Akshay Ankush Kapare

Akshay08@bu.edu

Abstract

This Document provides an DATAWARE HOSE for IPL (Indian Premier League) Ball by ball data about each match played in between year 2008 to 2017. The main purpose of the document is to find runs, wickets, winning of each team participated in IPL

IPL DATA WAREHOUSE

MET CS 689 Term Project

MET CS 689 Term Project

Table of Contents

[1. Introduction 2](#_Toc8753632)

[2. Design Process 2](#_Toc8753633)

[3. ETL 3](#_Toc8753636)

[1) Extract 3](#_Toc8753637)

[2) Transform 4](#_Toc8753638)

[3) Load 5](#_Toc8753638)

[4) Why this ETL 5](#_Toc8753638)

[4. Referrences 6](#_Toc8753639)

# Introduction

The Indian Premier League (IPL) is a professional Twenty20 cricket league in India usually contested between March and May of every year by eight teams representing eight different cities or states in India.

In this project I have taken data for every bowl bowled in match and what was the outcome for that bowl. Using that data, I have calculated total run per year, total extras ratio etc.

# Design Process

1. After taking Data from Kaggle I have worked on data to figure out what question I will be answering form Datawarehouse
2. According to following question I have created Star Schema :
   * 1. Total Run scored per season.
     2. Total Number of Sixes hit per season.
     3. Total run scored by captains per season.
     4. Which team won maximum number of matches per season.
     5. Each year Extra run ratio with Run scored

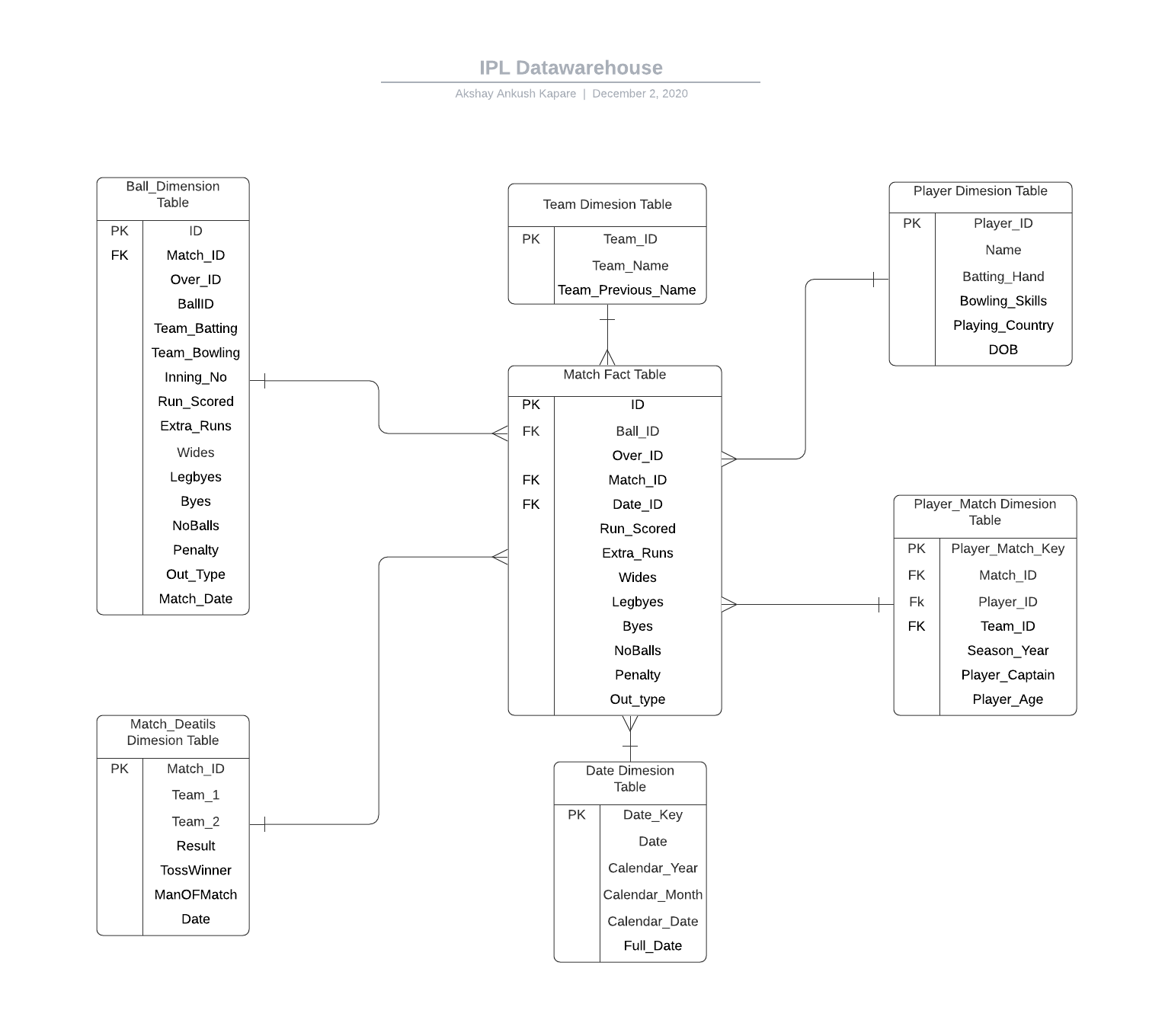
C.STAR SCHEMA:

1.Slowly changing Dimension:

In star Schema I have selected Team name as Slowly changing Dimension as Current Team Name and previous team Name Reason beginf slelcting it has slowly changing dimension is Team name Might change after 5 -6 years only if owner wish change it or Owner itself changed.

