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 № 1 0

Название: Spark

Дисциплина: Языки программирования для работы с большими данными

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

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

Москва, 2023

ЗАДАНИЕ

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

Решение

Для решения задания был выбран датасет еженедельных чартов Spotify. В датасете собраны данные со всех регионов присутствия Spotify. В нём перечислены исполнители, названия композиций, позиции в чартах, а также метрики Spotify, такие как `Danceability`, `Energy` и ряд других.

Создадим сессию Spark.

```
import org.apache.spark.sql.functions._
import org.apache.spark.sql.{SparkSession, DataFrame}
```

```
Intitializing Scala interpreter ...
```

```
Spark Web UI available at http://172.29.26.166:4040
SparkContext available as 'sc' (version = 3.4.0, master = local[*], app id =
local-1684503836092)
SparkSession available as 'spark'
```

```
import org.apache.spark.sql.functions._
import org.apache.spark.sql.{SparkSession, DataFrame}
```

Запуск сессии.

```
val spark = SparkSession.builder()
  .appName("Music")
  .getOrCreate()
```

```
spark: org.apache.spark.sql.SparkSession =
org.apache.spark.sql.SparkSession@5a35a0b3
```

Загрузка датасета.

```
val df = spark.read
  .format("csv")
  .option("header", "true")
  .load("final.csv")
```

```
df.show()
```

```
+---+--------------------+----+--------------+----------+-----------------+-----
---------------+-----------------+--------------------+------+--------------------
+------------+---------------+--------------------+--------------------+---------
+-------------+--------------+-------+----------+-----------------+-------------
```

```
----+----+----+------------------+------------------+-----------------+--------
--------+--------+------------------+------------------+--------+---------+------
-------+--------+-----+
|_c0|                 uri|rank|   artist_names|artists_num|artist_individual|
artist_id|     artist_genre|          artist_img|collab|
track_name|release_date|album_num_tracks|         album_cover|
source|peak_rank|previous_rank|weeks_on_chart|streams|      week|
danceability|            energy| key|mode|           loudness|        speechiness|
acousticness|instrumentalness|liveness|            valence|
tempo|duration|  country|       region|language|pivot|
+---+------------------+----+--------------+-----------+-----------------+-----
--------------+-----------------+--------------------+------+--------------------
+------------+---------------+--------------------+--------------------+---------
+-------------+-------------+-------+----------+-----------------+-------------
----+----+----+------------------+------------------+-----------------+--------
--------+--------+------------------+------------------+--------+---------+------
-------+--------+-----+
|  0|spotify:track:2gp...|   1|   Paulo Londra|        1.0|     Paulo
Londra|spotify:artist:3v...|argentine hip hop|https://i.scdn.co...|     0|
Plan A|  2022-03-23|             1.0|https://i.scdn.co...|          WEA Latina|
1|            1|            4|3003411|2022-04-14|
0.583|0.8340000000000001| 0.0| 1.0|             -4.875|             0.0444|
0.0495|             0.0|  0.0658|              0.557|
173.935|178203.0|Argentina|South America| Spanish|    0|
|  1|spotify:track:2x8...|   2|            WOS|        1.0|
WOS|spotify:artist:5Y...|  argentine indie|https://i.scdn.co...|     0|
ARRANCARMELO|  2022-04-06|             1.0|https://i.scdn.co...|DOGUITO Records
/...|        2|          129|             2|2512175|2022-04-14|             0.654|
0.354| 5.0| 1.0|             -7.358|             0.0738|0.7240000000000001|
0.0|   0.134|              0.262|            81.956|183547.0|Argentina|South
America| Spanish|    0|
|  2|spotify:track:2SJ...|   3|   Paulo Londra|        1.0|     Paulo
Londra|spotify:artist:3v...|argentine hip hop|https://i.scdn.co...|     0|
Chance|  2022-04-06|             2.0|https://i.scdn.co...|          WEA Latina|
3|           59|             2|2408983|2022-04-14|             0.721|
0.463| 1.0| 0.0|             -9.483|             0.0646|             0.241|
0.0|  0.0929|0.21600000000000005|           137.915|204003.0|Argentina|South
America| Spanish|    0|
