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 09.04.01/07 **Интеллектуальные системы анализа, обработки и интерпретации больших данных**

О Т Ч Е Т

по лабораторной работе № 10

Название: Spark

Дисциплина: Языки программирования для работы с большими данными

Студент ИУ6-23М (Группа) _____ (Подпись, дата) Д.В. Авдонин (И.О. Фамилия)

Преподаватель _____ (Подпись, дата) _____ (И.О. Фамилия)

Москва, 2022

Вариант 1:

Сделать выборку данных на ваше усмотрение

Решение:

```
# -*- coding: utf-8 -*-

from pyspark.sql import SparkSession

spark_session = SparkSession.builder\
    .appName('adv-banners-clicks')\
    .config('spark.sql.shuffle.partitions', 5)\
    .getOrCreate()

# spark_session.sparkContext.setLogLevel('WARN')

banners_df = spark_session.read.load('hdfs://localhost:9000/banners.csv', format='csv', inferSchema=True,
header=True)
clicks_df = spark_session.read.load('hdfs://localhost:9000/clicks.csv', format='csv', inferSchema=True,
header=True)

banners_df.createOrReplaceTempView('banners')
clicks_df.createOrReplaceTempView('clicks')

# spark_session.sql('SELECT b.id AS banner_id, c.id AS click_id FROM clicks c JOIN banners b ON
c.banner_id = b.id').show(40)
spark_session.sql("""
    SELECT banners.id AS banner_id, COUNT(banners.id) AS clicks_count
    FROM clicks
    JOIN banners ON clicks.banner_id = banners.id
    GROUP BY banners.id
    ORDER BY clicks_count DESC
""").show()

# Для удобства работы с монитором, чтобы скрипт автоматически не завершал работу
input('Ctrl C')
```

```
# -*- coding: utf-8 -*-

from elasticsearch import Elasticsearch
from pyspark.sql import SparkSession, Row

es_client = Elasticsearch([{'host': '127.0.0.1', 'port': 9200}])

banners = es_client.search(
    index='banners',
    body={"size": 30}
)['hits']['hits']

clicks = es_client.search(
    index='clicks',
    body={"size": 30}
)['hits']['hits']
```

```
# Возвращает тело объекта вместе с его id
def get_data_from_object(obj):
    result = obj['_source']
    result['id'] = obj['_id']
    return result


spark_session = SparkSession.builder.appName('adv-banners-
clicks').getOrCreate()

banners_df = spark_session.createDataFrame(Row(**x) for x in
map(get_data_from_object, banners))
banners_df\
    .withColumn('client_targeting_info',
banners_df['client_targeting_info'].cast('string'))\
    .write.csv(path='hdfs://localhost:9000/banners.csv', mode='overwrite',
header=True)

clicks_df = spark_session.createDataFrame(Row(**x) for x in
map(get_data_from_object, clicks))
clicks_df\
    .withColumn('utm_params', clicks_df['utm_params'].cast('string'))\
    .write.csv(path='hdfs://localhost:9000/clicks.csv', mode='overwrite',
header=True)
```