shadrin_iris_sink
student559_12
student559_8_lesson2
test-lesson2
test_lesson2_1
test_lesson_2_sapr
tolstykov_les4
tolstykov_les4_sink
us_navy
[student898_2@bigdataanalytics-worker-3 ~]$
```

Подписываемся на его обновления
/usr/hdp/current/kafka-broker/bin/kafka-console-consumer.sh --topic tolstykov_les4_sink --bootstrap-server bigdataanalytics-worker-3:6667

```
student559_8_lesson2
test-lesson2
test_lesson2_1
test_lesson_2_sapr
tolstykov_les4
tolstykov_les4_sink
us_navy
[student898_2@bigdataanalytics-worker-3 ~]$ /usr/hdp/current/kafka-broker/bin/kafka-console-consumer.sh --topic tolstykov_les4_sink --bootstrap-server
 bigdataanalytics-worker-3:6667
```

Запускаем поток в первой консоли
stream = kafka_sink(parsed_data, 5)
stream.stop()

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
-------------------------------------------
Batch: 5
-------------------------------------------
+-----+------------+----------+------+-----------+------+
|album|lyrics_title|lyrics_url|lyrics|track_views|offset|
+-----+------------+----------+------+-----------+------+
|null |null        |null      |null  |null       |6     |
+-----+------------+----------+------+-----------+------+

-------------------------------------------
Batch: 6
-------------------------------------------
+-----+------------+----------+------+-----------+------+
|album|lyrics_title|lyrics_url|lyrics|track_views|offset|
+-----+------------+----------+------+-----------+------+
|null |null        |null      |null  |null       |7     |
+-----+------------+----------+------+-----------+------+

-------------------------------------------
Batch: 7
-------------------------------------------
+-----+------------+----------+------+-----------+------+
|album|lyrics_title|lyrics_url|lyrics|track_views|offset|
+-----+------------+----------+------+-----------+------+
|null |null        |null      |null  |null       |8     |
+-----+------------+----------+------+-----------+------+

stream.stop()
>>> stream.stop()-------------------------------------------
Batch: 8
-------------------------------------------
+-----+------------+----------+------+-----------+------+
|album|lyrics_title|lyrics_url|lyrics|track_views|offset|
+-----+------------+----------+------+-----------+------+
|null |null        |null      |null  |null       |9     |
+-----+------------+----------+------+-----------+------+


>>> stream.stop()
>>>
```

Переключимся в json

```
def kafka_sink_json(df, freq):
        return df.selectExpr("CAST(null AS STRING) as key", "CAST(to_json(struct(*)) AS STRING) as value") \
                .writeStream \
                .format("kafka") \
                .trigger(processingTime='%s seconds' % freq) \
                .option("topic", "tolstykov_les4_sink") \
                .option("kafka.bootstrap.servers", kafka_brokers) \
                .option("checkpointLocation", "tolstykov_les4_kafka_checkpoint") \
                .start()
stream = kafka_sink_json(parsed_data, 5)
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> def kafka_sink_json(df, freq):
...     return df.selectExpr("CAST(null AS STRING) as key", "CAST(to_json(struct(*)) AS STRING) as value") \
...             .writeStream \
...             .format("kafka") \
...             .trigger(processingTime='%s seconds' % freq) \
...             .option("topic", "tolstykov_les4_sink") \
...             .option("kafka.bootstrap.servers", kafka_brokers) \
...             .option("checkpointLocation", "tolstykov_les4_kafka_checkpoint") \
...             .start()
...
>>> stream = kafka_sink_json(parsed_data, 5)
>>>
```

stream.stop()

Переходим к foreach_batch_sink

```
extended_data = parsed_data.withColumn("my_current_time", F.current_timestamp())
extended_data.printSchema()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> stream = kafka_sink_json(parsed_data, 5)
>>> stream.stop()
>>> extended_data = parsed_data.withColumn("my_current_time", F.current_timestamp())
>>> extended_data.printSchema()
root
 |-- album: string (nullable = true)
 |-- lyrics_title: string (nullable = true)
 |-- lyrics_url: string (nullable = true)
 |-- lyrics: string (nullable = true)
 |-- track_views: string (nullable = true)
 |-- offset: long (nullable = true)
 |-- my_current_time: timestamp (nullable = false)

>>>
```

