Group 4

Elizabeth Bandy
Robert Han
Romit Shah
Jeff Watson

IS – 6480 Data Warehousing, Summer 2018
Group Final Project

TABLE OF CONTENTS

EXECUTIVE SUMMARY

As a top-rated team within the Major League Soccer organization, the Vernal Velociraptors are the pride of northeastern Utah. Our graciously supportive enthusiasts and home-field attendees journey to see us from as far as Myton and Maeser, from Ballard City and Harpers Corner. To maintain appeal and to broaden geographical penetration, we must focus on winning matches.

OUR THEORY

Winning games has a large impact on revenue and fan loyalty.

VISION

To promote the art of in-person collaboration,
athletic prowess, and inter-community involvement.

MISSION

Attract and entertain guests, of all ages, with the excitement of soccer.

PRODUCTS & SERVICES

Our product is entertainment. Our primary form of entertainment is the *experience* of a home-field stadium match played against fellow members of the United States Soccer Federation (USSF). This is an on-site, sensory-rich occasion. Guests watch the on-field action and they hear the passion of the crowd. They feel the vibration of the cheering, smell the concessions, and taste the refreshments. Viewing a soccer match will generate a range of emotions, the most important of which is excitement for the game.

Our secondary form of entertainment is between-match spectator engagement. We accomplish this via social media and our interactive website. Occasionally, a pair of players will "crash" backyard cookouts of season ticket holders and arrive unannounced at children's birthday parties.[1] We also license the likeness and the logo of the Vernal Velociraptors to appear on third-party merchandise. This visual penetration includes apparel, stationery, beverageware, and toys. Our mascots, Velma & Victor Velociraptor, energize our guests pre-game and during halftime.

[1] When requested in advance by an adult family member and based on player and/or team availability.

Our service is community stewardship. Our players participate in public appearances to bolster reciprocity with local businesses and institutions. We promote youth athletics, not only team-based, but also one-on-one and individual. Our objective is to encourage the adults of tomorrow to be healthy and active today. Introducing youth to potential hobbies or unique interests is designed to broaden their opportunities for education as well as divert them from fallow activities such as gangs, drugs, and violence.

BENEFITS OF A DATA WAREHOUSE

A data warehouse will allow us to amass, over time, from disparate sources, the information required to tell the story of Vernal Velociraptors, humanly, financially, and operationally. We can include ticket sales data, merchandise sales data, and social media data.

Following each match, particularly each match we win, we can monitor tickets purchases. Turnstile data determines attendance at subsequent games. These two metrics allow us to measure not only box office, but also concessions revenue.

We can tally merchandise sales at the stadium at the end of a winning match, and all our authorized merchandise retailers are equipped to provide to us daily point-of-sales data.[2] Our marketing information systems will provide social media activity. These components will allow us to survey fan loyalty and engagement.

Our data warehouse will be a consortium of information, providing a view of our present, as well as lead our future.

[2] Using the EDI (electronic data interchange) 852 transaction set.