|  3|spotify:track:1O2...|   5|        Cris Mj|        1.0|          Cris
Mj|spotify:artist:1Y...|   urbano chileno|https://i.scdn.co...|     0|Una Noche en
Mede...|  2022-01-21|             1.0|https://i.scdn.co...|   Nabru Records LLC|
5|            5|             8|2080139|2022-04-14|
0.87|0.5479999999999999|10.0| 0.0| -5.252999999999999|             0.077|
0.0924|         4.6e-05|  0.0534| 0.8320000000000001|
96.018|153750.0|Argentina|South America| Spanish|    0|
|  4|spotify:track:1Tp...|   6|         Emilia|        1.0|
Emilia|spotify:artist:0A...|    pop argentino|https://i.scdn.co...|     0|
cuatro veinte|  2022-03-24|             1.0|https://i.scdn.co...|    Sony Music
Latin|        6|            9|             3|1923270|2022-04-
14|0.7609999999999999|             0.696| 7.0| 0.0|             -3.817|
0.0505|            0.0811|        6.25e-05|   0.101|             0.501|
95.066|133895.0|Argentina|South America| Spanish|    0|
```

```
|  5|spotify:track:4LR...|  11|   Harry Styles|        1.0|     Harry
Styles|spotify:artist:6K...|             pop|https://i.scdn.co...|     0|
As It Was|  2022-03-31|            1.0|https://i.scdn.co...|           Columbia|
6|           6|             2|1555631|2022-04-14|              0.52|
0.731| 6.0| 0.0| -5.337999999999999|             0.0557|             0.342|
0.00101|   0.311|              0.662|            173.93|167303.0|Argentina|South
America| Spanish|    0|
|  6|spotify:track:3Ec...|  17|      La K'onga|        1.0|        La
K'onga|spotify:artist:3g...|         cuarteto|https://i.scdn.co...|     0|
Te Mentiría|  2021-12-09|           14.0|https://i.scdn.co...|        Leader
Music|       14|           16|            47|1272870|2022-04-
14|0.6509999999999999|            0.731| 7.0| 1.0|-6.8889999999999985|
0.0549|             0.116|              0.0|  0.0708|
0.653|153.10399999999996|218431.0|Argentina|South America| Spanish|    0|
|  7|spotify:track:42G...|  20|  Maria Becerra|        1.0|    Maria
Becerra|spotify:artist:1D...|    pop argentino|https://i.scdn.co...|     0|
Felices x Siempre|  2022-02-22|            1.0|https://i.scdn.co...|   300
Entertainment|       11|           15|             8|1149499|2022-04-
14|0.7709999999999999|             0.467| 5.0| 0.0|             -6.653|
0.123|             0.375|         0.00974|   0.112|              0.256|
100.089|199657.0|Argentina|South America| Spanish|    0|
|  8|spotify:track:3Fk...|  23|         Anitta|        1.0|
Anitta|spotify:artist:7F...|         funk pop|https://i.scdn.co...|     0|
Envolver|  2022-04-12|           15.0|https://i.scdn.co...|      Warner Records|
13|           17|            6|1104997|2022-04-14|             0.812|
0.736| 4.0| 0.0|             -5.421|             0.0833|             0.152|
0.00254|  0.0914|              0.396|            91.993|193806.0|Argentina|South
America| Spanish|    0|
|  9|spotify:track:5Us...|  24|     LIT killah|        1.0|       LIT
killah|spotify:artist:1v...|argentine hip hop|https://i.scdn.co...|     0|    La
Trampa es Ley|  2022-02-10|             1.0|https://i.scdn.co...|          WEA
Latina|        2|           21|             9|1076236|2022-04-14|
0.596|              0.71| 6.0| 1.0|             -5.815|              0.136|
0.243|             0.0|   0.204|              0.632|
117.871|141864.0|Argentina|South America| Spanish|    0|
| 10|spotify:track:7on...|  25|      La K'onga|        1.0|        La
K'onga|spotify:artist:3g...|         cuarteto|https://i.scdn.co...|     0|
El Mismo Aire|  2020-11-13|           18.0|https://i.scdn.co...|Leader Music
bajo...|       16|           25|            70|1058521|2022-04-14|
0.762|             0.748| 0.0| 0.0|             -6.273|             0.0407|
0.0746|              0.0|   0.128| 0.8220000000000001|
149.985|209415.0|Argentina|South America| Spanish|    0|
| 11|spotify:track:1Ud...|  26|    Danny Ocean|        1.0|      Danny
Ocean|spotify:artist:5H...|            latin|https://i.scdn.co...|     0|   Fuera
del mercado|  2022-02-17|           16.0|https://i.scdn.co...|    Atlantic
Records|       26|           49|             3|1029069|2022-04-14|
0.453|0.6729999999999999| 8.0| 1.0|            -5.649|0.14400000000000002|
0.32|             0.0|   0.131|              0.266|
92.06|159849.0|Argentina|South America| Spanish|    0|
