2015

BIG DATA
IPL TWITTER ANALYSIS
PROJECT REPORT

SUBMITTED BY:

GAURANG PARMAR | DHRUMIL JOSHI

UNIVERSITY OF ILLINOIS AT CHICAGO

TABLE OF CONTENTS

1 SUMMARY

Twitter has become a leading social media platform where registered users can share messages with the world (These messages are known as tweets). These tweets are used by individuals to express their thoughts and feelings. Our aim for this project was to study extracting live data from Twitter and further identify Twitter reactions to specific events. We took the cricket extravaganza – IPL (Indian Premier League) as our topic of interest and further investigated tweet dynamics (count, popularity, retweets etc.) for various parameters that make the game. These parameters include team battles (indicating team popularity on Twitter), team trends (how the team tweet dynamics respond to specific events in match), popularity of players and specific hashtags used over the match. We achieved the results using Hadoop distributed programming and using the MapReduce model. The results were presented graphically for better understanding and easier co-relation.

2 PROJECT OBJECTIVES

Social networking sites such as Twitter, Facebook etc. nowadays are contributing a lot towards big data. In order to find the interesting patterns or trends from this huge data, data analyst / engineers need to clean, integrate, aggregate and further analyze the data to produce meaningful insights.

The core objective of the project is to understand the components and core technologies related to content retrieval, storage and data intensive analysis of large corpus of data collected over a specific period of time.

The topic of interest is IPL (Indian Premier League), which is an Indian Twenty20 cricket tournament. The IPL is the most-watched Twenty20 league in the world and is also known for its commercial success. In 2010, the IPL became the first sporting event to be broadcast live on YouTube. The brand value of the 2014 Indian Premier League was estimated to be around US$7.2 billion.

3 PROJECT SCOPE

The purpose of this project is to find out trends by aggregating the data in social networking site such as Twitter. We are specifically looking at data related to IPL (Indian Premier League) and further investing on the data collected during the 2015 edition of the game to derive trends / popularity as well as the scale of this sporting event.

Additionally, the project takes into account 5 matches, played between the top 5 teams of the tournament and provides trends on the matches / teams over a duration of one week.

Teams:

- Chennai Super Kings (CSK)
- Mumbai Indians (MI)
- Royal Challengers Bangalore (RCB)
- Kings XI Punjab (KXIP)
- Kolkata Knight Riders (KKR)

Matches:

- CSK v/s RCB
- MI v/s KKR

- MI v/s KXIP
- KKR v/s RCB
- MI v/s CSK

4 PROJECT DELIVERABLES

The below documentation and deliverables were / will be delivered over the course of the project.

- **Project Proposal Document**
  A project proposal document was submitted to gain approval on the project topic as well as the scope of the project. It included the project topic, description of the topic, features of the application and data source.

- **Project Report**
  The project report is delivered at the end of the project timeline and entails the objective, scope, deliverables and the project approach over the course of the project. Further it includes, data analysis and visualizations created on the collected data. It also details the technologies used and the any appendix indicating related words / phrases and other materials.

- **Project Code**
  The project source code is delivered with any ReadMe documents, if required.

5 PROJECT APPROACH

- **Content Retrieval:** The large amount of data is collected using Topsy Twitter API.

- **Data Processing:** Data collected over a period of time is processed by using parallel and distributed processing software framework developed by Apache Hadoop and using map reduce programming model.

- **Storage:** This data is stored in a certain format so as to form key value pair which is needed to feed to mapper in map-reduce programming approach. The data is stored in Hadoop Distributed File System.

- **Data Analysis:** The output obtained from reducer phase is further analyzed using by developing custom programs that provide various analysis such as word count, stopwords removal, keywords extraction etc.

- **Visualization:** Various ongoing trends on social networking sites are aesthetically represented using different graphs / tables in MS Excel.

6 DATA

Data Description:

Data is collected and aggregated from Topsy Twitter API. Topsy is a social media analytics website which specializes in analyzing global chatter on Twitter. This data is in the form of tweets where each tweet is 140 characters long. The language used within the tweets can be very informal with incorrect grammar, repetitive text and special symbols. To extract relevant information and generate underlying insights / trends pertaining to those tweets was a challenging task. The following are the major steps in the data retrieval and pre-processing:

- Tweets for the current trending topic "IPL" are collected using filter query in Topsy Twitter API.
- Data is collected for different ranges of dates (hour range, day-range, week-range) using specific start-time and end-time.
- Aggregated raw data is cleansed to some extend before analyzing it using custom MapReduce and Java programs.
  - Removal of hashtags
  - Removal of special symbols
  - Removal of stopwords
  - Removal of punctuations
- The final input file for analysis has each tweet separated by new line so that the processing in MapReduce is easy.

