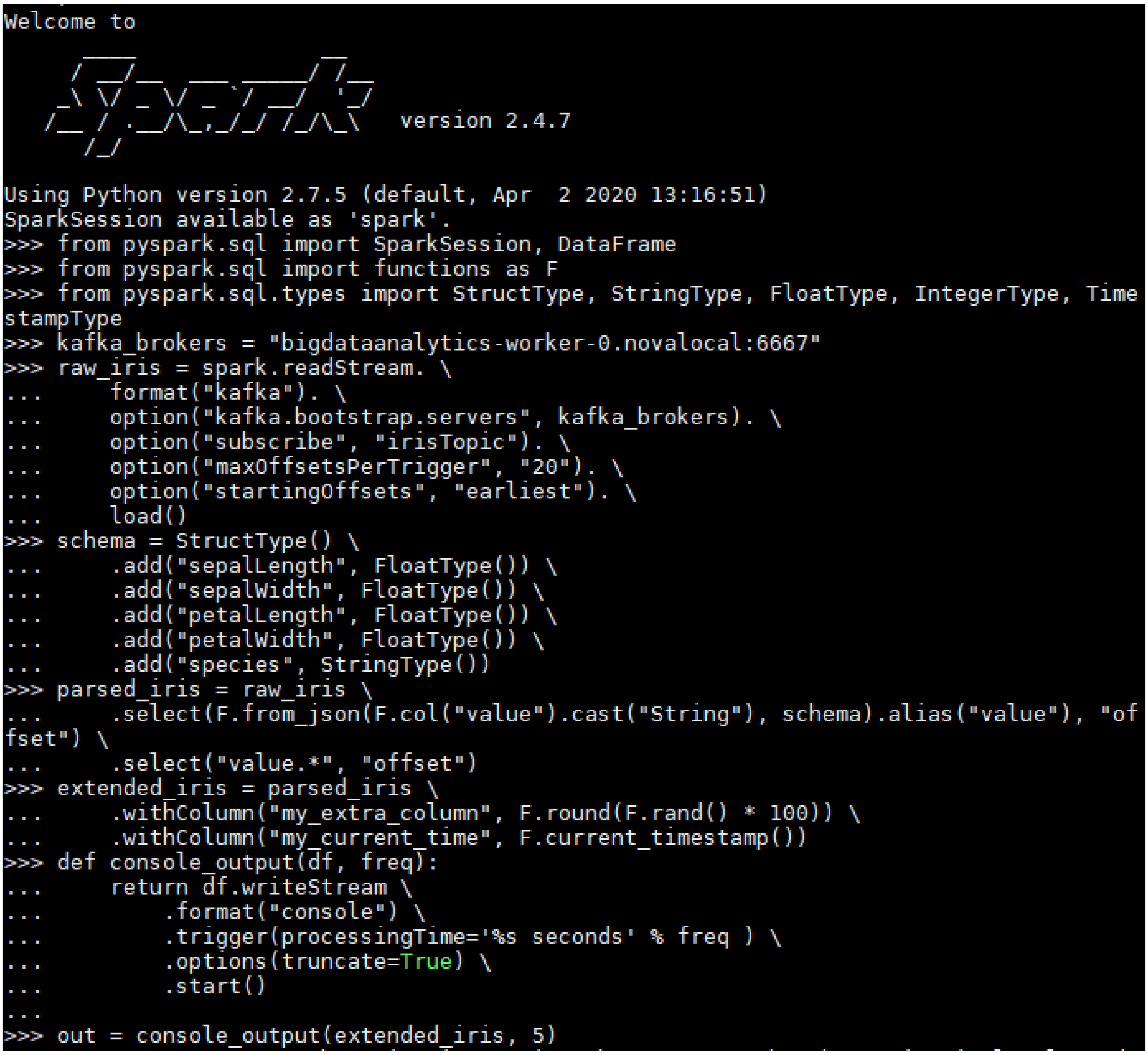
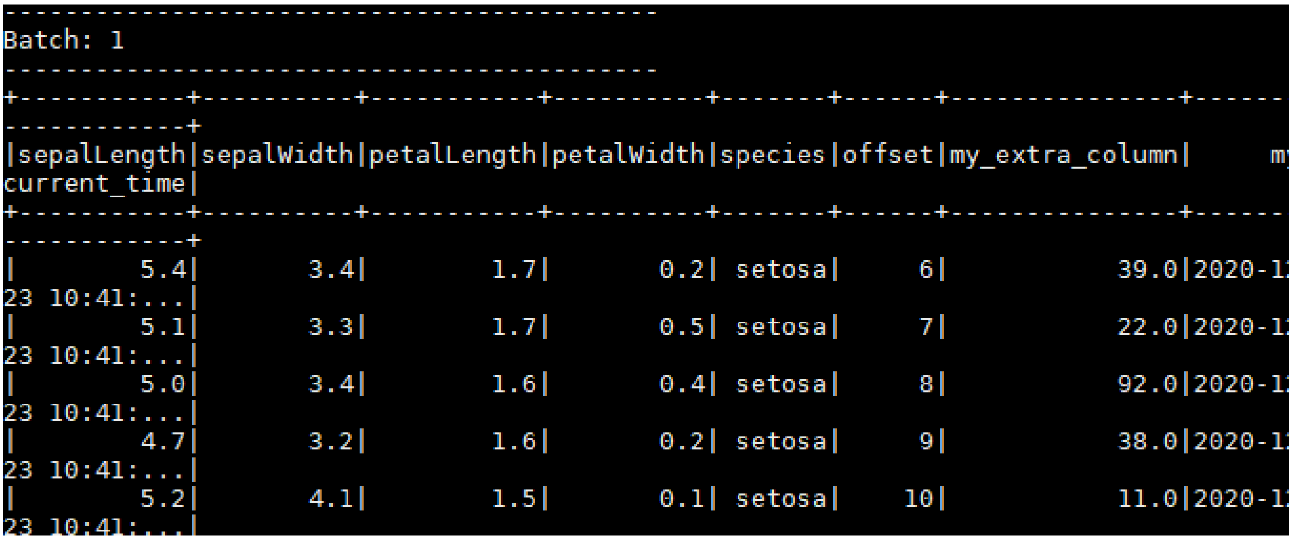
1. Запустил спарк. Спарсил irisTopic и сделал датафрейм дополненный двумя столбцами как на занятии.



