Big Data Analysis of Indian Premier League using Hadoop and MapReduce

Rajdeep Paul
Department of Computer Science & Engineering
National Institute of Technology, Tiruchirappalli
Tamil Nadu, India
rdp.rajdeeppaul@gmail.com

Abstract - **The exponential growth of the internet has changed the way people used to interact with each other. With the extensive use of social media platforms like Facebook, Instagram, Google+, Twitter etc., people nowadays share their emotions, opinions, and reactions quite actively. Twitter, for example, being one of the prime sources of real-time information, it's being used by millions of netizens all around the globe. Twitterati ardently share their political views, movies' experience as well as react to numerous sporting events on a daily basis. This paper analyzes one such sporting event, the final match of Indian Premier League 2015, using Hadoop and MapReduce programming paradigm. The objective of this paper is to assess the popularity of Indian Premier League 2015 as well as various metrics related to the game. The analysis found that Indian Premier League is not only popular in India but it has gained immense popularity throughout the world, ranging from Ahmedabad (India) to Zurich (Switzerland). Furthermore, the time interval in which cricket fans tweeted the most, most talked about players, and the dominant team – are the various metrics that have also been analyzed from the game's perspective.**

Keywords – Big Data, Flume, Hadoop, Indian Premier League, MapReduce, Social Networking Sites.

I. INTRODUCTION

The internet has changed the way people used to share data or information with each other. In the late 90s and early 2000s, people used to message, or rather email each other to communicate and exchange information. But now social media has evolved and expanded beyond anyone's expectation and imagination. Nowadays people extensively share their achievements, opinions and reactions on Facebook, Google+, Twitter, Weibo, VK and several other social networking platforms. Social media has become an integral part of everyone's day-to-day life. Several social networking websites like Facebook, Google+ and microblogging websites like Twitter, Tumblr have become quite popular among netizens. Twitter, for example, has gained immense popularity worldwide and probably the most popular microblogging website available today with over 310 million monthly active users as of March 2016 [1].

Twitter is also a primary source of real-time information. Twitterati share their daily activities, opinions, reactions in short text messages, known as tweets, on Twitter which can influence people on a large scale worldwide in a short span of time. Whenever something happens around the world, people start discussing about it on Twitter. The use of hashtags in tweets make it easy for people who want to search for certain events, places or things. Several twitter users often indulge in conversation by posting tweets, retweeting other tweets, which in turn identify the top trends in the Twitter world.

With this heavy use of Twitter, people around the world use this platform to share opinions about movies, analyze stock markets, study the political inclination of voters and react to sporting events. Moviegoers use Twitter to see the trending movies and others' opinions about those movies which help them in deciding which movie they're going to watch next, whereas movie analysts use Twitter to conduct polls and analyze tweets which help them in rating the movies accordingly. On the other hand, Twitter helps stock brokers in keeping an eye on the flow of stocks, and investors use Twitter to decide which companies it's safe and fruitful to invest on. Furthermore, political parties conduct surveys on Twitter to assess the political inclination of voters which help them in outlining future course of action.

Sports being in the core of everyone's life, sports fans are ardent for news and Twitter gives them a way to reach out to real-time information. Sports fans arduously react to every aspect of the game whether its pre-match discussion or post-match postmortem, and actively use Twitter to share emotional responses in the form of tweets. Their emotions like facial expressions, gestures, body and eye movements change during games, which also reflect in their tweets. Twitter captures these emotions worldwide in every sporting event, irrespective of whether fans are watching the game live, on TV or on any other handheld devices.

In this regard, this paper narrowed down the present investigation to a specific domain. The motivation behind this paper was to examine the popularity of Indian Premier League throughout the world and study cricket fans' behavior on Twitter through big data analysis using Hadoop framework and MapReduce programming paradigm [2, 3, 4]. This paper outlined the related past research works, the present research work followed by results and analysis.

978-1-5090-5595-1/17/$31.00 ©2017 IEEE

II. Related Work

Social networking and microblogging websites have become quite popular among online users. Websites like Facebook, Google+, Twitter, Weibo are nowadays the primary means of communication and networking tools. As stated earlier, Twitter is the most prominent microblogging platform available today. People use Twitter for a variety of purposes, like for example, for mining opinions of moviegoers, for assessing the political inclination of voters and for getting up-to-date news about sporting events. Several research works have been carried out on this domain over the past couple of years [5-10]. Few key research works are as follows.