Sample Data:

```
IPL 8: MI face tough challenge against formidable CSK http://t.co/Yv9qxhSQTf http://t.co/SmXCWhaU0w #HottParineeti
"Mumbai Indians vs Chennai Super Kings Live Score: IPL 8, Match 12 http://t.co/WDjudRAy5j #IPL8"
"RT @ESPNcricinfo: Numbers Game: Since 2008, Brendon McCullum has among the highest strike-rates in ODIs, but among the lowest in IPL http:/â€¦"
@SonySIX im@chirusanju a very big match between champions csk made all wins in this ipl and mi loose all matches can mum able to put a win
@SkyCricket Can u pls keep channel 1 or 3/4 for UK cricket Dart Golf tennis football .. PLS keep IPL on CH 2...uptil now @ITV4 was great fun
IPL 8: MI face tough challenge against formidable CSK http://t.co/waH3ZyLeqa http://t.co/I28cUkS1E6 #GossipLoveLady
IPL 8: MI face tough challenge against formidable CSK http://t.co/XQ2vYutkS3 http://t.co/5RZqEJp0Xy #HottSonam
IPL 8: MI face tough challenge against formidable CSK http://t.co/z9i141BmjB http://t.co/h3I6mGqp8l #HottRanbirK
IPL 8: MI face tough challenge against formidable CSK http://t.co/OJ8ZrQti0l http://t.co/Gb2LV11zjE #HottVarunDhawan
"RT @spice_mobiles: #Contest #SPL #IPL2015 #IPL8 @Contest_in @ContestsInIndia @Contest_Alert @ContestBoard @Contest_Hub IPL tickets, Smartphâ€¦"
IPL 8: MI face tough challenge against formidable CSK http://t.co/UINv7hzAN7 http://t.co/tikeOck9V0 #Hot_Girl4Love
RT @Iamankurgoyal: #FollowMe #FF #RT IPL: Focus on Yuvi as stars go under hammer: The biggest concern for the... http://t.co/Ktm7JRdoFZ #Foâ€¦
IPL 8: MI face tough challenge against formidable CSK http://t.co/WOVinwG2Hx http://t.co/ijxqhpvqb9 #SexxySunnyLeone
IPL 8: MI face tough challenge against formidable CSK http://t.co/o1jShtKMLd http://t.co/lzxUK6zfhp #HottAliaBhatt
RT @Iamankurgoyal: #FollowMe #FF #RT IPL: Focus on Yuvi as stars go under hammer: The biggest concern for the... http://t.co/Ktm7JRdoFZ #Foâ€¦
IPL 8: MI face tough challenge against formidable CSK http://t.co/IAr4i7txAw http://t.co/Jk0y9IwWz1 #HottYamiGautam
IPL 8: MI face tough challenge against formidable CSK http://t.co/C9BdErN7be http://t.co/UBxXHQw1cV #FunUnlimited420
```

Amount of Data collected and details:

- Data is collected for different ranges of dates / times.
- Data is collected over the period from April 8th, 2015 to April 15th, 2015.
- This includes daily match data aggregated as follows :
  - Pre-match hourly data ["Match start time –1" to "Match start"]
  - "Match start time" – "Match start time +1"
  - "Match start time +1" – "Match start time +2"
  - "Match start time +2" – "Match start time+3"
  - "Match start time +3" – "Match start time+4"
  - Post-match hourly data ["Match start time+4" – "Match end time +0:30"]

Note: Total duration of match, approximately = 3 hours 30 mins

- Additionally, twitter chatter over the same week is collected cumulatively as a separate file to perform overall analysis of the entire week.

7 IMPLEMENTATION & ANALYSIS

The data collected for analysis consists of 5 matches. These matches were further analyzed using different time intervals noted below:

1 – Pre-match hourly data ["Match start time –1" to "Match start"]
2 – "Match start time" – "Match start time +1"
3 – "Match start time +1" – "Match start time +2"
4 – "Match start time +3" – "Match start time+4"
5 – Post-match hourly data ["Match start time+4" – "Match end time +0:30"]