PRIORITIZED REQUIREMENTS

REQUIREMENTS SUMMARY

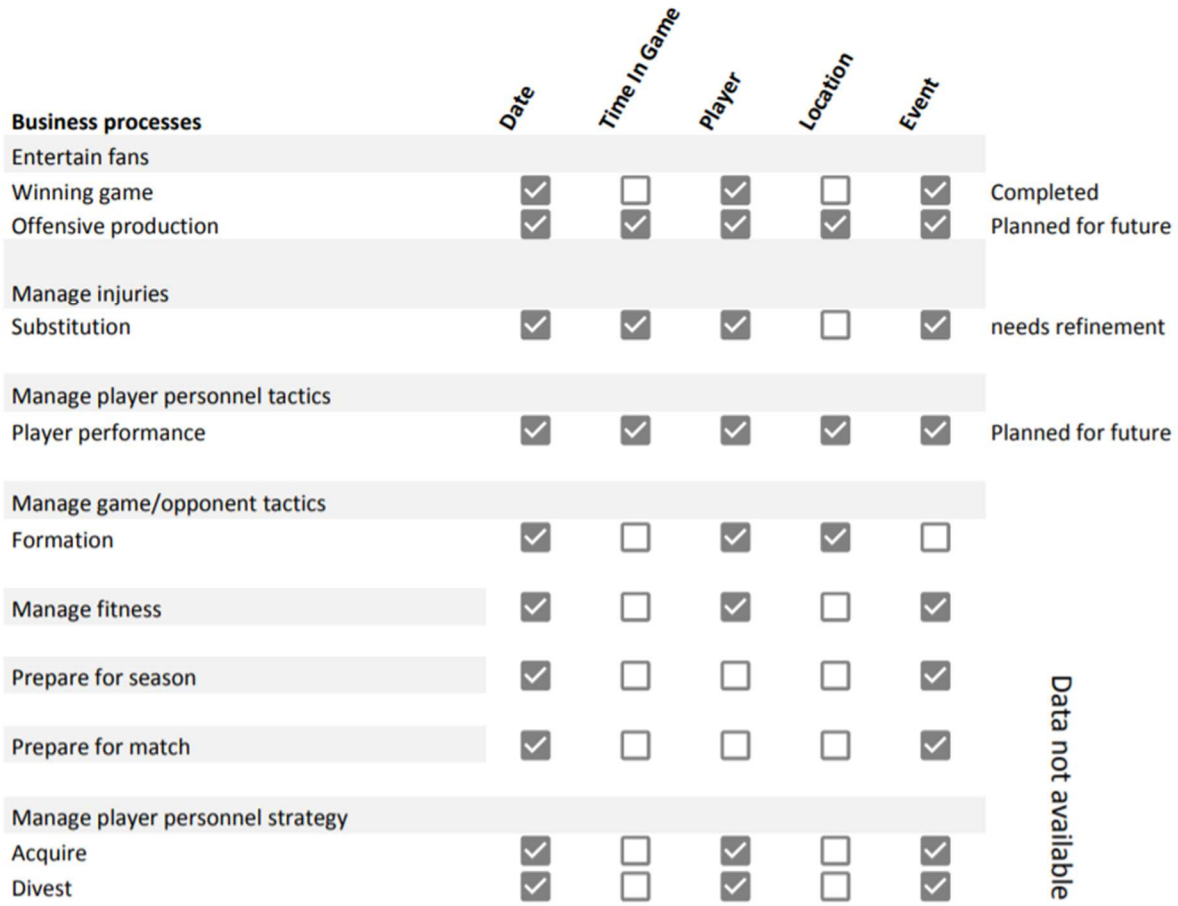
The image below is the data warehouse bus matrix categorizing the requirements at a summary level. Since our project targets the soccer operations of winning games, which has a large impact on revenue and fan loyalty, it tops the bus matrix as the first priority.

Now that we have determined the business requirement which has the largest potential business impact, we should concentrate on its feasibility. Is the core data available to execute on this business requirement? To answer this question, a data analysis on the given data files would be necessary.

We were provided two data files: an event dataset and a location dataset. The event dataset focuses on the events which occurred during the match like shots, goals, goalkeeper saves, etc. The location dataset contains a collection vast of records indicating the locations of the players throughout a match. Since both of them only concentrate on the time period during the match, we are unable to manage the business processes such as season preparation or player personnel strategy.

As mentioned above, our project focuses on the "winning game" soccer operations. We determined we could easily manage the outcomes of winning games with the data access to the number of goals in a certain match by the host team and its opponent. The event tab file contains the records of the "goal" event type with other necessary attributes like the player team ID attached to it. As a result, we decided to abandon the location tab file to simplify our data warehouse and concentrate on the target strategy of our project.

