7. Spark ML. Аналитика признаков в пакетном режиме. Подготовка, обучение ML-модели

Подключиться к кластеру, выполнить команду spark-submit с приложенными к занятию скриптами, приложить листинг консоли (запуск, результат).
Обучить модель на основании данных из хранилища Hive sint_sales, проверить сходимость и показатель ROC.
Дополнительно, спроектировать приложение по потоковой обработки данных на основании схемы предложенной на вебинаре - итоговая работа по курсу

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

ls for_stream/

ls
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
igor@igor-MS-7808:~$ ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176
Last login: Sat Jan 29 21:12:58 2022 from 109.252.19.10
[student898_2@bigdataanalytics-worker-3 ~]$ ls for_stream/
archive.csv  data.json    drake_data.json  iris.json          product_list2.csv  product_list4.csv
data.csv     dataset.csv  file1.json       product_list1.csv  product_list3.csv  product_list.csv
[student898_2@bigdataanalytics-worker-3 ~]$ ls
7.1_spark-submit_stream.py  7.2_spark-submit_stable.py  7.spark-submit-batch.py  for_stream
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
vi 7.spark-submit-batch.py
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
spark.sparkContext.setLogLevel("WARN")
schema = StructType() \
    .add("product_name", StringType()) \
    .add("product_category", StringType())

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

load_time = datetime.datetime.now().strftime("%Y%m%d%H%M%S")
print("START BATCH LOADING. TIME = " + load_time)

raw_files.withColumn("p_date", F.lit("load_time")) \
    .write \
    .mode("append") \
    .parquet("my_submit_parquet_files/p_date=" + str(load_time))

print("FINISHED BATCH LOADING. TIME = " + load_time)

spark.stop()
~
```

```
ls for_stream
```

```
cat for_stream/product_list.csv
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
[student898_2@bigdataanalytics-worker-3 ~]$ ls for_stream/
archive.csv  data.json    drake_data.json  iris.json          product_list2.csv  product_list4.csv
data.csv     dataset.csv  file1.json       product_list1.csv  product_list3.csv  product_list.csv
[student898_2@bigdataanalytics-worker-3 ~]$ cat 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'
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
изменим

vi 7.spark-submit-batch.py
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
spark.sparkContext.setLogLevel("WARN")
schema = StructType() \
    .add("product_name", StringType()) \
    .add("product_category", StringType())

raw_files = spark \
    .read \
    .format("csv") \
    .schema(schema) \
    .options(path="for_stream", header=True) \
    .load()

load_time = datetime.datetime.now().strftime("%Y%m%d%H%M%S")
print("START BATCH LOADING. TIME = " + load_time)

raw_files.withColumn("p_date", F.lit("load_time")) \
    .write \
    .mode("append") \
    .parquet("my_submit_parquet_files/p_date=" + str(load_time))

print("FINISHED BATCH LOADING. TIME = " + load_time)

spark.stop()
:wq
```

```
hdfs dfs -ls

hdfs dfs -ls for_stream
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 10 items
drwx------   - student898_2 student898_2          0 2022-01-30 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:31 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:15 checkpionts
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:38 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:49 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 ~]$ hdfs dfs -ls for_stream
Found 5 items
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:38 for_stream/product_list.csv
-rw-r--r--   2 student898_2 student898_2         98 2022-01-29 21:38 for_stream/product_list1.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:38 for_stream/product_list2.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:38 for_stream/product_list3.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:38 for_stream/product_list4.csv
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
spark-submit 7.spark-submit-batch.py
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
22/01/30 10:27:49 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.8
22/01/30 10:27:49 INFO BlockManagerMasterEndpoint: Registering block manager bigdataanalytics-worker-1.mcs.local:37297 with 366.3 MB RAM, BlockManager
Id(1, bigdataanalytics-worker-1.mcs.local, 37297, None)
22/01/30 10:27:49 INFO SharedState: loading hive config file: file:/etc/spark2/3.1.4.0-315/0/hive-site.xml
22/01/30 10:27:49 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('/apps/spark/warehouse').
22/01/30 10:27:49 INFO SharedState: Warehouse path is '/apps/spark/warehouse'.
22/01/30 10:27:49 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL.
22/01/30 10:27:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@2afd0a95{/SQL,null,AVAILABLE,@Spark}
22/01/30 10:27:49 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/json.
22/01/30 10:27:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@4742b091{/SQL/json,null,AVAILABLE,@Spark}
22/01/30 10:27:49 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution.
22/01/30 10:27:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@406a8146{/SQL/execution,null,AVAILABLE,@Spark}
22/01/30 10:27:49 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution/json.
22/01/30 10:27:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@544adf27{/SQL/execution/json,null,AVAILABLE,@Spark}
22/01/30 10:27:49 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
22/01/30 10:27:49 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7a96f9f6{/static/sql,null,AVAILABLE,@Spark}
22/01/30 10:27:49 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
START BATCH LOADING. TIME = 20220130102750
FINISHED BATCH LOADING. TIME = 20220130102750
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
hdfs dfs -ls