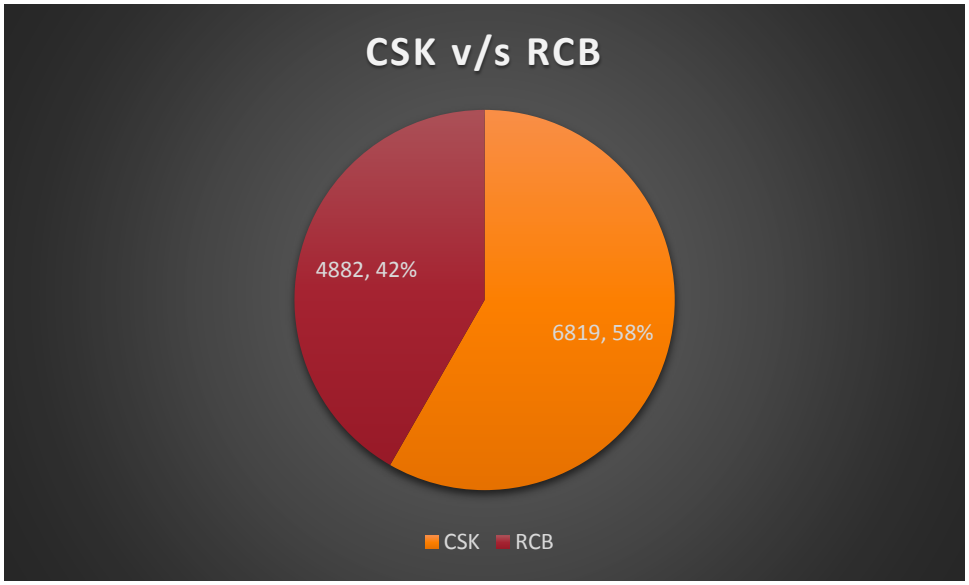
The mentions and hashtags for each of the playing teams was gathered over the duration of the time interval using relevant keywords and following analysis was done:

- **Team Twitter battle:**
  Signifies the overall (cumulative) twitter chatter for both the teams over the entire duration of the match. This gives us an insight as to which team was more popular amongst the tweeters and had most game changing moments (toss, wickets, catches, runs, win, etc.) across the duration of the match.

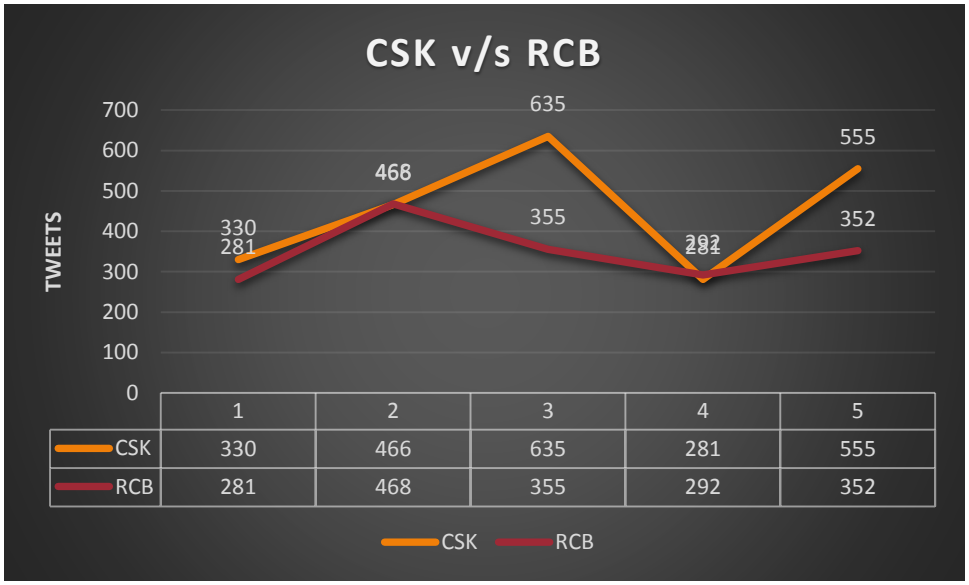
- **Team Trend analysis:**
  The team trend analysis gives an interval-by-interval graphical representation of which team was in talk amongst the tweeters and how the conversations about the playing teams shaped in regards to specific events within the match.