Figure 1 - Requirements Summary Bus Matrix

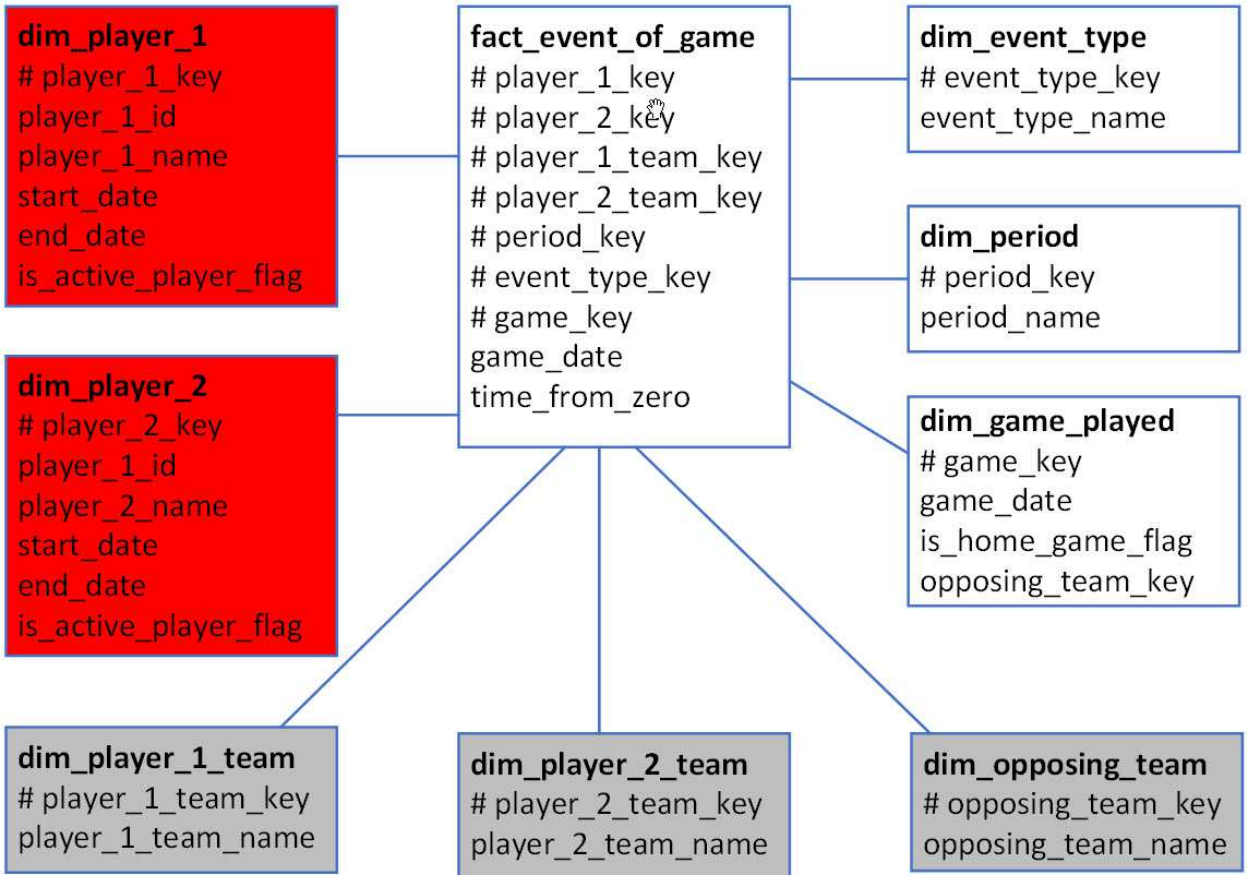


REQUIREMENTS DETAIL

#	Requirement Name	Requirement Description
1	System Availability	The data warehouse must be accessible 24/7/365 apart from scheduled maintenance.
2	Secure File Transfer	All files received from outside entities must be delivered via SFTP, Secure File Transfer Protocol.
3	Extract Automation	The ETL tool must have the ability to draw data from our existing systems on a scheduled basis. Similarly, files received from outside entities must be loaded and processes automatically.
4	Data Stewardship	Assign a Data Steward to be familiar with the source data, the transformation process, and the table(s) as presented within the data warehouse.

LOGICAL MODEL

Within our logical model are two role-playing dimensions which are represented as views. Highlighted in red, we have rendered our player dimension twice, first as player #1, and second as player #2. Highlighted in green, we have identified teams as player #1, player, #2, and opposing.



PHYSICAL MODEL

PHYSICAL DESIGN

The following two images represent the script we used to define our scheme, our dimension tables, and our fact tables.