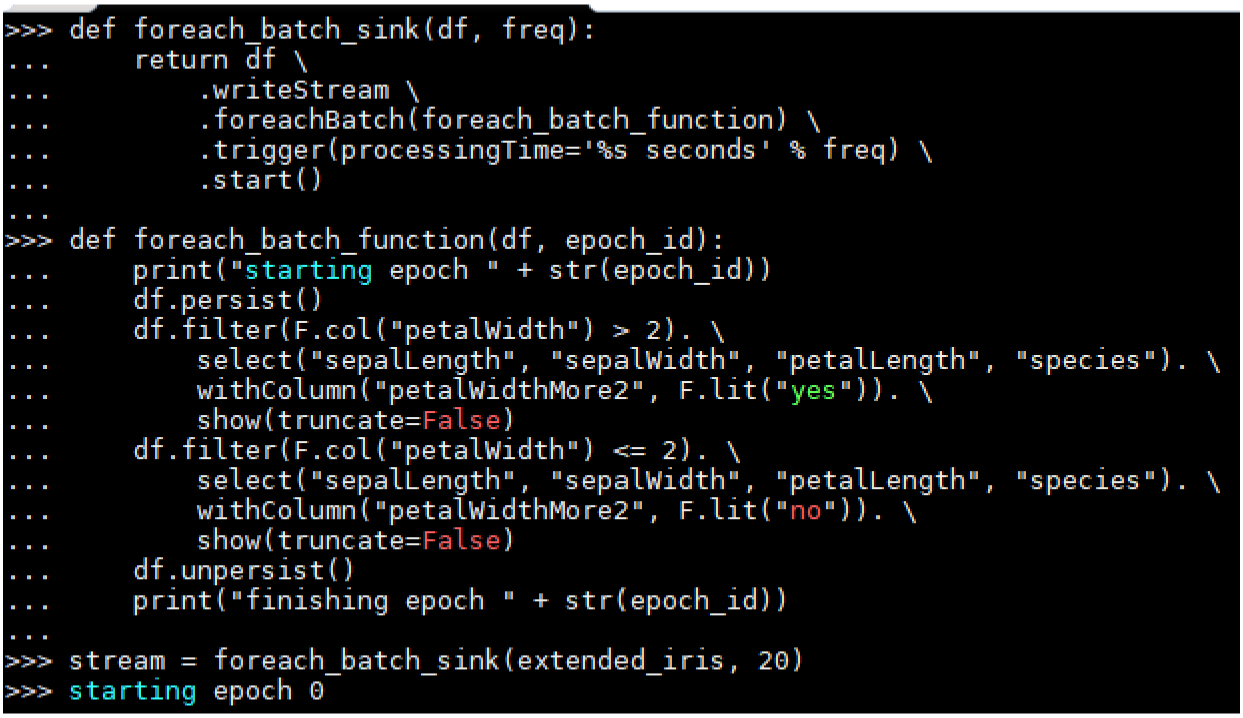
```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.7
      /_/

Using Python version 2.7.5 (default, Apr  2 2020 13:16:51)
SparkSession available as 'spark'.
