



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

## Вариант 6

**Дисциплина:** Языки программирования для работы с большими данными

Преподаватель	<hr/>	П.В. Степанов
	(Подпись, дата)	(И.О. Фамилия)

Москва, 2022

## Цель работы:

Получение навыков работы со Scala Spark.

## Выполнение:

### Задание:

1. Выбрать любой датасет (взяв датасет из курсового проекта, тема «Поликлиника»)
2. Сделать 10 выборки данных на ваше усмотрение

Листинг 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 sc = new SparkContext(conf);
    val spark = SparkSession.builder.appName("Test app").getOrCreate()
    val path = "/home/stalekc/click.csv"
    val df = spark.read.option("header", "true").csv(path)
    df.show()
    df.createOrReplaceTempView("click")
    spark.sql("select id, count(id) as counter from click group by
id").show()
    spark.stop()
  }
}
```

```
scala> val path = "/home/stalekc/click.csv"
path: String = /home/stalekc/click.csv

scala> val df = spark.read.option("header", "true").csv(path)
df: org.apache.spark.sql.DataFrame = [client_ip: string, utm_marks: string ... 3 more fields]

scala> df.show()
+-----+-----+-----+-----+
| client_ip|      utm_marks|click_date|site_url| id|
+-----+-----+-----+-----+
|192.168.0.1|['google', 'cpc',...|2021-01-01|site1.ru| 1|
|192.168.0.2|['yandex', 'cpc',...|2021-02-01|site2.ru| 2|
|192.168.0.2|['yandex', 'cpc',...|2021-02-01|site2.ru| 2|
|192.168.0.3|['google', 'email...|2021-03-01|site3.ru| 3|
|192.168.0.4|['google', 'cpc',...|2021-04-01|site4.ru| 4|
|192.168.0.5|['google', 'cpc',...|2021-05-01|site5.ru| 5|
|192.168.0.1|['google', 'cpc',...|2021-06-01|site6.ru| 4|
|192.168.0.1|['vk', 'cpc', 'pr...|2021-01-01|site7.ru| 3|
|192.168.0.1|['google', 'banne...|2021-02-01|site6.ru| 2|
|192.168.0.1|['google', 'cpc',...|2021-03-01|site5.ru| 1|
|192.168.0.2|['google', 'cpc',...|2021-04-01|site4.ru| 1|
|192.168.0.3|['google', 'cpc',...|2021-05-01|site3.ru| 2|
|192.168.0.4|['facebook', 'cpc...|2021-06-01|site2.ru| 3|
|192.168.0.5|['google', 'artic...|2021-01-01|site1.ru| 4|
|192.168.0.5|['vk', 'cpc', 'pr...|2021-02-01|site2.ru| 5|
|192.168.0.4|['vk', 'email', '...|2021-03-01|site3.ru| 4|
|192.168.0.3|['facebook', 'ema...|2021-04-01|site4.ru| 3|
|192.168.0.2|['vk', 'email', '...|2021-05-01|site5.ru| 2|
|192.168.0.2|['vk', 'email', '...|2021-06-01|site6.ru| 1|
|192.168.0.3|['facebook', 'ema...|2021-01-01|site7.ru| 2|
+-----+-----+-----+-----+
only showing top 20 rows

scala> df.createOrReplaceTempView("click")

scala> spark.sql("select id, count(id) as counter from click group by id").show()
+----+-----+
| id|counter|
+----+-----+
| 3|      4|
| 5|      2|
| 1|      4|
| 4|      4|
| 2|      7|
+----+-----+
```

Рисунок 1 - Результат выполнения запроса

### Вывод:

При выполнении лабораторной работы были получены навыки работы со Scala Spark.