```
/*
chcp 65001 && spark-shell -i \Users\Ekaterina\ETL\HW_2\HW_2_fifa.scala --conf
"spark.driver.extraJavaOptions=-Dfile.encoding=utf-8"
*/
import org.apache.spark.internal.Logging
import org.apache.spark.sql.functions.{col, collect_list, concat_ws}
import org.apache.spark.sql.{DataFrame, SparkSession}
import org.apache.spark.sql.functions._
import org.apache.spark.sql.expressions.Window
val t1 = System.currentTimeMillis()
val misqlcon = "jdbc:mysql://localhost:3306/spark?user=root&password=24082019:jhf"
val driver = "com.mysql.cj.jdbc.Driver"
if(1==1){
var df1  = spark.read
       .option("header", "true")
       .option("delimiter", ",")
                        .option("encoding", "utf-8")
       .csv("/Users/Ekaterina/ETL/HW_2/fifa_s2.csv")

                        df1=df1
            .withColumn("ID",col("ID").cast("int")).dropDuplicates()
                        .withColumn("Age",col("Age").cast("int"))
                        .withColumn("Overall",col("Overall").cast("int"))
                        .withColumn("Potential",col("Potential").cast("int"))
                        .withColumn("Value",col("Value").cast("float"))
                        .withColumn("Wage",col("Wage").cast("float"))
                        .withColumn("International Reputation",col("International
Reputation").cast("float"))
                        .withColumn("Skill Moves",col("Skill Moves").cast("float"))
                        .withColumn("Height",col("Height").cast("float"))
                        .withColumn("Weight",col("Weight").cast("float"))
                        .withColumn("Release Clause",col("Release Clause").cast("float"))

       .withColumn("Age_category",
                        when(col("Age") <= 20, "before 20 years")
                        .when(col("Age").between(20, 30), "20-30 years")
                        .when(col("Age").between(30, 36), "30-36 years")
                        .otherwise("after 36 years"))

            df1.write.format("jdbc")
            .option("url", misqlcon)
        .option("driver", driver)
            .option("dbtable", "HW_2_task_3")
   .option("user", "root")
   .option("password", "24082019:jhf")
   .option("charset", "utf8mb4_general_ci")
   .mode("overwrite")
   .save()

            df1.show()

val s = df1.columns.map(c => sum(col(c).isNull.cast("integer")).alias(c))
val df2 = df1.agg(s.head, s.tail:_*)
val t = df2.columns.map(c => df2.select(lit(c).alias("col_name"), col(c).alias("null_count")))
val df_agg_col = t.reduce((df1, df2) => df1.union(df2))
df_agg_col.show()
}

val s0 = (System.currentTimeMillis() - t1)/1000
val s = s0 % 60
```

```
val m = (s0/60) % 60
val h = (s0/60/60) % 24
println("%02d:%02d:%02d".format(h, m, s))


System.exit(0)
```

