Data Warehousing: Soccer statistics

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

***Abstract*—Within the content of the course "Data Warehousing" an analytical project was carried out concerned with Soccer statistics about the Bundesliga in Germany. The soccer statistics data for the Bundesliga involves a huge amount of data and also a huge load. To give the soccer interested fans and viewers a useful overview as summarization, to all games of their teams and rivals, we use the collected data and present the relevant information in compact form for publics understanding. To achieve this goal, our lab project presents a data warehouse and performs some analysis, which are explained in this paper, to get the relevant data from a understandable perspective.**

I. Introduction

The fact is that football is the most popular sport on this planet. Games from previous seasons are compared again and again to better follow the development of teams. The task is to enable useful information out of the data. It's fact that in no other sport are so many bets placed as in football. Among other things, the data which we have handled also find a certain use for this purpose. The assignment involved the collection and preparation of the open data for information- and reporting purposes by applying various techniques that were introduced over the course. During the project to create a Data Warehouse, data mart and worked following the CRISP-DM process.

A. Scenario

Several data from the past Soccer Bundeliga games are needed, to give the soccer interested an overview. The fans and viewers can get easily and compact understandable information about the matches, goals, red- and yellow cards and shots. These Soccer data are relevant for the Soccer follower and also for these who do bets. For the league position is the information about the matches, points and results 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 show where we have our data from and explain this. Then, in Section **??** we describe our tool we used. Section **??** we descripe the Data Preparation for our Data Warehouse and our Snowflakes. In Section **??** we explain the input our 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 achievments 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.

III. Tools

The tools we used are:

1) XAMPP to launch apache and 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 and 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. It's follows a describtion of some

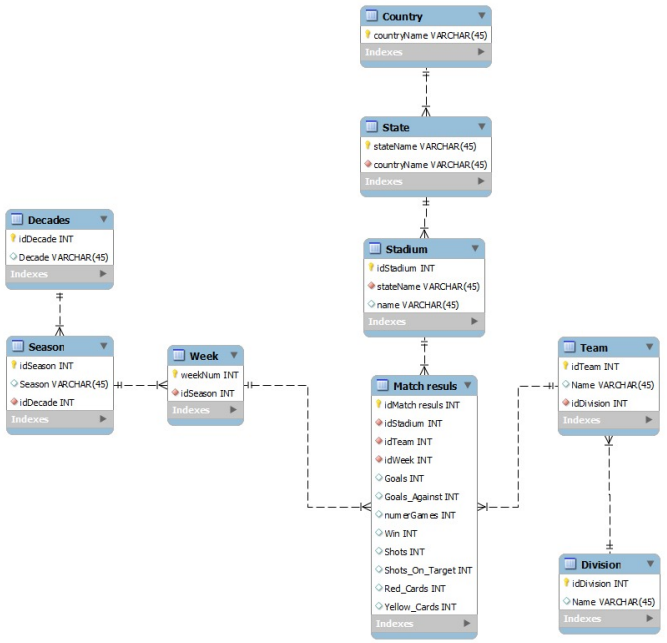

Fig. 1. Snowflake Schema

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

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, 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.

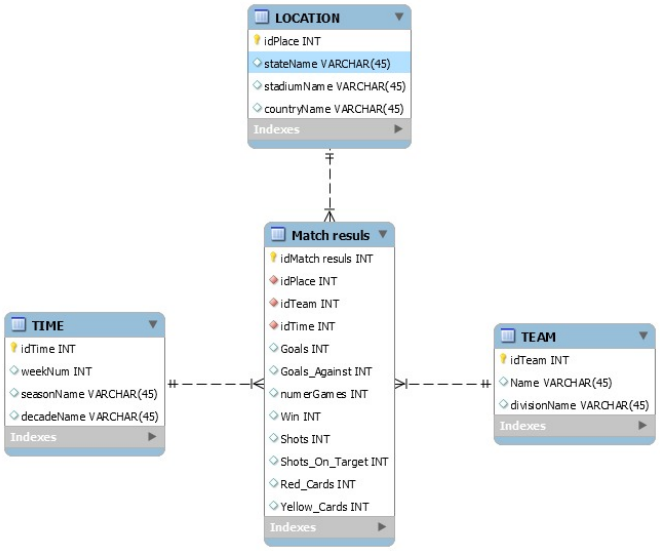

Fig. 2. Data Mart

VI. Analysis

Here i will descripe our Images!

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 Seasion

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 with MySQL. We have check the data quality and have been reduced to the relevance of data needed for the Project. The data can be managed effectively and the analysis can be created. With our Data Warehouse it's possible to create a forecast of a new

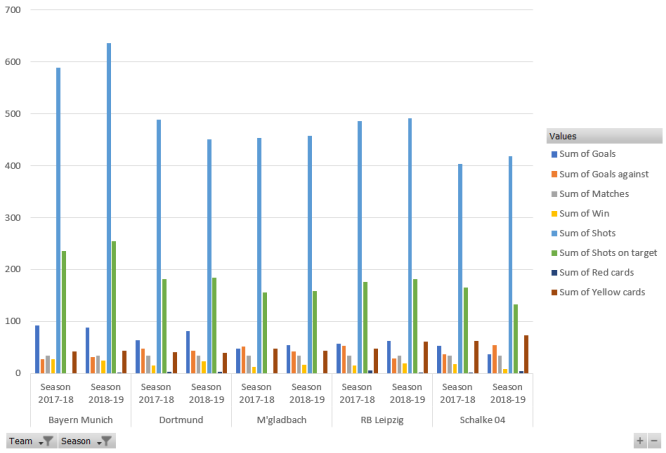

Fig. 4. Performance from 5 Teams in the last two Seasion

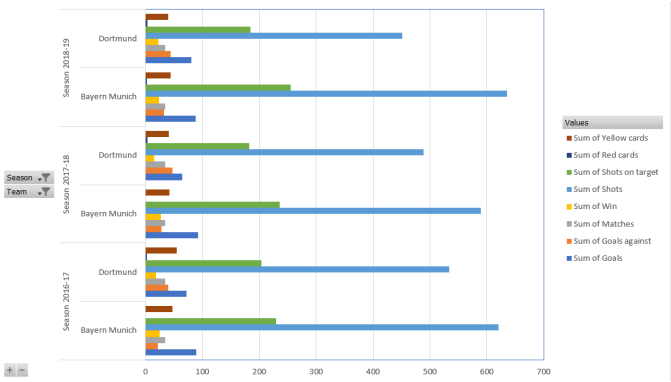

Fig. 5. Comparison between Dortmund and Bayern Munich over the last 3 season

game with information about the games from the last seasons. With this information we can get easily and compact 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.