



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

Дисциплина: Языки программирования для работы с большими данными.

Студент

ИУ6-23М

(Группа)

В.А Антонов

(Подпись, дата)

(И.О. Фамилия)

Преподаватель

П.В. Степанов

(Подпись, дата)

(И.О. Фамилия)

Москва, 2022

Лабораторная работа №10

Задание: Произвести анализ данных двух связанных таблиц “Группа” и “Кадет”

Ход работы: Код программы

```
from pyspark import SparkContext
from datetime import datetime
from pyspark.sql import SparkSession
import time
start_time = time.time()
spark = SparkSession.builder.appName('abc').getOrCreate()
df1=spark.read.csv('hdfs://localhost:9000/zadanie5/Cadets.csv',header=True)
df2=spark.read.csv('hdfs://localhost:9000/zadanie5/Group.csv',header=True)
my_table1 = df1.createOrReplaceTempView('Cadets')
my_table2 = df2.createOrReplaceTempView('Group')
#sql_table = spark.sql('SELECT * FROM Cadets')
#sql_table = spark.sql('SELECT
count(c1.id_group)*g1.cost_education FROM Cadets c1, Group g1
Where c1.id_group = g1.id_group;')
#sql_table = spark.sql('SELECT
c1.id_group,g1.cost_education,count(c1.id_group)*max(g1.cost_education) FROM Cadets c1, Group g1 Where c1.id_group =
g1.id_group;')
sql_table = spark.sql('SELECT g1.id_group as id,
count(c1.id_group)*g1.cost_education AS m FROM Cadets c1, Group
g1 Where c1.id_group = g1.id_group GROUP BY id,
g1.cost_education ORDER BY m desc LIMIT 1;')
sql_table.show()
print("-- %s seconds --" % (time.time() - start_time))
quit()
```

Запросы

```
sql_table = spark.sql('SELECT service FROM Group Where
id_group = 10 GROUP BY service;')

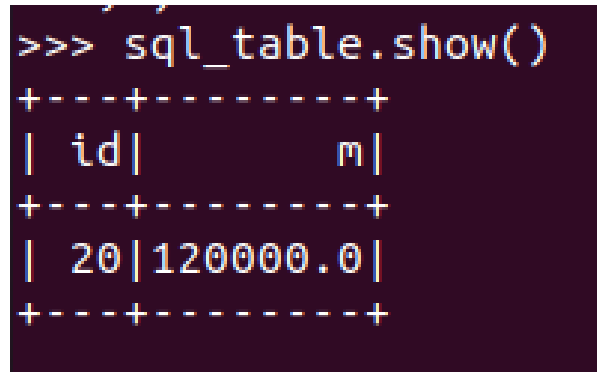
sql_table = spark.sql('SELECT cadets as cadet FROM Cadets
Where id_group = 1 GROUP BY cadet;')

sql_table = spark.sql('SELECT cadets as cadet FROM Cadets
Where id_group > 7 GROUP BY cadet;')

sql_table = spark.sql('SELECT id_group FROM Group Where
cost_education = 30000 GROUP BY id_group ORDER BY id_group;')
```

```
sql_table = spark.sql('SELECT id_group FROM Group Where  
cost_education < 30000 GROUP BY id_group ORDER BY id_group;')  
  
sql_table = spark.sql('SELECT id_teacher FROM Group Where  
id_group > 5 GROUP BY id_group, id_teacher ;')  
  
sql_table = spark.sql('SELECT id_group FROM Group Where  
number_of_classes < 35 GROUP BY id_group;')  
  
sql_table = spark.sql('SELECT id_teacher FROM Group Where  
start_date_of_classes = "2021-09-08" GROUP BY id_teacher;')  
  
sql_table = spark.sql('SELECT id_group FROM Group Where  
start date of classes > "2021-09-08" GROUP BY id group;')
```

Пример результата запроса



```
>>> sql_table.show()  
+---+-----+  
| id|      m|  
+---+-----+  
| 20|120000.0|  
+---+-----+
```

Рисунок 58 – Результат запроса

Вывод: лабораторная работа была выполнена в соответствии с заданием и полученные верные результаты работ программ