600.315/415 Databases Final Project Writeup

Jin Yong Shin (Jshin44) || Ji Won Shin (Jshin49)

Email: Jshin44@jhu.edu || Jshin49@jhu.edu

Section : 600.315

December 22, 2016

Abstract

Why do we care and how are we going to organize data?
These question questions were the biggest motivation for our project. On Nov 29th, a chartered plane carrying a Brazilian first division soccer team crashed into a Colombian hillside. The plane crash itself was disaster. However, the number of survivors and fatalities kept changing because it is hard to match names on malformed and inconsistent database. So my team designed database focusing on (1) WorldCup (2) Olympic and (3) European Top 3 Leagues. For each division, we have imported team squad, game result, and league or national standings. Major challenges that we have faced were:
(i) tuning all different languages into a uniform language, which was English in our case
(ii) converting player and club name into consistent form
(iii) keeping track of those raw data and their connection to the converted version in database
To implement such database, we extensively use string similarity check codes in Java platform and some database techniques.

1 Introduction: Data Reference and Query By Example

1.1 Data Reference

$TID \rightarrow$ Universal unique team ID

$Club\_original \rightarrow$ Club name with original language

$Club \rightarrow$ Club name

$NCode \rightarrow$ Unique national code for specific country, including country name variation

$LID \rightarrow$ Unique league ID

$CurChamp \rightarrow$ Most recent winning team of the league

$Pld \rightarrow$ Number of games played

$GF \rightarrow$ Goals Forward

$GA \rightarrow$ Goals Against

$GD \rightarrow$ Goals Difference

$Mdate \rightarrow$ Match Date

$Home\_T \rightarrow$ Home team

$Away\_T \rightarrow$ Away team

$Home\_S \rightarrow$ Home team score

$Away\_S \rightarrow$ Away team score

$Home\_ES \rightarrow$ Home team expected score

$Away\_ES \rightarrow$ Away team expected score

$Participate \rightarrow$ Number of World Cup participation

$F\_Score \rightarrow$ World Cup final score

$SF\_Score \rightarrow$ World Cup semi-final score

1.2 Query By Example(QBE)

QBE

#mapping of raw team names to TID
#which later will serve as a universal team ID that connects between refined and raw expressions of team names

Club_original	TID	Club_original
	T010	Atlético Madrid

Clubs	TID	Club
	T010	Atletico Madrid

#mapping of FIFA country code and country name

Country_code	NCode	Country
	BRA	Brazil

#Information about the "Big 3" Football leagues

League_List	LID	League_Name	NCode	Year	N_Teams	CurChampion	MostChampion
	a001	Bundesliga	GER	1963	18	Bayern Munich	Bayern Munich

#table shows standing of teams each year

League_standing	LID	Season	Rank	Club	Pld	Win	Draw	Lost	GF	GA	GD	Points
	a001	2011	1	Borussia Dortmund	34	12	4	1	67	22	+45	75

#every big3 footbal league all game from 2010 to 2014

League_games	LID	Season	Mdate	Home_T	Away_T	Home_S	Away_S	Home_ES	Away_ES
	a001	2011	2010-08-20	Bayern Munchen	VfL Wolfsburg	2	1	1	0

#every big3 football league team squad

League_squad	Season	LID	Team	Num	Player	Nation	Position	Height	Weight	Date_of_Birth
	2011	a001	Bayer Leverkusen	1	Rene Adler	GER	G	1.9	85	1985-01-15

Position_name	Position	Pos_Name
	F	Forward

Olympic_host	Year	City	NCode
	2012	London	ENG

Olympic_record	Year	Rank	Country	NOC
	1948	5	France	FRA

Olympic_squad	Year	Country	Position	FName	LName	Num	Club
	2014	uruguay	MF	Alvaro	Pereira	56	Sao Paulo

WorldCup_host	Year	Country	Continent	Performance
	2006	Germany	Europe	3rd

WorldCup_squad	Year	Num	Player	Nation	Position	Club
	2002	1	Erick Lonnis	CRC	G	Deportivo La Coruna