Match 1: CSK v/s RCB

- **Team Twitter battle (based on Total number of relevant tweets)**



- **Team Trend analysis**

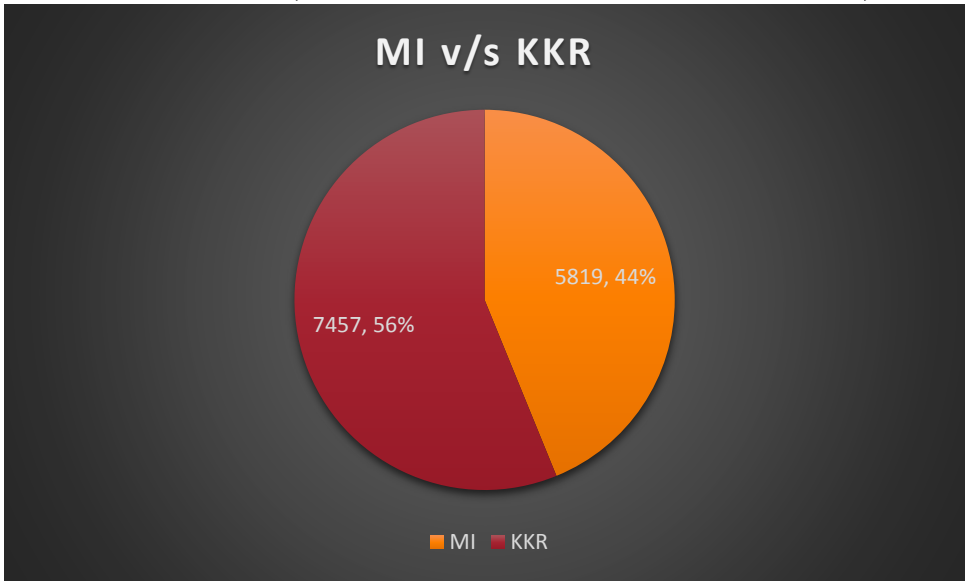


Time Interval	Event	Trending Team
1	CSK won the toss	CSK
2	CSK batting: RCB takes early wickets	RCB
3	Match is evenly poised	RCB

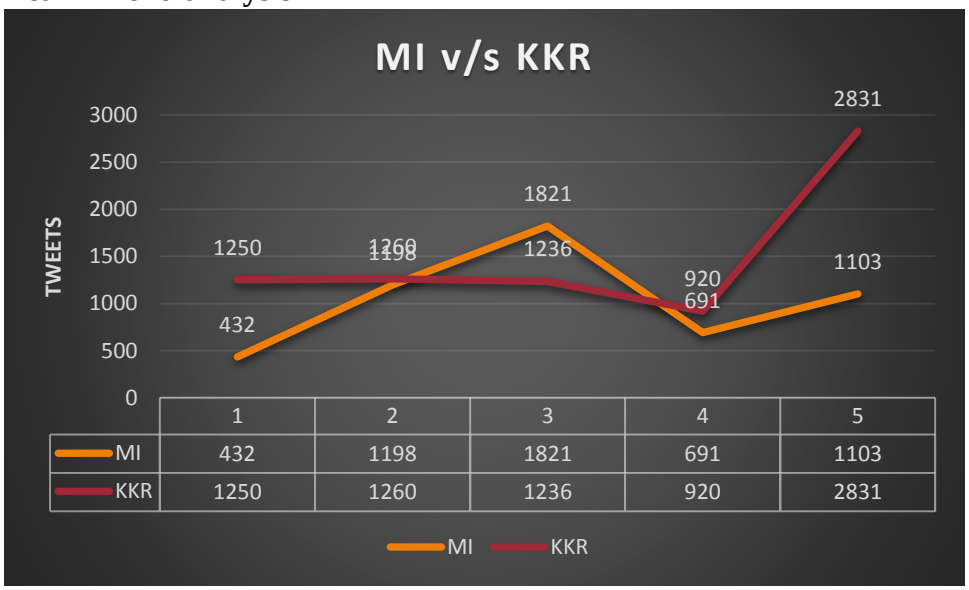
4	RCB batting: RCB loses early wickets and has a slow start to their chase	RCB / CSK – even
5	CSK wins the game	CSK

