5. Spark Streaming. Stateful streams

Загрузить в топик kafka свои данные, прочитать их в потоке, применить watermark и window. Повторить шаги выполненные на занятии.
Дополнительно, объединить статичный и динамичный потоки.
Задание на повышенный бал: Написать скрипт на python для конвертации файла csv в json.

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

Запускаем `pyspark`

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

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
        confs: [default]
        0 artifacts copied, 6 already retrieved (0kB/7ms)
22/01/23 17:49:41 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).
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 2.4.8
      /_/

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

```
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", "shadrin_iris"). \

    option("startingOffsets", "earliest"). \

    option("maxOffsetsPerTrigger", "6"). \

    load()
```

```
Задаём структуру для потока

schema = StructType() \

    .add("sepalLength", FloatType()) \

    .add("sepalWidth", FloatType()) \

    .add("petalLength", FloatType()) \

    .add("petalWidth", FloatType()) \

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

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", "shadrin_iris"). \
...     option("startingOffsets", "earliest"). \
...     option("maxOffsetsPerTrigger", "6"). \
...     load()
>>> raw_data = spark.readStream. \
...     format("kafka"). \
...     option("kafka.bootstrap.servers", kafka_brokers). \
...     option("subscribe", "shadrin_iris"). \
...     option("startingOffsets", "earliest"). \
...     option("maxOffsetsPerTrigger", "6"). \
...     load()
>>>
```

```
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()
...
>>>
```

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

```
забыли сделать плоскую схему

out.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
 61 22 7D 2C]|shadrin_iris|0        |7     |2022-01-17 17:58:40.669|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 35 2E 30 2C 20 22 73 65 70 61 6C 57 69 64 74 68 22 3A 20 33 2E 34 2C 20 22 70 65 74 61 6C
 4C 65 6E 67 74 68 22 3A 20 31 2E 35 2C 20 22 70 65 74 61 6C 57 69 64 74 68 22 3A 20 30 2E 32 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73
 61 22 7D 2C]|shadrin_iris|0        |8     |2022-01-17 17:58:40.669|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 34 2E 34 2C 20 22 73 65 70 61 6C 57 69 64 74 68 22 3A 20 32 2E 39 2C 20 22 70 65 74 61 6C
 4C 65 6E 67 74 68 22 3A 20 31 2E 34 2C 20 22 70 65 74 61 6C 57 69 64 74 68 22 3A 20 30 2E 32 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73
 61 22 7D 2C]|shadrin_iris|0        |9     |2022-01-17 17:58:40.669|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 34 2E 39 2C 20 22 73 65 70 61 6C 57 69 64 74 68 22 3A 20 33 2E 31 2C 20 22 70 65 74 61 6C
 4C 65 6E 67 74 68 22 3A 20 31 2E 35 2C 20 22 70 65 74 61 6C 57 69 64 74 68 22 3A 20 30 2E 31 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73
 61 22 7D 2C]|shadrin_iris|0        |10    |2022-01-17 17:58:40.669|0            |
|null|[20 20 7B 22 73 65 70 61 6C 4C 65 6E 67 74 68 22 3A 20 35 2E 34 2C 20 22 73 65 70 61 6C 57 69 64 74 68 22 3A 20 33 2E 37 2C 20 22 70 65 74 61 6C
 4C 65 6E 67 74 68 22 3A 20 31 2E 35 2C 20 22 70 65 74 61 6C 57 69 64 74 68 22 3A 20 30 2E 32 2C 20 22 73 70 65 63 69 65 73 22 3A 20 22 73 65 74 6F 73
 61 22 7D 2C]|shadrin_iris|0        |11    |2022-01-17 17:58:40.67 |0            |
+----+-------------------------------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------------------------------------------------------------------------------------------------
------------+------------+---------+------+----------------------+------------+

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

```
parsed_iris = 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

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

```
out = console_output(parsed_iris, 10)

out.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
Batch: 1
-------------------------------------------
+-----------+----------+-----------+----------+-------+------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|
+-----------+----------+-----------+----------+-------+------+
|5.4        |3.9       |1.7        |0.4       |setosa |6     |
|4.6        |3.4       |1.4        |0.3       |setosa |7     |
|5.0        |3.4       |1.5        |0.2       |setosa |8     |
|4.4        |2.9       |1.4        |0.2       |setosa |9     |
|4.9        |3.1       |1.5        |0.1       |setosa |10    |
|5.4        |3.7       |1.5        |0.2       |setosa |11    |
+-----------+----------+-----------+----------+-------+------+

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

```
extended_iris = raw_data \
```

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

    .select("value.*", "offset") \

    .withColumn("receive_time", F.current_timestamp())

extended_iris.printSchema()
```

Мы преобразуем джисон объект добавили офсет помимо этого добавляем timestamp, это поле является динамическим

