|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | **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 **09.04.01/07 Интеллектуальные системы анализа,**

**обработки и интерпретации больших данных.**

**Отчет**

|  |  |
| --- | --- |
| **по лабораторной работе №** | 10 |

**Название:**

Spark

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент | ИУ6-22М |  |  | Е.Е. Шморгун |
|  | (Группа) |  | (Подпись, дата) | (И.О. Фамилия) |
|  |  |  |  |  |
| Преподаватель |  |  |  | П.В. Степанов |
|  |  |  | (Подпись, дата) | (И.О. Фамилия) |

Москва, 2023

**Задача.** Выбрать любой датасет на kaggle.com. Cделать 10 выборок данных по выбранной предметной области

Main.java

import org.apache.spark.api.java.\*;

import org.apache.spark.SparkConf;

import org.apache.spark.api.java.function.Function;

import org.apache.spark.sql.SparkSession;

import org.apache.spark.sql.SparkSession.\*;

import org.apache.spark.sql.Dataset;

import org.apache.spark.sql.Row;

public class Main {

public static void main(String[] args) {

SparkConf conf = new SparkConf().setAppName("Lab10");

JavaSparkContext sc = new JavaSparkContext(conf);

SparkSession spark = SparkSession.builder().appName("Lab10").getOrCreate();

Dataset<Row> csv = spark.read().format("csv").option("header", "true").load("russian\_demography.csv");

csv.createOrReplaceTempView("demography");

spark.sql("select \* from demography limit 10").show();

spark.sql("select year, min(birth\_rate)\n" +

"from demography\n" +

"group by year\n" +

"order by year"

).show();

spark.sql("select count(\*) from demography").show();

spark.sql("select region, count(region) from demography group by region").show();

spark.sql("select region, avg(death\_rate), avg(birth\_rate) from demography group by region").show();

spark.sql("select region, urbanization from demography where urbanization is not null order by urbanization").show();

spark.sql("select year, region from demography where npg > 1").show();

spark.sql("select year, sum(npg) as sum\_npg from demography group by year having sum\_npg > 100").show();

spark.sql("select region, avg(gdw) as avg\_gdw from demography group by region having avg\_gdw > 10").show();

spark.sql("select region, max(gdw) as max\_gdw from demography group by region having max\_gdw >= 90").show();

sc.stop();

}

}

**Вывод:** В данной работе я познакомился с настройкой и использованием Spark для работы с большими наборами данных.