Parallel and Distributed Computing
Lab5
Basic SparkSQL

In this lab you will read in a datafile into an RDD, create a Schema and convert the RDD into a DataFrame, and finally do some simple queries using both SQL and DataFrame functions.

The Datafile

We will be use baseball data from https://www.kaggle.com/open-source-sports/baseball-databank/data. I have posted the zip file with all the baseball data on Canvas. You will only need Master.csv, which contains general player bio data. Player baseball stats are in other files that we won't be using for this lab.

Creating the DataFrame

We will be creating our DataFrame from RDD rows for this lab. Thus, you will need to read in the Master.csv file as a text file into an RDD, parse it, and prepare it with Rows.

Note that it includes a header line at the top that you will need to make sure not to include.

Also note that when you are creating the rows, you will not need all the data you read in. For the example queries you will be doing, you will only need a subset of this data. Thus, your Schema and Rows should only include the information you need. In particular, playerID, birthCountry, birthState, and height will be needed.

You will also need to do a bit of data cleanup. Some of the entries in the csv file do not contain height information. Please do not include these in the DataFrame.

You should create a Schema as well with StructType rather than inferring one from the data as it is read in. Use your Row RDD and Schema to create the DataFrame.

Colorado

The first query you should implement is:

Find the number of players who were born in the state of Colorado.

You should do this both using SQL and again with DataFrame functions.

Height by Country

The second query you should implement is:

List the average height by birth country of all players, ordered from highest to lowest.

You should do this both using SQL and again with DataFrame functions.

Note that by default .show() will cap the output at 20 rows. To see more you can pass a number to show to display more. For example: df.show(df.count()) will show all the rows.

Submission

You are to upload 2 documents as your submission. The first is your Python source code (PY). Make sure to have your name in a comment at the top of your source code. Second is a document that has a cut-paste of your results from running your code.