Scientific paper for the Laboratory project of the module Data Warehousing: Soccer statistics

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

***Abstract*—Within the contect 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 an useful overwiew 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 in 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 handeled 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 the create a Data Warehouse, data mart and worked following the CRISPM-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 II we present the availability of Open Data we use in this project. Then, in Section III we describe our tools, visualize our Data Warehouse and create a analysis with MySql. Section IV we descripe the Data Preparation for our Data Warehouse. Finally, we conclude our work in Section VI with our achievments and a feedback of this project.

II. Open Data: MyDataSource

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. Concept for Problem X

In this Section we descripe the concept of our problems and our definition. We want to give the fans a good view of the last seasons games.

The tools we used are:

1) Xampp to connect to MySQL Database
2) Libre Office to perform analytic report
3) Microsoft Excel to convert to a CSV-file
4) MySQL Workbench to create the ER-Diagram and the database
5) Anaylsis with Microsoft Excel

IV. Data Preparation

First step is to get an overview of the available data and their quality. The create an analysis with Excel and problems with the quality of data should be identified. In this case, we select the data that is relevant for the selected task (columns, rows). We step follows by the import of the data into the local MySQL RDBMS. More details following: The data preparation create a final data set that forms the bases for the next phase of modeling The goal is good data quality! We have to check amongst others the completeness: there should be not NULL's

for example. Correctness clears if have any typos or missing data. The check if the data is outdated orftom the peroid the need. The download the CSV-file in Microsoft Excel. We have to create a final data set, that forms the basis for the next phase of modeling. The next step is to create a Snowflake schema as you can see in figure 1. Our Snowflake schema consists of the fact table and the three dimensions time, location and team displayed in there finest granularity! This section deals with the implementation of the concept described in Section III following the *Data Preparation* phase of CRISP-DM.

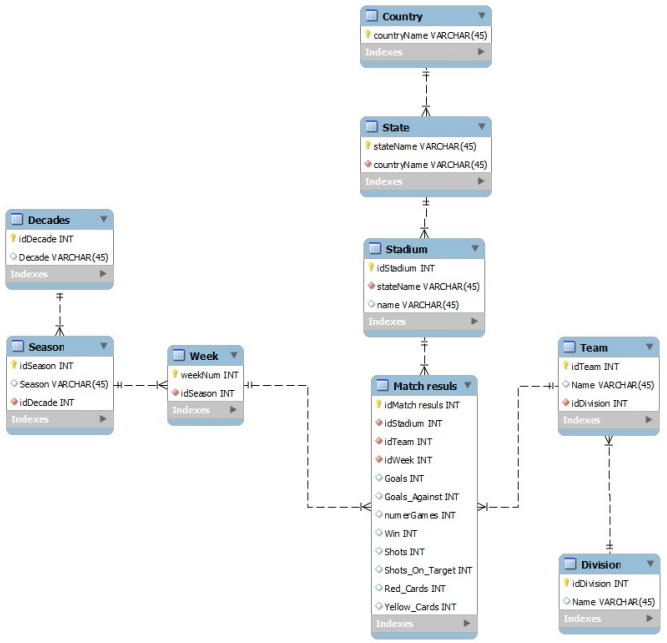

Fig. 1. Snowflake Schema

It's follows a describtion 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

V. More Sections

At the following points we describe our general solution:

- with the idMatch resuls we save all games resuls of a match
- with the idStadium you get the information, where the game is located

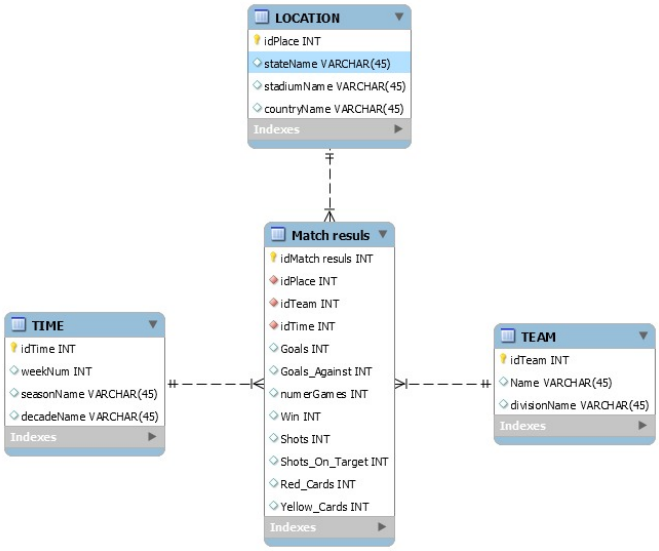

Fig. 2. Data Mart

- with the idTeam you can get the IDs of the participating teams
- with the idWeek you can see the play Week of the game

You can see in the figures X to X we performed the analysis of our Data Warehouse with the MySQL.

VI. 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 effectivly and the analysis can be created. With our Data Warehouse it's possible to create a forecast of a new game with information about the games from the last seasions. 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.