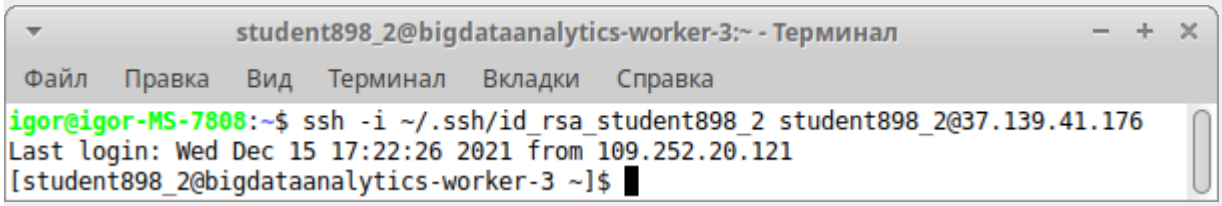
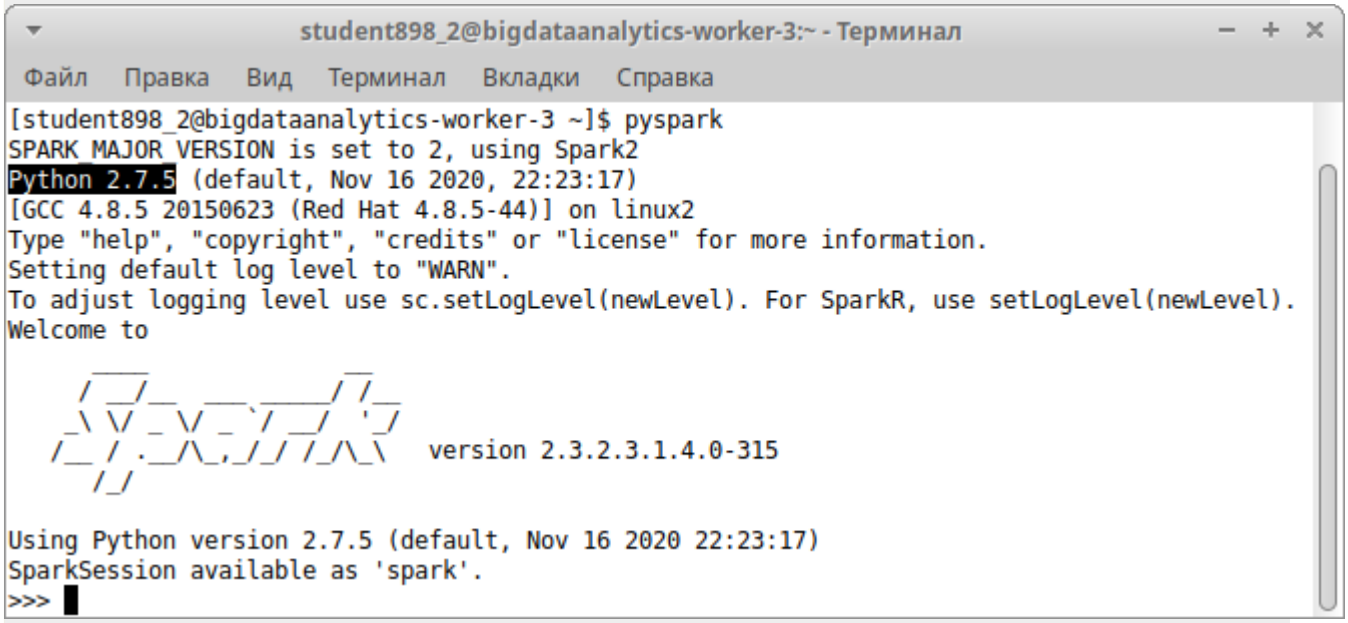
1. Spark Streaming. Тестовые стримы, чтение файлов в реальном времени.

Получить у преподавателя доступ к консоли, подключиться, выполнить команды рассмотренные на уроке. Дополнительно загрузить в консоль свои данные и выполнить команды с ними. В качестве отчета о проделанной работе приложить файл с листингом выполнения команд в консоли

```
cd /home/igor/Загрузки/Telegram\ Desktop/

cp id_rsa_student898_2 ~/.ssh/

chmod 600 ~/.ssh/id_rsa_student898_2

Подключаемся к серверу

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



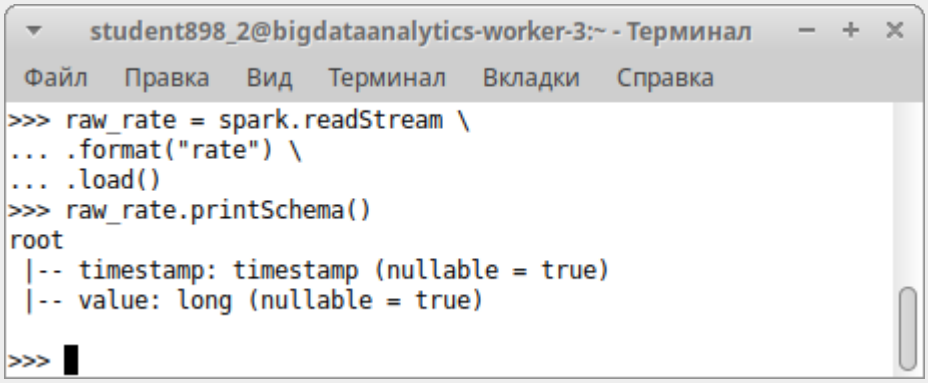

