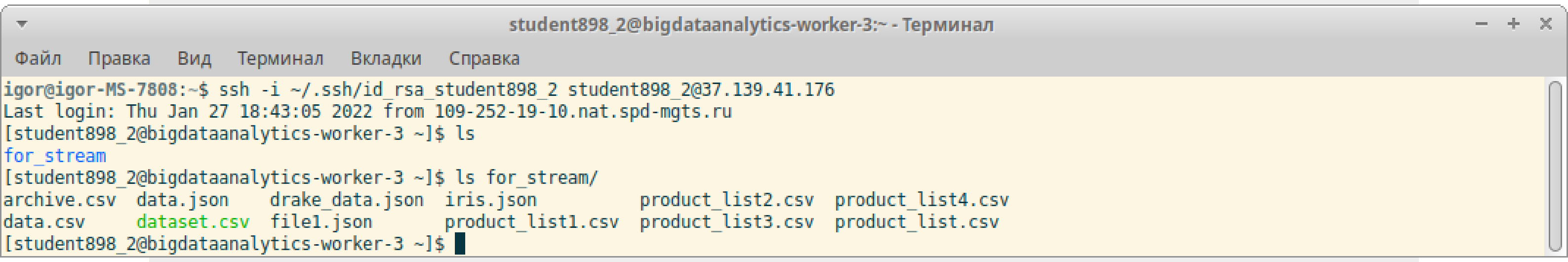
7. Spark ML. Аналитика признаков в пакетном режиме. Подготовка, обучение ML-модели

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

ls

ls for_stream/



```
igor@igor-MS-7808:~$ ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176
Last login: Thu Jan 27 18:43:05 2022 from 109-252-19-10.nat.spd-mgts.ru
[student898_2@bigdataanalytics-worker-3 ~]$ ls
for_stream
[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 ~]$ 
```

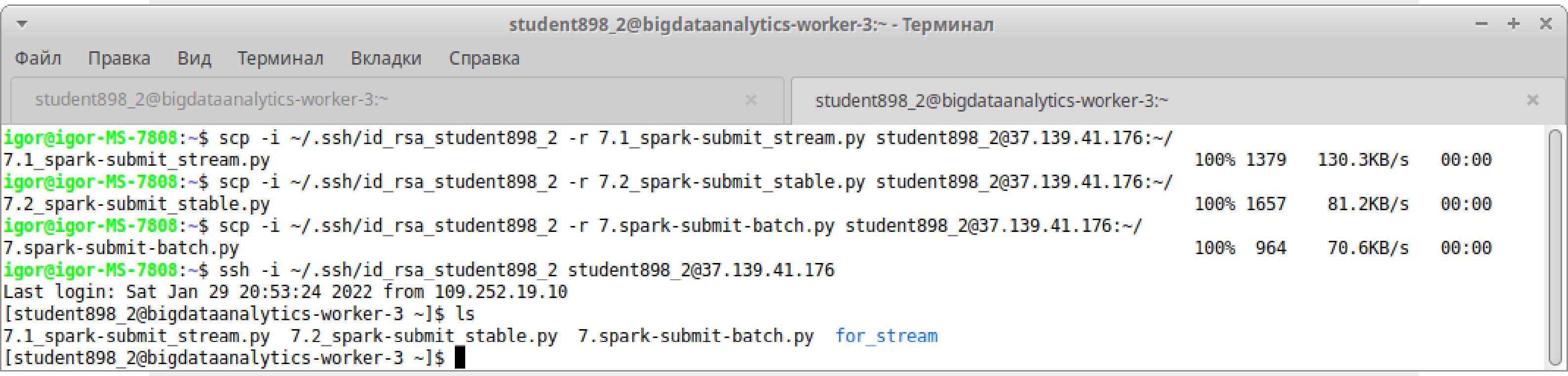
Скопируем подготовленные файлы на удаленный сервер с помощью команды `scp`. Эта команда запускается на локальном компьютере

scp -i ~/.ssh/id_rsa_student898_2 -r 7.1_spark-submit_stream.py student898_2@37.139.41.176:~/

scp -i ~/.ssh/id_rsa_student898_2 -r 7.2_spark-submit_stable.py student898_2@37.139.41.176:~/

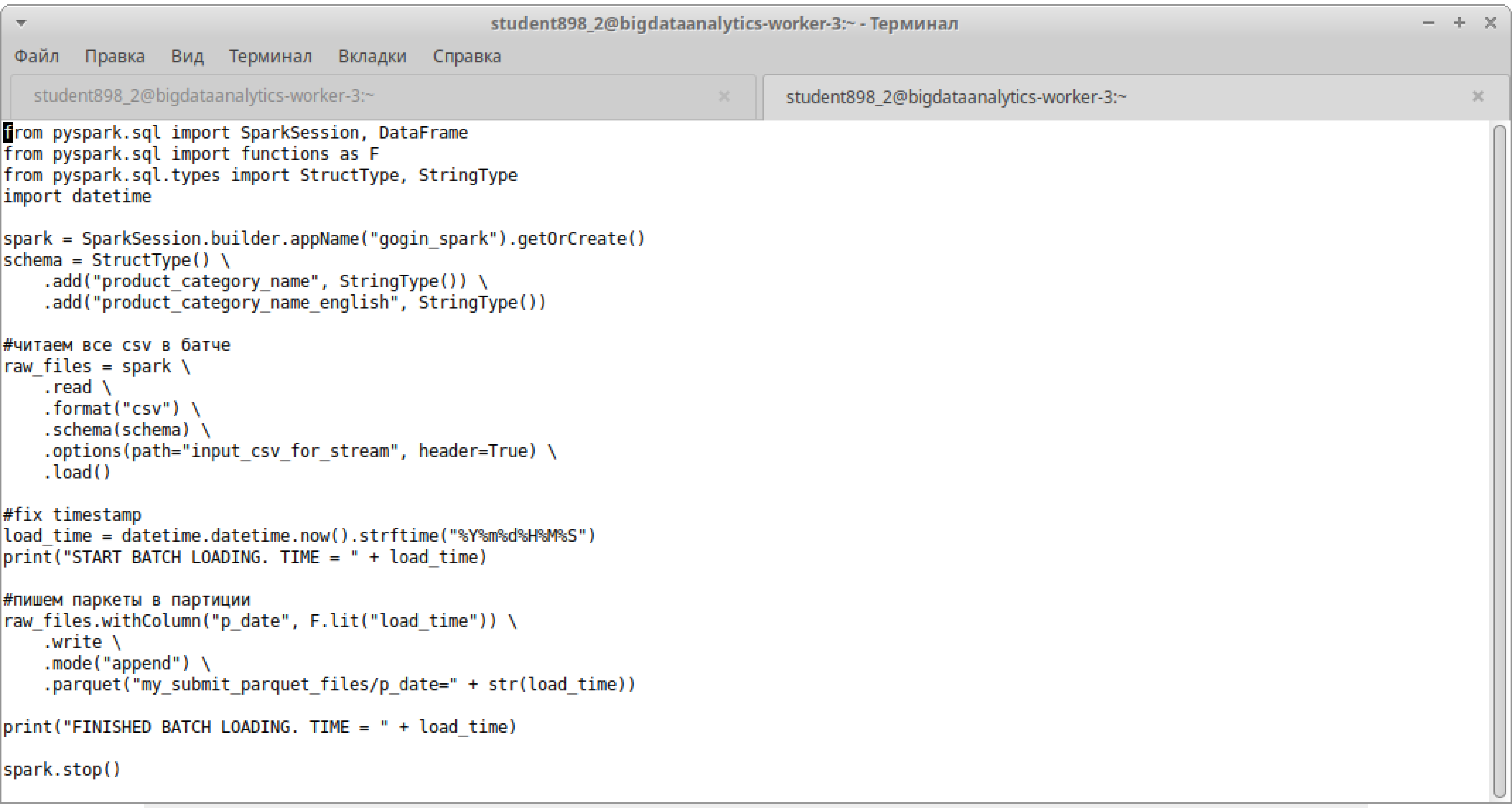
scp -i ~/.ssh/id_rsa_student898_2 -r 7.spark-submit-batch.py student898_2@37.139.41.176:~/

ls


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

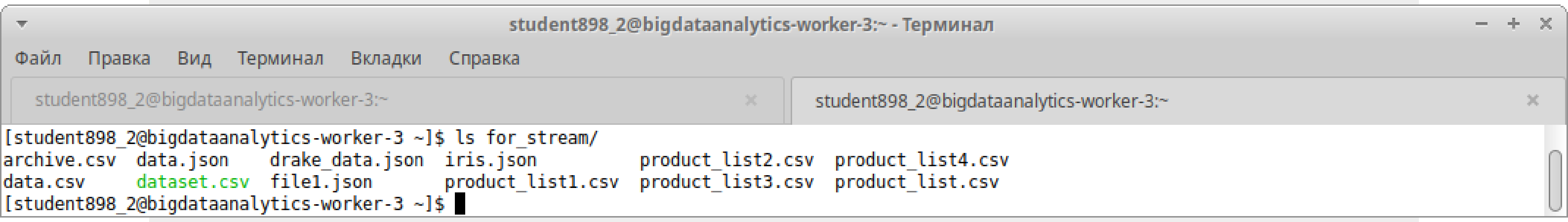
```
igor@igor-MS-7808:~$ scp -i ~/.ssh/id_rsa_student898_2 -r 7.1_spark-submit_stream.py student898_2@37.139.41.176:~/
7.1_spark-submit_stream.py                                                                        100% 1379   130.3KB/s   00:00
igor@igor-MS-7808:~$ scp -i ~/.ssh/id_rsa_student898_2 -r 7.2_spark-submit_stable.py student898_2@37.139.41.176:~/
7.2_spark-submit_stable.py                                                                        100% 1657    81.2KB/s   00:00
igor@igor-MS-7808:~$ scp -i ~/.ssh/id_rsa_student898_2 -r 7.spark-submit-batch.py student898_2@37.139.41.176:~/
7.spark-submit-batch.py                                                                           100%  964    70.6KB/s   00:00
igor@igor-MS-7808:~$ ssh -i ~/.ssh/id_rsa_student898_2 student898_2@37.139.41.176
Last login: Sat Jan 29 20:53:24 2022 from 109.252.19.10
[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
student898_2@bigdataanalytics-worker-3:~
student898_2@bigdataanalytics-worker-3:~

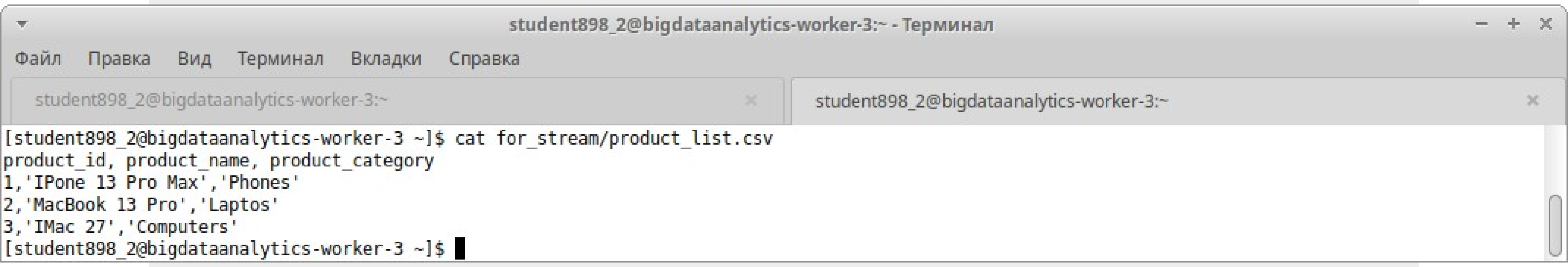
```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

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

#читаем все csv в батче
raw_files = spark \
    .read \
    .format("csv") \
    .schema(schema) \
    .options(path="input_csv_for_stream", header=True) \
    .load()

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

#пишем паркеты в партиции
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/
```


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

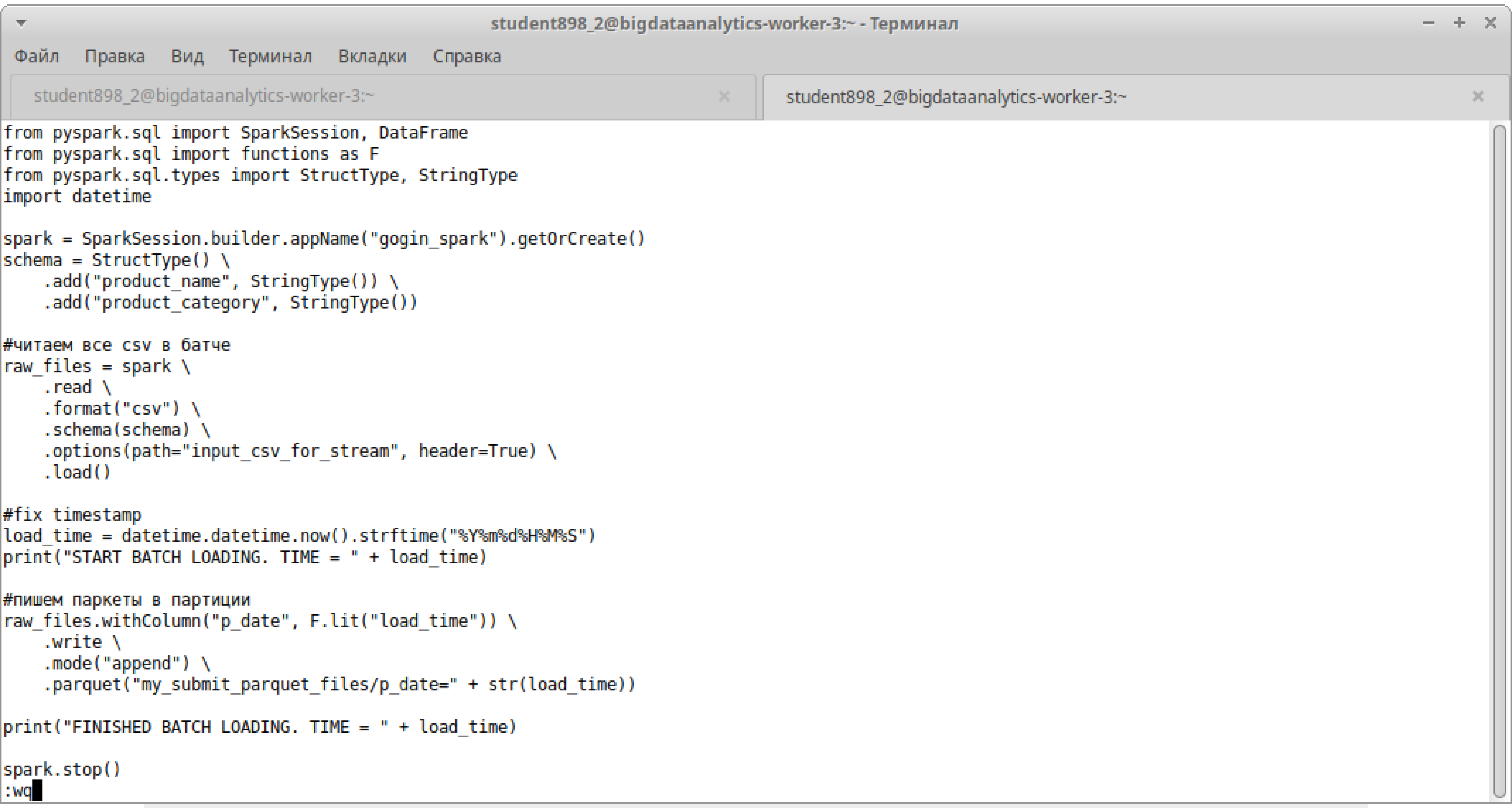
```
[student898_2@bigdataanalytics-worker-3 ~]$ ls for_stream/
archive.csv  data.json    drake_data.json  iris.json          product_list2.csv  product_list4.csv
data.csv     dataset.csv  file1.json       product_list1.csv  product_list3.csv  product_list.csv
[student898_2@bigdataanalytics-worker-3 ~]$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

