Data Warehousing: Soccer statistics

University of Applied Sciences Ulm
Onur Yavuz, Eugenio Donaque, Artur
Baliet, Hannes Daniel

***Abstract*—It's a fact that soccer is the most popular sport on this planet. Matches from previous seasons are compared again and again to understand the development of teams; and it's a fact that no other sport like soccer is source to so many bets in the world. Understanding how a team performs might give a better profiling and in turn, better profit for betting platforms or gamblers. Having understandable information that makes it easy -not only for fans and profit-seekers, but also for algorithms- to extract and gain valuable knowledge from is often not an straightforward solution to come upon. It is very likely that, to gain such knowledge, one may try to come up with solutions following inadequate processes that in turn lead to inconsistent or erroneous results. By following the CRISP-DM model and implementing a Data Warehouse, we have developed a solution that allows for better analyzing and reporting, producing the desired knowledge and providing a framework that lessens the hardship of recreating the results obtained consistently.**

I. Introduction

Within the content of the course "Data Warehousing" an analytical project was carried out that concerned with soccer statistics about the Bundesliga in Germany. The soccer statistics data for the Bundesliga involves a huge amount of data about matches, without directly giving information on the performance of a team. To give the soccer interested fans and viewers an useful overview to a team's performance, we use the collected data and transformed it into relevant information in a compact form for public's understanding. The task was then to enable useful information out of the data, and involved the collection and preparation of the open data for improving the quality and availability of the data for analytical and reporting purposes. In order to create a good model that would suit the goal in question, various techniques introduced over the course were applied to generate a Data Warehouse that would model the information gathered in a scheme that would answer to the question: What is the performance of teams in the past seasons?

A. Scenario

A big amount of data from the past Bundesliga games is needed to give the soccer interested parties an useful overview. The fans and viewers can get understandable information about the matches, goals, red and yellow cards and shots (on target and total). This data is relevant for the followers and also for these who do bets. For the league position the information about the matches, points and results is necessary.

B. Structure of the Paper

In order to achieve a logical structure of the project, it's sensible to orientate us by the CRISP-DM model. The following steps don't include only the CRISP-DM approach! The paper is structured as follows: in Section **??** we present the availability of Open Data we use in this project. Then, in Section **??** we describe our tools to create our Data Warehouse and to generate analysis from it. Section **??** we describe the Data Preparation for our Data Warehouse. In Section **??** we explain the how we input the data and the CDHW and the Data Mart. In Section **??** we analyze the result of our data. Finally, we conclude our work in Section **??** with our achievements and a feedback of this project.

II. Open Data: German Bundesliga

As part of *Data Understanding* phase for our project, the dataset we used is sourced from http://www.football-data.co.uk/ website and is open data. This dataset contains data for last 10 seasons of German Bundesliga including the current season. The data itself contained the results of the matches themselves, and did not specify the location of each game.

III. Concept for Problem

The goal is to give users an understandable view of the performance of a team. Part of that goal is that users can obtain answers based on questions such as when and where. The raw data did not contain specifically the parameters for one team only, nor did it contain explicit information of the location of the match, therefore solutions had to be found. Then was the question on how

to treat all of this data to achieve the other part of the goal: how to make it an understandable view? For that, the following tools have been used:

The tools we used are:

1) XAMPP to launch apache & MySQL servers on the local machine
2) Microsoft Excel to cleanse and upload data to the databases
3) MySQL Workbench to create and forward engineer ER-Diagrams into the databases
4) MySQL workbench & PHPMyAdmin to execute all SQL scripts to transform raw data into the data warehouse
5) Microsoft Excel to generate Pivot Tables and Charts for analytical purposes

IV. Data Preparation

After inspecting the csv files from the data source, it was realized that many columns were not needed. Thus keeping only the information of interest such as date and division of the match played, home and away team, and for both of them the amount of goals, shots, shots on target, red and yellow cards were kept too.

The date format in the data source didn't match the representation we preferred to talk about: Typically sports leagues are distributed into match weeks or rounds, where teams get to face one team per round and all teams play. The German Bundesliga has in total 34 match weeks, with a sum of 9 matches per match week. The Date information was then transformed to an integer value between [1, 34] for this representation.

Other information of interest that did not come from the data source was the location of each match. From the 10 seasons the data source supplied, a list of all teams that took part playing as home team was made and then information on the name of the stadium and the state in which the stadium is located were searched to compile another data source.