hdfs dfs -ls my_submit_parquet_files
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - student898_2 student898_2          0 2022-01-30 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-30 10:27 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:15 checkpionts
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:38 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:49 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 2022-01-30 10:27 my_submit_parquet_files
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 ~]$ hdfs dfs -ls my_submit_parquet_files
Found 1 items
drwxr-xr-x   - student898_2 student898_2          0 2022-01-30 10:27 my_submit_parquet_files/p_date=20220130102750
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
vi 7.1_spark-submit_stream.py
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
spark.sparkContext.setLogLevel("WARN")
schema = StructType() \
    .add("product_name", StringType()) \
    .add("product_category", StringType())

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

load_time = datetime.datetime.now().strftime("%Y%m%d%H%M%S")

def file_sink(df, freq):
    return df.writeStream.format("parquet") \
        .trigger(processingTime='%s seconds' % freq ) \
        .option("path","my_submit_parquet_files/p_date=" + str(load_time)) \
        .option("checkpointLocation", "checkpionts/my_parquet_checkpoint") \
        .start()

timed_files = raw_files.withColumn("p_date", F.lit("load_time"))

stream = file_sink(timed_files,10)

#will always spark.stop() at the end
~
:wq
```

```
hdfs dfs -ls

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

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - student898_2 student898_2          0 2022-01-30 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-30 10:27 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:15 checkpionts
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:38 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:49 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 2022-01-30 10:27 my_submit_parquet_files
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 ~]$ hdfs dfs -rm -r -f checkpoints
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -r -f checkpoints
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
spark-submit 7.1_spark-submit_stream.py
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
22/01/30 10:47:13 ERROR MicroBatchExecution: Query [id = 706fb042-d8a6-442f-9eb0-03bddf292d97, runId = 113581cf-601e-4db5-9f0e-01e51a333e7e] terminate
d with error
java.lang.IllegalStateException: Cannot call methods on a stopped SparkContext.
This stopped SparkContext was created at:

org.apache.spark.api.java.JavaSparkContext.<init>(JavaSparkContext.scala:58)
sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
java.lang.reflect.Constructor.newInstance(Constructor.java:423)
py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:247)
py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
py4j.Gateway.invoke(Gateway.java:238)
py4j.commands.ConstructorCommand.invokeConstructor(ConstructorCommand.java:80)
py4j.commands.ConstructorCommand.execute(ConstructorCommand.java:69)
py4j.GatewayConnection.run(GatewayConnection.java:238)
java.lang.Thread.run(Thread.java:748)

The currently active SparkContext was created at:

org.apache.spark.api.java.JavaSparkContext.<init>(JavaSparkContext.scala:58)
sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)
sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:62)
sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)
java.lang.reflect.Constructor.newInstance(Constructor.java:423)
py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:247)
py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
py4j.Gateway.invoke(Gateway.java:238)
py4j.commands.ConstructorCommand.invokeConstructor(ConstructorCommand.java:80)
py4j.commands.ConstructorCommand.execute(ConstructorCommand.java:69)
py4j.GatewayConnection.run(GatewayConnection.java:238)
java.lang.Thread.run(Thread.java:748)

        at org.apache.spark.SparkContext.assertNotStopped(SparkContext.scala:99)
        at org.apache.spark.sql.SparkSession.<init>(SparkSession.scala:91)
        at org.apache.spark.sql.SparkSession.cloneSession(SparkSession.scala:256)
        at org.apache.spark.sql.execution.streaming.StreamExecution.org$apache$spark$sql$execution$streaming$StreamExecution$$runStream(StreamExecutio
n.scala:268)
        at org.apache.spark.sql.execution.streaming.StreamExecution$$anon$1.run(StreamExecution.scala:189)
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
hdfs dfs -ls

hdfs dfs -du -h checkpionts

hdfs dfs -du -h checkpionts/my_parquet_checkpoint
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - student898_2 student898_2          0 2022-01-30 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-30 10:47 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:15 checkpionts
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:38 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 21:49 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 2022-01-30 10:47 my_submit_parquet_files
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 ~]$ hdfs dfs -du -h checkpionts
1.2 K  2.5 K  checkpionts/my_parquet_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h checkpionts/my_parquet_checkpoint
29   58     checkpionts/my_parquet_checkpoint/commits
45   90     checkpionts/my_parquet_checkpoint/metadata
422  844    checkpionts/my_parquet_checkpoint/offsets
761  1.5 K  checkpionts/my_parquet_checkpoint/sources
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
vi 7.2_spark-submit_stable.py
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
spark.sparkContext.setLogLevel("WARN")
schema = StructType() \
    .add("product_id", StringType()) \
    .add("product_name", StringType()) \
    .add("product_category", StringType())

