ETL: автоматизация подготовки данных

Урок 6. Обзор возможностей Airflow, установка и настройка

Задание

Установить спарк как показано на семинаре:

- Для этого переместите папку spark в home.

- Дайте права командой chmod -R 777 ./

- nano ~/.bashrc

- export SPARK_HOME=/home/spark && export

PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin

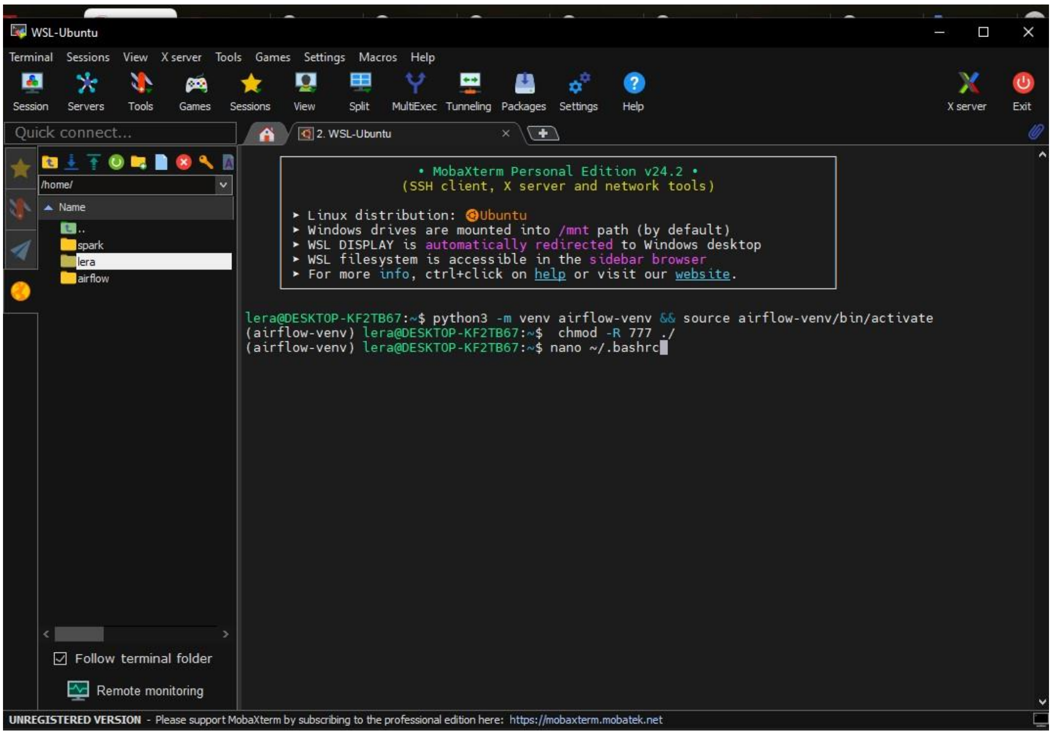
- source ~/.bashrc

- sudo apt-get install openjdk-8-jdkpip

- Указанные библиотеки нужно также установить и в виртуальную среду: python3-m venv airflow venv && source airflow venv /bin/activate

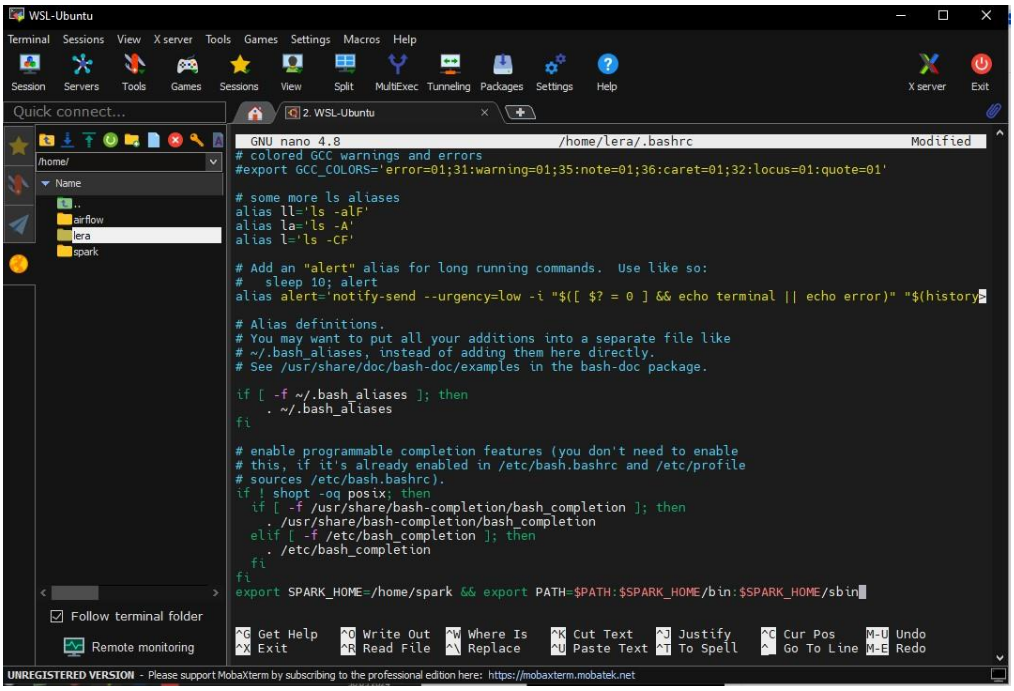
- pip install pyspark==3.2.4 && pip install pandas==1.5.3 && pip install SQLAlchemy==1.4.46

Используйте ДЗ которые вы мне высылали для 3-4 семинара. Запустите данные задачи ПОСЛЕДОВАТЕЛЬНО, одну за другой в аирфлоу. Пришлите мне скриншоты выполненных задач в аирфлоу, логов аирфлоу, скриншоты что у вас записались таблицы в БД mysql на WSL. По возможности доработайте код чтобы изображение с линии платежей генерировалось в указанную директорию. Скриншоты соберите в pdf.



```
           • MobaXterm Personal Edition v24.2 •
         (SSH client, X server and network tools)

 ▸ Linux distribution: Ubuntu
 ▸ Windows drives are mounted into /mnt path (by default)
 ▸ WSL DISPLAY is automatically redirected to Windows desktop
 ▸ WSL filesystem is accessible in the sidebar browser
 ▸ For more info, ctrl+click on help or visit our website.

lera@DESKTOP-KF2TB67:~$ python3 -m venv airflow-venv && source airflow-venv/bin/activate
(airflow-venv) lera@DESKTOP-KF2TB67:~$  chmod -R 777 ./
(airflow-venv) lera@DESKTOP-KF2TB67:~$ nano ~/.bashrc
```

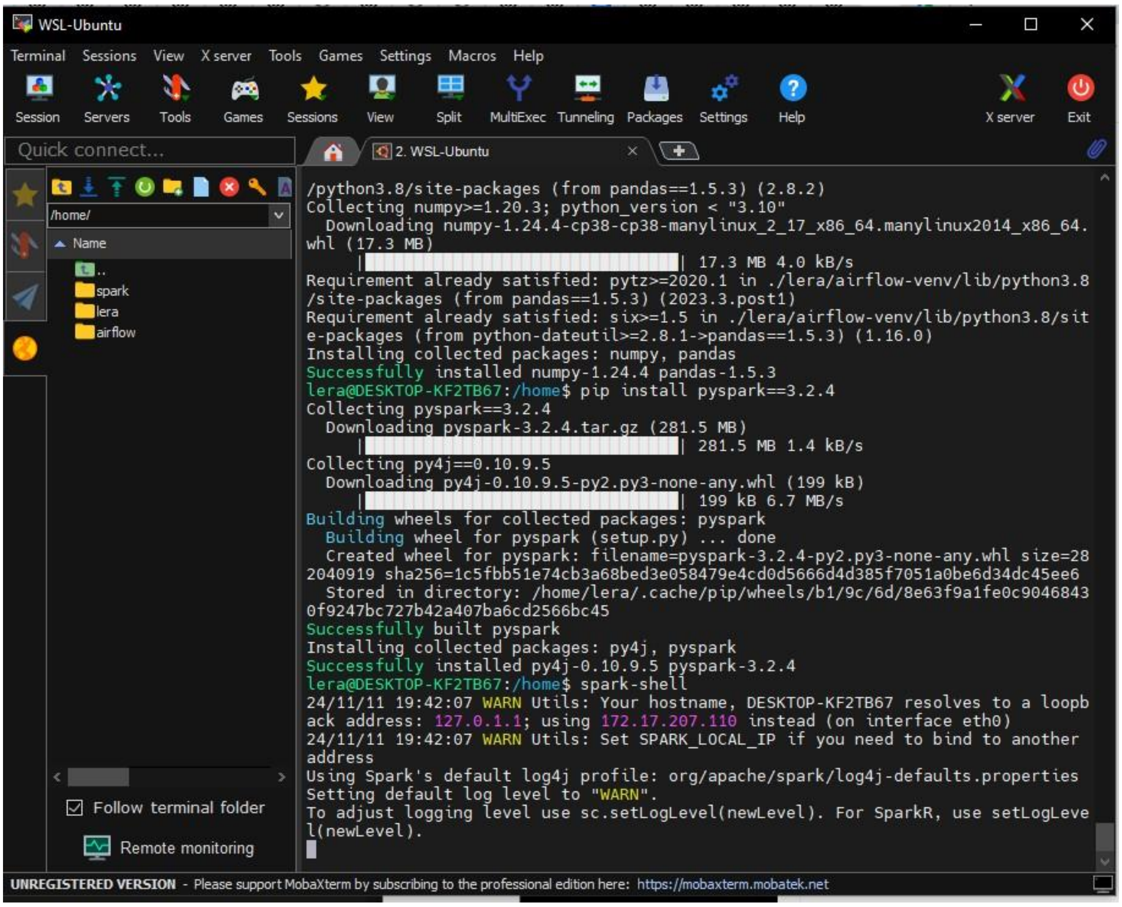
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net


WSL-Ubuntu
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
2. WSL-Ubuntu
/home/
Name
..
airflow
lera
spark
☑ Follow terminal folder
Remote monitoring
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

