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: Запросы в Spark

Дисциплина: Языки программирования для работы с большими данными

Студент	ИУ6-23М		Э.А. Гаджиев
	(Группа)	(Подпись, дата)	(И.О. Фамилия)
Преподаватель			П.В. Степанов
		(Подпись, дата)	(И.О. Фамилия)

Москва, 2022

```
from pyspark.sql import SparkSession

sparkSession = SparkSession.builder.appName("SQL Queries").getOrCreate()

sparkSession.read.load(path="russian_demography.csv",
                       format="csv", sep=",", header=True, inferSchema=True) \
                    .createOrReplaceTempView("demography")

# Average birth_rate and death_rate per year
sparkSession.sql(
    'SELECT year, avg(birth_rate), avg(death_rate)\
    FROM demography \
    GROUP BY year \
    ORDER BY year;'
).show(n=100)

sparkSession.sql(
    'SELECT region\
    FROM demography \
    WHERE year = 1995 AND birth_rate > death_rate \
    ;'
).show(n=100)

sparkSession.sql(
    'SELECT year, min(urbanization)\
    FROM demography \
    GROUP BY year \
    ORDER BY year;'
).show(n=100)

sparkSession.sql(
    'SELECT region, max(urbanization)\
    FROM demography \
    WHERE region="Kemerovo Oblast" \
    GROUP BY region;'
).show(n=100)

sparkSession.sql(
    'SELECT region, birth_rate \
    FROM demography \
    ORDER BY birth_rate DESC \
    LIMIT 5;'
).show(n=100)

sparkSession.sql(
    'SELECT region, avg(birth_rate), avg(death_rate), avg(urbanization) \
    FROM demography \
    GROUP by year, region \
    ORDER BY year \
    ;'
).show(n=100)

sparkSession.sql(
    'SELECT year, birth_rate \
    FROM demography \
    WHERE year >= 1990 and year <= 1993 \
    ORDER BY birth_rate DESC \
    LIMIT 5;'
).show(n=100)

sparkSession.sql(
    'SELECT region, death_rate \
    FROM demography \
    ORDER BY death_rate DESC \
    LIMIT 5;'
).show(n=100)

sparkSession.sql(
    'SELECT count(region) \
    FROM demography \
    WHERE gdw > 60 \
    GROUP BY year;'
).show(n=100)

sparkSession.sql(
    'SELECT region, min(gdw) \
    FROM demography \
    GROUP BY region \
    LIMIT 10;'
).show(n=100)
```