Gokulakrishnan et al., [11] pointed out that Twitter widely captures opinions and sentiments. They classified the sentiments in three broad categories, such as positive, negative and neutral, and analyzed the performance of several classifying algorithms by incorporating such emotions. They also pointed that these classifications are significantly different from classifications that are based on detailed and structured messages. Kumar et al., [12] took this even further and suggested that opinions can further be categorized based on five distinct emotions, such as Disgust, Sadness, Fear, Anger and Happiness, which are fairly defined in human psychology. In their experiment, they proposed a two-step process. First, the opinions words are extracted for each tweet, and then emotion values are calculated for each opinion word using a novel approach. Even though this mechanism puts straight forward tweets in the appropriate emotion buckets, it's quite difficult to assess the emotion values of sarcastic tweets. Nonetheless, this is a promising technique which might help in making decisions in business and even in government policy.

As discussed earlier, extracting the correct emotions in sarcastic tweets is quite difficult, but few researchers in [13] disagreed with this philosophy and believed that sarcastic tweets can be used to enhance the accuracy of sentiment analysis. They proposed a method where a set of textual and non-textual features are extracted from tweets, which helps in classifying the tweets irrespective of their topic.

Apart from analyzing the general sentiments of online users, several research works have also been done on key specific domains. Selvan et al., [14] studied how companies and organizations use Twitter to analyze customers' implicit feedbacks about their products and services. They used a dictionary based approach where each tweet is tokenized and compared against a set of predefined words. Based on the comparison the tokens' weights are aggregated, which subsequently helps in classifying the overall sentiment of the tweet.

Hodeghatta [15] studied the general behavior of moviegoers. He pointed out that Twitter actually reflects human behavior when it comes to movies. People living in different regions express different emotions depending on the nature of the movie. This proves that movies substantially affect the cultural sentiments of people.

There are several other research works available cognate to these key domains. However, this paper focuses on examining the recognition of Indian Premier League and understanding the behavior of cricket fans.

III. Proposed Work

On 24th May 2015, during the final game of Indian Premier League 2015, the Mumbai Indians played against the Chennai Super Kings at the Eden Gardens stadium in Kolkata. A total of 8 teams played in the tournament but only these two teams shined through the group and playoff stage and reached the final. On the final day of the league, Mumbai Indians beat Chennai Super Kings by 41 runs. The emotional responses of Chennai Super Kings' fans present in the stadium were broadcasted live. Their anticipation early in the match turned into surprise and to despair after Mumbai Indians won the match. Contrary to Mumbai Indians' ecstasy, the sadness and dejection of many Chennai Super Kings' fans can still be remembered by numerous people who watched the match.

The present work adopted a natural methodology to examine the popularity of Indian Premier League throughout the world and how cricket fans reacted before, during and after the final game when the Chennai Super Kings competed against the Mumbai Indians. In this project the analysis was limited to tweets in the English language originated from all across the globe instead of using all the tweets that also includes regional language. The present investigation answered the following four broad aspects:

- What is the popularity of Indian Premier League throughout the globe?
- In which time interval cricket fans tweeted the most?
- Who are the most talked about players during the finale of Indian Premier League 2015?
- Which team is more dominant in the Twitter world?

Due to its broad nature the present investigation was categorized into the following two parts.

1. Data Collection Procedure:

Data/tweets were collected from twitter.com using Twitter streaming API and Flume agent during the final match of Indian Premier League 2015, and single node Hadoop cluster was used to store the tweets. Twitter streaming API gives developers access to Twitter's global stream of real-time tweet data, and Flume agent on the other hand efficiently collects, aggregates and moves enormous amounts of data from source to sink. Apache's Flume agent was chosen to collect real-time English tweets using a customized configuration file. The configuration file contains a list of predefined keywords like "*Indian Premier League*", "*Chennai Super Kings*", "*Mumbai Indians*", ignoring the case sensitivity in tweets. Single node Hadoop cluster was set up for storage and big data analysis. Fig. 1 shows the workflow of data collection process.

