Data Warehousing Project Report

IS 6480 - Group 7

Abhijay Sharma Gaurav Kutemate Angel Yang

Table of Contents

Introduction

Premier League is arguably the best professional soccer league that represents the sport's highest level in England. It is contested by 20 clubs and operates on a system of promotion and relegation within English Football Leagues. The premier league trophy is one of the most heavily contested in professional sports today. With the teams having a loyal fan base and a high amount of ticket sales, there is always pressure to perform best in each match. While having a loyal fan base means that there will always be people watching the match in the stadiums or live streaming, but due to the global nature of this sport, there is always room to acquire more viewers and fans. A lot of this depends on how well a team plays in matches, specifically on how good they are in their offensive tactics. As goal wins matches and hearts, having a high offensive production is crucial for every team.

Vision

To help improve the performance of a team playing in the Premier League, UK.

Strategic Objective

Of the many strategic objectives of the organization, we have addressed:

1. Increase in Revenue
2. High Fan satisfaction

In short, Having high offensive production has a high impact on fan satisfaction and a small impact on the revenue.

How do a DW and/or analytics fit into the organization vision and objective?

- With the advancement and infusion of technology into the very fabric of sports, the need for centralized data storage is higher and crucial than ever.
- Premier League teams have a tremendous need for a data warehouse because to capture and store the information from all the different aspects of the organization and to be able to analyze that data in real-time is how a DW differs from a simple Database.

- Once a quality data is stored and ready to be used, Analytics will give an organization a strong support system to run various operations successfully such as team preparation, player and opposition analysis, player trading, scouting process, player health analysis.

Prioritized Requirements

Completed:

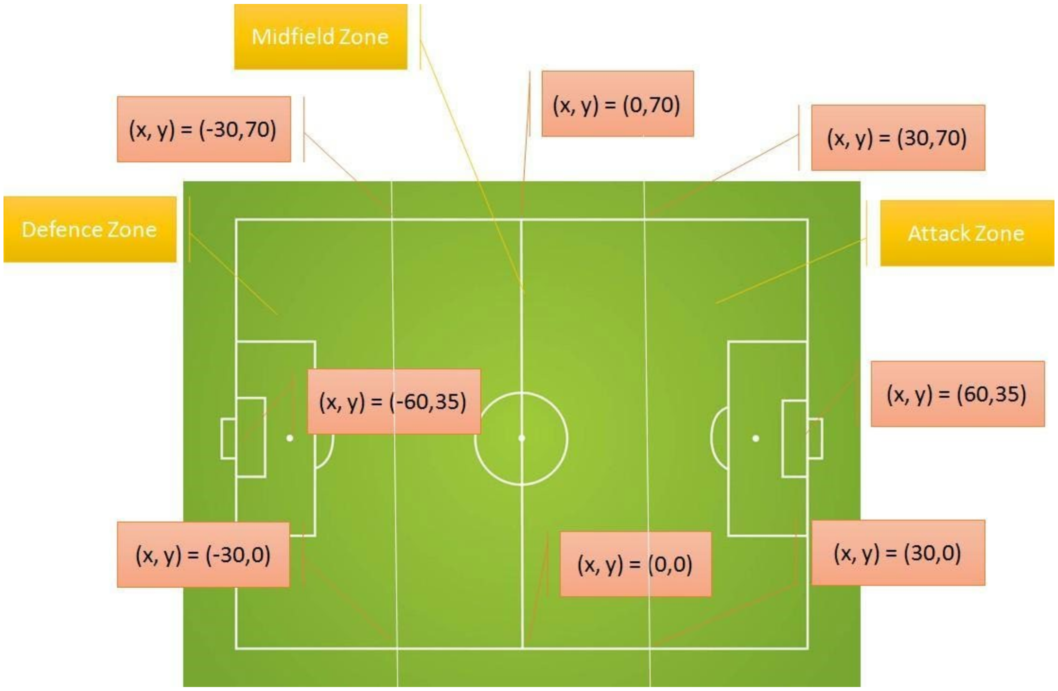
- Built a Dimensional model which helped conceptualize the project to achieve strategic objectives.
- Performed player analysis based on which team selection and preliminary strategy could be done.
- Derived which teams are offensive based on player performance and type of play (events).

Needs Refinement:

- With more data available on players physicality:
  - Injury management could be achieved.
  - Team utilization and strategy could be devised.
  - Overall performance output could be better.

Planned for future:

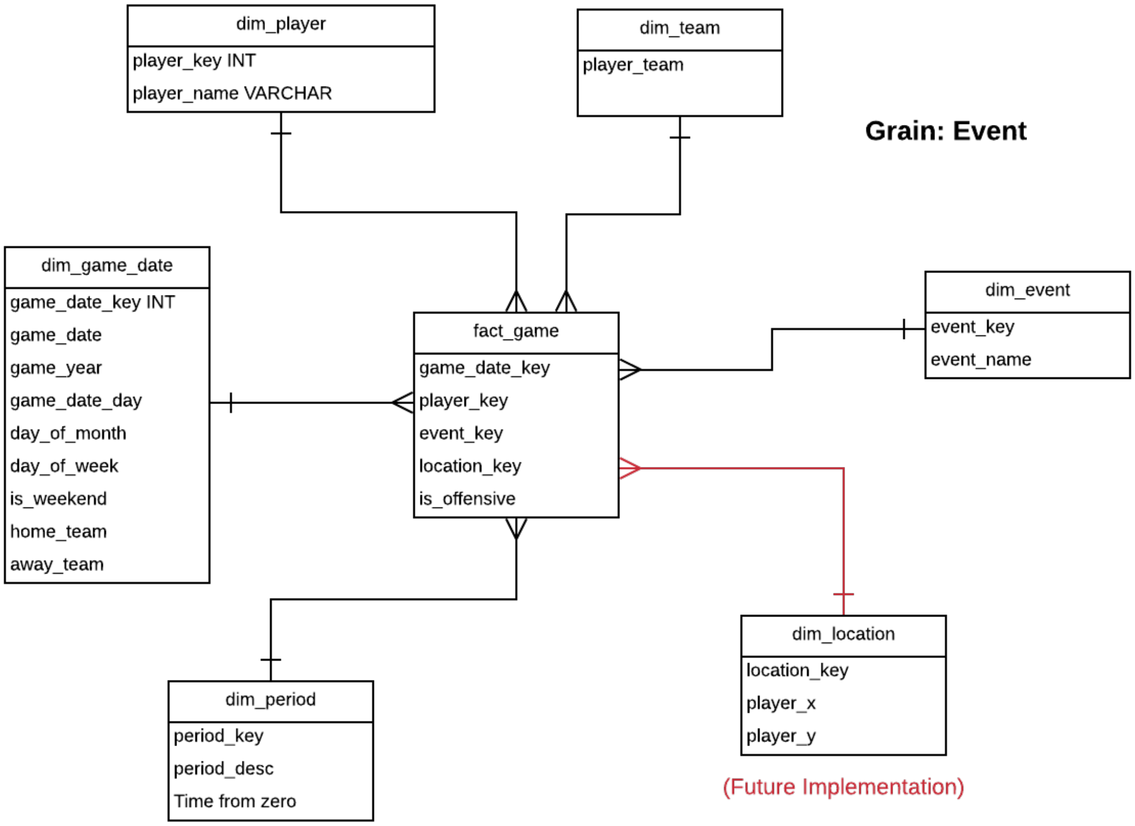
- To customize the existing dimensional model to be used to combine event and location data of the players to categorize players based on the zones (Forward, Midfielder, and Defender). Below is a snapshot of the 'zone' concept that could be implemented and integrated with events data.



