**Contents:**

1. Motivation
2. Thoughts for initial analysis
3. Description of the project
4. Explanation of World Cup stages
5. Formulation of Entity Relationship (ER) model
6. Loading Data into the Entity Relationship (ER) Model
7. Tables implemented in the database
8. Formulating queries for the data
9. Simple and complex queries implemented on database
10. Facts and important results got from the queries
11. Predicting winner of World Cup based on previous data
12. Difficulties
13. Future work - What we would do next:
14. References
15. **Motivation:**

Football World Cup is one of the most watched sporting events in the world. 32 teams play for the coveted trophy which is up for grabs once during a span of 4 years. A huge amount of data is generated by the events during the world cup which can give us interesting information about players, teams and managers involved in the tournament. Thus there is a lot of scope for analyzing the data from the event in different ways to get better statistics of each game held and the players involved in it. The game is popular across the globe; each and every statistic of the game is discussed debated and also used for predictions. A well formatted database with proper relations can yield many interesting perspectives about the game which can be used by football fans to know their team better.

**2.Thoughts for initial analysis:**

We can use aggregate data to make predictions on future outcomes. For example, individual player information such as the goals scored in the world cup, assists by a player, clean sheets kept by goalkeeper, successful passes completed by a player, shots on target and red/yellow cards received can be used to predict the current form of the player and how well he might perform in the next match. Analysis can also be carried out to see how the previous match form affected the player’s confidence level in the successful matches.

Each team consists of players who may be star players and novices. Its performance might be correlated to the form of their star players in the previous matches. Some other facts we could use for predicting future match outcomes are team age, yellow card count, goals scored and conceded, reliance on star players, frequently used tactics, frequency of injuries, foul count, emerging star players, mental strength (measured by goal frequency towards the end of the game), performance under certain referees and so on. The distribution of these figures helps us compare the teams at various levels.

We could also make recommendations on player, coach and manager selection based on aggregate stats we obtain from the above analysis.

1. **Description of the project:**

The project implements all the features that are described in the proposal. The project flow is as follows

* First, we formulated the ER model for the entire database.
* Secondly, we loaded basic data in the ER model.
* Thirdly, we formulated the queries for the data so that we get useful data out of it.

Lastly, we fitted a formula that correctly predicts the winners of the world cup based on all matches before the round of 8 and can be extended to any other world cup.

1. **Explanation of World Cup stages:**

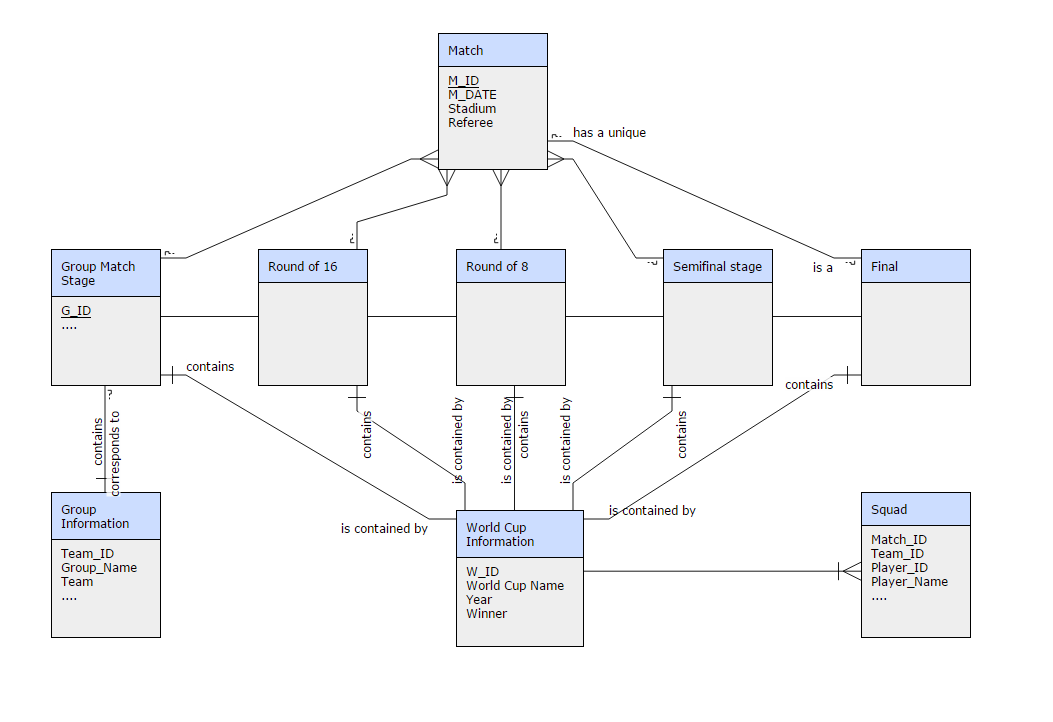
The various stages of the world cup are as follows:

* The first stage is the group stage that consists of 48 matches of 32 teams who were able to qualify for the world-cup in the preliminary play-offs or had a high enough rank that they automatically qualified.
* The second stage is the ‘round of 16’ stage that has 8 matches between various teams. This is decided in order of wins/defeats in the various groups. This stage marks the beginning of the ‘knockout’ stages where a win ensures passage into the next round and a loss means that the team in knocked out from the tournament.
* The third stage is the ‘round of 8’ stage that consists of 4 knockout matches. This stage is also known as the quarterfinal stage.
* The fourth stage is the semifinal with four teams and two matches that decide who qualifies for the finals.
* The finals consist of one match and decide who wins and who is the runner-up for the world cup! Always a tense occasion that is a do-or-die situation for both the qualifying teams.

