Apache Spark – Zestaw 6 – fifa-players

Misja główna

Cel przetwarzania

Dla następujących kategorii

- narodowość (opartej na `nationality_name`)
- klub (opartej na `club_name`)
- liga (opartej na `league_name`)

należy wyznaczyć trzy wartości nazw z piłkarzami o największych średnich zarobkach.

W analizach nie uwzględniamy piłkarzy grających w ligach, w których liczba grających zespołów mających co najmniej 11 piłkarzy jest mniejsza niż 10.

Wynik przetwarzania powinien zawierać następujące atrybuty:

- `category` – nazwa kategorii (`nationality`, `club`, `league`)
- `name` – nazwa narodowości/klubu/ligi
- `sum_value_eur` – sumaryczna wartość piłkarzy
- `avg_wage_eur` – średnie zarobki piłkarzy
- `avg_age` – średni wiek piłkarzy w lidze obliczony na podstawie wartości kolumny `age` (nie korzystamy z `dob`)
- `count_players` – liczba zawodników
- `player_positions` – lista skrótów nazw pozycji, na których grają piłkarze

Sugerowany schemat wyniku

```
root
 |-- category: string (nullable = false)
 |-- name: string (nullable = true)
 |-- sum_value_eur: double (nullable = true)
 |-- avg_wage_eur: double (nullable = true)
 |-- avg_age: double (nullable = true)
 |-- count_players: long (nullable = false)
 |-- player_positions: array (nullable = true)
 |    |-- element: string (containsNull = true)
```

Uwagi

- lista skrótów nazw pozycji, na których grają piłkarze nie może zawierać duplikatów
- oryginalne wartości kolumny `player_positions` mogą zawierać wiele skrótów nazw pozycji, należy je traktować jako zbiór
- wartości liczbowe mają być zaokrąglone do jedności

Misje poboczne

Misja 1

Znajdź trzy kluby, których zawodnicy otrzymują największe średnie płace. Dla każdego z takich klubów wyświetl dodatkowo liczbę zawodników, ich średni wiek (na podstawie age), sumaryczną wartość, liczbę zawodników preferujących prawą nogę

Wynik ma zawierać następujące kolumny:

- `club_name` – nazwa klubu
- `wage_eur_avg` – średnie zarobki zawodników
- `players_count` – liczba zawodników
- `age_avg` – średni wiek
- `value_eur_sum` – sumaryczna wartość
- `preferred_right_count` - liczba zawodników preferujących grę prawą nogą

Misja 2

Dla każdego poziomu ligi określ liczbę lig, liczbę klubów oraz liczbę zawodników. Uwzględniaj w obliczeniach tylko zawodników urodzonych po 1969 roku. Nie uwzględniaj w obliczeniach klubów mających mniej niż 11 zawodników urodzonych po 1969 roku. Nie uwzględniaj zawodników grających w odrzuconych klubach.

Wynik ma zawierać następujące kolumny:

- `league_level` – nazwa poziomu ligi
- `leagues_count` – liczba lig
- `teams_count` – liczba klubów
- `players_count` – liczba zawodników