hugomcp@di.ubi.pt, 2021/22

Data Science

Distributed Extract-Transform-Load Application

This exercise regards the development of a distributed Extract-Transform-Load (ETL) application.

Your application should load the data from a relational database system and use a distributed data processing tool (e.g., **Apache Hadoop** or **Apache Spark**) to compute some statistics and output them in a form that can be loaded into some destination storage system for further "consumption".

Please write an application in Python, that should run from the command-line.

1) Source database

Build the source database by using an instance of MySQL or equivalent. Use the SQL script (available at: http://www.di.ubi.pt/~hugomcp/dataScience/source_db.zip) to create the database schema and load it..

2) Extract data and compute statistics

Your application should extract data from the source database and perform the following computations:

1. Calculate the average salary for infielders and pitchers for each year;

2. Calculate the number of all star appearances for each Hall of Fame pitcher and their average ERA their all-star years and list the year they were inducted into the Hall of Fame;

3. Calculate the top 10 pitchers' average regular season and post-season ERAs and average win/loss (w/(w+l)) percentages:

    - The top 10 pitcher's ERAs ((0.65(player1) + 0.72(player2) + ...) / 10) and (win/loss of player1 + win/loss of player2 + ...)/10

- The pitchers in the top 10 may not have been on a team that made it to the post-season, so average the ERAs & win/loss of the pitchers that made it into the post-season

4. List the first and last place teams and their number of at-bats for each year.

3) Get data ready for loading

The application should format the output data set into a number of CSV files as specified in the below examples and upload them to a storage system of your choosing (e.g. S3) for daily download by consumers.

1. Average Salaries

   Year, Fielding, Pitching
   1985, "2,028,571", "1,713,333"
   1990, "2,100,000", "2,600,000"
   2000, "3,111,000", "4,500,000"

2. Hall of Fame All Star Pitchers

   Player, ERA, # All Star Appearances, Hall of Fame Induction Year
   abcdef01, 3.11, 8, 1999
   defghi01, 2.31, 8, 1988
   ghijkl01, 1.91, 11, 2006

3. Pitching

   Year, Player, Regular Season ERA, Regular Season Win/Loss, Post-season ERA, Post-season Win/Loss
   1990, defgei01, 1.74, 73, 1.14, 100
   1991, abcdhi01, 1.36, 71, 2.14, 85
   1992, fdwesi01, 2.06, 70, 1.85, 90
   1993, sdfwei01, 1.90, 65, 0.85, 87

4. Rankings

   Team ID, Year, Rank, At Bats
   PH1, 1871, 1, 1281
   RC1, 1871, 9, 1036
   LAA, 2014, 1, 5652
   CHN, 2014, 5, 5508