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, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ 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

Название: Spark

Дисциплина: Языки программирования для работы с большими данными

Студент ИУ6-22М (Группа) _____ (Подпись, дата) В.А. Трофимов (И.О. Фамилия)

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

Москва, 2023

Задания:

1) Выбрать любой датасет на kaggle.com
2) Сделать 10 выборок данных по выбранной предметной области

Код для решения задания:

```
package bdjava.lab10;

import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;
import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;
import org.apache.spark.sql.SQLContext;

public class Program {
        public static void main(String[] args) {
            // создаем конфигурацию Spark
            SparkConf conf = new SparkConf().setAppName("CSV
Reader").setMaster("local[*]");

            // создаем контекст Spark
            JavaSparkContext sc = new JavaSparkContext(conf);

            // создаем SQL-контекст Spark
            SQLContext sqlContext = new SQLContext(sc);

            // создаем DataFrame
            Dataset<Row> df = sqlContext.read().format("csv").option("header",
"true").option("inferSchema",
"true").load("src/bdjava/lab10/russian_demography.csv");

            df.createOrReplaceTempView("data");
            Dataset<Row> result1 = sqlContext.sql("SELECT * FROM data WHERE year =
1990");
            Dataset<Row> result2 = sqlContext.sql("SELECT * FROM data WHERE region
LIKE '%Oblast'");
            Dataset<Row> result3 = sqlContext.sql("SELECT region, birth_rate FROM
data WHERE birth_rate > 15");
            Dataset<Row> result4 = sqlContext.sql("SELECT year, AVG(birth_rate) AS
avg_birth_rate FROM data GROUP BY year");
            Dataset<Row> result5 = sqlContext.sql("SELECT region, death_rate FROM
data ORDER BY death_rate DESC LIMIT 10");
            Dataset<Row> result6 = sqlContext.sql("SELECT year, SUM(npg) AS
total_npg FROM data GROUP BY year");
            Dataset<Row> result7 = sqlContext.sql("SELECT region, urbanization
FROM data WHERE urbanization > 70 AND urbanization < 80");
            Dataset<Row> result8 = sqlContext.sql("SELECT region, gdw FROM data
WHERE gdw > 80 ORDER BY gdw DESC");
            Dataset<Row> result9 = sqlContext.sql("SELECT year, region, npg FROM
data WHERE npg > 5 ORDER BY npg DESC LIMIT 10");
            Dataset<Row> result10 = sqlContext.sql("SELECT year, region,
```

```
birth_rate, death_rate, npg FROM data WHERE birth_rate > death_rate");

            result1.show();
            result2.show();
            result3.show();
            result4.show();
            result5.show();
            result6.show();
            result7.show();
            result8.show();
            result9.show();
            result10.show();

            // останавливаем контекст Spark
            sc.stop();
        }
    }
}
```

Часть обрабатываемого файла csv:

```
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 Bashkortostan,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 Oblast,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
```

Вывод:

В процессе выполнения задания было необходимо выбрать датасет на kaggle.com и провести 10 выборок данных по выбранной предметной области. Это позволило разобраться в особенностях выбранного датасета, провести анализ и дать более точную оценку его содержанию. Кроме того, задание помогло понять, как работать с данными и проводить выборки, что может быть полезным для решения будущих задач. В итоге, выполнение задания позволило закрепить знания и навыки

работы с данными, а также использовать эту информацию для решения задач в будущем.