| 12|spotify:track:7JZ...|  27|           Duki|        1.0|
Duki|spotify:artist:1b...|argentine hip hop|https://i.scdn.co...|     0|
TOP 5|  2021-11-25|            7.0|https://i.scdn.co...|DALE PLAY Records...|
```

```
3|          23|           20|1004471|2022-04-14|
0.853|0.8240000000000001| 2.0| 1.0|            -3.384|             0.214|
0.0943|       2.68e-05|    0.11|             0.693|
100.05|146815.0|Argentina|South America| Spanish|    0|
| 13|spotify:track:3Ga...|  32|     C. Tangana|       1.0|       C.
Tangana|spotify:artist:5T...|   urbano espanol|https://i.scdn.co...|     0|
Demasiadas Mujeres|  2021-02-26|           14.0|https://i.scdn.co...|Sony Music
Entert...|       31|           31|            9| 870219|2022-04-
14|0.6579999999999999|            0.453| 9.0| 0.0| -7.377999999999999|
0.39|            0.131|          0.0002|  0.0848|             0.358|
126.043|153960.0|Argentina|South America| Spanish|    0|
| 14|spotify:track:2cR...|  35|    Salastkbron|       1.0|
Salastkbron|spotify:artist:3W...|    pop argentino|https://i.scdn.co...|     0|
TITAN|  2021-10-14|            1.0|https://i.scdn.co...|        WM Argentina|
2|          29|           20| 838194|2022-04-14|            0.725|
0.49| 5.0| 0.0|            -9.487|             0.187|            0.341|
0.0|  0.0978|              0.96|           180.042|122137.0|Argentina|South
America| Spanish|    0|
| 15|spotify:track:1Pm...|  37|    Salastkbron|       1.0|
Salastkbron|spotify:artist:3W...|    pop argentino|https://i.scdn.co...|     0|
TURROMANTIKO|  2022-01-12|            1.0|https://i.scdn.co...|        WM
Argentina|       28|           33|           11| 819773|2022-04-14|
0.795|             0.58| 7.0| 1.0|            -6.443|             0.132|
0.338|            0.0|   0.163|             0.657|
90.099|131000.0|Argentina|South America| Spanish|    0|
| 16|spotify:track:1kj...|  38|         Trueno|       1.0|
Trueno|spotify:artist:2x...|argentine hip hop|https://i.scdn.co...|     0|
DANCE CRIP|  2021-11-17|            1.0|https://i.scdn.co...|© Ⓟ 2022 Sur
Cap...|        3|           36|            21| 814714|2022-04-14|
0.857|0.7659999999999999| 0.0| 1.0|            -3.699|            0.0978|
0.108|            0.0|  0.0817| 0.8640000000000001|
106.024|165019.0|Argentina|South America| Spanish|    0|
| 17|spotify:track:6yl...|  39|Sebastian Yatra|       1.0|  Sebastian
Yatra|spotify:artist:07...|           latin|https://i.scdn.co...|     0|
Tacones Rojos|  2022-01-28|           17.0|https://i.scdn.co...|       UMLE -
Latino|       21|           38|           25| 805602|2022-04-14|
0.746|0.8440000000000001|11.0| 0.0|            -3.499|            0.0359|
0.062|            0.0|   0.149|             0.934|
123.014|189333.0|Argentina|South America| Spanish|    0|
| 18|spotify:track:1fK...|  40|  Ke Personajes|       1.0|    Ke
Personajes|spotify:artist:06...|               0|https://i.scdn.co...|     0|Si
No Te Tengo / ...|  2021-08-17|           1.0|https://i.scdn.co...|       Ke
Personajes|       29|           37|           31| 803350|2022-04-14|
0.419|            0.711|11.0| 0.0|            -4.083|            0.0488|
0.0513|            0.0|   0.243|             0.529|
82.18|428439.0|Argentina|South America| Spanish|    0|
| 19|spotify:track:74a...|  41|      Tiago PZK|       1.0|        Tiago
PZK|spotify:artist:5Y...|   trap argentino|https://i.scdn.co...|     0|
Hablando De Love|  2022-03-17|            1.0|https://i.scdn.co...|          WEA
Latina|       29|           34|            4| 786293|2022-04-
14|0.6579999999999999|           0.892| 7.0| 0.0|            -4.646|
0.146|            0.459|            0.0|   0.101|             0.413|
```

```
91.032|156090.0|Argentina|South America| Spanish|    0|
+---+------------------+----+--------------+----------+-----------------+-----
--------------+----------------+-------------------+------+------------------
+------------+---------------+--------------------+--------------------+---------
+-------------+-------------+-------+---------+-----------------+-------------
----+----+----+------------------+------------------+-----------------+--------
--------+--------+------------------+-----------------+--------+---------+------
-------+--------+-----+
only showing top 20 rows


df: org.apache.spark.sql.DataFrame = [_c0: string, uri: string ... 34 more fields]
```

