IPL Dataset Analysis

The dataset contains IPL data. This data can be used to analyse some key statistics of players, teams and match venues.

Understanding Data

The data format is '|' separated values. There are 5800 observations.

Schema Description

match_id,inning,batting_team,bowling_team,over,ball,batsman,non_striker,bowler,is_super_over,wide_runs,bye_runs,legbye_runs,noball_runs,penalty_runs,batsman_runs,extra_runs,total_runs

Problem Statements

1) Load the IPL file from HDFS and make it an RDD & Convert the RDD into dataframe.

```
$spark-shell

scala> val sqlContext = new org.apache.spark.sql.SQLContext(sc)

scala> import sqlContext.implicits._

scala> case class IplMatch(match_id: Int,innings: Int,batting_team: String,bowling_team: String,over: Int,ball: Int,batsman: String,non_striker: String,bowler: String,is_super_over: Int,wide_runs: Int,bye_runs: Int,legbye_runs: Int,noball_runs: Int,penalty_runs: Int,batsman_runs: Int,extra_runs: Int,total_runs: Int)

scala> val ipl_data = sc.textFile("file:///home/cloudera/khasimbabu/Spark_Excercise/ipl.txt").map(_.split("\\|")).map(i => IplMatch(i(0).trim.toInt,i(1).trim.toInt,i(2),i(3),i(4).trim.toInt,i(5).trim.toInt,i(6),i(7),i(8),i(9).trim.toInt,i(10).trim.toInt,i(11).trim.toInt,i(12).trim.toInt,i(13).trim.toInt,i(14).trim.toInt,i(15).trim.toInt,i(16).trim.toInt,i(17).trim.toInt)).toDF()
```

2) Register the temp table for the dataframe.

```
scala> ipl_data.registerTempTable("iplmatchtable")

scala> val allrecords = sqlContext.sql("select * from iplmatchtable")

scala> allrecords.show()

scala> allrecords.printSchema()
```

4) How many runs did RA Jadeja score in the match?

```
scala> val allrecords = sqlContext.sql("select batsman,sum(batsman_runs) as total_run from iplmatchtable where batsman='RA Jadeja' group by batsman")

scala> allrecords.show()

scala> val allrecords = sqlContext.sql("select batsman,sum(batsman_runs) as total_run from iplmatchtable group by batsman")

scala> allrecords.show()
```

5) How many wides did A Nehra bowled against Chennai Super Kings in match number 8?

```
scala> val allrecords = sqlContext.sql("select bowler,count(wide_runs) as no_of_wide_balls from
iplmatchtable where bowler='A Nehra' and batting_team='Chennai Super Kings' and wide_runs>0
group by bowler")
```

6) What is the final score of Mumbai Indians?

```
scala> val allrecords = sqlContext.sql("select batting_team,sum(total_runs) as final_score from
iplmatchtable where batting_team='Mumbai Indians' group by batting_team")
```

7) Highest run scored by batsman from Royal Challengers Bangalore in a IPL series

```
scala> val allrecords = sqlContext.sql("select batting_team,batsman,max(batsman_runs) as
high_score from iplmatchtable where batting_team='Royal Challengers Bangalore' group by
batting_team,batsman order by high_score desc limit 10")
```

8) Fastest 50 by a batsman in a IPL series

```
scala> val allrecords = sqlContext.sql("select batsman, batsman_runs, ball from iplmatchtable where
batsman_runs>49 order by ball desc")
```

```
scala> allrecords.show()
```