MIE1628: BIG DATA SCIENCE

ASSIGNMENT 2: Spark

Posted on March 15th

**Due on March 28th**

You can use either DataBricks or Queens University cluster resources to complete this assignment.

You also have a choice in terms of Spark APIs: feel free to use PySpark or Scala.

Each practical question requires submission of your code.

We will be working with the following files:

* game.csv
* team\_info.csv
* game\_team\_stats.csv

**Part A: General theoretical questions**

Question 1 (1 points)

List all Spark APIs

* Scala,
* PySPark,
* SQL
* Java
* R

Datatype APIs:

* Datasets
* DataFrames
* SQL
* RDDs
* Distributed Variables

Question 2 (2 points)

Define an executor

* An executor is a process that runs in a container, it is responsible for executing tasks. Usually, 1 executor is 1 container

Question 3 (2 points)

What makes dataframes different from RDDs?

* Unlike RDD, dataframes is organized into named columns, like a table in a relational database.

Question 4 (2 points)

What distinguishes narrow and wide transformations?

* Narrow transformations are operations with dependencies on **just one or a known set of partitions** in the parent RDD. Thus narrow transformations can be executed on an arbitrary subset of the data without any information about other partitions.
* Wide transformations are operations that **can’t be executed on a subset of data**. Wide transformations include sort, reduceByKey, groupBYKey, join, and anything that calls for repartition.
* In sum, each partition of the parent RDD is used by at most one partition of child RDD in narrow transformation but in wide transformation each partition may be depended on multiple child partitions.

**Part B: PySpark/Scala syntax**

Question 5 (38 points)

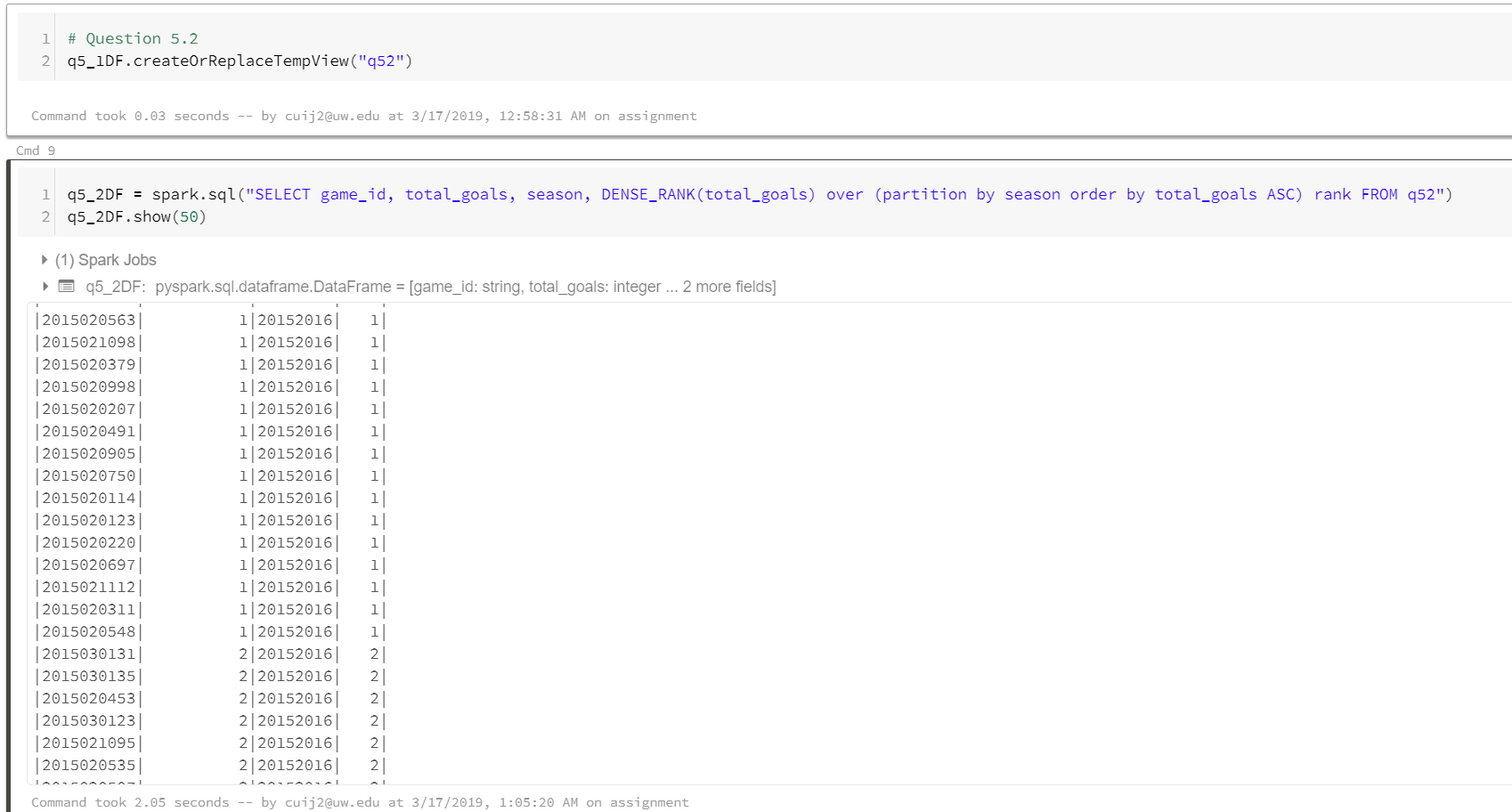
1. Select all games and seasons and add a column with total goals (sum of home and away goals). (Suggestion: use df.withColumn() function )- 4 points