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


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

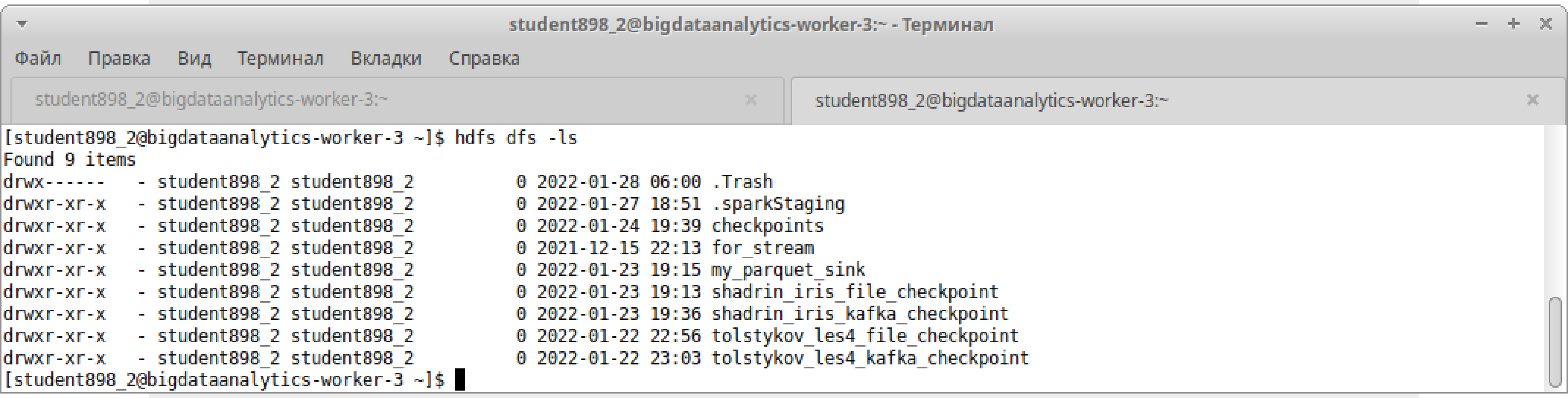
```
[student898_2@bigdataanalytics-worker-3 ~]$ cat for_stream/product_list.csv
product_id, product_name, product_category
1,'IPone 13 Pro Max','Phones'
2,'MacBook 13 Pro','Laptos'
3,'IMac 27','Computers'
[student898_2@bigdataanalytics-worker-3 ~]$
```

изменяем


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

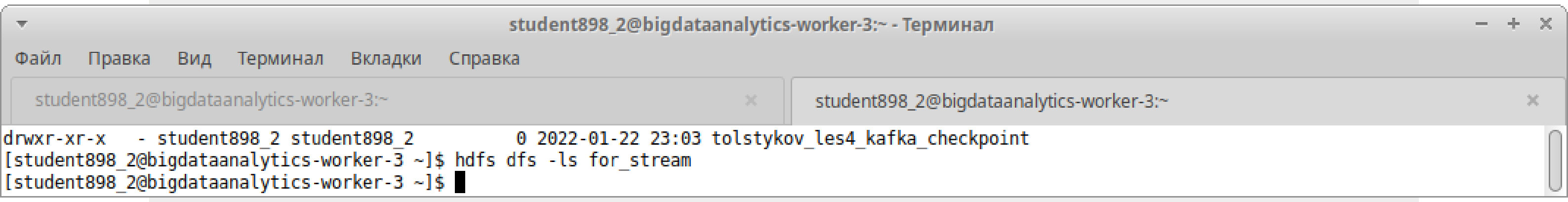
```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

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

#читаем все csv в батче
raw_files = spark \
    .read \
    .format("csv") \
    .schema(schema) \
    .options(path="input_csv_for_stream", header=True) \
    .load()

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

#пишем паркеты в партиции
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
```


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

```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 9 items
drwx------   - student898_2 student898_2          0 2022-01-28 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-27 18:51 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-24 19:39 checkpoints
drwxr-xr-x   - student898_2 student898_2          0 2021-12-15 22:13 for_stream
drwxr-xr-x   - student898_2 student898_2          0 2022-01-23 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 ~]$
```

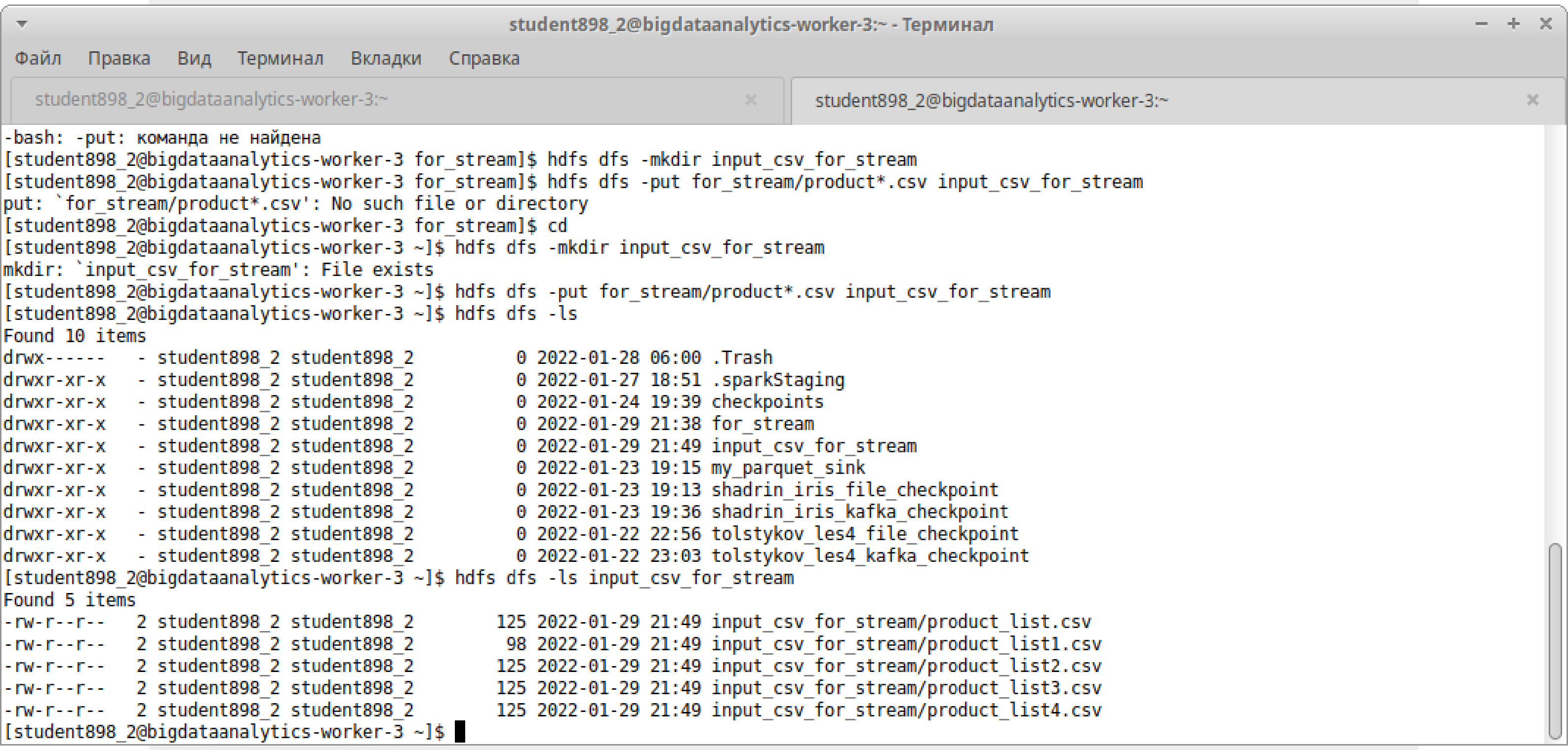
```
hdfs dfs -ls for_stream
```


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

```
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
[student898_2@bigdataanalytics-worker-3 ~]$
```

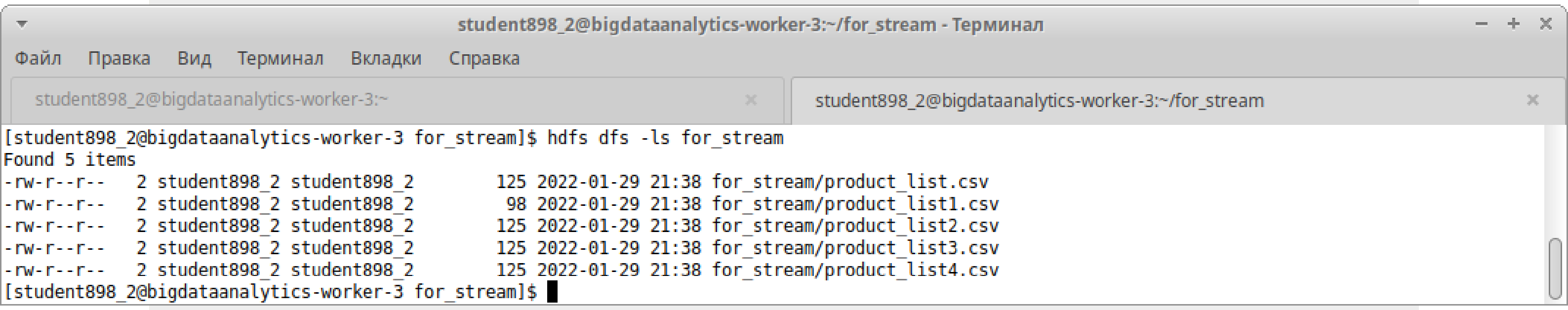
Перемещаем файлы

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

hdfs dfs -ls

hdfs dfs -ls input_csv_for_stream
```


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

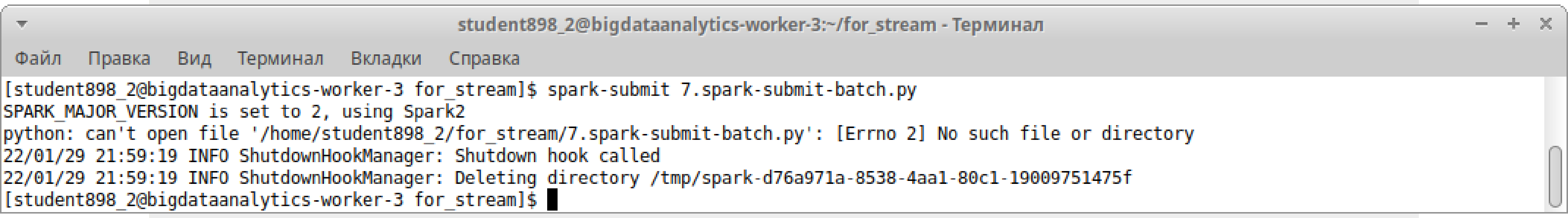
```
-bash: -put: команда не найдена
[student898_2@bigdataanalytics-worker-3 for_stream]$ hdfs dfs -mkdir input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 for_stream]$ hdfs dfs -put for_stream/product*.csv input_csv_for_stream
put: `for_stream/product*.csv': No such file or directory
[student898_2@bigdataanalytics-worker-3 for_stream]$ cd
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -mkdir input_csv_for_stream
mkdir: `input_csv_for_stream': File exists
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -put for_stream/product*.csv input_csv_for_stream
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 10 items
drwx------   - student898_2 student898_2          0 2022-01-28 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-27 18:51 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-24 19:39 checkpoints
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 input_csv_for_stream
Found 5 items
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list.csv
-rw-r--r--   2 student898_2 student898_2         98 2022-01-29 21:49 input_csv_for_stream/product_list1.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list2.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list3.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list4.csv
[student898_2@bigdataanalytics-worker-3 ~]$ 
```