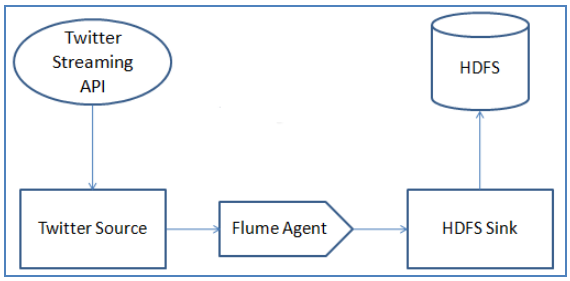

Fig. 1. Workflow of data collection process

As a steady flow of tweet streams comes from twitter.com, Flume agent analyzes and retrieves the relevant tweets, as explained in Algorithm 1.

With this process, approximately 4 lakh tweets were collected in three days(23rd, 24th, and 25th May). For each tweet, several key features were parsed such as user's ID, screen name, post time, retweet count as well as the content of the tweet. Furthermore, location and time zone data were also collected for each tweet, if available. During this data collection process, we ignored all the tweets in regional language which might affect our analysis on tweets' patterns and fans' behavior.

2. Data Analysis and Coding Procedure:

The current project analyzed each of the four metrics in the following way:

i. Designed and implemented MapReduce programs, that is, Driver class, Mapper class and Reducer class, packaged up the code as jar file and exported the package.
ii. Executed the jar package on the Hadoop cluster with relevant input.
iii. The output file of the MapReduce program was dumped in Hadoop cluster.

A more detailed description of each of these MapReduce programs is as follows:

Input: tweet stream

Output: a file with selected tweets in json format

Step 1: A steady flow of tweet streams comes to Flume agent.

Step 2: Flume agent compares individual tweet's textual data with the predefined keywords. If a match is found go to step 3, otherwise go to step 4.

Step 3: Flume agent encapsulates the tweet's textual and metadata as JSON format in a file and dumps it in Hadoop cluster.

Step 4: Tweet is simply ignored and the next tweet in the pipeline is fed to Flume agent.

Algorithm 1: Data collection process

Each of the MapReduce programs had 3 components, Main class or Driver class, Mapper class and Reducer class. The Driver class declared only the configuration settings for the MapReduce program, it didn't have any complex business logic. The Mapper class had the map function and the Reducer class had the *reduce* function. An internal *shuffle and sort* phase was used to send data from the Mapper class to Reducer class. The design and implementation of each of these classes and phase are explained below.

a) Design and Implementation of Driver Class:

The Driver class is the class where the execution of program begins and is almost same for all the MapReduce programs. It sets configuration values for the Main or Driver, Mapper and Reducer classes. The exact steps followed by the Driver class are explained in algorithm 2.

b) Design and Implementation of Mapper Class:

The Mapper classes were different for each MapReduce program, which is explained in algorithm 3, 4, 5 and 6.

c) Shuffle and Sort Phase

Shuffle and sort phase was performed internally by the Hadoop framework. <key,value> pairs from the Mapper class was grouped by the key and sent to the Reducer class. For example, the <key,value> pairs *<CHENNAI SUPER KINGS,1>* and *<CHENNAI SUPER KINGS,1>* is sent as *<CHENNAI SUPER KINGS,[1,1]>* to the Reducer class.

d) Design and Implementation of Reducer Class

The Reducer class was used to aggregate the values sent from *shuffle and sort* phase, and is quite same for all the MapReduce programs. The exact steps followed by the Reducer class are explained in algorithm 7.

IV. Results and Discussion

As stated earlier, approximately 4 lakh tweets were collected in the final three days(23rd, 24th and 25th May) of Indian Premier League 2015, and were used to analyze tweets' patterns and fans' behavior. Based on the analysis it is evident that, since its inception in 2008 Indian Premier League has gained immense popularity, and that too in such short span of time. The league is not only popular in India but it has gained immense fan-following from all over the world. The experimental result pointed out that, from Ahmedabad (India) to Zurich (Switzerland), cricket fans from all across the globe were engaged with the game, and were heavily active on Twitter, as depicted with colored dots in fig 2.

