UJIAN TENGAH SEMESTER GANJIL TA. 2022 / 2023

Fakultas	: Ilmu Komputer
Program Studi	: Informatika
Mata Kuliah dan kode mk	: BIGDATA & PREDICTIVE ANALYTICS LANJUT (ST153)
Sifat Ujian	: Buku terbuka
Tanggal Ujian	: 26 November 2022, 08.00
Dosen Pengampu	: Arif Dwi Laksito, M.Kom

NAMA	**: FATHURRAHMAN NUR AZIZ**
NIM	**: 20.11.3694**
KELAS	**: 20 IF 07**

Dengan dataset yang didapat dari https://www.kaggle.com/datasets/artimous/complete-fifa-2017-player-dataset-global/code?select=FullData.csv, lakukan:

1. Setup PySpark pada Google Colab (5, SCPMK 1534002)
2. Baca data dari FullData.csv, kemudian pecah setidaknya menjadi 2 RDD (players, clubs) (10, SCPMK 1534104)
3. Ubah RDD di atas menjadi PySpark Dataframe (5, SCPMK 1532205)
4. Lakukan agregasi untuk mendapatkan jumlah pemain untuk masing-masing kelompok berikut: (25, SCPMK 1534104)
    a. Berdasarkan Nationality
    b. Berdasarkan Club
5. Lakukan agregasi untuk mendapatkan nilai: (25, SCPMK 1534104)
    a. Rata-rata Rating berdasarkan Club_Position
    b. Maximum Rating berdasarkan National_Position
6. Lakukan setidaknya 3 operasi transform RDD pada data yang kalian punya, contoh reduceByKey, sortByKey, flatMap, etc. (30, SCPMK 1532206)

Setiap codeblock harus disertakan penjelasan proses yang terjadi/dilakukan.

Kumpulkan hasil pekerjaan notebook dalam format PDF.

Jawab

1

SOAL NO 1

Setup PySpark pada Google Colab (5, SCPMK 1534002)

```
[ ]  #install JDK
     !apt-get install openjdk-8-jdk-headless -qq > /dev/null

     #extract spark-hadoop from drive
     !tar xf "/content/drive/MyDrive/spark-3.0.0-bin-hadoop3.2.gz"
```

```
[ ]  #install findspark from pip
     !pip install -q findspark
```

```
[ ]  #setup env for JDK and SPARK
     import os
     os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
     os.environ["SPARK_HOME"] = "/content/spark-3.0.0-bin-hadoop3.2"
```

```
[ ]  #install pyspark from pip
     !pip install -q pyspark
```

```
     |████████████████████████████████| 281.4 MB 53 kB/s
     |████████████████████████████████| 199 kB 71.2 MB/s
  Building wheel for pyspark (setup.py) ... done
```

```
[1]  #importing SparkSession from pyspark.sql library
     from pyspark.sql import SparkSession

     #create new spark session with object
     spark = SparkSession.builder.getOrCreate()
```

```
[2]  #create sparkcontext variable.
     sc = spark.sparkContext
```