- With zones data available, a better play strategic approach could be used to improve play and against oppositions.

Logical Dimensional Model:

The logical model of our Business process:



The data on which this logical model is built on has been provided in the form of a tab separated file - 'events data'.

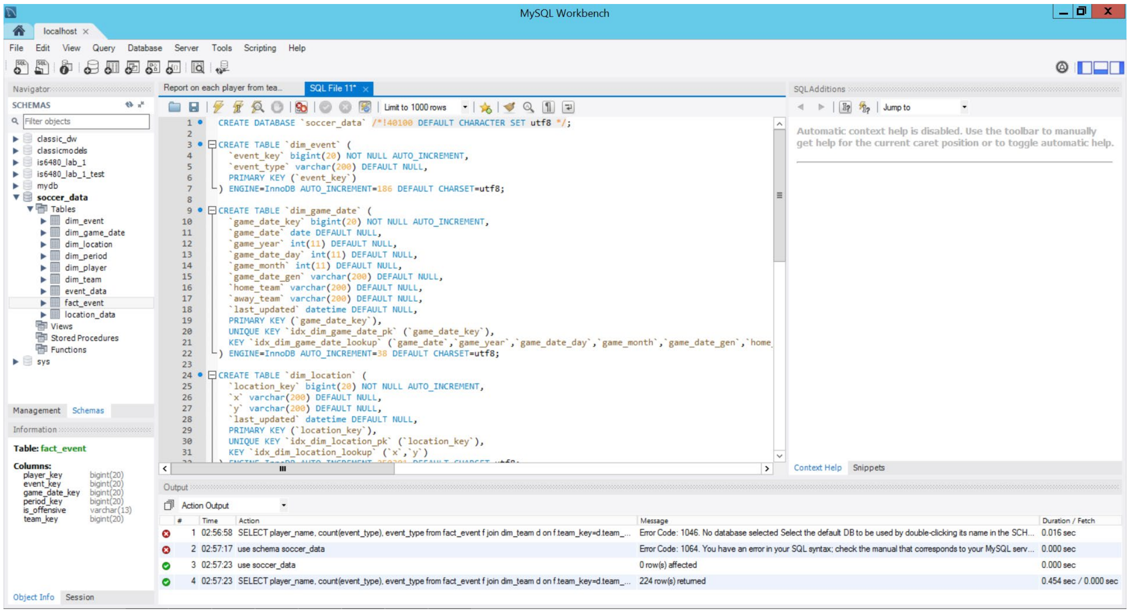
1. **Fact Table**: The fact table has been created based on the grain – an event. Currently, the fact table consists of one metric attribute(is_offensive) apart from all the foreign keys. We have categorized a subset of events those would be a clear indication of an offensive play. Based on the percentage of offensive events, we could infer that our team had an offensive play indicated by 'is_offensive'.
2. We have **5 dimensional tables** in the logical design –
   1. Team – This dimension is used to store all the distinct teams participating in the league.
   2. Player – This dimension stores all the players participating in the league.

3. Period – This dimension is used to store the period value which indicates whether it is first half or the second half of the match.
4. Game_Date – This dimension includes all the specific details about the day that a particular match is played; whether that day was a weekend, the month in which it was played, the day on which it was played, etc.
5. Event: This dimension is used to record every event committed by a player during the game.

Physical Design

The implementation of the logical design is described below:

1. Source Data: The source systems extract data from match videos and store them in file format. The source data is two ‘tab' separated files created from two separate source systems. From the files, we have currently used only one file-events data, to create our Physical model in a relational database. As mentioned previously, integration of location data would be the next phase.



```sql
CREATE DATABASE `soccer_data` /*!40100 DEFAULT CHARACTER SET utf8 */;

CREATE TABLE `dim_event` (
  `event_key` bigint(20) NOT NULL AUTO_INCREMENT,
  `event_type` varchar(200) DEFAULT NULL,
  PRIMARY KEY (`event_key`)
) ENGINE=InnoDB AUTO_INCREMENT=186 DEFAULT CHARSET=utf8;

CREATE TABLE `dim_game_date` (
  `game_date_key` bigint(20) NOT NULL AUTO_INCREMENT,
  `game_date` date DEFAULT NULL,
  `game_year` int(11) DEFAULT NULL,
  `game_date_day` int(11) DEFAULT NULL,
  `game_month` int(11) DEFAULT NULL,
  `game_date_gen` varchar(200) DEFAULT NULL,
  `home_team` varchar(200) DEFAULT NULL,
  `away_team` varchar(200) DEFAULT NULL,
  `last_updated` datetime DEFAULT NULL,
  PRIMARY KEY (`game_date_key`),
  UNIQUE KEY `idx_dim_game_date_pk` (`game_date_key`),
  KEY `idx_dim_game_date_lookup` (`game_date`,`game_year`,`game_date_day`,`game_month`,`game_date_gen`,`home
) ENGINE=InnoDB AUTO_INCREMENT=38 DEFAULT CHARSET=utf8;

CREATE TABLE `dim_location` (
  `location_key` bigint(20) NOT NULL AUTO_INCREMENT,
  `x` varchar(200) DEFAULT NULL,
  `y` varchar(200) DEFAULT NULL,
  `last_updated` datetime DEFAULT NULL,
  PRIMARY KEY (`location_key`),
  UNIQUE KEY `idx_dim_location_pk` (`location_key`),
  KEY `idx_dim_location_lookup` (`x`,`y`)
```

Automatic context help is disabled. Use the toolbar to manually get help for the current caret position or to toggle automatic help.