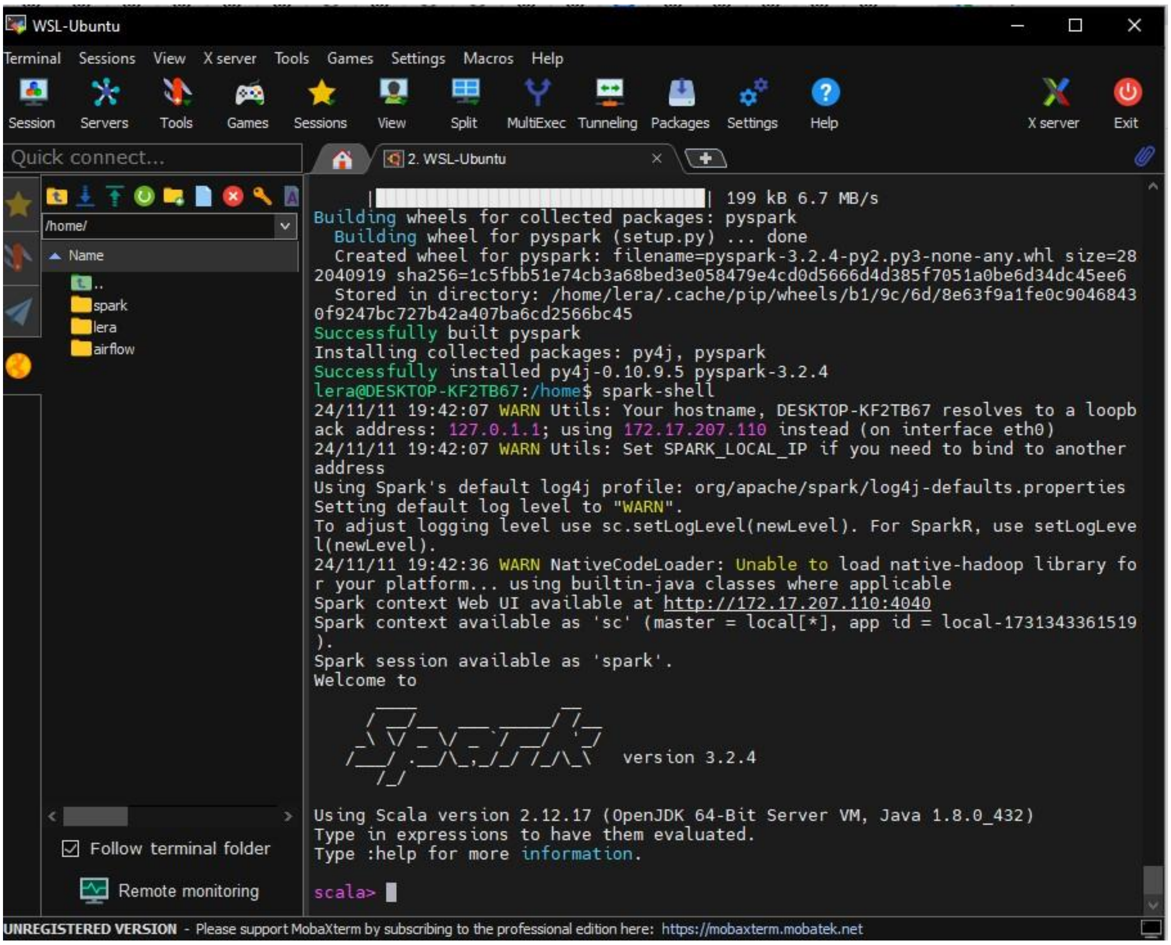
```
  GNU nano 4.8                        /home/lera/.bashrc                         Modified
# colored GCC warnings and errors
#export GCC_COLORS='error=01;31:warning=01;35:note=01;36:caret=01;32:locus=01:quote=01'

# some more ls aliases
alias ll='ls -alF'
alias la='ls -A'
alias l='ls -CF'

# Add an "alert" alias for long running commands.  Use like so:
#   sleep 10; alert
alias alert='notify-send --urgency=low -i "$([ $? = 0 ] && echo terminal || echo error)" "$(history>

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export SPARK_HOME=/home/spark && export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin

^G Get Help   ^O Write Out  ^W Where Is   ^K Cut Text   ^J Justify    ^C Cur Pos     M-U Undo
^X Exit       ^R Read File  ^\ Replace    ^U Paste Text ^T To Spell   ^_ Go To Line  M-E Redo
```

sudo apt-get install openjdk-8-jdkpip (не сработало, чтобы запустилось пришлось делать так:

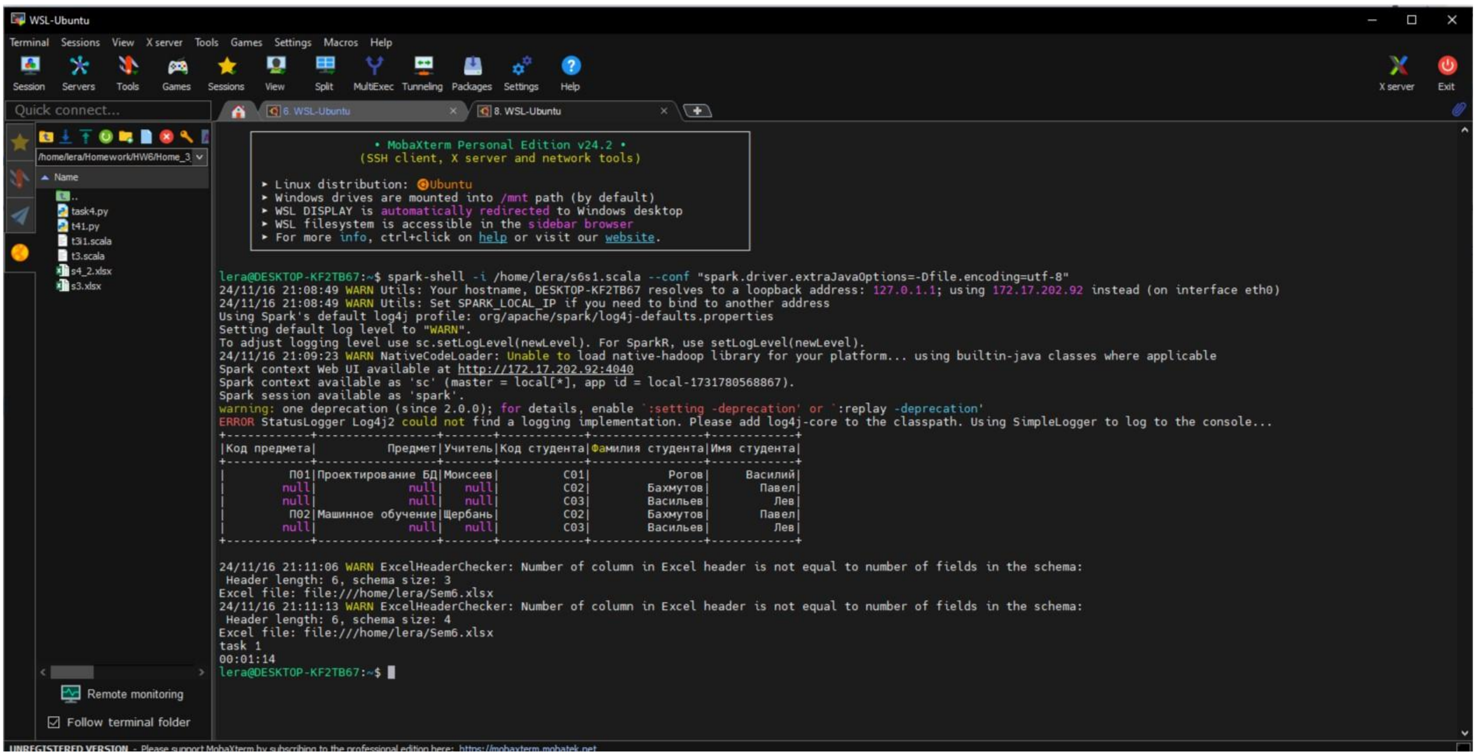
```
sudo add-apt-repository ppa:openjdk-r/ppa

sudo apt-get update

sudo apt-get install openjdk-8-jdk)
```


WSL-Ubuntu
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
2. WSL-Ubuntu
/home/
Name
..
spark
lera
airflow
☑ Follow terminal folder
Remote monitoring
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

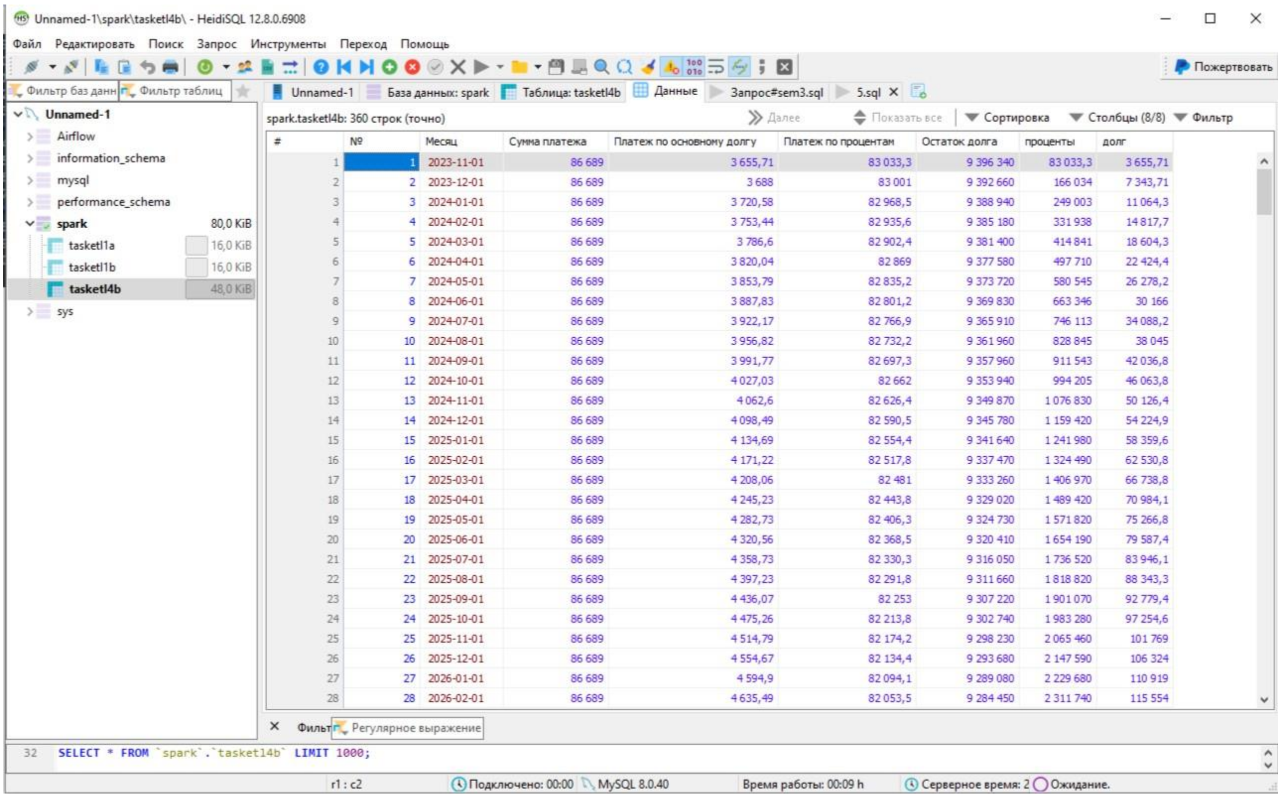
```
/python3.8/site-packages (from pandas==1.5.3) (2.8.2)
Collecting numpy>=1.20.3; python_version < "3.10"
  Downloading numpy-1.24.4-cp38-cp38-manylinux_2_17_x86_64.manylinux2014_x86_64.
whl (17.3 MB)
     |████████████████████████████████| 17.3 MB 4.0 kB/s
Requirement already satisfied: pytz>=2020.1 in ./lera/airflow-venv/lib/python3.8
/site-packages (from pandas==1.5.3) (2023.3.post1)
Requirement already satisfied: six>=1.5 in ./lera/airflow-venv/lib/python3.8/sit
e-packages (from python-dateutil>=2.8.1->pandas==1.5.3) (1.16.0)
Installing collected packages: numpy, pandas
Successfully installed numpy-1.24.4 pandas-1.5.3
lera@DESKTOP-KF2TB67:/home$ pip install pyspark==3.2.4
Collecting pyspark==3.2.4
  Downloading pyspark-3.2.4.tar.gz (281.5 MB)
     |████████████████████████████████| 281.5 MB 1.4 kB/s
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
     |████████████████████████████████| 199 kB 6.7 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.4-py2.py3-none-any.whl size=28
2040919 sha256=1c5fbb51e74cb3a68bed3e058479e4cd0d5666d4d385f7051a0be6d34dc45ee6
  Stored in directory: /home/lera/.cache/pip/wheels/b1/9c/6d/8e63f9a1fe0c9046843
0f9247bc727b42a407ba6cd2566bc45
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.5 pyspark-3.2.4
lera@DESKTOP-KF2TB67:/home$ spark-shell
24/11/11 19:42:07 WARN Utils: Your hostname, DESKTOP-KF2TB67 resolves to a loopb
ack address: 127.0.1.1; using 172.17.207.110 instead (on interface eth0)
24/11/11 19:42:07 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
█
```