Though online users were significantly active on Twitter during the final three days of the tournament, their activities were not consistent with time. A zigzag pattern in the number of tweets was observed over the period of three days, as depicted in fig. 3. From 24th May 7:45 P.M. to 25th May 12:45 A.M.(Indian Standard Time), the maximum number of tweets were observed with 1,36,776 tweets, whereas on 23rd May, between 8:45 A.M. to 01:45 P.M.(Indian Standard Time), a bare minimum of 409 tweets were received. Furthermore, a sudden surge in the number of tweets on 24th May, between 9:45 A.M. to 2:45 P.M.(Indian Standard Time) was also observed where a total of 1,07,163 tweets were received. Moreover, cricket fans used Twitter more on 24th May i.e. on the day of the match than on 23rd May(previous day) or 25th

May(next day), and as the match progressed towards the end, fans expressed their emotions more on Twitter.

The analysis also showed one unusual result. Even though Rohit Sharma led the team to victory and Mumbai Indians won the match by 41 runs, he was not the most talked about player. Surprisingly, the captain of Chennai Super Kings, Mahendra Singh Dhoni, was the most talked about player, followed by Rohit Sharma and Suresh Raina. Rohit Sharma received over 6 thousand targeted tweets, whereas Mahendra Singh Dhoni received a total of 12,019 targeted tweets, nearly twice as much as Rohit Sharma. This was probably due to the fact that as Chennai Super Kings lost the match, many fans tweeted in support of him, many expressed their anger and disappointment in the form of tweets and few others mocked him in their sarcastic tweets, but nonetheless he was, incomparably, the most dominant player in the Twitter world. Apart from Mahendra Singh Dhoni, several Caribbean players like Kieron Pollard, Dwayne Smith, Dwayne Bravo and Lendl Simmons were also popular among the cricket fans. The players' analysis is depicted in fig. 4.

Input: input *jar* file and output file name

Output: status of MapReduce program

Step 1: The Driver class first checks the number of argument passed from the command-line. If the number of arguments is less than two or greater than two go to step 2, otherwise go to step 3.

Step 2: The programs throws a fatal error and returns error status.

Step 3: All the configuration settings are extracted from the configuration file, which includes the number of mappers and reducers to run and are applied to the MapReduce program.

Step 4: A job name is declared and input and output file names are set for the MapReduce program.

Step5: The datatypes for the output <key,value> pairs for the Mapper and Reducer class are set.

Step 6: The Driver class submits the job on the Hadoop cluster to run and returns success status from the program.

Algorithm 2: Driver class

Finally, the experimental result also proved one general hypothesis where cricket fans talks more about the winning team than the losing team. During the final three days of the league, Mumbai Indians undoubtedly dominated the Twitter world receiving a total of 1,87,647 targeted tweets, whereas Chennai Super Kings received a total of 47,881 targeted tweets, almost one-fourth as much as Mumbai Indians. And fig. 5, obtained from table 1, also proved this fact and showed that a whopping 79.67% crickets fans were talking about Mumbai Indians whereas only 20.33% fans were talking about Chennai Super Kings.

V. CONCLUSION

This paper examined the popularity of Indian Premier League throughout the globe and assessed the behavior of cricket fans i.e. how they reacted before, during and after the match. Approximately 4 lakh tweets were collected and analyzed using Hadoop cluster and MapReduce Programming paradigm.

From the experimental results it is clear that Indian Premier League is exceptionally popular throughout the world. Moreover, instead of general sentiments of tweets, the overall behavior of cricket fans was analyzed. The results showed that cricket fans don't use Twitter consistently, which also reflects in their number of tweets. During the final game of the league, cricket fans tweeted the most during and just after the match. Furthermore, although Chennai Super Kings lost the match to Mumbai Indians, their captain, Mahendra Singh Dhoni, was the most talked about player. However, the results pellucidly showed that the Mumbai Indians fairly dominated the Twitter world.

Input: a JSON file with tweets

Output: a <key,value> pair

Step 1: The file is read line by line and each line is parsed into tokens based on the *time_zone* delimiter.