Table: fact_event

Columns:

```
player_key      bigint(20)
event_key       bigint(20)
game_date_key   bigint(20)
period_key      bigint(20)
is_offensive    varchar(13)
team_key        bigint(20)
```

```
#  Time      Action                                                                                      Message                                                                                         Duration / Fetch
1  02:56:58  SELECT player_name, count(event_type), event_type from fact_event f join dim_team d on f.team_key=d.team_...  Error Code: 1046. No database selected Select the default DB to be used by double-clicking its name in the SCH...  0.016 sec
2  02:57:17  use schema soccer_data                                                                      Error Code: 1064. You have an error in your SQL syntax; check the manual that corresponds to your MySQL serv...  0.000 sec
3  02:57:23  use soccer_data                                                                             0 row(s) affected                                                                               0.000 sec
4  02:57:23  SELECT player_name, count(event_type), event_type from fact_event f join dim_team d on f.team_key=d.team_...  224 row(s) returned                                                                             0.454 sec / 0.000 sec
```

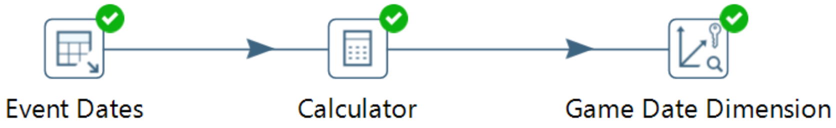
2. Using the MySQL Workbench, we loaded our source data files and analyzed the variables available to achieve our objective.

   Proceeding to implement the dimensional model, we applied some transformations in order to achieve our business process objective and to create a dimensional model.

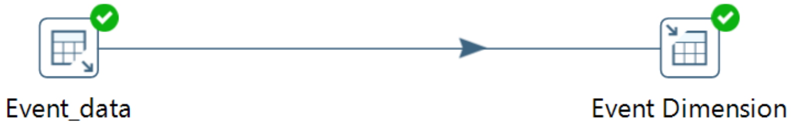
3. ETL: Pentaho – Spoon ETL tool was used to extract data from the tables.

   Six separate transformations were required in order to convert the input data to the required dimension and fact tables.

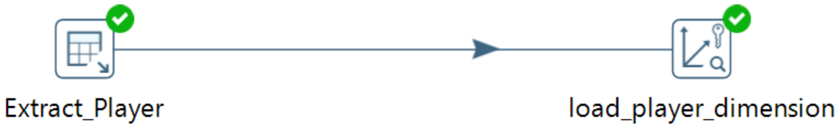
   Transformation 1: Load and calculate Game_Date Dimension



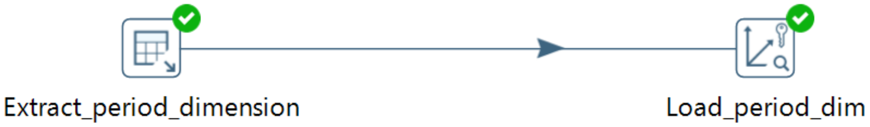
   Transformation 2: Load Event Dimension



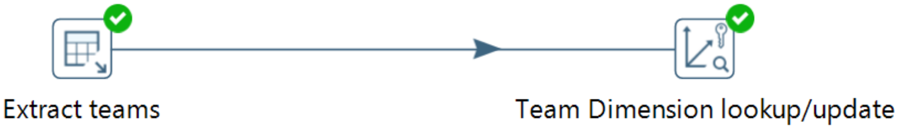
   Transformation 3: Load Player Dimension



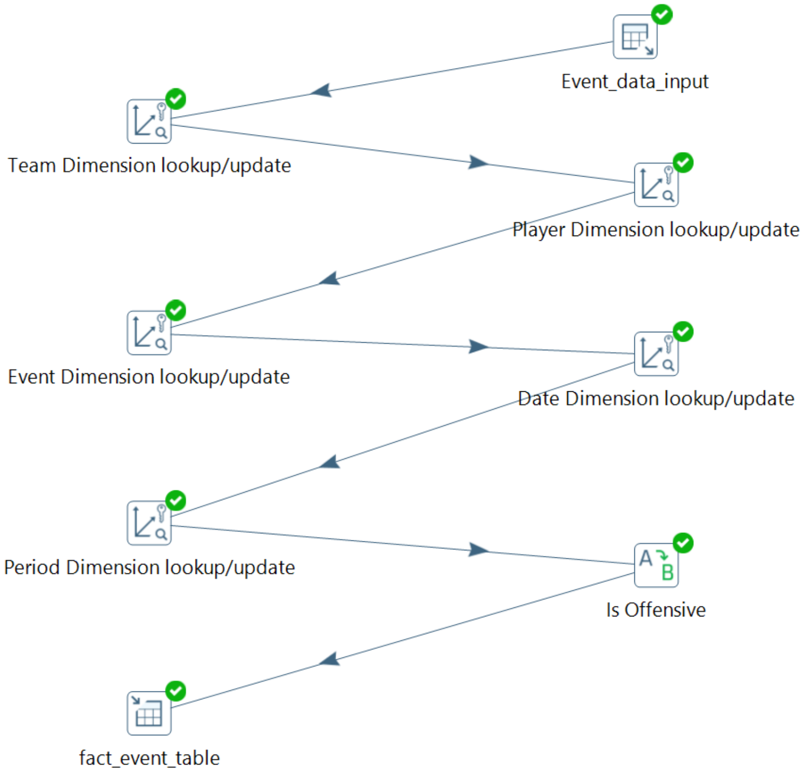
Transformation 4: Load Period Dimension.



Transformation 5: Load Team Dimension.



Transformation 6: Load fact_event fact table.


Event_data_input
Team Dimension lookup/update
Player Dimension lookup/update
Event Dimension lookup/update
Date Dimension lookup/update
Period Dimension lookup/update
Is Offensive
fact_event_table

This transformation diagram shows how the fact table was loaded using the dimension table lookup, Queries, and formulas to populate fact table fields.

