Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ **ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ**

КАФЕДРА **КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ 09.04.01/07 **Интеллектуальные системы анализа, обработки и интерпретации больших данных**

О Т Ч Е Т

по лабораторной работе № 10

Название: Spark

Дисциплина: Языки программирования для работы с большими данными

Студент ИУ6-23М (Группа) _____ (Подпись, дата) А.А. Аветисян (И.О. Фамилия)

Преподаватель _____ (Подпись, дата) _____ (И.О. Фамилия)

Москва, 2022

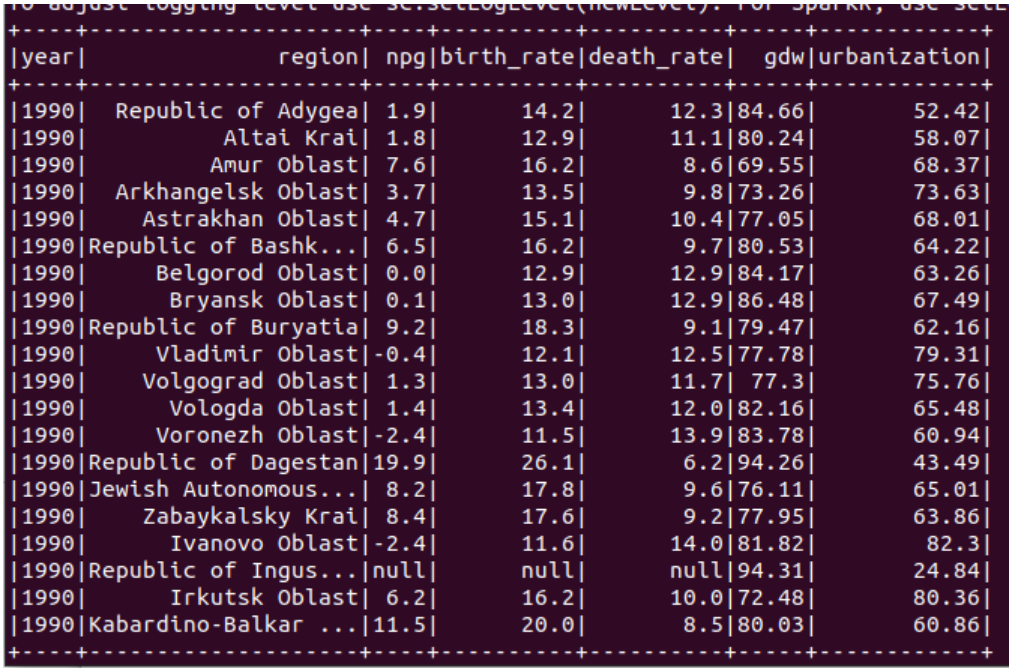
Лабораторная работа № 10

Задание:

1) Выбрать любой датасет на kaggle.com
2) Сделать 10 выборок данных на ваше усмотрение**Ход работы:**

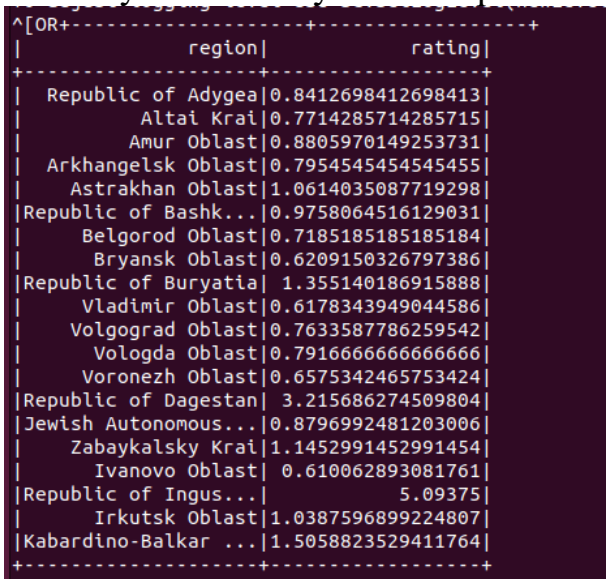
Код программы:

```
from pyspark.sql import SparkSession

spark=SparkSession \
.builder \
.appName("SQL") \
.config("spark.some.config.option", "some-value") \
.getOrCreate()


spark.sparkContext.setLogLevel("WARN")

data=spark.read.load("/home/hdoop/russian_demography.csv", format="csv",
sep=",",inferSchema="true", header="true")



data.registerTempTable("table")

#df = spark.sql("select * from table").show()
#df = spark.sql("select region, (birth_rate/death_rate) as rating from table WHERE year=2017").show()
#df = spark.sql("select region, MAX(npg) from table WHERE region!='Москва' AND year BETWEEN 2007 AND 2017 GROUP
BY region ORDER by MAX(npg) DESC LIMIT 1").show()
#df = spark.sql("select region, MAX(npg) from table GROUP BY region ORDER by MAX(npg) DESC LIMIT 5").show()
#df = spark.sql("select region, MIN(urbanization) from table WHERE year BETWEEN 2007 AND 2017 GROUP BY region
ORDER by MIN(urbanization) ASC LIMIT 1").show()
#df = spark.sql("select year, npg from table WHERE region = 'Chechen Republic' AND year BETWEEN 2007 AND 2017
ORDER by year").show()
#df = spark.sql("select region, MAX(npg) from table WHERE year=2017 GROUP BY region ORDER by MAX(npg) DESC LIMIT
1").show()
#df = spark.sql("select region, birth_rate from table WHERE year=2017 AND birth_rate BETWEEN 5 AND 9").show()
#df = spark.sql("select region, MAX(gdw) from table WHERE year=2017 GROUP BY region ORDER by MAX(gdw) DESC
LIMIT 1").show()
df = spark.sql("select distinct count(region) from table").show()
input()
```