```
Запускаем спарк-приложение

pyspark
```



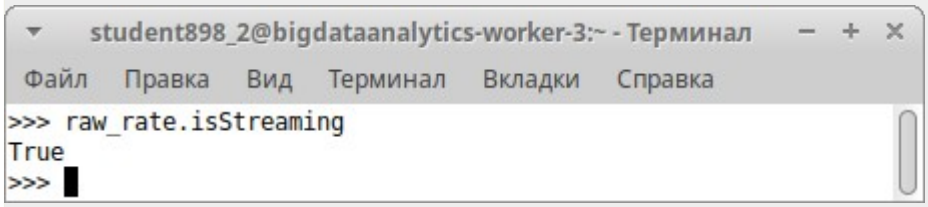

```
Пробуем выполнить команды из файла

raw_rate = spark.readStream \

… .format("rate") \
```

```
… .load()
```

```
raw_rate.printSchema()
```




```
raw_rate.isStreaming
```




```
Будем писать наш стрим в консоль с интервалом 30 секунд

stream = raw_rate.writeStream \

      .trigger(processingTime='30 seconds') \

      .format("console") \

      .options(truncate=False) \

      .start()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> stream = raw_rate.writeStream \
... .trigger(processingTime='30 seconds') \
... .format("console") \
... .options(truncate=False) \
... .start()
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+---------+-----+
|timestamp|value|
+---------+-----+
+---------+-----+

[Stage 0:>                                                  [Stage 0:=======
=====================>
                     -------------------------------------------
Batch: 1
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 20:23:44.906|0    |
|2021-12-15 20:23:45.906|1    |
|2021-12-15 20:23:46.906|2    |
|2021-12-15 20:23:47.906|3    |
|2021-12-15 20:23:48.906|4    |
|2021-12-15 20:23:49.906|5    |
|2021-12-15 20:23:50.906|6    |
|2021-12-15 20:23:51.906|7    |
|2021-12-15 20:23:52.906|8    |
|2021-12-15 20:23:53.906|9    |
|2021-12-15 20:23:54.906|10   |
|2021-12-15 20:23:55.906|11   |
|2021-12-15 20:23:56.906|12   |
|2021-12-15 20:23:57.906|13   |
|2021-12-15 20:23:58.906|14   |
+-----------------------+-----+

>>> -------------------------------------------
Batch: 2
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 20:23:59.906|15   |
|2021-12-15 20:24:00.906|16   |
|2021-12-15 20:24:01.906|17   |
|2021-12-15 20:24:02.906|18   |
|2021-12-15 20:24:03.906|19   |
|2021-12-15 20:24:04.906|20   |
|2021-12-15 20:24:05.906|21   |
|2021-12-15 20:24:06.906|22   |
|2021-12-15 20:24:07.906|23   |
|2021-12-15 20:24:08.906|24   |
|2021-12-15 20:24:09.906|25   |
|2021-12-15 20:24:10.906|26   |
|2021-12-15 20:24:11.906|27   |
|2021-12-15 20:24:12.906|28   |
|2021-12-15 20:24:13.906|29   |
|2021-12-15 20:24:14.906|30   |
|2021-12-15 20:24:15.906|31   |
|2021-12-15 20:24:16.906|32   |
|2021-12-15 20:24:17.906|33   |
|2021-12-15 20:24:18.906|34   |
+-----------------------+-----+
only showing top 20 rows

>>> 
```

```
stream.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> -------------------------------------------
Batch: 3
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 20:24:29.906|45   |
|2021-12-15 20:24:30.906|46   |
|2021-12-15 20:24:31.906|47   |
|2021-12-15 20:24:32.906|48   |
|2021-12-15 20:24:33.906|49   |
|2021-12-15 20:24:34.906|50   |
|2021-12-15 20:24:35.906|51   |
|2021-12-15 20:24:36.906|52   |
|2021-12-15 20:24:37.906|53   |
|2021-12-15 20:24:38.906|54   |
|2021-12-15 20:24:39.906|55   |
|2021-12-15 20:24:40.906|56   |
|2021-12-15 20:24:41.906|57   |
|2021-12-15 20:24:42.906|58   |
|2021-12-15 20:24:43.906|59   |
|2021-12-15 20:24:44.906|60   |
|2021-12-15 20:24:45.906|61   |
|2021-12-15 20:24:46.906|62   |
|2021-12-15 20:24:47.906|63   |
|2021-12-15 20:24:48.906|64   |
+-----------------------+-----+
only showing top 20 rows