All this information would be compiled and distributed into the data warehouse designed for the goal. Its ER Diagram looks like this:

It's follows a description of some columns:

- countyName: Name of the country → Germany
- stateName: Name of the state → Baden-Württemberg, Rheinland-Pfalz
- idStadium: Stadium identification → Allianz Arena
- idTeam: Unique team name → FC Bayern Münich
- idDivision: which division → 1. Bundesliga
- weekNum: Play day in german is called week → 2. Play day = 2. Week
- idSeason: season 18/19

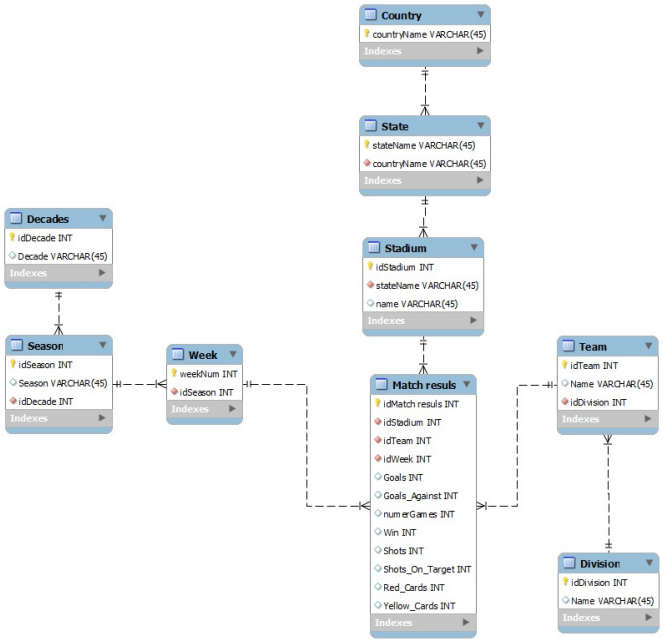

Fig. 1. Snowflake Schema

But since the purpose of a Data Warehouse is also to make queries quicker and simpler, we extracted the data from the data warehouse into a data mart which looks like this:

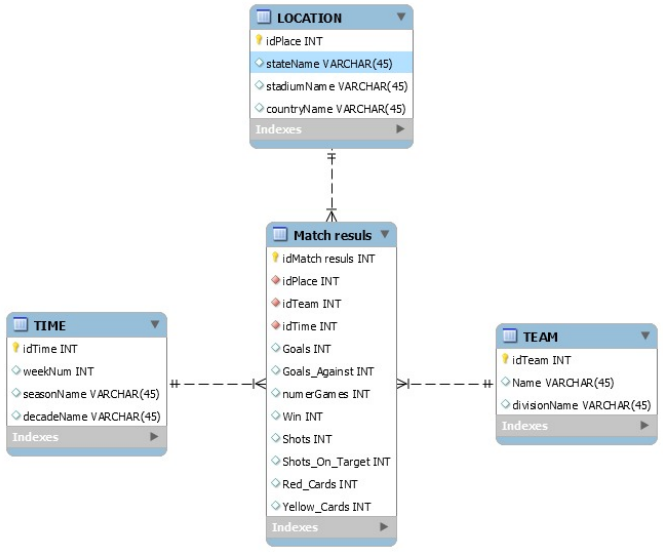

Fig. 2. Data Mart

V. DWH and Data Mart

Once the raw data had been uploaded to a MySQL server and transformed in a staging database, the CDWH and DM had to be set up and their tables properly structured and linked according to existing data contained in

the staging area and knowledge on range of the data set. SQL scripts were created and executed to carry out this labour.

On the data mart database, views were created to query for commonly interesting information, such as: How well does a team do depending on where they play? Are they better when playing home or is it neglectable? What teams scored the most goals per season?

The result of these queries were imported using Microsoft Excel to create pivot tables and generate charts to perform analytics.

VI. Analysis

In Figure 3 and 4 we analyze 5 teams in a period of the last 2 seasons. In the column Sum of Goals we see all Goals from a team and in the column Sum of Goals against we see the Goals from the Opposing team. In the next column Sum of Matches we can see that they all had the same number of matches. To determine the best team, we can see in the column Sum of Win, the number of games won. Also we can see in the column Sum of Shots how many shots were shot by the team and by the column Sum of Shots on target how many shots were shot at the goal. Finally we see the number of red and yellow cards.