WorldCup_stat	Rank	Country	M_Played	Won	Tie	Lost	GF	GA	Points	Participations	Championships
	14	Mexico	49	12	13	24	52	89	49	14	0

WorldCup_win	Year	Winner	F_Score	Runner_Up	Third	SF_Score	Fourth	Total_Team
	2006	Italy	1-1 (5-3p)	France	Germany	3-1	Portugal	32

2 Description

2.1 Modification from Phase I

Project description or database design has changed from phase I:

1. We mostly maintained the initial objectives mentioned in the Phase I part of the project. However, we focused more on cleaning up raw forms of data and the predictive aspect of data analysis as we developed the database.

2. We faced unexpected challenges regarding the processing of different representations of the same object. We made use of various Java programs that converts one form to another and detects similarities among inconsistent representations. After translation, we kept the original form and mapped raw data to refined data.

3. Also, after an intermediate meeting with Professor Yarowsky, we received recommendations to focus more on the predictive aspect of the data analysis. We developed stored predictions that calculates probabilities and predictions of the matches based upon the data.

2.2 Major / Minor Areas of Specialization

I. Complex extraction/processing of real data from online sources

- Since we were dealing with world-wide data sets that were expressed in a variety of languages, converting different representations of proper nouns − player names, team names, country names − into one standard format was a challenge. We wrote a Java program that de-accents all Slavic, Romance, and Germanic languages and converts each word into standard English form. Although our database was set to recognize UTF8 letters, our primary and preferable form of representations was the ASCII setting. While processing these proper nouns, we also kept a list of raw representations and developed a relation that maps original representation of team names to team IDs(Club_original.sql). This mapping was later utilized in stored procedures and enabled search using keywords in foreign representations.

- Inconsistent representations of club names in the data sets acquired from different sources also posed difficulty. For example, Manchester United was represented as Manchester United in an English Premiere League data source, while it was represented as Man U in a World Cup squad data source. A Java program was used to determine similarities among these variations and convert into one standard form. These raw representations were included in the mapping relation. Different representations of country names were also mapped to the ISO country code (CountryCode.sql). This mapping was later utilized in stored representations so that the user

can look up information about a certain country using a variety of keywords for the country name. For example, the ISO code for the United States, USA, was mapped to the ?United States,? ?USA,? and ?America.?

II. A specialized view or formed-based interface with sophisticated report generation.

- We implemented an HTML page that is connected to php files that call stored procedures on our database. We created a total of 10 stored procedures, all of which are included in the HTML page. Among many operations include "Find out the probability of club1 winning a match against club2," "Find out scores of all the games that a certain team has played from 2009/2010 season to 2013/2014 season," "List all the football clubs that a certain player has played for," and "Find out game history between two teams played." Some of these stored procedures requires specialized views. For example, procedure Winning_Odds receives a team name as an input and shows the percentage of wins, loses, and draws of soccer matches depending on the location of the game (home vs. away).

3 Limitations & Future Improvement

Given more data sets, we could develop more accurate and more diverse forms of predictions regarding results, expected rank, and expected performance of each individual player. More advanced predictive measures such as ELO or betting rates would also be calculated. A better data refinement program will significantly strengthen the database since this will allow us to collect data from more sources. Major challenges regarding data collection came from the fact that different sources had different representations of the same object. Instead of one standard name, different names, nicknames, names with accents, and different translations were used to represent the same object.

