Data Engineer Challenge

Name: Devanand Arul Pragasam

Date: 26.06.2020

1. Implement a solution to store the FIFA19 dataset to enable frequent query and efficient analysis?

Using spark, read the data as dataframe and save in hive using "saveAsTable". In hive, there are optimization techniques such as bucketing, partitioning, bucket map joins and so on to enable frequent query and efficiency analysis.

2. a. Which club has the most number of left footed midfielders under 30 years of age?

Let fifaDF be the dataframe read from the given data,

```
val fifaDF = spark.read.option("header", "true").option("inferSchema", "true").csv("C:\\Users\\Devanand\\Desktop\\datafifa.csv")

val preferredFootAge = fifaDF.where("PreferredFoot = 'Left' and age <= 30")

val clubWithMostLeftFooters = preferredFootAge.groupBy(col("club")).agg(count("club").alias("clubcount")).orderBy(desc("clubcount"))

clubWithMostLeftFooters.show(1)
```

b. The strongest team by overall rating for a 4-4-2 formation

```
val strongClub = fifaDF.groupBy(col("club")).agg(avg("overall").alias("cluboverallrating")).orderBy(desc("cluboverallrating"))

 strongClub.show(1)
```

c. Which team has the most expensive squad value in the world? Does that team also have the largest wage bill ?

```
val averageValueByClub = df.groupBy(col("club")).agg(avg("value").alias("averageValueByClub"))

val expensiveSquadValue = averageValueByClub.agg(max("averageValueByClub").alias("expensiveSquadValue"))

val averageWageByClub = df.groupBy(col("club")).agg(avg("wage").alias("averageWageByClub"))

val highestWageByClub = averageWageByClub.agg(max("averageWageByClub").alias("highestWageByClub"))
```

From the resultset, when a club has "highest wage by club" and "expensiveSquadValue" then the club also has largest wage bill otherwise not.

d. Which position pays the highest wage in average?

```
val highestWageByPosition = fifaDF.groupBy(col("Position")).agg(avg("Wage").alias("averageWageByPosition")).orderBy(desc("averageWageByPosition"))

  highestWageByPosition.show(1)
```

e. What makes a goalkeeper great? Share 4 attributes which are most relevant to becoming a good goalkeeper?

Good jumping ability, fantastic co-ordination, excellent distribution and fast reflexes are all attributes you will find in a great goalkeeper.

f. What makes a good Striker (ST)? Share 5 attributes which are most relevant to becoming a top striker ?

The ability to control a football or 'touch', Ability to run fast with football speed, Shot accuracy with consistent training and Positioning to make sure where to place yourself.

3. Steps to connect to "mysql" (not worked with Postgres database)

a. add "mysql" dependency to build.sbt

b. add the jar file "mysql-connector-java" in libraries in project structure.

c. Read the data as dataframe.

d. Create another dataframe by selecting the required fields from the dataset and write it in the mysql database using jdbc url connection,driver,username and password.

****