Топ-10 исполнителей по общему количеству прослушиваний.

```
df.select("artist_names", "track_name", "streams")
  .groupBy("artist_names")
  .agg(sum("streams").alias("total_streams"))
  .orderBy(desc("total_streams"))
  .limit(10)
  .show()
```

```
+--------------------+--------------+
|        artist_names|  total_streams|
+--------------------+--------------+
|          Ed Sheeran| 1.364386947E10|
|      Olivia Rodrigo|1.1741091136E10|
|           Bad Bunny|1.0379978355E10|
|       Billie Eilish|  9.870341236E9|
|          The Weeknd|  9.485489222E9|
|        Harry Styles|  9.124389514E9|
|       Ariana Grande|  8.190436653E9|
|               Drake|  7.812727326E9|
|The Kid LAROI, Ju...|   7.35161562E9|
|         Post Malone|  7.283596733E9|
+--------------------+--------------+
```

Топ 5 песен с наибольшим количеством недель в чартах по всему миру.

```
df.filter(col("country") === "Global")
  .orderBy(desc("weeks_on_chart"))
  .select("track_name", "artist_names", "weeks_on_chart")
  .limit(5)
  .show()
```

```
+--------------------+--------------------+--------------+
|          track_name|       artist_names|weeks_on_chart|
+--------------------+--------------------+--------------+
|          High Hopes|Panic! At The Disco|            99|
|when the party's ...|      Billie Eilish|            99|
|      Mr. Brightside|         The Killers|            99|
|             Heather|          Conan Gray|            99|
```

```
|          goosebumps|       Travis Scott|            99|
+--------------------+-------------------+--------------+
```