```
hdfs dfs -ls for_stream
```


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

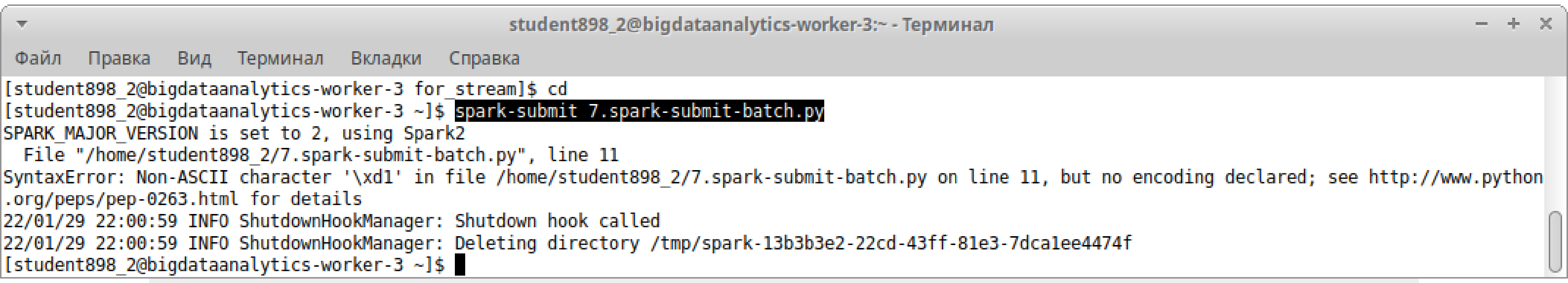
```
[student898_2@bigdataanalytics-worker-3 for_stream]$ hdfs dfs -ls for_stream
Found 5 items
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:38 for_stream/product_list.csv
-rw-r--r--   2 student898_2 student898_2         98 2022-01-29 21:38 for_stream/product_list1.csv
-rw-r--r--   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 for_stream]$ 
```

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


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

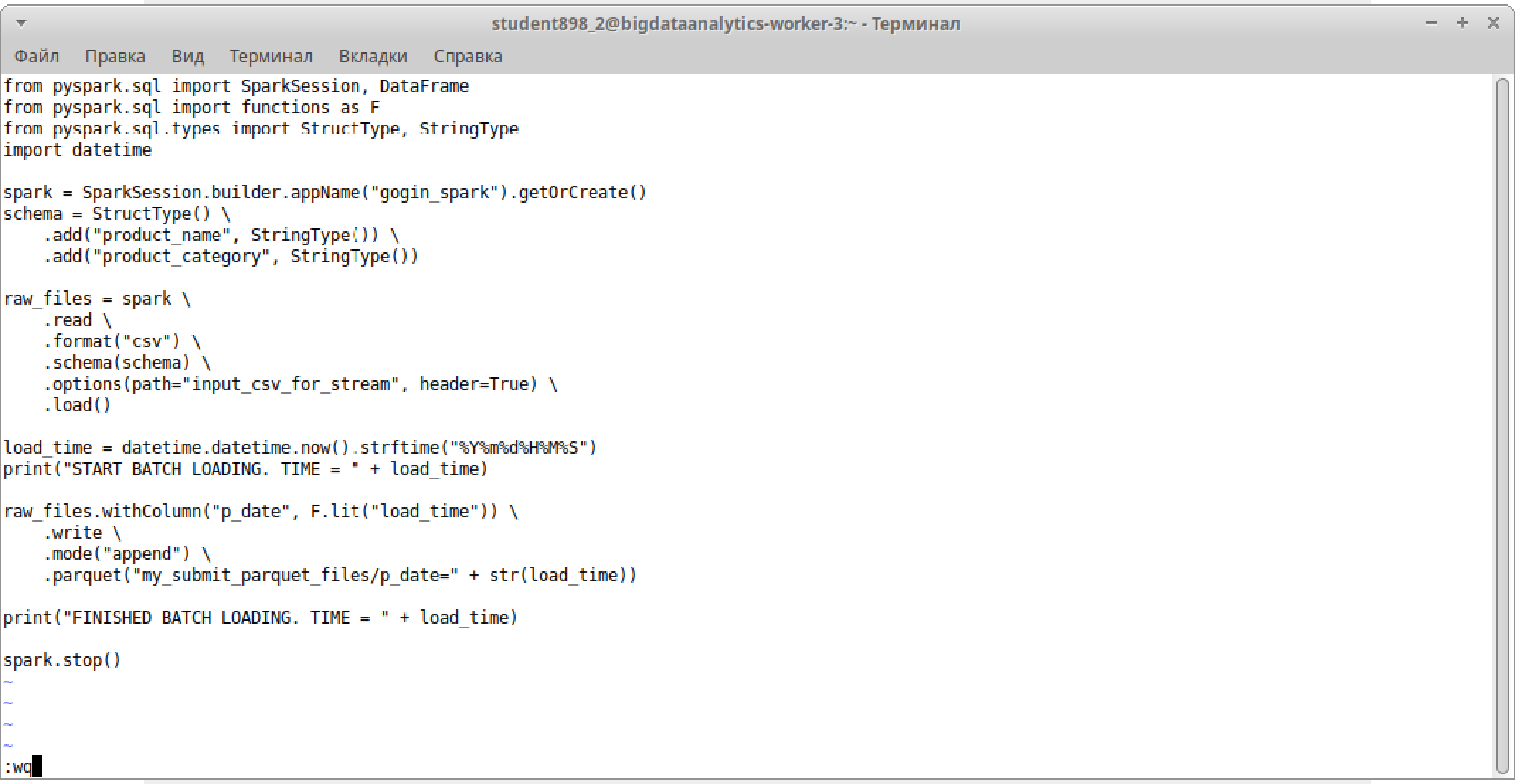
```
[student898_2@bigdataanalytics-worker-3 for_stream]$ spark-submit 7.spark-submit-batch.py
SPARK_MAJOR_VERSION is set to 2, using Spark2
python: can't open file '/home/student898_2/for_stream/7.spark-submit-batch.py': [Errno 2] No such file or directory
22/01/29 21:59:19 INFO ShutdownHookManager: Shutdown hook called
22/01/29 21:59:19 INFO ShutdownHookManager: Deleting directory /tmp/spark-d76a971a-8538-4aa1-80c1-19009751475f
[student898_2@bigdataanalytics-worker-3 for_stream]$ 
```

```
cd

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



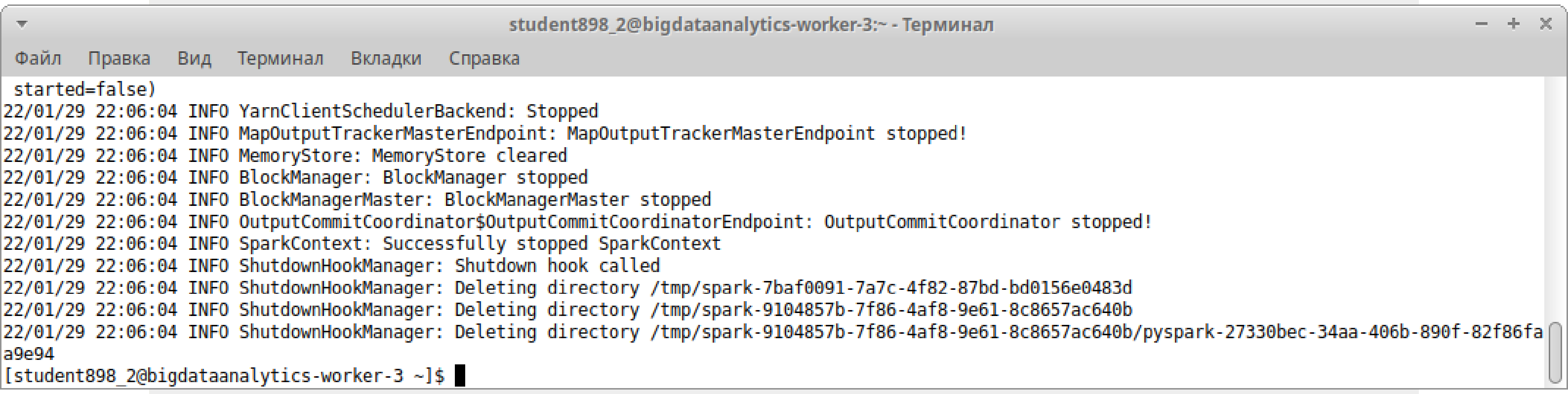
```
[student898_2@bigdataanalytics-worker-3 for_stream]$ cd
[student898_2@bigdataanalytics-worker-3 ~]$ spark-submit 7.spark-submit-batch.py
SPARK_MAJOR_VERSION is set to 2, using Spark2
  File "/home/student898_2/7.spark-submit-batch.py", line 11
SyntaxError: Non-ASCII character '\xd1' in file /home/student898_2/7.spark-submit-batch.py on line 11, but no encoding declared; see http://www.python
.org/peps/pep-0263.html for details
22/01/29 22:00:59 INFO ShutdownHookManager: Shutdown hook called
22/01/29 22:00:59 INFO ShutdownHookManager: Deleting directory /tmp/spark-13b3b3e2-22cd-43ff-81e3-7dca1ee4474f
[student898_2@bigdataanalytics-worker-3 ~]$ 
```

удаляю русские слова



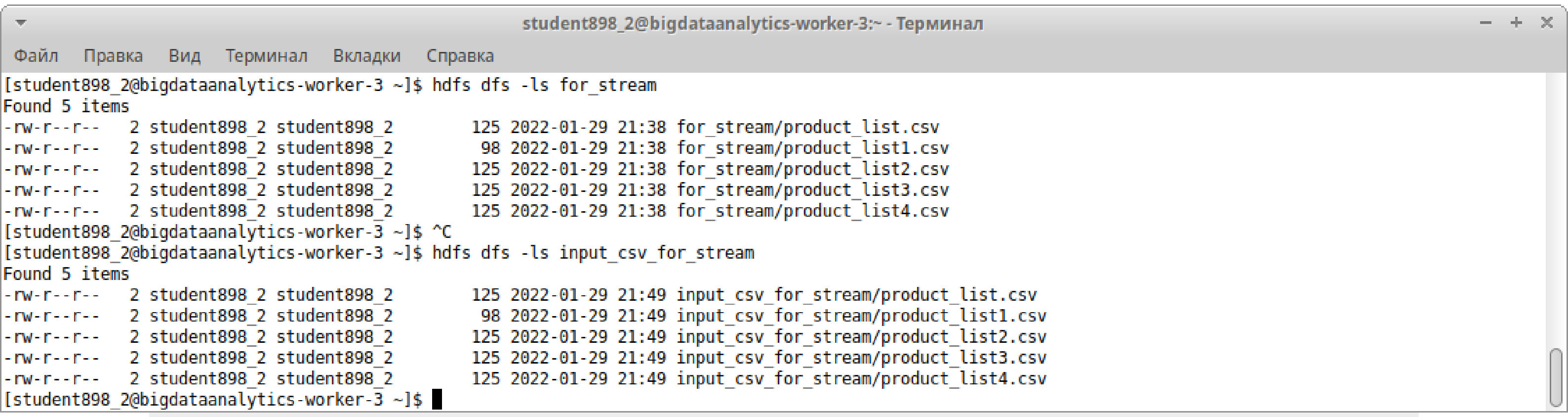
```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
schema = StructType() \
    .add("product_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()
~
~
~
~
:wq
```

запустился



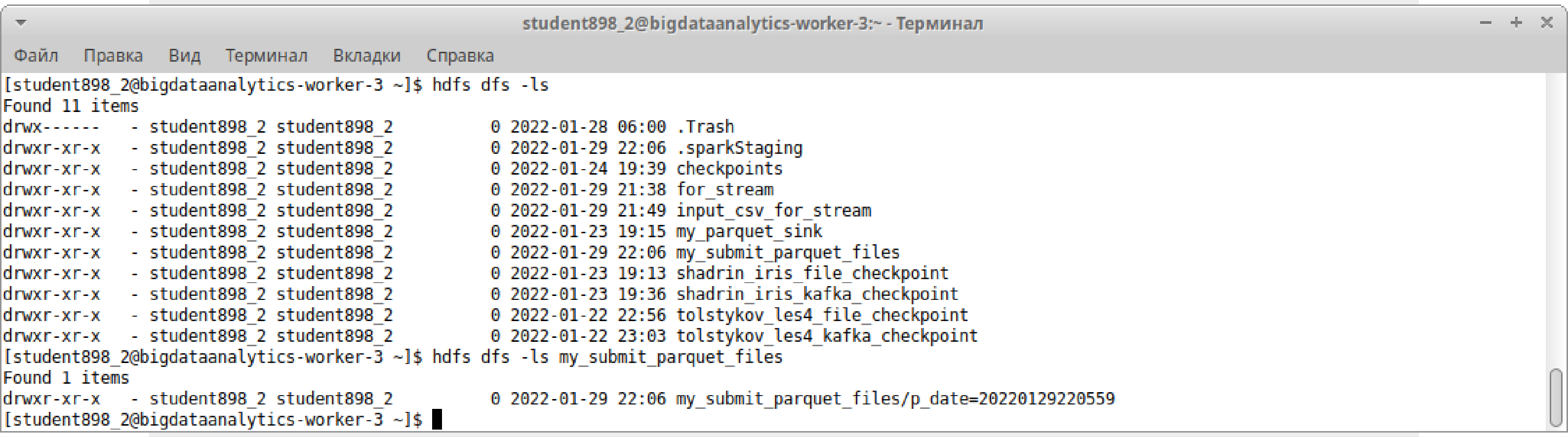
```
 started=false)
22/01/29 22:06:04 INFO YarnClientSchedulerBackend: Stopped
22/01/29 22:06:04 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/01/29 22:06:04 INFO MemoryStore: MemoryStore cleared
22/01/29 22:06:04 INFO BlockManager: BlockManager stopped
22/01/29 22:06:04 INFO BlockManagerMaster: BlockManagerMaster stopped
22/01/29 22:06:04 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/01/29 22:06:04 INFO SparkContext: Successfully stopped SparkContext
22/01/29 22:06:04 INFO ShutdownHookManager: Shutdown hook called
22/01/29 22:06:04 INFO ShutdownHookManager: Deleting directory /tmp/spark-7baf0091-7a7c-4f82-87bd-bd0156e0483d
22/01/29 22:06:04 INFO ShutdownHookManager: Deleting directory /tmp/spark-9104857b-7f86-4af8-9e61-8c8657ac640b
22/01/29 22:06:04 INFO ShutdownHookManager: Deleting directory /tmp/spark-9104857b-7f86-4af8-9e61-8c8657ac640b/pyspark-27330bec-34aa-406b-890f-82f86fa
a9e94
[student898_2@bigdataanalytics-worker-3 ~]$ 
```

```
hdfs dfs -ls for_stream

hdfs dfs -ls input_csv_for_stream
```