stream.stop()-------------------------------------------
Batch: 4
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 20:24:59.906|75   |
|2021-12-15 20:25:00.906|76   |
|2021-12-15 20:25:01.906|77   |
|2021-12-15 20:25:02.906|78   |
|2021-12-15 20:25:03.906|79   |
|2021-12-15 20:25:04.906|80   |
|2021-12-15 20:25:05.906|81   |
|2021-12-15 20:25:06.906|82   |
|2021-12-15 20:25:07.906|83   |
|2021-12-15 20:25:08.906|84   |
|2021-12-15 20:25:09.906|85   |
|2021-12-15 20:25:10.906|86   |
|2021-12-15 20:25:11.906|87   |
|2021-12-15 20:25:12.906|88   |
|2021-12-15 20:25:13.906|89   |
|2021-12-15 20:25:14.906|90   |
|2021-12-15 20:25:15.906|91   |
|2021-12-15 20:25:16.906|92   |
|2021-12-15 20:25:17.906|93   |
|2021-12-15 20:25:18.906|94   |
+-----------------------+-----+
only showing top 20 rows

>>>
```

Посмотрим что содержится

```
stream.explain()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> stream.explain()
== Physical Plan ==
WriteToDataSourceV2 org.apache.spark.sql.execution.streaming.sources.MicroBatchWriter@3db55b
9d
+- Scan ExistingRDD[timestamp#76,value#77L]
>>>
```

```
stream.lastProgress
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> stream.lastProgress
{u'stateOperators': [], u'name': None, u'timestamp': u'2021-12-15T20:25:30.000Z', u'processedRowsPe
rSecond': 119.04761904761905, u'inputRowsPerSecond': 1.0, u'numInputRows': 30, u'batchId': 4, u'sou
rces': [{u'description': u'RateSource[rowsPerSecond=1, rampUpTimeSeconds=0, numPartitions=2]', u'en
dOffset': 105, u'processedRowsPerSecond': 119.04761904761905, u'inputRowsPerSecond': 1.0, u'numInpu
tRows': 30, u'startOffset': 75}], u'durationMs': {u'queryPlanning': 7, u'getOffset': 0, u'addBatch'
: 195, u'getBatch': 9, u'walCommit': 38, u'triggerExecution': 252}, u'runId': u'696e9b1d-fb16-444d-
90a1-4f3480621614', u'id': u'4baf3bf7-ffc4-4f5f-99c6-43710e4ef2e8', u'sink': {u'description': u'org
.apache.spark.sql.execution.streaming.ConsoleSinkProvider@68eff2fc'}}
>>>
```

```
stream.status
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> stream.status
{u'message': u'Stopped', u'isTriggerActive': False, u'isDataAvailable': False}
>>>
```

Добавим метод для вывода стрима в консоль.

```
def console_output(df, freq):

      return df.writeStream \

            .format("console") \

            .trigger(processingTime='%s seconds' % freq ) \

            .options(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 ) \
...             .options(truncate=False) \
...             .start()
...
>>>
```

Смотрим результат

```
out = console_output(raw_rate, 10)
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> out = console_output(raw_rate, 10)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+---------+-----+
|timestamp|value|
+---------+-----+
+---------+-----+

>>> -------------------------------------------
Batch: 1
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 20:44:43.915|0    |
|2021-12-15 20:44:44.915|1    |
|2021-12-15 20:44:45.915|2    |
|2021-12-15 20:44:46.915|3    |
|2021-12-15 20:44:47.915|4    |
|2021-12-15 20:44:48.915|5    |
+-----------------------+-----+

>>>
```

```
out.stop()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
|2021-12-15 20:45:57.915|74   |
|2021-12-15 20:45:58.915|75   |
+-----------------------+-----+

-------------------------------------------
Batch: 9
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 20:45:59.915|76   |
|2021-12-15 20:46:00.915|77   |
|2021-12-15 20:46:01.915|78   |
|2021-12-15 20:46:02.915|79   |
|2021-12-15 20:46:03.915|80   |
|2021-12-15 20:46:04.915|81   |
|2021-12-15 20:46:05.915|82   |
|2021-12-15 20:46:06.915|83   |
|2021-12-15 20:46:07.915|84   |
|2021-12-15 20:46:08.915|85   |
+-----------------------+-----+

out.stop()
>>>
```

добавим фильтр к нашему потоку

```
from pyspark.sql import functions as F
```

Напишем сам фильтр к стриму

```
filtered_rate = raw_rate \

      .filter( F.col("value") % F.lit("2") == 0 )
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> filtered_rate = raw_rate \
... .filter( F.col("value") % F.lit("2") == 0 )
>>>
```