Reports and Analysis

Provide descriptions of one or more sample reports/analyses that are:

- Currently functional
- Could be supported by the current DW design but have yet to be developed

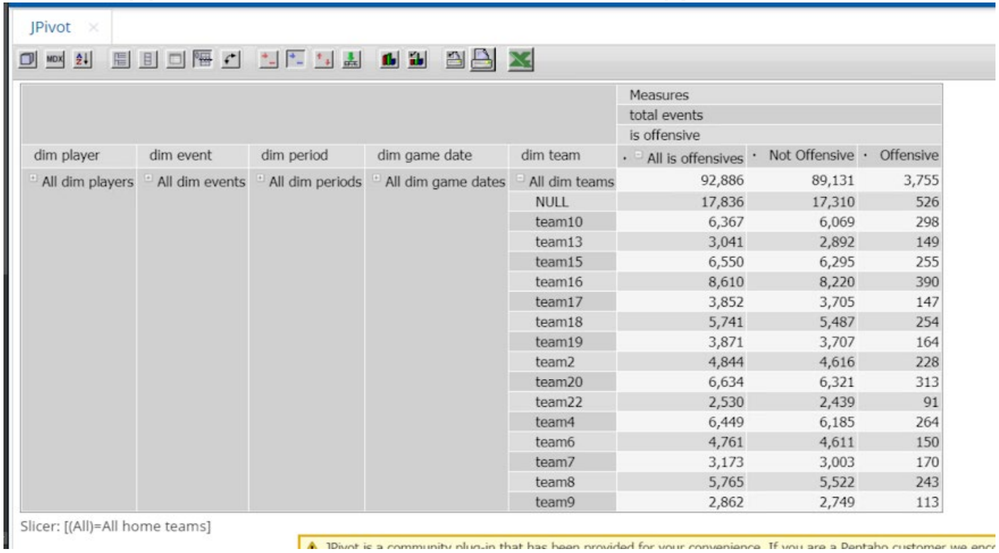
One of the most important objectives of this Data warehouse is to build metrics on the performance of players and teams that can be used to generate visually rich reports. The fact table can roll up or extended further as the grain we have chosen is fairly lowe. Ideal and insightful reports would be those that describe:

- All the stats of players who contribute towards our objective.
- Teams that exhibit offensive play and by what margin.
- Events that would help us understand which teams perform offensively and to what degree.
- Which events contribute towards the offensive play.

Analysis using OLAP cubes:

We used Pentaho Schema Workbench to setup dimensions and OLAP cube to publish them in Pentaho BI Server so that we can analyze the data using JPivot table to filter more detailed information on how offensive these teams are. Below are a few examples of OLAP cube analysis:

1. From the JPivot table below, we can tell that the team that has the largest number of offensive events is team 16. (We ignored NULL team which appeared due to data inconsistency)

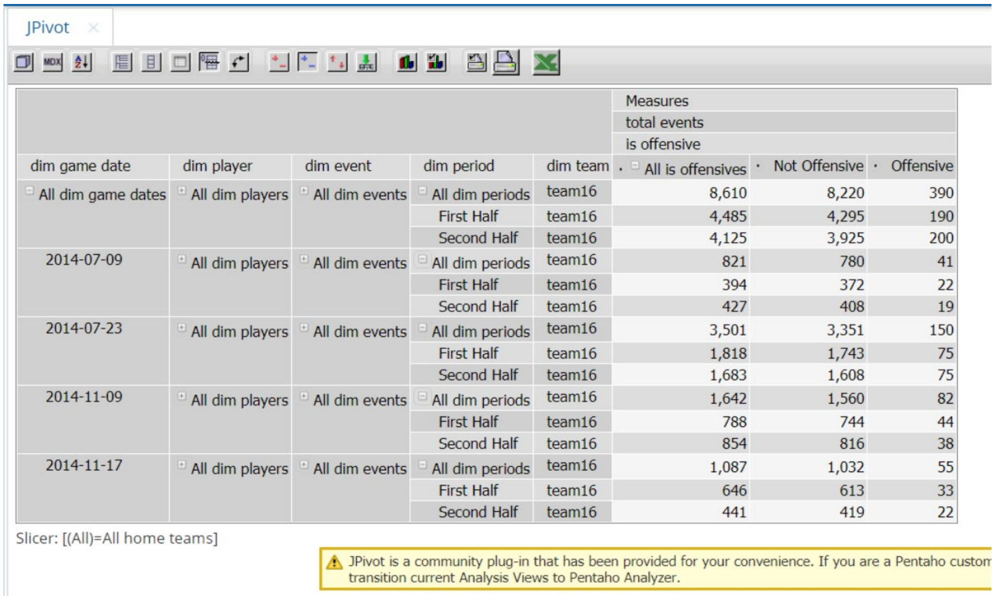


					Measures		
					total events		
					is offensive		
dim player	dim event	dim period	dim game date	dim team	All is offensives	Not Offensive	Offensive
All dim players	All dim events	All dim periods	All dim game dates	All dim teams	92,886	89,131	3,755
				NULL	17,836	17,310	526
				team10	6,367	6,069	298
				team13	3,041	2,892	149
				team15	6,550	6,295	255
				team16	8,610	8,220	390
				team17	3,852	3,705	147
				team18	5,741	5,487	254
				team19	3,871	3,707	164
				team2	4,844	4,616	228
				team20	6,634	6,321	313
				team22	2,530	2,439	91
				team4	6,449	6,185	264
				team6	4,761	4,611	150
				team7	3,173	3,003	170
				team8	5,765	5,522	243
				team9	2,862	2,749	113

Slicer: [(All)=All home teams]

⚠ JPivot is a community plug-in that has been provided for your convenience. If you are a Pentaho customer we enc

2. Drilling down into the cube, a more detailed JPivot table shows the dates and periods that team 16 played and how many offensive event they did on each date and period.

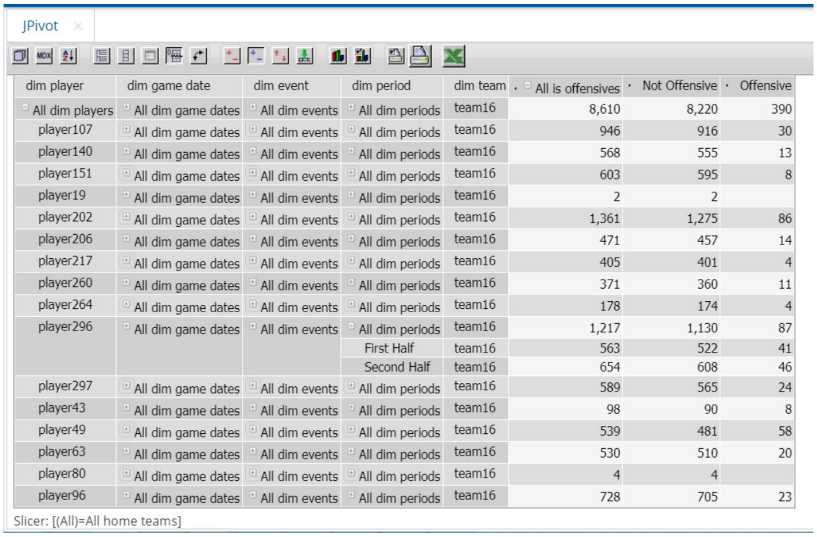


					Measures		
					total events		
					is offensive		
dim game date	dim player	dim event	dim period	dim team	All is offensives	Not Offensive	Offensive
All dim game dates	All dim players	All dim events	All dim periods	team16	8,610	8,220	390
			First Half	team16	4,485	4,295	190
			Second Half	team16	4,125	3,925	200
2014-07-09	All dim players	All dim events	All dim periods	team16	821	780	41
			First Half	team16	394	372	22
			Second Half	team16	427	408	19
2014-07-23	All dim players	All dim events	All dim periods	team16	3,501	3,351	150
			First Half	team16	1,818	1,743	75
			Second Half	team16	1,683	1,608	75
2014-11-09	All dim players	All dim events	All dim periods	team16	1,642	1,560	82
			First Half	team16	788	744	44
			Second Half	team16	854	816	38
2014-11-17	All dim players	All dim events	All dim periods	team16	1,087	1,032	55
			First Half	team16	646	613	33
			Second Half	team16	441	419	22