Средняя `danceability`, `energy`, и `valence` песен, выпущенных в 2022 году.

```
df.filter(year(col("release_date")) === 2022)
  .agg(avg("danceability").alias("avg_danceability"),
       avg("energy").alias("avg_energy"),
       avg("valence").alias("avg_valence"))
       .show()
```

```
+------------------+------------------+------------------+
|  avg_danceability|        avg_energy|       avg_valence|
+------------------+------------------+------------------+
|0.7109250632167036|0.6633135773008008|0.5631852331667105|
+------------------+------------------+------------------+
```

Топ 10 исполнителей с наибольшим средним количством прослушиваний в неделю.

```
df.groupBy("artist_names")
  .agg(avg("streams").alias("avg_streams"))
  .orderBy(desc("avg_streams"))
  .limit(10)
  .show()
```

```
+--------------------+--------------------+
|        artist_names|         avg_streams|
+--------------------+--------------------+
|StaySolidRocky, L...|        1.75541045E7|
|Luis Fonsi, Daddy...|   1.7091451234375E7|
|Anuel AA, Daddy Y...|1.4687811955555556E7|
|Taylor Swift, The...|         1.4385703E7|
| Nicky Jam, J Balvin|1.4219319205882354E7|
|DJ Khaled, Justin...|1.4134057901960785E7|
|Lil Nas X, Billy ...| 1.366073164864865E7|
|DJ Khaled, Justin...| 1.362010794736842E7|
|Drake, Michael Ja...|     1.34909444375E7|
|6ix9ine, Nicki Mi...|        1.34590289E7|
+--------------------+--------------------+
```

Наиболее популярные жанры исполнителей в датасете.

```
df.groupBy("artist_genre")
  .count()
  .orderBy(desc("count"))
  .show()
```

```
+--------------------+------+
|        artist_genre| count|
+--------------------+------+
|                 pop|156910|
|                   0|114293|
|         trap latino| 86138|
|           reggaeton| 73498|
|               latin| 73176|
|           dance pop| 49390|
|                 rap| 31621|
|              uk pop| 27131|
|        canadian pop| 20416|
|               k-pop| 19415|
|reggaeton colombiano| 19321|
|           latin pop| 16922|
|      reggaeton flow| 16003|
|             hip hop| 15572|
|       latin hip hop| 14813|
|             pop rap| 14719|
|                trap| 14044|
|      trap argentino| 13080|
|      german hip hop| 12529|
|           pop dance| 12021|
+--------------------+------+
only showing top 20 rows
```

Среднее количество прослушиваний в неделю для каждой страны.

```
df.groupBy("country")
  .agg(avg("streams").alias("avg_streams_per_week"))
  .orderBy(desc("avg_streams_per_week"))
  .show()
```

```
+--------------+--------------------+
|       country|avg_streams_per_week|
+--------------+--------------------+
|        Global|   8590443.78178008|
| United States|  2612468.3299010973|
|        Brazil|  1541791.7800724516|
|        Mexico|  1317617.0092200846|
|       Germany|    784578.343812709|
|         Spain|    713261.081671159|
|         India|   709459.8710540024|
|United Kingdom|   696611.6486762615|
|         Italy|   624623.5501046622|
|        France|   584984.9326468467|
|     Argentina|   547315.6179943004|
|     Indonesia|   542640.4211397746|
|        Turkey|  502700.14107988315|
|         Chile|   456507.3099201683|
|   Philippines|  455304.44186168537|
```

```
|         Japan|   437043.5067144136|
|     Australia|  394437.60274980456|
|        Canada|   385216.4609696333|
|   Netherlands|   369540.7430649392|
|        Poland|   356572.3758571201|
+--------------+--------------------+
only showing top 20 rows
```