4 Views & Stored Procedures Outputs

4.1 Views

- (1) Best_Ranking : This view outputs club name with its best ranking achieved along with season. By looking this data, observer can easily find team's basic performance because this is showing best performance of the team from 09/10 to 13/14 season

```
mysql> SELECT * FROM Best_Rank;
+-------------------------+-----------+--------+
| Club                    | Best_Rank | Season |
+-------------------------+-----------+--------+
| 1. FC Kaiserslautern    | 18        | 2011   |
| 1. FC Koln              | 10        | 2010   |
| 1. FC Nurnberg          | 10        | 2010   |
| 1. FSV Mainz 05         | 13        | 2010   |
| 1899 Hoffenheim         | 11        | 2010   |
| Almeria                 | 13        | 2010   |
| Arsenal                 | 3         | 2010   |
| Aston Villa             | 15        | 2010   |
| Athletic Bilbao         | 10        | 2010   |
| Atletico Madrid         | 1         | 2010   |
| Barcelona               | 1         | 2010   |
| Bayer Leverkusen        | 2         | 2010   |
| Bayern Munchen          | 1         | 2010   |
| Betis                   | 12        | 2012   |
| Birmingham              | 18        | 2011   |
| Birmingham City         | 9         | 2010   |
| Blackburn Rovers        | 10        | 2010   |
| Blackpool               | 19        | 2011   |
| Bolton Wanderers        | 14        | 2010   |
```

- (2) Most_Winning : This simply outputs top performance team at each league. This view will tell observer which team won the most national league along with the winning number among three leagues during 09/10 to 13/14 season

```
mysql> SELECT * FROM Most_Winning;
+---------------+-------------------+----------------+
| League        | Club              | Winning_Number |
+---------------+-------------------+----------------+
| Bundesliga    | Borussia Dortmund |              3 |
| Premier League | Manchester United |              2 |
| La Liga       | Real Madrid       |              3 |
+---------------+-------------------+----------------+
```

- (3) Game_Result : This table will show more win-lose focused table from *League_games* table. Based on the score given, this table displays whether each team has won or lose game.
  Result is telling with three numbers: 1 is for win, 2 is for lose, 3 is for draw. Additionally, H_Result refers to home team result and H_E_Result refers to home team expected result. Same rule applies to away team as well.

```
mysql> SELECT * FROM Game_Result;
+--------+------+-------------------------+----------+------------+-------------------------+----------+------------+
| Season | LID  | Home_T                  | H_Result | H_E_Result | Away_T                  | A_Result | A_E_Result |
+--------+------+-------------------------+----------+------------+-------------------------+----------+------------+
| 2010   | a001 | VfL Wolfsburg           |        1 |          3 | VfB Stuttgart           |        2 |          3 |
| 2010   | a001 | 1. FC Nurnberg          |        2 |          2 | FC Schalke 04           |        1 |          1 |
| 2010   | a001 | 1. FSV Mainz 05         |        3 |          2 | Bayer Leverkusen        |        3 |          1 |
| 2010   | a001 | 1899 Hoffenheim         |        3 |          3 | Bayern Munchen          |        3 |          3 |
| 2010   | a001 | Borussia Dortmund       |        1 |          3 | 1. FC Koln              |        2 |          3 |
| 2010   | a001 | Hertha BSC              |        1 |          3 | Hannover 96             |        2 |          3 |
| 2010   | a001 | Werder Bremen           |        2 |          3 | Eintracht Frankfurt     |        1 |          3 |
| 2010   | a001 | SC Freiburg             |        3 |          2 | Hamburger SV            |        3 |          1 |
| 2010   | a001 | VfL Bochum              |        3 |          2 | Borussia Monchengladbach |        3 |          1 |
| 2010   | a001 | 1. FC Koln              |        2 |          3 | VfL Wolfsburg           |        1 |          3 |
| 2010   | a001 | Bayer Leverkusen        |        1 |          3 | 1899 Hoffenheim         |        2 |          3 |
| 2010   | a001 | Bayern Munchen          |        3 |          2 | Werder Bremen           |        3 |          1 |
| 2010   | a001 | Eintracht Frankfurt     |        3 |          1 | 1. FC Nurnberg          |        3 |          2 |
| 2010   | a001 | Hamburger SV            |        1 |          1 | Borussia Dortmund       |        2 |          2 |
| 2010   | a001 | Hannover 96             |        3 |          3 | 1. FSV Mainz 05         |        3 |          3 |
| 2010   | a001 | VfB Stuttgart           |        1 |          3 | SC Freiburg             |        2 |          3 |
| 2010   | a001 | Borussia Monchengladbach |        1 |          1 | Hertha BSC              |        2 |          2 |
| 2010   | a001 | FC Schalke 04           |        1 |          1 | VfL Bochum              |        2 |          2 |
| 2010   | a001 | 1899 Hoffenheim         |        3 |          3 | FC Schalke 04           |        3 |          3 |
| 2010   | a001 | 1. FC Koln              |        3 |          3 | Eintracht Frankfurt     |        3 |          3 |
```