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

def file_sink(df, freq):
    return df.writeStream.foreachBatch(foreach_batch_function) \
        .trigger(processingTime='%s seconds' % freq ) \
        .option("checkpointLocation", "checkpionts/my_parquet_checkpoint") \
        .start()

def foreach_batch_function(df, epoch_id):
    load_time = datetime.datetime.now().strftime("%Y%m%d%H%M%S")
    print("START BATCH LOADING. TIME = " + load_time)
    df.withColumn("p_date", F.lit("load_time")) \
        .write \
        .mode("append") \
        .parquet("my_submit_parquet_files/p_date=" + str(load_time))
    print("FINISHED BATCH LOADING. TIME = " + load_time)

stream = file_sink(raw_files,10)

while(True):
    print("I'M STILL ALIVE")
    stream.awaitTermination(9)

#unreachable
spark.stop()
:wq
```

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

spark-submit 7.2_spark-submit_stable.py
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
22/01/30 10:56:52 INFO BlockManagerMasterEndpoint: Registering block manager bigdataanalytics-worker-1.mcs.local:40334 with 366.3 MB RAM, BlockManager
Id(1, bigdataanalytics-worker-1.mcs.local, 40334, None)
22/01/30 10:56:52 INFO SharedState: loading hive config file: file:/etc/spark2/3.1.4.0-315/0/hive-site.xml
22/01/30 10:56:52 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('/apps/spark/warehouse').
22/01/30 10:56:52 INFO SharedState: Warehouse path is '/apps/spark/warehouse'.
22/01/30 10:56:52 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL.
22/01/30 10:56:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@70ec71fc{/SQL,null,AVAILABLE,@Spark}
22/01/30 10:56:52 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/json.
22/01/30 10:56:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@6bfe05b5{/SQL/json,null,AVAILABLE,@Spark}
22/01/30 10:56:52 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution.
22/01/30 10:56:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7cdaa51c{/SQL/execution,null,AVAILABLE,@Spark}
22/01/30 10:56:52 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution/json.
22/01/30 10:56:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@3566f979{/SQL/execution/json,null,AVAILABLE,@Spark}
22/01/30 10:56:52 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
22/01/30 10:56:52 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@4e6263f0{/static/sql,null,AVAILABLE,@Spark}
22/01/30 10:56:53 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
Traceback (most recent call last):
  File "/home/student898_2/7.2_spark-submit_stable.py", line 35, in <module>
    stream = file_sink(raw_files,10)
  File "/home/student898_2/7.2_spark-submit_stable.py", line 21, in file_sink
    return df.writeStream.foreachBatch(foreach_batch_function) \
AttributeError: 'DataStreamWriter' object has no attribute 'foreachBatch'
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
/opt/spark-2.4.8/bin/spark-submit 7.2_spark-submit_stable.py
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/for_stream/product_list.csv
22/01/30 11:03:41 WARN csv.CSVDataSource: CSV header does not conform to the schema.
 Header: product_id,  product_name,  product_category
 Schema: product_id, product_name, product_category
Expected: product_name but found:  product_name
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/for_stream/product_list3.csv
FINISHED BATCH LOADING. TIME = 20220130110339
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
```

```
hdfs dfs -du -h checkpionts

hdfs dfs -du -h checkpionts/my_parquet_checkpoint
```

```
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл  Правка  Вид  Терминал  Вкладки  Справка
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h checkpionts
2.4 K  4.7 K  checkpionts/my_parquet_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h checkpionts/my_parquet_checkpoint
58     116    checkpionts/my_parquet_checkpoint/commits
45     90     checkpionts/my_parquet_checkpoint/metadata
844    1.6 K  checkpionts/my_parquet_checkpoint/offsets
1.4 K  2.9 K  checkpionts/my_parquet_checkpoint/sources
[student898_2@bigdataanalytics-worker-3 ~]$
```

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

```
hdfs dfs -rm -r -f my_submit_parquet_files
```

student898_2@bigdataanalytics-worker-3:~ - Терминал

Файл Правка Вид Терминал Вкладки Справка

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -r -f checkpoints
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -r -f my_submit_parquet_files
22/01/30 11:10:31 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/my_submit_parquet_files' to tras
h at: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/my_submit_parquet_files
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
hdfs dfs -ls