Определим функцию понятие формат заменяем на  foreach_batch

```
def foreach_batch_sink(df, freq):
    return df \
        .writeStream \
        .foreachBatch(foreach_batch_function) \
        .trigger(processingTime='%s seconds' % freq) \
        .start()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> def foreach_batch_sink(df, freq):
...     return df \
...         .writeStream \
...         .foreachBatch(foreach_batch_function) \
...         .trigger(processingTime='%s seconds' % freq) \
...         .start()
...
>>>
```

```
def foreach_batch_function(df, epoch_id):
    print("starting epoch " + str(epoch_id))
    print("averege values for batch:")
    df.groupBy("species").avg().show()
    print("finishing epoch " + str(epoch_id))
```

внутри этой функции можно работать как со статическим датасетом и порождать фильтрации, изминения, новый поток и т.д.

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
...
>>> def foreach_batch_function(df, epoch_id):
...     print("starting epoch " + str(epoch_id))
...     print("averege values for batch:")
...     df.groupBy("species").avg().show()
...     print("finishing epoch " + str(epoch_id))
...
>>>
```

```
stream = foreach_batch_sink(extended_data, 5)
stream.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
        at org.apache.spark.sql.execution.SQLExecution$.withNewExecutionId(SQLExecution.scala:75)
        at org.apache.spark.sql.execution.streaming.MicroBatchExecution$$anonfun$org$apache$spark$sql$execution$streaming$MicroBatchExecution$$runBatc
h$5.apply(MicroBatchExecution.scala:546)
        at org.apache.spark.sql.execution.streaming.ProgressReporter$class.reportTimeTaken(ProgressReporter.scala:351)
        at org.apache.spark.sql.execution.streaming.StreamExecution.reportTimeTaken(StreamExecution.scala:58)
        at org.apache.spark.sql.execution.streaming.MicroBatchExecution.org$apache$spark$sql$execution$streaming$MicroBatchExecution$$runBatch(MicroBa
tchExecution.scala:545)
        at org.apache.spark.sql.execution.streaming.MicroBatchExecution$$anonfun$runActivatedStream$1$$anonfun$apply$mcZ$sp$1.apply$mcV$sp(MicroBatchE
xecution.scala:198)
        at org.apache.spark.sql.execution.streaming.MicroBatchExecution$$anonfun$runActivatedStream$1$$anonfun$apply$mcZ$sp$1.apply(MicroBatchExecutio
n.scala:166)
        at org.apache.spark.sql.execution.streaming.MicroBatchExecution$$anonfun$runActivatedStream$1$$anonfun$apply$mcZ$sp$1.apply(MicroBatchExecutio
n.scala:166)
        at org.apache.spark.sql.execution.streaming.ProgressReporter$class.reportTimeTaken(ProgressReporter.scala:351)
        at org.apache.spark.sql.execution.streaming.StreamExecution.reportTimeTaken(StreamExecution.scala:58)
        at org.apache.spark.sql.execution.streaming.MicroBatchExecution$$anonfun$runActivatedStream$1.apply$mcZ$sp(MicroBatchExecution.scala:166)
        at org.apache.spark.sql.execution.streaming.ProcessingTimeExecutor.execute(TriggerExecutor.scala:56)
        at org.apache.spark.sql.execution.streaming.MicroBatchExecution.runActivatedStream(MicroBatchExecution.scala:160)
        at org.apache.spark.sql.execution.streaming.StreamExecution.org$apache$spark$sql$execution$streaming$StreamExecution$$runStream(StreamExecutio
n.scala:281)
        at org.apache.spark.sql.execution.streaming.StreamExecution$$anon$1.run(StreamExecution.scala:193)
>>> stream.stop()
>>>
```

Запись/сохранение данных в файл

```
# CSV
data.write.csv('dataset.csv')

# JSON
data.write.save('dataset.json', format='json')

# Parquet
data.write.save('dataset.parquet', format='parquet')
```