Slicer: [(All)=All home teams]

⚠ JPivot is a community plug-in that has been provided for your convenience. If you are a Pentaho custom
transition current Analysis Views to Pentaho Analyzer.

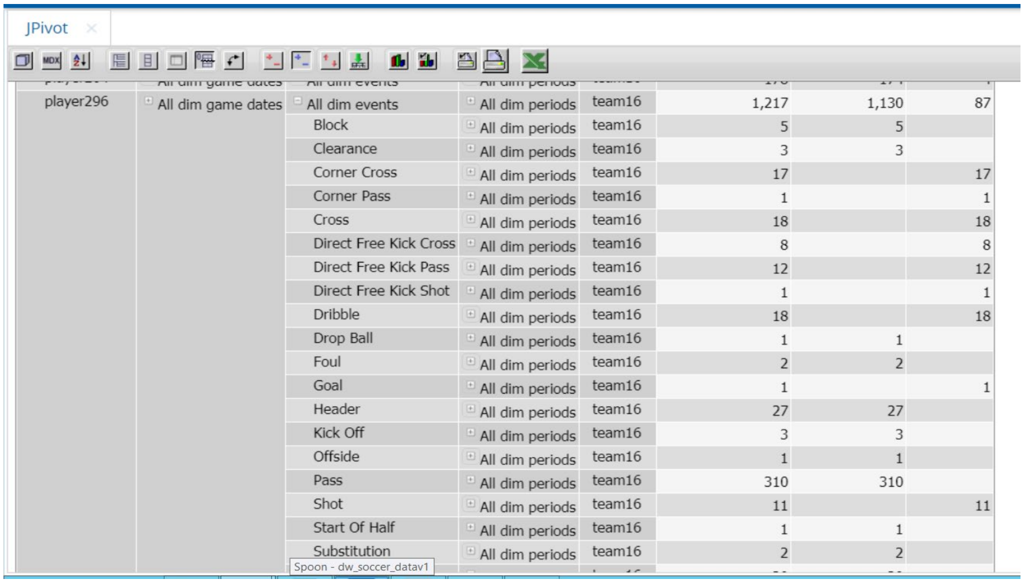
3. Then we looked at each player in team 16 with how many offensive events they performed as shown in the JPivot table below. It is easy to note that player 296 performed the most offensive events.



dim player	dim game date	dim event	dim period	dim team	All is offensives	Not Offensive	Offensive
All dim players	All dim game dates	All dim events	All dim periods	team16	8,610	8,220	390
player107	All dim game dates	All dim events	All dim periods	team16	946	916	30
player140	All dim game dates	All dim events	All dim periods	team16	568	555	13
player151	All dim game dates	All dim events	All dim periods	team16	603	595	8
player19	All dim game dates	All dim events	All dim periods	team16	2	2	
player202	All dim game dates	All dim events	All dim periods	team16	1,361	1,275	86
player206	All dim game dates	All dim events	All dim periods	team16	471	457	14
player217	All dim game dates	All dim events	All dim periods	team16	405	401	4
player260	All dim game dates	All dim events	All dim periods	team16	371	360	11
player264	All dim game dates	All dim events	All dim periods	team16	178	174	4
player296	All dim game dates	All dim events	All dim periods	team16	1,217	1,130	87
			First Half	team16	563	522	41
			Second Half	team16	654	608	46
player297	All dim game dates	All dim events	All dim periods	team16	589	565	24
player43	All dim game dates	All dim events	All dim periods	team16	98	90	8
player49	All dim game dates	All dim events	All dim periods	team16	539	481	58
player63	All dim game dates	All dim events	All dim periods	team16	530	510	20
player80	All dim game dates	All dim events	All dim periods	team16	4	4	
player96	All dim game dates	All dim events	All dim periods	team16	728	705	23

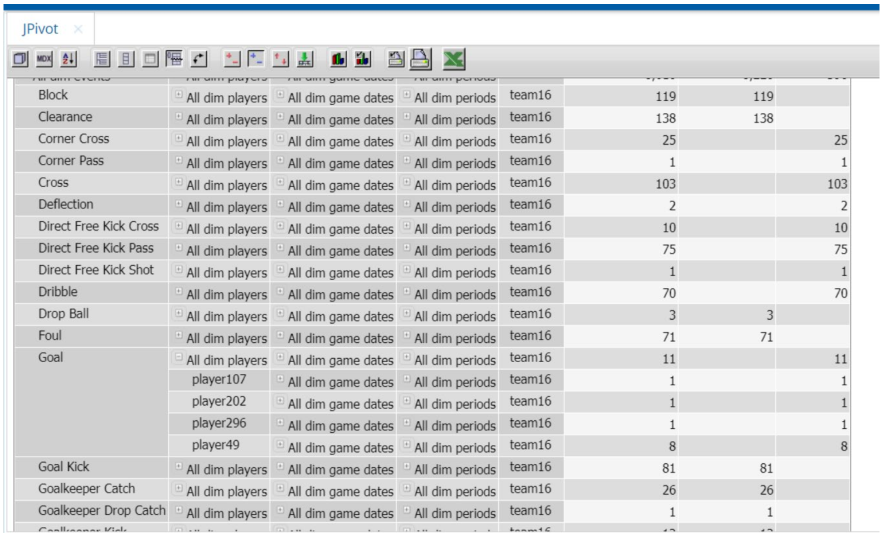
Slicer: [(All)=All home teams]

4. Looking at detailed performance of player 296 by filtering down to each event, it comes to a guess that player 296 should be the winger.