```
out = console_output(filtered_rate, 10)

out.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терм

Файл Правка Вид Терминал Вкладки Справка

```
>>> out = console_output(filtered_rate, 10)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+---------+-----+
|timestamp|value|
+---------+-----+
+---------+-----+

>>> -------------------------------------------
Batch: 1
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 21:05:30.692|0    |
|2021-12-15 21:05:32.692|2    |
|2021-12-15 21:05:34.692|4    |
|2021-12-15 21:05:36.692|6    |
|2021-12-15 21:05:38.692|8    |
+-----------------------+-----+

>>> -------------------------------------------
Batch: 2
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 21:05:40.692|10   |
|2021-12-15 21:05:42.692|12   |
|2021-12-15 21:05:44.692|14   |
|2021-12-15 21:05:46.692|16   |
|2021-12-15 21:05:48.692|18   |
+-----------------------+-----+

>>> -------------------------------------------
Batch: 3
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 21:05:50.692|20   |
|2021-12-15 21:05:52.692|22   |
|2021-12-15 21:05:54.692|24   |
|2021-12-15 21:05:56.692|26   |
|2021-12-15 21:05:58.692|28   |
+-----------------------+-----+

-------------------------------------------
Batch: 4
-------------------------------------------
+-----------------------+-----+
|timestamp              |value|
+-----------------------+-----+
|2021-12-15 21:06:00.692|30   |
|2021-12-15 21:06:02.692|32   |
|2021-12-15 21:06:04.692|34   |
|2021-12-15 21:06:06.692|36   |
|2021-12-15 21:06:08.692|38   |
+-----------------------+-----+

out.stop()
>>>
```

Добавим к стриму ещё одну колонку

```
extra_rate = filtered_rate \

      .withColumn("my_value", F.when((F.col("value") % F.lit(10) == 0),
F.lit("jubilee")).otherwise(F.lit("not yet")))
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
out.stop()
>>> extra_rate = filtered_rate \
... .withColumn("my_value",
...             F.when((F.col("value") % F.lit(10) == 0), F.lit("jubilee"))
...                     .otherwise(F.lit("not yet")))
>>>
```

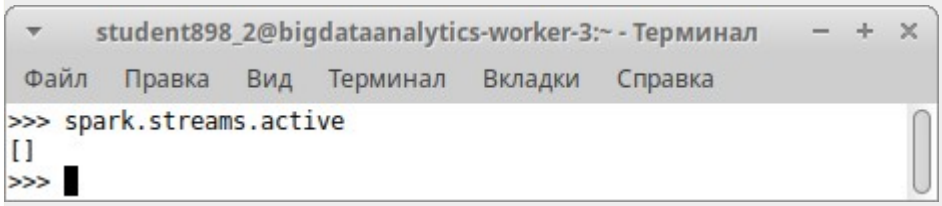
```
out = console_output(extra_rate, 10)
```

student898_2@bigdataanalytics-worker-3:~ - Te

Файл Правка Вид Терминал Вкладки Справка

```
>>> out = console_output(extra_rate, 10)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+---------+-----+--------+
|timestamp|value|my_value|
+---------+-----+--------+
+---------+-----+--------+

-------------------------------------------
Batch: 1
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:16:27.431|0    |jubilee |
+-----------------------+-----+--------+

>>> -------------------------------------------
Batch: 2
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:16:29.431|2    |not yet |
|2021-12-15 21:16:31.431|4    |not yet |
|2021-12-15 21:16:33.431|6    |not yet |
|2021-12-15 21:16:35.431|8    |not yet |
|2021-12-15 21:16:37.431|10   |jubilee |
+-----------------------+-----+--------+

-------------------------------------------
Batch: 3
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:16:39.431|12   |not yet |
|2021-12-15 21:16:41.431|14   |not yet |
|2021-12-15 21:16:43.431|16   |not yet |
|2021-12-15 21:16:45.431|18   |not yet |
|2021-12-15 21:16:47.431|20   |jubilee |
+-----------------------+-----+--------+

-------------------------------------------
Batch: 4
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:16:49.431|22   |not yet |
|2021-12-15 21:16:51.431|24   |not yet |
|2021-12-15 21:16:53.431|26   |not yet |
|2021-12-15 21:16:55.431|28   |not yet |
|2021-12-15 21:16:57.431|30   |jubilee |
+-----------------------+-----+--------+

out.stop()
>>>
```

```
spark.streams.active
```



Проверим как он отработает, если объект вывода стрима в консоль не сохранить в переменную.

```
console_output(extra_rate, 10)

spark.streams.active[0].stop()

spark.streams.active
```

