# SI 618 Fall 2021 Lab 6 - SparkSQL

In this lab, we will practice SparkSQL. We will be diving into a dataset of League of Legends matches and individual performance in the North American Region. This open dataset was created as part of Lee, C. S., & Ramler, I. (2017, August).

LoL is a multiplayer online battle arena game where two teams of 5 (or 3) players battle each other on a map called an arena and try to destroy each other's base (or Nexus) while protecting their own. Each player can choose from among over 100 heros (or champions) with different skills and play different roles (attack, support, top, bottom, jungle).


*League of Legends Arena*

To avoid traffic jams when logging into pyspark, please use command:
pyspark --master yarn --conf spark.ui.port="$(shuf -i 10000-60000 -n 1)"

## Question 1: (20 points)

Load the csv file *hdfs:///var/umsi618f21/lab6/na_ranked_team.csv* and register it as a table. This dataset includes player behavior data for a total of 112,301 ranked 5 vs. 5 matches in North America during the period from November 12, 2014, to November 11, 2015, corresponding to 1,123,010 rows (no. of matches x 10 players per match) and 320,468 players. The dataset contains 80 columns providing a great deal of information about the behavior of players during each match (see the ReadMe.txt file for description for all columns), but we will only be using a few of them in this lab.

In LoL, KDA or the kill-death-assist ratio is a popular measure of individual performance. A player scores a *kill* whenever they deal the final blow that defeats another player of the opposing team in combat and the opposing player suffers a *death.* Since LoL is a team game, a third player of the same team as the first player may *assist* them in defeating the second player. Based on these actions, the KDA is defined as (kills + assists)/(deaths+1). First, calculate the KDA for each player for each match and then output average KDA with the number of matches for players who have played at least 10 matches.

As an example, if a player X had played two matches and had the following performance statistics,

Match 1 3 kills 1 death 1 assist
Match 2 1 kills 3 deaths 1 assists

Their match 1 KDA would be (3 +1)/(1+1) = 2 and match 2 KDA would be (1+1)/(3+1) = ½. Their average KDA would be (2 + 0.5)/2 = 1.25

In the input, the column **summonerId** is a unique identifier for a player. Your output should have the format **summonerId<tab>matches<tab>avg. KDA**. It should be sorted by the decreasing order of avg. KDA and then decreasing order of the # matches.

Save it as UNIQUENAME_si618_lab6_output_1.tsv. Your output should match si618_lab6_desired_output_1.tsv.

## Question 2: (30 points)

While average KDA is a reasonable indicator of player performance, we can do better. A player's performance during a single match may be affected by the role they play, the champion they select, their experience, and the skill levels of the two teams. Let's try to improve the KDA measure to take some of this into account. Calculate a new adjusted KDA for each player's performance during a match by normalizing their match KDA by the average match KDA of their own team as follows; (player KDA)/(average team KDA +1). Similar to Q1, now calculate the average normalized KDA for each player who has played at least 10 matches across all matches.

Note that the ***winner*** column can be used to identify the team to which a player belonged in a match. The value 1 corresponds to the winning team and 0 corresponds to the losing team. Also, when you calculate the team KDAs for a match you **should** consider all players in a team irrespective of the total number of matches they have played.

Your output should include **summoner_id<tab># matches<tab>average normalized KDA** sorted by the decreasing order of avg. KDA and then decreasing order of the # matches.

| |
|---|
| Save it as UNIQUENAME si618_lab6_output_2.tsv. Your output should match si618_lab6_desired_output_2.tsv. |

## Question 3: (50 points)

In LoL, players select champions based on community knowledge about which of them are more appropriate for different roles as well as choices of the opposing team (in ranked games teams pick champions in stages). Let's try to discover which pairs of champions are chosen by opposing teams most often for the same role and how well they match up.

First find the match ups of unique pairs (by alphabetical order) of champions used in the same role in opposing teams during matches. Then only considering pairs who have had at least 10 match ups against each other in the same role, calculate average KDA ratio across all their match ups. The KDA ratio for a given match is calculated as (KDA of champion1)/(KDA of champion2) where champion1 is alphabetically ahead of champion2

Note that the ***championName*** column has the name of the champion and the column **predictedRole** specifies the role (0-top,1-jungle,2-middle,3-attack damage carry,4-support). You can write multiple separate SQL queries to get the output.

The output should contain rows of the form champion1<tab>champion2<tab>role<tab># matches<tab> average KDA-ratio. It should be sorted by champion1, role, the descending order of the number of match ups and finally champion2.

| |
|---|
| Save it as si618_lab6_output_3_UNIQUENAME.tsv. Your output should match si618_lab6_desired_output_3.tsv. |

You MUST use **SparkSQL** to do this lab. Other solutions will not get any credit.

## What to submit:

Submit your Python source code file, uniquename_si618_lab6_.py, and the 3 tsv files. You should submit these as separate files and **not as a zipped folder** (This helps the IAs to grade more efficiently)