Нам необходимо переработать метод

```
def console_output(df, freq):

    return df.writeStream \

        .format("console") \

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

        .option("checkpointLocation", "checkpoints/duplicates_console_chk") \

        .options(truncate=False) \

        .start()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

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

```
>>> extended_iris = raw_data \
...     .select(F.from_json(F.col("value").cast("String"), schema).alias("value"), "offset") \
...     .select("value.*", "offset") \
...     .withColumn("receive_time", F.current_timestamp())
>>> extended_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
 |-- receive_time: timestamp (nullable = false)

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

В другом окне

hdfs dfs -ls

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
student898_2@bigdataanalytics-worker-3:~   student898_2@bigdataanalytics-worker-3:~
igor@igor-MS-7808:~$ ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176
Last login: Sun Jan 23 20:34:33 2022 from 109-252-19-10.nat.spd-mgts.ru
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 9 items
drwx------   - student898_2 student898_2          0 2022-01-23 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 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 2022-01-23 19:15 my_parquet_sink
drwxr-xr-x   - student898_2 student898_2          0 2022-01-23 19:13 shadrin_iris_file_checkpoint
drwxr-xr-x   - student898_2 student898_2          0 2022-01-23 19:36 shadrin_iris_kafka_checkpoint
drwxr-xr-x   - student898_2 student898_2          0 2022-01-22 22:56 tolstykov_les4_file_checkpoint
drwxr-xr-x   - student898_2 student898_2          0 2022-01-22 23:03 tolstykov_les4_kafka_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$
```

удаляем checkpoints

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

hdfs dfs -ls
```

В первом окне запускаем

```
stream = console_output(extended_iris , 5)

stream.stop()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
student898_2@bigdataanalytics-worker-3:~   student898_2@bigdataanalytics-worker-3:~
+-----------+----------+-----------+----------+----------+------+----------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species   |offset|receive_time          |
+-----------+----------+-----------+----------+----------+------+----------------------+
|5.2        |2.7       |3.9        |1.4       |versicolor|60    |2022-01-23 20:59:00.003|
|5.0        |2.0       |3.5        |1.0       |versicolor|61    |2022-01-23 20:59:00.003|
|5.9        |3.0       |4.2        |1.5       |versicolor|62    |2022-01-23 20:59:00.003|
|6.0        |2.2       |4.0        |1.0       |versicolor|63    |2022-01-23 20:59:00.003|
|6.1        |2.9       |4.7        |1.4       |versicolor|64    |2022-01-23 20:59:00.003|
|5.6        |2.9       |3.6        |1.3       |versicolor|65    |2022-01-23 20:59:00.003|
+-----------+----------+-----------+----------+----------+------+----------------------+

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

Во втором окне наблюдаем обычное наполнение чекпоинта

```
hdfs dfs -du -h  checkpoints/duplicates_console_chk
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
student898_2@bigdataanalytics-worker-3:~   student898_2@bigdataanalytics-worker-3:~
45      90      checkpoints/duplicates_console_chk/metadata
3.0 K   5.9 K   checkpoints/duplicates_console_chk/offsets
28      56      checkpoints/duplicates_console_chk/sources
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h  checkpoints/duplicates_console_chk
232     464     checkpoints/duplicates_console_chk/commits
45      90      checkpoints/duplicates_console_chk/metadata
3.4 K   6.7 K   checkpoints/duplicates_console_chk/offsets
28      56      checkpoints/duplicates_console_chk/sources
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h  checkpoints/duplicates_console_chk
261     522     checkpoints/duplicates_console_chk/commits
45      90      checkpoints/duplicates_console_chk/metadata
3.8 K   7.6 K   checkpoints/duplicates_console_chk/offsets
28      56      checkpoints/duplicates_console_chk/sources
[student898_2@bigdataanalytics-worker-3 ~]$
```

Задаём вотермарку, которая должна очищать чекпоинт. Первый параметр - назване колонки, на которую смотрит вотермарка, второй параметр - гарантированное время жизни информации о сообщении в чекпойнте. Именно для этого мы добавляли столбец `receive_time`.

```
waterwarked_iris = extended_iris.withWatermark("receive_time", "30 seconds")

waterwarked_iris.printSchema()

extended_iris.printSchema()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> waterwarked_iris = extended_iris.withWatermark("receive_time", "30 seconds")
>>> waterwarked_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
 |-- receive_time: timestamp (nullable = false)

>>> extended_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
 |-- receive_time: timestamp (nullable = false)

>>>
```

Схема не поменялась. Вотермарка только следит за чекпойнтом, но никак не аффектит наши данные.

Теперь данные можно проверить на наличие дубликатов. Дубли проверяем по двум колонкам: `species` и `receive_time`. Таким образом будут отсеиваться дубли по полю `species` внутри одного микробатча, так как столбец `receive_time` для всех записей внутри этого микробатча одинаковый.

Для этого пишем новый датасет deduplicated_iris