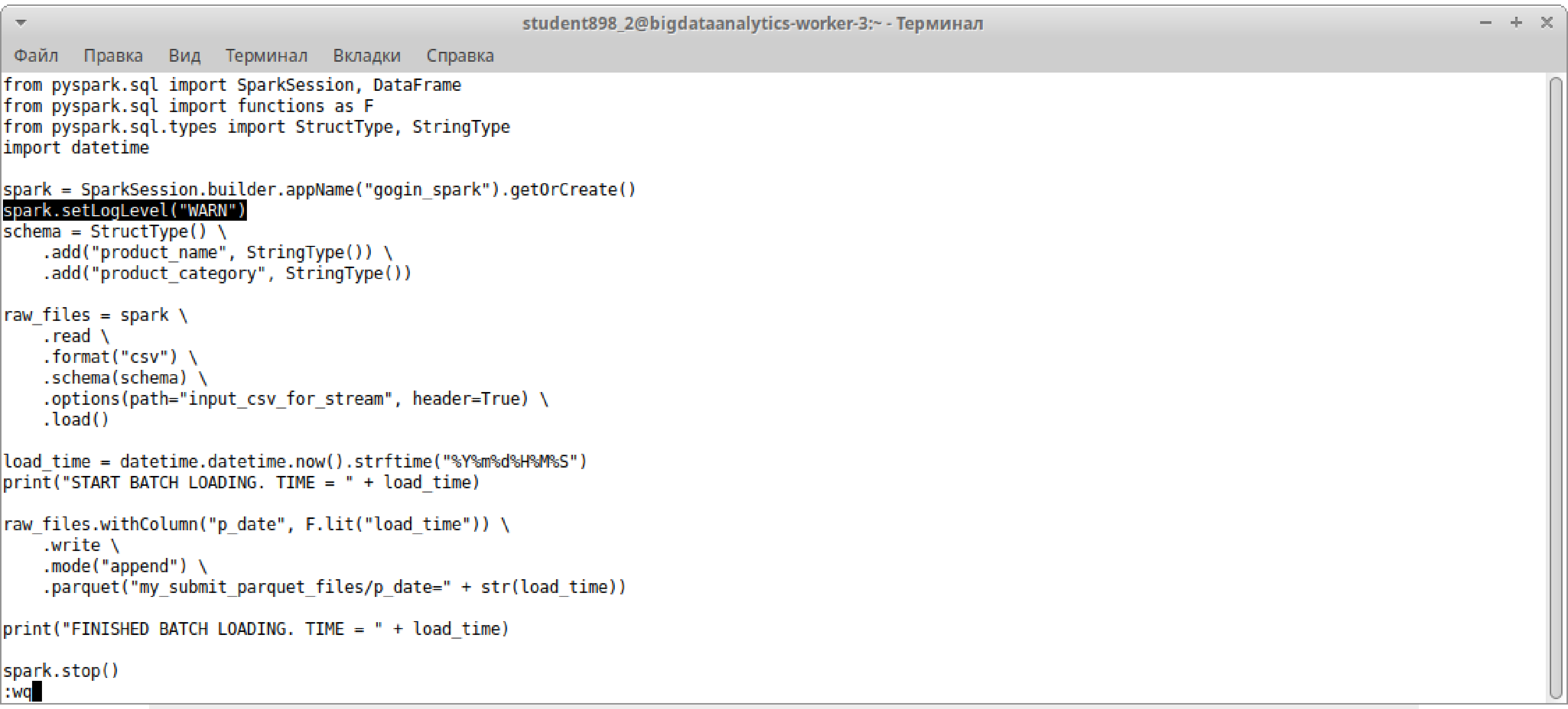
```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls for_stream
Found 5 items
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:38 for_stream/product_list.csv
-rw-r--r--   2 student898_2 student898_2         98 2022-01-29 21:38 for_stream/product_list1.csv
-rw-r--r--   2 student898_2 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 ~]$ ^C
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls input_csv_for_stream
Found 5 items
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list.csv
-rw-r--r--   2 student898_2 student898_2         98 2022-01-29 21:49 input_csv_for_stream/product_list1.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list2.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list3.csv
-rw-r--r--   2 student898_2 student898_2        125 2022-01-29 21:49 input_csv_for_stream/product_list4.csv
[student898_2@bigdataanalytics-worker-3 ~]$ 
```

```
hdfs dfs -ls

hdfs dfs -ls my_submit_parquet_files
```



```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - student898_2 student898_2          0 2022-01-28 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 22:06 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-24 19:39 checkpoints
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 2022-01-29 22:06 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 2022-01-29 22:06 my_submit_parquet_files/p_date=20220129220559
[student898_2@bigdataanalytics-worker-3 ~]$ 
```

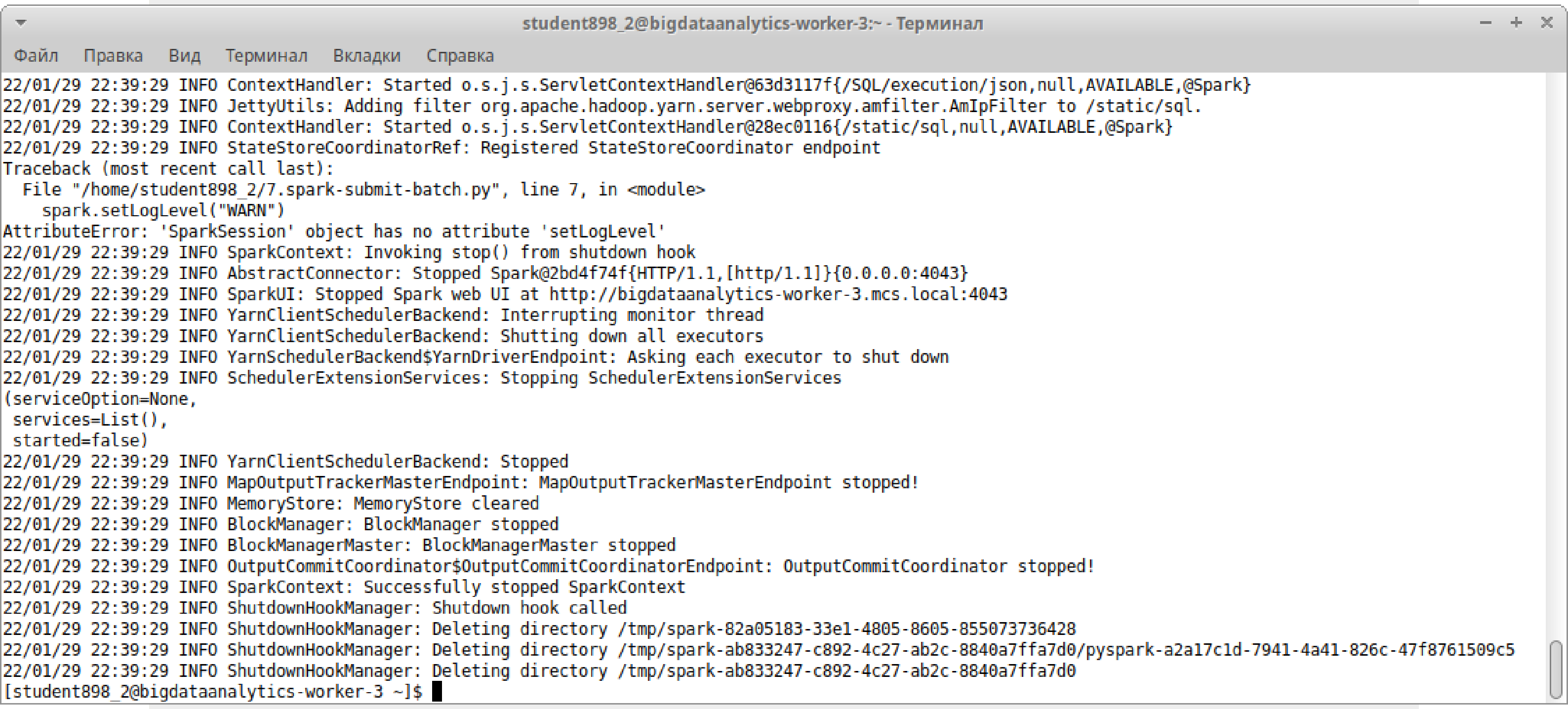
```
эксперементируем

vi 7.spark-submit-batch.py

вставляем

spark.setLogLevel("WARN")
```



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.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()
:wq
```

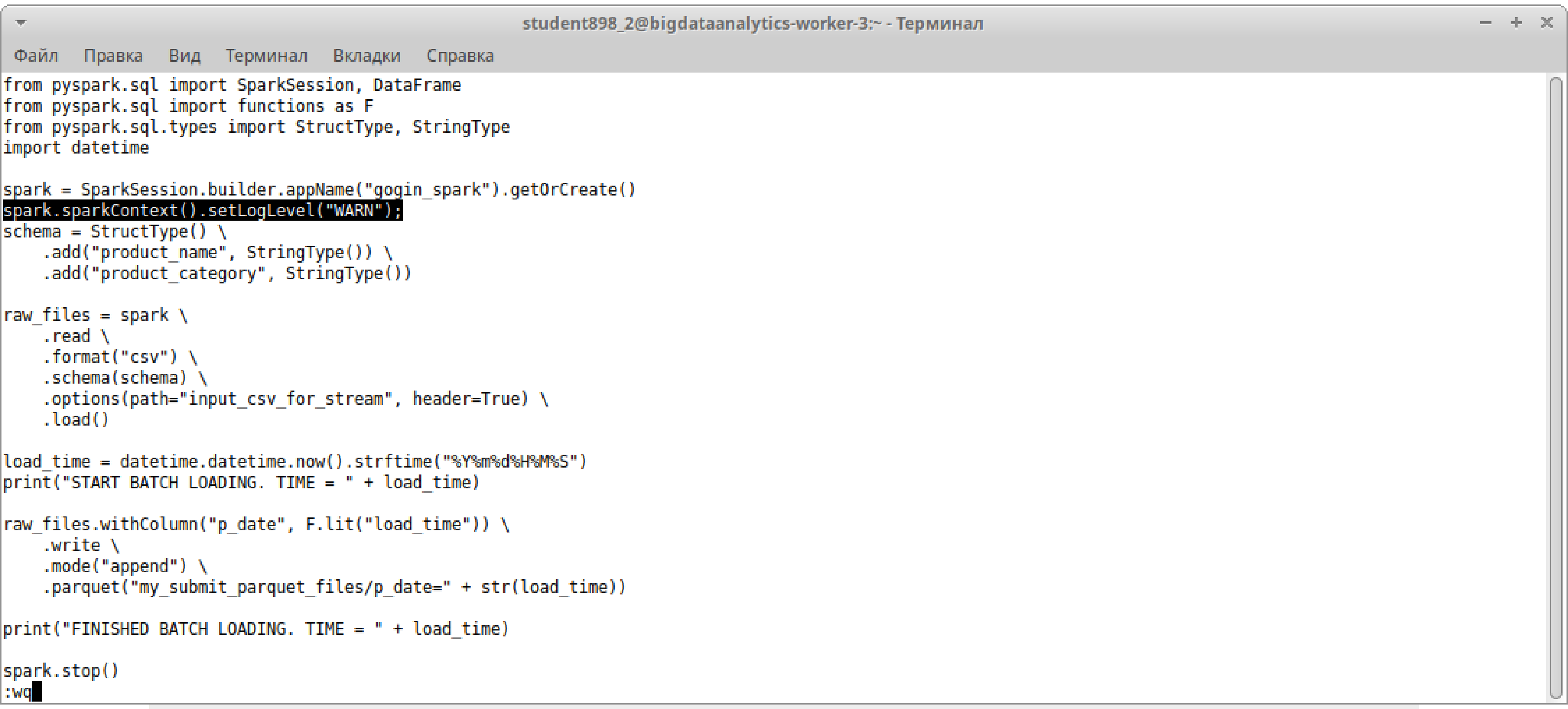
```
spark-submit 7.spark-submit-batch.py
```



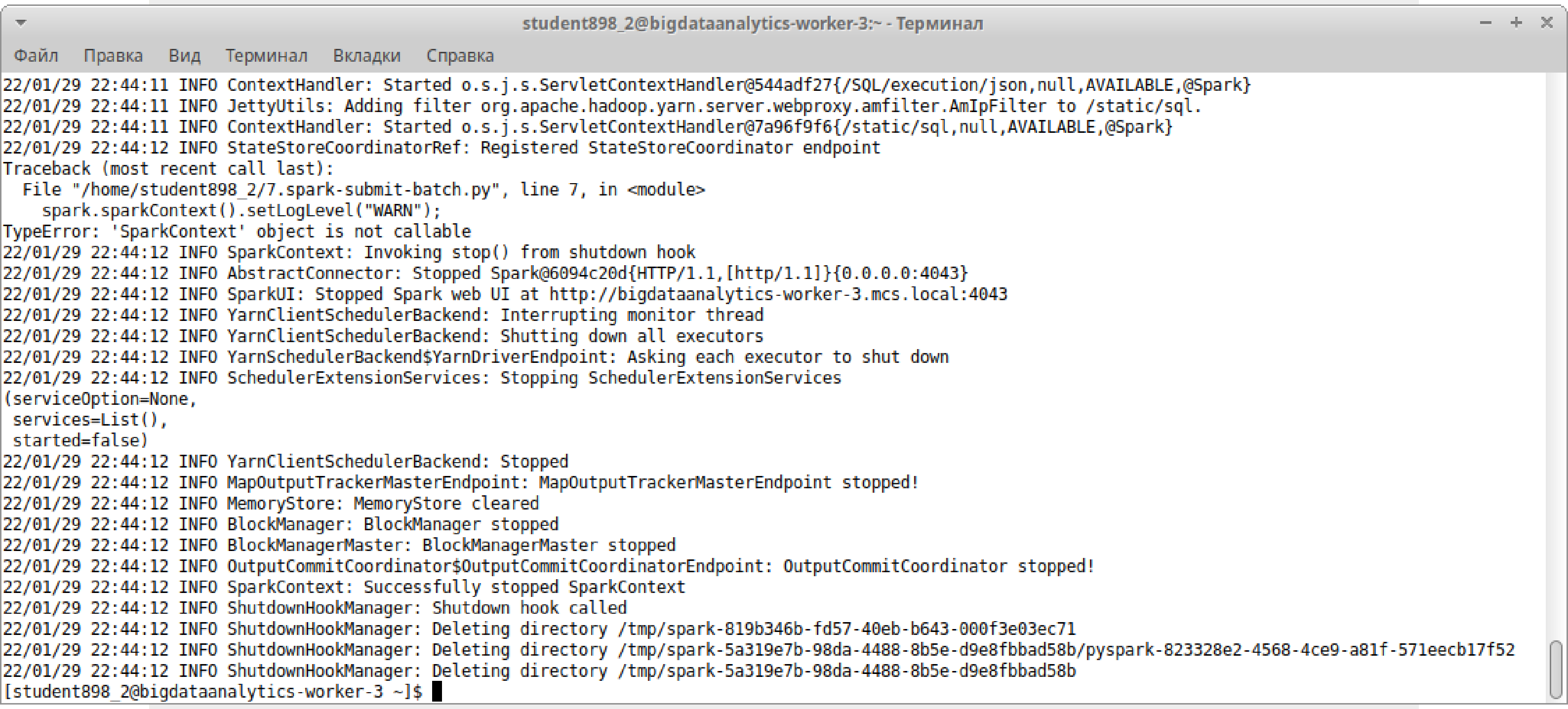
student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл Правка Вид Терминал Вкладки Справка