Match 2: MI v/s KKR

- **Team Twitter battle (based on Total number of relevant tweets)**



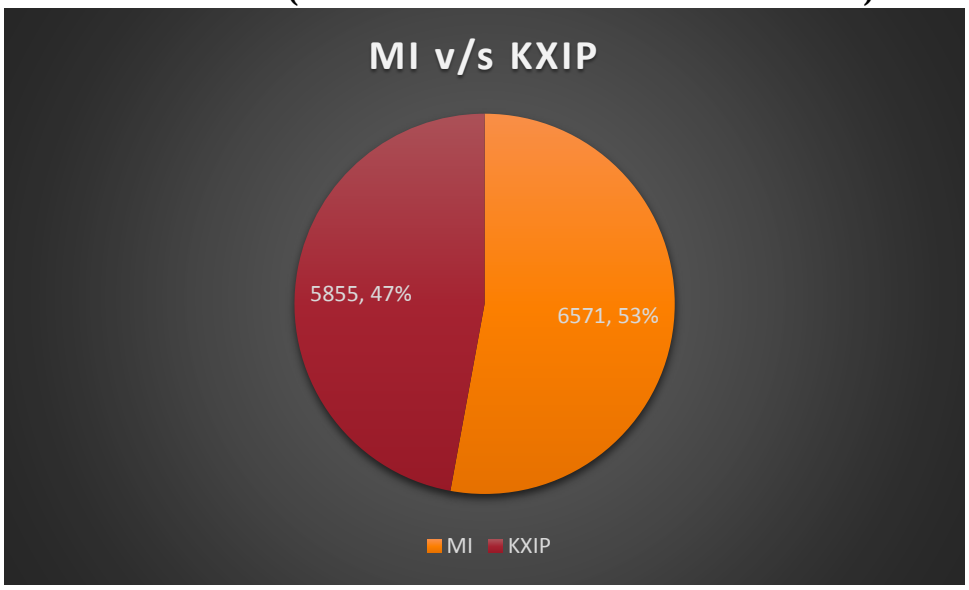
- **Team Trend analysis**



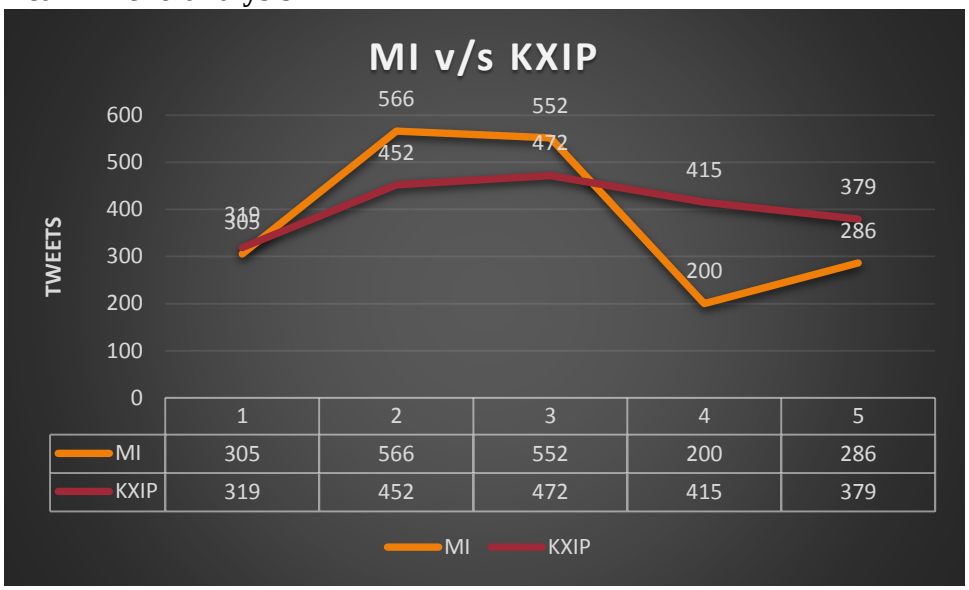
Time Interval	Event	Trending Team
1	1st match of IPL, KKR wins toss	KKR
2	MI batting: Early wickets for KKR	KKR
3	MI batting: Good final score for MI, Rohit scores an excellent hundred	MI
4	KKR batting: Excellent batting start for KKR	KKR
5	KKR wins the game	KKR