```
deduplicated_iris = waterwarked_iris.drop_duplicates(["species",
"receive_time"])
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> deduplicated_iris = waterwarked_iris.drop_duplicates(["species", "receive_time"])
>>>
```

В другом окне удаляем папку чекпоитс

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

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -f -r checkpoints
22/01/23 21:08:48 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 ~]$
```

```
stream = console_output(deduplicated_iris , 20)

stream.stop()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> stream = console_output(deduplicated_iris , 20)
-------------------------------------------
Batch: 0
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+----------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time          |
+------------+-----------+------------+-----------+-------+------+----------------------+
|null        |null       |null        |null       |null   |0     |2022-01-23 21:09:40.648|
|5.1         |3.5        |1.4         |0.2        |setosa |1     |2022-01-23 21:09:40.648|
+------------+-----------+------------+-----------+-------+------+----------------------+

-------------------------------------------
Batch: 1
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+----------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time          |
+------------+-----------+------------+-----------+-------+------+----------------------+
|5.4         |3.9        |1.7         |0.4        |setosa |6     |2022-01-23 21:09:43.953|
+------------+-----------+------------+-----------+-------+------+----------------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+----------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time          |
+------------+-----------+------------+-----------+-------+------+----------------------+
|4.8         |3.4        |1.6         |0.2        |setosa |12    |2022-01-23 21:10:00.004|
+------------+-----------+------------+-----------+-------+------+----------------------+

-------------------------------------------                                  stream.stop()
Batch: 3
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+----------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time          |
+------------+-----------+------------+-----------+-------+------+----------------------+
|5.1         |3.5        |1.4         |0.3        |setosa |18    |2022-01-23 21:10:20.004|
+------------+-----------+------------+-----------+-------+------+----------------------+

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

Создаём временное окно. В структуру датафрейма добавился новый столбец.

```
windowed_iris = extended_iris.withColumn("window_time",
F.window(F.col("receive_time"), "2 minutes"))

windowed_iris.printSchema()
```

Мы добавили колонку withColumn, сделали receive_time"), "2 minutes

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> windowed_iris = extended_iris.withColumn("window_time", F.window(F.col("receive_time"), "2 minutes"))
>>> windowed_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
 |-- receive_time: timestamp (nullable = false)
 |-- window_time: struct (nullable = false)
 |    |-- start: timestamp (nullable = true)
 |    |-- end: timestamp (nullable = true)

>>>
```

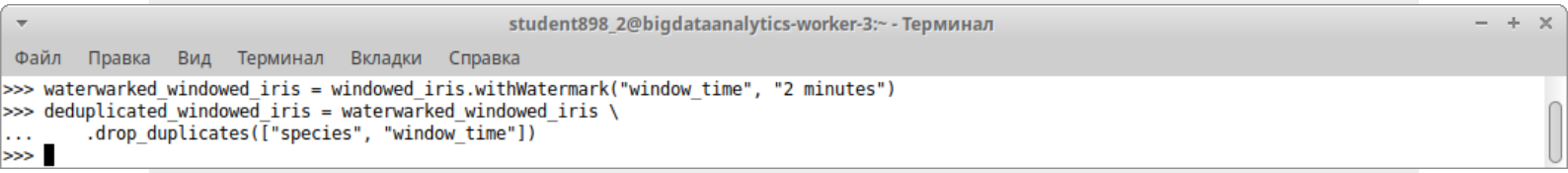
Ещё на это окно надо установить вотермарку

Устанавливаем вотермарку для очистки чекпоинта и удаляем дубли в каждом окне.

```
waterwarked_windowed_iris = windowed_iris.withWatermark("window_time", "2
minutes")

deduplicated_windowed_iris = waterwarked_windowed_iris \

    .drop_duplicates(["species", "window_time"])
```



Сначало надо удвлять чекпинты

```
hdfs dfs -rm -r checkpoints/duplicates_console_chk
```