2

SOAL NO 2

Baca data dari FullData.csv, kemudian pecah setidaknya menjadi 2 RDD (players, clubs) (10, SCPMK 1534104)

```
[27]  1  #Membaca Dataset FullData.csv
      2  FD = spark.read.csv("FullData.csv", header=True, inferSchema=True)
      3
      4  #Menampilkan 5 baris teratas pada Dataset
      5  FD.show(5)
```

```
+----------------+-----------+----------------+------------+------------+-------------+--------+--------
|            Name|Nationality|National_Position|National_Kit|        Club|Club_Position|Club_Kit|Club_Joi
+----------------+-----------+----------------+------------+------------+-------------+--------+--------
|Cristiano Ronaldo|   Portugal|              LS|         7.0| Real Madrid|           LW|     7.0|  07/01/
|     Lionel Messi|  Argentina|              RW|        10.0|FC Barcelona|           RW|    10.0|  07/01/
|          Neymar|     Brazil|              LW|        10.0|FC Barcelona|           LW|    11.0|  07/01/
|      Luis Suárez|    Uruguay|              LS|         9.0|FC Barcelona|           ST|     9.0|  07/11/
|     Manuel Neuer|    Germany|              GK|         1.0|   FC Bayern|           GK|     1.0|  07/01/
+----------------+-----------+----------------+------------+------------+-------------+--------+--------
only showing top 5 rows
```

```
      1  #Menambahkan kolom id pada dataset
      2
      3  #import library monotonically_increasing_id untuk membuat index secara auto increment
      4  from pyspark.sql.functions import monotonically_increasing_id
      5
      6  #menambahkan kolom id dengan value auto increment pada dataset
      7  FD = FD.select("*").withColumn("id", monotonically_increasing_id())
      8
      9  #melihat apakah kolom id sudah ditambahkan pada kolom terakhir
     10  FD.show(1)
```

```
--------+-------+-------+-------+-------+----------+---------+----------+-----+-----------------+---------
Strength|Balance|Agility|Jumping|Heading|Shot_Power|Finishing|Long_Shots|Curve|Freekick_Accuracy|Penalties
--------+-------+-------+-------+-------+----------+---------+----------+-----+-----------------+---------
      80|     63|     90|     95|     85|        92|        93|        90|   81|               76|       85
--------+-------+-------+-------+-------+----------+---------+----------+-----+-----------------+---------
```

```
[23]  1  #memecah dataset player dengan memanggil kolom yang sesuai
      2  players = FD['id','Name','Rating','Height','Weight','Preffered_Foot','Birth_Date','Age']
      3
      4  #membuat RDD dari dataset player
      5  players_rdd = players.rdd
      6
      7  #menampilan 5 data teratas dari players_rdd
      8  players_rdd.take(5)
```

```
[Row(id=0, Name='Cristiano Ronaldo', Rating=94, Height='185 cm', Weight='80 kg', Preffered_Foot='Right', Birth
 Row(id=1, Name='Lionel Messi', Rating=93, Height='170 cm', Weight='72 kg', Preffered_Foot='Left', Birth_Date=
 Row(id=2, Name='Neymar', Rating=92, Height='174 cm', Weight='68 kg', Preffered_Foot='Right', Birth_Date='02/0
 Row(id=3, Name='Luis Suárez', Rating=92, Height='182 cm', Weight='85 kg', Preffered_Foot='Right', Birth_Date=
 Row(id=4, Name='Manuel Neuer', Rating=92, Height='193 cm', Weight='92 kg', Preffered_Foot='Right', Birth_Date
```

```
[24]  1  #memecah dataset clubs dengan memanggil kolom yang sesuai
      2  clubs = FD['id','Club','Club_Position']
      3
      4  #membuat RDD dari dataset clubs
      5  clubs_rdd = clubs.rdd
      6
      7  #menampilkan 5 data teratas dari clubs_rdd
      8  clubs_rdd.take(5)
```

```
[Row(id=0, Club='Real Madrid', Club_Position='LW'),
 Row(id=1, Club='FC Barcelona', Club_Position='RW'),
 Row(id=2, Club='FC Barcelona', Club_Position='LW'),
 Row(id=3, Club='FC Barcelona', Club_Position='ST'),
 Row(id=4, Club='FC Bayern', Club_Position='GK')]
```

```
[26]  1  #memecah dataset nations dengan memanggil kolom yang sesuai
      2  nations = FD['id','Nationality','National_Position']
      3
      4  #membuat RDD dari dataset nations
      5  nations_rdd = nations.rdd
      6
      7  #menampilkan 5 data teratas dari nations_rdd
      8  nations_rdd.take(5)
```

```
[Row(id=0, Nationality='Portugal', National_Position='LS'),
 Row(id=1, Nationality='Argentina', National_Position='RW'),
 Row(id=2, Nationality='Brazil', National_Position='LW'),
 Row(id=3, Nationality='Uruguay', National_Position='LS'),
 Row(id=4, Nationality='Germany', National_Position='GK')]
```

