|  |  |
| --- | --- |
| **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 (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

**ОТЧЕТ**

**по лабораторной работе № 10**

|  |  |
| --- | --- |
| **Название:** | Spark |
| **Дисциплина:** | Языки программирования для работы с большими данными |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Студент | ИУ6-22М |  | 26.05.2023 | И.Б. Нуриддинов |
|  | (Группа) |  | (Подпись, дата) | (И.О. Фамилия) |
|  |  |  |  |  |
| Преподаватель |  |  |  | П.В. Степанов |
|  |  |  | (Подпись, дата) | (И.О. Фамилия) |

Москва, 2023

Цель работы: Освоение фреймворка Spark и его применение для разработки распределенных вычислительных приложений на языке Java.

*Задание:*

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

import org.apache.spark.sql.SparkSession

import org.apache.spark.{SparkConf, SparkContext}

object Main {

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()

println("Hello world!")

val datafile = spark.read

.format("com.databricks.spark.csv")

.option("header", true)

.load("/home/ismoil/projects/IdeaProjects/lab10/russian\_demography.csv")

// datafile.show()

datafile.createOrReplaceTempView("demography")

// spark.sql("SELECT COUNT(\*) FROM demography").show()

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

// // 1

spark.sql("SELECT \* FROM demography WHERE year = 2014").show()

// // 2

spark.sql("SELECT \* FROM demography WHERE year = 2015 AND region = 'Moscow'").show()

// 3

spark.sql("SELECT \* FROM demography WHERE year = 2017 ORDER BY npg DESC LIMIT 5").show()

// // 4

spark.sql("SELECT year, AVG(birth\_rate) AS avg\_birth\_rate, AVG(death\_rate) AS avg\_death\_rate FROM demography WHERE year = 2000 GROUP BY year").show()

// // 5

spark.sql("SELECT \* FROM demography WHERE year = 2001 AND npg < 0").show()

// // 6

spark.sql("SELECT \* FROM demography WHERE year = 2002 ORDER BY gdw DESC LIMIT 1").show()

// // 7

spark.sql("SELECT AVG(urbanization) AS avg\_urbanization FROM demography").show()

// // 8

spark.sql("SELECT year, MAX(birth\_rate) AS max\_birth\_rate FROM demography GROUP BY year").show()

// // 9

spark.sql("SELECT year, death\_rate, region FROM demography ORDER BY death\_rate DESC LIMIT 5").show()

// // 10

spark.sql("SELECT \* FROM demography WHERE year = 2004 ORDER BY urbanization DESC LIMIT 1").show()

spark.stop()

}

}

Вывод: Лабораторная работа позволила овладеть важными навыками использования фреймворка Spark для разработки распределенных вычислительных приложений. Spark предоставляет мощные инструменты для обработки и анализа больших объемов данных, позволяя эффективно решать сложные вычислительные задачи. Освоение Spark позволит студентам разрабатывать высокопроизводительные и масштабируемые приложения, обрабатывать большие объемы данных и извлекать ценную информацию из них.