Администратор: Командная строка

```
Microsoft Windows [Version 10.0.19045.4170]
(c) Корпорация Майкрософт (Microsoft Corporation). Все права защищены.

C:\WINDOWS\system32>chcp 65001 && spark-shell -i \Users\Ekaterina\ETL\HW_2\HW_2_fifa.scala --conf "spark.driver.extraJavaOptions=-Dfile.encoding=utf-8"
Active code page: 65001
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://MSI:4040
Spark context available as 'sc' (master = local[*], app id = local-1712206627821).
Spark session available as 'spark'.
24/04/04 14:57:20 WARN ProcfsMetricsGetter: Exception when trying to compute pagesize, as a result reporting of ProcessTree metrics is stopped
+------+--------------+---+------------------+-------+---------+--------------------+-------+-----+------------+----------------------+----------+--------+------+--------------------+---------+------+----------+---------------+
|    ID|          Name|Age|       Nationality|Overall|Potential|                Club|  Value| Wage|Preferred Foot|International Reputation|Skill Moves|Position|Joined|Contract Valid Until|   Height|Weight|Release Clause|   Age_category|
+------+--------------+---+------------------+-------+---------+--------------------+-------+-----+------------+----------------------+----------+--------+------+--------------------+---------+------+----------+---------------+
|167397|        Falcao| 32|          Colombia|     84|       84|          AS Monaco|25000.0|115.0|       Right|                   3.0|       3.0|      RS|  2013|          01/01/2020|5.8333335| 159.0|   47500.0|    30-36 years|
|171189|Aythami Artiles| 32|             Spain|     72|       72|         Córdoba CF| 1800.0|  7.0|       Right|                   1.0|       2.0|     LCB|  2018|          01/01/2020|      6.0| 172.0|    2800.0|    30-36 years|
|180903|     J. O'Shea| 29|Republic of Ireland|     65|       65|              Bury|  525.0|  5.0|       Right|                   1.0|       3.0|     RCM|  2017|          01/01/2019|      6.0| 154.0|     919.0|    20-30 years|
|183565|       A. Feick| 30|           Germany|     70|       70|1. FC Heidenheim ...| 1300.0|  7.0|        Left|                   1.0|       2.0|      LB|  2015|          01/01/2020|5.8333335| 174.0|    2000.0|    20-30 years|
|192431|    Mario Ortíz| 29|             Spain|     68|       68|    CF Reus Deportiu|  850.0|  4.0|       Right|                   1.0|       2.0|     LCM|  2018|          01/01/2020|5.8333335| 143.0|    1300.0|    20-30 years|
|193116|    M. Gonalons| 29|            France|     75|       75|         Sevilla FC| 5000.0| 47.0|       Right|                   3.0|       3.0|     CDM|  2016|          30/06/2019|6.1666665| 168.0|  4585.061|    20-30 years|
|196942|    Iñigo Pérez| 30|             Spain|     74|       74|         CA Osasuna| 4200.0|  9.0|        Left|                   1.0|       2.0|     LDM|  2018|          01/01/2022|5.8333335| 163.0|    6700.0|    20-30 years|
|201862|        M. Rojo| 28|         Argentina|     79|       79|   Manchester United|10000.0|115.0|        Left|                   3.0|       3.0|      CB|  2014|          01/01/2021|6.1666665| 176.0|   18500.0|    20-30 years|
|205469|  S. Kverkvelia| 26|           Georgia|     76|       80|    Lokomotiv Moscow| 8500.0|  1.0|       Right|                   1.0|       2.0|      CB|  2017|          01/01/2021|6.4166665| 192.0|   18500.0|    20-30 years|
|213823|       C. Rojas| 32|             Chile|     64|       64|      CD Antofagasta|  260.0|  1.0|       Right|                   1.0|       2.0|      CB|  2013|          01/01/2018|     5.75| 163.0|     351.0|    30-36 years|
|220677|     J. Morrell| 21|             Wales|     62|       71|       Bristol City|  475.0|  5.0|       Right|                   1.0|       3.0|      CM|  2014|          01/01/2019|      5.5| 141.0|     998.0|    20-30 years|
|222973|     M. Diakota| 27|            France|     64|       64|         AS Béziers|  475.0|  2.0|       Right|                   1.0|       3.0|      CM|  2017|          01/01/2023|5.9166665| 159.0|     831.0|    20-30 years|
|224933|     K. Lasagna| 25|             Italy|     77|       80|           Udinese|11500.0| 29.0|        Left|                   2.0|       3.0|      ST|  2017|          01/01/2020|6.0833335| 176.0|   21600.0|    20-30 years|
|235957|      J. Gaspar| 21|            France|     63|       75|         AS Monaco|  575.0|  7.0|       Right|                   1.0|       2.0|      RB|  2017|          01/01/2020|5.6666665| 128.0|    1200.0|    20-30 years|
|241160|   G. McEachran| 17|           England|     59|       81|           Chelsea|  350.0|  6.0|        Left|                   1.0|       3.0|     CAM|  2017|          01/01/2019|     5.75| 143.0|     998.0|before 20 years|
|242781|    R. Blumberg| 19|         Australia|     58|       72|   Charlton Athletic|  250.0|  1.0|        Left|                   1.0|       2.0|      CB|  2017|          01/01/2021|5.5833335| 143.0|     594.0|before 20 years|
```

