CS322:Big Data

Final Class Project Report

Project (FPL Analytics / YACS coding): **FPL Analytics** **Date: 4th December 2020**

SNo	Name	SRN	Class/Section
1	Sneha Jayaraman	PES1201802825	5J
2	Rahil N Modi	PES1201802826	5J
3	Sooryanath I. T	PES1201802827	5J
4	Himanshu Jain	PES1201802828	5J

Introduction

Fantasy Premier League (FPL) Analytics aims at processing and analyzing the live events of a football match. Streaming Spark and Spark MLlib provide the framework to achieve this and generate useful insights from the data.

Related work

Documentation of PySpark and Spark MLlib was referred. Setup of the environment to run streaming events was done.

Design

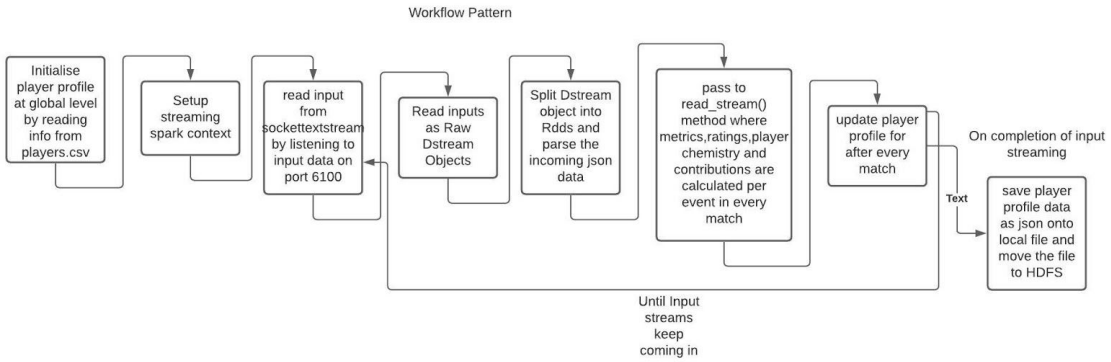
The match and events data are streamed continuously and are sent to the main program via a TCP socket. This stream of data comes in as a DStream object and is processed for every rdd using readstream (user-defined) method. Filtering of events is done based on the requirements of the problem.

Various metrics are computed for every match and stored in a local dictionary. The players ratings are then calculated. Chemistry of the players was computed by using a data structure of the dictionary where the keys are the two-pair tuple of player ID and values are the chemistry coefficient. Player Profile is also built during the streaming of data.

Users are more concerned regarding the prediction of the winning team. This has been implemented by using Spark MLlib (for Regression and Clustering) to find the team's strength and hence find their winning chances.

Prediction of winning chance, Player's Profile and Match Info are provided as output based on the query given as a JSON input

The following block diagram illustrates the our workflow of the project



Results

The winning chance of a team is predicted based on the team's strength. A record of the players profile is maintained as the matches progress and also the match info is maintained.

Unexpected values for the chemistry coefficient were observed during the training of the model. These were rectified by fine-tuning the parameters

The parameters for clustering and regression are defined. In K-Means clustering, we define k = 5 as the number of clusters to group the players. This is done to get the ratings of the player who has played less than five matches. Quadratic regression predicts the rating of a player based on the age.

Problems

Initial setup of streaming spark lead to various configuration issues such as unsupported versions, absence of required libraries.

Google results helped us in solving the problems faced and debugging suggestions and code corrections were done by StackOverflow results

Various other blogs and documentations were followed to complete the project

Conclusion

Working with streaming spark has now become easier. Performing the setup and handling the streaming data was a lesson learnt. Good insights can be drawn from the data using Spark MLlib. A major skill of debugging the code is also developed. The complexities involved in the real-time analysis of the data is understood and can be used to explore and solve various other problem statements

EVALUATIONS:

SNo	Name	SRN	Contribution (Individual)
1	Sneha Jayaraman	PES1201802825	Chemistry Calculation, Preprocessing, Clustering
2	Rahil N Modi	PES1201802826	Two Event Metrics, Player Profile, Match Details
3	Sooryanath I. T	PES1201802827	Setup, Streaming Input, Two Event Metrics, UI Tasks
4	Himanshu Jain	PES1201802828	Two Event Metrics, Winning Chances, Player Rating

(Leave this for the faculty)

Date	Evaluator	Comments	Score

CHECKLIST:

SNo	Item	Status
1.	Source code documented	
2.	Source code uploaded to GitHub – (access link for the same, to be added in status)	
3.	Instructions for building and running the code. Your code must be usable out of the box.	