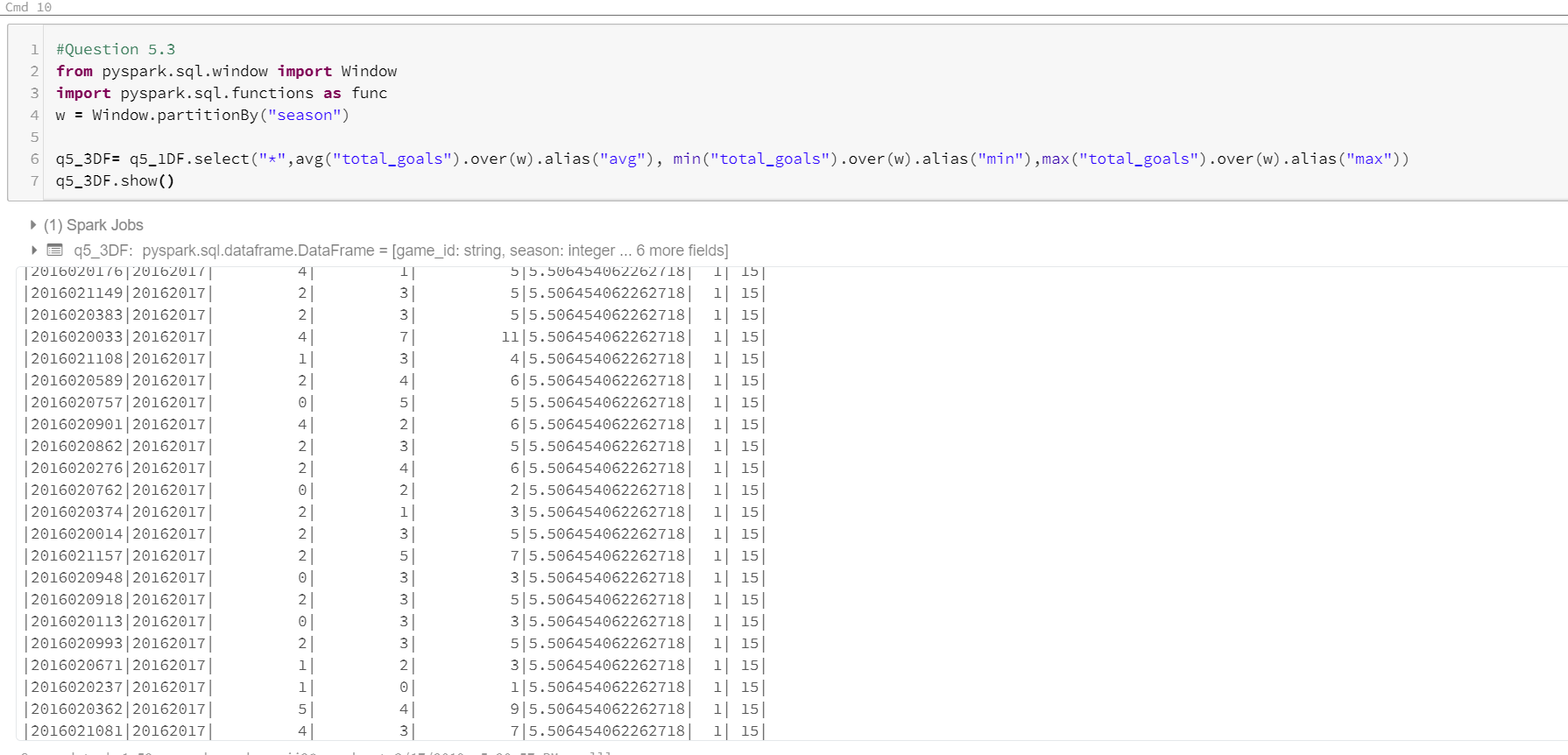
1. Organize records in ascending order (by season)- 2 point

