Group 1
Project Summary Report
Designing and Implementing a Data Warehouse for Soccer Clubs

Team Members

Monal Agnihotri| Aditya Kannan

Table of Contents

Executive Summary

Traditionally soccer and analytics were poles apart. Soccer clubs were unaware of the power of analytics, which could influence the soccer operations of any league, club or national soccer. There are currently 21 professional soccer leagues in the world. With the growing popularity of soccer in remote corners of the world, soccer clubs are facing a challenge to ensure that there is a balance between fan loyalty and revenues.

Most successful soccer clubs such as Arsenal, Real Madrid, Barcelona and Manchester United (to list a few) have efficiently managed to achieve an equilibrium between profits and fan loyalty by leveraging Big Data Analytics. Strategic decisions such as acquiring new players, match tactics, ticket and merchandise pricing and attract fans are made through analysis on lag and lead data.

While soccer clubs have moved to lead data, we have chosen to focus on lag data of a professional soccer league. Our business objective in the scope of this project are as follows

- To support analysis of fan loyalty within soccer operations.
- This would help soccer clubs price tickets depending on the most eventful match and increase their revenues.

To build our data warehouse, we started with logical dimensional model which was aligned with our business objective. We designed Physical dimensional model and performed ETL processing to load the dimensional structure. For reporting we selected Microsoft Power BI as, the reporting tool for creating reports and visualizations.

Pro Soccer Organisation Summary

Who we are

Our organisation ‘Soccer Maniacs’ is a non profit organisation which helps soccer clubs and independent teams to understand and make profitable decisions. We as an organisation we are pioneers in designing customized data warehouse on lag data. We design smart decision making roadmap for our clients.

Our Mission

To help soccer clubs understand the importance of balance between increasing revenue and maintaining fan loyalty.

What we do

We use a five step agile approach to deliver a perfectly built solution. Given below is a sample of what we do:

- We start by performing a short primitive study on the size of the club, league or team.
- After measuring this we do a thorough evaluation of ‘as is analytics system’ if the client is currently using one. By doing this we ensure that we do not duplicate features that are already existing in the system.
- Meet with stakeholders to understand their current problems or objectives.
- Using the above information we work on the ‘to be analytics system and deliver our solution.
- We also provide post implementation support for our solution with exhaustive documentation and training.

As we use agile SCRUM approach for building and delivering all our solutions, we provide the flexibility to our clients for scope change at any point between our sprint cycles.

Above and beyond a customized data warehouse, we also offer an option to ingest external data sources to the data warehouse. This external data is very valuable for a soccer club as it contains current average match statistics of various soccer leagues in the world.

Prioritized Requirements

Introduction:

Soccer analytics being a vast domain, we wanted to essentially narrow down and implement features that would support our business objective. Based on our objective ‘To analyze fan loyalty’ below mentioned are the parameters from lag data we have planned to consider in our ‘to be system’.

Summary:

- Match statistics is the most important aspect of any soccer game
- A soccer match comprises of many events but majorly of
    - Pass
    - Goals
    - Goal Keeper Saves
    - Fouls
    - Red Cards
    - Yellow Cards
    - Shots on target
    - Penalty
    - Tackle
- Using this summarized data, clubs can analyze on the most eventful match of the tournament.

Data Collection from Datasets:

To achieve our objective, we had access to two distinct datasets. Below given is brief description of the data we used

1. Event File Data: This file primarily had data about all the events occurred in a match.
2. Location File Data: In this file, for every event it had information about the player and his position during the event.

For our purpose, we plan to use both datasets, but extensively the event data file as our objective is to analyze major events in a soccer game. To collect distinct game date and matches between teams, we plan to combine unique data from both the files as this would give us an accurate picture.