3

SOAL NO 3

Ubah RDD di atas menjadi PySpark Dataframe (5, SCPMK 1532205)

```
[32]  1  #mengubah RDD players ke PySpark DataFrame
      2  players_df = players_rdd.toDF()
      3
      4  #menampilkan 5 baris teratas dari players_df
      5  players_df.show(5)
```

```
+---+----------------+------+------+------+--------------+----------+---+
| id|            Name|Rating|Height|Weight|Preffered_Foot|Birth_Date|Age|
+---+----------------+------+------+------+--------------+----------+---+
|  0|Cristiano Ronaldo|    94|185 cm| 80 kg|         Right|02/05/1985| 32|
|  1|     Lionel Messi|    93|170 cm| 72 kg|          Left|06/24/1987| 29|
|  2|           Neymar|    92|174 cm| 68 kg|         Right|02/05/1992| 25|
|  3|      Luis Suárez|    92|182 cm| 85 kg|         Right|01/24/1987| 30|
|  4|     Manuel Neuer|    92|193 cm| 92 kg|         Right|03/27/1986| 31|
+---+----------------+------+------+------+--------------+----------+---+
only showing top 5 rows
```

[33]
```
#mengubah RDD clubs ke PySpark DataFrame
clubs_df = clubs_rdd.toDF()

#menampilkan 5 baris teratas dari clubs_df
clubs_df.show(5)
```

```
+---+-----------+-------------+
| id|       Club|Club_Position|
+---+-----------+-------------+
|  0| Real Madrid|           LW|
|  1|FC Barcelona|           RW|
|  2|FC Barcelona|           LW|
|  3|FC Barcelona|           ST|
|  4|   FC Bayern|           GK|
+---+-----------+-------------+
only showing top 5 rows
```

[34]
```
#mengubah RDD nations ke PySpark DataFrame
nations_df = nations_rdd.toDF()

#menampilkan 5 baris teratas dari nations_df
nations_df.show(5)
```

```
+---+-----------+-----------------+
| id|Nationality|National_Position|
+---+-----------+-----------------+
|  0|   Portugal|               LS|
|  1|  Argentina|               RW|
|  2|     Brazil|               LW|
|  3|    Uruguay|               LS|
|  4|    Germany|               GK|
+---+-----------+-----------------+
only showing top 5 rows
```

4

SOAL NO 4

Lakukan agregasi untuk mendapatkan jumlah pemain untuk masing-masing kelompok berikut: (25, SCPMK 1534104)

- Berdasarkan Nationality
- Berdasarkan Club

```
#membuat dataframe grouping berdasarkan kolom 'Nationality'
#dari dataframe nations_df
by_nations = nations_df.groupBy('Nationality')

#menampilkan jumlah pemain dari hasil grouping 'Nationality'
by_nations.count().show()
```

```
+--------------+-----+
|   Nationality|count|
+--------------+-----+
|          Chad|    1|
|        Russia|  309|
|      Paraguay|   75|
|Chinese Taipei|    1|
|       Senegal|  119|
|        Sweden|  378|
|        Guyana|    3|
|       Eritrea|    1|
|   Philippines|    2|
|          Fiji|    1|
|        Turkey|  292|
|          Iraq|    8|
|       Germany|  689|
|St Kitts Nevis|    3|
|       Comoros|    9|
|   Afghanistan|    2|
|   Ivory Coast|   90|
|        France|  974|
|        Greece|   86|
|        Kosovo|   31|
+--------------+-----+
only showing top 20 rows
```

```
#membuat dataframe grouping berdasarkan kolom 'Club'
#dari dataframe clubs_df
by_club = clubs_df.groupBy('Club')

#menampilkan jumlah pemain dari hasil grouping 'Club'
by_club.count().show()
```

```
+---------------+-----+
|           Club|count|
+---------------+-----+
|        Palermo|   28|
|Shonan Bellmare|   24|
|    Yeovil Town|   24|
|     Sagan Tosu|   25|
|          Carpi|   26|
| Kaiserslautern|   28|
|   Sparta R'dam|   29|
|       FC Basel|   27|
|  Karlsruher SC|   28|
|Cheltenham Town|   25|
|             AZ|   29|
|    SC Freiburg|   27|
|  Al. Petrolera|   28|
|    GFC Ajaccio|   26|
|      FC Luzern|   24|
|  SC Heerenveen|   28|
|       Brighton|   30|
|            AIK|   20|
|         Santos|   20|
|  Sp. Charleroi|   27|
+---------------+-----+
only showing top 20 rows
```

5

SOAL NO 5

Lakukan agregasi untuk mendapatkan nilai: (25, SCPMK 1534104)

