|  |  |
| --- | --- |
|  | **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\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/12 Интеллектуальный анализ больших**

**данных в системах поддержки принятия решений.**

**Отчет**

**по лабораторной работе №10**

**Название:** Spark

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

Студент \_\_\_ИУ6-23М\_\_\_\_ **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_**А.Д. Зеленский**\_\_\_**

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_**П. В. Степанов**\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

**Цель:** ознакомиться с базовыми принципами языка Java для работы с большими данными.

**Вариант 1:**

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

**Решение:**

Ниже представлен код класса первой программы, а на рисунке 1 – результат работы программы.

import org.apache.spark.sql.Dataset;

import org.apache.spark.sql.Row;

import org.apache.spark.sql.SparkSession;

public class Main {

public static void main(String[] args) {

SparkSession spark = SparkSession.builder()

.appName("RandomSplitExample")

.master("local[\*]")

.getOrCreate();

// База данных Marijuana Arrests in Toronto: Racial Disparities

Dataset<Row> data = spark.read().format("csv")

.option("header", "true")

.load("C:\\Users\\artem.zelenskiy\\Downloads\\Arrests.csv");

data.createOrReplaceTempView("arrests");

spark.sql("select \* from arrests ").show(); // Выводи всю БД

spark.sql("select \* from arrests where year=2000 ").show(); // Выводим всё за 2000 год

spark.sql("select year, count (\*) from arrests group by 1").show(); // Выводи к-во приводов по годам

spark.sql("select colour, count (\*) from arrests group by 1").show(); // К-во приводов по цвету кожи

spark.sql("select colour, sex, count (\*) from arrests group by 1,2 order by 1,2").show(); //К-во приводов по цвету кожи и полу

spark.sql("select \* from arrests where citizen='Yes' ").show(); // Вся БД для жителей

spark.sql("select year, count (\*) as maxi from arrests group by 1 order by 2 desc limit 1").show();//Самый орестуемый год

spark.sql("select year, colour, sex, count (\*) from arrests group by 1,2,3 order by 1,2,3").show(); // агрег по году, цвету и полу

spark.sql("select year, count(\*) from arrests where sex='Male' group by 1 having count(\*)>10 ").show();// годы с 11+ приводами

spark.sql("select \* from arrests order by checks desc limit 10").show(); //10 самых опасных людей

// Останавливаем SparkSession

spark.stop();

}

}

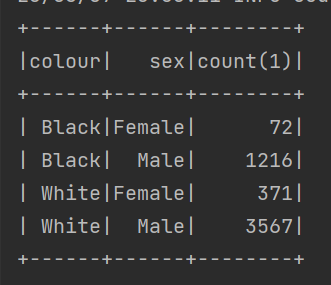


Рисунок 1 – Результат работы программы

**Вывод:** были получены базовые представления о работе языка java с Spark. Были написаны программы согласно вариантам.