```
22/01/29 22:39:29 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@63d3117f{/SQL/execution/json,null,AVAILABLE,@Spark}
22/01/29 22:39:29 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
22/01/29 22:39:29 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@28ec0116{/static/sql,null,AVAILABLE,@Spark}
22/01/29 22:39:29 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
Traceback (most recent call last):
  File "/home/student898_2/7.spark-submit-batch.py", line 7, in <module>
    spark.setLogLevel("WARN")
AttributeError: 'SparkSession' object has no attribute 'setLogLevel'
22/01/29 22:39:29 INFO SparkContext: Invoking stop() from shutdown hook
22/01/29 22:39:29 INFO AbstractConnector: Stopped Spark@2bd4f74f{HTTP/1.1,[http/1.1]}{0.0.0.0:4043}
22/01/29 22:39:29 INFO SparkUI: Stopped Spark web UI at http://bigdataanalytics-worker-3.mcs.local:4043
22/01/29 22:39:29 INFO YarnClientSchedulerBackend: Interrupting monitor thread
22/01/29 22:39:29 INFO YarnClientSchedulerBackend: Shutting down all executors
22/01/29 22:39:29 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/01/29 22:39:29 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
22/01/29 22:39:29 INFO YarnClientSchedulerBackend: Stopped
22/01/29 22:39:29 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/01/29 22:39:29 INFO MemoryStore: MemoryStore cleared
22/01/29 22:39:29 INFO BlockManager: BlockManager stopped
22/01/29 22:39:29 INFO BlockManagerMaster: BlockManagerMaster stopped
22/01/29 22:39:29 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/01/29 22:39:29 INFO SparkContext: Successfully stopped SparkContext
22/01/29 22:39:29 INFO ShutdownHookManager: Shutdown hook called
22/01/29 22:39:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-82a05183-33e1-4805-8605-855073736428
22/01/29 22:39:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-ab833247-c892-4c27-ab2c-8840a7ffa7d0/pyspark-a2a17c1d-7941-4a41-826c-47f8761509c5
22/01/29 22:39:29 INFO ShutdownHookManager: Deleting directory /tmp/spark-ab833247-c892-4c27-ab2c-8840a7ffa7d0
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
spark.sparkContext().setLogLevel("WARN");
```

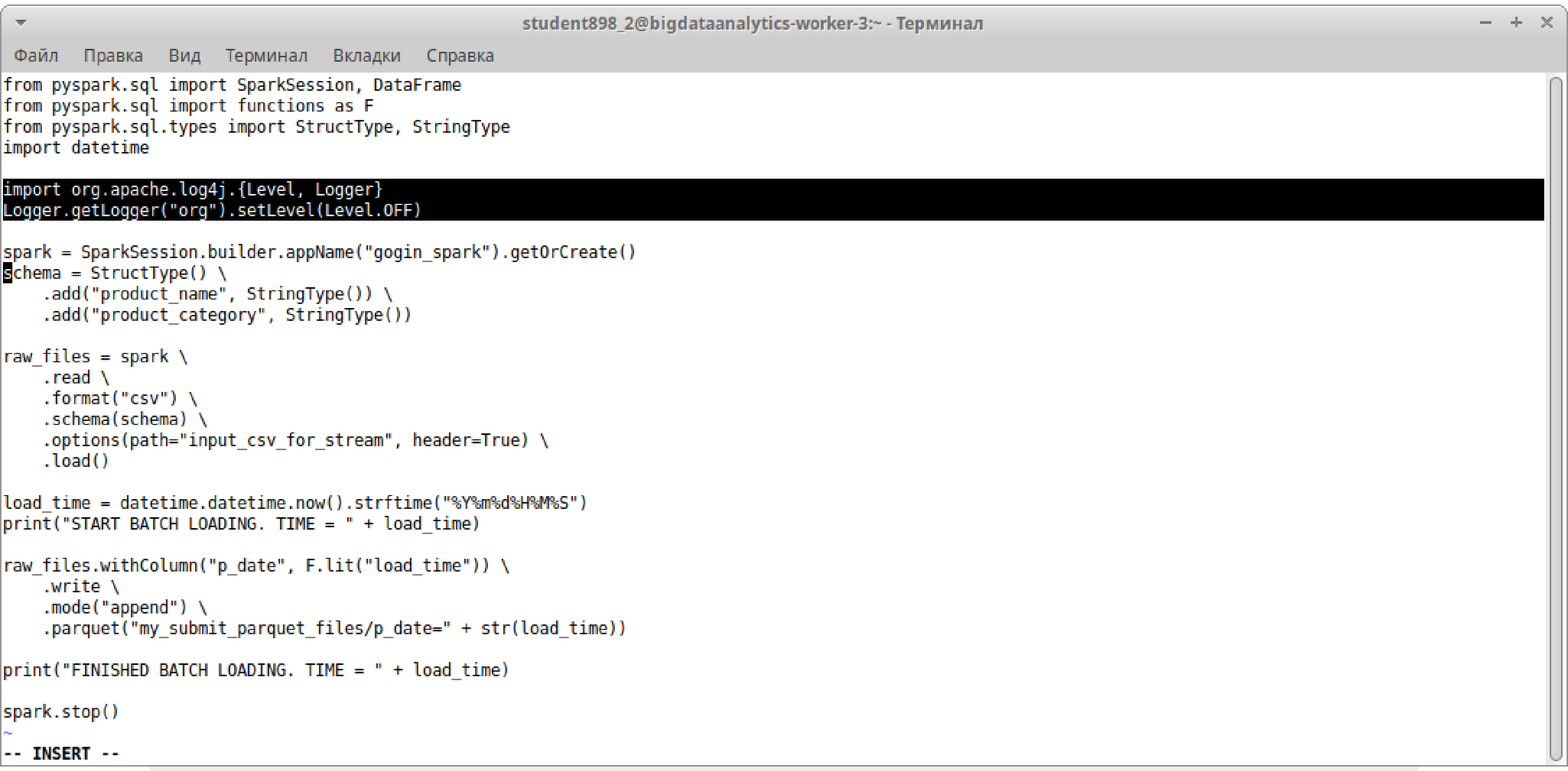


```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

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()
:wq
```

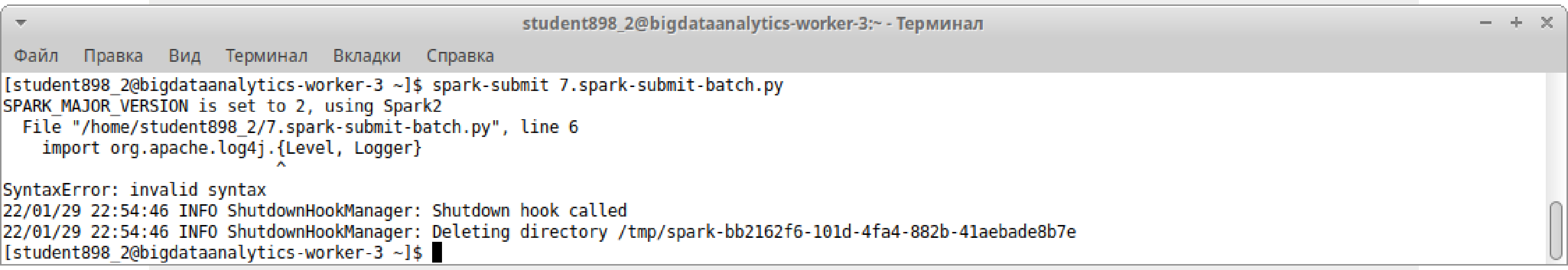


```
22/01/29 22:44:11 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@544adf27{/SQL/execution/json,null,AVAILABLE,@Spark}
22/01/29 22:44:11 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
22/01/29 22:44:11 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@7a96f9f6{/static/sql,null,AVAILABLE,@Spark}
22/01/29 22:44:12 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
Traceback (most recent call last):
  File "/home/student898_2/7.spark-submit-batch.py", line 7, in <module>
    spark.sparkContext().setLogLevel("WARN");
TypeError: 'SparkContext' object is not callable
22/01/29 22:44:12 INFO SparkContext: Invoking stop() from shutdown hook
22/01/29 22:44:12 INFO AbstractConnector: Stopped Spark@6094c20d{HTTP/1.1,[http/1.1]}{0.0.0.0:4043}
22/01/29 22:44:12 INFO SparkUI: Stopped Spark web UI at http://bigdataanalytics-worker-3.mcs.local:4043
22/01/29 22:44:12 INFO YarnClientSchedulerBackend: Interrupting monitor thread
22/01/29 22:44:12 INFO YarnClientSchedulerBackend: Shutting down all executors
22/01/29 22:44:12 INFO YarnSchedulerBackend$YarnDriverEndpoint: Asking each executor to shut down
22/01/29 22:44:12 INFO SchedulerExtensionServices: Stopping SchedulerExtensionServices
(serviceOption=None,
 services=List(),
 started=false)
22/01/29 22:44:12 INFO YarnClientSchedulerBackend: Stopped
22/01/29 22:44:12 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
22/01/29 22:44:12 INFO MemoryStore: MemoryStore cleared
22/01/29 22:44:12 INFO BlockManager: BlockManager stopped
22/01/29 22:44:12 INFO BlockManagerMaster: BlockManagerMaster stopped
22/01/29 22:44:12 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
22/01/29 22:44:12 INFO SparkContext: Successfully stopped SparkContext
22/01/29 22:44:12 INFO ShutdownHookManager: Shutdown hook called
22/01/29 22:44:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-819b346b-fd57-40eb-b643-000f3e03ec71
22/01/29 22:44:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-5a319e7b-98da-4488-8b5e-d9e8fbbad58b/pyspark-823328e2-4568-4ce9-a81f-571eecb17f52
22/01/29 22:44:12 INFO ShutdownHookManager: Deleting directory /tmp/spark-5a319e7b-98da-4488-8b5e-d9e8fbbad58b
[student898_2@bigdataanalytics-worker-3 ~]$
```

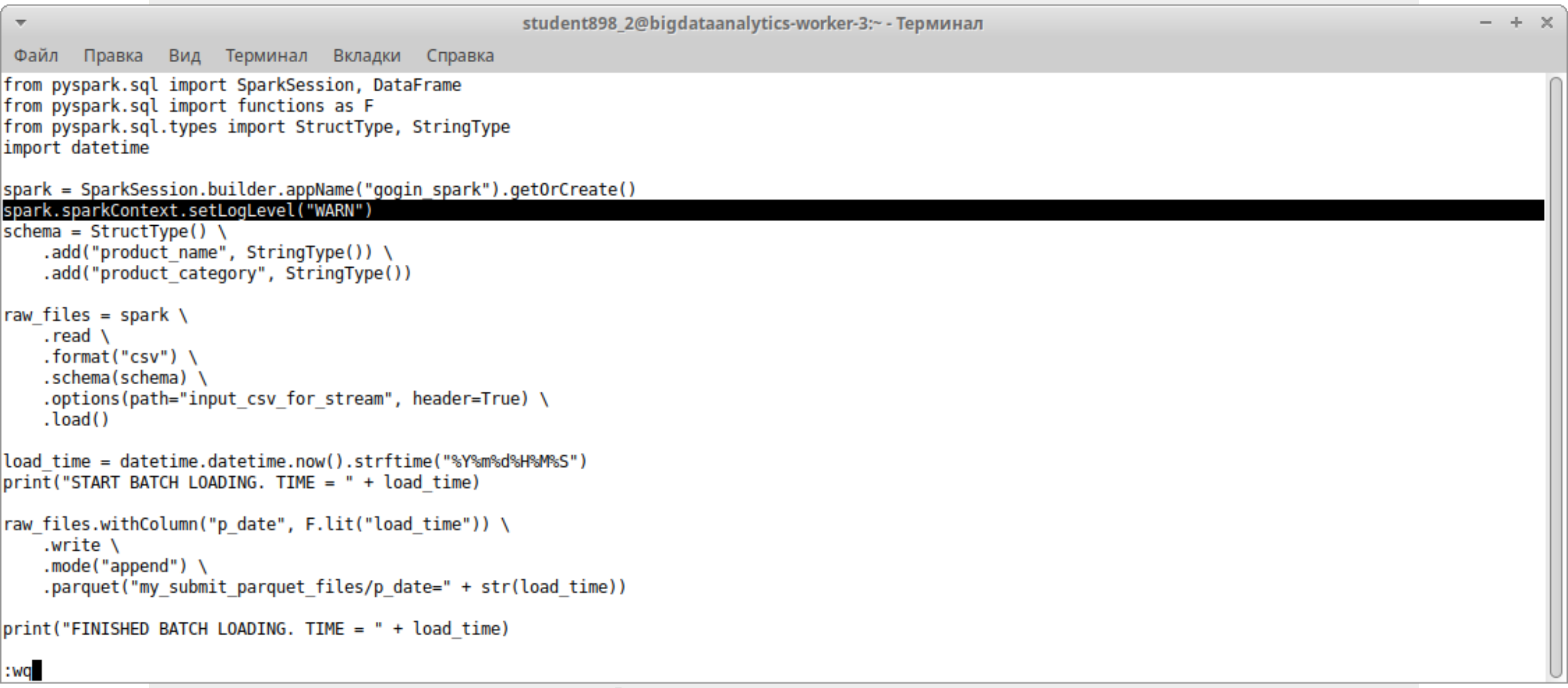
```
еще один вариант

import org.apache.log4j.{Level, Logger}

Logger.getLogger("org").setLevel(Level.OFF)
```


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

import org.apache.log4j.{Level, Logger}
Logger.getLogger("org").setLevel(Level.OFF)

spark = SparkSession.builder.appName("gogin_spark").getOrCreate()
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()
~
-- INSERT --
```