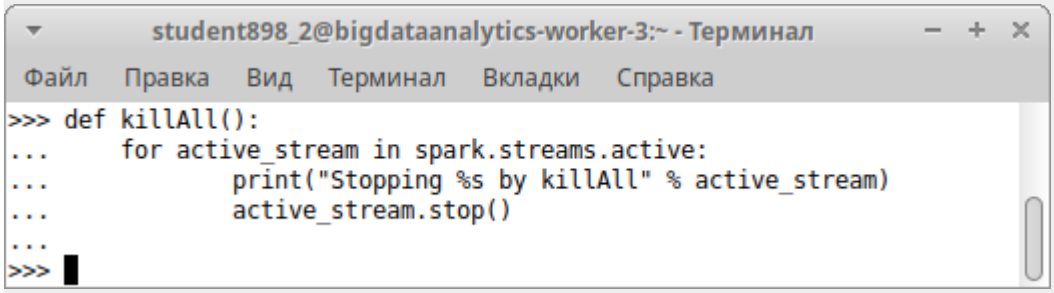
student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
-------------------------------------------
Batch: 40
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:28:00.997|384  |not yet |
|2021-12-15 21:28:02.997|386  |not yet |
|2021-12-15 21:28:04.997|388  |not yet |
|2021-12-15 21:28:06.997|390  |jubilee |
|2021-12-15 21:28:08.997|392  |not yet |
+-----------------------+-----+--------+

-------------------------------------------
Batch: 41
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:28:10.997|394  |not yet |
|2021-12-15 21:28:12.997|396  |not yet |
|2021-12-15 21:28:14.997|398  |not yet |
|2021-12-15 21:28:16.997|400  |jubilee |
|2021-12-15 21:28:18.997|402  |not yet |
+-----------------------+-----+--------+

-------------------------------------------
Batch: 42
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:28:20.997|404  |not yet |
|2021-12-15 21:28:22.997|406  |not yet |
|2021-12-15 21:28:24.997|408  |not yet |
|2021-12-15 21:28:26.997|410  |jubilee |
|2021-12-15 21:28:28.997|412  |not yet |
+-----------------------+-----+--------+

-------------------------------------------
Batch: 43
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:28:30.997|414  |not yet |
|2021-12-15 21:28:32.997|416  |not yet |
|2021-12-15 21:28:34.997|418  |not yet |
|2021-12-15 21:28:36.997|420  |jubilee |
|2021-12-15 21:28:38.997|422  |not yet |
+-----------------------+-----+--------+

spark.streams.active[0].stop()
>>> spark.streams.active[0].stop()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
IndexError: list index out of range
>>> spark.streams.active
[]
>>>
```

Метод для прекращения всех активных стримов

```
def killAll():

    for active_stream in spark.streams.active:

        print("Stopping %s by killAll" % active_stream)

        active_stream.stop()
```



Проверим как он отработает, если объект вывода стрима в консоль не сохранить в переменную.

```
console_output(extra_rate, 10)

killAll()

spark.streams.active
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:36:20.505|44   |not yet |
|2021-12-15 21:36:22.505|46   |not yet |
|2021-12-15 21:36:24.505|48   |not yet |
|2021-12-15 21:36:26.505|50   |jubilee |
|2021-12-15 21:36:28.505|52   |not yet |
+-----------------------+-----+--------+

-------------------------------------------
Batch: 7
-------------------------------------------
+-----------------------+-----+--------+
|timestamp              |value|my_value|
+-----------------------+-----+--------+
|2021-12-15 21:36:30.505|54   |not yet |
|2021-12-15 21:36:32.505|56   |not yet |
|2021-12-15 21:36:34.505|58   |not yet |
|2021-12-15 21:36:36.505|60   |jubilee |
|2021-12-15 21:36:38.505|62   |not yet |
+-----------------------+-----+--------+