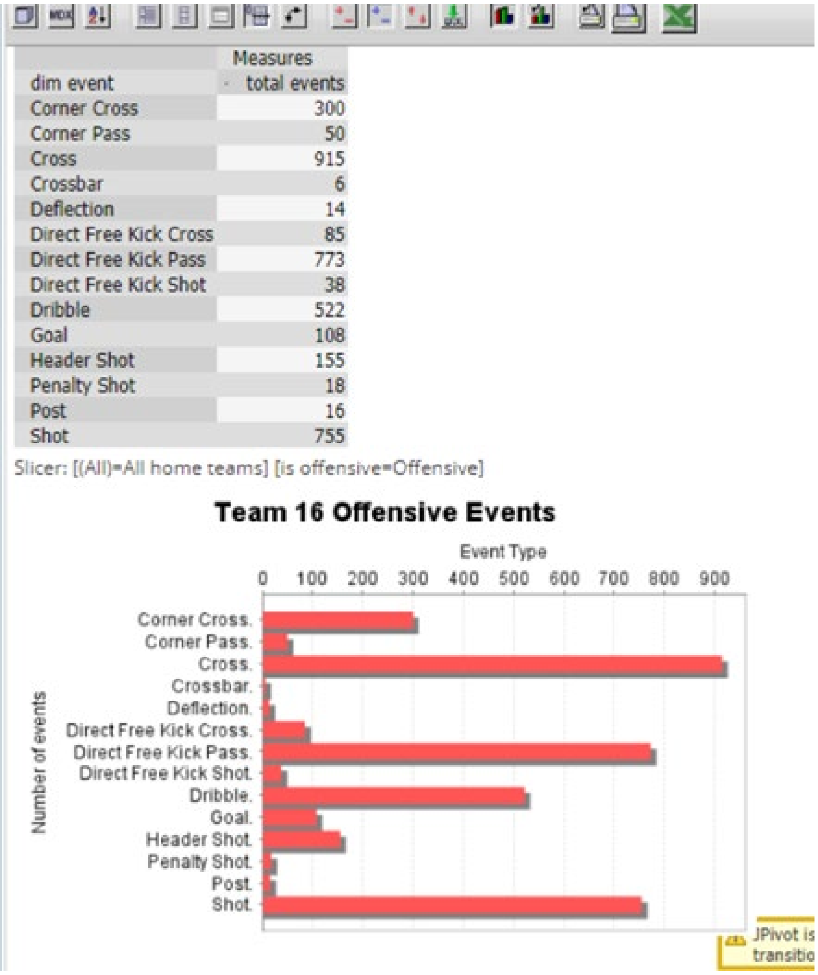
player296	All dim game dates	All dim events	All dim periods	team16	1,217	1,130	87
		Block	All dim periods	team16	5	5	
		Clearance	All dim periods	team16	3	3	
		Corner Cross	All dim periods	team16	17		17
		Corner Pass	All dim periods	team16	1		1
		Cross	All dim periods	team16	18		18
		Direct Free Kick Cross	All dim periods	team16	8		8
		Direct Free Kick Pass	All dim periods	team16	12		12
		Direct Free Kick Shot	All dim periods	team16	1		1
		Dribble	All dim periods	team16	18		18
		Drop Ball	All dim periods	team16	1	1	
		Foul	All dim periods	team16	2	2	
		Goal	All dim periods	team16	1		1
		Header	All dim periods	team16	27	27	
		Kick Off	All dim periods	team16	3	3	
		Offside	All dim periods	team16	1	1	
		Pass	All dim periods	team16	310	310	
		Shot	All dim periods	team16	11		11
		Start Of Half	All dim periods	team16	1	1	
		Substitution	All dim periods	team16	2	2	

5. Then if we only look at detailed data of event "Goal", we can see that player 49 has made the most goals, so he is probably a striker or a forward player.



Block	All dim players	All dim game dates	All dim periods	team16	119	119	
Clearance	All dim players	All dim game dates	All dim periods	team16	138	138	
Corner Cross	All dim players	All dim game dates	All dim periods	team16	25		25
Corner Pass	All dim players	All dim game dates	All dim periods	team16	1		1
Cross	All dim players	All dim game dates	All dim periods	team16	103		103
Deflection	All dim players	All dim game dates	All dim periods	team16	2		2
Direct Free Kick Cross	All dim players	All dim game dates	All dim periods	team16	10		10
Direct Free Kick Pass	All dim players	All dim game dates	All dim periods	team16	75		75
Direct Free Kick Shot	All dim players	All dim game dates	All dim periods	team16	1		1
Dribble	All dim players	All dim game dates	All dim periods	team16	70		70
Drop Ball	All dim players	All dim game dates	All dim periods	team16	3	3	
Foul	All dim players	All dim game dates	All dim periods	team16	71	71	
Goal	All dim players	All dim game dates	All dim periods	team16	11		11
	player107	All dim game dates	All dim periods	team16	1		1
	player202	All dim game dates	All dim periods	team16	1		1
	player296	All dim game dates	All dim periods	team16	1		1
	player49	All dim game dates	All dim periods	team16	8		8
Goal Kick	All dim players	All dim game dates	All dim periods	team16	81	81	
Goalkeeper Catch	All dim players	All dim game dates	All dim periods	team16	26	26	
Goalkeeper Drop Catch	All dim players	All dim game dates	All dim periods	team16	1	1	

6. We have also made a plot on team 16 offensive events in the Pentaho Server. This is easier to compare which offensive event they performed most.



	Measures
dim event	total events
Corner Cross	300
Corner Pass	50
Cross	915
Crossbar	6
Deflection	14
Direct Free Kick Cross	85
Direct Free Kick Pass	773
Direct Free Kick Shot	38
Dribble	522
Goal	108
Header Shot	155
Penalty Shot	18
Post	16
Shot	755