- (4) Unexpected_Home_Wins || Unexpected_Away_Wins: This is application of the previous view table. By comparing expected result with actual result, this table will displays club name with its unexpected number of wins.

```
mysql> SELECT * FROM Unexpected_Home_Wins;
+--------+---------------+-------------------------+-----------------+
| Season | League_name   | Home_T                  | Unexpected_Wins |
+--------+---------------+-------------------------+-----------------+
| 2010   | Bundesliga    | Bayern Munchen          |               7 |
| 2010   | Bundesliga    | FC Schalke 04           |               6 |
| 2010   | Bundesliga    | Eintracht Frankfurt     |               5 |
| 2010   | Bundesliga    | Bayer Leverkusen        |               5 |
| 2010   | Bundesliga    | VfB Stuttgart           |               4 |
| 2010   | Bundesliga    | VfL Wolfsburg           |               3 |
| 2010   | Bundesliga    | Werder Bremen           |               3 |
| 2010   | Bundesliga    | Borussia Dortmund       |               3 |
| 2010   | Bundesliga    | SC Freiburg             |               2 |
| 2010   | Bundesliga    | 1. FSV Mainz 05         |               2 |
| 2010   | Bundesliga    | VfL Bochum              |               1 |
| 2010   | Bundesliga    | Hertha BSC              |               1 |
| 2010   | Bundesliga    | 1. FC Nurnberg          |               1 |
| 2010   | Bundesliga    | Hannover 96             |               1 |
| 2010   | Bundesliga    | Hamburger SV            |               1 |
| 2011   | Bundesliga    | Borussia Dortmund       |               6 |
| 2011   | Bundesliga    | 1. FC Koln              |               6 |
```

```
mysql> SELECT * FROM Unexpected_Away_Wins;
+--------+---------------+-------------------------+-----------------+
| Season | League_name   | Home_T                  | Unexpected_Wins |
+--------+---------------+-------------------------+-----------------+
| 2010   | Bundesliga    | Hertha BSC              |               5 |
| 2010   | Bundesliga    | 1899 Hoffenheim         |               4 |
| 2010   | Bundesliga    | Eintracht Frankfurt     |               3 |
| 2010   | Bundesliga    | 1. FC Nurnberg          |               3 |
| 2010   | Bundesliga    | VfL Bochum              |               3 |
| 2010   | Bundesliga    | VfL Wolfsburg           |               3 |
| 2010   | Bundesliga    | 1. FC Koln              |               3 |
| 2010   | Bundesliga    | FC Schalke 04           |               2 |
| 2010   | Bundesliga    | Werder Bremen           |               2 |
| 2010   | Bundesliga    | 1. FSV Mainz 05         |               2 |
| 2010   | Bundesliga    | SC Freiburg             |               2 |
| 2010   | Bundesliga    | Hannover 96             |               2 |
| 2010   | Bundesliga    | Hamburger SV            |               2 |
| 2010   | Bundesliga    | Borussia Dortmund       |               2 |
| 2010   | Bundesliga    | Bayern Munchen          |               1 |
| 2010   | Bundesliga    | Borussia Monchengladbach |               1 |
| 2011   | Bundesliga    | FC St. Pauli            |               7 |
| 2011   | Bundesliga    | Borussia Monchengladbach |               6 |
```