Row Labels	Sum of Goals	Sum of Goals against	Sum of Matches	Sum of Win	Sum of Shots	Sum of Shots on target	Sum of Red cards	Sum of Yellow cards
Bayern Munich	**180**	**60**	**68**	**51**	**1225**	**491**	**2**	**86**
Season 2017-18	92	28	34	27	589	236	0	42
Season 2018-19	88	32	34	24	636	255	2	44
Dortmund	**145**	**91**	**68**	**38**	**940**	**366**	**6**	**81**
Season 2017-18	64	47	34	15	489	182	3	41
Season 2018-19	81	44	34	23	451	184	3	40
M'gladbach	**102**	**94**	**68**	**29**	**911**	**314**	**0**	**90**
Season 2017-18	47	52	34	13	454	156	0	47
Season 2018-19	55	42	34	16	457	158	0	43
RB Leipzig	**120**	**82**	**68**	**34**	**978**	**358**	**7**	**109**
Season 2017-18	57	53	34	15	486	176	6	48
Season 2018-19	63	29	34	19	492	182	1	61
Schalke 04	**90**	**92**	**68**	**26**	**822**	**298**	**7**	**135**
Season 2017-18	53	37	34	18	404	165	2	62
Season 2018-19	37	55	34	8	418	133	5	73
Grand Total	**637**	**419**	**340**	**178**	**4876**	**1827**	**22**	**501**

Fig. 3. Performance from 5 Teams in the last two seasons

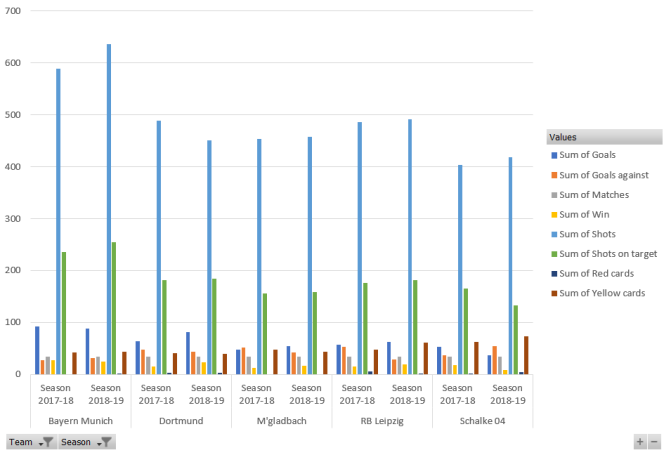

Fig. 4. Performance from 5 Teams in the last two seasons

In Figure 5 we compare the two teams Bayern Munich and Borussia Dortmund over the last 3 seasons to see which team was better in which season and which team has the better evolution. In the 2016 season you can see that Bayern Munich scored more goals than Borussia Dortmund and also won more games. In the 2017 season Bayern Munich improved and Borussia Dortmund deteriorated. In the last season 2018 the performance of Bayern Munich remains stable and Borussia Dortmund improved a lot.

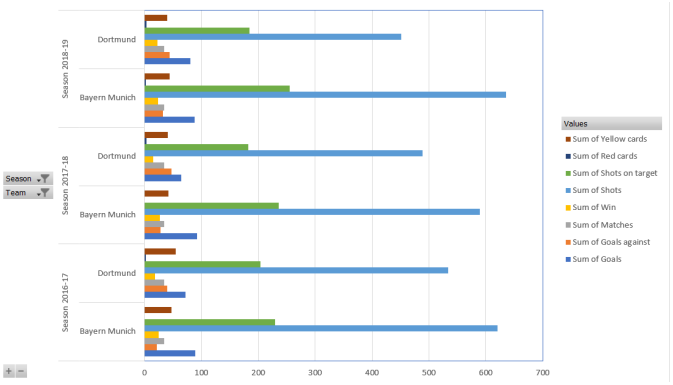

Fig. 5. Comparison between Dortmund and Bayern Munich over the last 3 seasons

VII. Conclusion

We have shown in this paper that the raw data is used to create a Data Warehouse, a data mart and also a detailed analysis, using with MySQL and other tools. We have checked the data quality and have reduced it to the relevant data needed for the project. The data can be managed effectively and the analysis can be created. Thanks to our Data Warehouse it's possible to create forecasts of a new game with information about the games from the last seasons. With this information we can get compact and understandable information about the matches, goals, red and yellow cards and shots. Additionally, it's necessary to say which team has a chance of winning a tournament or which team can win the game at home or away from home. The Project was for our team a new experience and the first insight in the development from a big dataset to make a data warehouse with the focus on the reduced data. This insight was for our team a good experience.