Step 2: The tokens are passed through the regular expression i.e. *([A-Za-z|\s,&()]{0,}).** for pattern matching.

Step 3: If a match is found go to step 4, otherwise go to step 5.

Step 4: Extract the relevant match, if available, from the token and the corresponding <time zone,value> pair, for example <*Greenland,1*> pair is sent to *shuffle and sort* phase.

Step 5: The token is simply ignored and next line in the pipeline is fed to the Mapper class.

Algorithm 3: Mapper class(analyzing the popularity of Indian Premier League)

This paper analyzed raw Twitter data collected from all across the globe but it is scalable to any social media platform. Data from other sources can also be incorporated into the existing system which can further improve the analysis.

Input: a JSON file with tweets

Output: a <key,value> pair

Step 1: A predefined set of 5-hour intervals are set for grouping the tweets in the appropriate time interval buckets.

Step 2: The file is read line by line and each line is parsed into tokens based on the *created_at* delimiter.
Step 3: The tokens are passed through the regular expression i.e. *([A-Za-z0-9|\s+:]{0,}).** for pattern matching.

Step 4: If a match is found go to step 5, otherwise go to step 7.

Step 5: Extract the relevant match from the token and compare it with the set of time intervals.

Step 6: The matched <time interval,value> pair, for example <*SUN MAY 24 09:45:00 IST 2015-SUN MAY 24 14:45:00 IST 2015,1*> pair is sent to *shuffle and sort* phase.

Step 7: The token is simply ignored and next line in the pipeline is fed to the Mapper class.

Algorithm 4: Mapper class(analyzing the tweets in time interval)

Input: a JSON file with tweets

Output: a <key,value> pair

Step 1: A predefined set of hashtags, twitter handle, player name and few other probable keywords are declared for each player of Mumbai Indians and Chennai Super Kings.

Step 2: The file is read line by line and each line is parsed into tokens based on the *text* delimiter.

Step 3: The tokens are passed through the regular expression i.e. *([^|"]*).** for pattern matching.

Step 4: If a match is found go to step 5, otherwise go to step 7.

Step 5: Extract the relevant textual part of tweet data from the token.

Step 6: The textual part of tweet data is compared with the predefined sets. If a match is found, the corresponding <player name,value> pair, for example <*MAHENDRA SINGH DHONI,1*> pair is sent to *shuffle and sort* phase, otherwise the whole tweet data is ignored.

Step 7: The token is simply ignored and next line in the pipeline is fed to the Mapper class.

Algorithm 5: Mapper class(analyzing tweets to find out who are the most talked about players)

Input: a JSON file with tweets

Output: a <key,value> pair

Step 1: A predefined set of hashtags, twitter handle, team name and few other probable keywords are declared for each team.

Step 2: The file is read line by line and each line is parsed into tokens based on the *text* delimiter.

Step 3: The tokens are passed through the regular expression i.e *([^|"]*).** for pattern matching.

Step 4: If a match is found go to step 5, otherwise go to step 7.

Step 5: Extract the relevant textual part of tweet data from the token.

Step 6: The textual part of tweet data is compared with the predefined sets. If a match is found, the corresponding <team name,value> pair, for example <*CHENNAI SUPER KINGS,1*> pair is sent to *shuffle and sort* phase, otherwise the whole tweet data is ignored.

Step 7: The token is simply ignored and next line in the pipeline is fed to the Mapper class.

Algorithm 6: Mapper class(analyzing tweets to find out which team is more dominant in the Twitter world)

Input: a <key, [value list]> pair

Output: a text file with <key,value> pair

Step 1: The Reducer class counts the number of times a value appears in the value list and generates a new <key,value> pair, for example <*CHENNAI SUPER KINGS,2*>.

Step 2: The generated <key,value> pair is encapsulated in a text file and is sent to the Hadoop cluster for storage.