killAll()
Stopping <pyspark.sql.streaming.StreamingQuery object at 0x7fac
87795350> by killAll
>>> killAll()
>>> spark.streams.active
[]
>>>
```

Открываю другое окно терминала

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

hdfs dfs -ls

hdfs dfs -mkdir input_csv_for_stream
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalytics-worker-3:~ | student898_2@bigdataanalytics-worker-3:~

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 1 items
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 20:04 .sparkStaging
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -mkdir input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 2 items
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 20:04 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 21:56 input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ ls
[student898_2@bigdataanalytics-worker-3 ~]$ lsinput_csv_for_stream/
-bash: lsinput_csv_for_stream/: Нет такого файла или каталога
[student898_2@bigdataanalytics-worker-3 ~]$ lsinput_csv_for_stream
-bash: lsinput_csv_for_stream: команда не найдена
[student898_2@bigdataanalytics-worker-3 ~]$ vi for_stream/product_[stu[studen[st[[s[st[s[[[[s[s[s[[stu[s[[s[s
tudent898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -put *.csv input_csv_for_stream
put: `*.csv': No such file or directory
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -mkdir for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ ls
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 20:04 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 21:56 input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
hdfs dfs -ls

mkdir for_stream

ls

ls for_stream/
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

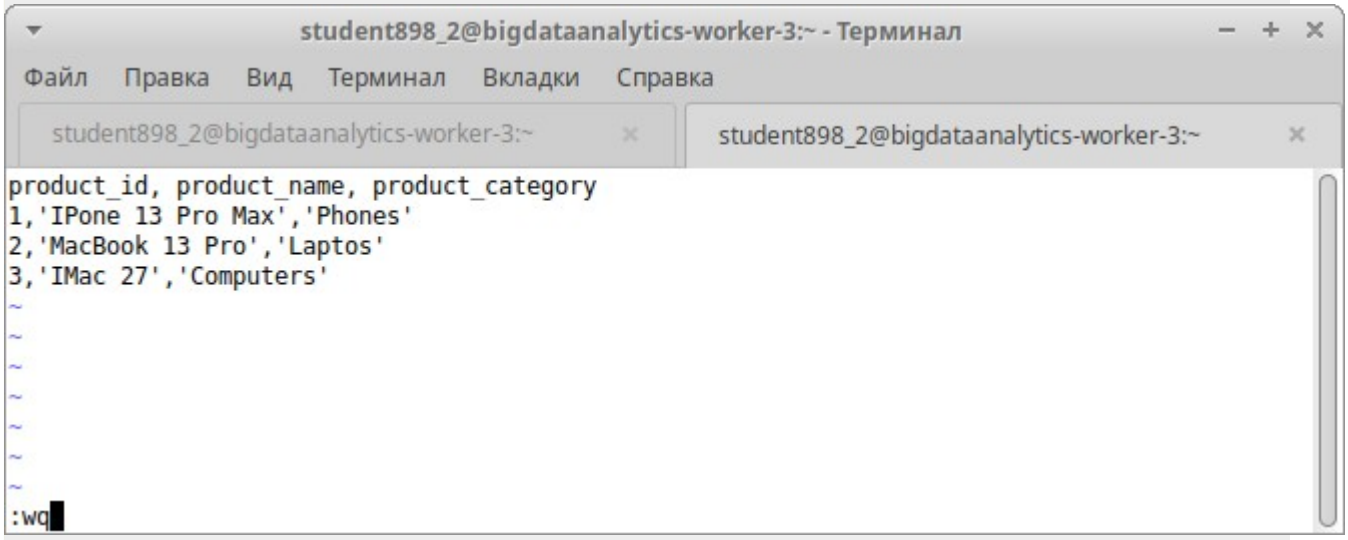
Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalytics-worker-3:~ | student898_2@bigdataanalytics-worker-3:~

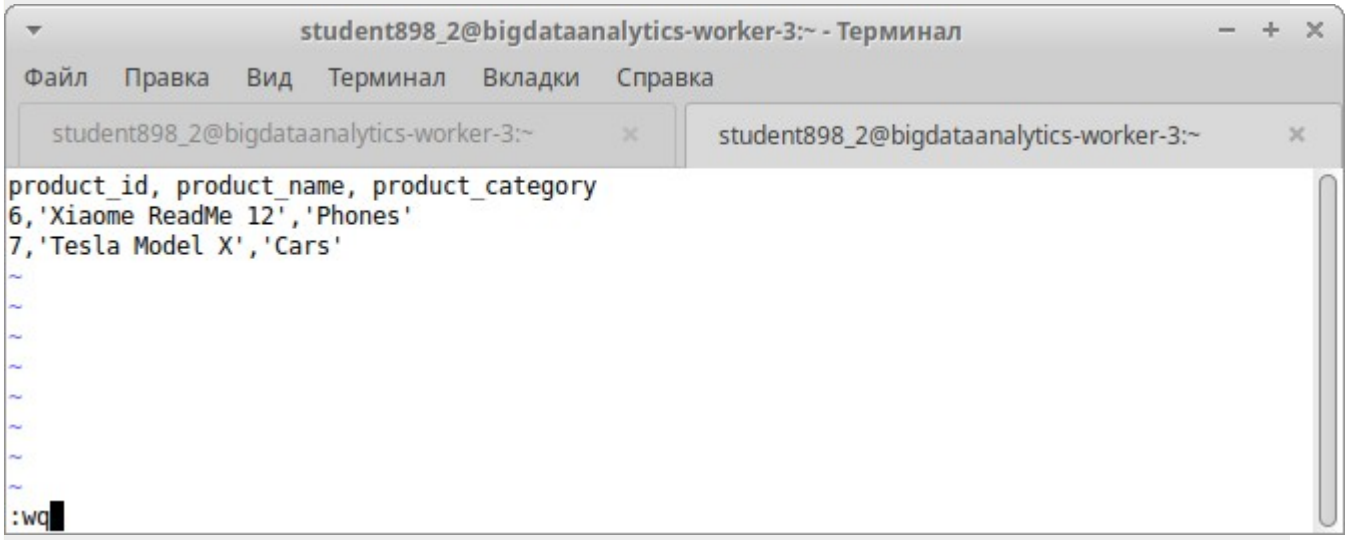
```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - student898_2 student898_2          0 2021-12-16 17:22 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 21:56 input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ mkdir for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ ls
for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ ls for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ ls for_stream/
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
vi for_stream/product_list.csv
```