WSL-Ubuntu
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
2. WSL-Ubuntu
/home/
Name
..
spark
lera
airflow
Follow terminal folder
Remote monitoring
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

```
     |████████████████████████████████| 199 kB 6.7 MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.2.4-py2.py3-none-any.whl size=28
2040919 sha256=1c5fbb51e74cb3a68bed3e058479e4cd0d5666d4d385f7051a0be6d34dc45ee6
  Stored in directory: /home/lera/.cache/pip/wheels/b1/9c/6d/8e63f9a1fe0c9046843
0f9247bc727b42a407ba6cd2566bc45
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.5 pyspark-3.2.4
lera@DESKTOP-KF2TB67:/home$ spark-shell
24/11/11 19:42:07 WARN Utils: Your hostname, DESKTOP-KF2TB67 resolves to a loopb
ack address: 127.0.1.1; using 172.17.207.110 instead (on interface eth0)
24/11/11 19:42:07 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
24/11/11 19:42:36 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Spark context Web UI available at http://172.17.207.110:4040
Spark context available as 'sc' (master = local[*], app id = local-1731343361519
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.4
      /_/

Using Scala version 2.12.17 (OpenJDK 64-Bit Server VM, Java 1.8.0_432)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```


WSL-Ubuntu
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings Help X server Exit
Quick connect...
6. WSL-Ubuntu
8. WSL-Ubuntu
/home/lera/Homework/HW6/Home_3
Name
..
task4.py
t41.py
t31.scala
t3.scala
s4_2.xlsx
s3.xlsx
Remote monitoring
Follow terminal folder
UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

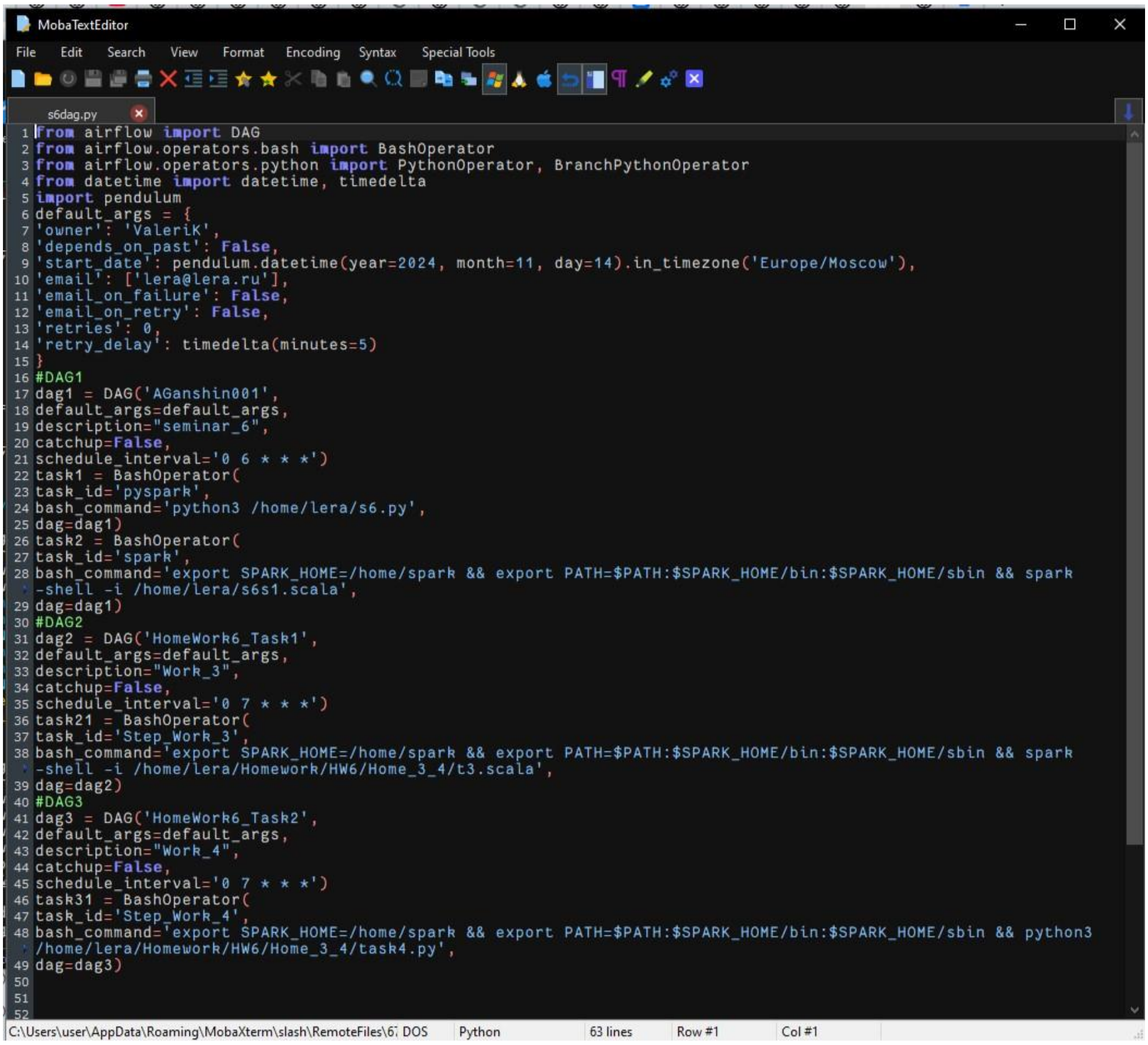
```
            • MobaXterm Personal Edition v24.2 •
          (SSH client, X server and network tools)

 ▸ Linux distribution: Ubuntu
 ▸ Windows drives are mounted into /mnt path (by default)
 ▸ WSL DISPLAY is automatically redirected to Windows desktop
 ▸ WSL filesystem is accessible in the sidebar browser
 ▸ For more info, ctrl+click on help or visit our website.

lera@DESKTOP-KF2TB67:~$ spark-shell -i /home/lera/s6s1.scala --conf "spark.driver.extraJavaOptions=-Dfile.encoding=utf-8"
24/11/16 21:08:49 WARN Utils: Your hostname, DESKTOP-KF2TB67 resolves to a loopback address: 127.0.1.1; using 172.17.202.92 instead (on interface eth0)
24/11/16 21:08:49 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/11/16 21:09:23 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://172.17.202.92:4040
Spark context available as 'sc' (master = local[*], app id = local-1731780568867).
Spark session available as 'spark'.
warning: one deprecation (since 2.0.0); for details, enable `:setting -deprecation' or `:replay -deprecation'
ERROR StatusLogger Log4j2 could not find a logging implementation. Please add log4j-core to the classpath. Using SimpleLogger to log to the console...
+------------+-----------------+-------+------------+----------------+------------+
|Код предмета|          Предмет|Учитель|Код студента|Фамилия студента|Имя студента|
+------------+-----------------+-------+------------+----------------+------------+
|         П01|Проектирование БД|Моисеев|         C01|           Рогов|     Василий|
|        null|             null|   null|         C02|        Бахмутов|       Павел|
|        null|             null|   null|         C03|        Васильев|         Лев|
|         П02|Машинное обучение|Щербань|         C02|        Бахмутов|       Павел|
|        null|             null|   null|         C03|        Васильев|         Лев|
+------------+-----------------+-------+------------+----------------+------------+

24/11/16 21:11:06 WARN ExcelHeaderChecker: Number of column in Excel header is not equal to number of fields in the schema:
 Header length: 6, schema size: 3
Excel file: file:///home/lera/Sem6.xlsx
24/11/16 21:11:13 WARN ExcelHeaderChecker: Number of column in Excel header is not equal to number of fields in the schema:
 Header length: 6, schema size: 4
Excel file: file:///home/lera/Sem6.xlsx
task 1
00:01:14
lera@DESKTOP-KF2TB67:~$
```



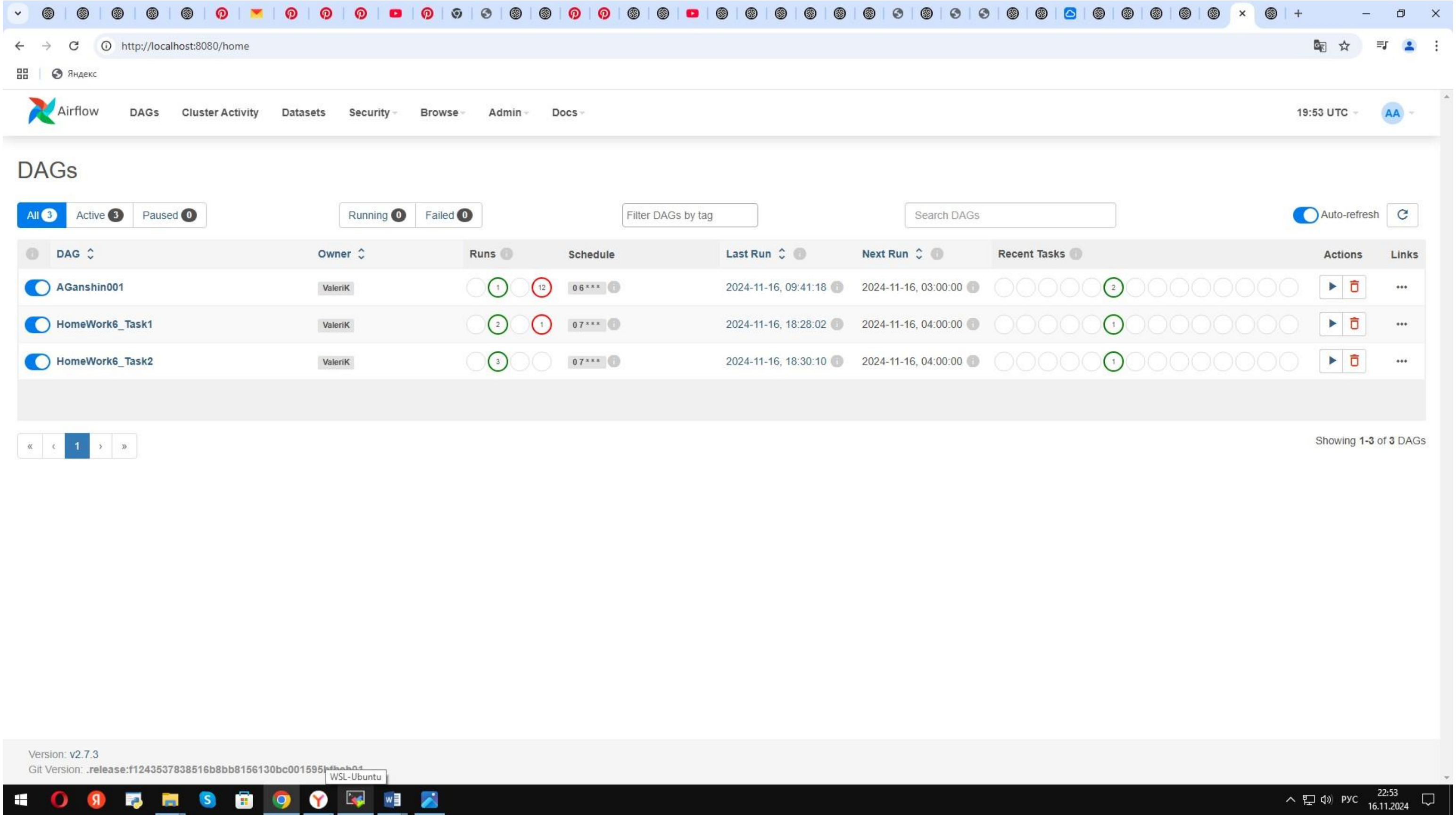
spark.tasketl4b: 360 строк (точно)

#	№	Месяц	Сумма платежа	Платеж по основному долгу	Платеж по процентам	Остаток долга	проценты	долг
1	1	2023-11-01	86 689	3 655,71	83 033,3	9 396 340	83 033,3	3 655,71
2	2	2023-12-01	86 689	3 688	83 001	9 392 660	166 034	7 343,71
3	3	2024-01-01	86 689	3 720,58	82 968,5	9 388 940	249 003	11 064,3
4	4	2024-02-01	86 689	3 753,44	82 935,6	9 385 180	331 938	14 817,7
5	5	2024-03-01	86 689	3 786,6	82 902,4	9 381 400	414 841	18 604,3
6	6	2024-04-01	86 689	3 820,04	82 869	9 377 580	497 710	22 424,4
7	7	2024-05-01	86 689	3 853,79	82 835,2	9 373 720	580 545	26 278,2
8	8	2024-06-01	86 689	3 887,83	82 801,2	9 369 830	663 346	30 166
9	9	2024-07-01	86 689	3 922,17	82 766,9	9 365 910	746 113	34 088,2
10	10	2024-08-01	86 689	3 956,82	82 732,2	9 361 960	828 845	38 045
11	11	2024-09-01	86 689	3 991,77	82 697,3	9 357 960	911 543	42 036,8
12	12	2024-10-01	86 689	4 027,03	82 662	9 353 940	994 205	46 063,8
13	13	2024-11-01	86 689	4 062,6	82 626,4	9 349 870	1 076 830	50 126,4
14	14	2024-12-01	86 689	4 098,49	82 590,5	9 345 780	1 159 420	54 224,9
15	15	2025-01-01	86 689	4 134,69	82 554,4	9 341 640	1 241 980	58 359,6
16	16	2025-02-01	86 689	4 171,22	82 517,8	9 337 470	1 324 490	62 530,8
17	17	2025-03-01	86 689	4 208,06	82 481	9 333 260	1 406 970	66 738,8
18	18	2025-04-01	86 689	4 245,23	82 443,8	9 329 020	1 489 420	70 984,1
19	19	2025-05-01	86 689	4 282,73	82 406,3	9 324 730	1 571 820	75 266,8
20	20	2025-06-01	86 689	4 320,56	82 368,5	9 320 410	1 654 190	79 587,4
21	21	2025-07-01	86 689	4 358,73	82 330,3	9 316 050	1 736 520	83 946,1
22	22	2025-08-01	86 689	4 397,23	82 291,8	9 311 660	1 818 820	88 343,3
23	23	2025-09-01	86 689	4 436,07	82 253	9 307 220	1 901 070	92 779,4
24	24	2025-10-01	86 689	4 475,26	82 213,8	9 302 740	1 983 280	97 254,6
25	25	2025-11-01	86 689	4 514,79	82 174,2	9 298 230	2 065 460	101 769
26	26	2025-12-01	86 689	4 554,67	82 134,4	9 293 680	2 147 590	106 324
27	27	2026-01-01	86 689	4 594,9	82 094,1	9 289 080	2 229 680	110 919
28	28	2026-02-01	86 689	4 635,49	82 053,5	9 284 450	2 311 740	115 554