Match 3: MI v/s KXIP

- **Team Twitter battle (based on Total number of relevant tweets)**



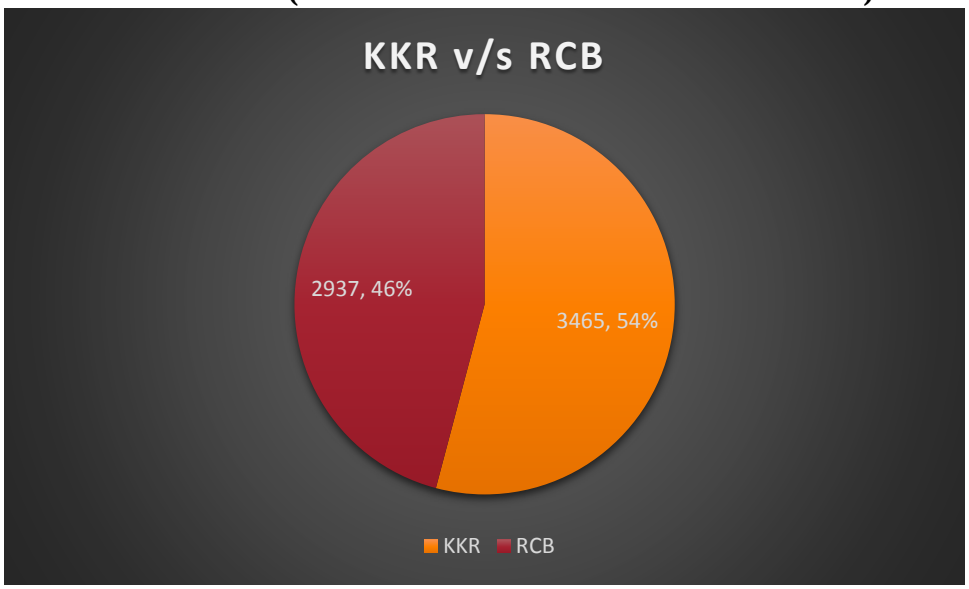
- **Team Trend analysis**



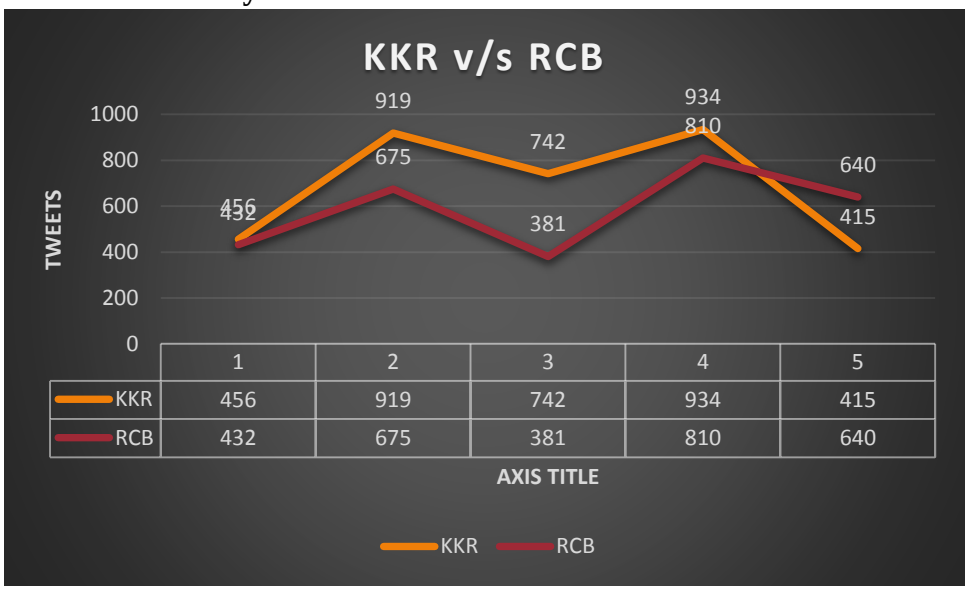
Time Interval	Event	Trending Team
1	MI won the toss and elected to bat first	MI / KXIP – even
2	MI batting: Excellent start with no loss of wickets	MI
3	MI batting: Good final score for MI. KXIP batting: Loses early wickets	MI
4	KXIP batting: Good partnership building for KXIP	KXIP
5	MI wins the game	KXIP

Match 4: KKR v/s RCB

- **Team Twitter battle (based on Total number of relevant tweets)**



- **Team Trend analysis**

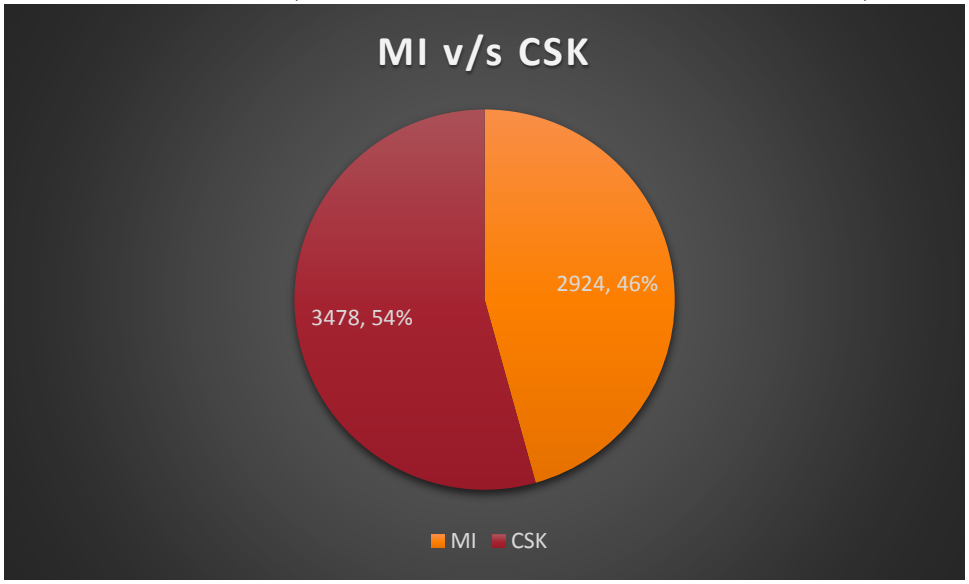


Time Interval	Event	Trending Team
1	KKR to bat, both are top teams in IPL	KKR / RCB – even
2	KKR batting: Excellent start with no loss of wickets	KKR
3	KKR batting: Andre Russell plays a short cameo	KKR