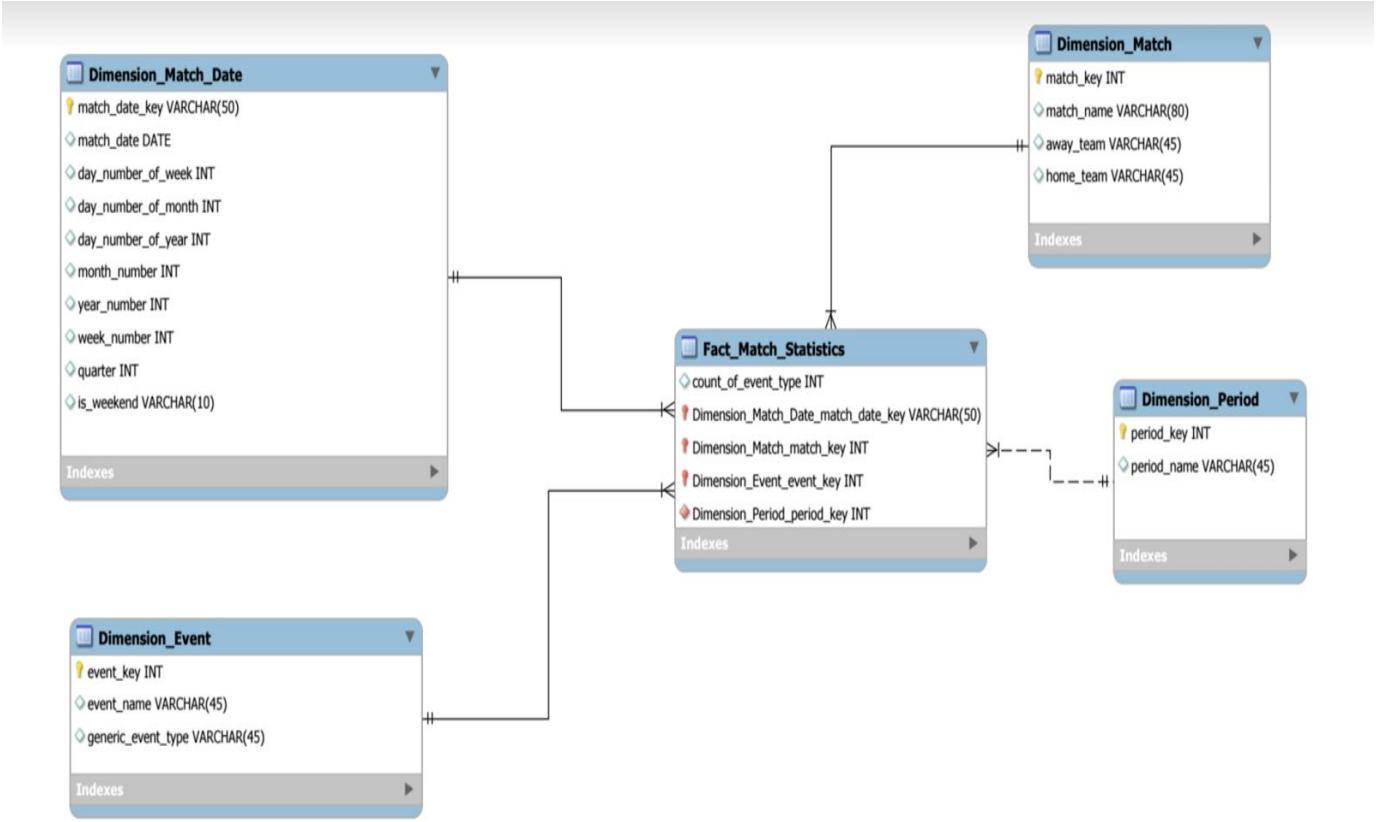
Conclusion:

Sentiments of fans in soccer world revolves largely around this factors. Even if a match results in a draw between, clubs and league federations would like to have an overall picture of match in terms of primary events of soccer.

Logical Dimensional Model

Overall Summary:

Our logical model is shown below. The features we chose to model are based on the single objective which was to summarize match statistics. We ended up with four dimension tables and one factless fact table.



Explanation of Dimensions:

A brief description of each dimension and attributes in it are given below:

Dimension_Match

a) Match_ Key: It is primary key for this dimension table .
b) Match_Name: It consists of name of two teams playing the match. For simplicity and to depict reality in reports we have separated the two names using ‘Vs’.
c) Away_Team: Visiting team details
d) Home_Team: The team hosting the match.

Dimension_Match _Date

a) Match_Date: Date on which match was played between two teams
b) Match_Date_Key: Primary Key for the table.

Other attributes are related to the analysis of different aspects of match date such as Year Number, Month Number and Day Number.

Dimension_Event

a) Event_key: It is the primary key .
b) Event_Name: It is name of the the event
c) Generic_Event: Will have either of three values ‘Aggressive,Defensive, Foul’.It is a part of future scope of the project.

Dimension_Period

a) Period_Key s assigned as primary key
b) Period_Name: It is name of the period.

Fact_Match_Statistics

This is a factless fact table mainly for analysis of match statistics to find out overall output of game.The table has three key dimensions and one analysis dimension.It consists of the

following fields:

a) Count _of_event_type: It will measure how many types of event happened in match/ per game . for example shot of target, goal scored, number of fouls, Number of passes
b) Match_Date_Key: Key dimension and is a foreign key.
c) Match_Key:Key dimension and is a foreign key
d) Event_Key:Key dimension and is a foreign key
e) Period_Key: Analysis dimension and is a foreign key

The primary key for this table consists of combination of Match_Date_Key,Match_Date and Event_Key.