Проверяем как удаляются дубли из каждого окна.

```
stream = console_output(deduplicated_windowed_iris , 20)

stream.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> stream = console_output(deduplicated_windowed_iris , 20)
-------------------------------------------
Batch: 0
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+----------------------+----------------------------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time          |window_time                             |
+------------+-----------+------------+-----------+-------+------+----------------------+----------------------------------------+
|5.1         |3.5        |1.4         |0.2        |setosa |1     |2022-01-23 21:17:38.861|[2022-01-23 21:16:00, 2022-01-23 21:18:00]|
|null        |null       |null        |null       |null   |0     |2022-01-23 21:17:38.861|[2022-01-23 21:16:00, 2022-01-23 21:18:00]|
+------------+-----------+------------+-----------+-------+------+----------------------+----------------------------------------+

-------------------------------------------
Batch: 1
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|window_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+----------------------+----------------------------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time          |window_time                             |
+------------+-----------+------------+-----------+-------+------+----------------------+----------------------------------------+
|4.8         |3.4        |1.6         |0.2        |setosa |12    |2022-01-23 21:18:00.004|[2022-01-23 21:18:00, 2022-01-23 21:20:00]|
+------------+-----------+------------+-----------+-------+------+----------------------+----------------------------------------+

-------------------------------------------
Batch: 3
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|window_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 4
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|window_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 5
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|window_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 6
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|window_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 7
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|window_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 8
-------------------------------------------
+------------+-----------+------------+-----------+----------+------+----------------------+----------------------------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species   |offset|receive_time          |window_time                             |
+------------+-----------+------------+-----------+----------+------+----------------------+----------------------------------------+
|4.6         |3.2        |1.4         |0.2        |setosa    |48    |2022-01-23 21:20:00.003|[2022-01-23 21:20:00, 2022-01-23 21:22:00]|
|7.0         |3.2        |4.7         |1.4        |versicolor|51    |2022-01-23 21:20:00.003|[2022-01-23 21:20:00, 2022-01-23 21:22:00]|
+------------+-----------+------------+-----------+----------+------+----------------------+----------------------------------------+

-------------------------------------------
Batch: 9
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|window_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

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

Аналогично предыдущему пункту создаём дополнительное поле `sliding_time`. В функции `F.window` первый аргумент это колонка (временная метка), по которой создаётся окно; второй аргумент - ширина окна; третий - сдвиг окна. Добавляем вотермарку и указываем колонки, по которым будем исключать дубли.

```
sliding_iris = extended_iris.withColumn("sliding_time",
F.window(F.col("receive_time"), "1 minute", "30 seconds"))

waterwarked_sliding_iris = sliding_iris.withWatermark("sliding_time", "2
minutes")

deduplicated_sliding_iris =
waterwarked_sliding_iris.drop_duplicates(["species", "sliding_time"])

deduplicated_sliding_iris.printSchema()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> sliding_iris = extended_iris.withColumn("sliding_time", F.window(F.col("receive_time"), "1 minute", "30 seconds"))
>>> waterwarked_sliding_iris = sliding_iris.withWatermark("sliding_time", "2 minutes")
>>> deduplicated_sliding_iris = waterwarked_sliding_iris.drop_duplicates(["species", "sliding_time"])
>>> deduplicated_sliding_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
 |-- receive_time: timestamp (nullable = false)
 |-- sliding_time: struct (nullable = true)
 |    |-- start: timestamp (nullable = true)
 |    |-- end: timestamp (nullable = true)

>>>
```

очищаем папку чекпоинтов. Запускаем стрим.

```
stream = console_output(deduplicated_sliding_iris , 5)

stream.stop()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
+------------+-----------+------------+-----------+---------+------+-----------------------+-----------------------------------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species  |offset|receive_time           |sliding_time                             |
+------------+-----------+------------+-----------+---------+------+-----------------------+-----------------------------------------+
|5.8        |2.7       |5.1        |1.9       |virginica|102   |2022-01-23 21:27:00.003|[2022-01-23 21:27:00, 2022-01-23 21:28:00]|
+------------+-----------+------------+-----------+---------+------+-----------------------+-----------------------------------------+

-------------------------------------------
Batch: 18
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|sliding_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 19
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|sliding_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 20
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|sliding_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 21
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|sliding_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+

-------------------------------------------
Batch: 22
-------------------------------------------
+------------+-----------+------------+-----------+-------+------+------------+------------+
|sepalLength|sepalWidth|petalLength|petalWidth|species|offset|receive_time|sliding_time|
+------------+-----------+------------+-----------+-------+------+------------+------------+
+------------+-----------+------------+-----------+-------+------+------------+------------+
```

```
Переопределяем метод `console_output` так, чтобы можно было задавать режим
вывода результата работы аггрегационных функций.

def console_output(df, freq, out_mode):

    return df.writeStream.format("console") \

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

        .options(truncate=False) \

        .option("checkpointLocation", "checkpoints/watermark_console_chk2") \

        .outputMode(out_mode) \

        .start()



waterwarked_windowed_iris.printSchema()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> def console_output(df, freq, out_mode):
...     return df.writeStream.format("console") \
...         .trigger(processingTime='%s seconds' % freq ) \
...         .options(truncate=False) \
...         .option("checkpointLocation", "checkpoints/watermark_console_chk2") \
...         .outputMode(out_mode) \
...         .start()
...
>>> waterwarked_windowed_iris.printSchema()
root
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- species: string (nullable = true)
 |-- offset: long (nullable = true)
 |-- receive_time: timestamp (nullable = false)
 |-- window_time: struct (nullable = false)
 |    |-- start: timestamp (nullable = true)
 |    |-- end: timestamp (nullable = true)

>>>
```

Сделаем новый датафрейм/стрим

```
count_iris = waterwarked_windowed_iris.groupBy("window_time").count()
```

очистим папку чекпоинтов

```
stream = console_output(count_iris , 10, "update")

stream.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|6    |
+-----------------------------------------+-----+