```
32  SELECT * FROM `spark`.`tasketl4b` LIMIT 1000;
```


MobaTextEditor
File Edit Search View Format Encoding Syntax Special Tools
s6dag.py
C:\Users\user\AppData\Roaming\MobaXterm\slash\RemoteFiles\67 DOS
Python
63 lines
Row #1
Col #1

```python
from airflow import DAG
from airflow.operators.bash import BashOperator
from airflow.operators.python import PythonOperator, BranchPythonOperator
from datetime import datetime, timedelta
import pendulum
default_args = {
'owner': 'ValeriK',
'depends_on_past': False,
'start_date': pendulum.datetime(year=2024, month=11, day=14).in_timezone('Europe/Moscow'),
'email': ['lera@lera.ru'],
'email_on_failure': False,
'email_on_retry': False,
'retries': 0,
'retry_delay': timedelta(minutes=5)
}
#DAG1
dag1 = DAG('AGanshin001',
default_args=default_args,
description="seminar_6",
catchup=False,
schedule_interval='0 6 * * *')
task1 = BashOperator(
task_id='pyspark',
bash_command='python3 /home/lera/s6.py',
dag=dag1)
task2 = BashOperator(
task_id='spark',
bash_command='export SPARK_HOME=/home/spark && export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin && spark-shell -i /home/lera/s6s1.scala',
dag=dag1)
#DAG2
dag2 = DAG('HomeWork6_Task1',
default_args=default_args,
description="Work_3",
catchup=False,
schedule_interval='0 7 * * *')
task21 = BashOperator(
task_id='Step_Work_3',
bash_command='export SPARK_HOME=/home/spark && export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin && spark-shell -i /home/lera/Homework/HW6/Home_3_4/t3.scala',