Physical Design, ETL Processes, and deployment to target environment

Overall Summary

To facilitate ETL processing for our project we used pentaho data integration tool extensively to load our dimensions and facts. Below given is the segmented explanation for loading each dimension and fact table with a pictorial representation of actual steps involved to achieve the result.

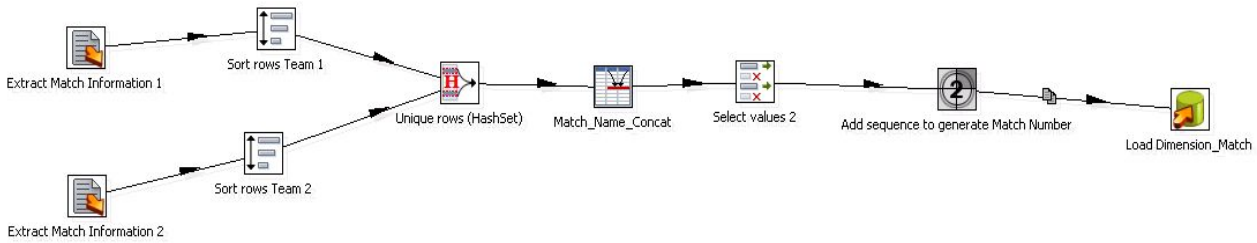
Explanation of ETL Processing

Dimension _Match :

It consists of data from event file as well as location file . Sorting is done to get exact values of match name , Home team and Away team. Through unique rows distinct values are evaluated and provided to next step.

Match name concat step is used to make sure that match name should fit in to team1 vs team2 format. Correct names of the teams are selected by select value2. Unique number key generation and arrangement of sequence is done by add sequence .

After sorting and arranging the data as per our needs we have loaded in the table.



Dimension_Event:

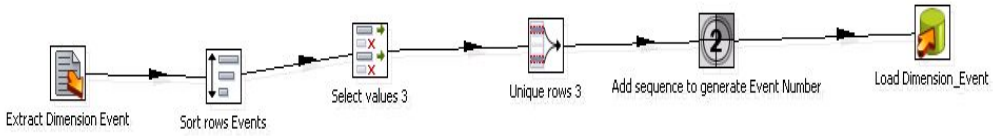
Similar to match _dimension here in case of event dimension we have extracted data from event file. Using sort rows, we have sorted the data per our need so that we should get information related to events.

Select values is used to select accurate events associated with match.

For getting distinct values of event unique rows has been used. Which will provide unique values of events avoiding repetition.

For providing unique sequence number and arranging them in sequential order we are generating event number.

Finally after making sure we have extracted the data accurately as per our needs we loaded into the dimension_event_ table.



Dimension_Match_Date:

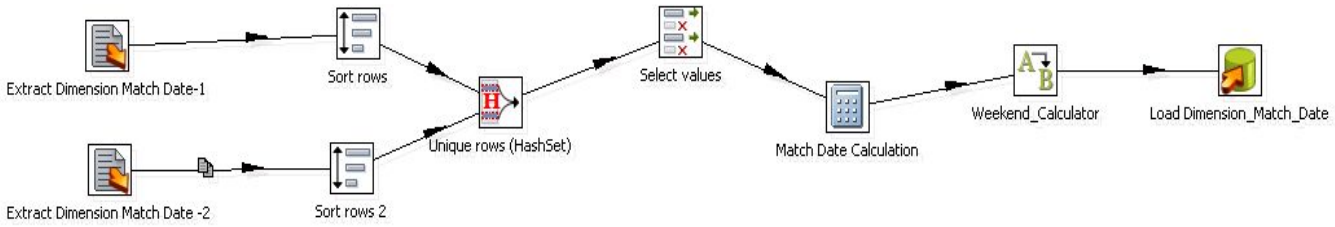
We have extracted data for Match Date from both the datasets, to ensure consistency.

Sort rows is applied for both the files to sort the data associated with each match as per date.

Unique rows will generate the distinct data avoiding repetition of dates.

Calculator is applied to calculate remaining attribute of time dimension apart from date like day_number_of_week, day_number_of_month, day_number_of_year etc.

After making sure we have cleaned the data according to our needs we loaded it into dimension table.

