Introduction

ALGORITHM/DESIGN

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. Using Big Data Technologies such as Hadoop, Hive, Spark etc, we simulate a 2016 IPL match using data from the previous editions of the IPL. Given the details of the teams along with the batting order and bowling order in the appropriate sequence, ball by ball prediction of the match can be done following which we can predict the winner of a given match.

1. The data collection process was divided into two parts. Player profiles of all cricketers were scraped from "www.espncricinfo.com" , cleansed & pre-processed into a format suitable for the follow-up steps. Ball by ball data was obtained from Cricsheet and using MapReduce on the Apache Hadoop framework, Player to Player scoring probabilities were generated. The data used spans IPL 2008 - 2015. The batsman-bowler combinations were generated in the mapper while the probabilities were computed in the Reducer. This resulted in obtaining combinations comprising the name of the batsman and corresponding bowler along with the probability of that batsman hitting a 0,1,2,3,4,6 against that bowler as well as the probability of not getting out.

2. In addition, player profiles were used for clustering the batsmen and bowlers to achieve a certain level of similarity and generalisation. Clustering was done using the K-Means algorithm on Spark MLLib. The criteria chosen for clustering batsman included Average Runs Scored and Average Strike Rate whereas the ones for bowlers included Average Economy Rate and Average Strike Rate. Number of clusters was set to 10 while the number of iterations was set to 200 to obtain an optimal split resulting in 10 clusters of batsmen and 10 clusters of bowlers.

3. The clustered data obtained was loaded into Hive and a Hive query was run to obtain batsman cluster vs bowler cluster probabilities resulting in 10x10 = 100 combinations. The output format comprises of Batsman Cluster Number, Bowler Cluster Number, Cluster to Cluster probabilities of scoring 0, 1, 2, 3, 4, 6 along with the probability of not getting out.

4. For simulating the match, we use the player-to-player probabilities from Step 1 along with the cluster-to-cluster probabilities from Step 3, when the player-to-player probabilities are not available. We perform a Hive Query to get the cluster numbers of all batsmen and bowlers in the current match. And for each batsman vs bowler instance, we predict the runs scored per ball and the probability of the batsman getting out against that bowler. To predict the runs scored for each ball bowled, we generate a random number between 0 and 1 and compare it with the cumulative probability distribution to simulate the runs scored for every ball. A batsman is initially assigned a value which is the probability of him not getting out against the bowler he is facing. Given a specific batsman vs bowler instance, we fetch the probability of the batsman not getting out against the bowler he is facing from the set of probabilities and multiply it with the current value of the batsman. This is done for every ball faced by the batsman. Once this value reduces to less than a threshold of 0.5, the batsman is declared out. These computations are done for all batsmen. The final score for the team is summed up and the target set for the opposing team. The same process is repeated for the other team also.

EXPERIMENTAL RESULTS

1. **Player-Player probabilities** were obtained in the following format -
   Eg: CH Gayle D Steyn 0.4 0.2 0.1 0 0.2 0.1 0.92
   Batsman Name Bowler Name P(0),P(1),P(2),P(3),P(4),P(6),P(Not Out)

```
A Ashish Reddy,SW Tait  0.25,0.0,0.25,0.0,0.5,0.0,1.0
A Ashish Reddy,Sandeep Sharma   0.0,0.3333333333333333,0.0,0.0,0.0,0.6666666666666666,1.0
A Ashish Reddy,TG Southee       0.25,0.25,0.25,0.0,0.0,0.25,1.0
A Ashish Reddy,UT Yadav 0.0,0.4,0.2,0.0,0.4,0.0,1.0
A Chandila,R Rampaul    0.5,0.5,0.0,0.0,0.0,0.0,1.0
A Chandila,R Vinay Kumar        0.5,0.5,0.0,0.0,0.0,0.0,1.0
A Chandila,RP Singh     0.0,1.0,0.0,0.0,0.0,0.0,1.0
A Chandila,V Pratap Singh       1.0,0.0,0.0,0.0,0.0,0.0,1.0
A Chopra,DP Vijaykumar  0.8,0.2,0.0,0.0,0.0,0.0,1.0
A Chopra,DW Steyn       0.5,0.25,0.0,0.0,0.0,0.0,0.75
A Chopra,GD McGrath     0.5,0.5,0.0,0.0,0.0,0.0,1.0
A Chopra,Harmeet Singh  0.875,0.0,0.0,0.0,0.125,0.0,1.0
A Chopra,JR Hopes       0.0,0.5,0.0,0.0,0.0,0.0,0.5
```

2. 10 batting clusters as well as 10 bowling clusters -
   Batting Cluster: Batsman Name, Avg. Score, Avg. Strike Rate, Cluster No.
   Bowling Cluster: Bowler Name, Avg. Economy, Avg. Strike Rate, Cluster No.