```
product_id, product_name, product_category

1,'IPone 13 Pro Max','Phones'

2,'MacBook 13 Pro','Laptos'

3,'IMac 27','Computers'
```




```
vi for_stream/product_list1.csv

product_id, product_name, product_category

6,'Xiaome ReadMe 12','Phones'

7,'Tesla Model X','Cars'
```




```
Перемещаем файлы

hdfs dfs -put for_stream/product_list.csv input_csv_for_stream
```

```
hdfs dfs -ls input_csv_for_stream
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalytics-worker-3:~ | student898_2@bigdataanalytics-worker-3:~

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -put for_stream/product_list.csv input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 3 items
drwxr-xr-x   - student898_2 student898_2          0 2021-12-16 17:22 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2021-12-16 18:24 input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls input_csv_for_stream
Found 1 items
-rw-r--r--   2 student898_2 student898_2        125 2021-12-16 18:24 input_csv_for_stream/product_list.csv
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
hdfs dfs -put for_stream/product_list1.csv input_csv_for_stream

hdfs dfs -ls input_csv_for_stream
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalytics-worker-3:~ | student898_2@bigdataanalytics-worker-3:~

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -put for_stream/product_list1.csv input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls input_csv_for_stream
Found 2 items
-rw-r--r--   2 student898_2 student898_2        125 2021-12-16 18:24 input_csv_for_stream/product_list.csv
-rw-r--r--   2 student898_2 student898_2         98 2021-12-16 18:45 input_csv_for_stream/product_list1.csv
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
Требуется схема

from pyspark.sql.types import StructType, StringType

df = spark.sql("select 1 as id, 'Big' as name")

df.show()

df.printSchema()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

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

```
>>> df = spark.sql("select 1 as id, 'Big' as name"
... )
>>> df.show
<bound method DataFrame.show of DataFrame[id: int, name: string]>
>>> df.show()
+---+----+
| id|name|
+---+----+
|  1| Big|
+---+----+

>>>
Traceback (most recent call last):
  File "/usr/hdp/current/spark2-client/python/pyspark/context.py", line 261, in signal_handler
    raise KeyboardInterrupt()
KeyboardInterrupt
>>> df.printShema()
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
  File "/usr/hdp/current/spark2-client/python/pyspark/sql/dataframe.py", line 1182, in __getattr__
    "'%s' object has no attribute '%s'" % (self.__class__.__name__, name))
AttributeError: 'DataFrame' object has no attribute 'printShema'
>>> df.printSchema()
root
 |-- id: integer (nullable = false)
 |-- name: string (nullable = false)

>>>
```

Создаю схему:

```
schema = StructType() \

    .add("product_name", StringType()) \

    .add("product_category", StringType())
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

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

```
>>> schema = StructType() \
... .add("product_name", StringType()) \
... .add("product_category", StringType())
>>> schema.show()
```

Возвращаю метод который пишет в консоль

```
def console_output(df, freq):

    return df.writeStream \
```

```
        .format("console") \

        .trigger(processingTime='%s seconds' % freq ) \

        .options(truncate=False) \

        .start()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalyt... student898_2@bigdataanalyt...

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

читаем все разом

```
raw_files = spark \

    .readStream \

    .format("csv") \

    .schema(schema) \

    .options(path="input_csv_for_stream", header=True) \

    .load()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalyt... student898_2@bigdataanalyt...

```
...
>>> raw_files = spark \
...     .readStream \
...     .format("csv") \
...     .schema(schema) \
...     .options(path="input_csv_for_stream", header=True) \
...     .load()
>>>
```

```
out = console_output(raw_files, 15)
```

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

...     .options(path="input_csv_for_stream", header=True) \
...     .load()
>>> out = console_output(raw_files, 15)
[Stage 14:>                                                         [Stag
e 14:============================>                          [Stage 14:
============================>
                                                    -------------------
----------------------
Batch: 0
-------------------------------------------
+------------+-----------------+
|product_name|product_category  |
+------------+-----------------+
|1           |'IPone 13 Pro Max'|
|2           |'MacBook 13 Pro'  |
|3           |'IMac 27'         |
|6           |'Xiaome ReadMe 12'|
|7           |'Tesla Model X'   |
+------------+-----------------+

>>>
```