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

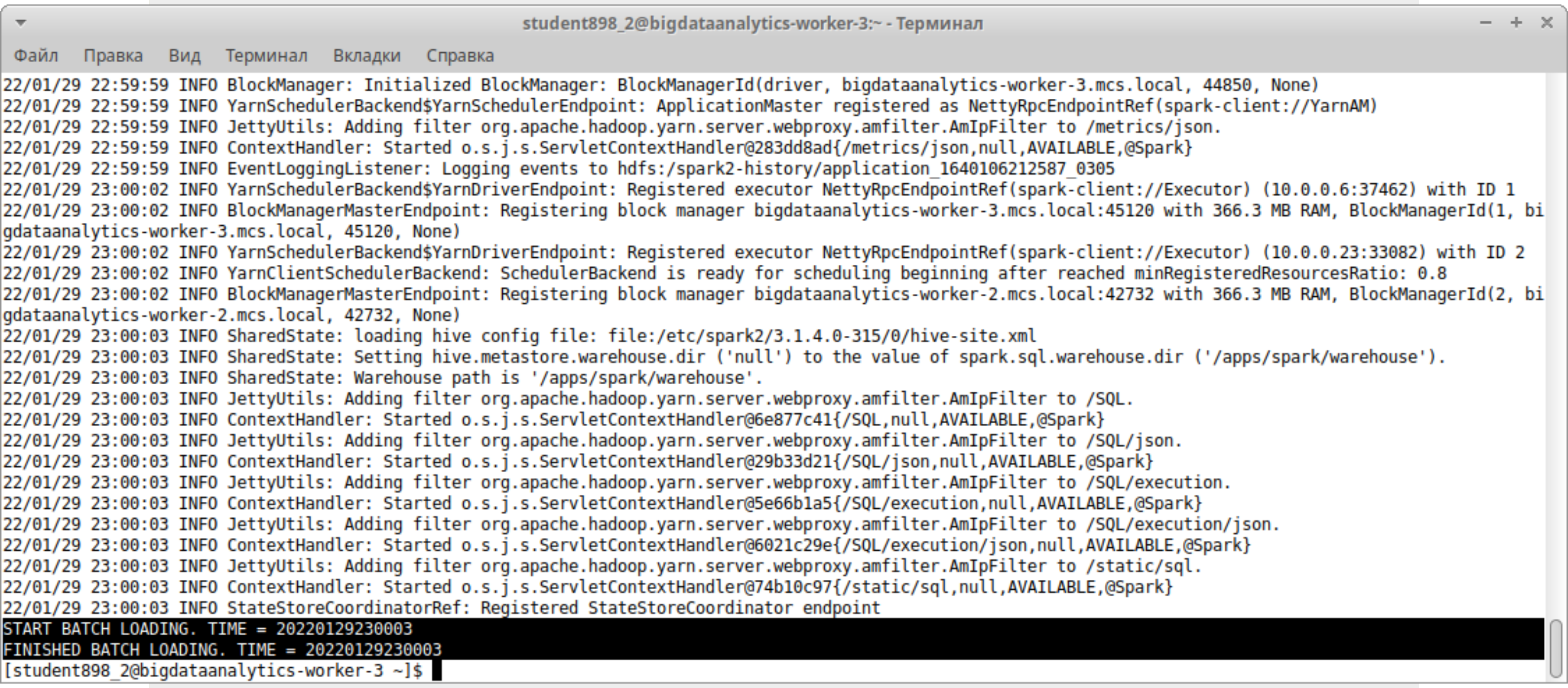
```
[student898_2@bigdataanalytics-worker-3 ~]$ spark-submit 7.spark-submit-batch.py
SPARK_MAJOR_VERSION is set to 2, using Spark2
  File "/home/student898_2/7.spark-submit-batch.py", line 6
    import org.apache.log4j.{Level, Logger}
                            ^

SyntaxError: invalid syntax
22/01/29 22:54:46 INFO ShutdownHookManager: Shutdown hook called
22/01/29 22:54:46 INFO ShutdownHookManager: Deleting directory /tmp/spark-bb2162f6-101d-4fa4-882b-41aebade8b7e
[student898_2@bigdataanalytics-worker-3 ~]$
```


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)

:wq
```

```
spark.sparkContext.setLogLevel("WARN")
```

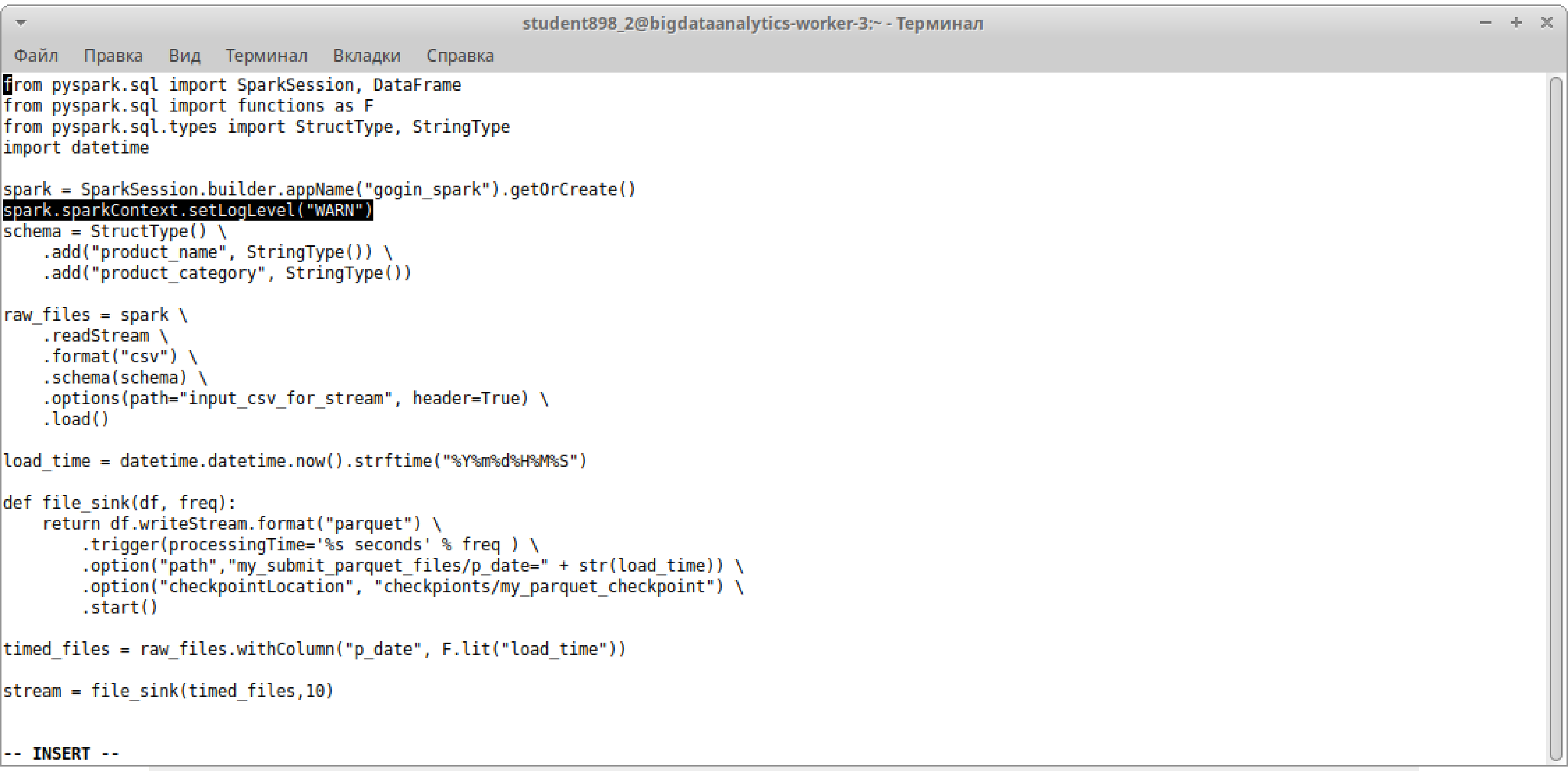
```
START BATCH LOADING. TIME = 20220129230003

FINISHED BATCH LOADING. TIME = 20220129230003
```



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

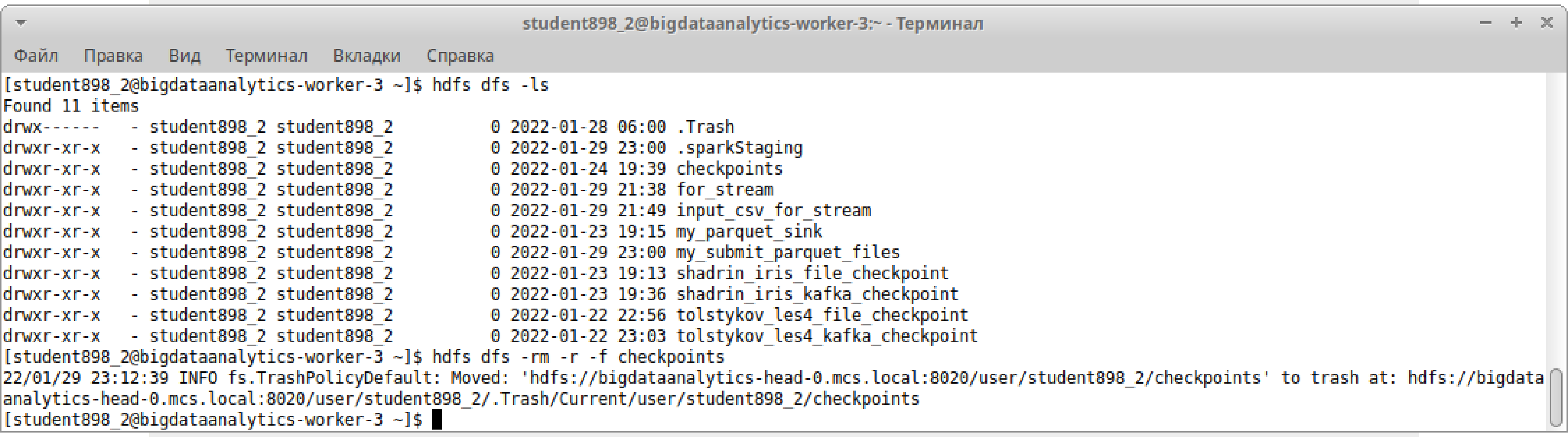
```
22/01/29 22:59:59 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, bigdataanalytics-worker-3.mcs.local, 44850, None)
22/01/29 22:59:59 INFO YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(spark-client://YarnAM)
22/01/29 22:59:59 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /metrics/json.
22/01/29 22:59:59 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@283dd8ad{/metrics/json,null,AVAILABLE,@Spark}
22/01/29 22:59:59 INFO EventLoggingListener: Logging events to hdfs:/spark2-history/application_1640106212587_0305
22/01/29 23:00:02 INFO YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (10.0.0.6:37462) with ID 1
22/01/29 23:00:02 INFO BlockManagerMasterEndpoint: Registering block manager bigdataanalytics-worker-3.mcs.local:45120 with 366.3 MB RAM, BlockManagerId(1, bi
gdataanalytics-worker-3.mcs.local, 45120, None)
22/01/29 23:00:02 INFO YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (10.0.0.23:33082) with ID 2
22/01/29 23:00:02 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.8
22/01/29 23:00:02 INFO BlockManagerMasterEndpoint: Registering block manager bigdataanalytics-worker-2.mcs.local:42732 with 366.3 MB RAM, BlockManagerId(2, bi
gdataanalytics-worker-2.mcs.local, 42732, None)
22/01/29 23:00:03 INFO SharedState: loading hive config file: file:/etc/spark2/3.1.4.0-315/0/hive-site.xml
22/01/29 23:00:03 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('/apps/spark/warehouse').
22/01/29 23:00:03 INFO SharedState: Warehouse path is '/apps/spark/warehouse'.
22/01/29 23:00:03 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL.
22/01/29 23:00:03 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@6e877c41{/SQL,null,AVAILABLE,@Spark}
22/01/29 23:00:03 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/json.
22/01/29 23:00:03 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@29b33d21{/SQL/json,null,AVAILABLE,@Spark}
22/01/29 23:00:03 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution.
22/01/29 23:00:03 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@5e66b1a5{/SQL/execution,null,AVAILABLE,@Spark}
22/01/29 23:00:03 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution/json.
22/01/29 23:00:03 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@6021c29e{/SQL/execution/json,null,AVAILABLE,@Spark}
22/01/29 23:00:03 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
22/01/29 23:00:03 INFO ContextHandler: Started o.s.j.s.ServletContextHandler@74b10c97{/static/sql,null,AVAILABLE,@Spark}
22/01/29 23:00:03 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
START BATCH LOADING. TIME = 20220129230003
FINISHED BATCH LOADING. TIME = 20220129230003
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
переношу код в остальные файлы

spark.sparkContext.setLogLevel("WARN")

vi 7.1_spark-submit_stream.py

удаляю русские слова
```



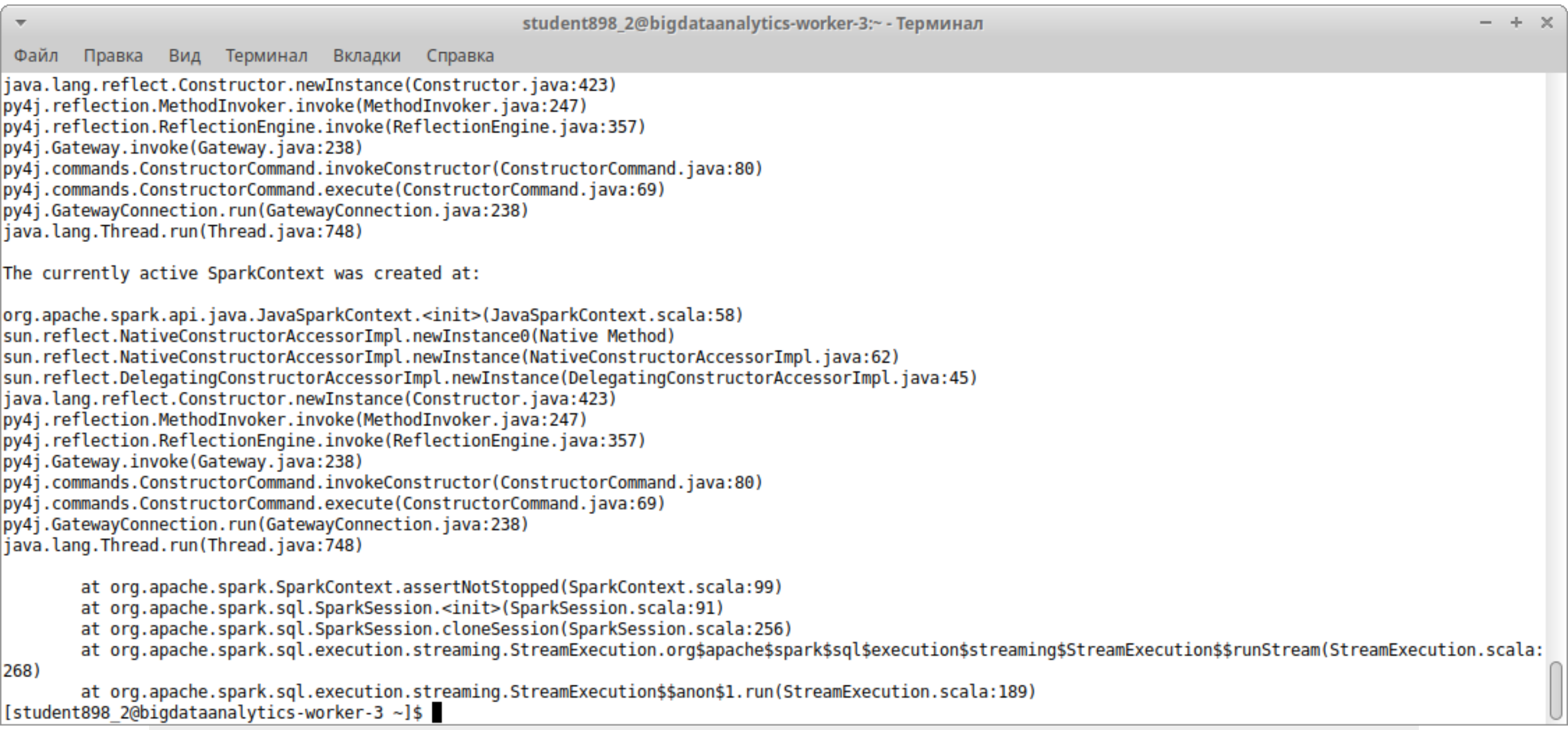
```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

