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 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/07 Интеллектуальные системы анализа, обработки и интерпретации больших данных.**

О Т Ч Е Т

по лабораторной работе № 1 0

Название: Scala Spark

Дисциплина: Языки программирования для работы с большими данными

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

Преподаватель _____ П.В. Степанов
(Подпись, дата) (И.О. Фамилия)

Москва, 2022

Цель работы: получение навыков работы с Scala Spark.

Задание:

1. Выбрать любой датасет на kaggle.com
2. Сделать 10 выборок данных на ваше усмотрение

Выполнение.

```
[1]: import $ivy.`org.apache.spark::spark-sql:3.0.0`;

[1]: import $ivy.$                                  ;

[2]: import org.apache.spark.sql._

     val spark = SparkSession.
                 builder().
                 appName("scala-spark-notebook").
                 master("spark://spark-master:7077").
                 config("spark.executor.memory", "512m").
                 getOrCreate()

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
22/05/20 14:54:17 INFO SparkContext: Running Spark version 3.0.0
22/05/20 14:54:18 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/05/20 14:54:18 INFO ResourceUtils: ==============================================================
22/05/20 14:54:18 INFO ResourceUtils: Resources for spark.driver:

22/05/20 14:54:18 INFO ResourceUtils: ==============================================================
22/05/20 14:54:18 INFO SparkContext: Submitted application: scala-spark-notebook
22/05/20 14:54:18 INFO SecurityManager: Changing view acls to: root
22/05/20 14:54:18 INFO SecurityManager: Changing modify acls to: root
22/05/20 14:54:18 INFO SecurityManager: Changing view acls groups to:
22/05/20 14:54:18 INFO SecurityManager: Changing modify acls groups to:
22/05/20 14:54:18 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(root); groups with view permissions: Set(); users  with modify permissi
ons: Set(root); groups with modify permissions: Set()
22/05/20 14:54:19 INFO Utils: Successfully started service 'sparkDriver' on port 36645.
22/05/20 14:54:19 INFO SparkEnv: Registering MapOutputTracker
22/05/20 14:54:19 INFO SparkEnv: Registering BlockManagerMaster
22/05/20 14:54:19 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/05/20 14:54:19 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/05/20 14:54:19 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
22/05/20 14:54:19 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-b2ac87c5-a8dd-4937-8e0f-aeb07a6da90b
22/05/20 14:54:19 INFO MemoryStore: MemoryStore started with capacity 1509.6 MiB
22/05/20 14:54:19 INFO SparkEnv: Registering OutputCommitCoordinator
22/05/20 14:54:20 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/05/20 14:54:20 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://8b3bcc328127:4040
22/05/20 14:54:20 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://spark-master:7077...
22/05/20 14:54:20 INFO TransportClientFactory: Successfully created connection to spark-master/172.18.0.3:7077 after 63 ms (0 ms spent in bootstraps)
22/05/20 14:54:21 INFO StandaloneSchedulerBackend: Connected to Spark cluster with app ID app-20220520145421-0000
22/05/20 14:54:21 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 43551.
22/05/20 14:54:21 INFO NettyBlockTransferService: Server created on 8b3bcc328127:43551
22/05/20 14:54:21 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
22/05/20 14:54:21 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 8b3bcc328127, 43551, None)
22/05/20 14:54:21 INFO BlockManagerMasterEndpoint: Registering block manager 8b3bcc328127:43551 with 1509.6 MiB RAM, BlockManagerId(driver, 8b3bcc328127, 43551, None)
```

[3]:
```
import org.apache.log4j.{Level, Logger};
Logger.getLogger("org").setLevel(Level.OFF);
```

[3]:
```
import org.apache.log4j.{Level, Logger};
```

[4]:
```
val data = spark.read.format("csv") \
            .option("sep", ",") \
            .option("header", "true") \
            load("russian_demography.csv")
```

[4]:
```
data: DataFrame = [year: string, region: string ... 5 more fields]
```

[6]:
```
data.count
```

[6]:
```
res5: Long = 2380L
```

[7]:
```
data.printSchema
```

```
root
 |-- year: string (nullable = true)
 |-- region: string (nullable = true)
 |-- npg: string (nullable = true)
 |-- birth_rate: string (nullable = true)
 |-- death_rate: string (nullable = true)
 |-- gdw: string (nullable = true)
 |-- urbanization: string (nullable = true)
```

[8]:
```
data.show()
```

```
+----+--------------------+----+----------+----------+-----+------------+
|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|
```

```
[10]: val req1 = data.where(data("region") === "Amur Oblast" && data("year") > 2007).show()
```

```
+----+-----------+----+----------+----------+-----+------------+
|year|     region| npg|birth_rate|death_rate|  gdw|urbanization|
+----+-----------+----+----------+----------+-----+------------+
|2008|Amur Oblast|-2.2|      12.9|      15.6|55.66|        66.6|
|2009|Amur Oblast|-1.4|      13.2|      15.1|56.86|        66.6|
|2010|Amur Oblast|-1.5|      13.8|      15.3|58.76|        66.5|
|2011|Amur Oblast|-1.2|      13.6|      14.8|60.63|        66.9|
|2012|Amur Oblast|-0.4|      14.3|      14.7|63.04|        67.0|
|2013|Amur Oblast| 0.2|      14.1|      13.9| 65.4|        67.1|
|2014|Amur Oblast|-0.2|      13.7|      13.9|68.21|        67.1|
|2015|Amur Oblast|-0.6|      13.3|      13.9|70.29|        67.3|
|2016|Amur Oblast|-0.8|      12.9|      13.7|72.97|        67.3|
|2017|Amur Oblast|-1.6|      11.8|      13.4|75.14|        67.3|
+----+-----------+----+----------+----------+-----+------------+
```

Ссылка на программное решение:

https://github.com/ArMaxik/BigDataLanguages/tree/main/lr10

Вывод: в ходе лабораторной работы были получены навыки работы с Spark Scala.