-------------------------------------------
Batch: 1
-------------------------------------------
+-----------------------------------------+-----+
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|12   |
+-----------------------------------------+-----+

-------------------------------------------
Batch: 2
-------------------------------------------
+-----------------------------------------+-----+
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|18   |
+-----------------------------------------+-----+

-------------------------------------------
Batch: 3
-------------------------------------------
+-----------------------------------------+-----+
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|24   |
+-----------------------------------------+-----+

-------------------------------------------
Batch: 4
-------------------------------------------
+-----------------------------------------+-----+
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:34:00, 2022-01-23 21:36:00]|6    |
+-----------------------------------------+-----+

stream.stop()
22/01/23 21:34:10 WARN hdfs.DFSClient: Caught exception
```

complete

```
stream = console_output(count_iris , 10, "complete")

stream.stop()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|24   |
|[2022-01-23 21:34:00, 2022-01-23 21:36:00]|12   |
|[2022-01-23 21:36:00, 2022-01-23 21:38:00]|24   |
+-----------------------------------------+-----+

-------------------------------------------
Batch: 10
-------------------------------------------
+-----------------------------------------+-----+
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|24   |
|[2022-01-23 21:34:00, 2022-01-23 21:36:00]|12   |
|[2022-01-23 21:36:00, 2022-01-23 21:38:00]|30   |
+-----------------------------------------+-----+

-------------------------------------------
Batch: 11
-------------------------------------------
+-----------------------------------------+-----+
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|24   |
|[2022-01-23 21:34:00, 2022-01-23 21:36:00]|12   |
|[2022-01-23 21:36:00, 2022-01-23 21:38:00]|36   |
+-----------------------------------------+-----+

-------------------------------------------
Batch: 12
-------------------------------------------
+-----------------------------------------+-----+
|window_time                              |count|
+-----------------------------------------+-----+
|[2022-01-23 21:32:00, 2022-01-23 21:34:00]|24   |
|[2022-01-23 21:34:00, 2022-01-23 21:36:00]|12   |
|[2022-01-23 21:36:00, 2022-01-23 21:38:00]|42   |
+-----------------------------------------+-----+

[Stage 121:=====================>                          (79 + 4) / 200]
22/01/23 21:37:11 ERROR v2.WriteToDataSourceV2Exec: Data source writer org.apache.spark.sql.execution.streaming.sources.MicroBatchWriter@2af2028b is a
```

append

Пишем все записи только один раз. Информация выводится один раз, когда окно заканчивается.

```
stream = console_output(count_iris , 10, "append")

stream.stop()
```

выходят пустые значения, агрегирующие функции не поддерживаются

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
-------------------------------------------
Batch: 13
-------------------------------------------
+-----------+-----+
|window_time|count|
+-----------+-----+
+-----------+-----+

-------------------------------------------
Batch: 14
-------------------------------------------
+-----------+-----+
|window_time|count|
+-----------+-----+
+-----------+-----+

-------------------------------------------
Batch: 15
-------------------------------------------
+-----------+-----+
|window_time|count|
+-----------+-----+
+-----------+-----+

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

Сджойнить стрим со статикой.

Создадим статический датафрейм, который будет расширять исходный датасет ирисов (объединение потоков)

```
static_df_schema = StructType() \

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

    .add("description", StringType())



static_df_data = (

    ("setosa", "Iris setosa has a deep violet blue flower. The sepals are deeply-veined dark purple with a yellow-white signal."),

    ("versicolor", "Iris versicolor is a flowering herbaceous perennial plant, growing 10-80 cm high. The well developed blue flower has 6 petals and sepals spread out nearly flat and have two forms."),

    ("virginica", "Iris virginica is a perennial plant. The plant has 2 to 4 erect or arching, bright green, lance-shaped leaves that are flattened into one plane at the base.")

)



static_df = spark.createDataFrame(static_df_data, static_df_schema)
```

```
static_joined = waterwarked_iris.join(static_df, "species", "left")

static_joined.isStreaming
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> static_df_schema = StructType() \
...     .add("species", StringType()) \
...     .add("description", StringType())
>>> static_df_data = (
...     ("setosa", "Iris setosa has a deep violet blue flower. The sepals are deeply-veined dark purple with a yellow-white signal."),
...     ("versicolor", "Iris versicolor is a flowering herbaceous perennial plant, growing 10-80 cm high. The well developed blue flower has 6 petals
and sepals spread out nearly flat and have two forms."),
...     ("virginica", "Iris virginica is a perennial plant. The plant has 2 to 4 erect or arching, bright green, lance-shaped leaves that are flattene
d into one plane at the base.")
... )
>>> static_df = spark.createDataFrame(static_df_data, static_df_schema)
>>> static_joined = waterwarked_iris.join(static_df, "species", "left")
>>> static_joined.isStreaming
True
>>>
```

После джойна стрима со статикой получаем стрим.