- (5) WorldCup and League Winning Player : This is player stat table which displays players who won both World Cup and League. If same player has displayed twice, it means he won World Cup and won league more than once with more than one league team.

```
mysql> SELECT * FROM WorldCup_League_Winning_Player;
+---------+-------------------+-----------------------+
| Country | Team              | Player                |
+---------+-------------------+-----------------------+
| Spain   | Real Madrid       | Alvaro Arbeloa        |
| Spain   | Barcelona         | Andres Iniesta        |
| Germany | Bayern Munchen    | Bastian Schweinsteiger |
| Brazil  | Chelsea           | Belletti              |
| Spain   | Barcelona         | Carles Puyol          |
| Spain   | Barcelona         | Cesc Fabregas         |
| Spain   | Manchester City   | David Silva           |
| Spain   | Atletico Madrid   | David Villa           |
| Spain   | Barcelona         | David Villa           |
| Spain   | Barcelona         | Gerard Pique          |
| Spain   | Real Madrid       | Iker Casillas         |
| Spain   | Bayern Munchen    | Javi Martinez         |
| Germany | Bayern Munchen    | Jerome Boateng        |
| Spain   | Manchester City   | Jesus Navas           |
| Brazil  | Real Madrid       | Kaka                  |
| Germany | Borussia Dortmund | Kevin Großkreutz      |
```

- (6) Winnig_Odds : This last view displays interesting statistical result of each team's performance. The table displays the percentage of win at home as well as away games. By looking at this data, observer can easily tell whether team is strong at home or not

```
mysql> SELECT * FROM Home_Winning_Odds LIMIT 10;
+-------------------+--------+--------+--------+
| Home_T            | Win    | Draw   | Lose   |
+-------------------+--------+--------+--------+
| VfL Wolfsburg     | 42.35% | 32.94% | 24.71% |
| 1. FC Nurnberg    | 35.29% | 40.00% | 24.71% |
| 1. FSV Mainz 05   | 48.24% | 29.41% | 22.35% |
| 1899 Hoffenheim   | 32.94% | 31.76% | 35.29% |
| Borussia Dortmund | 67.06% | 15.29% | 17.65% |
| Hertha BSC        | 21.57% | 54.90% | 23.53% |
| Werder Bremen     | 38.82% | 29.41% | 31.76% |
| SC Freiburg       | 38.82% | 37.65% | 23.53% |
| VfL Bochum        | 11.76% | 52.94% | 35.29% |
| 1. FC Koln        | 37.25% | 39.22% | 23.53% |
+-------------------+--------+--------+--------+
10 rows in set (0.19 sec)

mysql> SELECT * FROM Away_Winning_Odds LIMIT 10;
+-------------------------+--------+--------+--------+
| Away_T                  | Win    | Draw   | Lose   |
+-------------------------+--------+--------+--------+
| VfB Stuttgart           | 31.76% | 43.53% | 24.71% |
| FC Schalke 04           | 37.65% | 36.47% | 25.88% |
| Bayer Leverkusen        | 44.71% | 28.24% | 27.06% |
| Bayern Munchen          | 61.18% | 16.47% | 22.35% |
| 1. FC Koln              | 21.57% | 58.82% | 19.61% |
| Hannover 96             | 24.71% | 62.35% | 12.94% |
| Eintracht Frankfurt     | 26.47% | 48.53% | 25.00% |
| Hamburger SV            | 27.06% | 45.88% | 27.06% |
| Borussia Monchengladbach | 28.24% | 49.41% | 22.35% |
| VfL Wolfsburg           | 32.94% | 41.18% | 25.88% |
+-------------------------+--------+--------+--------+
10 rows in set (0.09 sec)
```

4.2 Stored Procedures

- (1) Event Country : This table shows that if we put the country, it will automatically finds you Olympic, World Cup, and League records accordingly. Good thing about this query is that we can either put the whole name of the country or abbreviated national code

Find out about events/big3 teams related to a certain country.
Country: Germany
Submit

