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 13

Название: Scala Spark

Дисциплина: Языки программирования для работы с большими данными

Студент ИУ6-13М (Группа) _____ (Подпись, дата) Е.К. Матвиенко (И.О. Фамилия)

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

Москва, 2022

Цель лабораторной работы: получение первичных навыков работы со Scala Spark с использованием языка программирования Java.

Ход работы:

Задание:

1. Выбрать любой датасет (взят датасет из курсового проекта, тема «Салон красоты»)
2. Сделать 10 выборок данных

Листинг вывода данных из датасета и выполнения одного из запросов (файл spark.scala):

```
import org.apache.spark.sql.SparkSession

object CounterDemo {
  def main(args: Array[String]): Unit = {
    val conf = new
        SparkConf().setAppName("CounterDemo").setMaster("local[*]")

    val path_1="hdfs://localhost:9000/procedure.csv"
    val path_2="hdfs://localhost:9000/master.csv"
    val path_3="hdfs://localhost:9000/client.csv"

    val df_procedure = spark.read.option("header", "true").csv(path_1)
    val df_master = spark.read.option("header", "true").csv(path_2)
    val df_client = spark.read.option("header", "true").csv(path_3)

    df_procedure.createOrReplaceTempView("procedures")
    df_master.createOrReplaceTempView("masters")
    df_client.createOrReplaceTempView("clients")

    spark.sql("SELECT * FROM procedures").show(5)
    spark.sql("SELECT * FROM masters").show(5)
    spark.sql("SELECT * FROM clients").show(5)

    spark.sql("SELECT DISTINCT procedures.master_id, COUNT(procedures.procedure_id) "
+
              "FROM procedures GROUP BY procedures.master_id").show()

    spark.stop()
  }
}
```

```
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.2.1
      /_/

Using Scala version 2.12.15 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_131)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val path_1="hdfs://localhost:9000/procedure.csv"
path_1: String = hdfs://localhost:9000/procedure.csv

scala> val path_2="hdfs://localhost:9000/master.csv"
path_2: String = hdfs://localhost:9000/master.csv

scala> val path_3="hdfs://localhost:9000/client.csv"
path_3: String = hdfs://localhost:9000/client.csv

scala>

scala> val df_procedure = spark.read.option("header", "true").csv(path_1)
df_procedure: org.apache.spark.sql.DataFrame = [procedure_id: string, client_id: string ... 5 more fields]

scala> val df_master = spark.read.option("header", "true").csv(path_2)
df_master: org.apache.spark.sql.DataFrame = [master_id: string, master_specialization: string ... 4 more fields]

scala> val df_client = spark.read.option("header", "true").csv(path_3)
df_client: org.apache.spark.sql.DataFrame = [client_id: string, client_personal_data: string ... 1 more field]

scala>
     | df_procedure.createOrReplaceTempView("procedures")

scala> df_master.createOrReplaceTempView("masters")

scala> df_client.createOrReplaceTempView("clients")
```

```
scala> spark.sql("SELECT * FROM procedures").show(5)
+------------+---------+--------------+---------+-----+--------------------+--------------------+
|procedure_id|client_id|procedure_date|master_id|price|             service|           cosmetics|
+------------+---------+--------------+---------+-----+--------------------+--------------------+
|        PR01|     5H3T|    2018-01-22|     SP03| 5067|[Ультразвуковая ч...|[Лак ART-VISAGE, ...|
|        PR02|     J4SI|    2018-02-03|     SP01| 2570|[Уход за руками, ...|    [Чабрец, Клевер]|
|        PR03|     S747|    2021-12-30|     SP02| 4289|[Маникюр Европейс...|[Крем для лица A`...|
|        PR04|     YK9N|    2016-06-26|     SP06| 5559|[Артдизайн, Параф...|[Лак OPI Nail Lac...|
|        PR05|     YK9N|    2020-10-09|     SP05| 2901|       [Лимфодренаж]|[Лак Benecos Happ...|
+------------+---------+--------------+---------+-----+--------------------+--------------------+
only showing top 5 rows


scala> spark.sql("SELECT * FROM masters").show(5)
+---------+--------------------+-----------------+--------------------+--------------------+--------------------+
|master_id|master_specialization|master_experience|master_personal_data|            schedule|             rewiews|
+---------+--------------------+-----------------+--------------------+--------------------+--------------------+
|     SP01| [Покрытие гель-ла...|                4|Шаповалова Мария ...|Чт,Пт,Сб 15:00-20:00|[Спасибо огромное...|
|     SP02| [Пилинг стоп, Дез...|               34|Трофимов Михаил О...|  Ср, Пт 15:00-20:00|[Люди будте бдите...|
|     SP03| [Маникюр Европейс...|               36|Белоусова Елизаве...|Пн, Вт, Сб, Вс 15...|[Делаю брови, рес...|
|     SP04| [Маникюр Европейс...|               40|Семина Анна Игоревна|Пн, Вт, Чт, Пт 10...|[Всегда выхожу от...|
|     SP05| [Педикюр Классиче...|               21|Матвеева Валерия ...|Пн, Вт, Чт, Пт 10...|[Прекрасный масте...|
+---------+--------------------+-----------------+--------------------+--------------------+--------------------+
only showing top 5 rows


scala> spark.sql("SELECT * FROM clients").show(5)
+---------+--------------------+----------+
|client_id|client_personal_data|client_age|
+---------+--------------------+----------+
|     5H3T|  Сорокин Илья Ильич|        37|
|     J4SI|Попов Иван Андреевич|        49|
|     S747|Филатова Елизавет...|        63|
|     YK9N|Миронова Кира Зах...|        64|
|     YK9N|Зимин Тимур Алекс...|        24|
+---------+--------------------+----------+
only showing top 5 rows
```

```
scala> spark.sql("SELECT DISTINCT procedures.master_id, COUNT(procedures.procedure_id) FROM procedures GROUP BY procedures.ma
ster_id").show()
+---------+-------------------+
|master_id|count(procedure_id)|
+---------+-------------------+
|     SP05|                  2|
|     SP19|                  3|
|     SP21|                  1|
|     SP24|                  2|
|     SP23|                  1|
|     SP02|                  2|
|     SP14|                  1|
|     SP10|                  2|
|       SP|                  1|
|     SP03|                  1|
|     SP18|                  1|
|     SP11|                  3|
|     SP16|                  2|
|     SP17|                  1|
|     SP13|                  1|
|     SP15|                  1|
|     SP06|                  1|
|     SP20|                  1|
|     SP07|                  1|
|     SP22|                  2|
+---------+-------------------+
only showing top 20 rows
```

Рисунок 1 - Результат выполнения запроса

Программное решение представлено в репозитории распределённой системы управления версиями Git: https://github.com/matvilen/BigDataLanguages/tree/main/lab10/src/main/scala

Вывод: при выполнении лабораторной работы были получены навыки работы со Scala Spark.