- Rata-rata Rating berdasarkan Club_Position
- Maximum Rating berdasarkan National_Position

```
#menggabungkan dataframe players dengan club
players_with_club = players_df.join(clubs_df, on='id')

#melakukan grouping dari hasil joining berdasarkan kolom "Club Position"
by_club_position = players_with_club.groupBy('Club_Position')

#melakukan dan menampilkan hasil aggregasi Rata-rata dari kolom rating
by_club_position.avg('Rating').show()
```

```
+-------------+-----------------+
|Club_Position|      avg(Rating)|
+-------------+-----------------+
|           RF|             72.0|
|          LWB|67.44444444444444|
|          LCM|68.77966101694915|
|           LM|68.81884057971014|
|          RDM|69.81954887218045|
|           LF|71.08333333333333|
|         null|             81.0|
|          CAM| 70.5904761904762|
|          RAM|71.22222222222223|
|           LB|68.10382513661202|
|           LW|71.84210526315789|
|          RCM|68.74787535410765|
|           GK|69.82594936708861|
|           RB|68.05656934306569|
|          Sub|65.37172984516818|
|           RS|68.63106796116504|
|          LCB| 69.1648177496038|
|           CM|69.16455696202532|
|           RW|70.93984962406014|
|          RCB|69.56714060031595|
+-------------+-----------------+
only showing top 20 rows
```

```
#menggabungkan dataframe players dengan nations
players_with_nation = players_df.join(nations_df,on='id')

#melakukan grouping dari hasil joining berdasarkan kolom "National_Position"
by_national_position = players_with_nation.groupBy('National_Position')

#melakukan dan menampilkan hasil aggregasi Nilai Maksimal dari kolom rating
by_national_position.max('Rating').show()
```

```
+-----------------+-----------+
|National_Position|max(Rating)|
+-----------------+-----------+
|               RF|         85|
|              LWB|         75|
|              LCM|         88|
|               LM|         87|
|              RDM|         84|
|               LF|         89|
|             null|         90|
|              CAM|         89|
|              RAM|         83|
|               LB|         86|
|               LW|         92|
|              RCM|         88|
|               GK|         92|
|               RB|         84|
|              Sub|         89|
|               RS|         90|
|              LCB|         89|
|               CM|         85|
|               RW|         93|
|              RCB|         89|
+-----------------+-----------+
only showing top 20 rows
```

6

SOAL NO 6

Lakukan setidaknya 3 operasi transform RDD pada data yang kalian punya, contoh reduceByKey, sortByKey, flatMap, etc. (30,

RDD PLAYER

```
[82] #membuat RDD dari dataframe baru dengan memanggil kolom birth_date dan age
     rdd1 = players['Birth_Date','Age'].rdd
```

```
#melakukan swap 'Age' menjadi keys dan 'Birth_Data menjadi values
rdd1_swap = rdd1.map(lambda x:(x[1],x[0]))

#menampilkan 10 baris teratas hasil swap
for i in rdd1_swap.take(10):
  print(f'{i[0]}, {i[1]}')
```

```
32, 02/05/1985
29, 06/24/1987
25, 02/05/1992
30, 01/24/1987
31, 03/27/1986
26, 11/07/1990
28, 08/21/1988
27, 07/16/1989
35, 10/03/1981
24, 05/11/1992
```

TRANSFORM 1 DENGAN CountByKey()

```
#menampilkan transformasi RDD
#menampilkan jumlah pemain berdasar umur dengan countByKey

for i,j in rdd1_swap.countByKey().items():
  print(f'pemain dengan umur {i} berjumlah {j}')
```

```
pemain dengan umur 32 berjumlah 567
pemain dengan umur 29 berjumlah 1104
pemain dengan umur 25 berjumlah 1447
pemain dengan umur 30 berjumlah 852
pemain dengan umur 31 berjumlah 668
pemain dengan umur 26 berjumlah 1195
pemain dengan umur 28 berjumlah 1071
pemain dengan umur 27 berjumlah 1134
pemain dengan umur 35 berjumlah 236
pemain dengan umur 24 berjumlah 1296
pemain dengan umur 33 berjumlah 598
pemain dengan umur 34 berjumlah 317
pemain dengan umur 39 berjumlah 24
pemain dengan umur 23 berjumlah 1356
pemain dengan umur 22 berjumlah 1283
pemain dengan umur 36 berjumlah 159
pemain dengan umur 21 berjumlah 1196
pemain dengan umur 20 berjumlah 1208
pemain dengan umur 37 berjumlah 102
pemain dengan umur 19 berjumlah 1004
pemain dengan umur 38 berjumlah 50
pemain dengan umur 40 berjumlah 16
pemain dengan umur 18 berjumlah 533
pemain dengan umur 44 berjumlah 3
pemain dengan umur 42 berjumlah 2
pemain dengan umur 17 berjumlah 157
pemain dengan umur 43 berjumlah 3
pemain dengan umur 41 berjumlah 6
pemain dengan umur 47 berjumlah 1
```