Общее количство уникальных исполнителей для каждого региона.

```
df.groupBy("region")
  .agg(countDistinct("artist_names").alias("unique_artists"))
  .orderBy(desc("unique_artists"))
  .show()
```

```
+---------------+--------------+
|         region|unique_artists|
+---------------+--------------+
|         Europe|         11085|
|           Asia|          3587|
|         Global|          2500|
|  South America|          2051|
|    Middle East|          2030|
|         Africa|          1796|
|  North America|          1502|
|Central America|           975|
|        Oceania|           812|
|      Caribbean|           663|
|        Ukraine|             2|
|       Paraguay|             1|
|           0.11|             1|
|         Taiwan|             1|
|         207394|             1|
|  United States|             1|
|         region|             1|
|         Mexico|             1|
|      Guatemala|             1|
|       Honduras|             1|
+---------------+--------------+
only showing top 20 rows
```

Средняя продолжительность песен на каждом языке.

```
df.groupBy("language")
  .agg(avg("duration").alias("avg_duration"))
  .orderBy(desc("avg_duration"))
  .show()
```

```
+----------+------------------+
|  language|      avg_duration|
```

```
+----------+------------------+
|      Zulu| 277846.9191161834|
|  Japanese|236926.38478066248|
|  Mandarin|229092.00780014182|
|   Spanish| 229090.1674511652|
| Cantonese| 225446.2013226366|
|Indonesian|223818.99283614007|
|   Tagalog|223522.45890338876|
|Vietnamese|222888.56846780164|
|     Hindi|222448.41670428895|
|Portuguese| 220139.6228857533|
|      Urdu|219468.59743918054|
|      Thai|216450.96938240537|
|    Global|211137.86207397297|
|    Korean|209461.11974094732|
| Icelandic|205727.16012183693|
|   English|205184.59067171466|
|     Malay|204851.87832109997|
|    Arabic|204682.33773191096|
|    French|203644.29232546827|
|    Hebrew|202510.22794735368|
+----------+------------------+
only showing top 20 rows
```

Наиболее частые источники песен (лэйблы) в датасете.

```
df.groupBy("source")
  .count()
  .orderBy(desc("count"))
  .show()
```

```
+--------------------+-----+
|              source|count|
+--------------------+-----+
|            Columbia|79751|
|Rimas Entertainme...|59518|
|    Sony Music Latin|49662|
|    Republic Records|40146|
| Atlantic Records UK|32368|
|    Atlantic Records|27894|
|       UMLE - Latino|24925|
|          WEA Latina|23544|
|      Warner Records|23409|
|Sony Music Entert...|22119|
|Kemosabe Records/...|21062|
|   RCA Records Label|20477|
|          Rich Music|18672|
|        RBMG/Def Jam|18278|
|   Olivia Rodrigo PS|16626|
|                 OVO|12814|
|Sony Music Latin/...|12369|
```

```
|     Polydor Records|12184|
|Ministry of Sound...|11942|
|                 EMI|11259|
+--------------------+-----+
only showing top 20 rows
```

Топ-10 исполнителей по числу коллабораций.

```
df.filter(col("collab") === 1)
  .groupBy("artist_names")
  .agg(countDistinct("track_name").alias("collaborations"))
  .orderBy(desc("collaborations"))
  .limit(10)
  .show()
```

```
+--------------------+--------------+
|        artist_names|collaborations|
+--------------------+--------------+
|          Ego, Coder|            20|
|       Bedoes, Lanek|            15|
|Marília Mendonça,...|            15|
|Big Baby Tape, ki...|            14|
|Kaaris, Kalash Cr...|            14|
|Benny Jamz, Gilli...|            12|
|      Azahriah, DESH|            12|
|     Anuel AA, Ozuna|            12|
|AKC Misi, AKC Kretta|            11|
|       Stormy, Tagne|            11|
+--------------------+--------------+
```

Закрытие сессии Spark.

```
spark.stop()
```

ВЫВОДЫ

Изучены принципы работы со Spark на языке Scala. Изучены способы построения поисковых запросов с помощью DSL библиотеки Spark.

Изучены способы использования Scala в среде Jupyter notebook.