Каждые 15 сек он обрабатывает, но поскольку новых данных не поступает, он ждет

Во второй консоли

```
cd for_stream/

cp product_list.csv product_list2.csv

cp product_list.csv product_list3.csv

cp product_list.csv product_list4.csv

ls
```

```
student898_2@bigdataanalytics-worker-3:~/for_stream - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
student898_2@bigdataanalytics-worker-3:~   student898_2@bigdataanalytics-worker-3:~/for_st...

[student898_2@bigdataanalytics-worker-3 ~]$ cd for_stream/
[student898_2@bigdataanalytics-worker-3 for_stream]$ ls
product_list1.csv  product_list.csv
[student898_2@bigdataanalytics-worker-3 for_stream]$ cp product_list.csv product_list2.csv
[student898_2@bigdataanalytics-worker-3 for_stream]$ cp product_list.csv product_list3.csv
[student898_2@bigdataanalytics-worker-3 for_stream]$ cp product_list.csv product_list4.csv
[student898_2@bigdataanalytics-worker-3 for_stream]$ ls
product_list1.csv  product_list2.csv  product_list3.csv  product_list4.csv  product_list.csv
[student898_2@bigdataanalytics-worker-3 for_stream]$
```

```
vi product_list3.csv
```

student898_2@bigdataanalytics-worker-3:~/for_stream - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalytics-worker-3:~ | student898_2@bigdataanalytics-worker-3:~/for_stream

```
product_id, product_name, product_category
10,'IPone 13 Pro Max','Phones'
20,'MacBook 13 Pro','Laptos'
30,'IMac 27','Computers'
~
~
~
~
~
~
~
~
:wq
```

```
vi product_list4.csv
```

student898_2@bigdataanalytics-worker-3:~/for_stream - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalytics-worker-3:~ | student898_2@bigdataanalytics-worker-3:~/for_stream

```
product_id, product_name, product_category
12,'IPone 13 Pro Max','Phones'
22,'MacBook 13 Pro','Laptos'
32,'IMac 27','Computers'
~
~
~
~
~
~
~
~
:wq
```

```
vi product_list2.csv
```

student898_2@bigdataanalytics-worker-3:~/for_stream - Терминал

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataanalytics-worker-3:~ | student898_2@bigdataanalytics-worker-3:~/for_stream

```
product_id, product_name, product_category
9,'IPone 13 Pro Max','Phones'
8,'MacBook 13 Pro','Laptos'
16,'IMac 27','Computers'
~
~
~
~
~
~
~
~
:wq
```

```
hdfs dfs -put for_stream/*.csv input_csv_for_stream

out.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Термин

Файл Правка Вид Терминал Вкладки Справка

student898_2@bigdataa... | student898_2@bigdataa...

```
+------------+-----------------+

>>> out.stop()
>>> out = console_output(raw_files, 15)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+------------+-----------------+
|product_name|product_category  |
+------------+-----------------+
|10          |'IPone 13 Pro Max'|
|20          |'MacBook 13 Pro'  |
|30          |'IMac 27'         |
|12          |'IPone 13 Pro Max'|
|22          |'MacBook 13 Pro'  |
|32          |'IMac 27'         |
|9           |'IPone 13 Pro Max'|
|8           |'MacBook 13 Pro'  |
|16          |'IMac 27'         |
|1           |'IPone 13 Pro Max'|
|2           |'MacBook 13 Pro'  |
|3           |'IMac 27'         |
|6           |'Xiaome ReadMe 12'|
|7           |'Tesla Model X'   |
+------------+-----------------+

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

```
#чтение по одному
```

```
raw_files = spark \

    .readStream \

    .format("csv") \

    .schema(schema) \

    .options(path="input_csv_for_stream", header=True,
maxFilesPerTrigger=1) \

    .load()

out = console_output(raw_files, 15)

out.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

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

```
>>> raw_files = spark \
...     .readStream \
...     .format("csv") \
...     .schema(schema) \
...     .options(path="input_csv_for_stream", header=True, maxFilesPerTrigger=1) \
...     .load()
>>> out = console_output(raw_files, 15)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+------------+------------------+
|product_name|product_category  |
+------------+------------------+
|1           |'IPone 13 Pro Max'|
|2           |'MacBook 13 Pro'  |
|3           |'IMac 27'         |
+------------+------------------+

-------------------------------------------
Batch: 1
-------------------------------------------
+------------+------------------+
|product_name|product_category  |
+------------+------------------+
|6           |'Xiaome ReadMe 12'|
|7           |'Tesla Model X'   |
+------------+------------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+------------+------------------+
|product_name|product_category  |
+------------+------------------+
|9           |'IPone 13 Pro Max'|
|8           |'MacBook 13 Pro'  |
|16          |'IMac 27'         |
+------------+------------------+

>>> -------------------------------------------
Batch: 3
-------------------------------------------
+------------+------------------+
|product_name|product_category  |
+------------+------------------+
|10          |'IPone 13 Pro Max'|
|20          |'MacBook 13 Pro'  |
|30          |'IMac 27'         |
+------------+------------------+

>>> out.stop()-------------------------------------------
Batch: 4
-------------------------------------------
+------------+------------------+
|product_name|product_category  |
+------------+------------------+
|12          |'IPone 13 Pro Max'|
|22          |'MacBook 13 Pro'  |
|32          |'IMac 27'         |
+------------+------------------+

out.stop()
  File "<stdin>", line 1
    out.stop()out.stop()
               ^
SyntaxError: invalid syntax
>>> out.stop()
>>>
```

Закрываем подключение к кластеру

```
exit
```