Figure 2 - Schema and Table Definition Script, 1 of 2

```
-- -----------------------------------------------------
--
-- IS-6480 Data Warehousing, Summer 2018
-- Team 4:  Elizabeth Bandy, Robert Han, Romit Shah, Jeff Watson
--
-- The objective of this script is to delete any instance of
-- schema `` and re-create it.  Four
-- dimension tables will be created, and one fact table will
-- be created.
--
-- -----------------------------------------------------

SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0;
SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS, FOREIGN_KEY_CHECKS=0;
SET @OLD_SQL_MODE=@@SQL_MODE, SQL_MODE='TRADITIONAL,ALLOW_INVALID_DATES';

DROP SCHEMA IF EXISTS `dw_soccer_project_group4`;

CREATE SCHEMA `dw_soccer_project_group4` DEFAULT CHARACTER SET utf8;
USE `dw_soccer_project_group4`;

-- drop table fact_event_of_game;
-- drop table dim_players;
-- drop table dim_event_type;
-- drop table dim_teams;
-- drop table dim_period;

-- -----------------------------------------------------
-- Table 'dim_players'
-- -----------------------------------------------------

CREATE TABLE dim_players
(
   player_key BIGINT AUTO_INCREMENT NOT NULL PRIMARY KEY
  ,player_id VARCHAR(12)
);
CREATE UNIQUE INDEX idx_dim_players_pk ON dim_players(player_key);

-- -----------------------------------------------------
-- Table 'dim_event_type'
-- -----------------------------------------------------

CREATE TABLE dim_event_type
(
  event_type_key BIGINT AUTO_INCREMENT NOT NULL PRIMARY KEY
, event_name VARCHAR(24)
, last_updated DATETIME
);
CREATE UNIQUE INDEX idx_dim_event_type_pk ON dim_event_type(event_type_key);

-- -----------------------------------------------------
-- Table 'dim_teams'
-- -----------------------------------------------------

CREATE TABLE dim_teams
(
  team_key BIGINT AUTO_INCREMENT NOT NULL PRIMARY KEY
, team_id VARCHAR(6)
);
CREATE UNIQUE INDEX idx_dim_teams_pk ON dim_teams(team_key);

```

Figure 3 - Schema and Table Definition Script, 2 of 2

```
-- ----------------------------------------------------
-- Table 'dim_period'
-- ----------------------------------------------------

CREATE TABLE dim_period
(
  period_key BIGINT AUTO_INCREMENT NOT NULL PRIMARY KEY
, period_name VARCHAR(24)
);
CREATE UNIQUE INDEX idx_dim_period_pk ON dim_period(period_key);

-- ----------------------------------------------------
-- Table 'fact_stage_event_of_game'
-- ----------------------------------------------------

 CREATE TABLE fact_stage_event_of_game
 (
   game_date_gen DATETIME
 , v_team_gen VARCHAR(12)
 , h_team_gen VARCHAR(12)
 , period_desc VARCHAR(24)
 , time_from_zero DOUBLE
 , event_type VARCHAR(64)
 , player_name_1_gen VARCHAR(12)
 , player_name_2_gen VARCHAR(12)
 , player_1_team_gen VARCHAR(12)
 , player_2_team_gen VARCHAR(12)
 , last_updated DATETIME
 );
CREATE INDEX idx_fact_stage_event_of_game_lookup ON fact_stage_event_of_game(game_date_gen, v_team_gen, h_team_gen
                                                                             ,period_desc, time_from_zero, event_type
                                                                             ,player_name_1_gen, player_name_2_gen
                                                                             ,player_1_team_gen, player_2_team_gen);

-- ----------------------------------------------------
-- Table 'fact_event_of_game'
-- ----------------------------------------------------

CREATE TABLE fact_event_of_game
(
  game_date VARCHAR(10)
, time_from_zero DOUBLE
, event_type_key BIGINT
, event_name VARCHAR(24)
, period_key BIGINT
, period_name VARCHAR(24)
, player_1_key BIGINT
, player_1_id VARCHAR(12)
, player_2_key BIGINT
, player_2_id VARCHAR(12)
, team_1_key BIGINT
, team_1_id VARCHAR(6)
, team_2_key BIGINT
, team_2_id VARCHAR(6)
)
;
CREATE INDEX idx_fact_event_of_game_lookup ON fact_event_of_game(game_date, event_type_key, period_key
                                                                ,player_1_key, player_2_key);

-- ----------------------------------------------------
-- End of Script
-- ----------------------------------------------------
```

Referential integrity will be applied by introducing foreign keys to the `FACT_EVENT_OF_GAME` for all fields which contain "_key" in the fieldname.

DESCRIPTION OF THE SAMPLE DATA

The sample data for event-related information is in a tab-delimited text file. The "Data Type" identified for each of column is a visual examination; the actual values appear in the file as character strings.

