

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 «data.csv» на удаленный сервер с помощью команды `scp`. Эта команда запускается на локальном компьютере

```
scp -i ~/.ssh/id_rsa_student898_2 -r data.csv student898_2@37.139.41.176:~/for_stream
```

Подключаемся и проверяем, что файл data.csv 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:~$ ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176
Last login: Sat Jan 22 22:21:23 2022 from 109.252.19.10
[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/data.csv
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
"time_id","ping_ms","temperature_c","humidity_p"
"2021-09-30 21:08:02","17.28","25","35"
"2021-09-30 21:09:02","17.73","23","40"
"2021-09-30 21:10:01","18.59","22","41"
"2021-09-30 21:12:02","16.73","22","42"
"2021-09-30 21:13:02","18.12","22","42"
for_stream/data.csv
```

```
hdfs dfs -ls
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
drwx----- - student898_2 student898_2      0 2022-01-22 06:00 .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
drwxr-xr-x - student898_2 student898_2      0 2022-01-21 22:32 tolstykov_les4_kafka_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$
```

Удаляю свои старые файлы

```
hdfs dfs -rm -f -r checkpoints
```

```
hdfs dfs -rm -f -r input_csv_for_stream
```

```
hdfs dfs -rm -f -r my_parquet_sink
```

```
hdfs dfs -rm -f -r tolstykov_les4_file_checkpoint
```

```
hdfs dfs -rm -f -r tolstykov_les4_kafka_checkpoint
```

```
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 checkpoints
22/01/22 22:30:51 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/checkpoints' to trash at: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/checkpoints
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -f -r input_csv_for_stream
22/01/22 22:31:26 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/input_csv_for_stream' to trash at: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -f -r my_parquet_sink
22/01/22 22:31:51 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/my_parquet_sink' to trash at: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/my_parquet_sink
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -f -r tolstykov_les4_file_checkpoint
22/01/22 22:32:21 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/tolstykov_les4_file_checkpoint' to trash at: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/tolstykov_les4_file_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -f -r tolstykov_les4_kafka_checkpoint
22/01/22 22:32:56 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/tolstykov_les4_kafka_checkpoint' to trash at: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/tolstykov_les4_kafka_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 3 items
drwx----- - student898_2 student898_2      0 2022-01-22 22:30 .Trash
drwxr-xr-x - student898_2 student898_2      0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x - student898_2 student898_2      0 2021-12-15 22:13 for_stream
[student898_2@bigdataanalytics-worker-3 ~]$

```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -mkdir input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 4 items
drwx----- - student898_2 student898_2      0 2022-01-22 22:30 .Trash
drwxr-xr-x - student898_2 student898_2      0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x - student898_2 student898_2      0 2021-12-15 22:13 for_stream
drwxr-xr-x - student898_2 student898_2      0 2022-01-22 22:34 input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
```

Файл Правка Вид Терминал Вкладки Справка

Welcome to

 version 2.4.8

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


```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
strap-server bigdataanalytics-worker-3:6667 --max-messages 10r/bin/kafka-console-consumer.sh --topic tolstykov_les4 --from-beginning --boot
"time_id","ping_ms","temperature_c","humidity_p"
"2021-09-30 21:08:02","17.28","25","35"
"2021-09-30 21:09:02","17.73","23","40"
"2021-09-30 21:10:01","18.59","22","41"
"2021-09-30 21:12:02","16.73","22","42"
"2021-09-30 21:13:02","18.12","22","42"
"2021-09-30 21:14:01","18.21","22","43"
"2021-09-30 21:15:01","17.92","22","43"
"2021-09-30 21:16:02","17.2","22","43"
"2021-09-30 21:17:02","18.16","22","43"
Processed a total of 10 messages
[student898_2@bigdataanalytics-worker-3 ~]$
```

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("time_id", StringType()) \
    .add("ping_ms", StringType()) \
    .add("temperature_c", StringType()) \
    .add("humidity_p", 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("time_id", StringType()) \
...     .add("ping_ms", StringType()) \
...     .add("temperature_c", StringType()) \
...     .add("humidity_p", 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()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка

>>> parsed_data.printSchema()
root
 |-- time_id: string (nullable = true)
 |-- ping_ms: string (nullable = true)
 |-- temperature_c: string (nullable = true)
 |-- humidity_p: 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()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка

>>> 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()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка

+-----+-----+-----+-----+
|time_id|ping_ms|temperature_c|humidity_p|offset|
+-----+-----+-----+-----+
|null   |null   |null        |null      |15    |
|null   |null   |null        |null      |16    |
|null   |null   |null        |null      |17    |
|null   |null   |null        |null      |18    |
|null   |null   |null        |null      |19    |
+-----+-----+-----+-----+

Batch: 4

+-----+-----+-----+-----+
|time_id|ping_ms|temperature_c|humidity_p|offset|
+-----+-----+-----+-----+
|null   |null   |null        |null      |20    |
|null   |null   |null        |null      |21    |
|null   |null   |null        |null      |22    |
|null   |null   |null        |null      |23    |
|null   |null   |null        |null      |24    |
+-----+-----+-----+-----+

out.stop()
>>> out.stop()
>>>
```

Данные не читаются data.csv

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

>>> out.stop()
>>> 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, 10)
```

