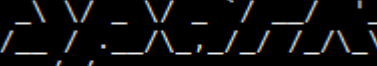


1. Запустил spark 2.4.7.

```
07/12/19 18:52:17 WARN Util.scala: Service SparkUI could not start.  
Welcome to  
 version 2.4.7  
Using Python version 2.7.5 (default, Apr 2 2020 13:16:51)  
SparkSession available as 'spark'.  
>>>
```

2. Сделал вывод в память и прочитал.

```
>>> 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_iris,5)
>>> spark.sql("select * from my_memory_sink_table").show()
+-----+-----+-----+-----+-----+-----+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|
+-----+-----+-----+-----+-----+-----+
|4.7|3.2|1.3|0.2|setosa|0|
|5.4|3.9|1.7|0.4|setosa|1|
|4.4|2.9|1.4|0.2|setosa|2|
|4.8|3.4|1.6|0.2|setosa|3|
|5.8|4.0|1.2|0.2|setosa|4|
|5.1|3.5|1.4|0.3|setosa|5|
|5.4|3.4|1.7|0.2|setosa|6|
|5.1|3.3|1.7|0.5|setosa|7|
|5.0|3.4|1.6|0.4|setosa|8|
|4.7|3.2|1.6|0.2|setosa|9|
|5.2|4.1|1.5|0.1|setosa|10|
|5.0|3.2|1.2|0.2|setosa|11|
|4.4|3.0|1.3|0.2|setosa|12|
|4.5|2.3|1.3|0.3|setosa|13|
|5.1|3.8|1.9|0.4|setosa|14|
|4.6|3.2|1.4|0.2|setosa|15|
|7.0|3.2|4.7|1.4|versicolor|16|
|5.5|2.3|4.0|1.3|versicolor|17|
|6.3|3.3|4.7|1.6|versicolor|18|
|5.2|2.7|3.9|1.4|versicolor|19|
+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

3. Сделал вывод в файл

```
>>> def file_sink(df, freq):
...     return df.writeStream.format("parquet") \
...         .trigger(processingTime='%s seconds' % freq) \
...         .option("path", "my_parquet_sink_iris") \
...         .option("checkpointLocation", "my_parquet_checkpoint") \
...         .start()
...
>>> stream = file_sink(parsed_iris, 1)
20/12/19 21:26:54 WARN streaming.ProcessingTimeExecutor: Current batch is falling behind. The trigger interval is 1000 milliseconds, but spent 4739 milliseconds
```

4. Убедился что записалось.

```
[BD_243_pstroganov@bigdataanalytics-worker-0 ~]$ hdfs dfs -ls
Found 5 items
drwx----- - BD_243_pstroganov BD_243_pstroganov      0 2020-12-16 12:00 .Trash
drwxr-xr-x - BD_243_pstroganov BD_243_pstroganov      0 2020-12-16 23:16 .sparkStaging
drwxr-xr-x - BD_243_pstroganov BD_243_pstroganov      0 2020-12-16 00:30 csv_stream
drwxr-xr-x - BD_243_pstroganov BD_243_pstroganov      0 2020-12-19 21:26 my_parquet_checkpoint
drwxr-xr-x - BD_243_pstroganov BD_243_pstroganov      0 2020-12-19 21:26 my_parquet_sink_iris
[BD_243_pstroganov@bigdataanalytics-worker-0 ~]$ hdfs dfs -ls my_parquet_sink_iris
Found 4 items
drwxr-xr-x - BD_243_pstroganov BD_243_pstroganov      0 2020-12-19 21:26 my_parquet_sink_iris/_spark_metadata
-rw-r--r--  3 BD_243_pstroganov BD_243_pstroganov    2491 2020-12-19 21:26 my_parquet_sink_iris/part-00000-6f5bfc73-014c-48c7-8efb-87ddfe645e46-c000.snappy.parquet
-rw-r--r--  3 BD_243_pstroganov BD_243_pstroganov    2467 2020-12-19 21:26 my_parquet_sink_iris/part-00001-8457d199-cc60-4370-8d69-66d7947ed478-c000.snappy.parquet
-rw-r--r--  3 BD_243_pstroganov BD_243_pstroganov    2505 2020-12-19 21:26 my_parquet_sink_iris/part-00002-14f61602-e67d-4288-8812-cf54af62c803-c000.snappy.parquet
[BD_243_pstroganov@bigdataanalytics-worker-0 ~]$
```

5. Включил перезапись.

```
>>> def file_sink(df, freq):
...     return df.writeStream.format("parquet") \
...         .trigger(processingTime='%s seconds' % freq) \
...         .option("path", "my_parquet_sink_iris") \
...         .option("checkpointLocation", "my_parquet_checkpoint") \
...         .start()
...
>>> stream = file_sink(parsed_iris, 1)
>>> 20/12/19 21:36:53 WARN streaming.ProcessingTimeExecutor: Current batch is falling behind. The trigger interval is 1000 milliseconds, but spent 3075 milliseconds
```

6. Убедился что работает.

```
Found 2 items
-rw-r--r--  3 BD_243_pstroganov BD_243_pstroganov      0 2020-12-19 21:33 my_parquet_sink_iris/_SUCCESS
-rw-r--r--  3 BD_243_pstroganov BD_243_pstroganov    3020 2020-12-19 21:32 my_parquet_sink_iris/part-00000-657d5e98-b56d-4f75-a331-55fd9a845e7f-c000.snappy.parquet
[BD_243_pstroganov@bigdataanalytics-worker-0 ~]$ hdfs dfs -ls my_parquet_sink_iris
Found 3 items
-rw-r--r--  3 BD_243_pstroganov BD_243_pstroganov      0 2020-12-19 21:34 my_parquet_sink_iris/_SUCCESS
drwxr-xr-x - BD_243_pstroganov BD_243_pstroganov      0 2020-12-19 21:34 my_parquet_sink_iris/_spark_metadata
-rw-r--r--  3 BD_243_pstroganov BD_243_pstroganov    3020 2020-12-19 21:34 my_parquet_sink_iris/part-00000-edf516c2-69e6-4b3f-a1f9-1ef6a242e91f-c000.snappy.parquet
```

7. Сделал слив в кафку.

```
>>> 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", "kafka_sink") \
...         .option("kafka.bootstrap.servers", kafka_brokers) \
...         .option("checkpointLocation", "my_kafka_checkpoint") \
...         .start()
...
>>> stream = kafka_sink(parsed_iris, 5)
```