Column Position	Header Name	Data Type	Description
1	game_date_gen	ccyy-mm-dd format Date (e.g. 2014-11-30)	The date the match was played.
2	v_team_gen	Alphanumeric text (e.g. team10)	Name of visiting team.
3	h_team_gen	Alphanumeric text (e.g. team10)	Name of home team.
4	period_desc	Text (e.g. Second Half)	Indicates either 1st or 2nd half
5	time_from_zero	A 7.2 precision decimal (e.g. 1764.23)	Duration in seconds from start time at which event occurred.
6	event_type	Text (e.g. Pass, Touch, Ball Out)	Name of type of event.
7	player_name_1_gen	Alphanumeric text (e.g. player199)	Name of player who caused the event.
8	player_name_2_gen	Alphanumeric text (e.g. player199)	Name of player to whom the event was directed, if applicable.
9	player_1_team_gen	Alphanumeric text (e.g. team10)	Name of team of player who caused the event.
10	player_2_team_gen	Alphanumeric text (e.g. team10)	Name of team of player to whom the event was directed, if applicable.

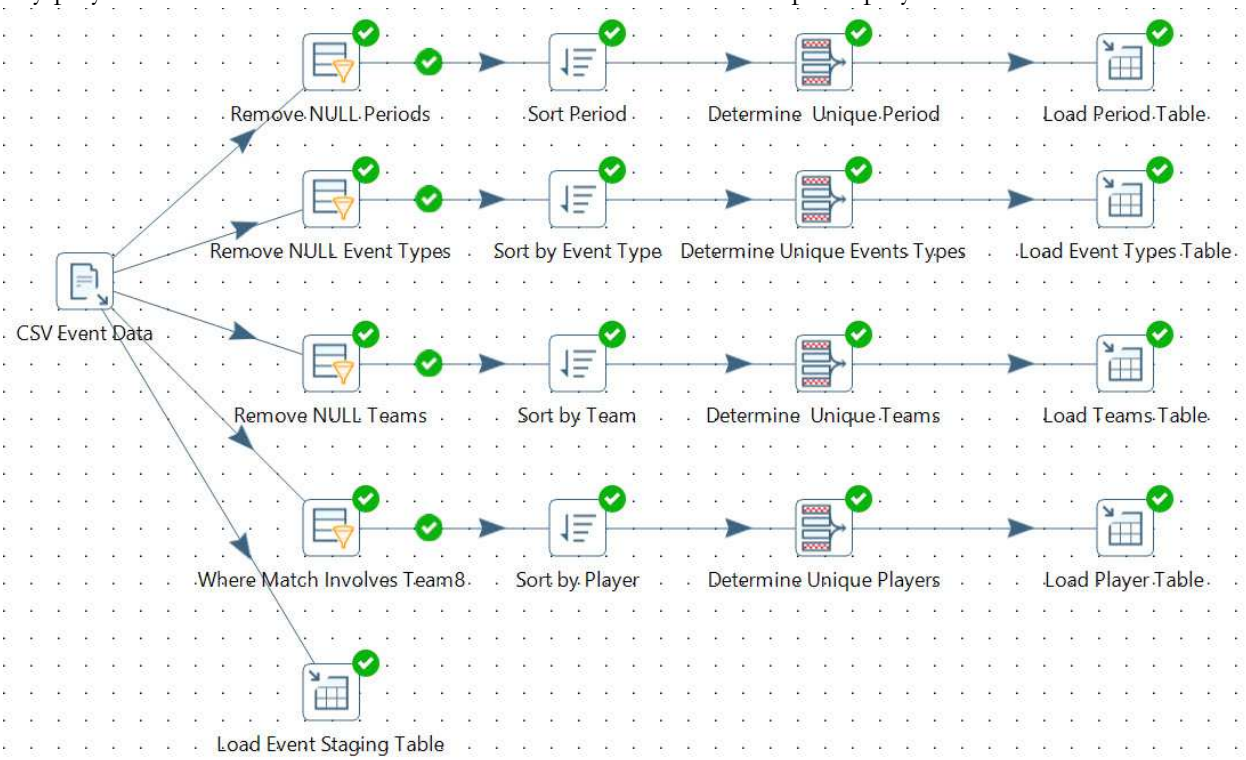
DEPLOYMENT TO TARGET ENVIRONMENT

The data warehouse will be deployed to a virtual server hosted by Amazon Web Services, instance type t2.2xlarge with 160 gigabytes of storage. The operating system will be Windows Server 2012 R2, the RDBMS MySQL, and the ETL tool Pentaho Data Integration.