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="input_csv_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)


-- INSERT --
```

```
hdfs dfs -ls

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



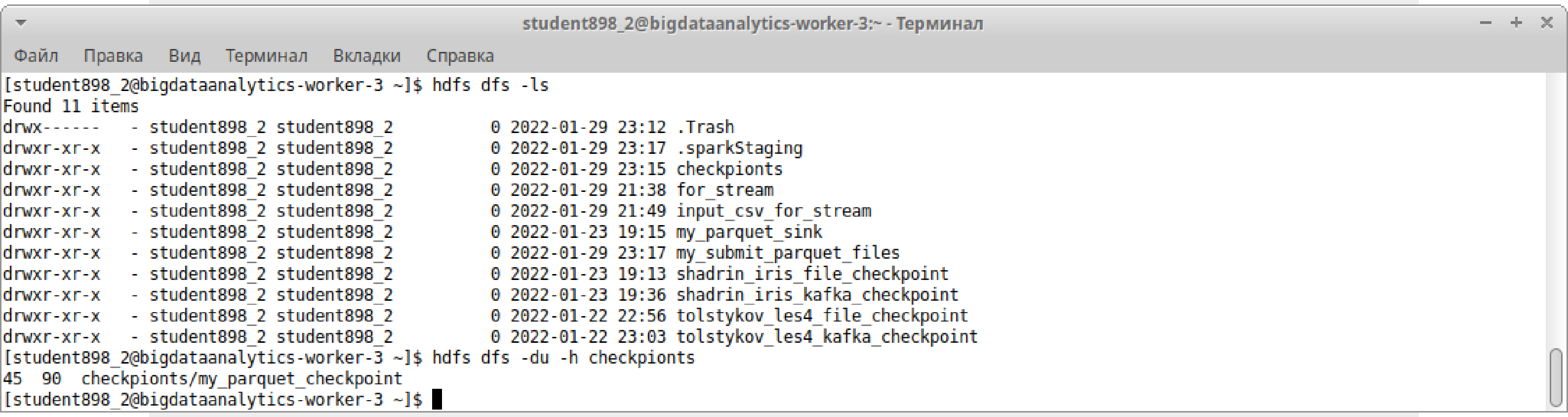
```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - student898_2 student898_2          0 2022-01-28 06:00 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:00 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-24 19:39 checkpoints
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 2022-01-29 23:00 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
22/01/29 23:12:39 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/checkpoints' to trash at: hdfs://bigdata
analytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/checkpoints
[student898_2@bigdataanalytics-worker-3 ~]$
```

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



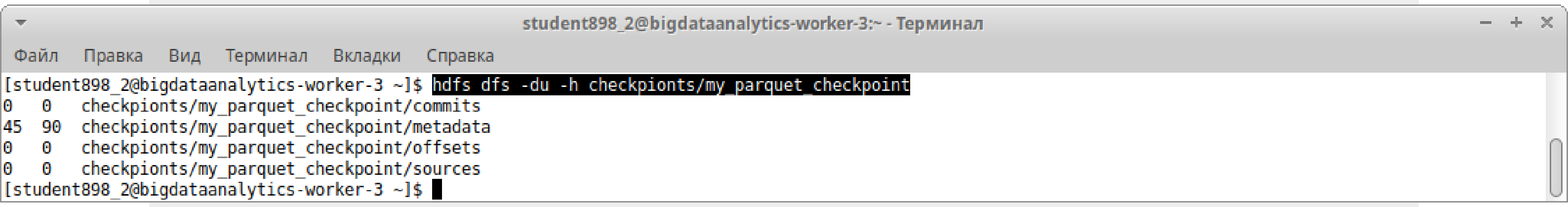
```
java.lang.reflect.Constructor.newInstance(Constructor.java:423)
py4j.reflection.MethodInvoker.invoke(MethodInvoker.java:247)
py4j.reflection.ReflectionEngine.invoke(ReflectionEngine.java:357)
py4j.Gateway.invoke(Gateway.java:238)
py4j.commands.ConstructorCommand.invokeConstructor(ConstructorCommand.java:80)
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$$runStream(StreamExecution.scala:
268)
        at org.apache.spark.sql.execution.streaming.StreamExecution$$anon$1.run(StreamExecution.scala:189)
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
hdfs dfs -ls

hdfs dfs -du -h checkpionts
```



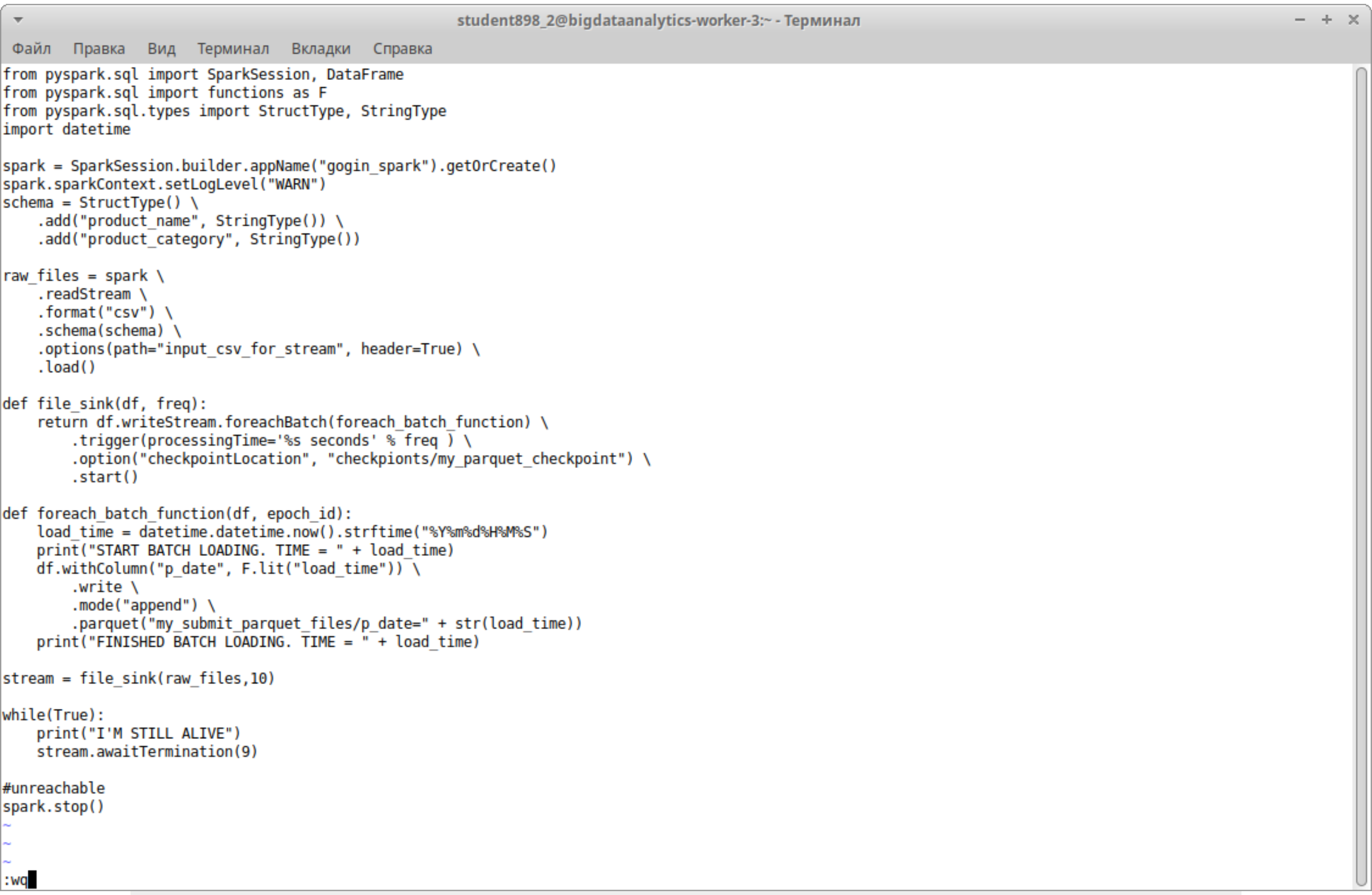
```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - student898_2 student898_2          0 2022-01-29 23:12 .Trash
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:17 .sparkStaging
drwxr-xr-x   - student898_2 student898_2          0 2022-01-29 23:15 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 2022-01-29 23:17 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
45  90  checkpionts/my_parquet_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$
```

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



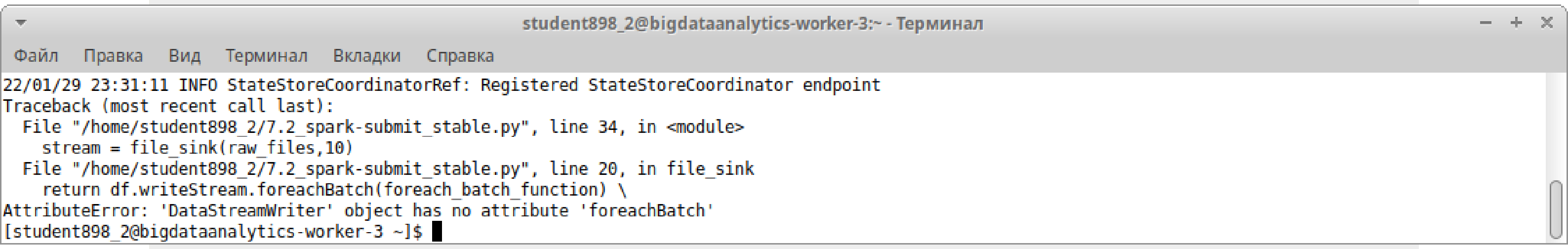
```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h checkpionts/my_parquet_checkpoint
0   0   checkpionts/my_parquet_checkpoint/commits
45  90  checkpionts/my_parquet_checkpoint/metadata
0   0   checkpionts/my_parquet_checkpoint/offsets
0   0   checkpionts/my_parquet_checkpoint/sources
[student898_2@bigdataanalytics-worker-3 ~]$
```

```
spark.sparkContext.setLogLevel("WARN")

vi 7.2_spark-submit_stable.py
```


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

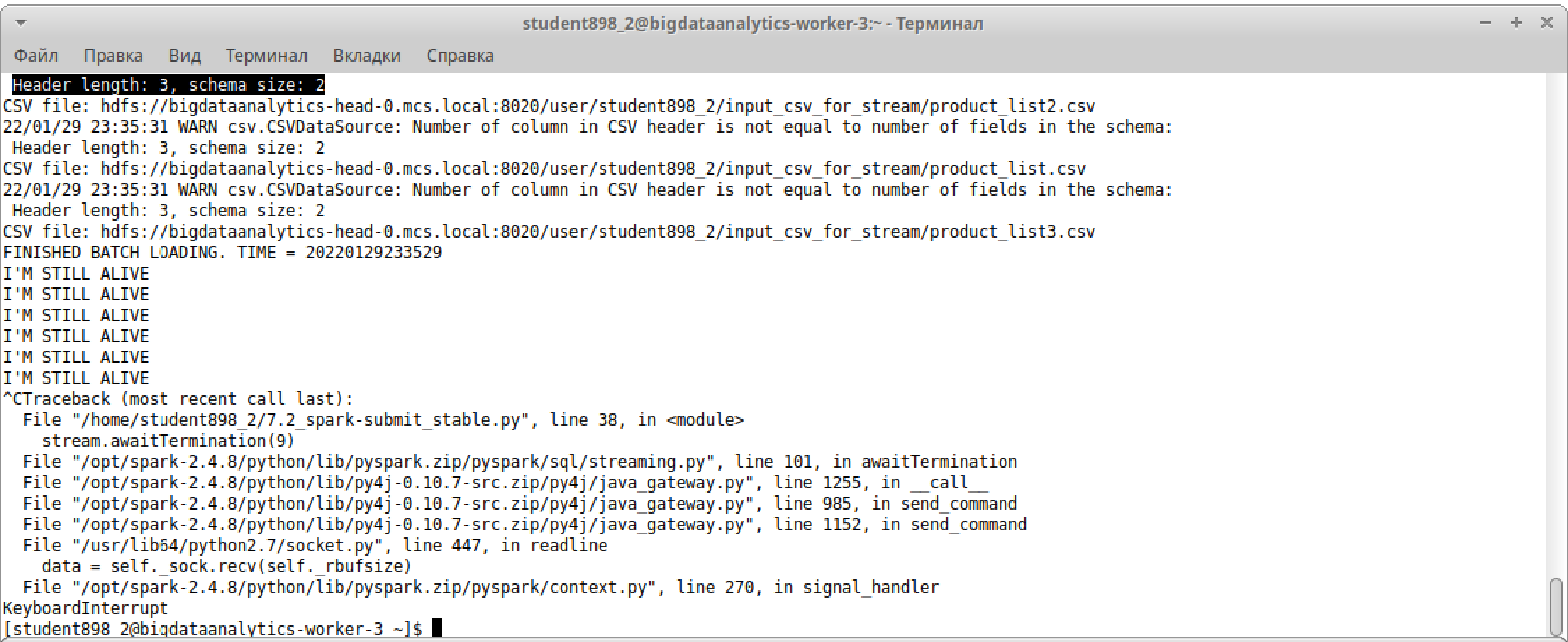
```
from pyspark.sql import SparkSession, DataFrame
from pyspark.sql import functions as F
from pyspark.sql.types import StructType, StringType
import datetime

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

raw_files = spark \
    .readStream \
    .format("csv") \
    .schema(schema) \
    .options(path="input_csv_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
```