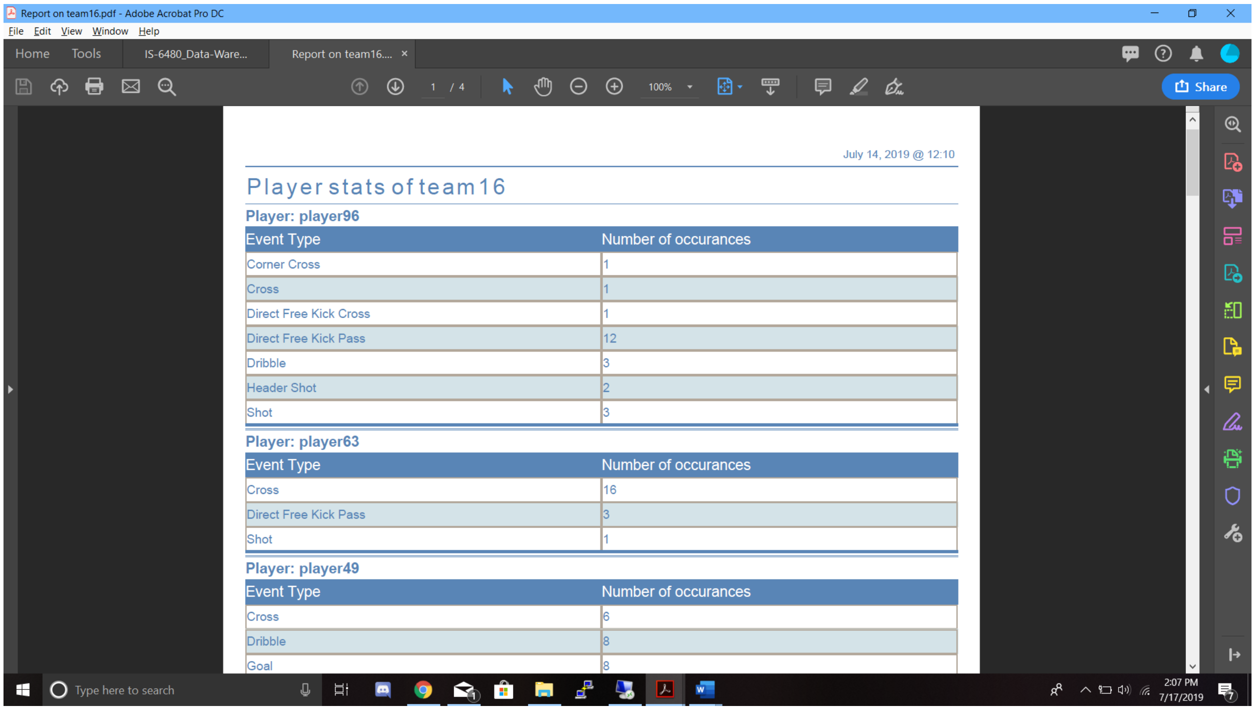
Slicer: [(All)=All home teams] [is offensive=Offensive]

JPivot is transitio

Plot and function in Pentaho Server

Currently functional reports:

Although the above tables are great for analysis, a visual representation is always more intuitive and easier to understand. Hence we used **Pentaho Report Designer** to present the above analysis along with visual representation that will help the team Coaches and the Manager to analyze how offensive or defensive their players are, what kind of approach to take when they come up against a tough opponent, how much ground players are covering throughout the game including the minutes spent on the field.


Report on team16.pdf - Adobe Acrobat Pro DC
File Edit View Window Help
Home Tools IS-6480_Data-Ware... Report on team16....
1 / 4
100%
Share
Type here to search
2:07 PM
7/17/2019

July 14, 2019 @ 12:10

Player stats of team16

Player: player96

Event Type	Number of occurances
Corner Cross	1
Cross	1
Direct Free Kick Cross	1
Direct Free Kick Pass	12
Dribble	3
Header Shot	2
Shot	3

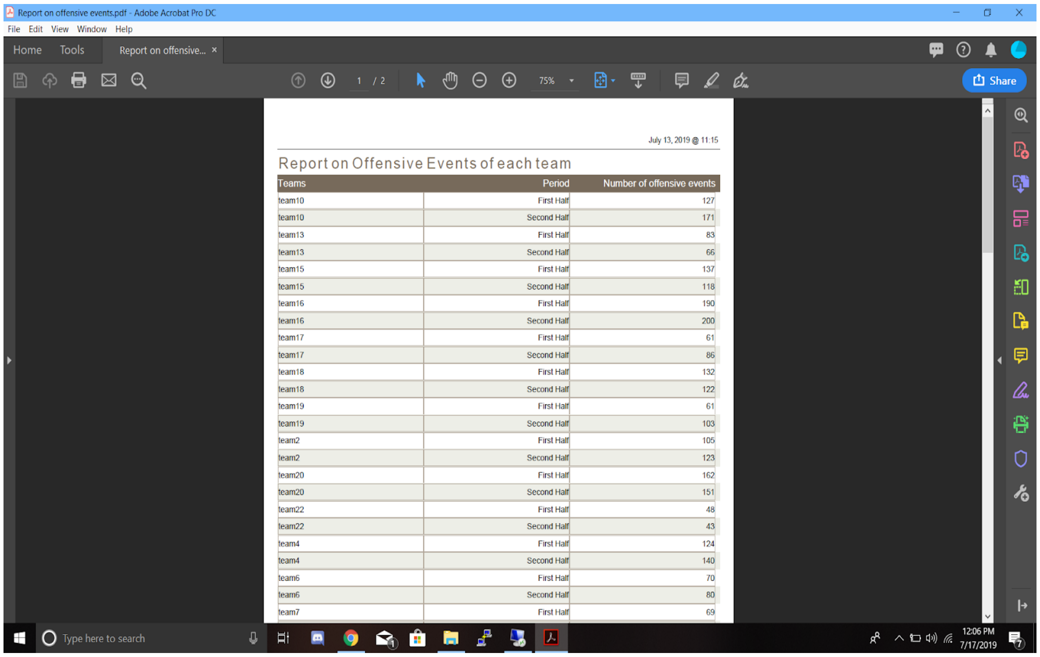
Player: player63

Event Type	Number of occurances
Cross	16
Direct Free Kick Pass	3
Shot	1

Player: player49

Event Type	Number of occurances
Cross	6
Dribble	8
Goal	8

Offensive stat of each Player from team 16

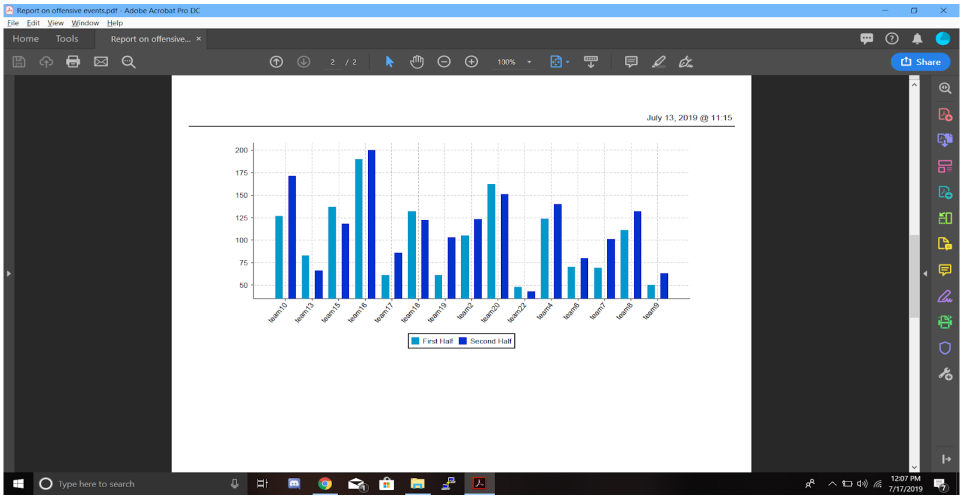


July 13, 2019 @ 11:15

Report on Offensive Events of each team

Teams	Period	Number of offensive events
team10	First Half	127
team10	Second Half	171
team13	First Half	83
team13	Second Half	66
team15	First Half	137
team15	Second Half	118
team16	First Half	190
team16	Second Half	200
team17	First Half	61
team17	Second Half	86
team18	First Half	132
team18	Second Half	122
team19	First Half	61
team19	Second Half	103
team2	First Half	105
team2	Second Half	123
team20	First Half	162
team20	Second Half	151
team22	First Half	48
team22	Second Half	43
team4	First Half	124
team4	Second Half	140
team6	First Half	70
team6	Second Half	80
team7	First Half	69