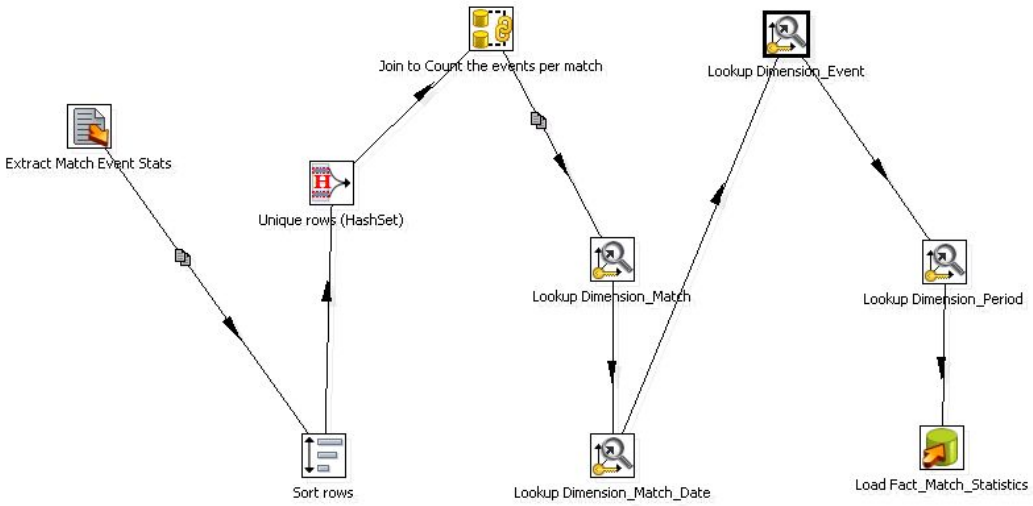


Dimension _Period:

For Dimesion_period, we have taken slightly different approach of extracting data from table instead of files. Value mapper will divide the period into respective half for example if value is 1 then its First Half.

Fact_Match_Statistics:

We have used Database join to populate values in factless fact table. In next step, we are joining the dimensional table with fact table through primary keys. In the final step we are loading all the aggregated data in fact table.

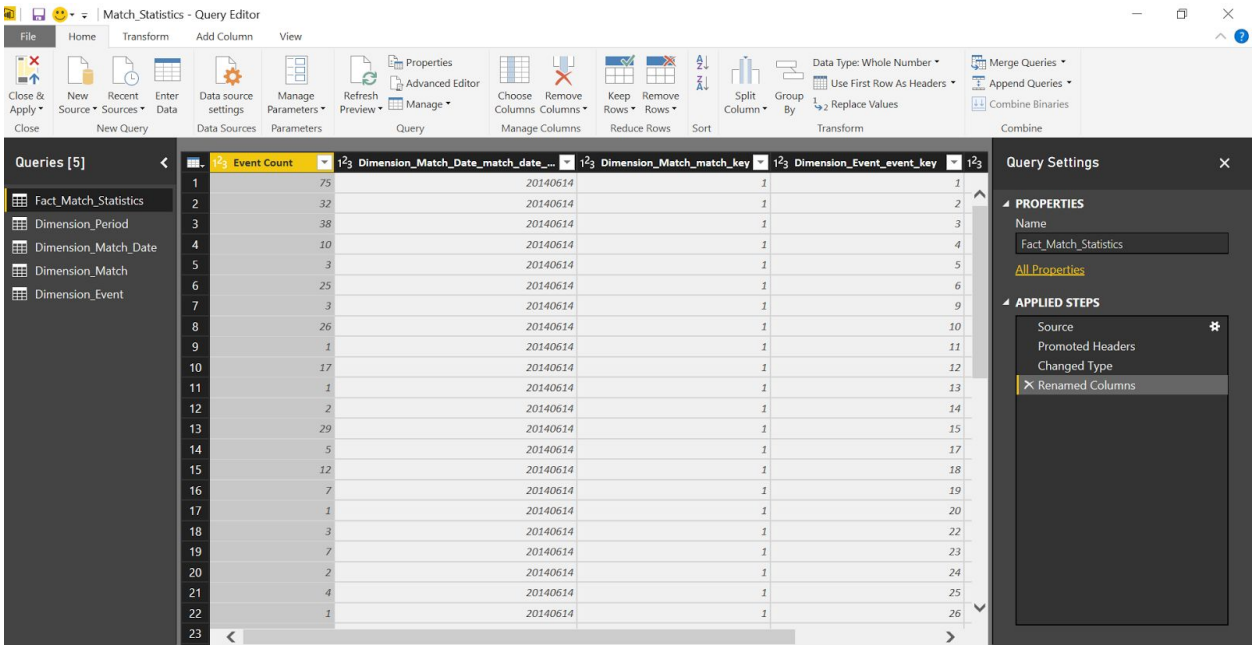

Join to Count the events per match
Lookup Dimension_Event
Extract Match Event Stats
Unique rows (HashSet)
Lookup Dimension_Match
Lookup Dimension_Period
Sort rows
Lookup Dimension_Match_Date
Load Fact_Match_Statistics

Current and Future Reports:

For the sake of simplicity, after loading our data to the dimensions and facts we extracted it in form of a csv file from mysql. To provide support for our analysis we used Power BI as our reporting tool to generate reports and visualizations from our dimensional model.