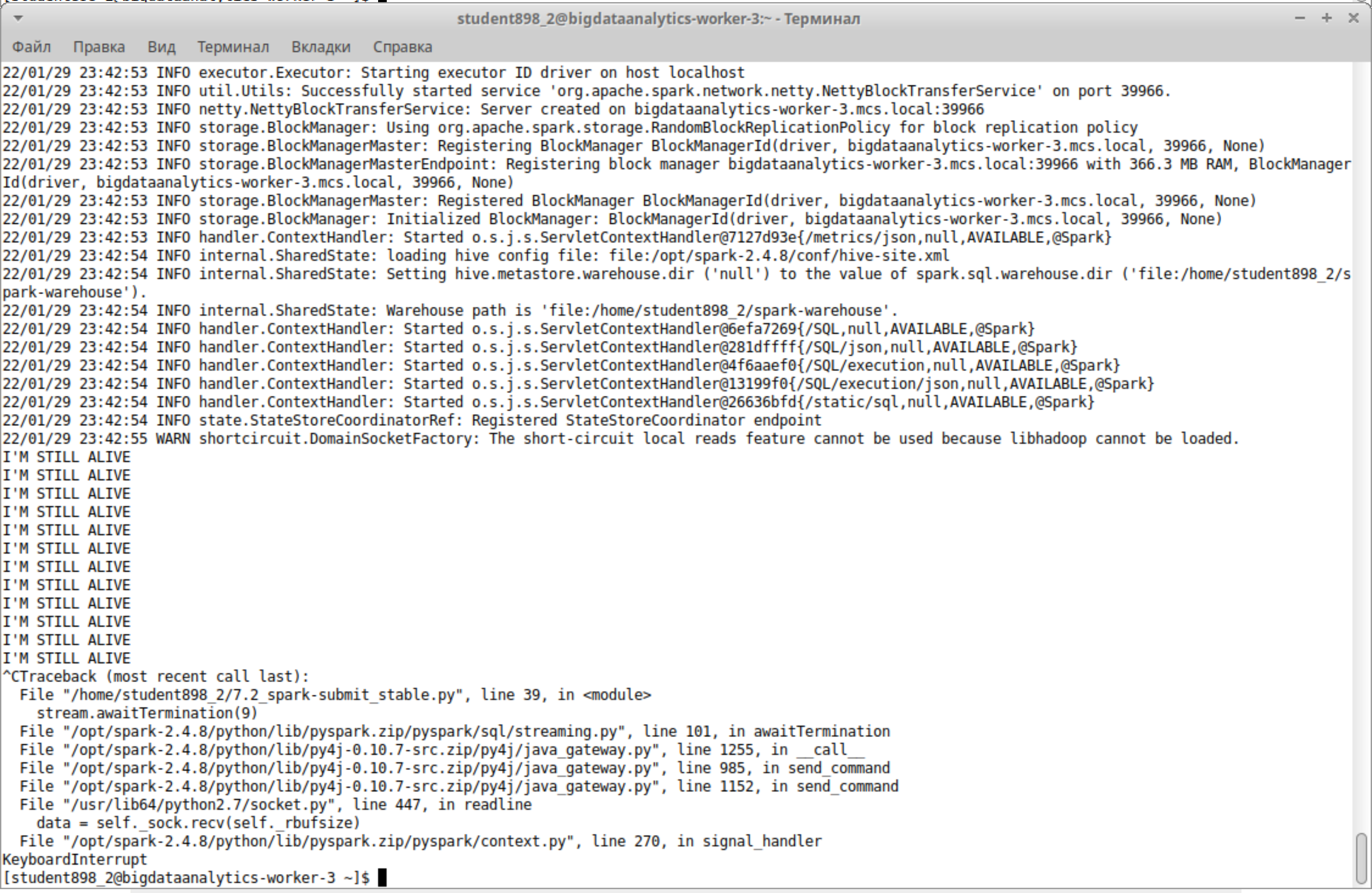

student898_2@bigdataanalytics-worker-3:~ - Терминал
Файл Правка Вид Терминал Вкладки Справка

```
22/01/29 23:31:11 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
Traceback (most recent call last):
  File "/home/student898_2/7.2_spark-submit_stable.py", line 34, in <module>
    stream = file_sink(raw_files,10)
  File "/home/student898_2/7.2_spark-submit_stable.py", line 20, 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
```

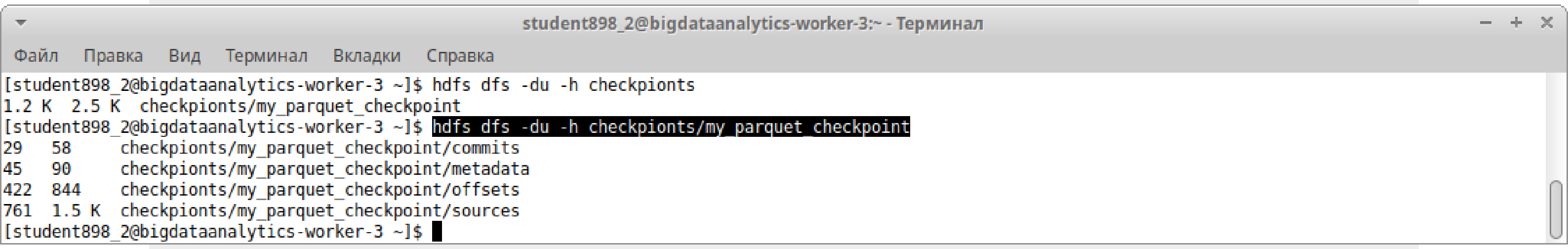


```
 Header length: 3, schema size: 2
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/input_csv_for_stream/product_list2.csv
22/01/29 23:35:31 WARN csv.CSVDataSource: Number of column in CSV header is not equal to number of fields in the schema:
 Header length: 3, schema size: 2
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/input_csv_for_stream/product_list.csv
22/01/29 23:35:31 WARN csv.CSVDataSource: Number of column in CSV header is not equal to number of fields in the schema:
 Header length: 3, schema size: 2
CSV file: hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/input_csv_for_stream/product_list3.csv
FINISHED BATCH LOADING. TIME = 20220129233529
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 38, 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 ~]$ 
```



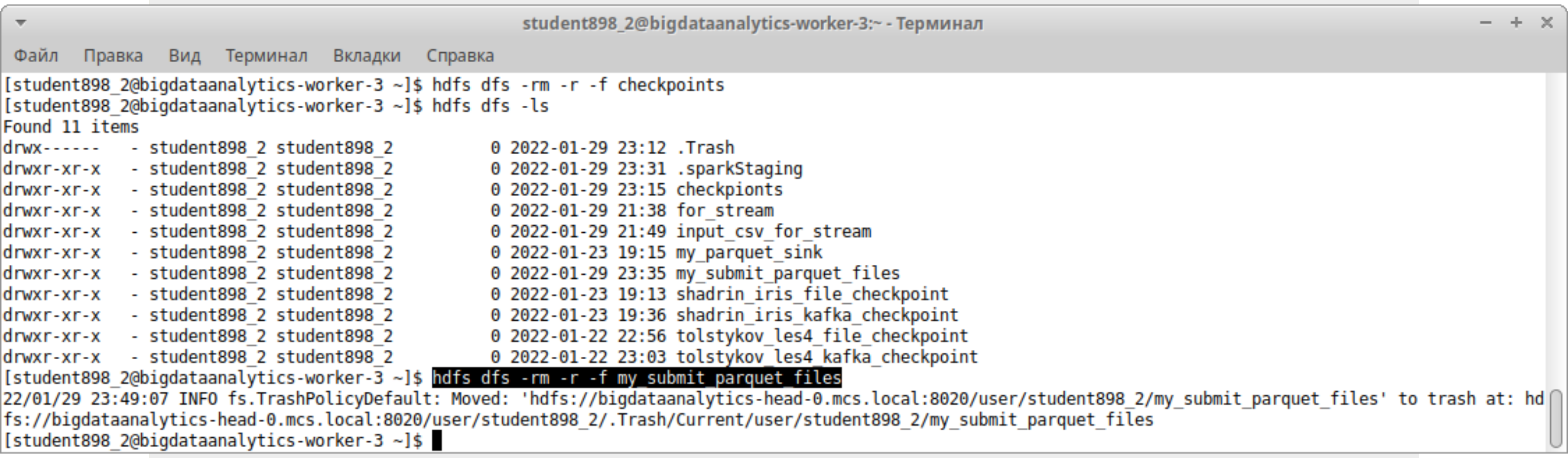
```
22/01/29 23:42:53 INFO executor.Executor: Starting executor ID driver on host localhost
22/01/29 23:42:53 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 39966.
22/01/29 23:42:53 INFO netty.NettyBlockTransferService: Server created on bigdataanalytics-worker-3.mcs.local:39966
22/01/29 23:42:53 INFO storage.BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/01/29 23:42:53 INFO storage.BlockManagerMaster: Registering BlockManager BlockManagerId(driver, bigdataanalytics-worker-3.mcs.local, 39966, None)
22/01/29 23:42:53 INFO storage.BlockManagerMasterEndpoint: Registering block manager bigdataanalytics-worker-3.mcs.local:39966 with 366.3 MB RAM, BlockManager
Id(driver, bigdataanalytics-worker-3.mcs.local, 39966, None)
22/01/29 23:42:53 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, bigdataanalytics-worker-3.mcs.local, 39966, None)
22/01/29 23:42:53 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, bigdataanalytics-worker-3.mcs.local, 39966, None)
22/01/29 23:42:53 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@7127d93e{/metrics/json,null,AVAILABLE,@Spark}
22/01/29 23:42:54 INFO internal.SharedState: loading hive config file: file:/opt/spark-2.4.8/conf/hive-site.xml
22/01/29 23:42:54 INFO internal.SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('file:/home/student898_2/s
park-warehouse').
22/01/29 23:42:54 INFO internal.SharedState: Warehouse path is 'file:/home/student898_2/spark-warehouse'.
22/01/29 23:42:54 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@6efa7269{/SQL,null,AVAILABLE,@Spark}
22/01/29 23:42:54 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@281dffff{/SQL/json,null,AVAILABLE,@Spark}
22/01/29 23:42:54 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4f6aaef0{/SQL/execution,null,AVAILABLE,@Spark}
22/01/29 23:42:54 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@13199f0{/SQL/execution/json,null,AVAILABLE,@Spark}
22/01/29 23:42:54 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@26636bfd{/static/sql,null,AVAILABLE,@Spark}
22/01/29 23:42:54 INFO state.StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
22/01/29 23:42:55 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
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 ~]$ 
```

```
hdfs dfs -rm -r -f checkpoints
hdfs dfs -du -h checkpionts/my_parquet_checkpoint
```



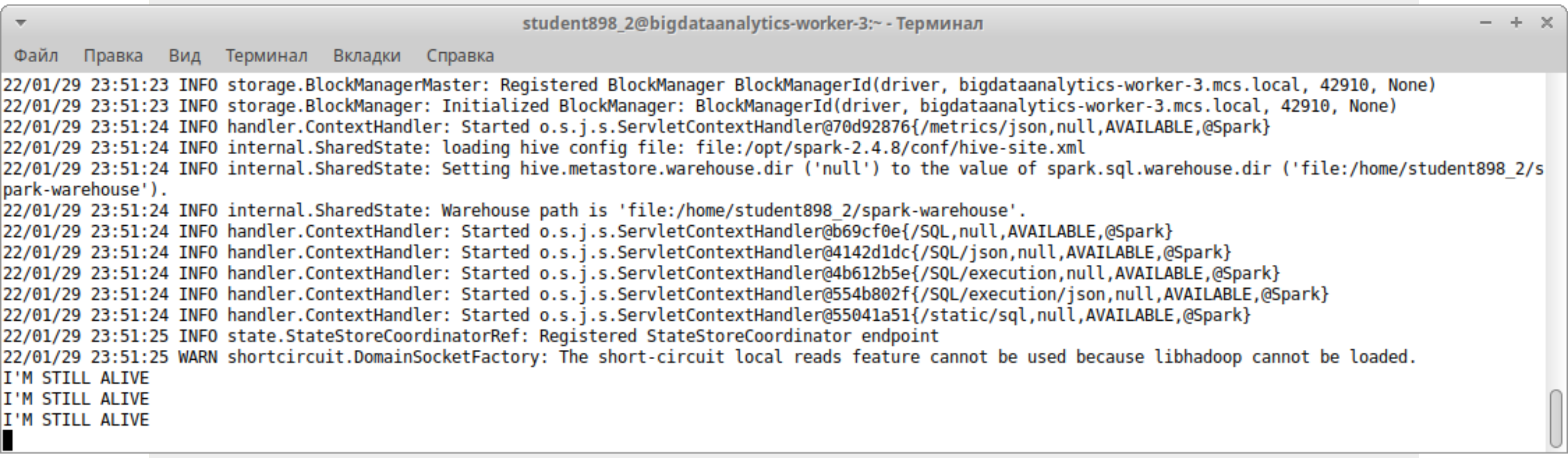
```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h checkpionts
1.2 K  2.5 K  checkpionts/my_parquet_checkpoint
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -du -h checkpionts/my_parquet_checkpoint
29   58     checkpionts/my_parquet_checkpoint/commits
45   90     checkpionts/my_parquet_checkpoint/metadata
422  844    checkpionts/my_parquet_checkpoint/offsets
761  1.5 K  checkpionts/my_parquet_checkpoint/sources
[student898_2@bigdataanalytics-worker-3 ~]$ 
```

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

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



```
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -rm -r -f checkpoints
[student898_2@bigdataanalytics-worker-3 ~]$ hdfs dfs -ls
Found 11 items
drwx------   - student898_2 student898_2          0 2022-01-29 23:12 .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 2022-01-29 23:35 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 my_submit_parquet_files
22/01/29 23:49:07 INFO fs.TrashPolicyDefault: Moved: 'hdfs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/my_submit_parquet_files' to trash at: hd
fs://bigdataanalytics-head-0.mcs.local:8020/user/student898_2/.Trash/Current/user/student898_2/my_submit_parquet_files
[student898_2@bigdataanalytics-worker-3 ~]$ 
```

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



```
22/01/29 23:51:23 INFO storage.BlockManagerMaster: Registered BlockManager BlockManagerId(driver, bigdataanalytics-worker-3.mcs.local, 42910, None)
22/01/29 23:51:23 INFO storage.BlockManager: Initialized BlockManager: BlockManagerId(driver, bigdataanalytics-worker-3.mcs.local, 42910, None)
22/01/29 23:51:24 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@70d92876{/metrics/json,null,AVAILABLE,@Spark}
22/01/29 23:51:24 INFO internal.SharedState: loading hive config file: file:/opt/spark-2.4.8/conf/hive-site.xml
22/01/29 23:51:24 INFO internal.SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('file:/home/student898_2/s
park-warehouse').
22/01/29 23:51:24 INFO internal.SharedState: Warehouse path is 'file:/home/student898_2/spark-warehouse'.
22/01/29 23:51:24 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@b69cf0e{/SQL,null,AVAILABLE,@Spark}
22/01/29 23:51:24 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4142d1dc{/SQL/json,null,AVAILABLE,@Spark}
22/01/29 23:51:24 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@4b612b5e{/SQL/execution,null,AVAILABLE,@Spark}
22/01/29 23:51:24 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@554b802f{/SQL/execution/json,null,AVAILABLE,@Spark}
22/01/29 23:51:24 INFO handler.ContextHandler: Started o.s.j.s.ServletContextHandler@55041a51{/static/sql,null,AVAILABLE,@Spark}
22/01/29 23:51:25 INFO state.StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
22/01/29 23:51:25 WARN shortcircuit.DomainSocketFactory: The short-circuit local reads feature cannot be used because libhadoop cannot be loaded.
I'M STILL ALIVE
I'M STILL ALIVE
I'M STILL ALIVE
```