We formulate an ER model that correctly captures all the matches in various stages of the world cup.

1. **Formulation of Entity Relationship (ER) model:**

First, we formulated the ER model so that it makes sense in English and in the correct syntax of the ER model discussed in class. As seen below, the model is entirely hierarchical in nature. The entire model corresponds to a single world cup and can be replicated if we want to represent data for multiple world cups. The first (top level) entity is the ‘World Cup’ entity that contains a world cup identifier. Each world cup entity has a number of output relations that corresponds to the output stages. The matches contain the data about all the matches in that particular stage. For example, the matches in group A of a particular stage contain data about all the groups in that stage and the attribute value ‘group\_A’. Some of the other attributes contained in the group stage are match id, home team, away team, home score away score among others. Each group stage entity is linked to the ‘round of 16’ stage using a single link relation. This is because each round of 8 stage of a world cup match is uniquely identified by the round of 16 stages that preceded it. The round of 8 stage match is then followed by the semifinal and final stage. Note that each stage entity is linked to multiple matches which are identified by it, except for the final stage which is linked to only one match. Thus we have formulated the ER model that captures all the features of the world cup matches mentioned in the previous section. The match entity itself contains the attributes of the match id, the home team, the away team, goals scores, number of yellow cards, number of red cards and the winner of the match or whether it resulted in a draw. In addition to these basic things, it also contains the stadium where the match was played and the referee who officiated the match.



**Figure 1 – Entity Relationship (ER) Model for the world cup**

1. **Loading Data into the Entity Relationship (ER) Model:**

We need to get the answers in a number of tables. There are 8 tables that are present in the database. They are WORD\_CUP\_INFORMATION, GROUP\_INFORMATION, SQUAD, FINAL\_STAGE, SEMIFINAL\_STAGE, ROUND\_OF\_8, ROUND\_OF\_16, GROUP\_MATCH\_STAGE.

We obtained data from three sources.

1. The first source was <https://github.com/openfootball/world-cup>, which is the open football database found on github. This repository contains data for each world cup match, namely the people who scored in each match, the time at which they scored, squads, match results and final winners of each world cup.
2. The second source was [www.fifa.com](http://www.fifa.com), which contains data for all players and statistics of each player in all matches of the world cups. This was useful for computing the contribution of each player to a match. We loaded the data of players only for the world cup finals in this case because the amount of data was quite big. We wrote a parser in perl to do this job for us.
3. The last source was Wikipedia, where we obtained statistics about each match and populated the match table. The Match table contained the players, the winners, losers, home team, away team, referees and generally all the information described by the attributes of the match entity.
4. **Tables implemented in the database:**

We created 10 tables. Here are the following details of the tables-

* **Groups (13 columns)** - This consists of the W\_ID, GROUP\_ID, TEAM\_ID, GROUP\_NAME, TEAM\_NAME, MATCHES\_PLAYED, MATCHES\_WON, MATCHES\_DRAWN, MATCHES\_LOST, GOALS\_FOR, GOALS\_AGAINST, GOAL\_DIFFERENCE, POINTS.

**Primary Key – W\_ID, GROUP\_ID, TEAM\_ID**

**Foreign Key – W\_ID**

* **Matches (15 columns)** - This consists of the W\_ID, GROUP\_ID, M\_ID, M\_DATE, STADIUM, HOME\_TEAM, HOME\_TEAM\_SCORE, AWAY\_TEAM\_SCORE, AWAY\_TEAM, REFEREE, WINNER, YELLOW\_CARD\_HOME, YELLOW\_CARD\_AWAY, RED\_CARD\_HOME, RED\_CARD\_AWAY.

**Primary Key – W\_ID, M\_ID**

**Foreign Key – W\_ID**

* **Round of 16 (14 columns)** - This consists of the W\_ID, GROUP\_ID, M\_ID, M\_DATE, STADIUM, HOME\_TEAM, HOME\_TEAM\_SCORE, AWAY\_TEAM\_SCORE, AWAY\_TEAM, REFEREE, WINNER, YELLOW\_CARD\_HOME, YELLOW\_CARD\_AWAY, RED\_CARD\_HOME, RED\_CARD\_AWAY.