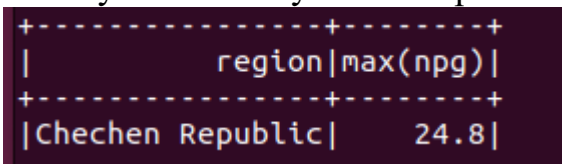
year	region	npg	birth_rate	death_rate	gdw	urbanization
1990	Republic of Adygea	1.9	14.2	12.3	84.66	52.42
1990	Altai Krai	1.8	12.9	11.1	80.24	58.07
1990	Amur Oblast	7.6	16.2	8.6	69.55	68.37
1990	Arkhangelsk Oblast	3.7	13.5	9.8	73.26	73.63
1990	Astrakhan Oblast	4.7	15.1	10.4	77.05	68.01
1990	Republic of Bashk...	6.5	16.2	9.7	80.53	64.22
1990	Belgorod Oblast	0.0	12.9	12.9	84.17	63.26
1990	Bryansk Oblast	0.1	13.0	12.9	86.48	67.49
1990	Republic of Buryatia	9.2	18.3	9.1	79.47	62.16
1990	Vladimir Oblast	-0.4	12.1	12.5	77.78	79.31
1990	Volgograd Oblast	1.3	13.0	11.7	77.3	75.76
1990	Vologda Oblast	1.4	13.4	12.0	82.16	65.48
1990	Voronezh Oblast	-2.4	11.5	13.9	83.78	60.94
1990	Republic of Dagestan	19.9	26.1	6.2	94.26	43.49
1990	Jewish Autonomous...	8.2	17.8	9.6	76.11	65.01
1990	Zabaykalsky Krai	8.4	17.6	9.2	77.95	63.86
1990	Ivanovo Oblast	-2.4	11.6	14.0	81.82	82.3
1990	Republic of Ingus...	null	null	null	94.31	24.84
1990	Irkutsk Oblast	6.2	16.2	10.0	72.48	80.36
1990	Kabardino-Balkar ...	11.5	20.0	8.5	80.03	60.86

Рисунок 1 – Результат запроса



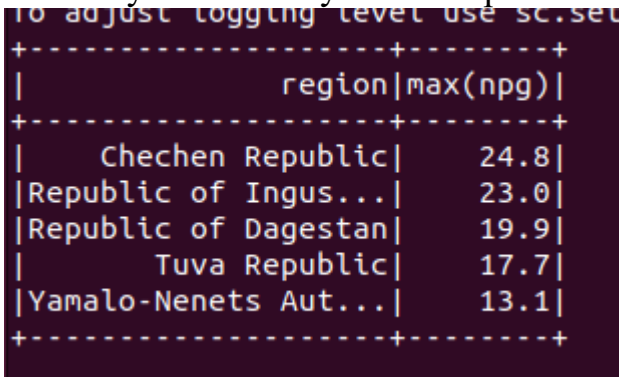
region	rating
Republic of Adygea	0.8412698412698413
Altai Krai	0.7714285714285715
Amur Oblast	0.8805970149253731
Arkhangelsk Oblast	0.7954545454545455
Astrakhan Oblast	1.0614035087719298
Republic of Bashk...	0.9758064516129031
Belgorod Oblast	0.7185185185185184
Bryansk Oblast	0.6209150326797386
Republic of Buryatia	1.355140186915888
Vladimir Oblast	0.6178343949044586
Volgograd Oblast	0.7633587786259542
Vologda Oblast	0.7916666666666666
Voronezh Oblast	0.6575342465753424
Republic of Dagestan	3.215686274509804
Jewish Autonomous...	0.8796992481203006
Zabaykalsky Krai	1.1452991452991454
Ivanovo Oblast	0.610062893081761
Republic of Ingus...	5.09375
Irkutsk Oblast	1.0387596899224807
Kabardino-Balkar ...	1.5058823529411764