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, 10)
>>> spark.sql("select * from my_memory_sink_table").show()
+-----+-----+-----+-----+-----+
|time_id|ping_ms|temperature_c|humidity_p|offset|
+-----+-----+-----+-----+-----+
| null  | null  | null       | null     | 0    |
| null  | null  | null       | null     | 1    |
| null  | null  | null       | null     | 2    |
| null  | null  | null       | null     | 3    |
| null  | null  | null       | null     | 4    |
| null  | null  | null       | null     | 5    |
| null  | null  | null       | null     | 6    |
| null  | null  | null       | null     | 7    |
| null  | null  | null       | null     | 8    |
| null  | null  | null       | null     | 9    |
| null  | null  | null       | null     | 10   |
| null  | null  | null       | null     | 11   |
| null  | null  | null       | null     | 12   |
| null  | null  | null       | null     | 13   |
| null  | null  | null       | null     | 14   |
+-----+-----+-----+-----+-----+

>>> █
```

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)|
+-----+
|      40|
+-----+

>>> █
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка

>>> spark.sql('select count(*) from my_memory_sink_table').show()
+-----+
|count(1)|
+-----+
|      50|
+-----+

>>> █
```

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 4 items
drwx----- - student898_2 student898_2      0 2022-01-22 22:30 .Trash
drwxr-xr-x - student898_2 student898_2      0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x - student898_2 student898_2      0 2021-12-15 22:13 for_stream
drwxr-xr-x - student898_2 student898_2      0 2022-01-22 22:34 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
SyntaxError: invalid syntax
>>> 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 6 items
drwx----- - student898_2 student898_2      0 2022-01-22 22:30 .Trash
drwxr-xr-x - student898_2 student898_2      0 2022-01-20 19:25 .sparkStaging
drwxr-xr-x - student898_2 student898_2      0 2021-12-15 22:13 for_stream
drwxr-xr-x - student898_2 student898_2      0 2022-01-22 22:34 input_csv_for_stream
drwxr-xr-x - student898_2 student898_2      0 2022-01-22 22:57 my_parquet_sink
drwxr-xr-x - student898_2 student898_2      0 2022-01-22 22:56 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
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls my_parquet_sink
Found 22 items
drwxr-xr-x - student898_2 student898_2      0 2022-01-22 22:58 my_parquet_sink/ spark metadata
1161 2022-01-22 22:56 my_parquet_sink/part-00000-000f9c04-324d-4c9c-8f5b-17de0f9044af-c000.snappy.parquet
1159 2022-01-22 22:56 my_parquet_sink/part-00000-05d0eccb-068d-45bd-ba5b-5387ed289f42-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-20c1ad7c-a3ce-4aeb-9063-95f99ac945ad-c000.snappy.parquet
1161 2022-01-22 22:56 my_parquet_sink/part-00000-3f2ded81-819f-45c7-9c5a-877b3ebfa8e6-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-41917a25-353e-4b30-8501-452a130d531f-c000.snappy.parquet
1161 2022-01-22 22:56 my_parquet_sink/part-00000-50de8cc2-e6af-43e4-88c7-db7fca6d2380-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-5f17403d-9f08-415a-8598-ffac6433f585-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-628a4957-502d-47c4-ab20-0fd9c1ee1e82-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-674a3d1e-5f96-4198-a4ca-d10e0b60960c-c000.snappy.parquet
1161 2022-01-22 22:56 my_parquet_sink/part-00000-6c34bc17-79a7-4996-8491-d6d32ce394af-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-72c91aae-01d2-447b-bf82-fd4d49debc59-c000.snappy.parquet
1161 2022-01-22 22:56 my_parquet_sink/part-00000-7526aa4b-a7ec-4a47-ab0f-5c7534a7bd15-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-75f743f1-87e6-4824-b479-47b40e7b5c18-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-98ea3120-493c-4dd2-872c-87ef926a9490-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-9dffd920-a5cd-433f-9e0d-538fb7927d0b-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-a7ce5a1a-0c9f-4655-8356-356e5828260a-c000.snappy.parquet
1161 2022-01-22 22:56 my_parquet_sink/part-00000-a85232c1-e239-4691-bffa-1dd409f59f4e-c000.snappy.parquet
1161 2022-01-22 22:58 my_parquet_sink/part-00000-ac22b12b-c481-453c-8567-744696690439-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-dd92da9e-ff54-4e88-88c7-556c8bcfb036-c000.snappy.parquet
1161 2022-01-22 22:58 my_parquet_sink/part-00000-e4c05f86-8c50-40e6-b414-9b0ff34bf7ee-c000.snappy.parquet
1161 2022-01-22 22:57 my_parquet_sink/part-00000-f91bf203-ff15-4d01-a108-c2dc3ca4a2e3-c000.snappy.parquet
[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()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> 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
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ /usr/hdp/current/kafka-broker/bin/kafka-topics.sh --create --topic tolstykov_les4_sink --zookeeper bigdataanal
ytics-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, but n
ot 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
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
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
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
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 big
dataanalytics-worker-3:6667
```

Запускаем поток в первой консоли

```
stream = kafka_sink(parsed_data, 5)
```

ВИСИТ НЕ ОТВЕЧАЕТ

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
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
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
...   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()
...
>>> stream = kafka_sink(parsed_data, 5)
>>> 
```

```
stream.stop()
```

Записи/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')
```