Администратор: Командная строка

```
    998.0|    20-30 years|
|222973|     M. Diakota| 27|           France|     64|     64|         AS Béziers|  475.0|  2.0|       Right|                1.0|       3.0|     CM|  2017|         01/01/2023|5.9166665| 159.0|
    831.0|    20-30 years|
|224933|     K. Lasagna| 25|            Italy|     77|     80|           Udinese|11500.0| 29.0|        Left|                2.0|       3.0|     ST|  2017|         01/01/2020|6.0833335| 176.0|
  21600.0|    20-30 years|
|235957|      J. Gaspar| 21|           France|     63|     75|          AS Monaco|  575.0|  7.0|       Right|                1.0|       2.0|     RB|  2017|         01/01/2020|5.6666665| 128.0|
   1200.0|    20-30 years|
|241160|   G. McEachran| 17|          England|     59|     81|            Chelsea|  350.0|  6.0|        Left|                1.0|       3.0|    CAM|  2017|         01/01/2019|     5.75| 143.0|
    998.0|before 20 years|
|242781|    R. Blumberg| 19|        Australia|     58|     72| Charlton Athletic|  250.0|  1.0|        Left|                1.0|       2.0|     CB|  2017|         01/01/2021|5.5833335| 143.0|
    594.0|before 20 years|
|243009|   D. Cornelius| 20|           Canada|     64|     74|               null|   null|  0.0|        Left|                1.0|       2.0|     CB|  2016|               null|6.1666665| 172.0|
  4585.061|before 20 years|
|211361|     D. Cataldi| 23|            Italy|     75|     79|             Lazio| 8500.0| 35.0|       Right|                2.0|       3.0|     CM|  2011|         01/01/2020|5.9166665| 154.0|
  15100.0|    20-30 years|
|215477|   I. Kecojević| 30|       Montenegro|     67|     67|           Cádiz CF|  575.0|  3.0|       Right|                1.0|       2.0|     CB|  2017|         01/01/2020|     6.25| 187.0|
    920.0|    20-30 years|
|224251|     R. Zentner| 23|          Germany|     69|     76|    1. FSV Mainz 05| 1100.0|  8.0|       Right|                1.0|       1.0|     GK|  2017|         01/01/2021|6.3333335| 212.0|
   2200.0|    20-30 years|
+------+--------------+---+-----------------+-------+-------+-------------------+-------+-----+------------+--------------------+----------+-------+------+-------------------+---------+------+---
---------+---------------+
only showing top 20 rows

+-------------------+----------+
|           col_name|null_count|
+-------------------+----------+
|                 ID|         0|
|               Name|         0|
|                Age|         0|
|        Nationality|         7|
|            Overall|         0|
|          Potential|         0|
|               Club|       241|
|              Value|       257|
|               Wage|         0|
|      Preferred Foot|         0|
|International Rep...|        10|
|        Skill Moves|         6|
|           Position|        21|
|             Joined|         0|
|Contract Valid Until|       273|
|             Height|         0|
|             Weight|         0|
|      Release Clause|         3|
|       Age_category|         0|
+-------------------+----------+

00:00:13

C:\WINDOWS\system32>
```

Unnamed\spark\hw_2_task_3 - HeidiSQL 12.6.0.6765

Файл Редактировать Поиск Запрос Инструменты Переход Помощь

Donate

Фильтр баз д Фильтр табли Хост: 127.0.0.1 База данных: spark Таблица: hw_2_task_3 Данные Запрос*

spark.hw_2_task_3: 2 421 строк (приблизительно), показано 1 000 Далее Показать все Сортировка Столбцы (19/19) Фильтр

- lecture_5
- lecture_6
- mysql
- performan...
- sakila
- seminar_1
- seminar_2
- seminar_3
- seminar_4
- seminar_5 0 B
- seminar_6
- **spark** 5,0 MiB
  - 1 nf 16,0 KiB
  - 2 nf_1 16,0 KiB
  - 2 nf_2 16,0 KiB
  - 3 nf_1 16,0 KiB
  - 3 nf_2 16,0 KiB
  - 3 nf_3 16,0 KiB
  - 3 nf_4 16,0 KiB
  - **hw_2_ta...** 384,0 KiB
  - task4 16,0 KiB
  - tasketl1a 16,0 KiB
  - tasketl2b 4,5 MiB
- sys
- tps_db
- tps_db_var_1
- world

