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 № 10

Вариант 14

Название: Scala Spark

Дисциплина: Языки программирования для работы с большими данными

Студент ИУ6-23М _____ В.Е. Санталов
(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
22/05/26 15:58:25 INFO SparkContext: Running Spark version 3.0.0
22/05/26 15:58:26 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/05/26 15:58:26 INFO ResourceUtils: ==============================================================
22/05/26 15:58:26 INFO ResourceUtils: Resources for spark.driver:

22/05/26 15:58:26 INFO ResourceUtils: ==============================================================
22/05/26 15:58:26 INFO SparkContext: Submitted application: scala-spark-notebook
22/05/26 15:58:26 INFO SecurityManager: Changing view acls to: root
22/05/26 15:58:26 INFO SecurityManager: Changing modify acls to: root
22/05/26 15:58:26 INFO SecurityManager: Changing view acls groups to:
22/05/26 15:58:26 INFO SecurityManager: Changing modify acls groups to:
22/05/26 15:58:26 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(root); groups with view permissions: Set();
users  with modify permissions: Set(root); groups with modify permissions: Set()
22/05/26 15:58:27 INFO Utils: Successfully started service 'sparkDriver' on port 39517.
22/05/26 15:58:27 INFO SparkEnv: Registering MapOutputTracker
22/05/26 15:58:27 INFO SparkEnv: Registering BlockManagerMaster
22/05/26 15:58:27 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/05/26 15:58:27 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/05/26 15:58:27 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
22/05/26 15:58:27 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-ab5f6b90-4be5-4a50-a7bd-aa2a85d952b6
22/05/26 15:58:27 INFO MemoryStore: MemoryStore started with capacity 1509.6 MiB
22/05/26 15:58:27 INFO SparkEnv: Registering OutputCommitCoordinator
22/05/26 15:58:28 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/05/26 15:58:28 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://8b3bcc328127:4040
22/05/26 15:58:28 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://spark-master:7077...
22/05/26 15:58:28 INFO TransportClientFactory: Successfully created connection to spark-master/172.18.0.2:7077 after 64 ms (0 ms spent in bootstraps)
22/05/26 15:58:29 INFO StandaloneSchedulerBackend: Connected to Spark cluster with app ID app-20220526155829-0000
22/05/26 15:58:29 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 41517.
22/05/26 15:58:29 INFO NettyBlockTransferService: Server created on 8b3bcc328127:41517

22/05/26 15:58:29 INFO StandaloneSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
[2]:       org.apache.spark.sql._
import
spark: SparkSession = org.apache.spark.sql.SparkSession@2201bab

[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_passenger_air_service_2.csv")

[4]: data: DataFrame = [Airport name: string, Year: string ... 14 more fields]

[5]: data.count

[5]: res4: Long = 3961L

[6]: data.printSchema

root
 |-- Airport name: string (nullable = true)
 |-- Year: string (nullable = true)
 |-- January: string (nullable = true)
 |-- February: string (nullable = true)
 |-- March: string (nullable = true)
 |-- April: string (nullable = true)
 |-- May: string (nullable = true)
 |-- June: string (nullable = true)
 |-- July: string (nullable = true)
 |-- August: string (nullable = true)
 |-- September: string (nullable = true)
 |-- October: string (nullable = true)
 |-- November: string (nullable = true)
 |-- December: string (nullable = true)
 |-- Whole year: string (nullable = true)
 |-- Airport coordinates: string (nullable = true)
```

[7]:
```
data.show()
```

```
+--------------------+----+-------+--------+-----+-----+---+----+----+------+---------+-------+--------+--------+----------+--------------------+
|        Airport name|Year|January|February|March|April|May|June|July|August|September|October|November|December|Whole year| Airport coordinates|
+--------------------+----+-------+--------+-----+-----+---+----+----+------+---------+-------+--------+--------+----------+--------------------+
|              Abakan|2020|13495.0| 14940.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('91.3997...|
|              Aikhal|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('111.543...|
|                Loss|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('125.398...|
|             Amderma|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('61.5774...|
|     Anadyr (Carbon)|2020| 4255.0|  4565.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('177.738...|
|   Anapa (Vitjazevo)|2020|43359.0| 33653.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('37.3415...|
|   Apatite (Khibiny)|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('33.5819...|
|Arkhangelsk (Vask...|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('40.7067...|
|Arkhangelsk (Talagy)|2020|62698.0| 61408.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('40.7148...|
|Astrakhan (Narima...|2020|47384.0| 46387.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('47.9998...|
|                Trip|2020|  157.0|   165.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('138.042...|
|              Baykit|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('96.3667...|
|Barnaul (Titov Name)|2020|34657.0| 33369.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('83.5477...|
|             In Salah|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('130.399...|
|      white Mountain|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('146.228...|
|            Belgorod|2020|30337.0| 23907.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|        Novy Urengoy|2020| 4472.0|  4032.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('66.6945...|
|             Belushi|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('47.6234...|
|              Usinsk|2020|    0.0|     0.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('65.0461...|
|        Beringovskiy|2020|   52.0|    64.0|  0.0|  0.0|0.0| 0.0| 0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('179.293...|
+--------------------+----+-------+--------+-----+-----+---+----+----+------+---------+-------+--------+--------+----------+--------------------+
only showing top 20 rows
```

[9]:
```
val req1 = data.where(data("Airport name") === "Belgorod").show()
```

```
+------------+----+-------+--------+-------+-------+-------+-------+-------+------+---------+-------+--------+--------+----------+--------------------+
|Airport name|Year|January|February|  March|  April|    May|   June|   July|August|September|October|November|December|Whole year| Airport coordinates|
+------------+----+-------+--------+-------+-------+-------+-------+-------+------+---------+-------+--------+--------+----------+--------------------+
|    Belgorod|2020|30337.0| 23907.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|    Belgorod|2019|18809.0| 17591.0|23221.0|30448.0|45872.0|55952.0|57922.0|58609.0|  55720.0|43963.0| 32577.0|    7.83|  468672.0|(Decimal('36.5705...|
|    Belgorod|2018|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|54722.0|  49487.0|39644.0| 26707.0|    6.34|       0.0|(Decimal('36.5705...|
|    Belgorod|2017|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|    Belgorod|2016|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|    Belgorod|2015|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|    Belgorod|2014|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|    Belgorod|2013|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|    Belgorod|2012|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
|    Belgorod|2011|    0.0|     0.0|    0.0|    0.0|    0.0|    0.0|    0.0|   0.0|      0.0|    0.0|     0.0|     0.0|       0.0|(Decimal('36.5705...|
```

Ссылка на программное решение:

https://github.com/Time2HackJS/BigDataLanguages/tree/master/lr10

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