Below given is the sample snapshot of the power bi desktop application, where we ingested the following csv files:

1) Dimension_Match: Contains information of all matches related to the project
2) Dimension_Match_Date_Key: It consists of unique match dates from the both the datasets
3) Dimension_Period: It consists of period information of all the matches.
4) Dimension_Event: It consists of all the events occurred in all the matches.
5) Fact_Match_Statistics: It consists of total of all events



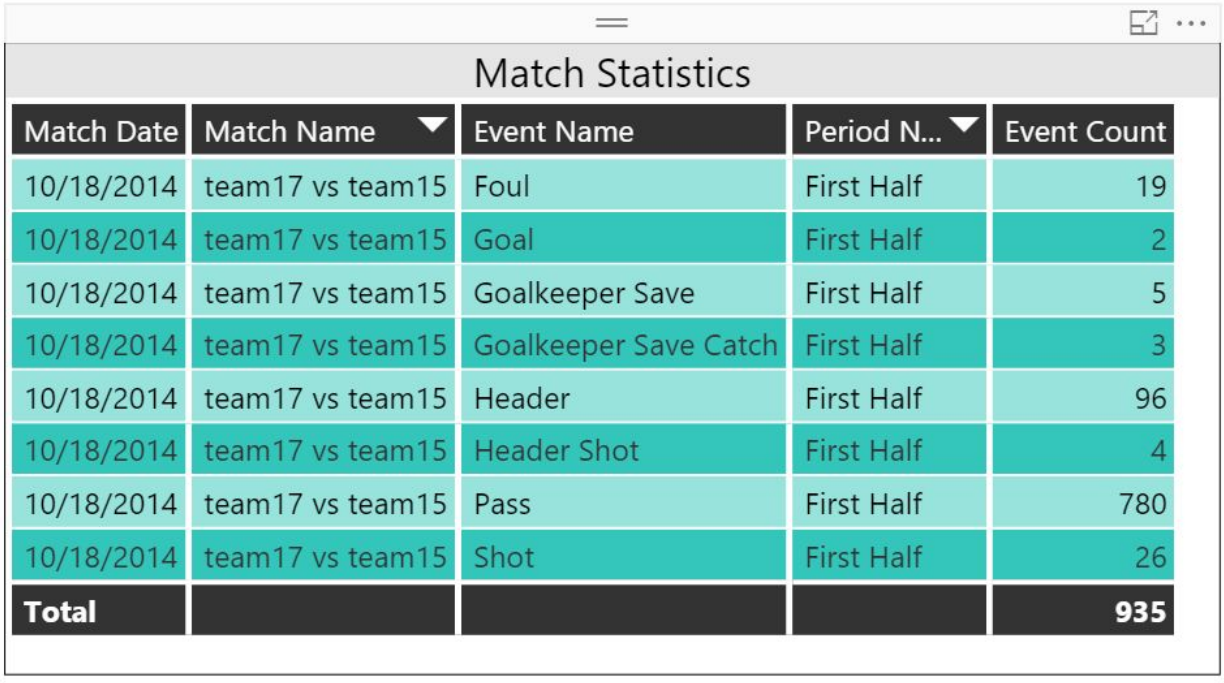
Using data modeling functionality of power bi, the dimensions and facts were linked to each other.Post that, we generated three reports. For each view of the report Power BI provides assistance to export data in form of a CSV file to check the data.

Below given is the detailed explanation of each report.

Match Statistics Report:

This report consists of information regarding a particular match between team 17 and team 15.For simplicity the current view is filtered on popular event type, particular match and particular date.

As shown below the report perfectly summarizes all major events in a match such as Goals, Fouls, Shots on target.



Match Statistics

Match Date	Match Name	Event Name	Period N...	Event Count
10/18/2014	team17 vs team15	Foul	First Half	19
10/18/2014	team17 vs team15	Goal	First Half	2
10/18/2014	team17 vs team15	Goalkeeper Save	First Half	5
10/18/2014	team17 vs team15	Goalkeeper Save Catch	First Half	3
10/18/2014	team17 vs team15	Header	First Half	96
10/18/2014	team17 vs team15	Header Shot	First Half	4
10/18/2014	team17 vs team15	Pass	First Half	780
10/18/2014	team17 vs team15	Shot	First Half	26
Total				**935**

Overview Report of multiple matches on same date:

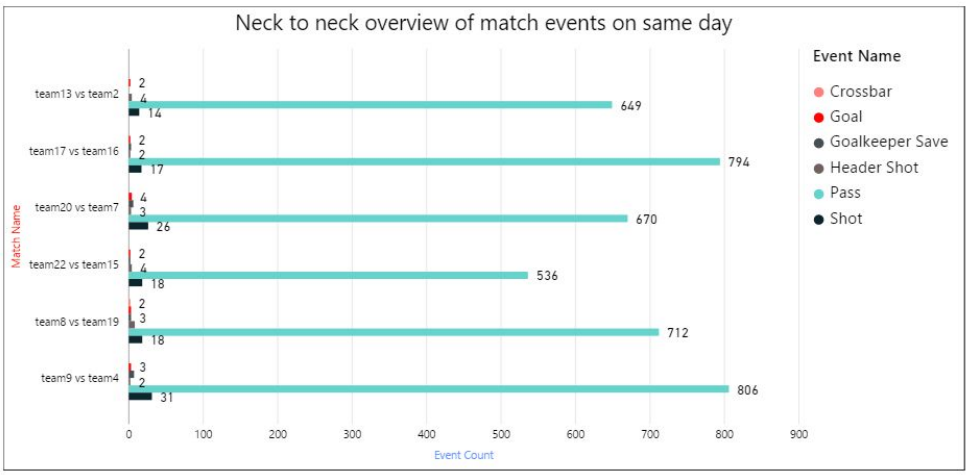
This report comprises of comparison of different matches held on the same date. It is drill down report where user can select a particular match and analyze details of the events occurred in

both the halves (period 1 & 2).

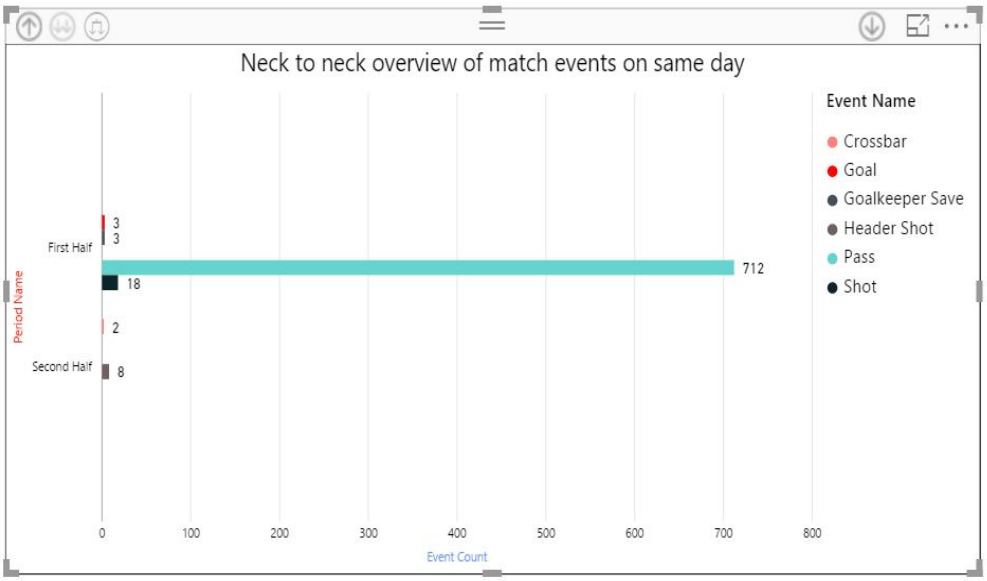
Using this report a business analyst can identify the most eventful match on a given day.

The current view of the report is filtered on a particular Match Date and few popular events.

Drill down level one



Drill down level two on Period Name



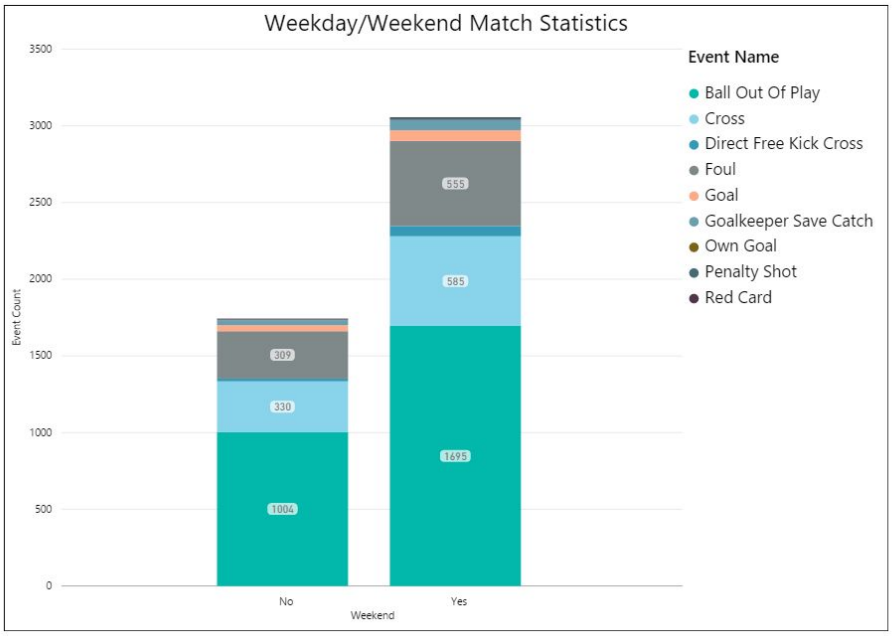
Comparison of Weekend/Weekday Match Statistics Report:

This is another drill down report. Here we have compared the match statistics on day type basis.

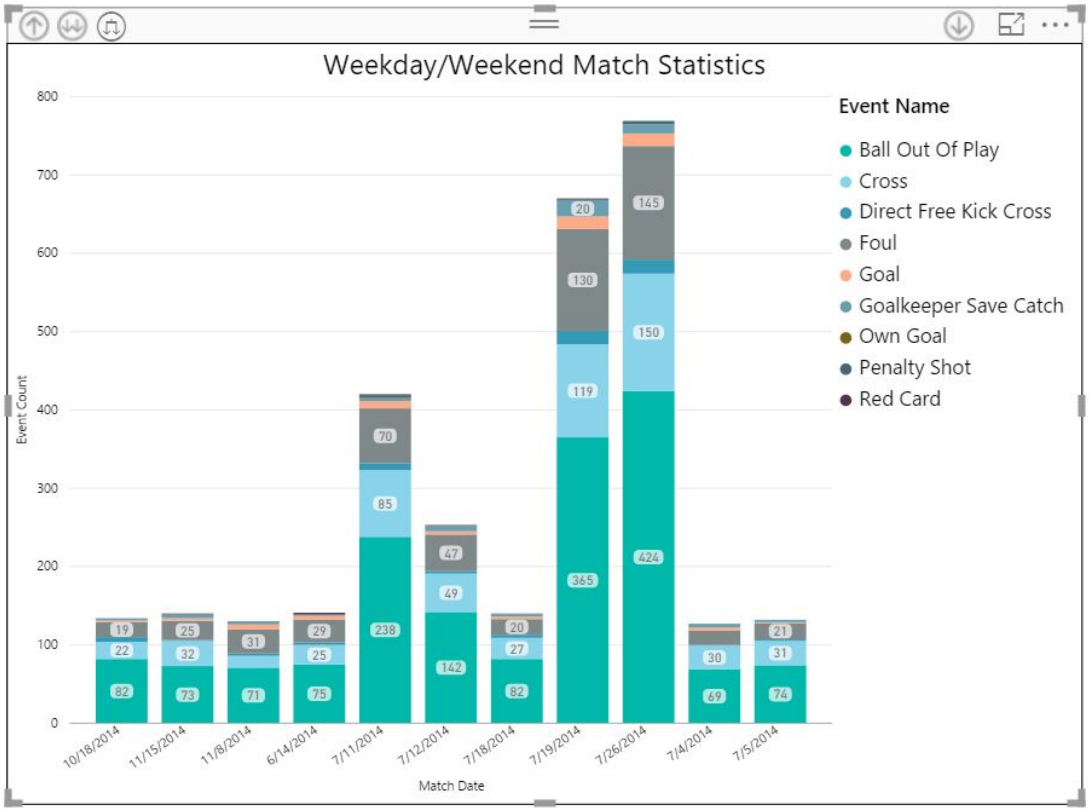
Using this report, a soccer club can efficiently manage ticket prices by analyzing trends of summarized events as per day type.

The view of report is filtered on the parameters Event Name.