dag=dag2)
#DAG3
dag3 = DAG('HomeWork6_Task2',
default_args=default_args,
description="Work_4",
catchup=False,
schedule_interval='0 7 * * *')
task31 = BashOperator(
task_id='Step_Work_4',
bash_command='export SPARK_HOME=/home/spark && export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin && python3 /home/lera/Homework/HW6/Home_3_4/task4.py',
dag=dag3)



```

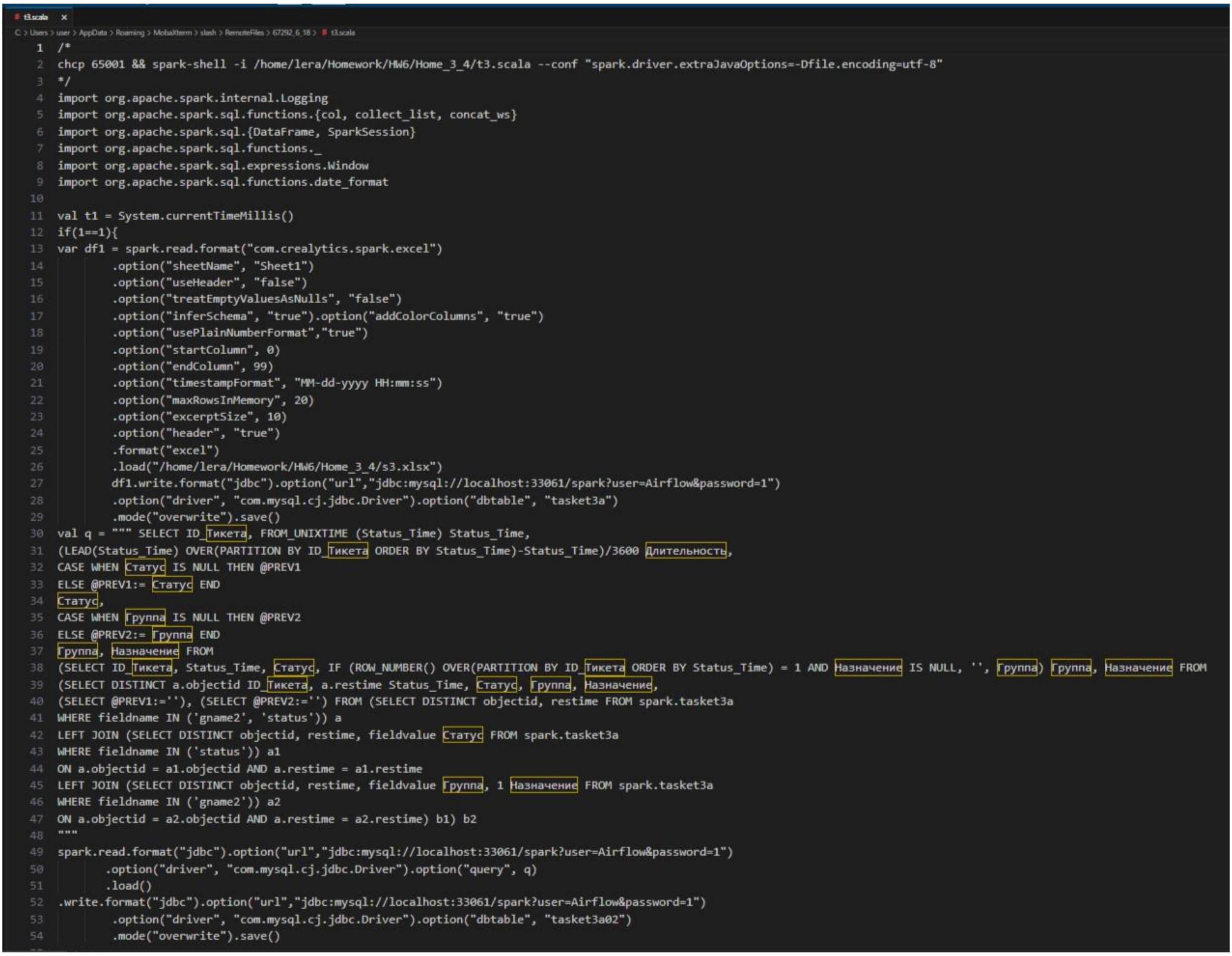


DAGs

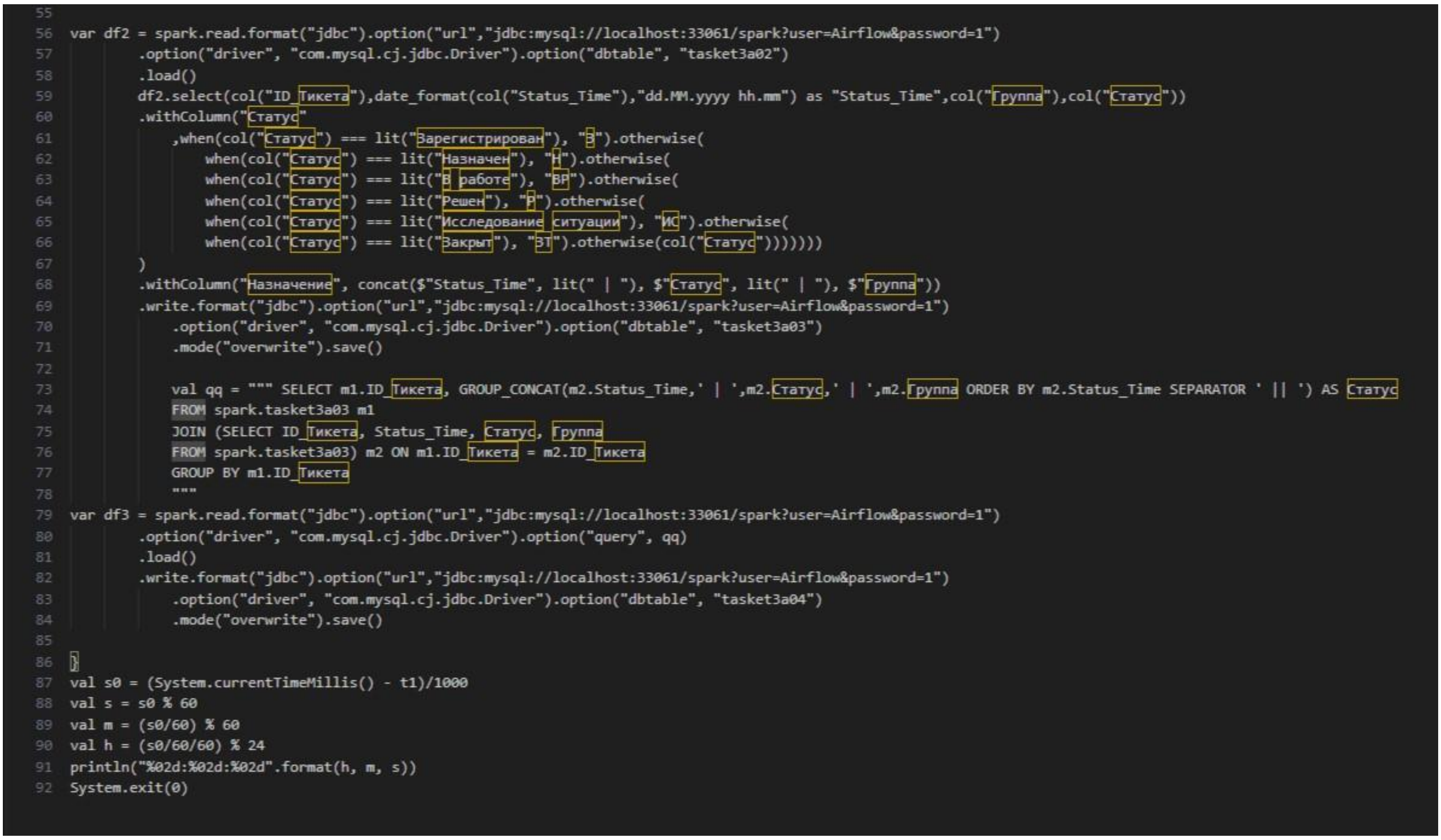
DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
AGanshin001	ValeriK	1 12	0 6 * * *	2024-11-16, 09:41:18	2024-11-16, 03:00:00	2	▶ 🗑	···
HomeWork6_Task1	ValeriK	2 1	0 7 * * *	2024-11-16, 18:28:02	2024-11-16, 04:00:00	1	▶ 🗑	···
HomeWork6_Task2	ValeriK	3	0 7 * * *	2024-11-16, 18:30:10	2024-11-16, 04:00:00	1	▶ 🗑	···

Showing **1-3** of **3** DAGs

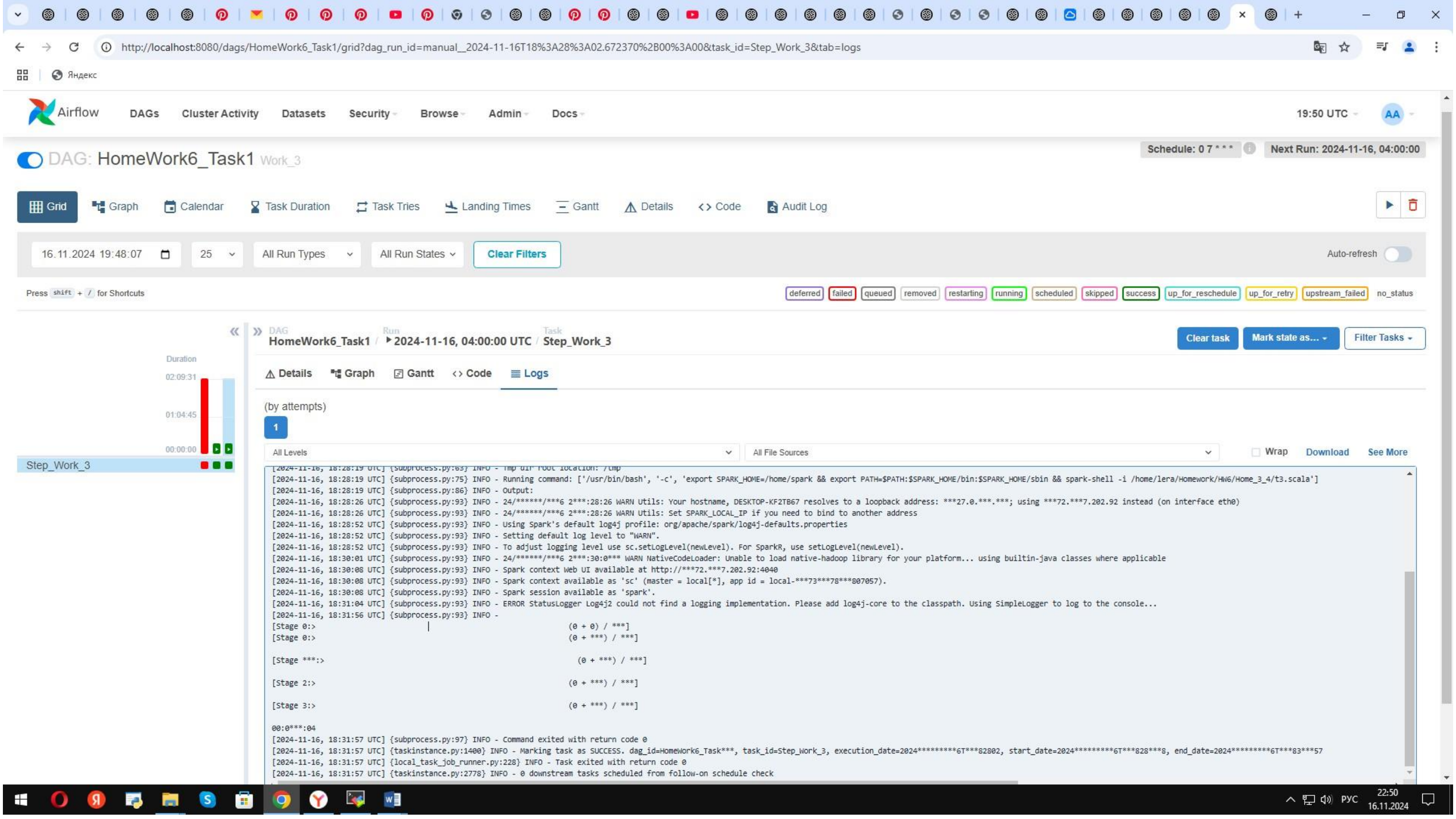
Version: v2.7.3
Git Version: .release:f1243537838516b8bb8156130bc001595


t3.scala
C > Users > user > AppData > Roaming > MobaXterm > slash > RemoteFiles > 67292_6_18 > t3.scala

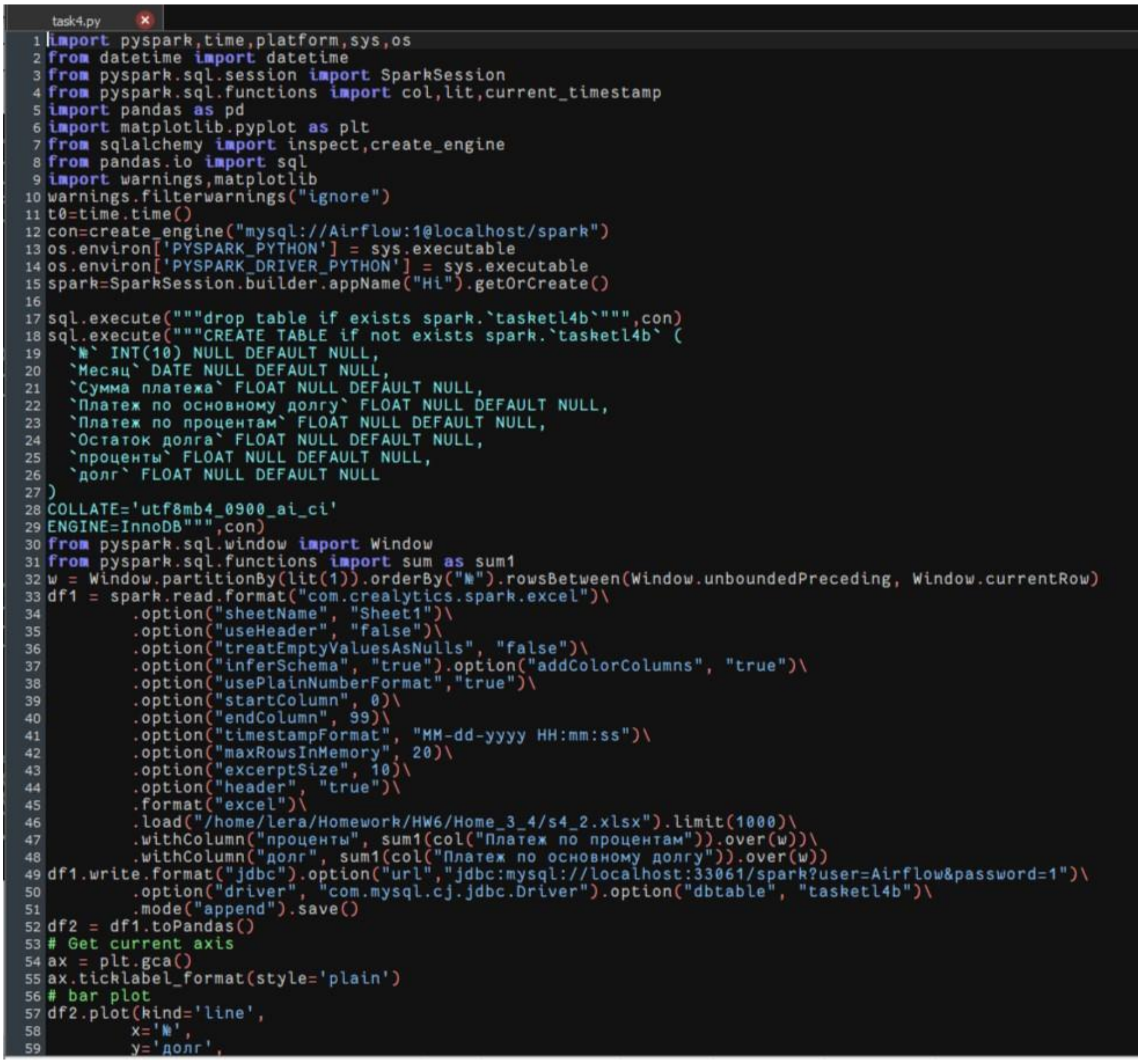
```
/*
chcp 65001 && spark-shell -i /home/lera/Homework/HW6/Home_3_4/t3.scala --conf "spark.driver.extraJavaOptions=-Dfile.encoding=utf-8"
*/
import org.apache.spark.internal.Logging
import org.apache.spark.sql.functions.{col, collect_list, concat_ws}
import org.apache.spark.sql.{DataFrame, SparkSession}
import org.apache.spark.sql.functions._
import org.apache.spark.sql.expressions.Window
import org.apache.spark.sql.functions.date_format