图片包含 屏幕截图

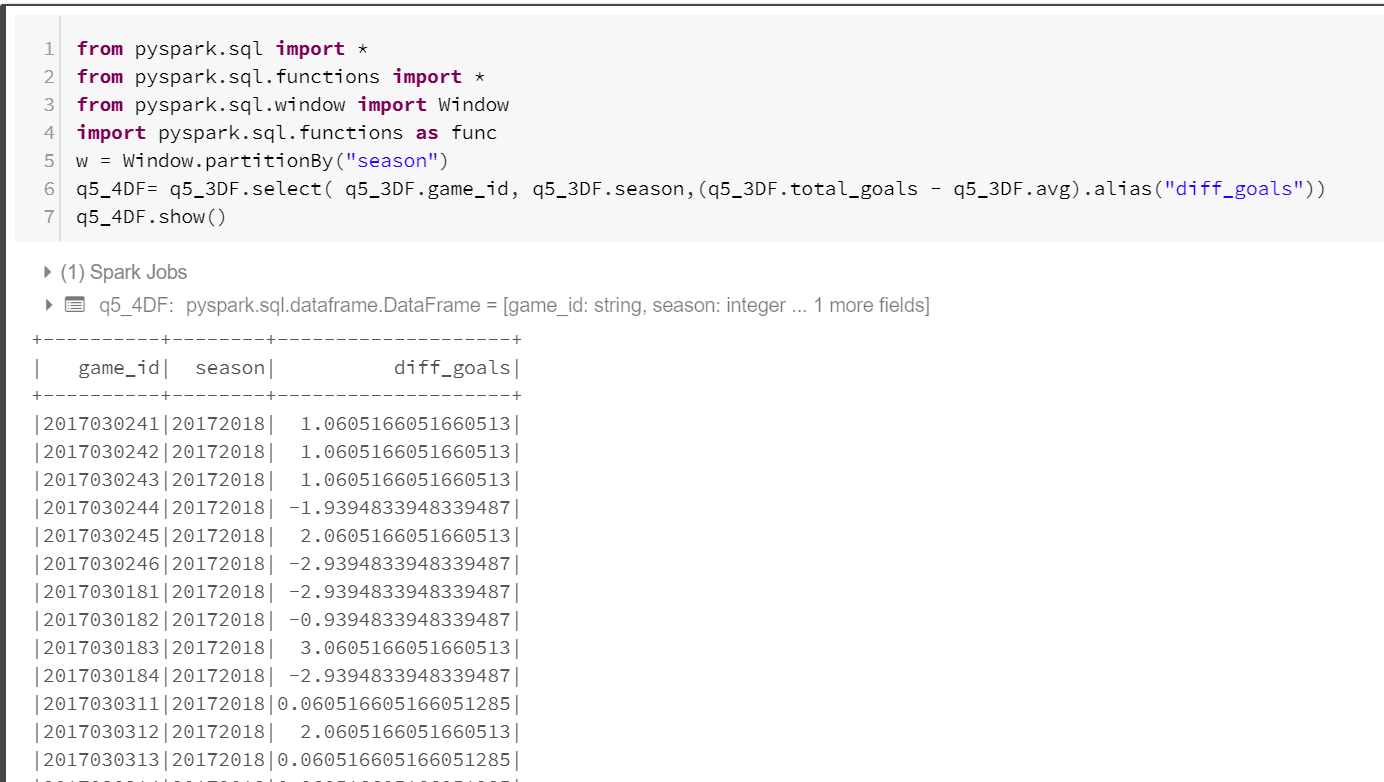
描述已自动生成



1. Add a column with an average, min and max total score for each season. Suggestion: use Window function - 15 points



1. Add a column that finds a difference between each game’s total score and average for that season. Suggestion: use Window function - 15 points



1. Print top 10 records. (Output: Your code and a snapshot of the top 10 records) -2 point



Question 6 (15 points)

1. List all team names (teamName) for teams that played as away team at TD Garden during seasons 2012-2013 and 2013-2014 - 10 points



1. How many unique teams are on the list? - 5 points