Year	Location	Champion
1936 Summer Olympics	Berlin,Germany	Italy
1972 Summer Olympics	Munich,Germany	Poland
2006 World Cup	Germany,Europe	Home Team: 3rd
founded in 1963	Bundesliga	Bayern Munchen

- (2) Country Player : With country code or country input, this query will find observer the all players in top three leagues

Find out information about all the big3 soccer players from a certain country.
Country: Korea
Submit

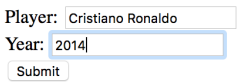
Player	Team	Season	League
Cha Du-Ri	SC Freiburg	2010	Bundesliga
Lee Chung-Yong	Bolton Wanderers	2010	Premier League
Seol Ki-Hyeon	Fulham	2010	Premier League
Park Ji-Sung	Manchester United	2010	Premier League
Cho Won-Hee	Wigan Athletic	2010	Premier League
Son Heung-Min	Hamburger SV	2011	Bundesliga
Koo Ja-Cheol	VfL Wolfsburg	2011	Bundesliga
Koo Ja-Cheol	FC Augsburg	2012	Bundesliga
Son Heung-Min	Hamburger SV	2012	Bundesliga
Koo Ja-Cheol	VfL Wolfsburg	2012	Bundesliga
Koo Ja-Cheol	FC Augsburg	2013	Bundesliga
Ji Dong-Won	FC Augsburg	2013	Bundesliga
Cha Du-Ri	Fortuna Dusseldorf	2013	Bundesliga
Park Jung-Bin	SpVgg Greuther Fuerth	2013	Bundesliga
Son Heung-Min	Hamburger SV	2013	Bundesliga
Hong Jeong-Ho	FC Augsburg	2014	Bundesliga
Ji Dong-Won	FC Augsburg	2014	Bundesliga
Son Heung-Min	Bayer Leverkusen	2014	Bundesliga
Ryu Seung-Woo	Bayer Leverkusen	2014	Bundesliga
Koo Ja-Cheol	1. FSV Mainz 05	2014	Bundesliga
Park Joo-Ho	1. FSV Mainz 05	2014	Bundesliga
Koo Ja-Cheol	VfL Wolfsburg	2014	Bundesliga

- (3) Player Team History : With player name input, this query will find where the player played from 09/10 to 13/14 season if exists

List all the football clubs that a certain player has played for (2009/2010 season ~ 2013/2014 season).
Player: Park Ji-sung
Submit

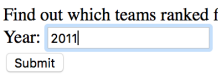
Player	Season	League	Team	No.
Park Ji-Sung	2010	Premier League	Manchester United	13
Park Ji-Sung	2011	Premier League	Manchester United	13
Park Ji-Sung	2012	Premier League	Manchester United	13
Park Ji-Sung	2013	Premier League	Queens Park Rangers	7

- (4) Player In Year : With player name and year input, this query will find which team the player had played if exists



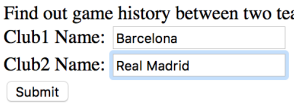
Player	Season	League	Team	No.
Cristiano Ronaldo	2014	La Liga	Real Madrid	7

- (5) Best In Year : With year input, this query will find the best performing (which is winning team) in specified year at all three leagues



Club	League	Season
Borussia Dortmund	Bundesliga	2011
Manchester United	Premier League	2011
Barcelona	La Liga	2011

- (6) Game_History : With input of two different teams, this query will find game history between these two teams if exist



Season	Home	Away	HomeScore	AwayScore
2010	Barcelona	Real Madrid	1	0
2010	Real Madrid	Barcelona	0	2
2011	Barcelona	Real Madrid	5	0
2011	Real Madrid	Barcelona	1	1
2012	Real Madrid	Barcelona	1	3
2012	Barcelona	Real Madrid	1	2
2013	Barcelona	Real Madrid	2	2
2013	Real Madrid	Barcelona	2	1
2014	Barcelona	Real Madrid	2	1
2014	Real Madrid	Barcelona	3	4