Рисунок 2 – Результат запроса



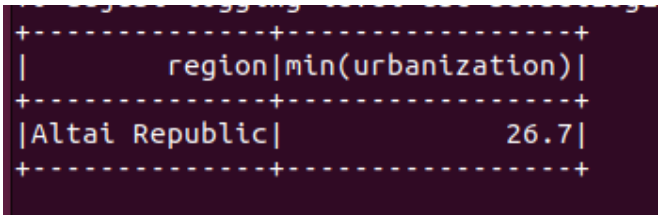
region	max(npg)
Chechen Republic	24.8

Рисунок 3 – Результат запроса



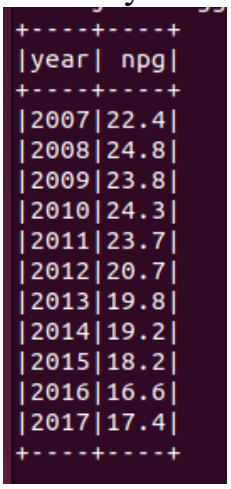
region	max(npg)
Chechen Republic	24.8
Republic of Ingus...	23.0
Republic of Dagestan	19.9
Tuva Republic	17.7
Yamalo-Nenets Aut...	13.1

Рисунок 4 – Результат запроса



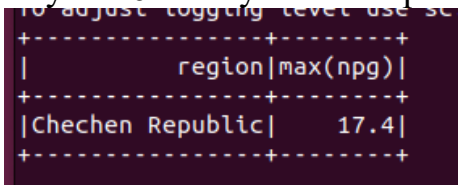
```
+--------------+-----------------+
|        region|min(urbanization)|
+--------------+-----------------+
|Altai Republic|             26.7|
+--------------+-----------------+
```

Рисунок 5 – Результат запроса



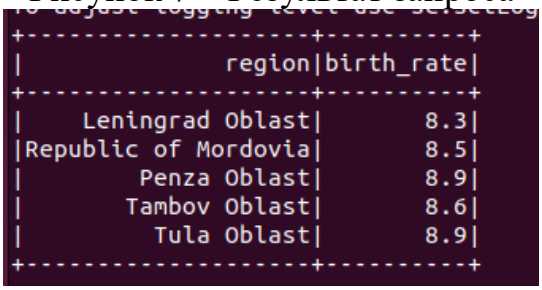
```
+----+----+
|year| npg|
+----+----+
|2007|22.4|
|2008|24.8|
|2009|23.8|
|2010|24.3|
|2011|23.7|
|2012|20.7|
|2013|19.8|
|2014|19.2|
|2015|18.2|
|2016|16.6|
|2017|17.4|
+----+----+
```

Рисунок 6 – Результат запроса



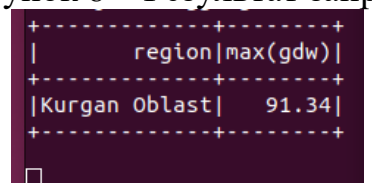
```
+----------------+--------+
|          region|max(npg)|
+----------------+--------+
|Chechen Republic|    17.4|
+----------------+--------+
```

Рисунок 7 – Результат запроса



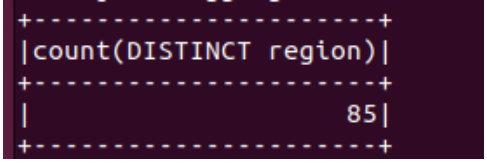
```
+-------------------+----------+
|             region|birth_rate|
+-------------------+----------+
|    Leningrad Oblast|       8.3|
|Republic of Mordovia|       8.5|
|        Penza Oblast|       8.9|
|       Tambov Oblast|       8.6|
|         Tula Oblast|       8.9|
+-------------------+----------+
```

Рисунок 8 – Результат запроса



```
+------------+--------+
|      region|max(gdw)|
+------------+--------+
|Kurgan Oblast|   91.34|
+------------+--------+
```

Рисунок 9 – Результат запроса



```
+----------------------+
|count(DISTINCT region)|
+----------------------+
|                    85|
+----------------------+
```

Рисунок 10 – Результат запроса

Вывод: лабораторная работа выполнена в соответствии с заданием и вариантом.