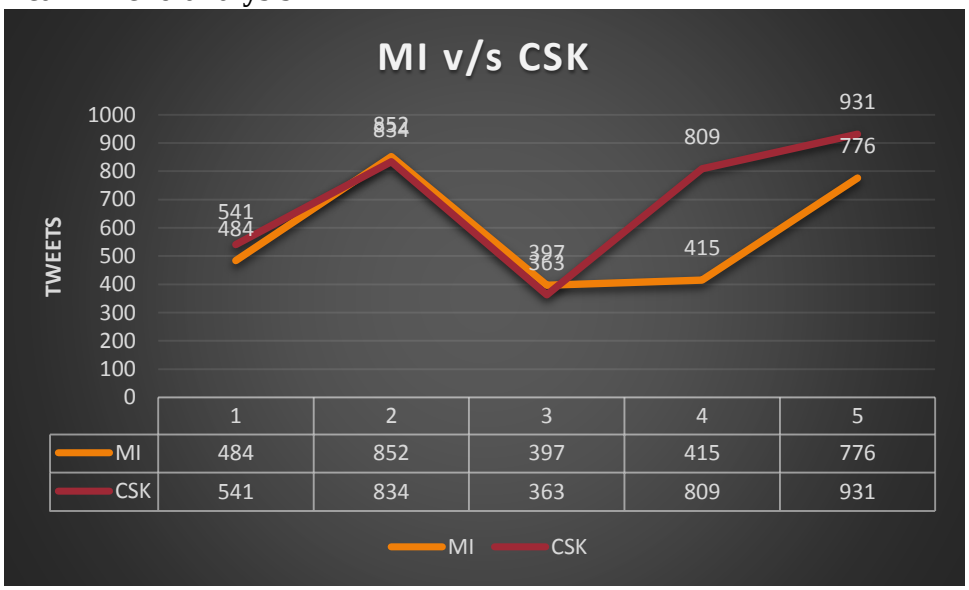
	RCB batting: Good start for RCB, but lose a couple of wickets	
4	RCB batting: Chris Gayle scores 100 and RCB lose some wickets on other end	RCB
5	RCB wins the game	RCB

Match 5: MI v/s CSK

- **Team Twitter battle (based on Total number of relevant tweets)**



- **Team Trend analysis**

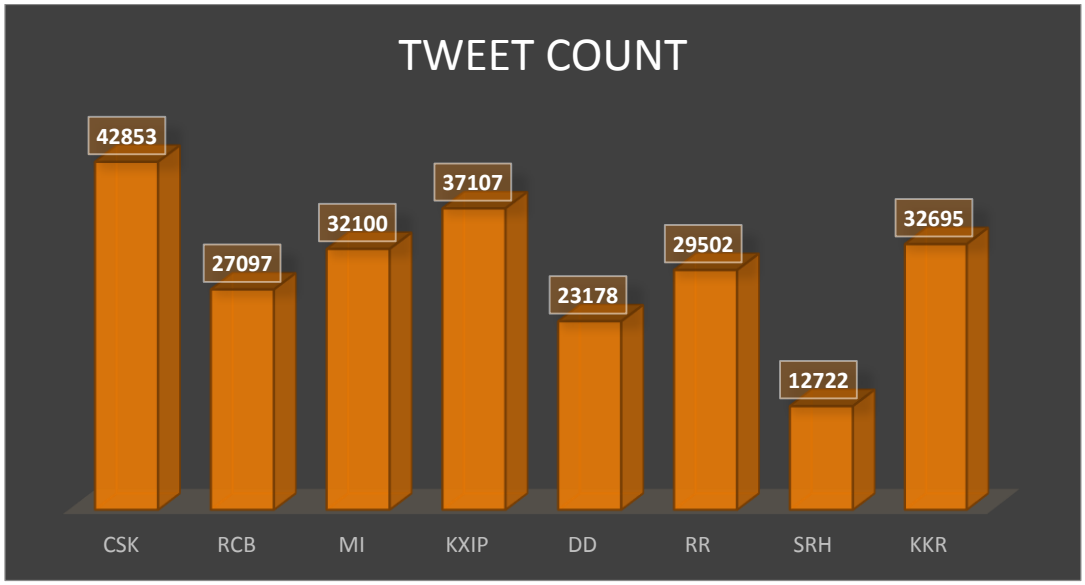


Time Interval	Event	Trending Team
1	MI to bat, both are top teams in IPL.	MI / CSK – even
2	MI batting: MI loses early wickets, captain scores 50.	MI / CSK – even
3	MI batting: Pollard scores a quick 50 and poses a high score CSK batting: Good start for CSK	MI / CSK – even
4	CSK batting: 2 players score a quick-fire 50.	CSK
5	CSK wins the game	CSK

