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**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 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 (Группа) _____ (Подпись, дата) М.Э.Хабаров (И.О. Фамилия)

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

2024 г

Цель: изучить работу с Spark на языке Java.

Задание №1

Формулировка задания и код программы представлены в листинге 1:

```
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) {
        // Создание сессии Spark
        SparkSession spark = SparkSession.builder()
                .appName("KaggleDataset")
                .master("local")
                .getOrCreate();

        // Загрузка данных из выбранного датасета
        Dataset<Row> dataset = spark.read()
                .option("header", true)
                .option("inferSchema", true)
                .csv("C:/Users/Ксения/OneDrive - MOEX/Рабочий
стол/Марк/Java/lr_10/onlinefoods.csv");
        dataset.createOrReplaceTempView("data");


        // 10 выборок данных:

        // 1) Первые 10 строк датасета
        Dataset<Row> result1 = spark.sql("SELECT * FROM data LIMIT 10");
        result1.show();

        // 2) Уникальные значение занятости
        Dataset<Row> result2 = spark.sql("SELECT DISTINCT Occupation AS
occupations FROM data");
        result2.show();

        // 3) Количество людей по гендерам
        Dataset<Row> result3 = spark.sql("SELECT Gender, COUNT(Gender) AS
gender_count FROM data GROUP BY Gender");
        result3.show();

        // 4) Схема датасета
        dataset.printSchema();

        // 5) Средний возраст клиентов
        Dataset<Row> result5 = spark.sql("SELECT avg(Age) FROM data");
        result5.show();

        // 6) Максимальный возраст по гендеру
        Dataset<Row> result6 = spark.sql("SELECT Gender, MAX(Age) AS max_age
FROM data GROUP BY Gender");
        result6.show();

        // 7) Отсортировать данные по определенному столбцу
        Dataset<Row> result7 = spark.sql("SELECT * FROM data ORDER BY Age
DESC Limit 5");
        result7.show();

        // 8) Оконная
        Dataset<Row> result8 = spark.sql("SELECT Gender, Age, RANK() OVER
(PARTITION Gender ORDER BY Age) as rk FROM data");
```

```
        result8.show();

        // 9) Вывести количество уникальных возрастов
        Dataset<Row> result9 = spark.sql("SELECT COUNT(DISTINCT Age) AS
unique_age FROM data");
        result9.show();

        // 10) Вывести 5 самых загруженных аэропортов
        Dataset<Row> result10 = spark.sql("SELECT Occupation, SUM(January) AS
total_load FROM data GROUP BY AirportName ORDER BY total_load DESC LIMIT 5");
        result10.show();

        spark.stop();
    }
}
```

Вывод: в результате выполнения лабораторной работы были освоены основные принципы Spark на языке Java.