ETL PROCESSES

Loading the Dimension Tables

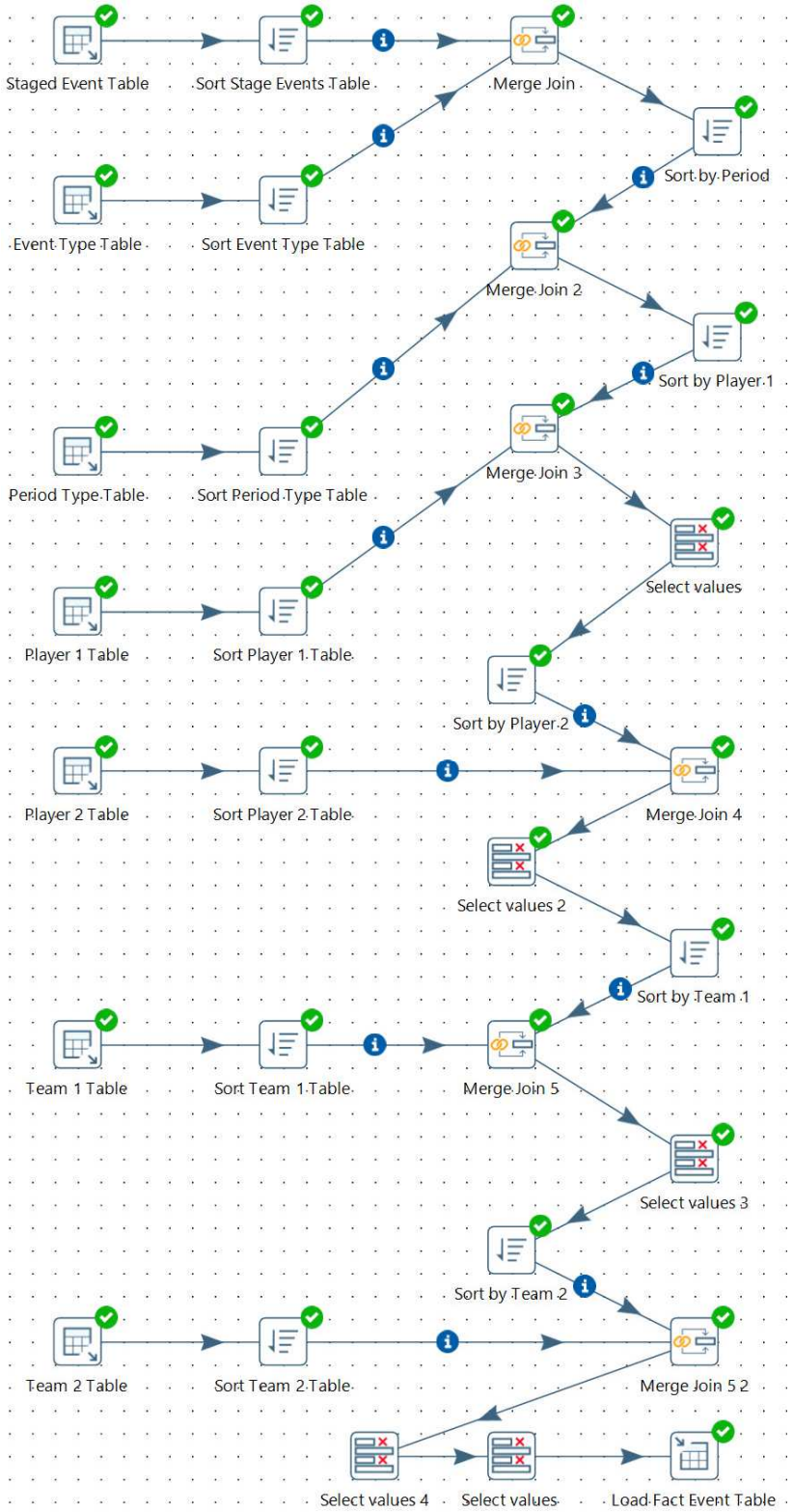
The first three lines of ETL process below will create dimension tables for periods, event types, and teams. In doing so, they will remove nulls and duplicates. The fourth line will create a unique list of any player associated with a match in which the Vernal Velociraptors played.



Loading an Event Staging Table

The fifth line in the ETL process above, merely imports the CSV-based event data into a database table named `FACT_STAGE_EVENT_OF_GAME`. No field renaming, no data type changing, no column inserts, no column deletions; rather, it is intended to increase processing efficiency.

Loading the Event Table



The image to the left is the ETL process for loading data warehouse table `FACT_EVENT_OF_GAME`. Beginning with the event staging table, the process will use a series of Sort and Merge Join tools to add relative keys for attributes. The Select Values tools will rename fields for clarity and remove excess columns.