val t1 = System.currentTimeMillis()
if(1==1){
var df1 = spark.read.format("com.crealytics.spark.excel")
        .option("sheetName", "Sheet1")
        .option("useHeader", "false")
        .option("treatEmptyValuesAsNulls", "false")
        .option("inferSchema", "true").option("addColorColumns", "true")
        .option("usePlainNumberFormat","true")
        .option("startColumn", 0)
        .option("endColumn", 99)
        .option("timestampFormat", "MM-dd-yyyy HH:mm:ss")
        .option("maxRowsInMemory", 20)
        .option("excerptSize", 10)
        .option("header", "true")
        .format("excel")
        .load("/home/lera/Homework/HW6/Home_3_4/s3.xlsx")
        df1.write.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")
        .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasket3a")
        .mode("overwrite").save()
val q = """ SELECT ID_Тикета, FROM_UNIXTIME (Status_Time) Status_Time,
(LEAD(Status_Time) OVER(PARTITION BY ID_Тикета ORDER BY Status_Time)-Status_Time)/3600 Длительность,
CASE WHEN Статус IS NULL THEN @PREV1
ELSE @PREV1:= Статус END
Статус,
CASE WHEN Группа IS NULL THEN @PREV2
ELSE @PREV2:= Группа END
Группа, Назначение FROM
(SELECT ID_Тикета, Status_Time, Статус, IF (ROW_NUMBER() OVER(PARTITION BY ID_Тикета ORDER BY Status_Time) = 1 AND Назначение IS NULL, '', Группа) Группа, Назначение FROM
(SELECT DISTINCT a.objectid ID_Тикета, a.restime Status_Time, Статус, Группа, Назначение,
(SELECT @PREV1:=''), (SELECT @PREV2:='') FROM (SELECT DISTINCT objectid, restime FROM spark.tasket3a
WHERE fieldname IN ('gname2', 'status')) a
LEFT JOIN (SELECT DISTINCT objectid, restime, fieldvalue Статус FROM spark.tasket3a
WHERE fieldname IN ('status')) a1
ON a.objectid = a1.objectid AND a.restime = a1.restime
LEFT JOIN (SELECT DISTINCT objectid, restime, fieldvalue Группа, 1 Назначение FROM spark.tasket3a
WHERE fieldname IN ('gname2')) a2
ON a.objectid = a2.objectid AND a.restime = a2.restime) b1) b2
"""
spark.read.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")
       .option("driver", "com.mysql.cj.jdbc.Driver").option("query", q)
       .load()
.write.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")
        .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasket3a02")
        .mode("overwrite").save()
```



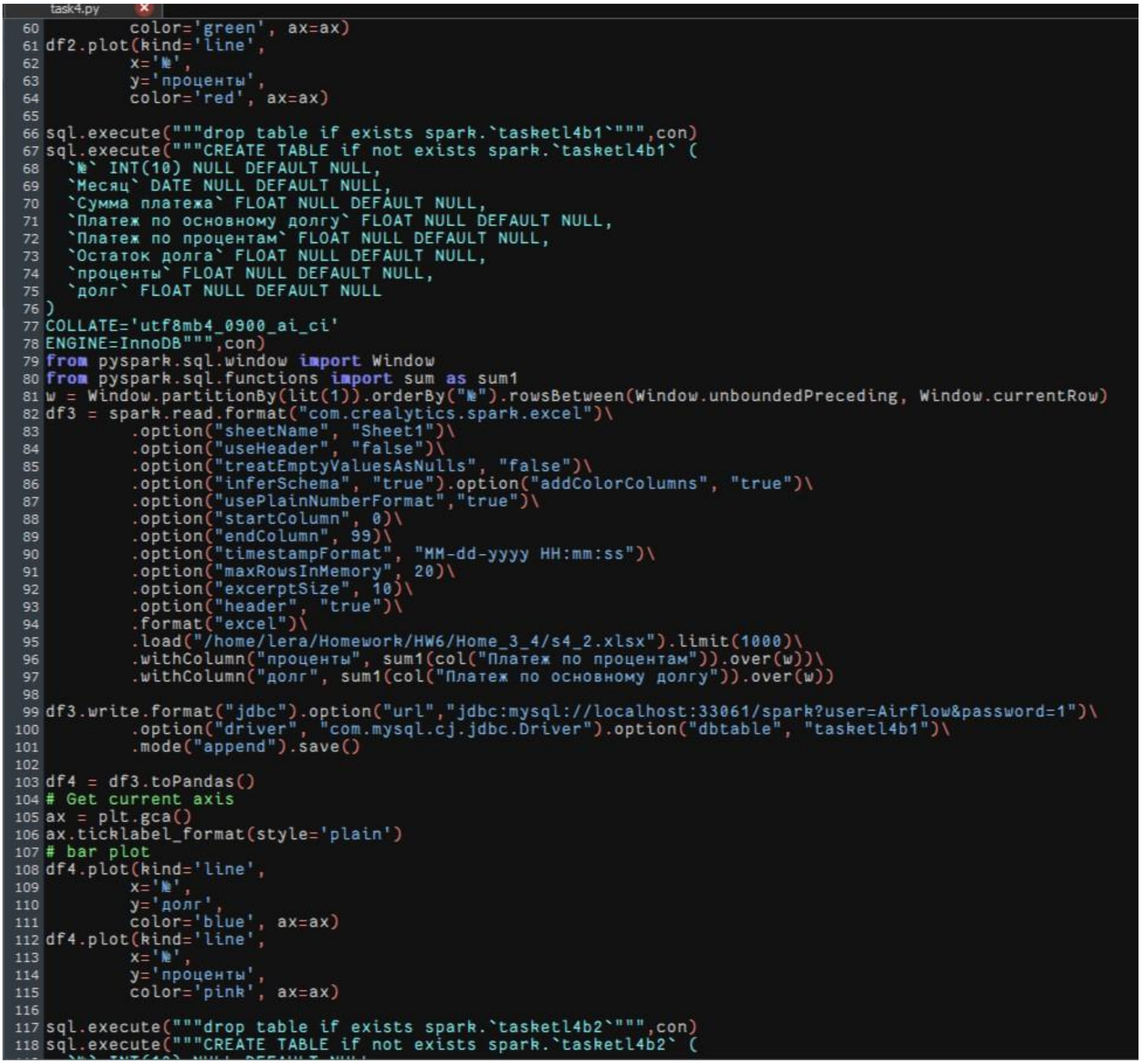
```scala

var df2 = spark.read.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")
        .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasket3a02")
        .load()
        df2.select(col("ID_Тикета"),date_format(col("Status_Time"),"dd.MM.yyyy hh.mm") as "Status_Time",col("Группа"),col("Статус"))
        .withColumn("Статус"
            ,when(col("Статус") === lit("Зарегистрирован"), "З").otherwise(
                when(col("Статус") === lit("Назначен"), "Н").otherwise(
                when(col("Статус") === lit("В работе"), "ВР").otherwise(
                when(col("Статус") === lit("Решен"), "Р").otherwise(
                when(col("Статус") === lit("Исследование ситуации"), "ИС").otherwise(
                when(col("Статус") === lit("Закрыт"), "ЗТ").otherwise(col("Статус")))))))
        )
        .withColumn("Назначение", concat($"Status_Time", lit(" | "), $"Статус", lit(" | "), $"Группа"))
        .write.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")
            .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasket3a03")
            .mode("overwrite").save()

            val qq = """ SELECT m1.ID_Тикета, GROUP_CONCAT(m2.Status_Time,' | ',m2.Статус,' | ',m2.Группа ORDER BY m2.Status_Time SEPARATOR ' || ') AS Статус
            FROM spark.tasket3a03 m1
            JOIN (SELECT ID_Тикета, Status_Time, Статус, Группа
            FROM spark.tasket3a03) m2 ON m1.ID_Тикета = m2.ID_Тикета
            GROUP BY m1.ID_Тикета
            """
var df3 = spark.read.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")
        .option("driver", "com.mysql.cj.jdbc.Driver").option("query", qq)
        .load()
        .write.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")
            .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasket3a04")
            .mode("overwrite").save()

}
val s0 = (System.currentTimeMillis() - t1)/1000
val s = s0 % 60
val m = (s0/60) % 60
val h = (s0/60/60) % 24
println("%02d:%02d:%02d".format(h, m, s))
System.exit(0)
```



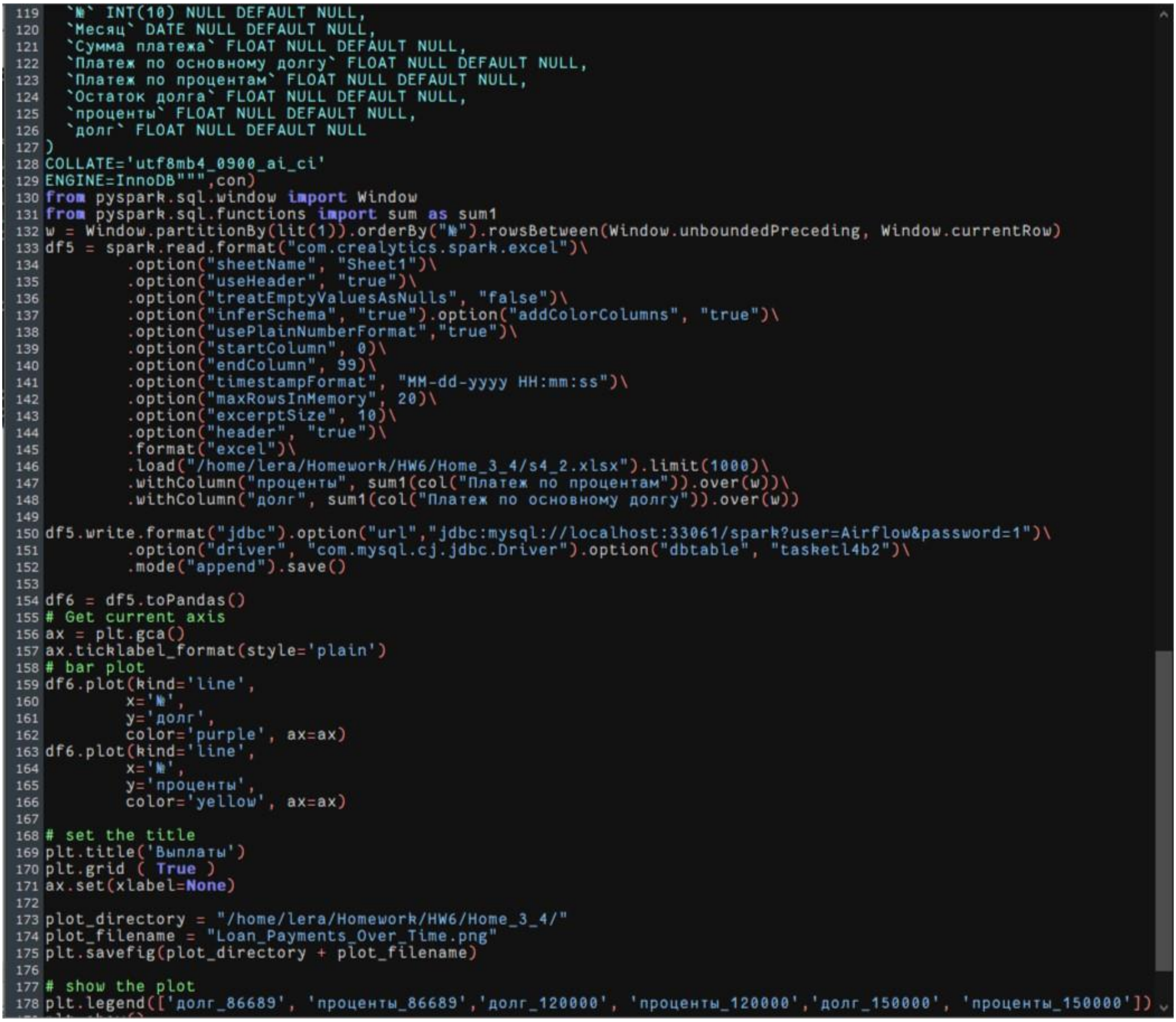
```
[2024-11-16, 18:28:19 UTC] {subprocess.py:63} INFO - Tmp dir root location: /tmp
[2024-11-16, 18:28:19 UTC] {subprocess.py:75} INFO - Running command: ['/usr/bin/bash', '-c', 'export SPARK_HOME=/home/spark && export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin && spark-shell -i /home/lera/Homework/HW6/Home_3_4/t3.scala']
[2024-11-16, 18:28:19 UTC] {subprocess.py:86} INFO - Output:
[2024-11-16, 18:28:26 UTC] {subprocess.py:93} INFO - 24/******/***6 2***:28:26 WARN Utils: Your hostname, DESKTOP-KF2TB67 resolves to a loopback address: ***27.0.***.***; using ***72.***7.202.92 instead (on interface eth0)
[2024-11-16, 18:28:26 UTC] {subprocess.py:93} INFO - 24/******/***6 2***:28:26 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
[2024-11-16, 18:28:52 UTC] {subprocess.py:93} INFO - Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
[2024-11-16, 18:28:52 UTC] {subprocess.py:93} INFO - Setting default log level to "WARN".
[2024-11-16, 18:28:52 UTC] {subprocess.py:93} INFO - To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
[2024-11-16, 18:30:01 UTC] {subprocess.py:93} INFO - 24/******/***6 2***:30:0*** WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[2024-11-16, 18:30:08 UTC] {subprocess.py:93} INFO - Spark context Web UI available at http://***72.***7.202.92:4040
[2024-11-16, 18:30:08 UTC] {subprocess.py:93} INFO - Spark context available as 'sc' (master = local[*], app id = local-***73***78***807057).
[2024-11-16, 18:30:08 UTC] {subprocess.py:93} INFO - Spark session available as 'spark'.
[2024-11-16, 18:31:04 UTC] {subprocess.py:93} INFO - ERROR StatusLogger Log4j2 could not find a logging implementation. Please add log4j-core to the classpath. Using SimpleLogger to log to the console...
[2024-11-16, 18:31:56 UTC] {subprocess.py:93} INFO -
[Stage 0:>                                                          (0 + 0) / ***]
[Stage 0:>                                                          (0 + ***) / ***]