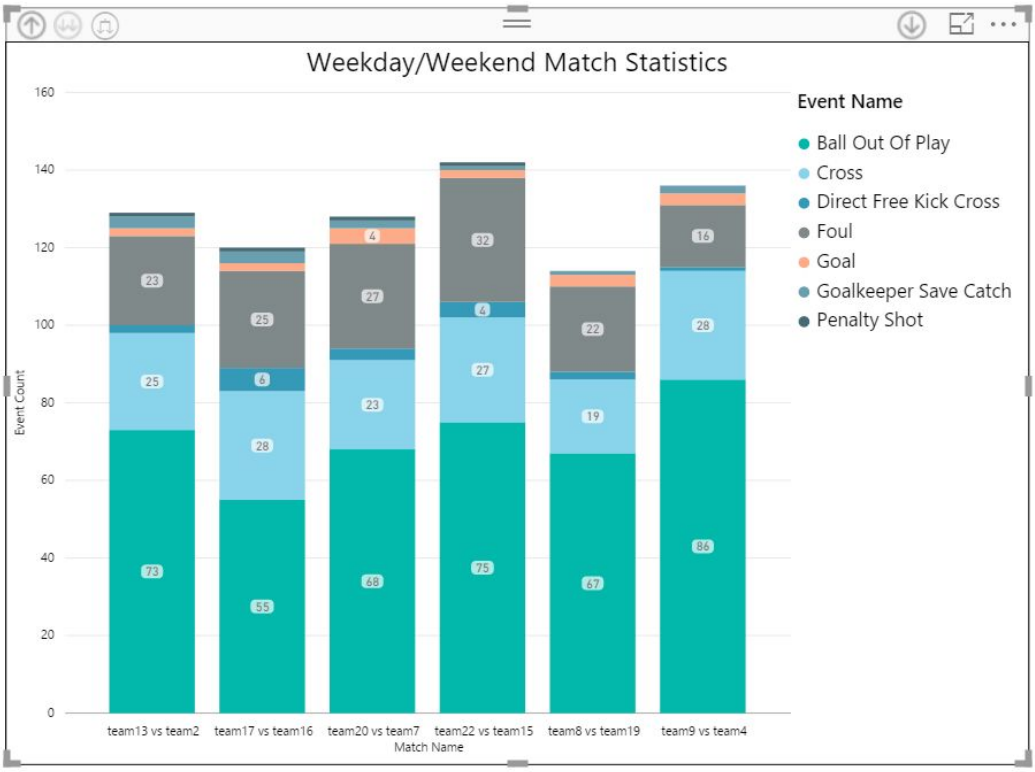
Drill down Level 1



Drill down Level 2 on Match Date


Weekday/Weekend Match Statistics
Event Count
800
700
600
500
400
300
200
100
0
Match Date
10/18/2014
11/15/2014
11/8/2014
6/14/2014
7/11/2014
7/12/2014
7/18/2014
7/19/2014
7/26/2014
7/4/2014
7/5/2014
19
22
82
25
32
73
31
71
29
25
75
70
85
238
47
49
142
20
27
82
20
130
119
365
145
150
424
21
30
69
31
74
Event Name
Ball Out Of Play
Cross
Direct Free Kick Cross
Foul
Goal
Goalkeeper Save Catch
Own Goal
Penalty Shot
Red Card

Drill down level 3 on different matches for Same Date:



Future Reports:

As a future scope, we plan to analyze on match statistics on individual team basis. This would help in understanding how actively a team was involved in a game and if the match was a close aggressive or a defensive match.

We also would like to modify our dimensional model, to accommodate another aggregated fact and dimension table which would summarize match event based on a dimensional attribute.

References:

1) https://en.wikipedia.org/wiki/Football
2) http://www.sloansportsconference.com/content/beyond-the-4-4-2-soccer-analytics/
3) https://www.tnooz.com/article/big-data-germany-world-cup/
4) http://sports.stackexchange.com/questions/4996/how-is-distance-covered-tracked-in-football-soccer-at-the-world-cup-and-duri

Appendix:

Course Material used:

1) Planning
2) Requirements
3) Logical Design Part 1 & 2
4) Handouts 1 & 2
5) Labs 0, 1 & 2

Project Timesheet:

Below given is a detailed timesheet of our project.

Date	Team Member	Hours Spent	Description of work	Additional Comments
2017/03/08	Aditya	2.0	Understanding soccer analytics	NA
2017/03/09	Aditya	1.0	Understanding the dataset	Loading the relational model in Mysql tables
2017/03/09	Monal	3.0	Basic understanding soccer game	
2017/03/10	Monal	2.5	- setting up development environment - ETL technology research	NA
2017/03/12	Aditya	1.0	- ETL proof of concept	NA
2017/03/13	Monal	3.0	- Summary	NA

			Report	
2017/03/14	Aditya	5.0	Analyzing Dimensional model from business objective	creating rough sketch of a dimensional model
2017/03/14	Monal	4.0	Using Mysql workbench to create dimensional model	
2017/03/16	Aditya	3.0	Querying existing relational database tables on events and location	
2017/03/17	Monal	6.0	Preparing project Summary and Presentation	
2017/03/17	Aditya	7.0	Loading data in dimensional model	ETL processing
2017/03/18	Monal	3.0	Finalizing Project presentation	
2017/03/18	Aditya	3.0	Generating reports in Power BI	
2017/03/19	Aditya	4.0	Creating Project deliverables	
2017/03/20	Aditya	2.0	Review of Project summary report.	
2017/03/21	Aditya and	5.0	Preparing Video	Project

	Monal		and uploading	presentation video.