**Primary Key – W\_ID, M\_ID**

**Foreign Key – W\_ID**

* **Round of 8 (14 columns)** - This consists of the W\_ID, GROUP\_ID, M\_ID, M\_DATE, STADIUM, HOME\_TEAM, HOME\_TEAM\_SCORE, AWAY\_TEAM\_SCORE, AWAY\_TEAM, REFEREE, WINNER, YELLOW\_CARD\_HOME, YELLOW\_CARD\_AWAY, RED\_CARD\_HOME, RED\_CARD\_AWAY.

**Primary Key – W\_ID, M\_ID**

**Foreign Key – W\_ID**

* **Semis (14 columns)** – This consists of the W\_ID, GROUP\_ID, M\_ID, M\_DATE, STADIUM, HOME\_TEAM, HOME\_TEAM\_SCORE, AWAY\_TEAM\_SCORE, AWAY\_TEAM, REFEREE, WINNER, YELLOW\_CARD\_HOME, YELLOW\_CARD\_AWAY, RED\_CARD\_HOME, RED\_CARD\_AWAY.

**Primary Key – W\_ID, M\_ID**

**Foreign Key – W\_ID**

* **Finals (15 columns)** - This consists of the W\_ID, GROUP\_ID, M\_ID, M\_DATE, STADIUM, HOME\_TEAM, HOME\_TEAM\_SCORE, AWAY\_TEAM\_SCORE, AWAY\_TEAM, REFEREE, WINNER, YELLOW\_CARD\_HOME, YELLOW\_CARD\_AWAY, RED\_CARD\_HOME, RED\_CARD\_AWAY.

**Primary Key – W\_ID, M\_ID**

**Foreign Key – W\_ID**

* **Team\_Stat (28 columns)** – W\_ID, M\_ID, T\_ID, PLAYER\_ID, PLAYER\_NAME, MANAGER\_NAME, MANAGER\_ID, MATCH\_PLAYED, TOTAL\_TIME\_PLAYED, LOW\_ACTIVITY\_SPENT, MEDIUM\_ACTIVITY\_SPENT, HIGH\_ACTIVITY\_SPENT, LOW\_ACTIVITY\_DISTANCE\_COVERED, M\_ACTIVITY\_DISTANCE\_COVERED, HIGH\_ACTIVITY\_DISTANCE\_COVERED, TOP\_SPEED, SPRINT, ASSISTS, OFFSIDE, FOUL\_COMMITTED, CLEARANCE\_COMPLETED, SHOTS, GOALS, SAVES, YELLOW\_CARD, RED\_CARD, POSITION\_OFF\_PLAY, STAR\_STATUS.

**Primary Key – W\_ID, PLAYER\_ID**

**Foreign Key – W\_ID**

* **World\_Cup\_Name (4 columns)** – W\_ID, WORLD\_CUP\_NAME, YEAR, WINNER.

**Primary Key – W\_ID and Foreign Key constraint – WC\_PK\_W\_ID**

1. **Formulating queries for the data:**

We formulated queries for the data for the ultimate purpose of obtaining an ability to compute the final winners of the next world cup. In order to do so, we first started out with simple queries, such as obtaining the team which scored the maximum/minimum number of goals in a match, teams who had the most number of yellow cards or read cards and the least number of red/yellow cards.

Lastly we calculated the form of each team by weighting each of these factors and obtained a formula for the solution. The prediction works for all world cups we tested except for world cup 2006.

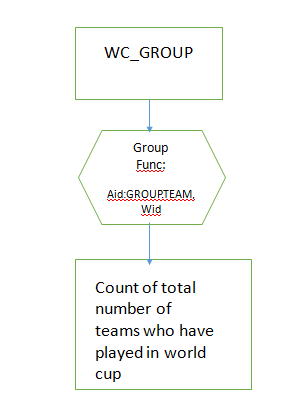
1. **Simple and complex queries implemented on database:**

We have listed out some queries to find interesting facts and results from the database. Some of the queries have been listed below:

**1. Number of distinct teams who played in all the WCs in group stages**

**SELECT COUNT(DISTINCT WC\_GROUP.TEAM) FROM WC\_GROUP**

Result: 64



**Figure 2 – Relational algebra format for Query 1**

**Other queries follow a similar format of relational algebra flowchart**

**2. Number of HOME\_TEAM who played in all the WCs in group stages**

**SELECT MATCH\_INFO.HOME\_TEAM FROM MATCH\_INFO GROUP BY MATCH\_INFO.HOME\_TEAM**

Result: 62

**3. Total number of points won by each team in all WCs combined together in group stages**

**SELECT SUM(WC\_GROUP.PTS), WC\_GROUP.TEAM FROM WC\_GROUP GROUP BY WC\_GROUP.TEAM**

Result:

SUM(WC\_GROUP.PTS) TEAM

6 Poland

45 Brazil

14 Croatia

1. South Africa

**4. Team statistics for all WCs – number of matches played, won, lost and draw**

**SELECT WC\_GROUP.TEAM,**

**SUM(WC\_GROUP.PLAYED),**

**SUM(WC\_GROUP.W),**

**SUM(WC\_GROUP.L),**

**SUM(WC\_GROUP.D)**

**FROM WC\_GROUP**

**GROUP BY WC\_GROUP.TEAM**

Result:

TEAM SUM(WG.PLAYED) SUM(WG.W) SUM(WG.L) SUM(WG.D)

Poland 6 2 4 0

Brazil 18 12 1 3

Croatia 12 4 6 2

South Africa 9 2 3 4

**5. Teams which have played <= 3 group matches in all WCs and their corresponding group points**

**SELECT WC\_GROUP.TEAM,**

**SUM(WC\_GROUP.PLAYED),**

**SUM(WC\_GROUP.W),**

**SUM(WC\_GROUP.L),**

**SUM(WC\_GROUP.D),**

**SUM(WC\_GROUP.PTS),**

**SUM(WC\_GROUP.PTS)**

**FROM WC\_GROUP**

**GROUP BY WC\_GROUP.TEAM**

**HAVING SUM(WC\_GROUP.PLAYED) <= 3**

**ORDER BY SUM(WC\_GROUP.PTS) DESC**

Result:

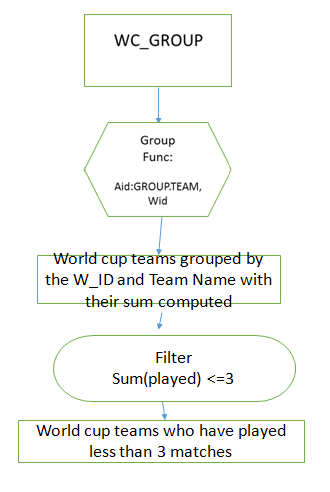
TEAM SUM(WG.PLAYED) SUM(WG.W) SUM(WG.L) SUM(WG.D) SUM(WG.PTS) SUM(WG.PTS)\_1

Yugoslavia 3 2 0 1 7 7

Ukraine 3 2 1 0 6 6

Senegal 3 1 0 2 5 5

Turkey 3 1 1 1 4 4



**Figure 3 – Relational algebra format for Query 5**

**Similar queries having group by and followed by filter will have similar relational algebra flowchart**

**6. Teams which have played > 6 WC matches and achieved < 7 group points**

**SELECT WC\_GROUP.TEAM,**

**SUM(WC\_GROUP.PLAYED),**

**SUM(WC\_GROUP.W),**

**SUM(WC\_GROUP.L),**

**SUM(WC\_GROUP.D),**

**SUM(WC\_GROUP.PTS),**

**SUM(WC\_GROUP.PTS)**

**FROM WC\_GROUP**

**GROUP BY WC\_GROUP.TEAM**

**HAVING SUM(WC\_GROUP.PLAYED) > 6**

**AND SUM(WC\_GROUP.PTS) < 7**

**AND SUM(WC\_GROUP.PTS) < 7**

**ORDER BY SUM(WC\_GROUP.PTS) DESC**

Result:

TEAM SUM(WG.PLAYED) SUM(WG.W) SUM(WG.L) SUM(WG.D) SUM(WG.PTS) SUM(WG.PTS)\_1

Iran 9 1 6 2 5 5

Tunisia 9 0 6 3 3 3

**7. Average number of wins, draws and losses per team per WC in group matches**

**SELECT WC\_GROUP.TEAM,**

**AVG(WC\_GROUP.GF),**

**SUM(WC\_GROUP.GA)**

**FROM WC\_GROUP**

**GROUP BY WC\_GROUP.TEAM**

**ORDER BY SUM(WC\_GROUP.GF) DESC,**

**AVG(WC\_GROUP.GF) DESC,**

**SUM(WC\_GROUP.GA) DESC**

Result:

TEAM AVG(WC\_GROUP.D) AVG(WC\_GROUP.W) AVG(WC\_GROUP.L)

Poland 0 1 2

Brazil 0.5 2 0.16667

Croatia 0.5 1 1.5

South Africa 1.33333 0.66667 1

**8. Average number of wins, draws and losses per team for all WCs in group matches**

**SELECT WC\_GROUP.TEAM,**

**AVG(WC\_GROUP.D),**

**AVG(WC\_GROUP.W),**

**AVG(WC\_GROUP.L)**

**FROM WC\_GROUP**

**GROUP BY WC\_GROUP.TEAM**

Result:

AVG(WC\_GROUP.W) AVG(WC\_GROUP.D) W\_ID

1.15625 0.6875 W\_1

1.0625 0.875 W\_2

1.21875 0.5625 W\_3

1.16667 0.66667 W\_4

1 1 W\_5

1 0.875 W\_6

**9. Total number of wins, draws and losses (in descending order in that precedence) for each WC in group matches**

**SELECT WC\_GROUP.W\_ID,**

**SUM(WC\_GROUP.W),**

**SUM(WC\_GROUP.D),**

**SUM(WC\_GROUP.L)**

**FROM WC\_GROUP**

**GROUP BY WC\_GROUP.W\_ID**

**ORDER BY SUM(WC\_GROUP.W) DESC,**

**SUM(WC\_GROUP.D) DESC,**

**SUM(WC\_GROUP.L) DESC**

Result:

W\_ID SUM(WC\_GROUP.W) SUM(WC\_GROUP.D) SUM(WC\_GROUP.L)

W\_3 39 18 39

W\_1 37 22 37

W\_2 34 28 34

W\_5 32 32 32

W\_6 32 28 34

W\_4 28 16 28

**10. Aggregate number of wins, red and yellow cards received by a particular team in each world cup in group stages (note it can be implemented for all stages of world cup also)**

**(Note: This is used in Match prediction along with results obtained from earlier simple queries)**

**SELECT MATCH\_INFO.HOME\_TEAM,**

**MATCH\_INFO.W\_ID,**

**SUM(MATCH\_INFO.YELLOW\_CARD\_HOME),**

**SUM(MATCH\_INFO.RED\_CARD\_HOME)**

**FROM MATCH\_INFO**

**WHERE MATCH\_INFO.HOME\_TEAM = 'Germany'**

**GROUP BY MATCH\_INFO.HOME\_TEAM,**

**MATCH\_INFO.W\_ID**

**SELECT MATCH\_INFO.AWAY\_TEAM,**

**MATCH\_INFO.W\_ID,**

**SUM(MATCH\_INFO.YELLOW\_CARD\_AWAY),**

**SUM(MATCH\_INFO.RED\_CARD\_AWAY)**

**FROM MATCH\_INFO**

**WHERE MATCH\_INFO.AWAY\_TEAM = 'Uruguay'**

**GROUP BY MATCH\_INFO.AWAY\_TEAM,**

**MATCH\_INFO.W\_ID**

Result:

Germany W\_2 5 1

Germany W\_6 2 0

Germany W\_4 6 0

Germany W\_3 2 0

Germany W\_5 6 0

Germany W\_1 4 0

Uruguay W\_6 9 0

Uruguay W\_3 2 0

Uruguay W\_2 1 0

**11. How can the players of Germany and Argentina compare among themselves and how can we predict which team wins according to the player statistics**

**SELECT WC\_MATCH\_TEAM\_STAT.T\_ID,**

**AVG(WC\_MATCH\_TEAM\_STAT.LOW\_ACTIVITY\_SPENT),**

**AVG(WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_SPENT),**

**AVG(WC\_MATCH\_TEAM\_STAT.MEDIUM\_ACTIVITY\_SPENT),**

**AVG(WC\_MATCH\_TEAM\_STAT.LOW\_ACTIVITY\_DISTANCE\_COVERED),**

**AVG(WC\_MATCH\_TEAM\_STAT.M\_ACTIVITY\_DISTANCE\_COVERED),**

**SUM(WC\_MATCH\_TEAM\_STAT.SHOTS),**

**SUM(WC\_MATCH\_TEAM\_STAT.GOALS),**

**SUM(WC\_MATCH\_TEAM\_STAT.FOUL\_COMMITTED),**

**AVG(WC\_MATCH\_TEAM\_STAT.SPRINT),**

**AVG(WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED)**

**FROM WC\_MATCH\_TEAM\_STAT**

**GROUP BY WC\_MATCH\_TEAM\_STAT.T\_ID**

Result:

W3G1T25 84.2143 8.64286 7.14286 5971.5 1416.64 10 0 16 34.7857 2497.36

W3G1T26 81.2857 10.2143 8.5 6108.86 1748.29 10 1 20 39.2857 2891.93

**12. The laziest player in the final match between Germany and Argentina**

**SELECT MIN(WC\_MATCH\_TEAM\_STAT.LOW\_ACTIVITY\_SPENT),**

**MIN(WC\_MATCH\_TEAM\_STAT.LOW\_ACTIVITY\_DISTANCE\_COVERED),**

**WC\_MATCH\_TEAM\_STAT.PLAYER\_ID,**

**WC\_MATCH\_TEAM\_STAT.PLAYER\_NAME,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_SPENT,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED**

**FROM WC\_MATCH\_TEAM\_STAT**

**GROUP BY WC\_MATCH\_TEAM\_STAT.PLAYER\_ID,**

**WC\_MATCH\_TEAM\_STAT.PLAYER\_NAME,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_SPENT,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED**

**ORDER BY MIN(WC\_MATCH\_TEAM\_STAT.LOW\_ACTIVITY\_SPENT) DESC,**

**MIN(WC\_MATCH\_TEAM\_STAT.LOW\_ACTIVITY\_DISTANCE\_COVERED) DESC,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_SPENT,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED**

Result:

ROMERO 98 3657.0 P1 1 389

NEUER 96 5583.0 P15 2.0 841

.

.

.

.