```
static_joined.printSchema()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> static_joined.printSchema()
root
 |-- species: string (nullable = true)
 |-- sepalLength: float (nullable = true)
 |-- sepalWidth: float (nullable = true)
 |-- petalLength: float (nullable = true)
 |-- petalWidth: float (nullable = true)
 |-- offset: long (nullable = true)
 |-- receive_time: timestamp (nullable = false)
 |-- description: string (nullable = true)

>>>
```

Добавлась колонка `description`.

```
stream = console_output(static_joined , 10, "update")

stream.stop()
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
>>> stream = console_output(static_joined , 10, "update")
-------------------------------------------
Batch: 0
-------------------------------------------
+-------+-----------+----------+-----------+----------+------+---------------------+-----------------------------------------------------------------
-----------------------------------------------+
|species|sepalLength|sepalWidth|petalLength|petalWidth|offset|receive_time         |description
                                                |
+-------+-----------+----------+-----------+----------+------+---------------------+-----------------------------------------------------------------
-----------------------------------------------+
|null   |null       |null      |null       |null      |0     |2022-01-23 22:02:34.65|null
                                                |
|setosa |5.1        |3.5       |1.4        |0.2       |1     |2022-01-23 22:02:34.65|Iris setosa has a deep violet blue flower. The sepals are deeply-
veined dark purple with a yellow-white signal.|
|setosa |4.9        |3.0       |1.4        |0.2       |2     |2022-01-23 22:02:34.65|Iris setosa has a deep violet blue flower. The sepals are deeply-
veined dark purple with a yellow-white signal.|
|setosa |4.7        |3.2       |1.3        |0.2       |3     |2022-01-23 22:02:34.65|Iris setosa has a deep violet blue flower. The sepals are deeply-
veined dark purple with a yellow-white signal.|
|setosa |4.6        |3.1       |1.5        |0.2       |4     |2022-01-23 22:02:34.65|Iris setosa has a deep violet blue flower. The sepals are deeply-
veined dark purple with a yellow-white signal.|
|setosa |5.0        |3.6       |1.4        |0.2       |5     |2022-01-23 22:02:34.65|Iris setosa has a deep violet blue flower. The sepals are deeply-
veined dark purple with a yellow-white signal.|
+-------+-----------+----------+-----------+----------+------+---------------------+-----------------------------------------------------------------
-----------------------------------------------+

-------------------------------------------
Batch: 1
-------------------------------------------
+-------+-----------+----------+-----------+----------+------+---------------------+-----------------------------------------------------------------
-----------------------------------------------+
|species|sepalLength|sepalWidth|petalLength|petalWidth|offset|receive_time         |description
                                                |
+-------+-----------+----------+-----------+----------+------+---------------------+-----------------------------------------------------------------
-----------------------------------------------+
|setosa |5.4        |3.9       |1.7        |0.4       |6     |2022-01-23 22:02:40.003|Iris setosa has a deep violet blue flower. The sepals are deeply
-veined dark purple with a yellow-white signal.|
|setosa |4.6        |3.4       |1.4        |0.3       |7     |2022-01-23 22:02:40.003|Iris setosa has a deep violet blue flower. The sepals are deeply
-veined dark purple with a yellow-white signal.|
|setosa |5.0        |3.4       |1.5        |0.2       |8     |2022-01-23 22:02:40.003|Iris setosa has a deep violet blue flower. The sepals are deeply
-veined dark purple with a yellow-white signal.|
|setosa |4.4        |2.9       |1.4        |0.2       |9     |2022-01-23 22:02:40.003|Iris setosa has a deep violet blue flower. The sepals are deeply
-veined dark purple with a yellow-white signal.|
|setosa |4.9        |3.1       |1.5        |0.1       |10    |2022-01-23 22:02:40.003|Iris setosa has a deep violet blue flower. The sepals are deeply
-veined dark purple with a yellow-white signal.|
|setosa |5.4        |3.7       |1.5        |0.2       |11    |2022-01-23 22:02:40.003|Iris setosa has a deep violet blue flower. The sepals are deeply
-veined dark purple with a yellow-white signal.|
+-------+-----------+----------+-----------+----------+------+---------------------+-----------------------------------------------------------------
-----------------------------------------------+

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

Сджойнить стрим со стримом.

Это задание сделаем на примере заранее созданных датасетов товаров и заказов.