SECURITY FEATURES

Access to the data warehouse will be multi-tiered. Only Data Architects will possess full-access into Pentaho within AWS. Developers will have access to Tableau Server within AWS. Data Analysts will have rights to the Tableau application and the Pentaho report development tool within AWS. Viewers of the Tableau reports will access the output via a browser.

Database schemas and data mart will be constructed and roles will be associated with relative tables. The non-analyst community will be assigned to restrictive roles.

The above positions will require a user ID and 12-character password to attach to the local area network. Each application/environment will require an additional user ID and 12-character password.

ADMINISTRATION FEATURES

We have used the following tools, products, and utilities to achieve our goals with our data warehouse system.

MySQL Workbench: with the help of this tool we created the logical/dimensional model by creating a fact table and dimension tables, and joined them by creating 1:n identifying relationships. Also, we used this tool to generate the physical model of our schema.

Pentaho Data Integration: this tool was used to carry out the ETL processes. We extracted the data from the MySQL data source and loaded them into fact and dimension tables.

Schema workbench: the OLAP cube was generated by using the schema file. This schema file was created in the schema workbench.

Tableau: our analysis is transformed into various beautiful visualizations in Tableau BI tool.

SOLUTION APPROACH

DATA REVIEW

Our initial step was to review the data in its rawest form. As a result, we were able the determine the nature of the data as can be seen in the *Description of the Sample Data* section on page 9. WE noticed missing data, for example, no player #1 associated with a play. We also determined a missing player #2 is valid with some types of plays.

FIELD NAMES

With numerous systems using different fieldnames within their exported files inbound to our data warehouse, we developed a naming convention for the most frequently encountered fields. As additional systems and their data streams are introduced to our data warehouse, new fields will adhere to a similar nomenclature.

CREATING THE DATA WAREHOUSE ENVIRONMENT

To create the schema, we will execute the script will appears within the *Physical Design* section on page 7. Secondly, we will populate the master files using the ETL process identified within the *Loading the Dimension Tables* section on page 10. Thirdly, we will backfill the `FACT_EVENT_OF_GAME` table within the process displayed on page 11 within the *Loading the Event Table* section.

Although the dimension tables need be populated only one, the portion which populates the event staging table portion can be repurposed and executed on a routine basis, as match data becomes available, followed by the *Loading the Event Table* process.

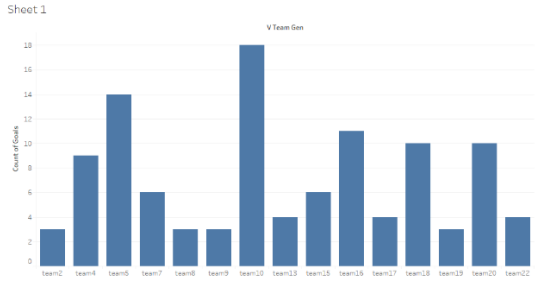
REPORT & ANALYSIS CAPABILITY

REPORTS

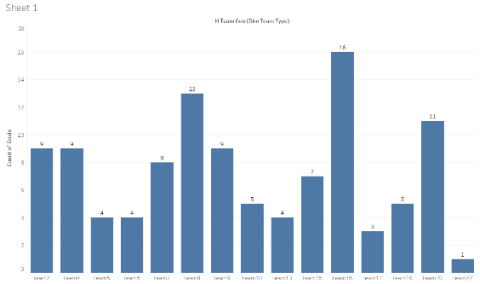
The report executed pulls data for various entities responsible for attracting fans and generating revenue. For our analysis, we have generated visualizations in Tableau where our main focus is on the following features. These features indirectly correspond to ticket sales, merchandising product sales, social media value, and entertainment for fans.