MUELLER 80 8355.0 P21 11.0 4386.0

OEZIL 80 7437.0 P19 12.0 4582.0

**13. Performance of players grouped according to their position of play**

**SELECT WC\_MATCH\_TEAM\_STAT.POSITION\_OF\_PLAY,**

**SUM(WC\_MATCH\_TEAM\_STAT.CLEARANCE\_COMPLETED),**

**SUM(WC\_MATCH\_TEAM\_STAT.FOUL\_COMMITTED),**

**SUM(WC\_MATCH\_TEAM\_STAT.OFFSIDE),**

**AVG(WC\_MATCH\_TEAM\_STAT.SPRINT),**

**AVG(WC\_MATCH\_TEAM\_STAT.TOP\_SPEED),**

**AVG(WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED),**

**AVG(WC\_MATCH\_TEAM\_STAT.M\_ACTIVITY\_DISTANCE\_COVERED),**

**AVG(WC\_MATCH\_TEAM\_STAT.LOW\_ACTIVITY\_DISTANCE\_COVERED)**

**FROM WC\_MATCH\_TEAM\_STAT**

**GROUP BY WC\_MATCH\_TEAM\_STAT.POSITION\_OF\_PLAY**

Result:

FWW 0 2 0 30 31.07 1888 1022

GK 2 0 0 7 29.84 615 374

DF 15 6 0 37.5 28.08 2694.37 1733.37

FW 3 13 4 41.25 29.35 2602.62 1254.62

MF 6 15 1 40.3 27.46 3328.44 2070.55

**14. Player ranking depending on all the statistics (based on given order of priority) in the match**

**SELECT WC\_MATCH\_TEAM\_STAT.STAR\_STATUS,**

**WC\_MATCH\_TEAM\_STAT.GOALS,**

**WC\_MATCH\_TEAM\_STAT.SHOTS,**

**WC\_MATCH\_TEAM\_STAT.FOUL\_COMMITTED,**

**WC\_MATCH\_TEAM\_STAT.CLEARANCE\_COMPLETED,**

**WC\_MATCH\_TEAM\_STAT.ASSSTS,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_SPENT,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED,**

**WC\_MATCH\_TEAM\_STAT.PLAYER\_NAME**

**FROM WC\_MATCH\_TEAM\_STAT**

**GROUP BY WC\_MATCH\_TEAM\_STAT.STAR\_STATUS,**

**WC\_MATCH\_TEAM\_STAT.GOALS,**

**WC\_MATCH\_TEAM\_STAT.SHOTS,**

**WC\_MATCH\_TEAM\_STAT.FOUL\_COMMITTED,**

**WC\_MATCH\_TEAM\_STAT.CLEARANCE\_COMPLETED,**

**WC\_MATCH\_TEAM\_STAT.ASSSTS,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_SPENT,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED,**

**WC\_MATCH\_TEAM\_STAT.PLAYER\_NAME**

**ORDER BY WC\_MATCH\_TEAM\_STAT.GOALS DESC,**

**WC\_MATCH\_TEAM\_STAT.SHOTS DESC,**

**WC\_MATCH\_TEAM\_STAT.FOUL\_COMMITTED,**

**WC\_MATCH\_TEAM\_STAT.CLEARANCE\_COMPLETED DESC,**

**WC\_MATCH\_TEAM\_STAT.ASSSTS,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_SPENT,**

**WC\_MATCH\_TEAM\_STAT.HIGH\_ACTIVITY\_DISTANCE\_COVERED,**

**WC\_MATCH\_TEAM\_STAT.STAR\_STATUS**

Result:

4 1 2 1 1 0 19.0 2239.0 GOETZE

1 0 4 1 0 0 5 2066.0 MESSI

4 0 3 3 0 0 9.0 3272.0 KROOS

4 0 2 1 0 0 10 2411.0 HIGUAIN

5 0 2 2 0 0 7 1888.0 AGUERO

4 0 2 2 0 1 12.0 3740.0 SCHUERRLE

.

.

.

.

**15. Implementing JOIN operator - Group statistics for a particular team given as input by the user along with the WC general statistics for all the WCs in which the team has played**

**SELECT WORLD\_CUP\_NAME.W\_ID,**

**WORLD\_CUP\_NAME.YEAR,**

**WORLD\_CUP\_NAME.WINNER,**

**MATCH\_INFO.WINNER AS WINNER1,**

**COUNT(MATCH\_INFO.WINNER) AS WINNER2,**

**SUM(MATCH\_INFO.RED\_CARD\_HOME),**

**SUM(MATCH\_INFO.YELLOW\_CARD\_HOME),**

**WORLD\_CUP\_NAME.WORLD\_CUP\_N**

**FROM MATCH\_INFO**

**INNER JOIN WORLD\_CUP\_NAME**

**ON WORLD\_CUP\_NAME.W\_ID = MATCH\_INFO.W\_ID**

**WHERE MATCH\_INFO.WINNER = 'Brazil'**

**GROUP BY WORLD\_CUP\_NAME.W\_ID,**

**WORLD\_CUP\_NAME.YEAR,**

**WORLD\_CUP\_NAME.WINNER,**

**MATCH\_INFO.WINNER,**

**WORLD\_CUP\_NAME.WORLD\_CUP\_N**

Result:

W\_2 2010 Spain South Africa Brazil 2 1 2

W\_4 1994 Brazil United States Brazil 2 0 1

W\_3 2014 Germany Brazil Brazil 2 0 2

W\_1 2006 Italy Germany Brazil 3 0 5

W\_6 2002 Brazil South Korea Japan Brazil 3 0 3

W\_5 1998 France France Brazil 2 0 4

.