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

```
I'M STILL ALIVE
START BATCH LOADING. TIME = 20220130111421
22/01/30 11:14:22 WARN csv.CSVDataSource: CSV header does not conform to the schema.
 Header: product_id,  product_name,  product_category
 Schema: product_id, product_name, product_category
Expected: product_name but found:  product_name
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/for_stream/product_list1.csv
22/01/30 11:14:22 WARN csv.CSVDataSource: CSV header does not conform to the schema.
 Header: product_id,  product_name,  product_category
 Schema: product_id, product_name, product_category
Expected: product_name but found:  product_name
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/for_stream/product_list4.csv
22/01/30 11:14:22 WARN csv.CSVDataSource: CSV header does not conform to the schema.
 Header: product_id,  product_name,  product_category
 Schema: product_id, product_name, product_category
Expected: product_name but found:  product_name
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/for_stream/product_list.csv
22/01/30 11:14:22 WARN csv.CSVDataSource: CSV header does not conform to the schema.
 Header: product_id,  product_name,  product_category
 Schema: product_id, product_name, product_category
Expected: product_name but found:  product_name
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/for_stream/product_list2.csv
22/01/30 11:14:22 WARN csv.CSVDataSource: CSV header does not conform to the schema.
 Header: product_id,  product_name,  product_category
 Schema: product_id, product_name, product_category
Expected: product_name but found:  product_name
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/for_stream/product_list3.csv
FINISHED BATCH LOADING. TIME = 20220130111421
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
^CTraceback (most recent call last):
  File "/home/student898_2/7.2_spark-submit_stable.py", line 39, in <module>
    stream.awaitTermination(9)
  File "/opt/spark-2.4.8/python/lib/pyspark.zip/pyspark/sql/streaming.py", line 101, in awaitTermination
  File "/opt/spark-2.4.8/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1255, in __call__
  File "/opt/spark-2.4.8/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 985, in send_command
  File "/opt/spark-2.4.8/python/lib/py4j-0.10.7-src.zip/py4j/java_gateway.py", line 1152, in send_command
  File "/usr/lib64/python2.7/socket.py", line 447, in readline
    data = self._sock.recv(self._rbufsize)
  File "/opt/spark-2.4.8/python/lib/pyspark.zip/pyspark/context.py", line 270, in signal_handler
KeyboardInterrupt
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
/opt/spark-2.4.8/bin/spark-submit 7.2_spark-submit_stable.py
```

```
если в vi 7.2_spark-submit_stable.py
```

Изменить чтение потока из кафки

/home/igor/4.1 Потоковая обработка данных/7. Spark ML. Аналитика признаков в пакетном режиме. Подготовка, обучение ML-модели/7.2_spark-submit_stable.py - Mousepad

Файл Правка Поиск Вид Документ Справка

```python
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
schema = StructType() \
    .add("product_category_name", StringType()) \
    .add("product_category_name_english", StringType())

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

#пишем стрим в foreachBatch, чтобы делать логику в зависимости от каждого микробатча
```

/home/igor/4.1 Потоковая обработка данных/7. Spark ML. Аналитика признаков в пакетном режиме. Подготовка, обучение ML-модели/7.2_spark-submit_stable.py - Mousepad

Файл Правка Поиск Вид Документ Справка

```python
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
schema = StructType() \
    .add("product_category_name", StringType()) \
    .add("product_category_name_english", StringType())

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


```python
#пишем стрим в foreachBatch, чтобы делать логику в зависимости от каждого микробатча
def file_sink(df, freq):
    return df.writeStream.foreachBatch(foreach_batch_function) \
        .trigger(processingTime='%s seconds' % freq ) \
        .option("checkpointLocation", "checkpionts/my_parquet_checkpoint") \
        .start()

#в каждом микробатче фиксируем время,  логируем на экран, пишем файлы в свою директорию
def foreach_batch_function(df, epoch_id):
    load_time = datetime.datetime.now().strftime("%Y%m%d%H%M%S")
    print("START BATCH LOADING. TIME = " + load_time)
    df.withColumn("p_date", F.lit("load_time")) \
        .write \
        .mode("append") \
        .parquet("my_submit_parquet_files/p_date=" + str(load_time))
    print("FINISHED BATCH LOADING. TIME = " + load_time)

stream = file_sink(raw_files,10)

#запускаем бесконечный цикл
while(True):
    print("I'M STILL ALIVE")
    stream.awaitTermination(9)

#unreachable
spark.stop()
```

А сюда помимо записи форич-бач, сюда сделать запись в касандру

А перед касандрой сделать ML-lib

```
это будет наша работа

towardsdatascience 7. ml lib train.py
```