>>> from pyspark.sql import SparkSession, DataFrame
>>> from pyspark.sql import functions as F
>>> from pyspark.sql.types import StructType, StringType, FloatType, IntegerType, Time
stampType
>>> kafka_brokers = "bigdataanalytics-worker-0.novalocal:6667"
>>> raw_iris = spark.readStream. \
...     format("kafka"). \
...     option("kafka.bootstrap.servers", kafka_brokers). \
...     option("subscribe", "irisTopic"). \
...     option("maxOffsetsPerTrigger", "20"). \
...     option("startingOffsets", "earliest"). \
...     load()
>>> schema = StructType() \
...     .add("sepalLength", FloatType()) \
...     .add("sepalWidth", FloatType()) \
...     .add("petalLength", FloatType()) \
...     .add("petalWidth", FloatType()) \
...     .add("species", StringType())
>>> parsed_iris = raw_iris \
...     .select(F.from_json(F.col("value").cast("String"), schema).alias("value"), "of
fset") \
...     .select("value.*", "offset")
>>> extended_iris = parsed_iris \
...     .withColumn("my_extra_column", F.round(F.rand() * 100)) \
...     .withColumn("my_current_time", F.current_timestamp())
>>> def console_output(df, freq):
...     return df.writeStream \
...         .format("console") \
...         .trigger(processingTime='%s seconds' % freq ) \
...         .options(truncate=True) \
...         .start()
...
>>> out = console_output(extended_iris, 5)
```



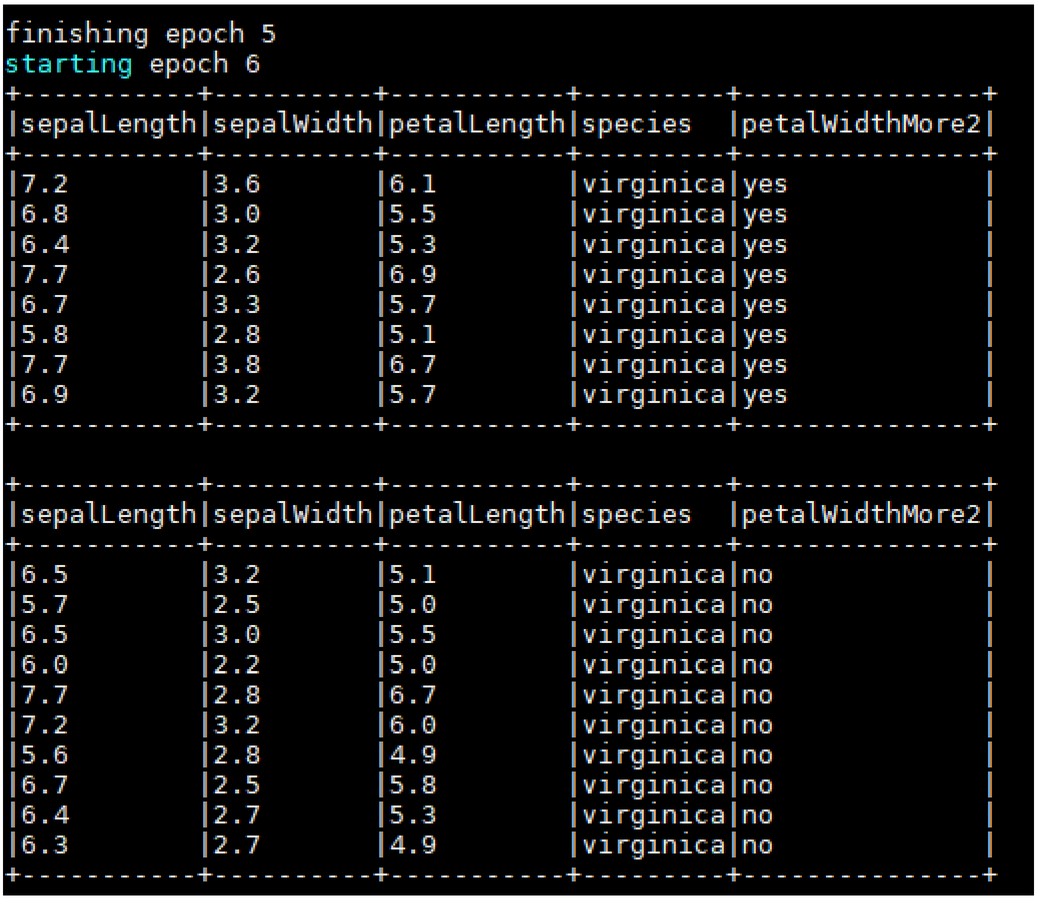
```
-------------------------------------------
Batch: 1
-------------------------------------------
+-----------+----------+-----------+----------+-------+------+---------------+------
------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|my_extra_column|      my
