



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 (ИУ6)

НАПРАВЛЕНИЕ ПОДГОТОВКИ 09.04.01 Информатика и вычислительная техника
МАГИСТЕРСКАЯ ПРОГРАММА 09.04.01/07 Интеллектуальные системы анализа,
обработки и интерпретации больших данных

О Т Ч Е Т

по лабораторной работе №10

Название: Работа со Spark

Дисциплина: Языки программирования для работы с большими
данными

Студент

ИУ6-22М

(Группа)

(Подпись, дата)

М.А. Зотов

(И.О. Фамилия)

Преподаватель

(Подпись, дата)

П.В. Степанов

(И.О. Фамилия)

Москва, 2023 г.

Цель: получить опыт работы со Spark.

Задание:

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

Выполнение:

Был выбран датасет «Статистика по пользователям интернета в мире» (<https://www.kaggle.com/datasets/ashishraut64/internet-users>).

Подключение Spark и предобработка датасета:

```
from pyspark.sql import SparkSession, functions as f

spark = SparkSession.builder.getOrCreate()
spark.sparkContext.setLogLevel("ERROR")
raw_data = spark.read.format("csv").option("header", True).load("Final.csv")
raw_data.filter((f.col('year') > 1989) &
                (~ f.col('entity').isin('Low and middle income', 'Low
income', 'Lower middle income',
                                     'Middle income', 'Upper middle
income'))).createOrReplaceTempView("src")
```

Запросы:

1. Среднее количество пользователей интернета по годам

```
spark.sql("select year, round(avg(internet_users_num), 2) as
avg_internet_users_num "
"from src "
"where code = 'Region' "
"group by 1 "
"order by 1").show(truncate=False)
```

year	avg_internet_users_num
1990	227402.89
1991	343425.78
1992	519349.44
1993	693947.0
1994	1499437.33
1995	2877366.78
1996	5163554.67

Рисунок 1 – Результат работы 1 запроса

2. Прирост количества пользователей интернета относительно прошлого года

```
spark.sql("with tmp as ("
"    select year, avg(internet_users_num) as avg_internet_users_num
"
"    from src "
"    where code = 'Region' "
"    group by 1) "
"select year, "
"    round(coalesce(avg_internet_users_num - "
"                    lag(avg_internet_users_num) over (order by
year), avg_internet_users_num) "
```

```
"      , 2) as internet_users_num_increase "
"from tmp "
"order by 1").show()
```

year	internet_users_num_increase
1990	227402.89
1991	116022.89
1992	175923.67
1993	174597.56
1994	805490.33
1995	1377929.44
1996	2286187.89
1997	1974325.11

Рисунок 2 – Результат работы 2 запроса

3. Страны, где доля пользователей интернета с 2000 по 2020 год была выше 40%

```
spark.sql("select entity, round(avg(internet_users_perc), 2)
avg_internet_users_perc "
"from src "
"where internet_users_perc > 40 "
"   and code <> 'Region' "
"   and year between 2000 and 2020 "
"group by 1 "
"having count(*) = 21 "
"order by 2 desc").show(truncate=False)
```

entity	avg_internet_users_perc
Iceland	88.66
Norway	87.32
Sweden	84.83
Netherlands	82.26
South Korea	80.11
Canada	78.99

Рисунок 3 – Результат работы 3 запроса

4. Список регионов по максимальной доле пользователей мобильной связи

```
spark.sql("with tmp as ("
"      select entity, round(avg(mobile_users), 2) as avg_mobile_users
"
"      from src "
"      where code = 'Region' "
"      group by 1) "
"select "
"      row_number() over (order by cast(avg_mobile_users as int)
desc) as num, "
"      tmp.* "
"from tmp "
"order by 1").show(truncate=False)
```

	num	entity	avg_mobile_users
1		European Union	74.98
2		High income	71.85
3		Europe and Central Asia	70.25
4		North America	61.98
5		Latin America and Caribbean	52.76

Рисунок 4 – Результат работы 4 запроса

5. Страны с долей пользователей интернета за 2020 год выше среднего

```
spark.sql("select entity, internet_users_perc "
"from src "
"where year = 2020"
" and code <> 'Region' "
" and internet_users_perc > (select avg(internet_users_perc)
from src where year = 2020) "
"order by 2 desc").show(truncate=False)
```

entity	internet_users_perc
Bahrain	99.66999817
Qatar	99.65284729
Kuwait	99.10588074
Iceland	99
Luxembourg	98.82242584
Saudi Arabia	97.86000061

Рисунок 5 – Результат работы 5 запроса

6. Года, когда в России доля пользователей интернета было ниже среднего по миру

```
spark.sql("with tmp as ("
" select year, avg(internet_users_perc) as "
avg_internet_users_perc "
" from src "
" group by 1) "
"select src.year "
"from src join tmp "
"on src.year = tmp.year "
"where entity = 'Russia' "
" and internet_users_perc < avg_internet_users_perc "
"order by 1").show()
```

year
1990
1991
1992
1993
1994
1995

Рисунок 6 – Результат работы 6 запроса

7. Топ 3 страны по доле пользователей мобильной связи

```
spark.sql("with tmp as (
  select entity, max(mobile_users) as mobile_users
  from src
  where code <> 'Region'
  group by 1)
select src.entity, src.year, src.mobile_users
from src join tmp
on src.entity = tmp.entity and src.mobile_users = tmp.mobile_users

order by 3 desc
limit 3").show(truncate=False)
```

entity	year	mobile_users
Brazil	2018	99.97170258
Mongolia	2013	99.85401154
Uzbekistan	2020	99.75439453

Рисунок 7 – Результат работы 7 запроса

8. Количество стран, у которых доля пользователей интернета была выше 30, по годам

```
spark.sql("select year, count(*) as count_of_countries
from src
where internet_users_perc > 30
and code <> 'Region'
group by 1
order by 1").show()
```

year	count_of_countries
1998	4
1999	12
2000	20
2001	27
2002	29
2003	36

Рисунок 8 – Результат работы 8 запроса

9. Последние страны подключившиеся к интернету

```
spark.sql("with tmp as (
  select entity, max(year) + 1 as connect_year
  from src
  where internet_users_num = 0
  and code <> 'Region'
  group by 1
  having max(year) < (select max(year) from src)
select entity, cast(connect_year as int)
from tmp
where connect_year = (select max(connect_year) from tmp)
order by 1").show()
```

```

+-----+-----+
|entity|connect_year|
+-----+-----+
| Sudan|          2016|
+-----+-----+

```

Рисунок 9 – Результат работы 9 запроса

10. Страны-рекордсменки по доле пользователей интернета по годам

```

spark.sql("select distinct year, "
         "    first_value(entity) over (partition by year order by "
         "internet_users_perc desc) as country, "
         "    max(internet_users_perc) over (partition by year) as "
         "internet_users_perc "
         "from src "
         "order by 1").show(truncate=False)

```

```

+---+-----+-----+-----+
|year|country|internet_users_perc|
+---+-----+-----+-----+
|1990|United States|0.784728527|
|1991|Thailand|5.22E-05|
|1992|South Asia|8.55E-05|
|1993|Norway|2.78399086|
|1994|Iceland|6.794811726|
|1995|United States|9.237088203|
|1996|Netherlands|9.649068832|

```

Рисунок 10 – Результат работы 10 запроса

Вывод: в ходе выполнения лабораторной работы был получен опыт работы со Spark.