**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()