.

.

.

**16. Venues hosting number of matches in descending order for all WCs in group matches**

**SELECT WORLD\_CUP\_NAME.W\_ID,**

**MATCH\_INFO.STADIUM,**

**COUNT(MATCH\_INFO.STADIUM) AS STADIUM1,**

**WORLD\_CUP\_NAME.YEAR**

**FROM MATCH\_INFO**

**INNER JOIN WORLD\_CUP\_NAME**

**ON WORLD\_CUP\_NAME.W\_ID = MATCH\_INFO.W\_ID**

**GROUP BY WORLD\_CUP\_NAME.W\_ID,**

**MATCH\_INFO.STADIUM,**

**WORLD\_CUP\_NAME.YEAR**

**ORDER BY STADIUM1 DESC**

Result:

W\_2 Loftus Versfeld Stadium, Pretoria 5 2010

W\_2 Cape Town Stadium, Cape Town 5 2010

W\_5 Stade Geoffroy-Guichard, Saint-Étienne 5 1998

W\_2 Ellis Park Stadium, Johannesburg 5 2010

.

.

.

.

**17. Referee statistics for each WC in group matches**

**SELECT WORLD\_CUP\_NAME.W\_ID,**

**WORLD\_CUP\_NAME.YEAR,**

**MATCH\_INFO.REFREE,**

**COUNT(MATCH\_INFO.REFREE) AS REFREE1**

**FROM MATCH\_INFO**

**INNER JOIN WORLD\_CUP\_NAME**

**ON WORLD\_CUP\_NAME.W\_ID = MATCH\_INFO.W\_ID**

**WHERE WORLD\_CUP\_NAME.W\_ID = 'W\_2'**

**GROUP BY WORLD\_CUP\_NAME.W\_ID,**

**WORLD\_CUP\_NAME.YEAR,**

**MATCH\_INFO.REFREE**

**ORDER BY REFREE1 DESC**

Result:

W\_2 2010 Ravshan Irmatov 3

W\_2 2010 Jorge Larrionda 3

W\_2 2010 Yuichi Nishimura 3

W\_2 2010 Héctor Baldassi 3

.

.

.

.

**18. Date statistics for each WCs in group matches**

**SELECT MATCH\_INFO.W\_ID,**

**MATCH\_INFO.M\_DATE,**

**COUNT(MATCH\_INFO.M\_DATE) AS M\_DATE1,**

**MATCH\_INFO.W\_ID AS W\_ID1**

**FROM MATCH\_INFO**

**WHERE MATCH\_INFO.W\_ID = 'W\_1'**

**GROUP BY MATCH\_INFO.W\_ID,**

**MATCH\_INFO.M\_DATE**

Result:

W\_1 13.06.2014 3

W\_1 10.06.2014 3

W\_1 16.06.2014 3

W\_1 23.06.2014 4

.

.

.

.

1. **Facts and important results got from the queries:**
2. Iran and Tunisia have played in 3 WCs (9 matches) and have not accumulated as many points as Yugoslavia did in 1 WC.
3. Most entertaining results WC in group matches - 2014 WC having the maximum number of wins in group matches. Most boring results WC in group matches - 1998 WC having maximum number of draws in group matches. 1998 WC in group matches has equal number of wins (or losses) and draws = 32
4. Referee Statistics (Format: Year - Maximum number of matches officiated by a referee, Number of referees officiating in that world cup)

1994 – (2, 28), 1998 – (2, 43), 2002 – (2, 37)

2006 – (3, 24), 2010 – (3, 25), 2014 – (4, 26)