2.Hierarchies

In Date Dimension Table Hierarchies will be Calendar Year, Calendar Month, Calendar Day

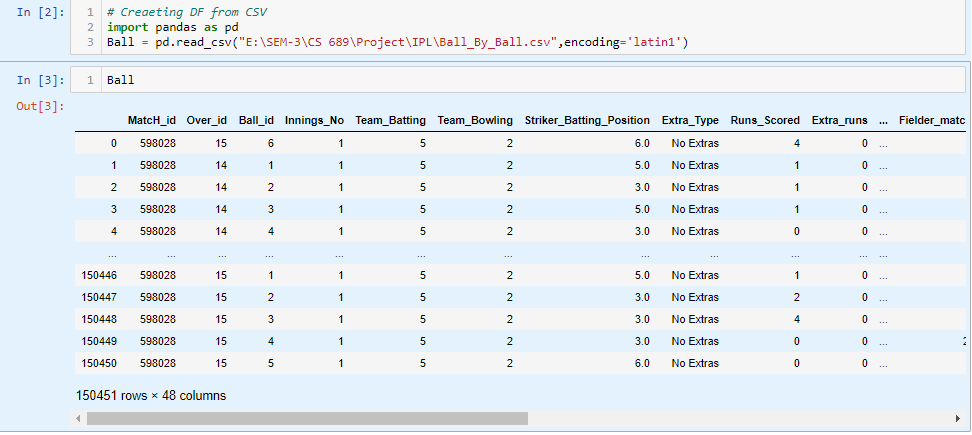


# ETL

1. Extract:

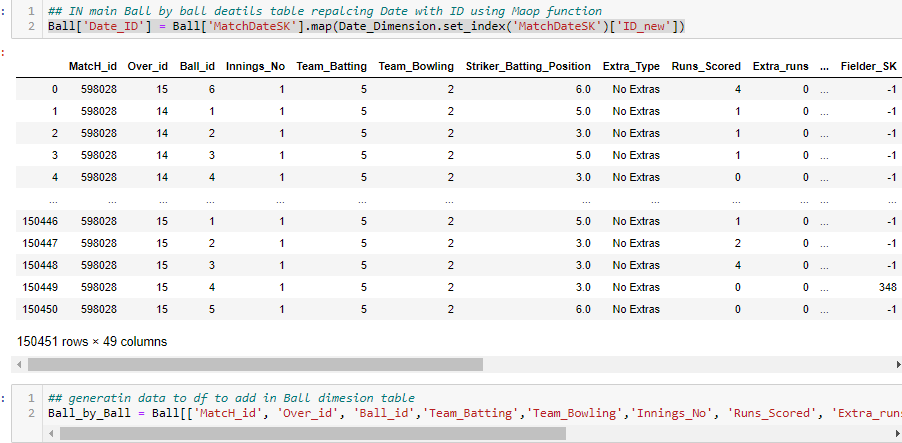
I have load data to jupyter using python pandas command as shown below:

I have used Pd.read\_csv function to convert CSV to data frame.



1. Transform:

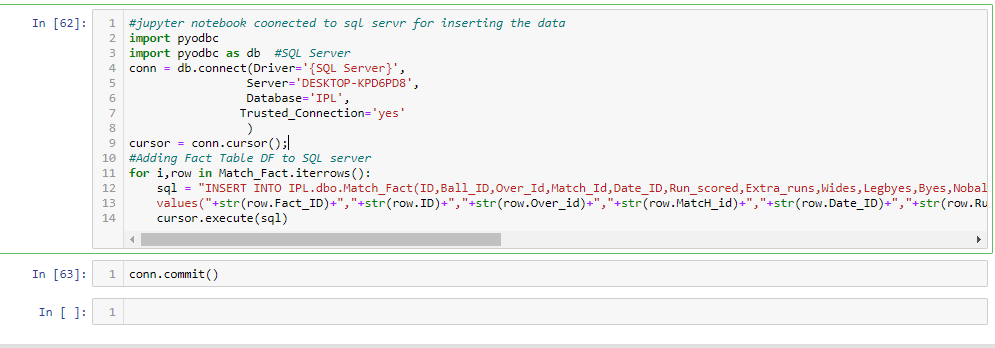
To work on data in DF I have used again python to remove the supplicates value , Remove NaN, Replace some string with int





1. Load:

To Load the data to database I have use pandas library pyodbc as shown below:



1. Why this method

The problem with native database scripts is it is not heterogeneous; you can't handle XML, flat files (not easily), different Databases, etc. It's much easier to handle ETL within the same kind of database (oracle to oracle, sql server to sql server). Managing the scripts become very difficult, managing failure path is not easy that’s why I feel python is easy to write a code and represent it

So, I have decided to work with python to load my data to database.

# References

[1] <https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-home.htm>

[2] <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to_excel.html>

[3] The Data Warehouse Toolkit(3rd edition) – Ralph Kimball Margy Ros

[4] https://www.kaggle.com/raghu07/ipl-data-till-2017?select=Match.csv

# 