current_time|
+-----------+----------+-----------+----------+-------+------+---------------+------
------------+
|        5.4|       3.4|        1.7|       0.2| setosa|     6|           39.0|2020-1
23 10:41:...|
|        5.1|       3.3|        1.7|       0.5| setosa|     7|           22.0|2020-1
23 10:41:...|
|        5.0|       3.4|        1.6|       0.4| setosa|     8|           92.0|2020-1
23 10:41:...|
|        4.7|       3.2|        1.6|       0.2| setosa|     9|           38.0|2020-1
23 10:41:...|
|        5.2|       4.1|        1.5|       0.1| setosa|    10|           11.0|2020-1
23 10:41:...|
```

2. Далее для каждого батча сделал вывод в два новых датафрейма, в которых в зависимости от значения(больше 2 или нет) petalWidth записывается yes или no в новую колоку petalWidthMore2 .



```
>>> 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))
...     df.persist()
...     df.filter(F.col("petalWidth") > 2). \
...         select("sepalLength", "sepalWidth", "petalLength", "species"). \
...         withColumn("petalWidthMore2", F.lit("yes")). \
...         show(truncate=False)
...     df.filter(F.col("petalWidth") <= 2). \
...         select("sepalLength", "sepalWidth", "petalLength", "species"). \
...         withColumn("petalWidthMore2", F.lit("no")). \
...         show(truncate=False)
...     df.unpersist()
...     print("finishing epoch " + str(epoch_id))
...
>>> stream = foreach_batch_sink(extended_iris, 20)
>>> starting epoch 0
```

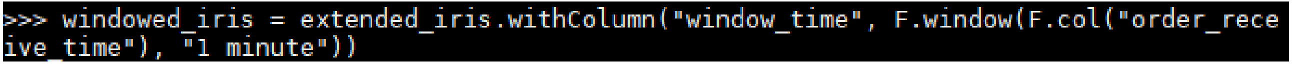
3. Пример разделения данных.



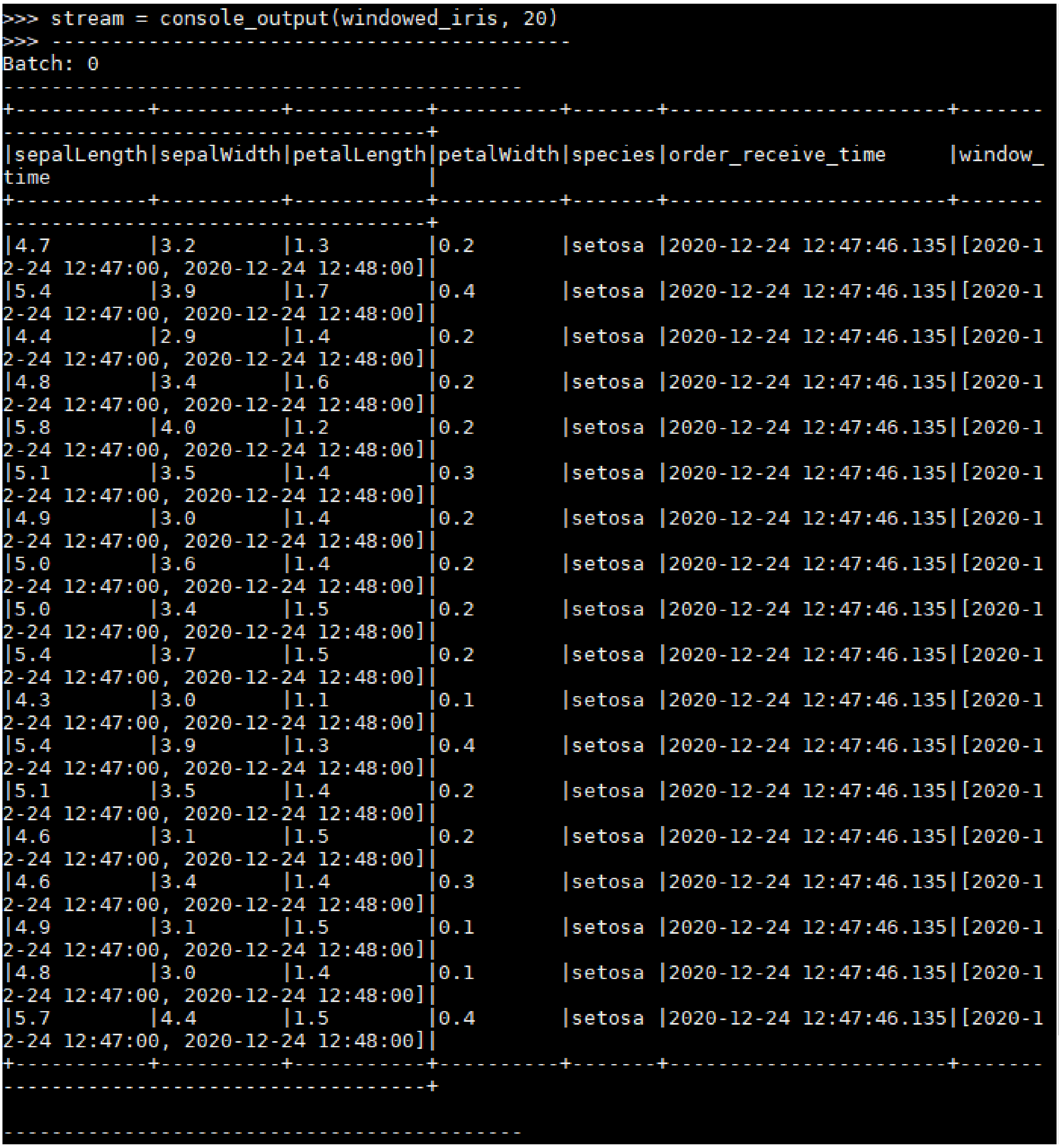
```
finishing epoch 5
starting epoch 6