- (7) Winning Odds : With team input, this query will tell you the winning odds of specified team both in home and away

Find out Home vs Away winning odds for a certain team.
Club Name: Manchester United
Submit

Home	Club	Win	Draw	Lose
Home	Manchester United	77.89%	14.74%	7.37%
Away	Manchester United	53.68%	20.00%	26.32%

- (8) Individual Club Game History : With given team name, this query will find all games that specified team had played from 09/10 to 13/14 season

Find out scores for all the games that a certain team has played from 2009/2010 season to 2013/2014 season.
Club Name: Barcelona
Submit

Season	Club	Against	Score
2010	Barcelona	Mallorca	1:0
2010	Barcelona	Osasuna	2:0
2010	Barcelona	Espanyol	0:0
2010	Barcelona	La Coruna	3:0
2010	Barcelona	Santander	4:1
2010	Barcelona	Sp Gijon	3:0
2010	Barcelona	Sevilla	3:2
2010	Barcelona	Tenerife	4:1
2010	Barcelona	Valencia	0:0
2010	Barcelona	Almeria	1:0
2010	Barcelona	Athletic Bilbao	1:1
2010	Barcelona	Mallorca	4:2
2010	Barcelona	La Coruna	3:1
2010	Barcelona	Espanyol	1:0

- (9) Winning Prediction : With given two team names, this query will tell the prediction of winning percentage

Find out the probability of club1 winning a match against club2.
Club1 Name: Barcelona
Club2 Name: Real Madrid
Submit

Team1	Team2	Winning Probability
Barcelona	Real Madrid	60.00%

Jin Yong Shin (jshin44)
Ji Won Shin (jshin49)
600.315(01)
Databases Final Project Phase I: Proposal

(1) Team Members

Jin Yong Shin. JHED ID: jshin44
Ji Won Shin. JHED ID: jshin49

(2) Target Domain : World Soccer Records

(a) We will include Olympic Soccer/World Cup records
(b) We will include records of European Soccer Leagues (such as UEFA Champions League, EUROPA Champions Leagues, etc)
(c) We will include national soccer league records of several popular countries (i.e. Germany, England, Italy, France, etc)
(d) We will include records of awards winning soccer players
(e) We will include physical data (i.e. height, weight, etc.) of players of major national league teams

(3) English Questions

(a) List countries which won Olympic Soccer more than once
(b) List countries which won World-Cup more than once
(c) List country/countries which has/have won the most Olympic Soccer medals
(d) List country/countries which has/have won same number of World-Cup winning records as Germany
(e) List country/countries which has/have won Olympic Soccer medal or World-Cup consecutively (twice in a row)
(f) List countries which have won more than two Olympic medals or World Cup and its domestic team has won UEFA Champions League more than twice
(g) List team with player who won the most awards
(h) List teams which has won either UEFA Champions League or EUROPA Champions League and have player who won the award more than once
(i) List teams with most league winning records by its own national league
(j) List team name and awards type with the most awards
(k) List team name and number of awards from the team with the most awards from its current players
(l) List player who are older than 30 and play at German league
(m) List player who won more than one award and is currently playing at the team with most winning records at its own national league
(n) List the oldest and the youngest player and its team and league name at each national league
(o) List team name, league, and total number of winning from the team with all current national players
(p) List the name of players whose league belong to the nation which has won the gold medal in the Olympics in the past 30 years
(q) List the name of the country that has won the most World Cup championship titles

(r) List the name of clubs that Lionel Messi played for at least two seasons

(4) Relational data model (composition and design subject to change)

League	LeagueID	LeagueName	Country
	1	English Premier League	England
	2	La Liga	Spain
	100	Olympic	NULL
	1000	WorldCup	NULL

Olympic	LeagueID	year	Country	medal	Continents
	100	2016	Brazil	Gold	Latin America
	100	2016	Germany	Silver	Europe

WorldCup	LeagueID	year	Country	Final Score	Continents
	1000	2014	Germany	1-0	Europe
	1000	2010	Spain	1-0	Europe