Algorithm 7: Reducer class

Fig. 2. World map showing the popularity of Indian Premier League

Fig. 3. Tweets in time interval

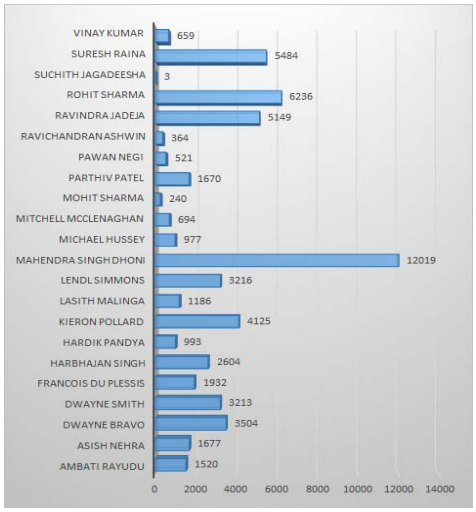

Fig. 4. Players' analysis

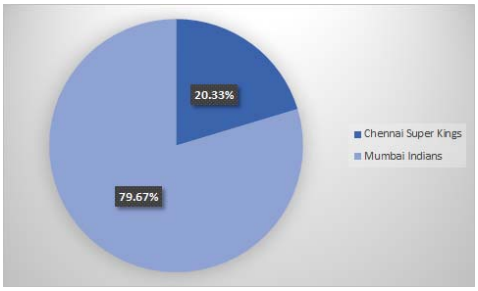

Fig. 5. Team analysis

REFERENCES

[1] https://en.wikipedia.org/wiki/Twitter

[2] S. G. Manikandan and S. Ravi, "Big Data Analysis Using Apache Hadoop", IT Convergence and Security (ICITCS), 2014 International Conference, pp. 1-4, October 2014.

[3] A. Saldhi, D. Yadav, D. Saksena, A. Goel, A. Saldhi and S. Indu, "Big data analysis using Hadoop cluster", Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference, pp. 1-6, December 2014.

[4] S. B. Elagib, A. R. Najeeb, A. H. Hashim and R. F. Olanrewaju, "Big Data Analysis Solutions Using MapReduce Framework", Computer and Communication Engineering (ICCCE), 2014 International Conference, pp. 127-130, September 2014.

[5] I. Guellil and K. Boukhalfa, "Social big data mining: A survey focused on opinion mining and sentiments analysis", Programming and Systems (ISPS), 2015 12th International Symposium, pp. 1-10, April 2015.

[6] A. Molla, Y. Biadgie and K. Sohn, "Network-Based Visualization of Opinion Mining and Sentiment Analysis on Twitter", IT Convergence and Security (ICITCS), 2014 International Conference, pp. 1-4, October 2014.

[7] L. Lin, J. Li, R. Zhang, W. Yu and C. Sun, "Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-Aware Approach", Utility and Cloud Computing (UCC), 2014 IEEE/ACM 7th International Conference, pp. 890-895, December 2014.

[8] K. Jedrzejewski and M. Morzy, "Opinion Mining and Social Networks: A Promising Match", Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference, pp. 599-604, July 2011.

[9] B. Singh, S. Kushwah, S. Das and P. Johri, "Issue and challenges of online user generated reviews across social media and e-commerce website", Computing, Communication & Automation (ICCCA), 2015 International Conference, pp. 818-822, May 2015.

[10] I. Khozyainov, E. Pyshkin and V. Klyuev, "Spelling out opinions: Difficult cases of sentiment analysis", Awareness Science and Technology and Ubi-Media Computing (iCAST-UMEDIA), 2013 International Joint Conference, pp. 231-237, November 2013.

[11] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath and A. Perera, "Opinion mining and sentiment analysis on a Twitter data stream", Advances in ICT for Emerging Regions (ICTer), 2012 International Conference, pp. 182-188, December 2012.

[12] A. Kumar, P. Dogra and V. Dabas, "Emotion analysis of Twitter using opinion mining", Contemporary Computing (IC3), 2015 Eighth International Conference, pp. 285-290, August 2015.

[13] M. Bouazizi and T. Ohtsuki, "Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis", Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference, pp. 1594-1597, August 2015.

[14] L. G. S. Selvan and T. Moh, "A framework for fast-feedback opinion mining on Twitter data streams", Collaboration Technologies and Systems (CTS), 2015 International Conference, pp. 314-318, June 2015.

[15] U. R. Hodeghatta, "Sentiment analysis of Hollywood movies on Twitter", Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference, pp. 1401-1404, August 2013.