+-----------+----------+-----------+---------+---------------+
|sepalLength|sepalWidth|petalLength|species  |petalWidthMore2|
+-----------+----------+-----------+---------+---------------+
|7.2        |3.6       |6.1        |virginica|yes            |
|6.8        |3.0       |5.5        |virginica|yes            |
|6.4        |3.2       |5.3        |virginica|yes            |
|7.7        |2.6       |6.9        |virginica|yes            |
|6.7        |3.3       |5.7        |virginica|yes            |
|5.8        |2.8       |5.1        |virginica|yes            |
|7.7        |3.8       |6.7        |virginica|yes            |
|6.9        |3.2       |5.7        |virginica|yes            |
+-----------+----------+-----------+---------+---------------+

+-----------+----------+-----------+---------+---------------+
|sepalLength|sepalWidth|petalLength|species  |petalWidthMore2|
+-----------+----------+-----------+---------+---------------+
|6.5        |3.2       |5.1        |virginica|no             |
|5.7        |2.5       |5.0        |virginica|no             |
|6.5        |3.0       |5.5        |virginica|no             |
|6.0        |2.2       |5.0        |virginica|no             |
|7.7        |2.8       |6.7        |virginica|no             |
|7.2        |3.2       |6.0        |virginica|no             |
|5.6        |2.8       |4.9        |virginica|no             |
|6.7        |2.5       |5.8        |virginica|no             |
|6.4        |2.7       |5.3        |virginica|no             |
|6.3        |2.7       |4.9        |virginica|no             |
+-----------+----------+-----------+---------+---------------+
```

4. С окошками пробовал, получилось.