EnglishPremierLeague	LeagueID	Season	Champion
	1	14-15	Leicester City
	1	13-14	Chelsea

UEFA_CHAMPIONS	League_id	Season	Champion
	2	15-16	Real Madrid

SoccerClub	Club_Name	LeagueID
	Chelsea	1
	Barcelona	2

Squad	Club_Name	Season	PName

	Chelsea	14-15	Diego Costa
	Chelsea	14-15	Didier Drogba

Player	PName	Height	Weight
	Diego Costa	6’ 2”	179

UEFA_CHAMPIONS	League_id	Season	Champion
	2	2015-16	Real Madrid

Result	LeagueID	Season	Team1	Team2	Result
	1	14-15	Leicester City	Manchester United	1-0

Player_stat	AwardID	PName	Year	Club_name	Type
	1	Jamie Vardy	15-16	Leicester City	Footballer of Year
	2	Lionel Messi	15-16	Barcelona	FIFA Ballon d’OR

Awards	AwardID	Awards_name
	1	FWA
	2	Ballon d’OR

(5) SQL Statements

List the name of players whose league belong to the nation which has won the gold medal in the Olympics in the past 30 years

```
SELECT S.PName
FROM Squad as S, SoccerClub as SC, League as L
WHERE S.Club_Name = SC.Club_Name and SC.LeagueID = L.LeagueID
        and L.Country in (SELECT Country
                                FROM Olympic
                                WHERE model = “Gold” and year > 1986);
```

List the name of the country that has won the most World Cup championship titles

```
SELECT Country
FROM WorldCup
GROUP BY Country
HAVING count(year) >= all (SELECT count(year)
```

```
FROM WorldCup
GROUP BY Country);
```

List the name of clubs that Lionel Messi played for at least two seasons

```
SELECT Club_Name
FROM Squad
WHERE PName = "Lionel Messi"
GROUP BY Club_Name
HAVING COUNT(Season) >= 2;
```

(6) How to load the databases with values

From an online website named Football-data.co.uk(http://www.football-data.co.uk/), we were able to receive the results of every game played in every major national soccer league such as the English Premiere League, the French Ligue 1, the Italian Serie, the Spanish La Liga, and more. We sent an official request to FIFA asking for the official records of every World Cup game played. From the International Olympic Committee, we could download the official Olympics soccer game results. We also plan on getting the list of players in the roaster for every season from every major national league team.

Since most of these files are csv files, we will make a Java or Python program that receives comma separated values as input and outputs prints to a .sql format file. We also found out that most of these values require processing since they all come from different sources, and these different sources use different formats. We will write a Java or Python program that converts certain representation format of values into another format and thus allow for uniformed representation of values.

(7) Type of reports you plan to generate or any special user interface issues that you plan to implement

From our database that we will create, we will be able to establish all the relevant soccer records data. For example, we can create views with most World Cup / Olympic Soccer winning countries and also most UEFA/EUROPA Champions League winning teams. With these data, we can also find some relevant statistics of which countries are showing most presence in world soccer leagues

(8) Specialized/advanced topics

We will work on complex data extraction issues from online sources. As most of our data comes from different sources (i.e. each respective soccer club, IOC, FIFA), we will have to convert between different formats and create uniform representation of values. We will also integrate advanced SQL topics to reduce complex data extraction issues.

(9) Describe the database platform you plan to use

We will implement our project with mySQL platform and we will also write a Java or Python program to convert csv data files to .sql formatted files. Considering the large amounts of datasets, we will use local machines.

(a) Implement Java/Python algorithm to read csv data files and output to .sql file
(b) Importing .sql file into mysql
(c) Challenging issues:

(i) Data mining → mostly our data files are open source from FIFA or national leagues but need to handle some private data sources such as team statistics or player statistics

(ii) Need to implement appropriate and complex program to read csv file since one csv file does not provide all attributes that are required to build individual table