[Stage ***:>                                                          (0 + ***) / ***]

[Stage 2:>                                                          (0 + ***) / ***]

[Stage 3:>                                                          (0 + ***) / ***]

00:0***:04
[2024-11-16, 18:31:57 UTC] {subprocess.py:97} INFO - Command exited with return code 0
[2024-11-16, 18:31:57 UTC] {taskinstance.py:1400} INFO - Marking task as SUCCESS. dag_id=HomeWork6_Task***, task_id=Step_Work_3, execution_date=2024*********6T***82802, start_date=2024*********6T***828***8, end_date=2024*********6T***83***57
[2024-11-16, 18:31:57 UTC] {local_task_job_runner.py:228} INFO - Task exited with return code 0
[2024-11-16, 18:31:57 UTC] {taskinstance.py:2778} INFO - 0 downstream tasks scheduled from follow-on schedule check
```



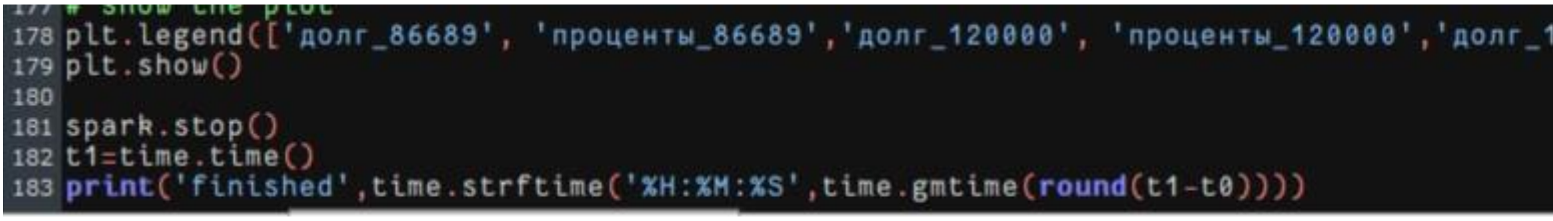
```
import pyspark,time,platform,sys,os
from datetime import datetime
from pyspark.sql.session import SparkSession
from pyspark.sql.functions import col,lit,current_timestamp
import pandas as pd
import matplotlib.pyplot as plt
from sqlalchemy import inspect,create_engine
from pandas.io import sql
import warnings,matplotlib
warnings.filterwarnings("ignore")
t0=time.time()
con=create_engine("mysql://Airflow:1@localhost/spark")
os.environ['PYSPARK_PYTHON'] = sys.executable
os.environ['PYSPARK_DRIVER_PYTHON'] = sys.executable
spark=SparkSession.builder.appName("Hi").getOrCreate()

sql.execute("""drop table if exists spark.`tasketl4b`""",con)
sql.execute("""CREATE TABLE if not exists spark.`tasketl4b` (
  `№` INT(10) NULL DEFAULT NULL,
  `Месяц` DATE NULL DEFAULT NULL,
  `Сумма платежа` FLOAT NULL DEFAULT NULL,
  `Платеж по основному долгу` FLOAT NULL DEFAULT NULL,
  `Платеж по процентам` FLOAT NULL DEFAULT NULL,
  `Остаток долга` FLOAT NULL DEFAULT NULL,
  `проценты` FLOAT NULL DEFAULT NULL,
  `долг` FLOAT NULL DEFAULT NULL
)
COLLATE='utf8mb4_0900_ai_ci'
ENGINE=InnoDB""",con)
from pyspark.sql.window import Window
from pyspark.sql.functions import sum as sum1
w = Window.partitionBy(lit(1)).orderBy("№").rowsBetween(Window.unboundedPreceding, Window.currentRow)
df1 = spark.read.format("com.crealytics.spark.excel")\
        .option("sheetName", "Sheet1")\
        .option("useHeader", "false")\
        .option("treatEmptyValuesAsNulls", "false")\
        .option("inferSchema", "true").option("addColorColumns", "true")\
        .option("usePlainNumberFormat","true")\
        .option("startColumn", 0)\
        .option("endColumn", 99)\
        .option("timestampFormat", "MM-dd-yyyy HH:mm:ss")\
        .option("maxRowsInMemory", 20)\
        .option("excerptSize", 10)\
        .option("header", "true")\
        .format("excel")\
        .load("/home/lera/Homework/HW6/Home_3_4/s4_2.xlsx").limit(1000)\
        .withColumn("проценты", sum1(col("Платеж по процентам")).over(w))\
        .withColumn("долг", sum1(col("Платеж по основному долгу")).over(w))
df1.write.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")\
        .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasketl4b")\
        .mode("append").save()
df2 = df1.toPandas()
# Get current axis
ax = plt.gca()
ax.ticklabel_format(style='plain')
# bar plot
df2.plot(kind='line',
        x='№',
        y='долг',
```



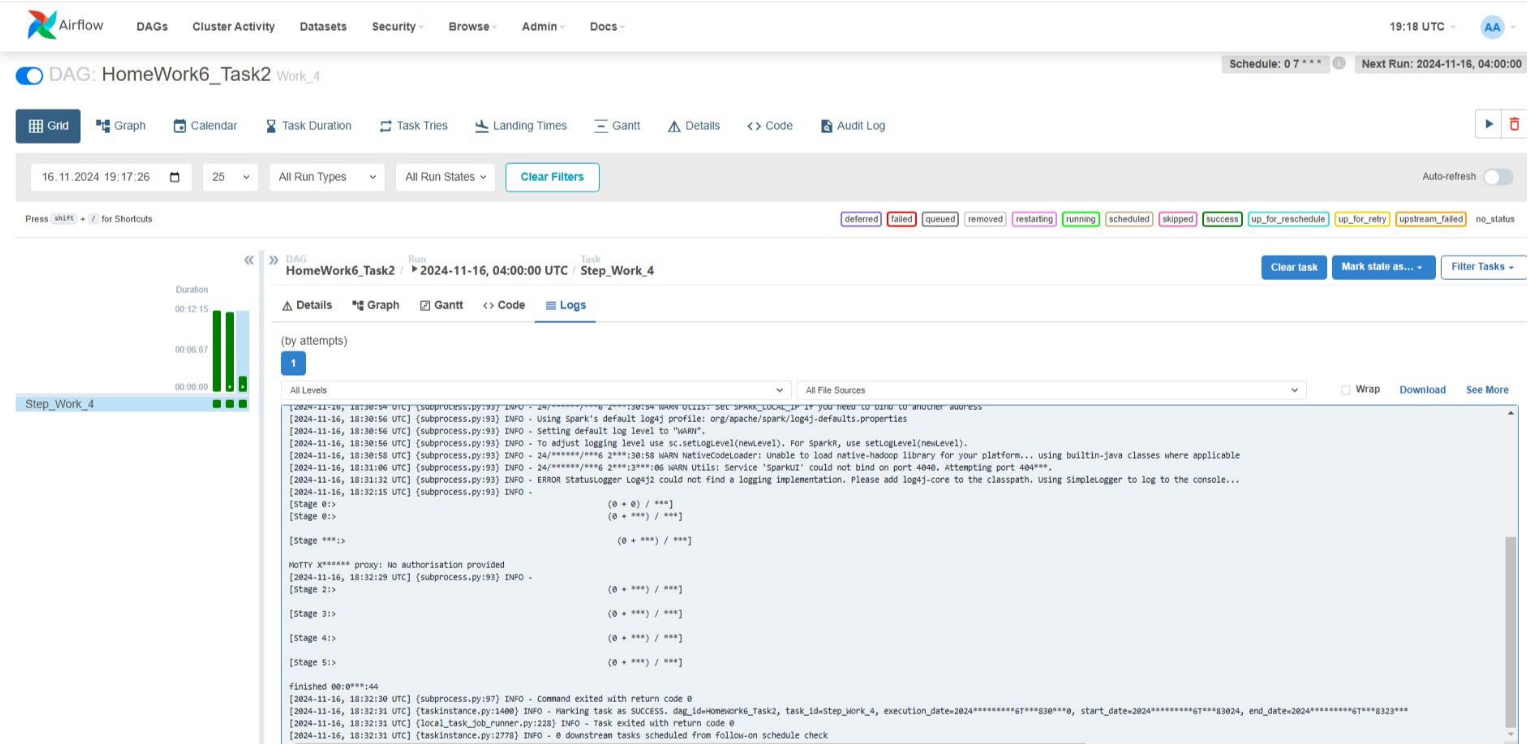
```
        color='green', ax=ax)
df2.plot(kind='line',
        x='№',
        y='проценты',
        color='red', ax=ax)

sql.execute("""drop table if exists spark.`tasketl4b1`""",con)
sql.execute("""CREATE TABLE if not exists spark.`tasketl4b1` (
  `№` INT(10) NULL DEFAULT NULL,
  `Месяц` DATE NULL DEFAULT NULL,
  `Сумма платежа` FLOAT NULL DEFAULT NULL,
  `Платеж по основному долгу` FLOAT NULL DEFAULT NULL,
  `Платеж по процентам` FLOAT NULL DEFAULT NULL,
  `Остаток долга` FLOAT NULL DEFAULT NULL,
  `проценты` FLOAT NULL DEFAULT NULL,
  `долг` FLOAT NULL DEFAULT NULL
)
COLLATE='utf8mb4_0900_ai_ci'
ENGINE=InnoDB""",con)
from pyspark.sql.window import Window
from pyspark.sql.functions import sum as sum1
w = Window.partitionBy(lit(1)).orderBy("№").rowsBetween(Window.unboundedPreceding, Window.currentRow)
df3 = spark.read.format("com.crealytics.spark.excel")\
        .option("sheetName", "Sheet1")\
        .option("useHeader", "false")\
        .option("treatEmptyValuesAsNulls", "false")\
        .option("inferSchema", "true").option("addColorColumns", "true")\
        .option("usePlainNumberFormat","true")\
        .option("startColumn", 0)\
        .option("endColumn", 99)\
        .option("timestampFormat", "MM-dd-yyyy HH:mm:ss")\
        .option("maxRowsInMemory", 20)\
        .option("excerptSize", 10)\
        .option("header", "true")\
        .format("excel")\
        .load("/home/lera/Homework/HW6/Home_3_4/s4_2.xlsx").limit(1000)\
        .withColumn("проценты", sum1(col("Платеж по процентам")).over(w))\
        .withColumn("долг", sum1(col("Платеж по основному долгу")).over(w))

df3.write.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")\
        .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasketl4b1")\
        .mode("append").save()

df4 = df3.toPandas()
# Get current axis
ax = plt.gca()
ax.ticklabel_format(style='plain')
# bar plot
df4.plot(kind='line',
        x='№',
        y='долг',
        color='blue', ax=ax)
df4.plot(kind='line',
        x='№',
        y='проценты',
        color='pink', ax=ax)