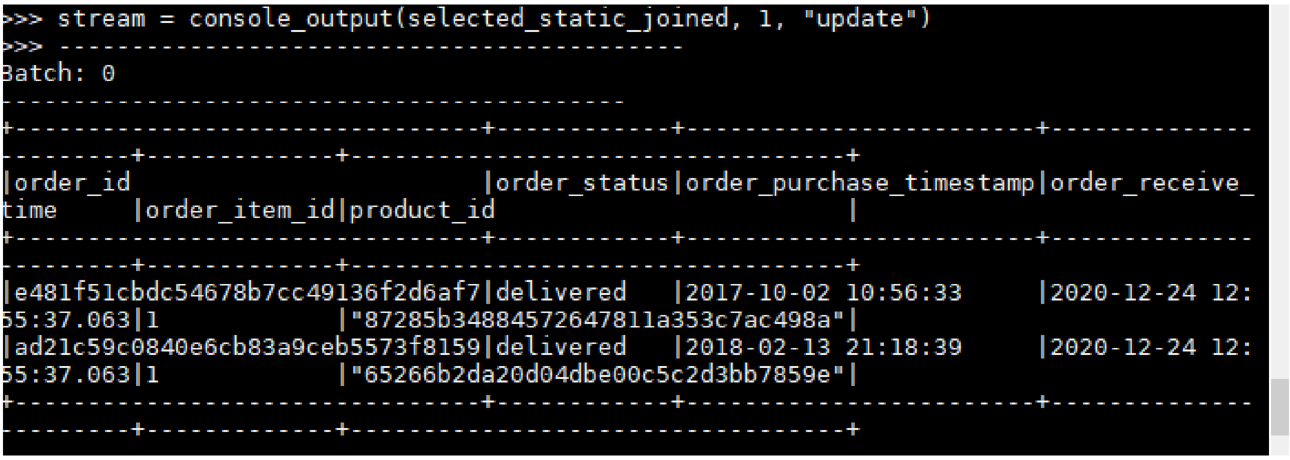
```
>>> windowed_iris = extended_iris.withColumn("window_time", F.window(F.col("order_rece
ive_time"), "1 minute"))
```



```
>>> stream = console_output(windowed_iris, 20)
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+-----------+----------+-----------+----------+-------+-----------------------+-------
-----------------------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|order_receive_time     |window_
time                               |
+-----------+----------+-----------+----------+-------+-----------------------+-------
-----------------------------------+
|4.7        |3.2       |1.3        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.4        |3.9       |1.7        |0.4       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.4        |2.9       |1.4        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.8        |3.4       |1.6        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.8        |4.0       |1.2        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.1        |3.5       |1.4        |0.3       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.9        |3.0       |1.4        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.0        |3.6       |1.4        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.0        |3.4       |1.5        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.4        |3.7       |1.5        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.3        |3.0       |1.1        |0.1       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.4        |3.9       |1.3        |0.4       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.1        |3.5       |1.4        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.6        |3.1       |1.5        |0.2       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.6        |3.4       |1.4        |0.3       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.9        |3.1       |1.5        |0.1       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|4.8        |3.0       |1.4        |0.1       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
|5.7        |4.4       |1.5        |0.4       |setosa |2020-12-24 12:47:46.135|[2020-1
2-24 12:47:00, 2020-12-24 12:48:00]|
+-----------+----------+-----------+----------+-------+-----------------------+-------
-----------------------------------+


-------------------------------------------
```

5. Джойн со статикой пробовал как на уроке на табличке orders. Не успел загрузить свой датасет, поэтому попробовал просто как на уроке с уже загруженным.



```
>>> stream = console_output(selected_static_joined, 1, "update")
>>> -------------------------------------------
Batch: 0
-------------------------------------------
+--------------------------------+------------+------------------------+--------------
---------+-------------+----------------------------------+
|order_id                        |order_status|order_purchase_timestamp|order_receive_
time     |order_item_id|product_id                        |
+--------------------------------+------------+------------------------+--------------
---------+-------------+----------------------------------+
|e481f51cbdc54678b7cc49136f2d6af7|delivered   |2017-10-02 10:56:33     |2020-12-24 12:
55:37.063|1            |"87285b34884572647811a353c7ac498a"|
|ad21c59c0840e6cb83a9ceb5573f8159|delivered   |2018-02-13 21:18:39     |2020-12-24 12:
55:37.063|1            |"65266b2da20d04dbe00c5c2d3bb7859e"|
+--------------------------------+------------+------------------------+--------------
---------+-------------+----------------------------------+
```