Датасет, соотносящий товары и заказы читаем из кафки, топик `order_items`.

```
raw_orders_items = spark.readStream. \
    format("kafka"). \
    option("kafka.bootstrap.servers", kafka_brokers). \
    option("subscribe", "order_items"). \
    option("startingOffsets", "earliest"). \
```

```
    load()
```

Разбираем value и добавляем окно.

```
schema_orders_items = StructType() \

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

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

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

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

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

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

    .add("freight_value", StringType())



extended_orders_items = raw_orders_items \

    .select(F.from_json(F.col("value").cast("String"),
schema_orders_items).alias("value")) \

    .select("value.*") \

    .withColumn("order_items_receive_time", F.current_timestamp()) \

    .withColumn("window_time",F.window(F.col("order_items_receive_time"),"2
minutes"))



extended_orders_items.printSchema()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> raw_orders_items = spark.readStream. \
...     format("kafka"). \
...     option("kafka.bootstrap.servers", kafka_brokers). \
...     option("subscribe", "order_items"). \
...     option("startingOffsets", "earliest"). \
...     load()
>>> schema_orders_items = StructType() \
...     .add("order_id", StringType()) \
...     .add("order_item_id", StringType()) \
...     .add("product_id", StringType()) \
...     .add("seller_id", StringType()) \
...     .add("shipping_limit_date", StringType()) \
...     .add("price", StringType()) \
...     .add("freight_value", StringType())
>>> extended_orders_items = raw_orders_items \
...     .select(F.from_json(F.col("value").cast("String"), schema_orders_items).alias("value")) \
...     .select("value.*") \
...     .withColumn("order_items_receive_time", F.current_timestamp()) \
...     .withColumn("window_time",F.window(F.col("order_items_receive_time"),"2 minutes"))
>>> extended_orders_items.printSchema()
root
 |-- order_id: string (nullable = true)
 |-- order_item_id: string (nullable = true)
 |-- product_id: string (nullable = true)
 |-- seller_id: string (nullable = true)
 |-- shipping_limit_date: string (nullable = true)
 |-- price: string (nullable = true)
 |-- freight_value: string (nullable = true)
 |-- order_items_receive_time: timestamp (nullable = false)
 |-- window_time: struct (nullable = false)
 |    |-- start: timestamp (nullable = true)
 |    |-- end: timestamp (nullable = true)

>>>
```

Второй датасет списка заказов читаем из кафки, топик `orders_json`.

```
raw_orders = spark.readStream. \

    format("kafka"). \

    option("kafka.bootstrap.servers", kafka_brokers). \

    option("subscribe", "orders_json"). \

    option("maxOffsetsPerTrigger", "5"). \

    option("startingOffsets", "earliest"). \

    load()
```

Разбираем value, добавляем колонку со временем получения сообщения,создаём по ней окно и добавляем вотермарку.

```
schema = StructType() \

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

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

    .add("order_status", StringType()) \
```

```
    .add("order_purchase_timestamp", StringType()) \

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

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

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

    .add("order_estimated_delivery_date", StringType())



waterwarked_windowed_orders = raw_orders \

    .select(F.from_json(F.col("value").cast("String"), schema).alias("value"),
"offset") \

    .select("value.order_id", "value.order_status",
"value.order_purchase_timestamp") \

    .withColumn("order_receive_time", F.current_timestamp()) \

    .withColumn("window_time",F.window(F.col("order_receive_time"),"2
minutes")) \

    .withWatermark("window_time", "2 minutes")

waterwarked_windowed_orders.printSchema()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> raw_orders = spark.readStream. \
...     format("kafka"). \
...     option("kafka.bootstrap.servers", kafka_brokers). \
...     option("subscribe", "orders_json"). \
...     option("maxOffsetsPerTrigger", "5"). \
...     option("startingOffsets", "earliest"). \
...     load()
>>> schema = StructType() \
...     .add("order_id", StringType()) \
...     .add("customer_id", StringType()) \
...     .add("order_status", StringType()) \
...     .add("order_purchase_timestamp", StringType()) \
...     .add("order_approved_at", StringType()) \
...     .add("order_delivered_carrier_date", StringType()) \
...     .add("order_delivered_customer_date", StringType()) \
...     .add("order_estimated_delivery_date", StringType())
>>> waterwarked_windowed_orders = raw_orders \
...     .select(F.from_json(F.col("value").cast("String"), schema).alias("value"), "offset") \
...     .select("value.order_id", "value.order_status", "value.order_purchase_timestamp") \
...     .withColumn("order_receive_time", F.current_timestamp()) \
...     .withColumn("window_time",F.window(F.col("order_receive_time"),"2 minutes")) \
...     .withWatermark("window_time", "2 minutes")
>>> waterwarked_windowed_orders.printSchema()
root
 |-- order_id: string (nullable = true)
 |-- order_status: string (nullable = true)
 |-- order_purchase_timestamp: string (nullable = true)
 |-- order_receive_time: timestamp (nullable = false)
 |-- window_time: struct (nullable = false)
 |    |-- start: timestamp (nullable = true)
 |    |-- end: timestamp (nullable = true)

>>>
```

Делаем джойн двух датасетов.

```
streams_joined = waterwarked_windowed_orders \

    .join(extended_orders_items, ["order_id", "window_time"] , "inner") \

    .select("order_id", "order_item_id", "product_id", "window_time")
```

Тип отображения `update`не подходит для `inner` джойна.