TRANSFORM 2 DENGAN filter()

```
#membuat RDD dari dataframe baru dengan memanggil kolom Name dan Rating
rdd2 = players['Name','Rating'].rdd

#inisialiasi variabel rating yang akan digunakan
rating = 89

#melakukan tranformasi filter RDD
#melakukan filter untuk pemain dengan rating lebih dari variabel yang ditentukan
rdd2_filter = rdd2.filter(lambda x: x[1]>rating)

print(f'Pemain dengan rating lebih dari {rating} :')

#menampilkan nama pemain hasil filter
for x in rdd2_filter.collect():
  print(x[0])
```

```
Pemain dengan rating lebih dari 89 :
Cristiano Ronaldo
Lionel Messi
Neymar
Luis Suárez
Manuel Neuer
De Gea
Robert Lewandowski
Gareth Bale
Zlatan Ibrahimović
```

TRANSFORM 3 DENGAN CountByValue()

RDD CLUB

```
[150]  1  #membuat RDD dari dataframe baru dengan memanggil kolom Club dan Club_Position
       2  rdd3 = clubs['Club','Club_Position'].rdd
```

```
1  #melakukan transform RDD
2  #menghitung jumlah pemain dengan role tertentu pada sebuah club
3  #kemudian di simpan ke dalam list agar bisa dilakukan iterasi
4
5  rdd3_list=list(rdd3.countByValue().items())
6
7  #menampilkan 100 data teratas dari hasil transform diatas
8  for i in range(100):
9     print(f"pemain dengan role '{rdd3_list[i][0][1]}' \
10 pada club '{rdd3_list[i][0][0]}' ada sejumlah {rdd3_list[i][1]}")
11
12 #rdd3_list[i][0][1] merujuk pada value ROLE
13 #rdd3_list[i][0][0] merujuk pada value NAMA CLUB
14 #rdd3_list[i][1] merujuk pada value JUMLAH PEMAIN hasil transform RDD
```

```
pemain dengan role 'RM' pada club 'FC Bayern' ada sejumlah 1
pemain dengan role 'CAM' pada club 'Juventus' ada sejumlah 1
pemain dengan role 'RCM' pada club 'PSG' ada sejumlah 1
pemain dengan role 'LB' pada club 'FC Bayern' ada sejumlah 1
pemain dengan role 'RW' pada club 'Manchester Utd' ada sejumlah 1
pemain dengan role 'GK' pada club 'Bayer 04' ada sejumlah 1
pemain dengan role 'CAM' pada club 'FC Bayern' ada sejumlah 1
pemain dengan role 'CDM' pada club 'FC Barcelona' ada sejumlah 1
pemain dengan role 'RDM' pada club 'FC Bayern' ada sejumlah 1
pemain dengan role 'LCB' pada club 'FC Barcelona' ada sejumlah 1
pemain dengan role 'LW' pada club 'Liverpool' ada sejumlah 1
pemain dengan role 'RW' pada club 'PSG' ada sejumlah 1
pemain dengan role 'ST' pada club 'Chelsea' ada sejumlah 1
pemain dengan role 'ST' pada club 'PSG' ada sejumlah 1
pemain dengan role 'LF' pada club 'Roma' ada sejumlah 1
pemain dengan role 'LB' pada club 'Real Madrid' ada sejumlah 1
pemain dengan role 'Sub' pada club 'Juventus' ada sejumlah 12
pemain dengan role 'RCB' pada club 'Inter' ada sejumlah 1
pemain dengan role 'ST' pada club 'Real Madrid' ada sejumlah 1
pemain dengan role 'Sub' pada club 'Chelsea' ada sejumlah 12
pemain dengan role 'Sub' pada club 'Arsenal' ada sejumlah 12
pemain dengan role 'RS' pada club 'Spurs' ada sejumlah 1
```