sql.execute("""drop table if exists spark.`tasketl4b2`""",con)
sql.execute("""CREATE TABLE if not exists spark.`tasketl4b2` (
```



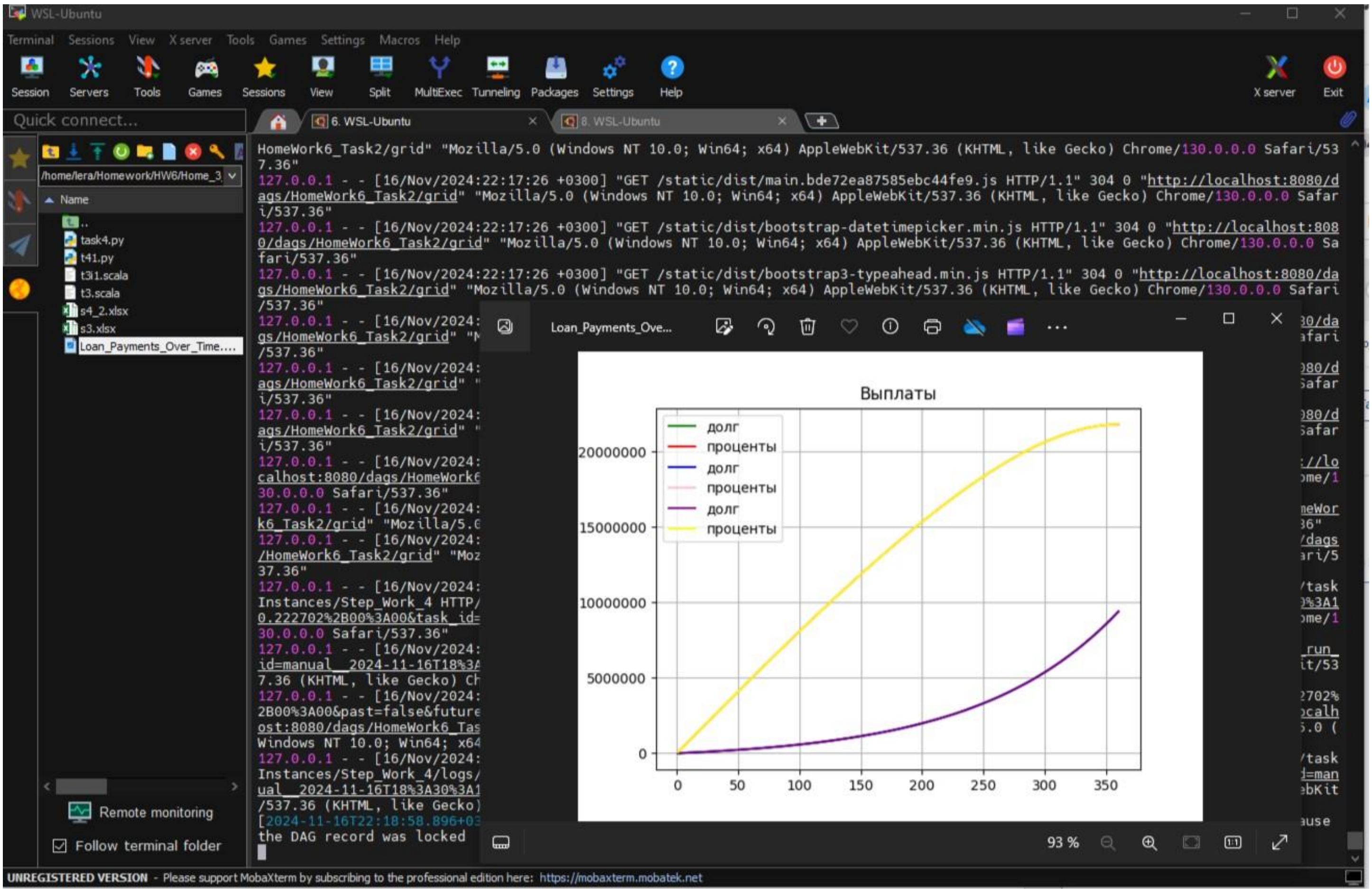
```
  `№` INT(10) NULL DEFAULT NULL,
  `Месяц` DATE NULL DEFAULT NULL,
  `Сумма платежа` FLOAT NULL DEFAULT NULL,
  `Платеж по основному долгу` FLOAT NULL DEFAULT NULL,
  `Платеж по процентам` FLOAT NULL DEFAULT NULL,
  `Остаток долга` FLOAT NULL DEFAULT NULL,
  `проценты` FLOAT NULL DEFAULT NULL,
  `долг` FLOAT NULL DEFAULT NULL
)
COLLATE='utf8mb4_0900_ai_ci'
ENGINE=InnoDB""",con)
from pyspark.sql.window import Window
from pyspark.sql.functions import sum as sum1
w = Window.partitionBy(lit(1)).orderBy("№").rowsBetween(Window.unboundedPreceding, Window.currentRow)
df5 = spark.read.format("com.crealytics.spark.excel")\
          .option("sheetName", "Sheet1")\
          .option("useHeader", "true")\
          .option("treatEmptyValuesAsNulls", "false")\
          .option("inferSchema", "true").option("addColorColumns", "true")\
          .option("usePlainNumberFormat","true")\
          .option("startColumn", 0)\
          .option("endColumn", 99)\
          .option("timestampFormat", "MM-dd-yyyy HH:mm:ss")\
          .option("maxRowsInMemory", 20)\
          .option("excerptSize", 10)\
          .option("header", "true")\
          .format("excel")\
          .load("/home/lera/Homework/HW6/Home_3_4/s4_2.xlsx").limit(1000)\
          .withColumn("проценты", sum1(col("Платеж по процентам")).over(w))\
          .withColumn("долг", sum1(col("Платеж по основному долгу")).over(w))

df5.write.format("jdbc").option("url","jdbc:mysql://localhost:33061/spark?user=Airflow&password=1")\
          .option("driver", "com.mysql.cj.jdbc.Driver").option("dbtable", "tasketl4b2")\
          .mode("append").save()

df6 = df5.toPandas()
# Get current axis
ax = plt.gca()
ax.ticklabel_format(style='plain')
# bar plot
df6.plot(kind='line',
          x='№',
          y='долг',
          color='purple', ax=ax)
df6.plot(kind='line',
          x='№',
          y='проценты',
          color='yellow', ax=ax)

# set the title
plt.title('Выплаты')
plt.grid ( True )
ax.set(xlabel=None)

plot_directory = "/home/lera/Homework/HW6/Home_3_4/"
plot_filename = "Loan_Payments_Over_Time.png"
plt.savefig(plot_directory + plot_filename)

# show the plot
plt.legend(['долг_86689', 'проценты_86689','долг_120000', 'проценты_120000','долг_150000', 'проценты_150000'])
```



```
# show the plot
plt.legend(['долг_86689', 'проценты_86689','долг_120000', 'проценты_120000','долг_1
plt.show()

spark.stop()
t1=time.time()
print('finished',time.strftime('%H:%M:%S',time.gmtime(round(t1-t0))))
```



```
[2024-11-16, 18:30:54 UTC] {subprocess.py:93} INFO - 24/******/***6 2***:30:54 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
[2024-11-16, 18:30:56 UTC] {subprocess.py:93} INFO - Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
[2024-11-16, 18:30:56 UTC] {subprocess.py:93} INFO - Setting default log level to "WARN".
[2024-11-16, 18:30:56 UTC] {subprocess.py:93} INFO - To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
[2024-11-16, 18:30:58 UTC] {subprocess.py:93} INFO - 24/******/***6 2***:30:58 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
[2024-11-16, 18:31:06 UTC] {subprocess.py:93} INFO - 24/******/***6 2***:3***:06 WARN Utils: Service 'SparkUI' could not bind on port 4040. Attempting port 404***.
[2024-11-16, 18:31:32 UTC] {subprocess.py:93} INFO - ERROR StatusLogger Log4j2 could not find a logging implementation. Please add log4j-core to the classpath. Using SimpleLogger to log to the console...
[2024-11-16, 18:32:15 UTC] {subprocess.py:93} INFO -
[Stage 0:>                                                          (0 + 0) / ***]
[Stage 0:>                                                          (0 + ***) / ***]

[Stage ***:>                                                          (0 + ***) / ***]

MoTTY X****** proxy: No authorisation provided
[2024-11-16, 18:32:29 UTC] {subprocess.py:93} INFO -
[Stage 2:>                                                          (0 + ***) / ***]

[Stage 3:>                                                          (0 + ***) / ***]

[Stage 4:>                                                          (0 + ***) / ***]

[Stage 5:>                                                          (0 + ***) / ***]

finished 00:0***:44
[2024-11-16, 18:32:30 UTC] {subprocess.py:97} INFO - Command exited with return code 0
[2024-11-16, 18:32:31 UTC] {taskinstance.py:1400} INFO - Marking task as SUCCESS. dag_id=HomeWork6_Task2, task_id=Step_Work_4, execution_date=2024*********6T***830***0, start_date=2024*********6T***83024, end_date=2024*********6T***8323***
[2024-11-16, 18:32:31 UTC] {local_task_job_runner.py:228} INFO - Task exited with return code 0
[2024-11-16, 18:32:31 UTC] {taskinstance.py:2778} INFO - 0 downstream tasks scheduled from follow-on schedule check
```



```
HomeWork6_Task2/grid" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/130.0.0.0 Safari/53
7.36"
127.0.0.1 - - [16/Nov/2024:22:17:26 +0300] "GET /static/dist/main.bde72ea87585ebc44fe9.js HTTP/1.1" 304 0 "http://localhost:8080/d
ags/HomeWork6_Task2/grid" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/130.0.0.0 Safar
i/537.36"
127.0.0.1 - - [16/Nov/2024:22:17:26 +0300] "GET /static/dist/bootstrap-datetimepicker.min.js HTTP/1.1" 304 0 "http://localhost:808
0/dags/HomeWork6_Task2/grid" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/130.0.0.0 Sa
fari/537.36"
127.0.0.1 - - [16/Nov/2024:22:17:26 +0300] "GET /static/dist/bootstrap3-typeahead.min.js HTTP/1.1" 304 0 "http://localhost:8080/da
gs/HomeWork6_Task2/grid" "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/130.0.0.0 Safari
/537.36"
127.0.0.1 - - [16/Nov/2024:                                                                                            30/da
gs/HomeWork6_Task2/grid" "M                                                                                            afari
/537.36"
127.0.0.1 - - [16/Nov/2024:                                                                                            980/d
ags/HomeWork6_Task2/grid" "                                                                                            Safar
i/537.36"
127.0.0.1 - - [16/Nov/2024:                                                                                            980/d
ags/HomeWork6_Task2/grid" "                                                                                            Safar
i/537.36"
127.0.0.1 - - [16/Nov/2024:                                                                                            ://lo
calhost:8080/dags/HomeWork6                                                                                            ome/1
30.0.0.0 Safari/537.36"
127.0.0.1 - - [16/Nov/2024:                                                                                            neWor
k6_Task2/grid" "Mozilla/5.0                                                                                            36"
127.0.0.1 - - [16/Nov/2024:                                                                                            /dags
/HomeWork6_Task2/grid" "Moz                                                                                            ari/5
37.36"
127.0.0.1 - - [16/Nov/2024:                                                                                            /task
Instances/Step_Work_4 HTTP/                                                                                            0%3A1
0.222702%2B00%3A00&task_id=                                                                                            ome/1
30.0.0.0 Safari/537.36"
127.0.0.1 - - [16/Nov/2024:                                                                                            _run_
id=manual__2024-11-16T18%3A                                                                                            it/53
7.36 (KHTML, like Gecko) Ch
127.0.0.1 - - [16/Nov/2024:                                                                                            2702%
2B00%3A00&past=false&future                                                                                            ocalh
ost:8080/dags/HomeWork6_Tas                                                                                            5.0 (
Windows NT 10.0; Win64; x64
127.0.0.1 - - [16/Nov/2024:                                                                                            /task
Instances/Step_Work_4/logs/                                                                                            d=man
ual__2024-11-16T18%3A30%3A1                                                                                            ebKit
/537.36 (KHTML, like Gecko)
[2024-11-16T22:18:58.896+03                                                                                            ause
the DAG record was locked
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net