#	erall	Potential	Club	Value	Wage	Preferred Foot	International Reputation	Skill Moves	Position	Joined	Contract Valid Until	Height	Weight	Release Clause	Age_category
129	67	70	CD Tondela	800	2	Left	1	2	CDM	2017	01/01/2019	6,16667	187	1 700	20-30 years
130	68	71	Brescia	950	2	Left	1	3	LB	2018	01/01/2019	6	165	1 500	20-30 years
131	53	62	Changchun Yatai FC	90	1	Right	1	2	CM	2015	01/01/2018	6	161	212	20-30 years
132	69	79	(NULL)	(NULL)	0	Right	1	2	CB	2016	(NULL)	5,66667	159	4 585,06	20-30 years
133	77	78	Burnley	10 500	56	Right	1	2	ST	2017	01/01/2021	6,16667	203	20 700	20-30 years
134	77	79	Arsenal	9 000	76	Right	1	3	RDM	2016	01/01/2022	5,91667	163	17 800	20-30 years
135	57	65	Cardiff City	160	5	Left	1	2	ST	2018	01/01/2020	6,25	187	352	20-30 years
136	64	64	Chapecoense	425	4	Right	1	2	CAM	2018	01/01/2021	5,91667	165	808	20-30 years
137	61	70	Brescia	400	1	Right	1	2	ST	2017	01/01/2020	5,91667	161	700	before 20 years
138	67	83	FC Barcelona	1 600	21	Right	1	3	ST	2017	01/01/2021	6	161	4 200	before 20 years
139	68	73	1. FC Nürnberg	1 200	9	Left	1	3	ST	2017	01/01/2021	6	174	2 300	20-30 years
140	64	69	CD Tondela	600	1	Right	1	3	CM	2016	01/01/2021	6	161	1 300	20-30 years
141	60	71	AS Béziers	325	1	Left	1	3	LM	2018	01/01/2021	5,83333	157	634	before 20 years
142	67	79	Real Madrid	1 200	24	Right	1	2	CB	2016	01/01/2019	6,16667	170	2 700	20-30 years
143	63	71	CD Tenerife	525	1	Right	1	3	ST	2018	01/01/2020	5,83333	152	945	before 20 years
144	83	83	Manchester United	24 500	160	Left	3	4	RM	2014	(NULL)	5,58333	139	45 300	20-30 years
145	72	74	Bristol City	3 000	27	Left	1	2	CB	2017	01/01/2021	6,16667	181	6 000	20-30 years
146	84	86	Inter	(NULL)	88	Right	3	2	CB	2018	01/01/2023	6,16667	172	55 900	20-30 years
147	69	69	(NULL)	(NULL)	0	Right	1	2	LCB	2016	(NULL)	5,91667	163	4 585,06	30-36 years
148	68	75	CD Everton de Viña del Mar	1 300	11	Right	1	3	ST	2016	31/12/2018	6	159	4 585,06	20-30 years
149	61	67	Bury	290	2	Right	1	2	RCB	2018	01/01/2020	5,83333	154	566	20-30 years
150	74	81	Lokomotiv Moscow	8 000	1	Right	1	3	CAM	2012	01/01/2019	6	159	18 400	20-30 years
151	64	66	CD Aves	550	3	Right	1	2	CM	2018	01/01/2023	5,75	150	1 200	20-30 years
152	67	75	CD Huachipato	1 100	2	Right	1	3	CM	2015	01/01/2023	5,75	157	1 800	20-30 years
153	62	68	AS Béziers	425	1	Right	1	2	LM	2017	01/01/2023	6	176	829	20-30 years
154	69	76	(NULL)	(NULL)	0	Right	1	2	RCM	2016	(NULL)	5,83333	157	4 585,06	20-30 years
155	62	81	Brighton & Hove Albion	625	4	Right	1	2	ST	2017	01/01/2019	5,83333	170	1 600	before 20 years