1) The number of goals made by the visiting/home teams

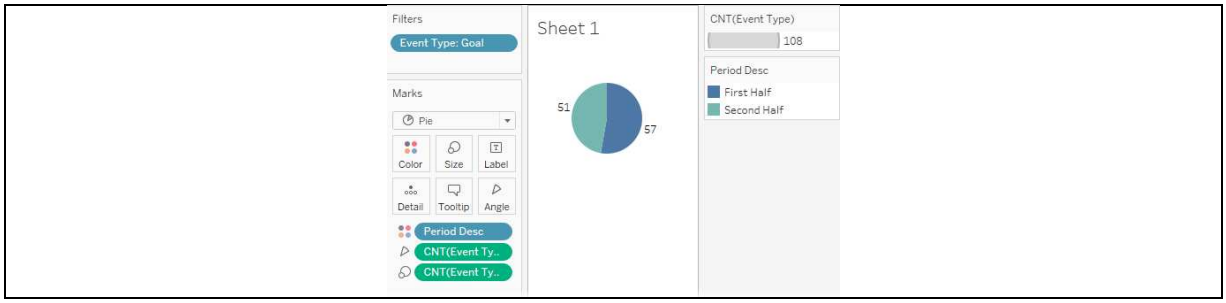
Visiting Team Goal Summary



Home Team Goal Summary

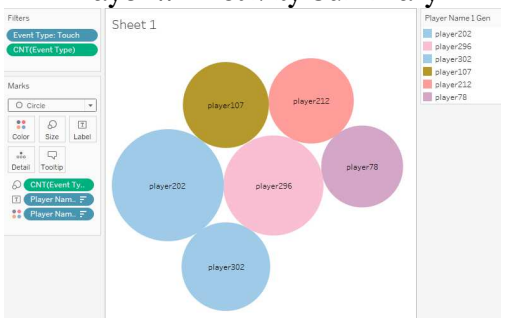


2) The number of goals made in the first half/second half



3) The names of top players performing well by having the maximum count of the touch event.

Player #1 Activity Summary



Player #2 Activity Summary

DATA CONNECTIONS

Initially, data connections from the data warehouse will be limited to Tableau Server and the Pentaho report development tool. Both of these options will allow for export to desktop applications. Initially, analytic applications will be limited to Excel.

FUTURE FEATURES

With respect to the scope of this project, our current report focuses on the large impact of winning games on revenue and on fan loyalty. There are many features that could be supported by our current data warehouse design. Below is a list of the reports for future development.

- Teams are more likely to gain fan loyalty with high offensive production.
  --- A report on the fouls and yellow cards received by the teams and players
- Teams who win more games in the home fields would gain more revenue and fan loyalty.
  --- A report on the game wined by the teams in the home courts.
- Teams are more likely to win the games with a better defense.
  --- A report on the goalkeeper saves and shots by the opposing team

These features have not yet been developed. However, if our data warehouse is well architected, new reports could be designed easily to satisfy business requirements in the future.

CONCLUSION

We, the Business Intelligence Development Team, believe strongly a data warehouse will allow the Vernal Velociraptors Soccer Organization to associate match outcomes with revenue and fan loyalty. Using predictive analytics tools, such as R, we can forecast attendance and staff our stadium accordingly. In subsequent phases we can implement athlete performance and injury data to predict player potential.

BIBLIOGRAPHY

Casters, Matt, Roland Bouman, and Jos van Dongen. 2010. *Pentaho Kettle Solutions.* Indianapolis, IN: Wiley Publishing, Inc.

Kimball, Ralph, and Margy Ross. 2013. *Data Warehouse Toolkit, The.* Indianapolis, IN: John Wiley & Sons, Inc.

Knaflic, Cole Nussbaumer. 2015. *Storytelling with Data.* Hoboken, NJ: John Wiley & Sons, Inc.

Roldán, María Carina. 2013. *Pentaho Datad Integration, Second Edition.* Birmingham, UK: Pakt Publishing Ltd.

APPENDIX

Team Member Time Participation

Date	Team Member(s)	Hours Spent	Description of Work
2018/6/16	All	2	-plan for the project -setting up development environment
2018/6/17	Robert	3	Business requirements gathering
2018/6/23	Romit, Jeff	3	Dimensional model design
2018/6/23	Robert	1.5	ETL research
2018/6/23	Elizabeth	3	Presentation slides and project summary report design
2018/6/30	Jeff	4	-Create and load the dimensional model - examine physical design
2018/6/30	Romit	1	Dimensional model description
2018/6/30	Elizabeth	2.5	Organization description
2018/6/31	Robert, Elizabeth	1.5	Requirements report
2018/7/5	All	1.5	Stage summary and discussion
2018/7/14	Romit	3.5	OLAP schema design and creation
2018/7/14	Jeff	4	Physical design, ETL processes description
2018/7/15	Elizabeth	4	Presentation slides creation
2018/7/17	Robert	2	Technologies and sample data used
2018/7/21	Romit	3	Reports design and creation
2018/7/21	Romit, Robert	3	Analyses
2018/7/21	Jeff, Elizabeth	2	Review and summary
2018/7/24	All	2.5	-Final discussion -Rehearse presentation