A Bali	24.0575	113.43	6
A Baloria	17.33	98.11	5
A Bhatt	12.8	98.46	5
A Bhattarai	0.66666667	22.22	6
A Bonora	21.8	99.175	7
A Bramble	8.75	107.69	2
A Carter	0	25	3
A Chandila	7.915	78.7875	2
A Chopra	21.3675	85.1025	6
A Choudhary	6.33333333	53.17	7
A Dananjaya	6.1	23.18	1

3. **Cluster-Cluster probabilities** (10x10 = 100 combinations) -
   Eg: 1 5 0.4 0.2 0.1 0 0.2 0.1 0.92
   Batsman Cluster Bowler Cluster P(0),P(1),P(2),P(3),P(4),P(6),P(Not Out)

4. **The IPL match simulation :**

```
                                    IPL MATCH SIMULATION
-------------------------------------------------------------------------------------------------------------------------
                                  BIG DATA (UE14CS314) PROJECT
-------------------------------------------------------------------------------------------------------------------------

                              HEMKESH V KUMAR         01FB14ECS083
                              POOJA BALUSANI          01FB14ECS145
                              PRATIK CHATTERJEE       01FB14ECS161
                              PRIYANSHA PATHAK        01FB14ECS168

-------------------------------------------------------------------------------------------------------------------------

0.0     Batsman: LMP Simmons            Bowler: A Nehra         Ball Outcome:   2
Score:  2/0
0.1     Batsman: LMP Simmons            Bowler: A Nehra         Ball Outcome:   1
Score:  3/0
0.2     Batsman: PA Patel               Bowler: A Nehra         Ball Outcome:   4
Score:  7/0
0.3     Batsman: PA Patel               Bowler: A Nehra         Ball Outcome:   0
Score:  7/0
0.4     Batsman: PA Patel               Bowler: A Nehra         Ball Outcome:   0
Score:  7/0
0.5     Batsman: PA Patel               Bowler: A Nehra         Ball Outcome:   1
Score:  8/0
```

4a. First Innings Result :

```
Score:  200/4
19.0    Batsman: AT Rayudu              Bowler: DJ Bravo                Ball Outcome:   0
Score:  200/4
19.1    Batsman: AT Rayudu              Bowler: DJ Bravo                Ball Outcome:   1
Score:  201/4
19.2    Batsman: HH Pandya              Bowler: DJ Bravo                Ball Outcome:   1
Score:  202/4
19.3    Batsman: AT Rayudu              Bowler: DJ Bravo                Ball Outcome:   6
Score:  208/4
19.4    Batsman: AT Rayudu              Bowler: DJ Bravo                Ball Outcome:   1
Score:  209/4
19.5    Batsman: HH Pandya              Bowler: DJ Bravo                Ball Outcome:   1
Score:  210/4

-------------------------------------------------------------------------------------------------------------------------
                                FIRST INNINGS SCORE: 210/4 in 20.0 overs
-------------------------------------------------------------------------------------------------------------------------
```

4b. Second Innings Result:

```
Score:  153/2
19.1    Batsman: MS Dhoni               Bowler: R Vinay Kumar           Ball Outcome:   0
Score:  153/2
19.2    Batsman: MS Dhoni               Bowler: R Vinay Kumar           Ball Outcome:   1
Score:  154/2
19.3    Batsman: SK Raina               Bowler: R Vinay Kumar           Ball Outcome:   6
Score:  160/2
19.4    Batsman: SK Raina               Bowler: R Vinay Kumar           Ball Outcome:   0
Score:  160/2
19.5    Batsman: SK Raina               Bowler: R Vinay Kumar           Ball Outcome:   4
Score:  164/2

------------------------------------------------------------------------------------------------------------------------
                                        SECOND INNINGS SCORE: 164/2 in 20.0 overs
------------------------------------------------------------------------------------------------------------------------

MATCH RESULT:   TEAM 1 WON BY 46 RUNS
```

The Match Result is Team 1 won by 46 runs as shown in the output above.

FUTURE ENHANCEMENTS

1. We would like to improve the accuracy of clustering the batsmen and bowlers for better predictive analysis of matches by exploring better models and algorithms for clustering.
2. We would also like to integrate Hive commands into Hadoop.
3. Another potential area for improvement would be in achieving the right decision boundary while deciding whether to pick batsman-batsman probability or cluster-cluster probability for a particular batsman-bowler combination. The goal is to fine tune the model for a better overall accuracy.
4. Taking into consideration the extras such as wides and LBW's.

REFERENCES / RELATED WORK

1. Hadoop: The Definitive Guide – Tom White. This book gives a spectacular introduction into the world of Big Data
2. Big Data Analytics Beyond Hadoop: Real-Time Applications with Storm, Spark, and More Hadoop Alternatives, Vijay Srinivasa Agneeswaran
3. Mining of Massive Datasets, Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman
4. Spark: cluster computing with working sets ,Zaharia M, Chowdhury M, Franklin MJ, Shekhar S, Stoica I.. HotCloud
5. http://spark.apache.org/docs/latest/mllib-clustering.html#k-means
6. https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
7. K-means algorithm is best suitable in this case for clustering. Spark MLlib (Stanford edu notes on Spark MLLib) technology was used as a platform to perform clustering
8. The paper “Hive: a warehousing solution over a MapReduce framework.” was very insightful and a strong guide in understanding Hive and applying this technology.
9. http://www.folkstalk.com/2011/11/conditional-functions-in-hive.html?m=1
10. Player statistics data is scraped from : http://stats.espncricinfo.com/ipl2010/engine/records/averages/batting.html?id=5319;type=tournament