1. Number of different match dates for group matches per WC

1994 – 14, 1998 – 17, 2002 – 15, 2006 – 15, 2010 – 15, 2014 – 14

Largest ongoing WC group matches is 1998

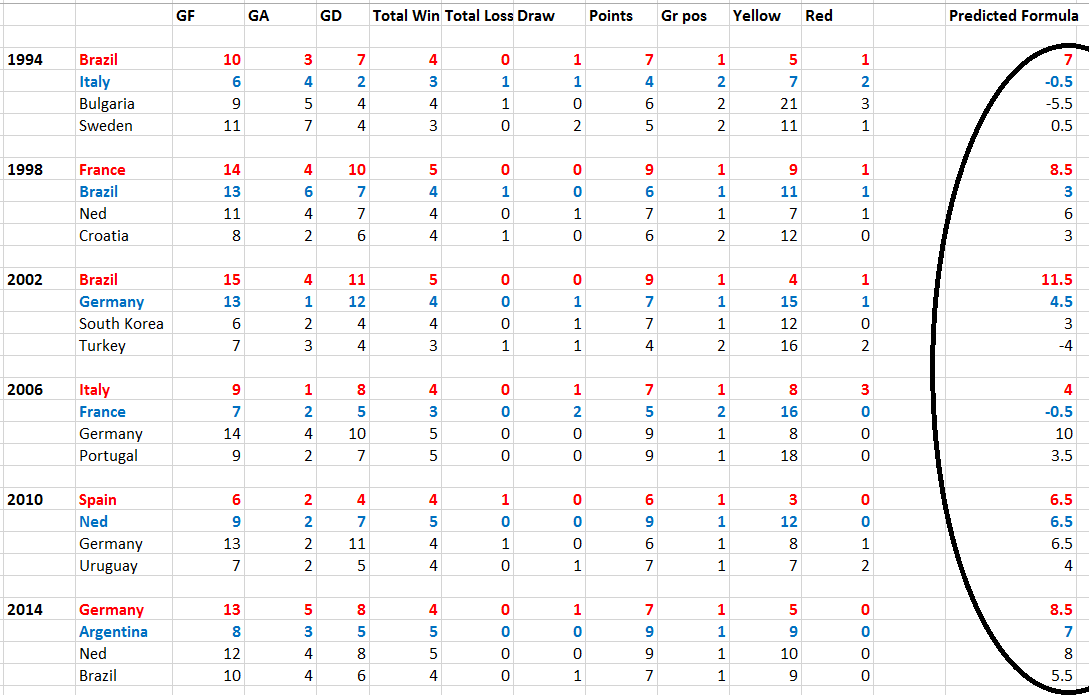
1. Brazil and Germany scored 42 goals each in total 6 WCs. Brazil conceded 12 while Germany conceded 11. China PR and Trinidad and Tobago scored 0 each (least) and conceded 9 and 4 goals respectively.
2. **Predicting winner of World Cup based on previous data:**

We consider the following things as parameters for predicting the **world cup winner from the four semifinalists** we have at a certain world cup. It also includes the statistics from previous world cups. The prediction formula is given below:

**Group Points + (Goal Difference)/2.0 – Total red cards – (Total Yellow)/2**

If there is tie of points we consider a team having previous world cup semifinal or final experience as a probable qualifier. For predicting the other finalist, we can extend this formula and use regression. 2006 world cup is the only case which does not fit this scenario as Germany (10) has more points than Italy (4) albeit Italy winning the world cup and Germany merely settling as the 3rd team.

The formula was derived using the facts that the team form is based on the matches won, lost and drawn in group stages. Also we have certain points negated for the yellow and red cards as they affect the overall morale of the team, may also cause suspensions. Goal difference is considered in all the previous matches as it directly reflects the playing style of the team in context. Previous world cup finals experience gives an odd on confidence factor which gives an added advantage, this parameter is only considered when there is a tie in 2 teams after calculating points from above formula.



**Figure 4 – Results of prediction formula satisfying most of the winners of the 6 WCs considered**

1. **Difficulties:**

We encountered the following difficulties:

* Entering data was time consuming, as it was collected in raw for from various sources.
* Analyzing data from different tables.
* While loading data to tables, difference in format (like including spaces before names) caused problems such as changing uniformity of data in a column.
* Some world cup has fewer teams participating than other world cups teams affecting structure of database.
* A team can play both as HOME\_TEAM and AWAY\_TEAM making aggregate analysis of wins/losses/draws/yellow and red cards difficult for a particular team.

1. **Future work - What we would do next:**

We want to use a more complete database by loading data of more world cups and also player data for all matches of the world cup. This was not possible so far due to a lack of time. We also could use a more complex prediction engine based on machine learning and use a number of complex relations in order to obtain the final performance score of likelihood of winning of a team. For example, the ‘Castrol-Index’ found on the FIFA webpage uses a large number of factors to calculate likelihood of a team performing well/seeding the team. Every tackle and sprint of a player is considered while weighting the final performance index. We can extend the database to include local, regional, national and popular football leagues across the world (like EPL, Champions League, Spanish La Liga). We could also find out more interesting facts from the database with more complex queries.

Some of the more complex techniques that could be used include cluster analysis, Principal Component Analysis, Stochastic modeling and Regression analysis that are used commercially today.

1. **References:**

Data collected to the database from [www.wikipedia.com](http://www.wikipedia.com) and [www.fifa.com](http://www.fifa.com). Other two sources we shall use are [https://github.com/openfootball](https://github.com/openfootball/world-cup) and <https://www.worldcup-history.com>. The specific source on github was <https://github.com/openfootball/world-cup>, which is the open football database found on github.