



**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**

**МАГИСТЕРСКАЯ ПРОГРАММА 09.04.01/12 Интеллектуальный анализ больших  
данных в системах поддержки принятия решений.**

**О Т Ч Е Т**

**по лабораторной работе № 10**

**Вариант № 5**

**Название:** Spark

**Дисциплина:** Языки программирования для работы с большими данными

Студент

ИУ6-23М  
(Группа)

\_\_\_\_\_  
(Подпись, дата)

А.О.Крейденко  
(И.О. Фамилия)

Преподаватель

\_\_\_\_\_  
(Подпись, дата)

П.В. Степанов  
(И.О. Фамилия)

Москва, 2024

**Цель:** ознакомиться с работой Spark в языке программирования Java.

**Задание:** сделать 10 выборки данных по выбранной предметной области

Был выбран датасет с данными об оценках приложений в Google Store.

Код класса Main:

```
package org.example;

import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;
import org.apache.spark.sql.Session;

public class Main {
    public static void main(String[] args) {
        // Инициализация Spark сессии
        Session spark = Session.builder()
            .appName("GooglePlayStoreAnalysis")
            .master("local")
            .getOrCreate();

        // Чтение данных из CSV файла
        Dataset<Row> df = spark.read()
            .option("header", "true")
            .option("inferSchema", "true")
            .csv("googleplaystore.csv");

        // Удаление строк с NaN значениями в любых столбцах
        df = df.na().drop();

        // Создание представления для SQL-запросов
        df.createOrReplaceTempView("googleplaystore");

        // 1. Показать первые 5 строк
        spark.sql("SELECT * FROM googleplaystore LIMIT 5").show();
    }
}
```

// 2. Подсчитать количество приложений для каждой категории

```
spark.sql("SELECT Category, COUNT(*) as count FROM googleplaystore GROUP BY Category ORDER BY count DESC").show();
```

// 3. Найти топ-10 самых популярных приложений по количеству установок

```
spark.sql("SELECT * FROM (SELECT *, CAST(REGEXP_REPLACE(Installs, '[+,]', '' ) AS INT) AS InstallsInt FROM googleplaystore) ORDER BY InstallsInt DESC LIMIT 10").show();
```

// 4. Найти топ-10 бесплатных приложений с наибольшим количеством установок

```
spark.sql("SELECT * FROM (SELECT *, CAST(REGEXP_REPLACE(Installs, '[+,]', '' ) AS INT) AS InstallsInt FROM googleplaystore WHERE Type = 'Free') ORDER BY InstallsInt DESC LIMIT 10").show();
```

// 5. Найти топ-10 платных приложений с наибольшим количеством установок

```
spark.sql("SELECT * FROM (SELECT *, CAST(REGEXP_REPLACE(Installs, '[+,]', '' ) AS INT) AS InstallsInt FROM googleplaystore WHERE Type = 'Paid') ORDER BY InstallsInt DESC LIMIT 10").show();
```

// 6. Подсчитать количество приложений для каждого уровня контента (Content Rating)

```
spark.sql("SELECT `Content Rating`, COUNT(*) as count FROM googleplaystore GROUP BY `Content Rating` ORDER BY count DESC").show();
```

// 7. Подсчитать общее количество установок для каждой категории

```
spark.sql("SELECT Category, SUM(CAST(REGEXP_REPLACE(Installs, '[+,]', '' ) AS INT)) as TotalInstalls FROM googleplaystore GROUP BY Category ORDER BY TotalInstalls DESC").show();
```

// 8. Найти приложения с самой высокой и самой низкой оценкой (по количеству установок)

```

        spark.sql("SELECT      *      FROM      (SELECT      *,
CAST(REGEXP_REPLACE(Installs, '[+,]', '' ) AS INT) AS InstallsInt
FROM googleplaystore) ORDER BY InstallsInt DESC LIMIT 1").show();

        spark.sql("SELECT      *      FROM      (SELECT      *,
CAST(REGEXP_REPLACE(Installs, '[+,]', '' ) AS INT) AS InstallsInt
FROM googleplaystore) ORDER BY InstallsInt ASC LIMIT 1").show();

        // 9. Подсчитать средний размер (Size) приложения для
каждой категории

        spark.sql("SELECT Category, AVG(CASE " +
                "WHEN Size LIKE '%k' THEN CAST(SUBSTRING(Size, 1,
LENGTH(Size)-1) AS FLOAT) / 1024 " +
                "WHEN Size LIKE '%M' THEN CAST(SUBSTRING(Size, 1,
LENGTH(Size)-1) AS FLOAT) " +
                "ELSE NULL END) as AverageSize " +
                "FROM googleplaystore GROUP BY Category ORDER BY
AverageSize DESC").show();

        // 10. Подсчитать общее количество отзывов (Reviews) для
каждой категории

        spark.sql("SELECT Category, SUM(CAST(Reviews AS INT)) as
TotalReviews FROM googleplaystore GROUP BY Category ORDER BY
TotalReviews DESC").show();

        // Остановка Spark сессии

        spark.stop();

    }
}

```

Часть вывода программы показана на рисунке 1.

```

24/06/17 04:49:15 INFO SparkContext: SparkContext is stopping with exitCode 0.
+-----+-----+
|      Category|TotalReviews|
+-----+-----+
|      GAME| 1585422349|
| COMMUNICATION| 815462260|
|      SOCIAL| 621241422|
|      FAMILY| 410226330|
|      TOOLS| 273185044|
| PHOTOGRAPHY| 213516650|
|     SHOPPING| 115041222|
| PRODUCTIVITY| 114116975|
| VIDEO_PLAYERS| 110380188|
| PERSONALIZATION| 89346140|
|      SPORTS| 70830169|
| TRAVEL_AND_LOCAL| 62617919|
| ENTERTAINMENT| 59178154|
| NEWS_AND_MAGAZINES| 54400863|
|      EDUCATION| 39595786|
| HEALTH_AND_FITNESS| 37891234|
| MAPS_AND_NAVIGATION| 30557006|
| BOOKS_AND_REFERENCE| 21959069|
|      FINANCE| 17550728|
|      WEATHER| 14604735|
+-----+-----+
only showing top 20 rows

24/06/17 04:49:16 INFO SparkUI: Stopped Spark web UI at http://192.168.56.1:4040
24/06/17 04:49:16 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
24/06/17 04:49:16 INFO MemoryStore: MemoryStore cleared
24/06/17 04:49:16 INFO BlockManager: BlockManager stopped
24/06/17 04:49:16 INFO BlockManagerMaster: BlockManagerMaster stopped
24/06/17 04:49:16 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
24/06/17 04:49:16 INFO SparkContext: Successfully stopped SparkContext
24/06/17 04:49:16 INFO ShutdownHookManager: Shutdown hook called
24/06/17 04:49:16 INFO ShutdownHookManager: Deleting directory C:\Users\User\AppData\Local\Temp\spark-1a2da24c-78e8-414f-a029-d6e9f8ba9818

Process finished with exit code 0

```

Рисунок 1 – Работа программы

**Вывод:** был изучен Spark в языке программирования Java.