Total offensive events in each half by each Team



Visualization of the above report

Reports that could be supported by the current DW design but have yet to be developed:

1. Team that has poorest offensive play and why is that so.
2. Detailed report on each player for the entire season.
3. Which is the most offensive play contributing player in the entire league or for the team.
4. Player acquisition for a team. For example, if a team has very few crosses, which heavily contributes to an offensive play, then this team could target/acquire a player who has a 'cross' output which means a winger.
5. Best positional play of a player in the entire league which would essentially mean selecting best combined team of the league.
6. Team performance based on Half of the game.
7. Which team has the lowest Fan satisfaction.
8. Which is the one event that occurs the most in an offensive display. Based on this, teams could focus on generating other offensive events in their play.

Note: After the inclusion of Location dimension to the current DW model, we could develop reports that would detail on:

1. 'Zones' i.e. Forward, Midfield, and Defence in conjunction with Player stats. Which player suits what position and subsequent strategic changes.
2. Overall team play could be determined to be offensive or defensive. without having to go through game videos each time.

Conclusion

The most important objective of the project is to create a system that enables organisations to analyse the data with utmost flexibility and visualize these analysis using rich insightful reports. More specifically, the idea was to create and analyze performance metrics of teams and players to see the effectiveness of strategies and impact of players in a match and how they react to challenges before, during and after each game. Sports Coaches and support staff benefit more from visual content than statistics numbers. We used a host of different tools and applications from Pentaho to create this Data warehouse environment. Through this project, we were able to take small but a crucial step towards creating such an analytics environment.

References

1) Introduction: https://en.wikipedia.org/wiki/Premier_League
2) Lab materials – IS 6480 – Data Warehousing – by Prof. Michael J Boyle
3) The Data Warehouse Toolkit, 3rd Edition
4) Insights from Guest lecturers during the course
5) Tool Manuals: Pentaho Report Designer and Pentaho BI Server

Appendix A

Course materials and their applications

No.	Topic	Context
1	Class Lecture on Logical Design	Identifying Dimensions & Facts. Design of the Data warehouse
2	Lab 1	Creating logical dimensional model
3	Lab 2	Creating data integration and using Pentaho Spoon
4	Lab 3	Creating JPivot table using Schema Workbench
5	Lab 4	Report Designing using Pentaho Report Designer

Appendix B

Detailed Hours spent on different project tasks by each team member.

Date	Team Member	Hours Spent	Description of Work	Additional Comments
2019/06/10	Abhijay	2	Researching AWS environment	NA
2019/06/10	Angel	2	Going through tutorial about AWS	
2019/06/11	Gaurav	1	Creating bus matrix	NA
2019/06/11	Abhijay	1	Creating bus matrix	NA
2019/06/11	Angel	1	Creating bus matrix	
2019/06/12	Abhijay	1	Setting up development environment	NA
2019/06/12	Gaurav	1	Setting up development environment	NA
2019/06/12	Angel	1	Setting up development environment	NA
2019/06/18	Abhijay	0.5	Creating tables from the tab files	NA
2019/06/18	Gaurav	1	Creating tables from the tab files	NA
2019/06/18	Angel	0.5	Creating tables from the tab files	NA
2019/06/22	Gaurav	0.5	Gathering business requirements	NA
2019/06/25	Abhijay	1	Design of Logical Model	NA
2019/06/25	Gaurav	1	Design of Logical Model	NA
2019/06/25	Angel	1	Design of Logical Model	NA

2019/07/02	Abhijay	1	Preparing & cleansing data for ETL	NA
2019/07/02	Gaurav	1	Preparing & cleansing data for ETL	NA
2019/07/02	Angel	1	Preparing & cleansing data for ETL	NA
2019/07/06	Abhijay	2	Creating Transformations scripts for ETL	NA
2019/07/06	Gaurav	2	Creating Transformations scripts for ETL	NA
2019/07/06	Angel	2	Creating Transformations scripts for ETL	NA
2019/07/09	Abhijay	2	Reports and visualizations	NA
2019/07/09	Gaurav	2	Reports and visualizations	NA
2019/07/09	Angel	2	Reports and visualizations	NA
2019/07/13	Abhijay	1	Documentations	NA
2019/07/13	Angel	1	Documentations	NA
2019/07/13	Gaurav	3	Project report	NA
2019/07/13	Angel	3	Project report	NA
2019/07/16	Abhijay	1	Preparing Presentation	NA
2019/07/16	Gaurav	0.5	Preparing Presentation	NA
2019/07/17	Angel	1	Preparing Presentation	NA
2019/07/17	Abhijay	2	Preparing Presentation	NA

Appendix C

Bus Matrix

Business processes	**Game Date**	**Team**	**Player**	**Location**	**Period**	**Event**	**Status**
Manage player personnel strategy							NR
Acquire players		🟨	🟨				NR
Divest players		🟨	🟨				NR
Develop players							NR
On pitch, game time training	🟨	🟨	🟨			🟨	NR
Off pitch Training			🟨	🟨		🟨	NR
Manage injuries							PF
Physical condition management			🟨				PF
Manage fitness							PF
Diet management			🟨				PF
Manage player personnel tactics							C
Goalie Tactics			🟨		🟨		C
Forward Tactics			🟨	🟨	🟨		C
Midfiled Tactics			🟨	🟨	🟨		C
Defensive Tactics			🟨	🟨	🟨		C
Manage game/opponent tactics							NR
Formations managements			🟨	🟨			NR
Player instructions			🟨	🟨			NR
Fouls management	🟨					🟨	NR
Entertain fans							C
Offense productions	🟨	🟨	🟨	🟨	🟨	🟨	C
Goal productions	🟨		🟨			🟨	C

Abbrevations:
Completed - C
Needs refinement - NR
Planned for future - PF