Фильт Регулярное выражение

```
110 USE `spark`;
111 SELECT * FROM `information_schema`.`COLUMNS` WHERE TABLE_SCHEMA='spark' AND TABLE_NAME='hw_2_task_3' ORDER BY ORDINAL_POSITION;
112 SHOW INDEXES FROM `hw_2_task_3` FROM `spark`;
113 SELECT * FROM information_schema.REFERENTIAL_CONSTRAINTS WHERE   CONSTRAINT_SCHEMA='spark'   AND TABLE_NAME='hw_2_task_3'   AND REFERENCED_TABLE_NAME IS NOT NULL;
114 SELECT * FROM information_schema.KEY_COLUMN_USAGE WHERE   TABLE_SCHEMA='spark'   AND TABLE_NAME='hw_2_task_3'   AND REFERENCED_TABLE_NAME IS NOT NULL;
115 SHOW CREATE TABLE `spark`.`hw_2_task_3`;
116 SELECT * FROM `spark`.`hw_2_task_3` LIMIT 1000;
117 SHOW TABLE STATUS LIKE 'hw_2_task_3';
```

r1 : c2 Подключено: 00:54 h MySQL 5.7.40 Время работы: 02:54 h Серверное время: Ожидание.

Unnamed\spark\hw_2_task_3\ - HeidiSQL 12.6.0.6765

Файл Редактировать Поиск Запрос Инструменты Переход Помощь

Donate

Фильтр баз д Фильтр табли

Хост: 127.0.0.1 | База данных: spark | Таблица: hw_2_task_3 | Данные | Запрос*

```
1  SELECT Age_category, COUNT(*) AS count_value
2  FROM spark.hw_2_task_3
3  WHERE Age_category IN ('before 20 years', '20-30 years', '30-36 years', 'after 36 years')
4  GROUP BY Age_category
```

Filter ...

- Столбцы hw_2_tas...
- Функции SQL
- Ключевые слова ...
- Заготовки
- История запросов
- Профилирова...
- Привязать пар...

- lecture_5
- lecture_6
- mysql
- performan...
- sakila
- seminar_1
- seminar_2
- seminar_3
- seminar_4
- seminar_5 0 B
- seminar_6
- **spark** 5,0 MiB
  - 1 nf 16,0 KiB
  - 2 nf_1 16,0 KiB
  - 2 nf_2 16,0 KiB
  - 3 nf_1 16,0 KiB
  - 3 nf_2 16,0 KiB
  - 3 nf_3 16,0 KiB
  - 3 nf_4 16,0 KiB
  - **hw_2_ta...** 384,0 KiB
  - task4 16,0 KiB
  - tasketl1a 16,0 KiB
  - tasketl2b 4,5 MiB
- sys
- tps_db
- tps_db_var_1
- world

hw_2_task_3 (4r × 2c)

#	Age_category	count_value
1	20-30 years	1 633
2	30-36 years	340
3	after 36 years	15
4	before 20 years	411

Фильт Регулярное выражение

```
123 SHOW TABLE STATUS LIKE 'hw_2_task_3';
124 SELECT * FROM `spark`.`hw_2_task_3` ORDER BY `Age_category` ASC LIMIT 1000;
125 SHOW TABLE STATUS LIKE 'hw_2_task_3';
126 SELECT * FROM `spark`.`hw_2_task_3` ORDER BY `Age_category` DESC LIMIT 1000;
127 SHOW TABLE STATUS LIKE 'hw_2_task_3';
128 SHOW CREATE TABLE `spark`.`1 nf`;SHOW CREATE TABLE `spark`.`2 nf_1`;SHOW CREATE TABLE `spark`.`2 nf_2`;SHOW CREATE TABLE `spark`.`3 nf_1`;SHOW CREATE TABLE `spark`.`3 nf_2`;SHOW CREATE TABLE `spark`.`3 nf_3`;SHOW CREATE TABLE `spa
129 SELECT Age_category, COUNT(*) AS count_value  FROM spark.hw_2_task_3  WHERE Age_category IN ('before 20 years', '20-30 years', '30-36 years', 'after 36 years')  GROUP BY Age_category;
130 /* Затронуто строк: 0  Найденные строки: 4  Предупреждения: 0  Длительность  1 запрос: 0,015 сек. */
```