очистим папку чекпоинтов

```
stream = console_output(streams_joined , 10, "append")

stream.stop()
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
>>> stream = console_output(streams_joined , 10,  'append')
-------------------------------------------
Batch: 0
-------------------------------------------
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|order_id                        |order_item_id|product_id                      |window_time                             |
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|53cdb2fc8bc7dce0b6741e2150273451|1            |595fac2a385ac33a80bd5114aec74eb8|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|order_id                        |order_item_id|product_id                      |[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|949d5b44dbf5de918fe9c16f97b45f8a|1            |d0b61bfb1de832b15ba9d266ca96e5b0|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|47770eb9100c2d0c44946d9cf07ec65d|1            |aa4383b373c6aca5d8797843e5594415|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|e481f51cbdc54678b7cc49136f2d6af7|1            |87285b34884572647811a353c7ac498a|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
+--------------------------------+-------------+--------------------------------+----------------------------------------+

22/01/23 22:11:31 WARN streaming.ProcessingTimeExecutor: Current batch is falling behind. The trigger interval is 10000 milliseconds, but spent 14327
milliseconds
-------------------------------------------
Batch: 1
-------------------------------------------
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|order_id                        |order_item_id|product_id                      |window_time                             |
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|76c6e866289321a7c93b82b54852dc33|1            |ac1789e492dcd698c5c10b97a671243a|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|6514b8ad8028c9f2cc2374ded245783f|1            |4520766ec412348b8d4caa5e8a18c464|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|a4591c265e18cb1dcee52889e2d8acc3|1            |060cb19345d90064d1015407193c233d|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|136cce7faa42fdb2cefd53fdc79a6098|1            |a1804276d9941ac0733cfd409f5206eb|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|ad21c59c0840e6cb83a9ceb5573f8159|1            |65266b2da20d04dbe00c5c2d3bb7859e|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
+--------------------------------+-------------+--------------------------------+----------------------------------------+

-------------------------------------------
Batch: 2
-------------------------------------------
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|order_id                        |order_item_id|product_id                      |window_time                             |
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|e6ce16cb79ec1d90b1da9085a6118aeb|1            |08574b074924071f4e201e151b152b4e|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|e6ce16cb79ec1d90b1da9085a6118aeb|2            |08574b074924071f4e201e151b152b4e|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|82566a660a982b15fb86e904c8d32918|1            |72a97c271b2e429974398f46b93ae530|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|34513ce0c4fab462a55830c0989c7edb|1            |f7e0fa615b386bc9a8b9eb52bc1fff76|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|e69bfb5eb88e0ed6a785585b27e16dbf|1            |9a78fb9862b10749a117f7fc3c31f051|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|5ff96c15d0b717ac6ad1f3d77225a350|1            |10adb53d8faa890ca7c2f0cbcb68d777|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
+--------------------------------+-------------+--------------------------------+----------------------------------------+

-------------------------------------------
Batch: 3
-------------------------------------------
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|order_id                        |order_item_id|product_id                      |window_time                             |
+--------------------------------+-------------+--------------------------------+----------------------------------------+
|116f0b09343b49556bbad5f35bee0cdf|1            |a47295965bd091207681b541b26e40a5|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|85ce859fd6dc634de8d2f1e290444043|1            |cce679660c66e6fbd5c8091dfd29e9cd|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|dcb36b511fcac050b97cd5c05de84dc3|1            |009c09f439988bc06a93d6b8186dce73|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|403b97836b0c04a622354cf531062e5f|1            |638bbb2a5e4f360b71f332ddfebfd672|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
|432aaf21d85167c2c86ec9448c4e42cc|1            |72d3bf1d3a790f8874096fcf860e3eff|[2022-01-23 22:10:00, 2022-01-23 22:12:00]|
+--------------------------------+-------------+--------------------------------+----------------------------------------+

-------------------------------------------
Batch: 4
-------------------------------------------
+--------+-------------+----------+-----------+
|order_id|order_item_id|product_id|window_time|
+--------+-------------+----------+-----------+
+--------+-------------+----------+-----------+

-------------------------------------------
Batch: 5
-------------------------------------------
+--------+-------------+----------+-----------+
|order_id|order_item_id|product_id|window_time|
+--------+-------------+----------+-----------+
+--------+-------------+----------+-----------+

[Stage 154:==========>                                          (39 + 4) / 200]stream.stop()
```

Здесь не увидел результата, так как в топике `order_items` не было данных. По факту этот топик вычитывается целиком за раз, поэтому в первом окне можно наблюдать микробатчи сджойненого датасета. Для остальных окон микробатчи пустые, так как `window_time` уже различаются. Из топика `order_items` новые данные не приходят.

```
import csv
```

```
import json



with open('test.csv') as f:

    reader = csv.DictReader(f)

    rows = list(reader)



with open('test.json', 'w') as f:

    json.dump(rows, f)
```