Additionally, we based on the top tags / tokens generated for each of the above matches we also discovered the **Most Popular Player of the Match** and the **Most Popular Hashtags during the Match.**

Match	Most Popular Player of the Match	Most Popular Hashtags during the Match
CSK v/s RCB	Suresh Raina (#suresh, #raina) AB de Villiers (#ab, #villiers)	#csk #rcb #chennai #super
MI v/s KKR	Rohit Sharma (#rohit, #sharma) Gautam Gambhir (#gautam, #gambhir)	#kkr #mi #rohit #kolkata
MI v/s KXIP	Lendl Simmons (#simmons)	#mumbai #punjab #kxip #indians
KKR v/s RCB	Gautam Gambhir (#gambhir) Brendon McCullum (#brendon, #mccullum)	#kkr #rcb #kings #brendon
MI v/s CSK	Rohit Sharma (#sharma) Dwayne Smith (#smith)	#mi #csk #mumbai #smith

We further generated a single input file that contained all tweets for the week of April 15th, 2015 to April 22nd, 2015 and performed data aggregation / analysis on the same to generate descriptive statistics (count) as well as identify the most popular hashtags / teams / players of IPL 2015.



In addition, based on the mentions / counts, we identified the following players / hashtags as most prevalent on Twitter.

Most popular Players for the week:

- Rohit Sharma
- Chris Gayle
- Virat Kohli
- Gautam Gambhir

Most Popular Hashtags for the week:

- #mumbai
- #ipl
- #csk
- #mumbai
- #cricket
- #kkr

8 LESSONS LEARNED

This project gave us hands on experience of handling and parallel processing of huge amount of data. Data collection process introduced us to Twitter Streaming API. It was very interesting to gather and then aggregate live social networking data so as to extract interesting patterns and recent trends from it. We got exposure to work with prominent parallel data processing tool: Hadoop.

Apache Hadoop framework is gaining significant momentum from both industry and academia as the volume of data to analyze grow rapidly. This project helped us not only to gain knowledge about installation and configuration of Hadoop distributed file system but also map reduce programming model. At the end of analysis phase data visualization was performed using tools such as MS Excel, Tableau.

9 APPENDIX

9.1 READ ME DOCUMENTATION

Directory Structure:

- **Project:** TwitterIPL
  - **Package 1:** twitterData
    - **Files:** TwitterDriver.java, TwitterMapper.java, TwitterReducer.java, WebUtils.java, InvalidInputException.java
  - **Package 2:** iplAnalysis
    - **Files:** Analysis.java, GetMostKeywords.java, KeywordTwitterDriver.java, KeywordTwitterMapper.java, KeywordTwitterReducer.java, StopWord.java

Steps:

Step I:
Download the BigDataProject folder and extract the TwitterIPL folder.
Run TwitterDriver.java file with the following argument:
Input Argument Example: -q ipl -st 2015-04-07;21:30:00 -et 2015-04-13;15:30:00 -i 10
where,
-q: filter criteria (ipl in our case)
-st: startdate (format: yyyy-mm-dd;hh:mm:ss)
-et: enddate (format: yyyy-mm-dd;hh:mm:ss)
-i: interval (in milliseconds)

Step II:
Run KeywordTwitterDriver.java file with the following parameters
Input path Example: C://BigDataProject/input/CSKRCB-1
Output path Example: C://BigDataProject/output/Twitter_extract_1.txt)

Step III:
Run GetMostKeywords.java file and change the following parameters
File path br Example: C://BigDataProject/output/Twitter_extract_1.txt
File path bw Example: C://BigDataProject/output/Descorder_1.txt

Step IV:
Run StopWord.java file and change the following parameters
File path br1 Example: C://BigDataProject/stopwords.txt
File path br Example: C://BigDataProject/output/Descorder_1.txt
File path bw Example: C://BigDataProject/output/KeyWord_extract_1.txt

Step V:
Run Analysis.java file and change the following parameters
File path br1 Example: C://BigDataProject/keyword.txt
File path br Example- C://BigDataProject/output/KeyWord_extract_1.txt
File path bw Example – C://BigDataProject/output/Final_count_1.txt